Inférence bayésienne pour la détection des activités de la vie quotidienne pour faciliter le maintien à domicile des personnes âgées

Yassine El Khadiri

École doctorale IAEM Lorraine

Inférence bayésienne pour la détection des activités de la vie quotidiennes pour faciliter le maintien à domicile des personnes âgées

THÈSE

présentée et soutenue publiquement le 31 Août 2021

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention informatique)

par

Yassine El Khadiri

Composition du jury

Président du jury :	Perret-Guillaume Christine	Professeur des universités Professeur hospitalier	Université de Lorraine CHRU Nancy Brabois
Rapporteurs :	Fleury Anthony Hewson David	Maître de conférences (HDR) Professeur	IMT Lille Douai University of Bedfordshire
Examinateur :	Lago Paula A.	Lecturer	Universidad de Los Andes
Directeur de thèse :	Charpillet François	Directeur de recherche	Inria Nancy
Invité :	Rose Cédric	Ingénieur	Diatelic Pharmagest

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Mis en page avec la classe thesul.

Remerciements

Cette thèse[1] ne serait jamais arrivée à son terme sans l'aide et l'encouragement de nombreuses personnes.

Je voudrais tout d'abord grandement remercier mon directeur de thèse, François Charpillet, Directeur de recherche au centre de l'Inria Nancy, pour son encadrement, ses encouragements, son support pendant les moments les plus difficiles et le partage des moments les plus agréables et joyeux. La thèse n'a pu être accomplie qu'à travers les efforts colossaux qu'il a portés à travers son aide, encore un grand et chaleureux merci.

Je remercie Cédric Rose, Gabriel Corona et tous mes collègues de Diatelic pour leurs encadrement, support et aide. Cette thèse est bien sûr le fruit de la collaboration entre Diatelic et Inria. J'ai pu donc acquérir pendant les trois ans passés à Diatelic une bonne expérience du monde industriel et de l'entreprise.

Je tiens à chaleureusement remercier mes amis et collègues à l'Inria pour leur immense support et aide moral lors des moments les plus bas sans lequel je n'aurai pu continuer, mais aussi pour les meilleurs moments partagés ensembles. Lucien Renaud, Jimmy Etienne, Nicolas Gauville, Adrien Malaisé, Oriane Dermy, Dorian Goepp... Désolé, mais ne serait-ce que la moitié de la liste en ferait déborder cette page.

Mes remerciements à Fleury Anthony, Maître de conférence à l'IMT Lille Douai et Hewson David, Professeur à l'University of Bedfordshire de l'honneur qu'ils m'ont fait en acceptant d'être rapporteurs de cette thèse. Ma gratitude à Lago Paula, Lecturer à l'Universidad de Los Andes et Perret-Guillaume Christine, Professeur des universités de l'Université de Lorraine et Professeur hospitalier au CHRU Nancy Brabois qui ont bien voulu être examinateurs.

Et enfin un grand merci à mes parents et ma famille. Mon père Abdelhafed Elkhadiri, Directeur du département de mathématiques de l'Université Ibn Tofail de Kénitra pour m'avoir encouragé à emprunter ce chemin et découvrir le monde de la recherche. Ma mère Souad Dhimni, PDG du cabinet d'étude et conseil Mast Engineering et Professeur à l'Université Ibn Tofail pour son support moral et ses encouragements.

1. This work was partially achieved through access to equipments funded by the CPER IT2MP (Contrat Plan État Région, Innovations Technologiques, Modélisation Médecine Personnalisée) and FEDER (Fonds européen de développement régional)

Sommaire

Table des figures

1

Introduction

Sommaire

1.1 Préambule

Cette thèse Cifre s'inscrit dans une collaboration de longue date entre la société Diatelic, l'Inria et le Loria. Diatélic est en effet une société Lorraine créée en 2002 par des chercheurs de l'Inria Nancy Grand Est et du Loria ainsi que par des médecins de l'ALTIR (Association Lorraine pour le Traitement de l'Insuffisance Rénale) et un médecin intervenant en tant que conseiller scientifique. Aujourd'hui, cette entreprise est une filiale du Groupe Pharmagest qui est un des grands groupes français spécialistes des infrastructures logicielles pour les pharmacies. Il y a une dizaine d'années, convaincue que les pharmaciens allaient jouer un rôle de plus en plus important dans le suivi des pathologies chroniques, la société Pharmagest a créé en son sein, un secteur e-santé destiné à répondre aux défis sociétaux que posent le vieillissement de la population ou la désertification médicale. C'est dans cet objectif que Pharmagest fait alors l'acquisition de Diatelic, société spécialisée dans l'intelligence artificielle appliquée à la télésurveillance et au suivi de l'observance des patients. La société Diatelic s'est d'abord fait connaître par ses solutions innovantes de télémédecine pour le traitement de l'insuffisance rénale et le suivi de patient sous dialyse à domicile [48]. Il s'agit d'une des premières entreprises en France à proposer un service de télémédecine opérationnel dès 2001. Plus récemment, la société Diatelic s'est orientée vers le développement de solutions pour le maintien à domicile des personnes âgées en perte

d'autonomie, afin de permettre aux personnes âgées de rester chez elles plus longtemps qu'elles ne peuvent le faire actuellement.

De manière concomitante, l'équipe Maia puis aujourd'hui l'équipe Larsen communes à l'Inria et au Loria, ont développé et développent un projet scientifique autour de la robotique, de l'intelligence ambiante et de l'intelligence artificielle dont une des applications est la conception de solutions qui facilitent le maintien à domicile des personnes en situation de fragilité, en perte d'autonomie ou dépendantes. De nombreux résultats ont été obtenus en particulier sur l'analyse du mouvement humain par caméras de profondeur, ou dalles sensibles [25; 34; 28; 4; 31; 83; 42; 84; 14]. Ces résultats résultent de soutiens divers en particulier de l'ANR (projets Depic, Parachute, Predica). Ensuite, la région Lorraine (aujourd'hui la région Grand est) ayant inscrit dans son plan stratégique la Silver économie parmi l'un des axes qu'elle souhaitait soutenir, elle a fortement financé différentes opérations autour des Contrats de plan État-Région (CPERs) et de projets d'Appel à Manifestation d'Intérêt (AMI) avec Pharmagest. L'Inria a financé d'autre part un projet d'envergure National PAL (personal Assisted Living), ce qui fait qu'aujourd'hui l'équipe Larsen a pu développer une activité de recherche soutenue sur l'assistance à la personne. Les solutions s'appuient sur différentes technologies à l'interface entre intelligence artificielle, robotique et domotique. L'équipe dispose d'une plate-forme de recherche : *l'Habitat intelligent pour la santé*. Il s'agit d'un lieu dédié à l'expérimentation comprenant 4 pièces meublées, 700 capteurs (caméras de profondeur, dalles intelligentes, capteurs de présence, microphones, etc.) et des robots mobiles. Cette plate-forme d'expérimentation offre un cadre idéal pour inventer les technologies de demain en matière d'aide à l'autonomie et de maintien à domicile des personnes fragiles ou dépendantes.

1.2 Contexte

1.2.1 Démographie et population senior

Le vieillissement de la population confronte les sociétés contemporaines à une transformation démographique sans précédent.

D'après la base de données de la Banque mondiale, le ratio de dépendance des personnes âgées, qui est le rapport entre le nombre de personnes âgées de plus de 64 ans et la population en âge de travailler est passé de 8,6% en 1960 à 13,94% en 2019 [12] sur la population mondiale, de 19% en 1960 à 33% en France. Selon le bureau du recensement des États-Unis par exemple, environ 29% des plus de 65 ans vivent seuls [19] avec un coût médian national d'une chambre semi-privée en maison de retraite qui devrait croître de 243% entre 2017 et 2047 [38].

En France, selon les projections de l'INSEE (Insee Première N^o1089 - juillet 2006), en 2050, près d'un habitant sur trois aura plus de 60 ans, contre un sur cinq en 2005. Ce ratio aura presque doublé en 45 ans. Une telle évolution pose, au-delà du problème du financement des retraites et de la dépendance, un allongement de la durée de vie qui ne fera qu'accroître le nombre de personnes souffrant soit d'une perte d'autonomie soit de maladies chroniques. Cela va fortement impacter l'organisation de notre système de santé. Déjà aujourd'hui l'impact des maladies chroniques et de la dépendance est considérable :

15 à 20 millions de personnes en France sont concernées. Le coût total de ces maladies est très élevé et augmente avec l'accroissement de la demande de soins. La perte d'autonomie touche de nombreuses familles nécessitant le placement en institution des personnes concernées. Par ailleurs, la démographie médicale et paramédicale diminue quantitativement et sa répartition sur le territoire n'est pas uniforme. En d'autres termes, l'offre de soin tend à se raréfier, et ce de manière non homogène, alors que la demande va considérablement augmenter dans les années qui viennent. En effet **(author?)** [8] constatent dans leur étude une baisse de l'offre globale de soins d'une ampleur plus importante que celle des effectifs qui, mise en parallèle avec le vieillissement de la population, fait que l'offre médicale devrait croître moins vite que la demande, au cours des dix prochaines années, c.-à-d. à l'horizon 2027. Pour faire face à cette situation, les progrès des technologies de l'information et de la communication permettent le développement de nouvelles solutions de prise en charge des personnes qui souffrent de pathologies chroniques graves ou d'un vieillissement nécessitant un suivi. Ces nouvelles solutions vont radicalement changer l'organisation des soins en particulier parce qu'elles permettent d'augmenter de manière significative la prise en charge à domicile des personnes. Cette dernière est source de confort pour les patients et est souvent beaucoup plus pertinente comme l'a démontré la société Diatelic dans les domaines de l'insuffisance rénale. Au-delà de l'insuffisance rénale, sont aussi concernées les maladies chroniques comme l'insuffisance cardiaque, l'insuffisance respiratoire, le diabète, l'hypertension et les problèmes liés au sommeil. Par ailleurs, les personnes âgées vivant seules à domicile sont sujettes à des risques du type chutes, baisse d'activité ou comportements pathologiques divers. La possibilité de détecter ces problèmes et d'intervenir rapidement grâce à des dispositifs et à des services innovants est prometteuse dans l'optique de permettre aux personnes âgées ou en situation de handicap de vivre plus facilement et plus longtemps dans leur logement en profitant de leur environnement social.

1.2.2 Habitat intelligent pour le maintien à domicile

Une des réponses à cette problématique sociétale du vieillissement de la population est le développement de technologies qui facilitent le maintien à domicile des personnes âgées.

L'état de l'art du domaine regorge de projets qui s'intéressent à cette question. Parmi eux, beaucoup consistent à développer des systèmes de télésurveillance à domicile, on parle aussi souvent dans la littérature d'habitats intelligents ou d'habitat connectés. L'objectif dans ce cadre est d'équiper un logement de capteurs afin de détecter, voire de prévenir l'occurrence de situations inquiétantes ou critiques et d'évaluer l'état physique et la fragilité des personnes suivies. Différents types de capteurs peuvent être installés à domicile : des capteurs environnementaux ou des capteurs portés par la personne. Les premiers sont souvent privilégiés, car ils sont intégrés au lieu de vie et sont autonomes, les personnes ne sont donc pas sollicitées pour les recharger ou à les porter. Les seconds peuvent aussi être envisagés : parmi ceux-ci il y a par exemple les montres connectées qui peuvent relever l'activité des personnes équipées. Un avantage de ces capteurs portés est qu'ils sont forcément associés à une personne donnée ce qui n'est pas le cas des capteurs environnementaux qui eux posent une problématique lorsque plusieurs locataires coexistent au

sein du même logement. Lorsque l'on veut équiper des logements à large échelle, se pose également la question du coût de l'installation qui peut fortement varier en fonction du type de capteur choisi, et leur nombre.

D'autres projets s'intéressent au rôle que pourraient jouer les robots dans la téléassistance et l'interaction, ainsi que l'acceptabilité des personnes vis-à-vis de ce type d'outils dans leur environnement de vie.

1.2.3 Apprentissage automatique et suivi des habitudes de vie

L'état de l'art de la recherche en assistance à l'autonomie à domicile repose en majeure partie sur des technologies dites d'«intelligence ambiante». Il s'agit d'un paradigme où les personnes sont assistées et guidées dans leurs activités quotidiennes par des systèmes capables de suivi, d'anticipation et de conseil tout en étant non-invasifs. Ces services d'assistance à l'autonomie à domicile englobent différentes technologies comprenant capteurs, effecteurs, boîtiers de communication et interfaces. Ils servent aussi différents tiers d'utilisateurs dont les demandes et attentes seront tout aussi variées.

1.3 Objectif de la thèse

La société Diatelic développe actuellement une offre de télésurveillance à destination des personnes âgées et des personnes fragiles. Cette offre s'articule autour d'une Box connectée et un ensemble de capteurs d'activité. Autour de cette offre matérielle, Diatélic propose un ensemble de services d'analyse et d'interprétation des données. C'est ici que se positionne la thèse dont la contribution est la recherche et développement d'algorithmes capables d'apprendre les habitudes de vie pour en déduire des incohérences qui peuvent potentiellement révéler l'apparition d'une fragilité. La solution doit donc assurer un suivi personnalisé et améliorer la coordination des intervenants, que ce soit à leur domicile ou en maison de retraite médicalisée ou non. Elle doit permettre également de maintenir le lien avec la famille et les aidants grâce à sa connexion à un réseau social dédié. Ce travail rentre dans le cadre du projet « 36 mois de plus à domicile » développé par le Groupe Pharmagest avec le soutien du Conseil Régional Grand Est et du FEDER (Fonds Européens de Développement Régional). L'objectif est d'explorer et de développer des solutions pour le suivi de résidents seniors dans leurs habitats à travers des données rapportées par des capteurs discrets, peu coûteux et non invasifs.

Plus précisément, cette thèse a pour objectif de proposer des solutions pour pouvoir évaluer le niveau d'autonomie d'une personne fragile en relevant ce qu'on appelle les Activités de la Vie Quotidienne (AVQ) (par exemple le sommeil, la visite de salle de bain, la préparation de repas, etc.). Cette évaluation se fera à travers un bilan automatique journalier comprenant la fréquence, durée et moment de la journée des activités relevées. Ce bilan sera envoyé au personnel soignant et au surveillant de la personne. Elle doit aussi pouvoir se faire sans aucune nécessité de mobiliser ou d'interférer en quoique ce soit avec la vie habituelle que la personne mènerait sans un tel dispositif de suivi.

L'automatisation du relevé des AVQ est une problématique de recherche qui mobilise de nombreux chercheurs à travers le monde. La question qui se pose est de savoir comment

reconnaître les activités de la vie quotidienne à partir des données issues de capteurs installés à domicile. Une difficulté majeure est que, souvent pour des raisons de coût, les capteurs sont en nombres limités et relativement peu informant (capteurs binaires le plus souvent).

Ces solutions se concrétisent principalement en différents modèles et algorithmes d'inférence des activités de la vie quotidienne en partant des données capteurs. Le cœur de la thèse est de démontrer la faisabilité de la mise en place d'algorithmes d'apprentissage non supervisé pour le suivi d'activité chez les personnes équipées. Et la visée est le développement d'une solution commerciale clé en main capable de fonctionner immédiatement après son installation.

Pour la recherche et le développement de ces solutions, nous nous appuyons sur des bases de données académiques ainsi que des données réelles recueillies sur le terrain. L'avantage est que nous avons un retour immédiat des besoins et remarques du personnel soignant au fur et à mesure de l'avancée des travaux.

1.3.1 Positionnement

Le cadre des travaux se situe au sein du projet *«36 mois de plus»* de Diatélic. Le but est de proposer une solution de télésurveillance facile à mettre en place et de maintenir avec des rapports immédiats après installation. Ceci impose certaines restrictions sur la nature et les propriétés des données collectées par les capteurs (nature, précision et fréquence des données). Ainsi, les algorithmes que nous proposons ici utilisent principalement les données événementielles issues de capteurs de mouvement environnementaux.

1.3.2 Approche

En raison de la pauvreté des données dont nous disposons en amont pour effectuer des inférences, et des contraintes imposées par la spécification de la solution de télésurveillance, la classe d'algorithmes que nous avons explorés devait y être tout aussi adaptée :

Non supervisé : Nous aurons tendance à préférer les algorithmes non supervisés qui n'ont pas impérativement besoin d'une période d'entrainement avant de pouvoir évaluer et retourner des résultats.

Modèles : Nous aurons ainsi besoin d'adopter une approche principalement articulée sur des modèles.

Activités : Les propriétés temporelles des capteurs imposent une limite sur les types d'activités possiblement inférées.

1.3.3 Organisation du mémoire

Dans ce travail de thèse, nous commençons par présenter un état de l'art des différentes techniques et technologies utilisées dans la littérature dans le cadre du suivi d'activité : La corrélation entre l'état de santé des personnes, leur degré d'autonomie et leurs activités de la vie quotidienne. Les outils de détection de ces activités, c.-à-d. les différents types de capteurs et données que l'on peut récupérer sur la personne suivie. Les différents

algorithmes développés et employés pour inférer et prédire ces activités en se basant sur ces mêmes données. Et enfin une revue des outils et plates-formes logiciels les plus répandus dans ce milieu.

Étant donné qu'il s'agit d'une thèse Cifre, le but est d'intégrer et déployer les contributions recherchées et développées dans un cadre industriel bien précis. En considérant tout un lot de plates-formes et de services et surtout des contraintes particulières qui vont orienter les sujets abordés. Nous allons donc brièvement présenter les différents services fournis aux patients, leurs structures, les capteurs utilisés avec leur nomenclature de données et les connaissances *a priori*.

S'en suivront nos deux contributions. La première concerne l'analyse et l'application de méthodes de segmentation et de classification de séries temporelles pour l'inférence d'AVQ avec un exemple sur la détection des phases de sommeil d'un résident. Nous abordons les choix et les compromis sur les approches étudiées ainsi que les potentielles évolutions et améliorations qu'offre chaque approche.

La deuxième présente une méthode de présentation et de visualisation des habitudes de vies d'un patient permettant à un expert de formuler un avis informé sur l'état de son rythme de vie. Nous y présentons la technique élaborée ainsi que plusieurs cas d'études ou nous montrons les lectures possibles offertes par cette technique.

L'apport de ce travail de thèse se situe à la fois sur le plan méthodologique et applicatif. Sur le plan méthodologique, nous proposons différentes techniques d'inférence explorées en se basant sur la structure et les données disponibles sur le terrain. Sur le plan applicatif, tout le travail effectué a été continuellement intégré à l'infrastructure existante et déployée sur le terrain avec tout le travail d'ingénierie qui en incombe.

2

État de l'art

Sommaire

2.1 Télésurveillance médicale à domicile

La télésurveillance médicale à domicile est définie comme un acte de télémédecine en France par le décret n° 2010-1229 du 19 octobre 2010 [68] dont l'objet est de permettre à un professionnel médical d'interpréter à distance les données nécessaires au suivi médical d'un résident et, le cas échéant, de prendre des décisions relatives à la prise en charge de celui-ci. L'enregistrement et la transmission des données peuvent être automatisés ou réalisés par le résident lui-même ou par un professionnel de santé.

Les outils offrant un tel service auront ainsi la vocation à communiquer le plus clairement et efficacement que possible aux professionnels de santé par le biais de rapports sur les habitudes de vie (AVQ) et l'état de santé de leurs résidents, pour qu'ils puissent prendre des décisions et agir au plus vite quand cela est nécessaire. Sachant que les AVQs sont hautement corrélées avec l'état physique [44] et cognitif [59], les équipements de télésurveillance médicale grand public et générique se concentrent principalement sur la détection et le suivi de ces AVQs à l'aide de capteurs et d'algorithmes d'apprentissage.

2.2 Détection des activités de la vie quotidienne

Ainsi, nous nous intéresserons à la détection et le suivi d'activités de la vie quotidienne (abrégé AVQ) en tant qu'indicateur de l'état de santé du résident placé sous télésurveillance à domicile.

Toute activité physique réalisée par une personne émet une multitude de signaux dans l'environnement. Ces signaux peuvent donner des indices sur l'activité physique menée. Par exemple, l'information de présence dans une pièce peut être détectée de différentes manières. En effet, la manière la plus directe serait d'équiper la pièce avec une caméra qui détecterait la présence d'une silhouette humaine ou non[32]. Une autre serait d'utiliser un capteur de mouvement et/ou des capteurs de contact sur les portes d'accès à la pièce[62]. Ou encore de trianguler sa position spatiale en localisant un appareil émetteur porté par la personne à l'aide de balises réceptrices des signaux émis. Typiquement des signaux Bluetooth et WiFi échangé par un Smartphone ou autres appareils et des bornes WiFi[94] ou balises Bluetooth[91].

Ainsi, tout effort dont le but est l'apprentissage et/ou l'inférence d'activités de la vie quotidienne sera amené à se restreindre à un sous-groupe d'activités délimité par l'intérêt du projet et les types de capteurs utilisés. En effet, le domaine applicatif de la détection d'activités est très vaste [78] et couvre transversalement plusieurs problématiques comme le suivi de l'activité physique de la personne [96], ses habitudes de vie [67] ou bien plus spécifiquement l'ergonomie de ses gestes sur des tâches spécifiques [65]. C'est pour cela que dans toute étude de l'état de l'art se base sur ou fournis une taxonomie des classes d'activités qu'elle étudie ou cherche à trouver dans un domaine précis [9]. Les grandes catégories d'activités que l'on rencontre en général sont :

— Les activités physiques sportives : marche, course, saut, natation
— Les activités physiques normales : s'asseoir, s'allonger, se déplacer, monter et descendre des escaliers, s'accroupir, chuter, perdre conscience
— Les activités d'interaction avec l'environnement : ouvrir un tiroir, une porte ou une fenêtre, prendre, porter, poser un objet, utiliser un objet sur soit, quelqu'un d'autre ou un autre objet
— Les activités d'interaction sociale : parler, applaudir, émettre des émotions
— Les activités de subsistance : préparation et prise de repas, visite de salle de bain, visite des w.c., sommeil

Chaque action contenue dans une classe ou sous-classe d'activité peut potentiellement générer des données similaires à une autre action d'une classe ou sous-classe différente, c'est pour cela qu'il est important de fixer et énumérer toutes les classes d'activités qui concernent une étude pour correctement traiter ces cas.

Les activités peuvent aussi se présenter avec des propriétés de composition et d'imbrication [23], c.-à-d. qu'une activité de préparation de repas par exemple peut être inférée par la reconnaissance d'une suite d'activités primitives chaînées à l'instar de la reconnaissance de chaînes de caractères qui peuvent se traduire ici soit par des activités, soit directement par des activations de capteurs.

2.3 Technologies de Capteurs pour la Détection d'AVQ

Pour proposer un suivi continu de l'état de santé d'une personne dans son milieu de vie quotidienne, nous avons ainsi vu qu'un moyen simple et efficace est de suivre ses AVQs. Pour se faire, il faut pouvoir physiquement capter les signaux physiques émis et radiés par les activités de la personne pour pouvoir en déduire les AVQs correspondants.

Un capteur est un système qui convertit une grandeur physique (température, humidité, lumière, ondes électromagnétiques de diverses fréquences) vers une autre qui est plus facilement mesurable, en général un signal électrique analogique ou numérique. Le domaine de recherche pour la détection des activités de la vie quotidienne regorge de capteurs divers et variés [92]. Ceux-ci peuvent être environnementaux tout comme portée, invasifs ou discrets.

2.3.1 Capteurs environnementaux

Capteurs de mouvement PIR

Figure 2.1 – Capteurs de mouvement

L'un des capteurs environnementaux les plus utilisés pour la détection de mouvement est le capteur PIR [57] [62]. Basé sur la détection passive des radiations infrarouges qu'un corps peut émettre. Ces capteurs émettent une simple information booléenne suivant s'ils ont détecté une variation infrarouge. Ils sont souvent utilisés pour allumer ou ouvrir automatiquement des portes ou bien enclencher une alarme. Leurs avantages sont leur simplicité, leur faible consommation électrique, le fait qu'ils peuvent détecter du mouvement dans des environnements peu éclairés et leur champ de vision réglable. Leurs inconvénients sont qu'ils demandent d'être soigneusement placés pour relayer des informations fiables : les obstructions généreront des faux négatifs et d'autres sources de variation de chaleur peuvent générer de faux positifs.

Capteurs de température, humidité et lumière

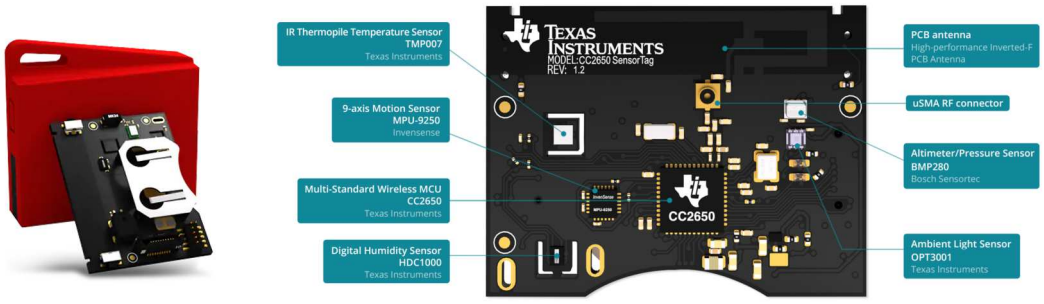


FIGURE 2.2 – Unité de mesure de température, humidité et lumière

Ces capteurs génèrent une série temporelle soit échantillonnée sur le temps soit événementiel suivant leur réglage des grandeurs physiques mesuré sur la pièce et plus précisément aux alentours de leur emplacement [41] [18] [11]. En général, au moins deux voir trois grandeurs sont mesurées par un seul composant et ils sont souvent vendu avec un capteur PIR en tant qu'un seul produit tout intégré. L'un des principaux avantages de ces capteurs et la riche information qu'ils peuvent délivrer sur l'état de la pièce ou du logement sur le long terme. Mais l'inconvénient est la nécessité de choisir une bonne période d'échantillonnage qui n'impactera pas la durée de charge de la batterie et un bon placement dans le logement pour ne pas être brouillée par des éléments perturbateurs, par exemple des sources de chaleur telles que la télévision ou le four. Cet inconvénient peut tout de même être tourné en un avantage. En effet, ils peuvent être utilisés pour détecter des actions plus spécifiques : l'utilisation de l'eau chaude dans un évier via le capteur de température ou humidité, l'ouverture d'un tiroir ou placard via le capteur de luminosité et la préparation d'un repas nécessitant de faire bouillir de l'eau via le capteur d'humidité.

Capteur magnétique

FIGURE 2.3 – Capteur d'ouverture et fermeture de porte

Ces capteurs sont souvent composés d'un aimant et d'un détecteur de variations de champs magnétiques [58] ou bien tout simplement de deux lames ferromagnétiques. Ils

sont souvent utilisés pour détecter une ouverture ou fermeture de porte, fenêtre ou tiroir. Souvent utilisé aussi dans les systèmes de sécurité. Assez simples dans leur conception et utilisation, ils offrent une information en général très fiable sur l'état de l'équipement instrumenté.

Capteurs gyroscopiques et accéléromètres

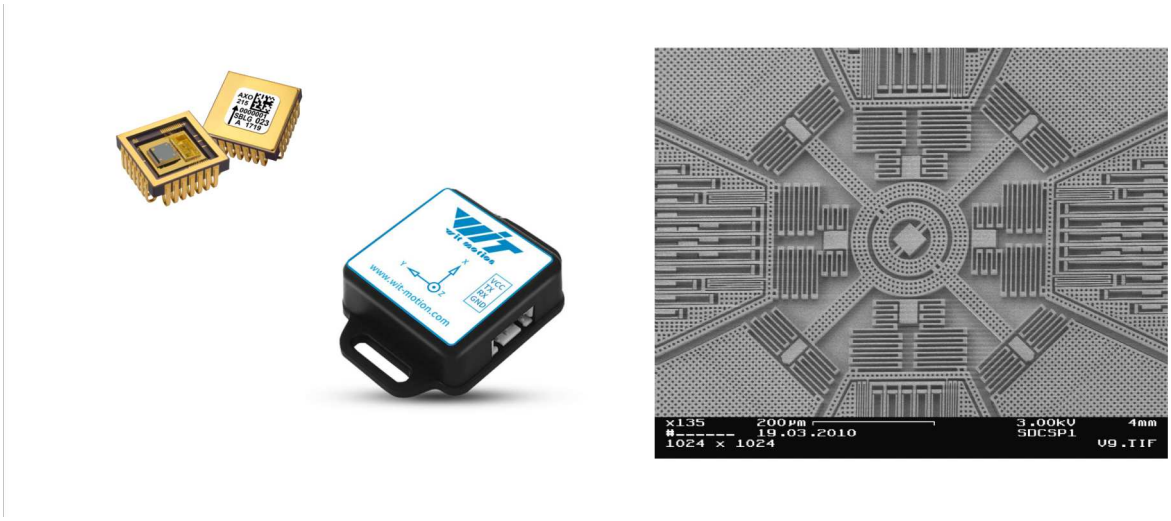


FIGURE 2.4 – Capteur équipé d'une centrale inertielle MEMS

Ces capteurs peuvent être employés en tant que capteurs portatifs peuvent être utile comme capteur environnemental pour détecter des ouvertures et fermetures de portes, fenêtres et tiroirs où même les accélérations d'un véhicule [87] [74].

Capteurs de pression

FIGURE 2.5 – Capteur de pression installé sur une chaise (à gauche) ou un lit (à droite)

Ces capteurs sensibles aux forces et au touché tactile sont généralement installé sur des surfaces prônes à être souvent en contact sur la durée avec des parties du corps tels les lits, chaises et tapis [6]. Suivant les types d'installation, ils peuvent permettre un suivi

régulier du poids de la personne jusqu'à pouvoir en extraire les paramètres de posture au repos de celle-ci. L'inconvénient étant l'encombrement et l'invasivité sur le milieu de vie instrumenté.

Capteurs RFID

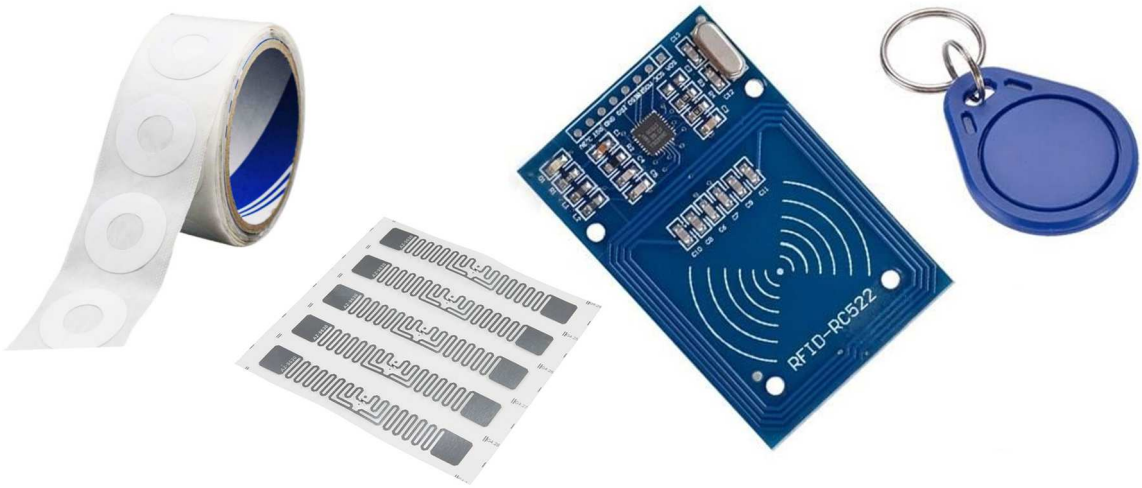


FIGURE 2.6 – Puce RFID et unité de détection

Une solution simple, mais demandant un grand effort de préparation et de maintenance sur le long terme pour la détection des AVQ instrumentaux est d'équiper chaque objet d'intérêt avec une puce RFID et le résident avec un bracelet lecteur [3].

Caméras

(a) Caméra thermique (b) Caméra RGB (c) Caméra infrarouge

FIGURE 2.7 – Différents types de caméra

"Caméra" désignant toute unité équipée d'un capteur de signal électromagnétique assisté par de l'optique. Les signaux captés et filtrés les plus répandus du spectre sont

l'infrarouge (caméras infrarouges et thermiques) [50] et le visible (caméras RGB) [20]. Certaines caméras de profondeur opèrent sur le principe du temps de vol où est calculé le temps mis par un laser à effectuer le trajet entre un obstacle et le capteur de la caméra [32].

Tout capteur équipé d'une caméra, quel que soit le type, est conjointement accompagné d'une unité de calcul pour compresser ou sélectionner les caractéristiques et/ou composantes intéressantes du signal qui sera transféré et fusionné aux autres données[60]. Ces capteurs présentent les désavantages suivants : haute consommation, nécessité de rester sur le secteur et problèmes d'acceptabilité par les résidents. Les avantages sont : la richesse d'information extractible de ces données (posture, squelette, démarche)[24], le suivi de plusieurs personnes en parallèle et le plus faible nombre de faux positifs (animaux de compagnie, invités, famille).

Autres signaux électromagnétiques

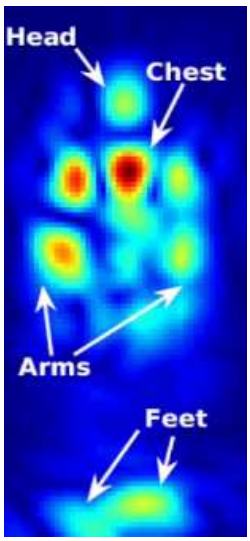


FIGURE 2.8 – Représentation d'une figure de personne capturée par signal électromagnétique

Certaines installations permettent le suivi de l'activité humaine à l'aide de la mesure des signaux radio WiFi réfléchis par le corps dû au fait que celui-ci est principalement composé d'eau[61]. Les systèmes de suivi d'activité par signaux radio sans assistance optique étant perçu comme étant moins intrusif [43] sont mieux acceptés. Ces systèmes offrent également une bien meilleure couverture spatiale de l'environnement surveillé. En effet, ils peuvent capter à travers les murs et meubles et ainsi minimiser les problèmes classiques d'occlusion qu'imposent les caméras.

FIGURE 2.9 – Une montre connectée contiendra un MEMS pour les fonctionnalités de détection d'activité

2.3.2 Capteurs portés

Capteurs à centrale inertielle MEMS

Se présentant pratiquement tout le temps sous la forme d'un microsystème électromécanique facilement intégrable dans tout système électronique, ces capteurs sont d'ores et déjà omniprésents dans notre quotidien. Ils sont installés au sein d'accessoires de la vie courante tels que les traqueurs de santé, podomètres numériques, montres connectées et les Smartphones. De ce fait, ces deux capteurs constituent une plateforme offrant une grande accessibilité aux données terrain.

Gyroscope

Un gyroscope mesure sa vitesse angulaire autour de ses trois axes. Couplé avec un accéléromètre ils peuvent permettre de déduire le mouvement de l'objet support, mais avec une certaine erreur qui peut s'accumuler sur le temps.

Accéléromètre

Un accéléromètre mesure une accélération linéaire. Si l'accéléromètre est immobile et posé sur une surface horizontale, il relèvera une accélération nulle sur ses deux axes horizontaux et une accélération de 1G sur l'axe vertical dû à la gravité. Pour les applications où il est nécessaire de mesurer la position et le mouvement d'un objet, les données accéléromètriques sont souvent couplé avec les données gyroscopiques par l'intermédiaire d'un filtre complémentaire ou de Kalman pour déduire une position angulaire et rapporter une accélération linéaire sans la composante de gravité.

Capteurs de fréquence cardiaque

Un capteur de fréquence cardiaque permet de suivre la fréquence cardiaque d'une personne en temps réel. Ces capteurs peuvent se baser sur deux méthodes de mesure différentes :

FIGURE 2.10 – Équipement pour effectuer une électrocardiographie

FIGURE 2.11 – Capteur de fréquence cardiaque utilisé dans les montres et bracelets connectés

— L'électrocardiographie, où il s'agit de mesurer l'activité électrique dans le cœur, un signal de l'ordre du millivolt est mesuré avec une fréquence d'échantillonnage d'environs 15kHz à l'aide d'électrodes à coller sur le torse.
— La photopléthysmographie, qui est une méthode basée sur un capteur optique appliqué sur la peau qui permet d'estimer le flux sanguin dû aux battements du cœur.

C'est cette deuxième méthode qui est la plus répandue dans les bracelets et montres connectés que l'on trouve dans le commerce grand public.

2.4 Algorithmes pour la Détection d'AVQ

Les algorithmes de détection des activités de la vie quotidienne sont très utiles pour la surveillance de personnes nécessitant une assistance à l'autonomie à domicile. Ces algorithmes ont principalement pour objectif soit la reconnaissance des activités, soit la prédiction de caractéristiques spécifiques dans divers signaux collectés sur la personne ou l'environnement.

Du fait de la nature des capteurs utilisés, l'équipement des logements et le type d'application désiré, le domaine applicatif des algorithmes visant l'assistance à l'autonomie à domicile est très divers. Ce domaine contient des sous-axes de recherche allant de la reconnaissance d'activité à la localisation et l'identification en passant par la détection d'anomalie. Il existe aussi des axes liés à la modélisation du contexte applicatif, la sémantique intrinsèque aux données, la topologie des placements de capteurs ainsi que les connaissances expertes.

2.4.1 Reconnaissance d'activité humaine

La reconnaissance d'activité est un domaine qui fait appel à des problématiques diverses que l'on peut classer selon que l'on utilise des capteurs portés par l'utilisateur, des capteurs ambiants ou encore des capteurs optiques tels des caméras [2].

Les capteurs portés et ceux basés sur la vision sont très efficaces pour identifier certaines situations. Ils peuvent permettre la prévention d'accidents corporels comme les chutes et surtout, ils sont adaptés à des environnements où plusieurs sujets co-habitent. Les capteurs ambiants, d'un autre côté, sont plus adaptés pour un suivi macroscopique des activités et habitudes de vie de la personne. Ils sont plus fiables, moins invasifs et mieux acceptés que les capteurs portés car aucune ou peu de participation de la part de l'usager est nécessaire à leur fonctionnement. Ils sont constitués généralement de réseaux communicants couvrant l'ensemble du logement et pouvant être maintenus par un opérateur à distance.

Les données issues de capteurs portés ou ambiants forment ce qu'on appelle des séries temporelles mono ou multidimensionnelles. Leur analyse nécessite différentes étapes de traitement :

— Collecte et transmission vers un centre de traitement des données,

2. Les traitements associés à des caméras ne seront pas abordés dans cette thèse car ils sortent du champ d'application des méthodes que nous avons développées.

— Filtrage et élaboration de caractéristiques discriminantes de ces données (moyenne, moyenne mobile, spectrogramme...),
— Reconnaissance d'activité,
— Prise de décision et diffusion d'information vers les différents intervenants et usagers.

La prise de décision et la reconnaissance d'activité peuvent s'appuyer sur des modèles obtenus par apprentissage automatique. De tels modèles nécessitent un gros travail d'entraînement en amont ce qui suppose l'établissement d'une base de données annotée sur laquelle se base l'apprentissage. Ces modèles et algorithmes sont divers et variés. Parmi les plus connus, on trouve les arbres de décision, les réseaux de neurones, les modèles probabilistes et modèles graphiques comme les chaînes de Markov et réseaux bayésiens. Parmi ceux-ci les modèles de Markov caché (souvent abrégé en HMM) et ses dérivés ont largement dominé la littérature de la reconnaissance d'activité de la dernière décennie.

L'apprentissage peut être supervisé ou non supervisé. Les techniques non supervisés sont plus exploratoires mais très prometteuses. La reconnaissance d'activités complexes ou composées participe à la prise de décision. Elle nécessite l'utilisation d'algorithmes d'apprentissage supervisé s'appuyant sur les séries d'activation des différents capteurs.

2.4.2 Apprentissage supervisé

Un problème d'apprentissage se pose comme suit :

Étant donné un ensemble de variables d'état $\{x_1, x_2, \ldots, x_n\}$ et un ensemble d'observations $\{y_1, y_2, \ldots, y_n\}$ lié par une fonction inconnue $y = f(x)$.

L'objectif est de trouver la meilleure approximation h de la fonction f.

Ainsi cette fonction h peut servir à soit classifier des observations $x = h^{-1}(y)$, soit à prédire de futures observations $y = h(x)$.

L'apprentissage supervisé considère l'existence d'une base d'entraînement qui est un ensemble d'entrées et sorties liées $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ permettant de faciliter la recherche de la fonction h et de l'évaluer.

L'exemple le plus connu et utilisé pour illustrer ces méthodes est celui de la décision de jouer ou non au golf suivant des données météorologiques **(author?)** [76]. Le tableau 2.12 représente une base de données de plusieurs jours de décisions qui ont été prises avec la météo de la journée correspondante. La même chose peut être envisagée en remplaçant, par exemple, les traits météorologiques par des symptômes de maladies et la décision de s' il y'a infection ou pas, etc. Nous utilisons cette base pour illustrer quelques méthodes d'apprentissage par la suite.

Classification naïve bayésienne

Il s'agit d'un classifieur qui utilise le théorème de Bayes pour calculer la probabilité d'appartenance à une classe en fonction des variables d'observations [70].

$$p(X \mid Y_1, \ldots, Y_n) = \frac{p(X)p(Y_1, \ldots, Y_n \mid X)}{p(Y_1, \ldots, Y_n)} \tag{2.1}$$

Jour	Ciel	Température	Humidité	Vent	Sortir jouer
1	ensoleillé	chaude	haute	faible	non
2	ensoleillé	chaude	haute	fort	non
3	couvert	chaud	haut	faible	oui
4	pluvieux	tiède	haute	faible	oui
5	pluvieux	froide	normale	faible	oui
6	pluvieux	froide	normale	fort	non
7	couvert	froid	normal	fort	oui
8	ensoleillé	tiède	haute	faible	non
9	ensoleillé	froide	normale	faible	oui
10	pluvieux	tiède	normale	faible	oui
11	ensoleillé	tiède	normale	fort	oui
12	couvert	tiède	haut	fort	oui
13	couvert	chaud	normal	faible	oui
14	pluvieux	tiède	haute	fort	non

FIGURE 2.12 – Caractéristiques de la météo et la décision correspondante

Il est considéré comme naïf, car il suppose que ses variables observées sont indépendantes pour simplifier le calcul de la vraisemblance. En effet, en appliquant la formule des probabilités composées et l'hypothèse d'indépendance des observations, le calcul de la vraisemblance devient :

$$p(Y_1, \ldots, Y_n \mid X) = \prod_{i=1}^{n} p(Y_i \mid X) \tag{2.2}$$

ainsi la probabilité conditionnelle de X sachant les observations $Y_1, \ldots, Y_n$:

$$p(X|Y_1, \ldots, Y_n) \propto p(X) \prod_{i=1}^{n} P(Y_i|X) \tag{2.3}$$

L'inférence d'appartenance peut être ainsi évaluée avec un classificateur comme le maximum *a posteriori* :

$$x = \underset{X}{\operatorname{argmax}}\ p(X|Y_1, \ldots, Y_n)p(X) \tag{2.4}$$

Pour un premier cas d'étude simple, prenons la base de données 2.12 et sélectionnons par exemple que l'attribut «Ciel» avec les observations sur la sortie pour jouer. On cherche à répondre à la question : "Est-ce qu'une sortie pour jouer est possible étant donné les conditions climatiques ?", c.-à-d. qu'elle est la probabilité $P(sortie = Oui \mid Ciel = X)$.

Chose qui est simplement faite avec un calcul de fréquences :

Ciel	Oui	Non		
ensoleillé	2	3	$P(ensoleillé)$	$5/14 = 0.36$
couvert	4	0	$P(couvert)$	$4/14 = 0.29$
pluvieux	3	2	$P(pluvieux)$	$5/14 = 0.36$
	$P(Oui)$	$P(Non)$		
	$9/14 = 0.64$	$5/14 = 0.36$		

$$P(sortie = Oui \mid Ciel = ensoleillé) = \frac{P(Ciel = ensoleillé \mid sortie = Oui)P(Oui)}{P(ensoleillé)} = \frac{2/9 \times 9/14}{5/14} = \frac{2}{5} = 0.4 \tag{2.5}$$

$$P(sortie = Non \mid Ciel = couvert) = \frac{P(Ciel = ensoleillé \mid sortie = Non)P(Non)}{P(ensoleillé)} = \frac{3/5 \times 5/14}{5/14} = \frac{3}{5} = 0.6 \tag{2.6}$$

Ainsi la valeur de l'attribut *Ciel* pour laquelle la distribution *a posteriori* est maximisée est $sortie = Non$. Donc, l'estimateur bayésien répond à la question par "les joueurs ne sortiront probablement pas jouer si le temps est ensoleillé".

Bien sûr, l'estimateur peut tout aussi s'appliquer avec plusieurs voire même tous les attributs c.-à-d. $P(sortie = s \mid Ciel = c, Température = t, Humidité = h, Vent = v)$:

$$\begin{aligned} P(s \mid c,t,h,v) &= \frac{P(c,t,h,v \mid s)P(s)}{P(c,t,h,v)} \\ &= \frac{P(s) \times \prod_{a \in \{c,t,h,v\}} P(a \mid s)}{\sum_{x \in \{Oui,Non\}} P(c,t,h,v \mid x)P(x)} \\ &= \frac{P(s) \times \prod_{a \in \{c,t,h,v\}} P(a \mid s)}{\sum_{x \in \{Oui,Non\}} \prod_{a \in \{c,t,h,v\}} P(a \mid x)P(x)} \end{aligned} \tag{2.7}$$

On remarque que la présence du produit $\prod_{a \in \{c,t,h,v\}} P(a \mid s)$ fait que si le modèle est utilisé pour la prédiction classe d'attributs non représentés dans les données d'entrée, alors, il lui affectera une probabilité de 0. Ceci est connu sous le nom du problème de la fréquence nulle. Il peut être résolu par des méthodes dites d'*additive smoothing*.

La méthode naïve bayésienne est très répandue dans les domaines de classification textuelle souvent complétés par des méthodes pour améliorer et nuancer son hypothèse naïve d'indépendance des variables [49]

Avantages	Inconvénients
✓ Simple et facile à comprendre et implémenter	× Très mauvais estimateur en général
✓ Marche bien sur des données qualitatives	× Marche moins bien sur des données quantitatives ou ils sont considérés comme normalement distribués
✓ Très rapide quel que soit le nombre d'attributs ce qui en fait un bon candidat pour toute application nécessitant de la prédiction en temps réel	× Présuppose l'indépendance des attributs
✓ Meilleur taux de réussite que d'autres algorithmes pour la classification de texte, filtrage d'emails indésirables et analyse de sentiment sur les réseaux sociaux	× Problème de la fréquence nulle pour la prédiction d'une classe d'attribut non présente dans les données de départ

Forêts d'arbres décisionnels

Aussi connues sous le nom de *forêts d'arbres aléatoires*, les forêts d'arbres décisionnels désignent une méthode d'apprentissage qui cherche à construire des arbres de décision. C'est une approche très utilisée, y compris pour des problèmes difficiles. Les premiers algorithmes développés par Microsoft pour le calcul de postures humaines à partir d'images de profondeur reposaient sur ce paradigme.

Un arbre de décision est constitué de :

— Nœuds : Représentant des attributs ou caractéristiques,

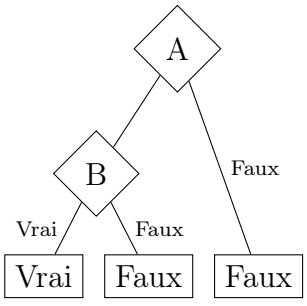
— Branchements : Représentant les différentes évolutions de l'arbre en fonction des décisions ou règles appliquées aux nœuds dont ils sont issus,

— Feuilles : Représentant les résultats de la décision.

La figure 2.13 est un exemple d'arbre de décision qui représente la table de vérité de la fonction booléenne `ET`.

A	B	A `ET` B
Faux	Faux	Faux
Faux	Vrai	Faux
Vrai	Faux	Faux
Vrai	Vrai	Vrai

(a) Table de vérité pour la fonction `ET`



(b) Arbre de décision pour la fonction `ET`

FIGURE 2.13 – Exemple d'un arbre de décision pour l'opération `ET` entre deux booléens

Les arbres de décision sont bien évidemment utilisés pour modéliser des fonctions plus

complexes qu'une simple fonction booléenne. Le principe de l'apprentissage par arbre de décision tourne autour de sa construction nœud par nœud en calculant pour chacun la variable d'entrée qui maximise un critère donné (par exemple l'information de Shannon ou le coefficient de Gini).

Pour mener cette construction, il existe plusieurs algorithmes : CART, C4.5, ID3, CHAID, ID4.5 etc [88].

Pour illustrer notre propos, reprenons la base de données 2.12 cette fois au complet. Le but est de construire un arbre de décision pour la variable "sortie au Golf". La méthode consiste à, itérativement pour chaque nœud en commençant par la racine, choisir sur quelle variable ce nœud va s'articuler en utilisant le critère de segmentation. Par exemple, on peut calculer ici soit le coefficient de Gini soit l'entropie. L'entropie pour chaque caractéristique est calculée en regroupant les valeurs de celle-ci et en réduisant avec une somme les instances des valeurs de la décision et le choix de la caractéristique du nœud en se basant sur la somme pondérée de ces valeurs.

La formule pour calculer le coefficient de Gini pour une valeur de caractéristique (*feature*) est donnée par :

$$Gini(feature = f) = 1 - \sum_{d \in décisions} p(décision = d \mid feature = f)^2 \tag{2.8}$$

et celle pour l'entropie par :

$$Entropie(feature = f) = - \sum_{d \in décisions} p(d \mid f) \log_2(p(d \mid f)) \tag{2.9}$$

La fonction pour calculer la valeur pour chaque caractéristique est ainsi donnée par le foncteur :

$$F(critère) = feature \mapsto \sum_{f \in feature} p(feature = f) critère(feature = f) \tag{2.10}$$

Par exemple pour $feature =$ Ciel :

$$\begin{aligned}
Gini(\text{Ciel} = \text{ensoleillé}) &= 1 - \left(\frac{2}{2+3}\right)^2 - \left(\frac{3}{5}\right)^2 = \frac{12}{25} = 0.48 \\
Gini(\text{Ciel} = \text{couvert}) &= 1 - \left(\frac{4}{4+0}\right)^2 - \left(\frac{4}{4}\right)^2 = 1 - 1 - 0 = 0 \\
Gini(\text{Ciel} = \text{pluvieux}) &= 1 - \left(\frac{3}{3+2}\right)^2 - \left(\frac{2}{5}\right)^2 = \frac{12}{25} = 0.48
\end{aligned}$$

Ainsi :

$$F(Gini)(\text{Ciel}) = \frac{5}{5+4+5} \times 0.48 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.48 = 0.34$$

Ciel	Oui	Non	Gini	Entropie
ensoleillé	2	3	0.48	0.97
couvert	4	0	0	0
pluvieux	3	2	0.48	0.97
Total			0.34	0.69

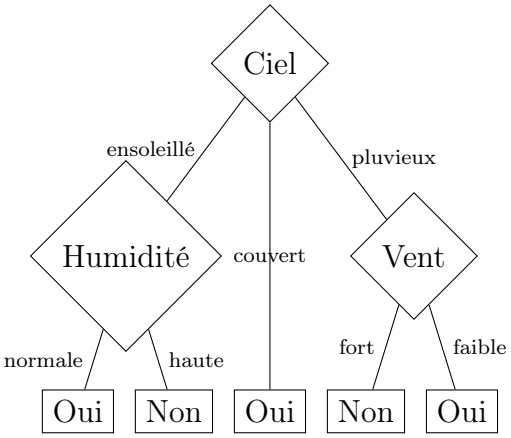
Température	Oui	Non	Gini	Entropie
chaude	2	2	0.5	1
tiède	3	1	0.37	0.81
froide	4	2	0.44	0.92
Total			0.44	0.91

Humidité	Oui	Non	Gini	Entropie
haute	3	4	0.49	0.99
normale	6	1	0.24	0.59
Total			0.37	0.79

Vent	Oui	Non	Gini	Entropie
faible	6	2	0.38	0.81
fort	3	3	0.5	1
Total			0.43	0.89

Remarque pour le calcul de l'entropie, on admet par convention que $0 \times \log 0 = 0$ car $x \log x \xrightarrow[x \to 0]{} 0$

Ainsi, pour cet exemple, la caractéristique dont le critère de segmentation est le plus bas est «Ciel» ce qui en fait le choix pour le nœud courant qui est la racine de l'arbre de décision. Ensuite la même méthode est appliquée pour trouver quelle caractéristique sera sur les trois branchements issus de ce nouveau nœud (de par les valeurs possibles de sa caractéristique) pour obtenir l'arbre de décision suivant :



Comme on peut le constater quand «Ciel = couvert» le nœud devient une feuille contenant la valeur de la décision lorsque le critère de segmentation est nul. C.-à-d. qu'il n'existe qu'une seule issue relative à cette valeur d'attribut quel que soit les valeurs des autres attributs.

La méthode que nous venons de présenter génère un arbre unique, ce qui présente plusieurs inconvénients, comme le surapprentissage ou le biais de sélection. C'est pour pallier ces inconvénients que **(author?)** [17] propose une méthode d'apprentissage dite par forêts d'arbres aléatoires.

Dans cette approche on ne construit plus un seul arbre mais une multitude d'arbres de décision partiellement indépendants, entraînés séparément sur des échantillons tirés

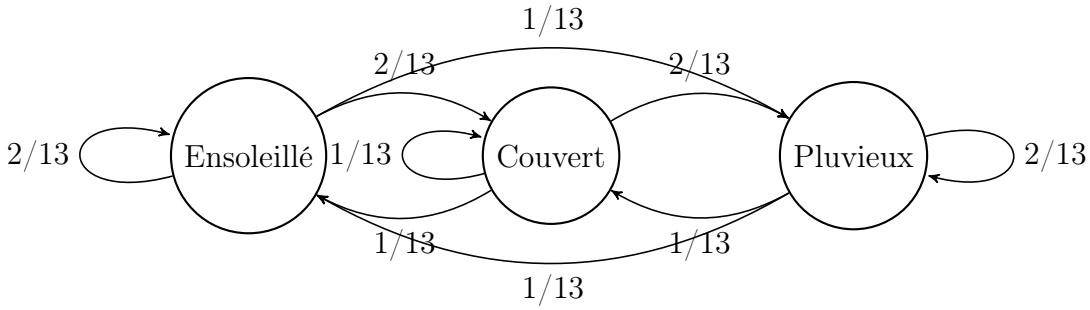


FIGURE 2.14 – Chaîne de Markov décrivant les changements d'état de l'attribut «Ciel»

aléatoirement des données d'entrées. Cet ensemble d'arbres constitue la «forêt» d'arbres décisionnels et la prédiction de la forêt résulte d'un simple vote majoritaire.

L'aspect aléatoire de ces arbres est caractérisé par l'échantillonnage des données d'entrée et celui des attributs choisis pour la segmentation des nœuds. Les données d'entrée sont tirées aléatoirement avec remise (bootstraping ou réplication multiple). Le nombre de prédicteurs utilisés pour la segmentation de chaque nœud est lui aussi tiré aléatoirement parmi tous les prédicteurs disponibles. Typiquement si l'on dispose de N prédicteurs, on retient $p = \sqrt{N}$ prédicteurs pour segmenter un nœud d'un arbre de la forêt.

Les forêts d'arbres décisionnels sont très répandues dans le domaine de la reconnaissance de mouvement ou d'activité basée sur des données issues de capteurs portés. On utilise typiquement l'accéléromètre [21] mais on peut le compléter avec l'ajout d'autres signaux physiques tel que la fréquence cardiaque [66].

Avantages	Inconvénients
✓ Quasi sans pareil pour résoudre des problèmes de classification	× Pas très adapté pour les problèmes de régression et plus particulièrement si les données sont de nature linéaire
✓ Marche très bien sur des données qualitatives et quantitatives	× Tend à être biaisé envers les classes d'attributs les plus présentes dans la base de données d'entrée
✓ Préparation des données simplissimes ne nécessitant ni mise à l'échelle ni normalisation	× Nécessite pas mal de puissance de calcul suivant le nombre d'arbres et d'attributs choisis

Modèle de Markov caché

Une chaîne de Markov est un processus stochastique discret qui vérifie la propriété de Markov qui dit que la probabilité conditionnelle de transition vers un état à l'instant $t+1$ ne dépend que de l'état à l'instant t, c.-à-d. $P(X_{n+1} \mid X_0, X_1, \ldots, X_t) = P(X_{n+1} \mid X_t)$. La figure 2.14 décrit la chaîne qui régit les changements de classes de l'attribut «Ciel» selon les observations de la base de données 2.12.

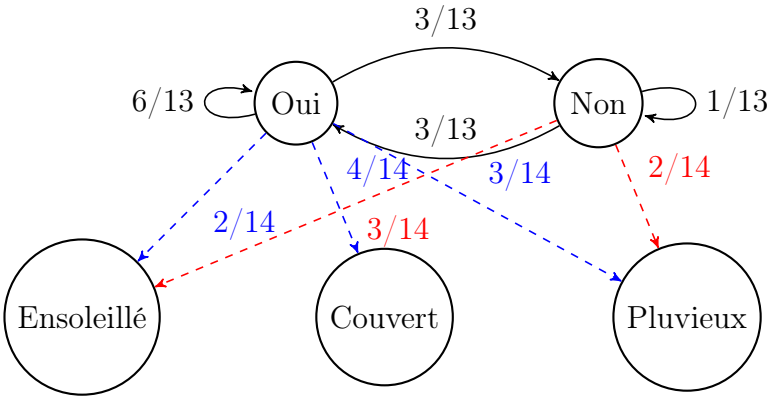


FIGURE 2.15 – HMM décrivant les changements d'état de l'attribut «Ciel»

Une chaîne de Markov caché (HMM) rajoute des états non observables au modèle en supposant que l'on peut observer des états dont les réalisations dépendent de ces états cachés. Par exemple, en reprenant la base de données 2.12 précédente nous pouvons nous poser la question de connaître ce qui régit le processus de «Sortir jouer» en observant l'évolution du temps dans le ciel. La figure 2.15 représente le HMM du lien entre «Sortir jouer» et les observations effectuées sur «Ciel».

Un HMM permet de répondre à trois principales questions :

— La probabilité qu'une séquence Y soit générée par le modèle M : $P(Y \mid M)$

— Quelle est la séquence d'états cachés la plus probable ayant permis de générer une séquence d'observations donnée $P(X \mid Y, M)$

— Quel est le modèle M le plus probable qui a généré une séquence Y : $P(M \mid Y)$

Les HMMs sont souvent utilisés dans les modèles visant la reconnaissance de parole [51] [1].

Avantages	Inconvénients
✓ Offre de meilleurs résultats qu'un modèle de Markov classique	× Bien plus coûteux en calcul que ce soit en temps ou en espace
✓ Le principe de fonction d'un HMM est applicable à une pléthore de modèles dans des domaines d'études divers et variés	× Reste contraignant en termes de corrélations des états et données par l'hypothèse de Markov
✓ Bibliographie très fournie avec des algorithmes bien renseignés et étudiés	× Apprentissage plus complexe
✓ Nécessite peu de données en comparaison avec des techniques de *deep learning*	

Réseaux bayésiens

Les réseaux bayésiens sont des modèles graphiques probabilistes représentant les relations de dépendance conditionnelles entre les variables aléatoires.

Par exemple, "sortir jouer" est une variable aléatoire dépendante de l'état du ciel et de la température. Cette relation est représentée par le réseau 2.16. Après avoir construit ce graphe de dépendance, il est possible de définir les tables des probabilités pour chaque variable si celle-ci est qualitative et que l'on dispose d'un moyen de les calculer. Pour les variables «Ciel» et «Température» c'est assez simple vu qu'elles ne sont dépendantes d'aucune autre variable, mais pour la variable «Sortie» la table des valeurs dépend des valeurs de ces dernières.

Si les variables sont qualitatives, alors elles sont définies par des distributions dont les paramètres sont appelés «hyper paramètres». Dans ce cas, l'inférence est une estimation de ces hyper paramètres.

L'inférence bayésienne sur des modèles complexes se base principalement sur les avancées en programmation probabiliste et les méthodes d'échantillonnage aléatoire comme le Monte-Carlo par chaînes de Markov [82], dont la plus connue est le Hamiltonian Monte-Carlo [15] qui se base sur la différentiation pour l'inférence de modèles complexes.

Avantages	Inconvénients
✓ Permet une meilleure estimation du risque et du hasard que les autres modèles qui ne se basent que sur les espérances	× Graphe acyclique signifie qu'il n'y a pas de support pour des boucles de rétroaction
✓ Étant basé sur le théorème de Bayes, le modèle permet d'inférer les conséquences des causes, mais aussi les causes des conséquences	× Nécessite de structurer les connaissances expertes et prior
✓ Adapté à de petites bases de données et supporte des données incomplètes	× Potentiellement assez sensibles aux priors
✓ Permet de facilement incorporer des priors et connaissance experte directement dans le modèle	

2.4.3 Apprentissage non-supervisé

Contrairement à l'apprentissage supervisé, il n'est plus nécessaire de disposer de base d'entraînement étiquetées $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ et la problématique se traduit plutôt comme suit :

Étant donné un ensemble d'observations $\{y_1, y_2, \ldots, y_n\}$ quel est l'ensemble des parties de celui-ci $\{\{y_i, \ldots, y_j\}, \ldots, \{y_n\}, \ldots, \{y_k, y_{k+m}, y_m\}\}$ qui présente la meilleure segmentation entre ses éléments suivant une fonction de critère h ?

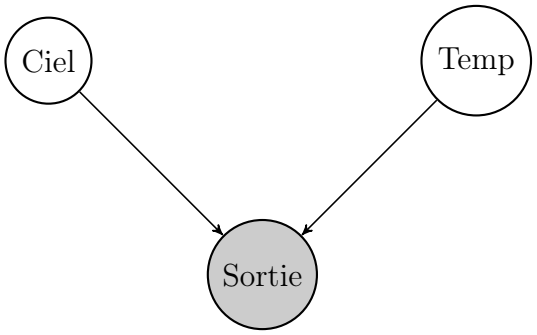
Classification

La classification est un exemple typique d'apprentissage non supervisé. Le but de la classification est de pouvoir former des collections d'objets qui vérifient les deux principales propriétés suivantes :

— Les objets appartenant à la même collection sont similaires.
— Les objets appartenant à des collections différentes sont dissimilaires.

Ciel		
ensoleillé	couvert	pluvieux
5/14	4/14	5/14

Température		
chaude	tiède	froide
4/14	6/14	4/14



		Sortie	
Ciel	Temp	Oui	Non
ensoleillé	chaude	0	2/14
	tiède	1/14	1/14
	froide	1/14	0
couvert	chaude	2/14	0
	tiède	1/14	0
	froide	1/14	0
pluvieux	chaude	0	0
	tiède	2/14	1/14
	froide	1/14	1/14

FIGURE 2.16 – Réseau bayésien décrivant la dépendance de la variable de sortie sur l'état du Ciel et la Température : $P(Sortie \mid Ciel,\ Température)$

Les collections formées ne sont pas prédéfinies, d'où le caractère non supervisé de cette méthode. On ne connaît pas la sémantique des classes qui vont être formées à l'issue du processus de classification.

L'efficacité d'un regroupement en classse repose sur la mesure de similarité/dissimilarité ainsi que sur l'algorithme utilisé pour regrouper les objets.

Formellement, un problème de regroupement en classse de données peut être posé de la sorte :

Étant donné un échantillon $E = \{w_1, w_2, \ldots, w_N\}$ d' N objets.

À chaque objet est associées $w_i = x_i^1, x_i^2, \ldots, x_i^p$ p variables observées/mesurées. Elles peuvent être qualitatives et/ou quantitatives.

Si l'on s'équipe d'une distance $d : E \times E \to \mathrm{R}$, et d'une fonction d'attribution $S : [\![1, N]\!] \to [\![1, K]\!]$ avec $K \geq 2$ le nombre de classes souhaité.

L'objectif sera de trouver la fonction d'attribution qui minimise une fonction de coût C décrivant les propriétés voulues pour les classes formées par cette première.

Un exemple de fonction de coût est la variabilité intra classe [2]. Pour regrouper les données avec cette dernière, il suffit d'évaluer l'expression suivante :

$$S_{\text{objectif}} = \underset{S}{\operatorname{argmin}}\ C(S) = \underset{S}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{S(i)=k} \sum_{S(i')=k} d(w_i, w_{i'}) \tag{2.11}$$

La littérature regorge de différentes méthodes et d'algorithmes adaptés aux données et aux applications. **(author?)** [16] proposent dans leur papier une vue d'ensemble des différents algorithmes de regroupement avec leurs avantages et inconvénients. Parmi les plus connus, on trouve :

Algorithme	Avantages	Inconvénients
Regroupement par Partitionnement	✓ Relativement simple et passe facilement à l'échelle	× Dégradation de l'efficacité dans les espaces de grande dimension
	✓ Adapté pour les données avec des classes bien marquées	× Nécessite de préciser le nombre de classes à l'avance
		× Très sensible à la phase d'initialisation, au bruit et aux valeurs aberrantes
Regroupement par Hiérarchisation	✓ Ne nécessite pas de fixer le nombre de classes à l'avance	× Ne permet pas d'effectuer des modifications ou des corrections une fois une décision de merger ou séparer est prise
	✓ Calcule toute une hiérarchie de regroupements possibles	× Critère de terminaison assez vague
	✓ Utilise un dendrogramme comme représentation graphique qui est visuellement parlante	× Efficacité sévèrement dégradée sur les espaces de grande dimension
Regroupement par modélisation	✓ Chaque classe peut être caractérisée par un nombre restreint de paramètres	× Intensif en calcul si le nombre de distributions est grand ou quand il n'y a que très peu de données.
	✓ le résultat respectera les hypothèses faites sur les modèles génératifs	× A besoin de grosses bases de données en entrée
		× Difficile d'estimer le nombre de classes sous-jacent.

2.5 Conclusion

La télésurveillance à domicile couvre un vaste domaine de recherche et de développement comprenant les problématiques suivantes : santé, capteurs, systèmes embarqués et algorithmiques. Les AVQ peuvent être classées en plusieurs catégories d'activités (sportives, normales, interaction avec l'environnement, etc.). Les activités de chaque catégorie nécessitent typiquement une instrumentalisation différente pour pouvoir les capter. Par exemple, les activités sportives sont plutôt suivies à l'aide de médaillons portés alors que le déplacement dans un logement ou l'interaction avec des tiroirs se prête plus à être détecté par des capteurs environnementaux.

Ces capteurs utilisent plusieurs types de technologies pour mesurer les signaux physiques résultant de l'activité des personnes. Ces signaux peuvent être optiques, mécaniques, thermiques et mêmes radios.

Plusieurs méthodes d'inférence sont donc applicables sur ces signaux pour en extraire

l'activité sous-jacente. Ces algorithmes sont classés principalement en deux catégories : les algorithmes d'apprentissage supervisé qui nécessitent une base de données étiquetée pour l'entraînement de leurs modèles, et les autres dits d'apprentissage semi ou non supervisé quand c'est possible.

3

Contexte Applicatif

Sommaire

3.1 Le Projet «36 mois de plus»

En 2012 Pharmagest créa la division *«innovation & e-santé»* pour se positionner sur le marché de l'observance, la télémédecine, le maintien à domicile connecté et l'intelligence artificielle. C'est dans cette optique qu'elle fit l'acquisition de Diatelic pour son expertise dans le domaine de l'intelligence artificielle et les systèmes experts prédictifs de l'évolution de l'état de santé des patients.

Une étroite collaboration avec deux équipes-projets Inria (Maia devenu ensuite Larsen, Trio), l'université de Lorraine et l'ancienne région Lorraine maintenant la région Grand Est avait pris forme sous le nom de projet Satelor [30].

Le projet *«36 mois de plus»* [73] se plaçait comme la suite logique au niveau du transfert des connaissances, résultats et délivrables de Satelor. La finalité étant de passer vers une seconde phase d'expérimentations sur le terrain à l'aide de partenaires industriels et médico-sociaux.

3.1.1 Objectifs

Le principal objectif du projet est de délivrer une solution commerciale multiservice à destination des personnes âgées capables de :

— Prévenir les chutes.
— Prévenir, le plus en amont possible le risque de décompensation de la personne âgée, afin d'éviter le recours aux services d'urgence.
— Détecter les signes de perte d'autonomie.
— Renforcer les liens sociaux entre les personnes âgées et leurs proches.
— Aménager et sécuriser l'habitation.

L'offre est composée de plusieurs logiciels, dispositifs et services (illustrés dans la figure 3.1) qui conjointement permettent de mener à bien cette mission. Parmi les dispositifs il y a bien sûr les capteurs et la *box* associés, les serveurs, mais aussi des outils plus classiques tels que des tablettes tactiles, téléphones et aides techniques classiques tels que les barres d'appui et du mobilier adapté. Les logiciels comprendront évidemment les systèmes experts d'analyse de données déployées soit sur les *box*, soit centralisés sur les serveurs et les logiciels de coordination et de *reporting*. Les services comprennent une plateforme d'appel, un service d'évaluation du logement et d'installation des capteurs et le suivi par du personnel médical.

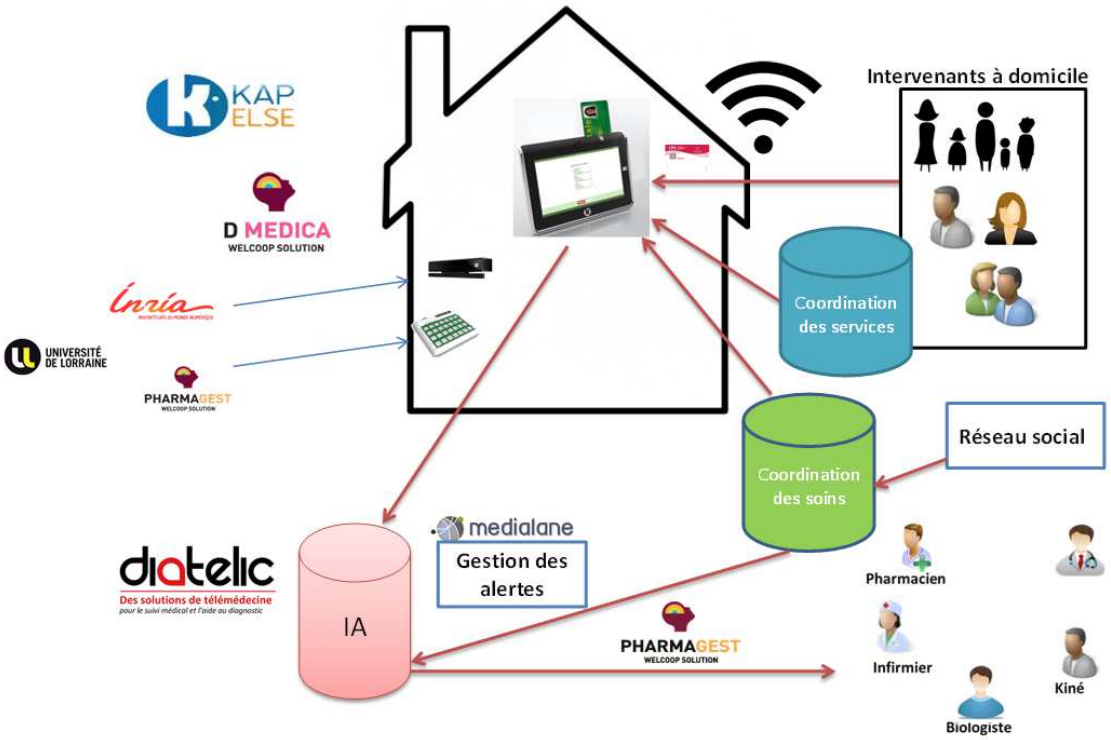


FIGURE 3.1 – Illustrations des composantes de l'offre commerciale du projet «36 mois de plus»

Le nom du projet «36 mois de plus» est basé sur la thèse dictant qu'étant donné un suivi et traitement adéquat de l'état de santé d'un senior, celui-ci peut se permettre de continuer à vivre dans son logement personnel 36 mois de plus que la majorité des seniors en perte d'autonomie. C'est pour cela que l'un des principaux objectifs du projet est la détection des signes de perte d'autonomie permettant une intervention rapide avec les

bons traitements pour bloquer toute dégradation potentielle de son état de santé.

L'un des principaux outils utilisés pour l'évaluation de la perte d'autonomie est la grille AGGIR (Autonomie, Gérontologie, Groupes Iso-Ressources) [80]. C'est l'outil de référence national français pour évaluer le degré de dépendance des personnes âgées. Habituellement le médecin traitant remplit cette grille en se basant sur des observations et un questionnaire pour déterminer le niveau de dépendance appelé *«GIR»* pour Groupes Iso Ressources. Celui-ci peut être classé de 1 à 6, le niveau 1 correspondant à une perte d'autonomie totale et le 6 à une parfaite autonomie.

Le GIR est fixé à travers 10 critères d'évaluation :

Cohérence : Conserver ou se comporter de façon sensée.

Orientation : Se repérer dans le temps et l'espace.

Toilette : Se laver seul.

Habillage : S'habiller, se déshabiller, se présenter.

Alimentation : Manger des aliments préparés.

Élimination : Assumer l'hygiène urinaire et fécale.

Transferts : Se lever, se coucher, s'asseoir.

Déplacements à l'intérieur : Mobilité spontanée, y compris avec un appareillage.

Déplacement à l'extérieur : Se déplacer à partir de la porte d'entrée sans moyen de transport.

Communication à distance : Utiliser les moyens de communication, téléphone, sonnette, alarme.

Vu que la majorité de ces critères correspondent partiellement ou totalement à des activités de la vie quotidienne, déterminer le niveau d'autonomie de la personne suivi revient ainsi à la détection de ses AVQs.

3.2 Installations

Dans le cadre du projet 36 mois de plus, Pharmagest audit et équipe des appartements et chambres en EPAD de résidents seniors. Deux types d'offres sont proposées. Les deux comprennent une base commune qui contient une *box* et des capteurs dits *universels*. La différence étant qu'une des deux offres embarque en plus un capteur de profondeur. La *box* de celle ci embarque plus de fonctionnalités de suivi et de communication.

Lors de l'installation, un technicien met en place et configure ces capteurs universels en renseignant des informations sémantiques relatives au placement de chaque unité (le type de la pièce : cuisine chambre salle de bain...). Après l'installation, les données sont soit traitées sur place pour lever des alertes urgentes (chutes etc...), soit remontées sur un serveur où elles seront traitées pour être affichées sur des interfaces d'accès.

3.2.1 Représentation Formelle de la Topologie du Logement

La topologie du logement est représentée comme un graphe définissant les transitions possibles entre chaque pièce (Figure 3.4 étant un exemple illustrateur de la représentation du logement montré sur la figure 3.2). Lors de l'installation, le plan est construit

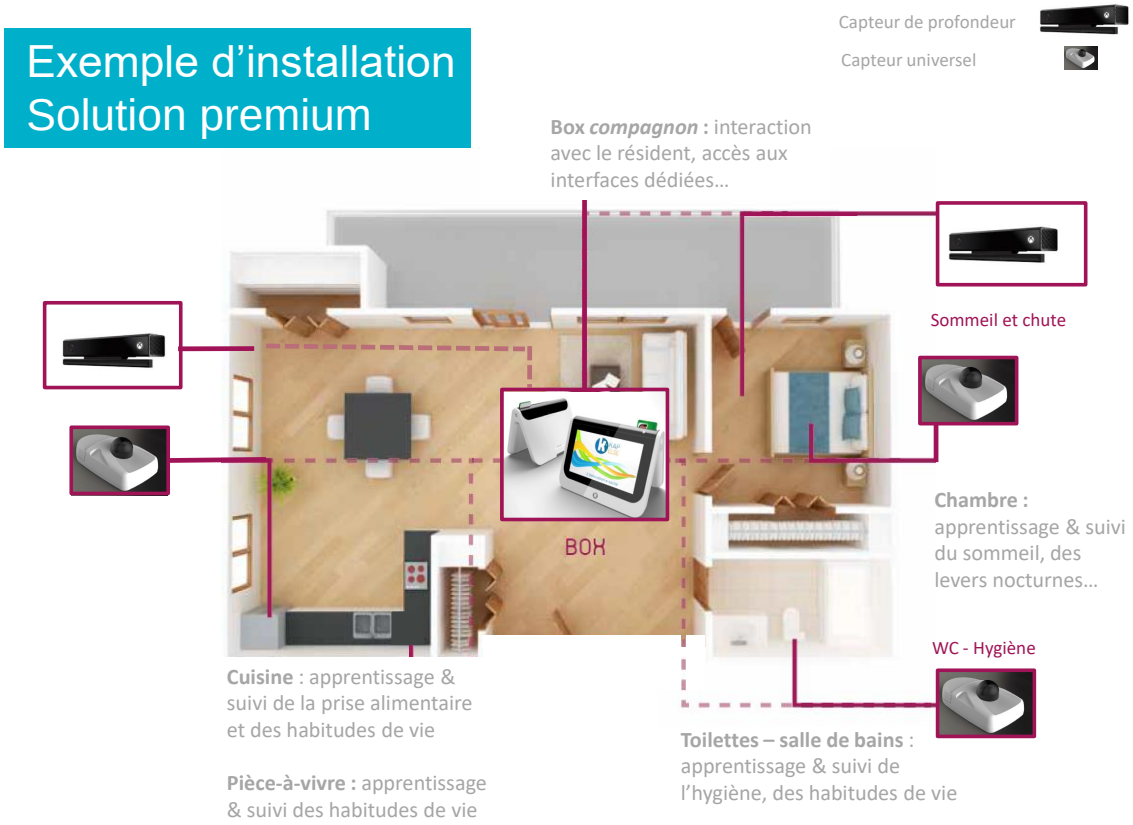


FIGURE 3.2 – Illustration du plan d'un studio avec les emplacements des capteurs

pour que chaque capteur ajouté puisse être associé à une pièce donnée. Chaque pièce et chaque capteur peuvent être marqués ou catégorisés suivant les activités qui peuvent s'y passer. Par exemple un capteur placé sur un tiroir de la Cuisine sera marqué comme étant un capteur rapportant de l'information spécifique aux activités culinaires. La figure 3.3 schématise cette structure hiérarchique des données disponibles pour une installation.

On associe une carte à chaque installation. Une carte comporte une ou plusieurs chambres étiquetées par un identifiant unique et caractérisé par un nom de classe normalisé à tous les logements. Les principales pièces d'une résidence auxquelles on s'intéresse sont le `Salon`, la `Cuisine`, la `Chambre à coucher`, les `WC`, la `Salle de bain` et la `Salle à manger`.

Chaque pièce peut être caractérisée par un type d'activité tel que les Soins personnels, le Repos/Sommeil, les Activités sociales et l'Alimentation/Cuisine. Elle contient aussi une structure décrivant quelles pièces sont directement accessibles depuis celle-ci et leurs types. Enfin, une pièce contient une structure regroupant tous les capteurs l'équipant.

Un capteur est caractérisé principalement par : un identifiant unique, son type et son profil d'utilisation (Entrée/Sortie, Détection de rotation, Détection de translation ou Présence). D'autres informations complémentaires sont disponibles suivant le profil d'utilisation du capteur. Par exemple, la nature de la pièce adjacente si jamais le capteur fonctionne en profil Entrée/Sortie par exemple.

Chaque profil d'utilisation entraîne un mode de fonctionnement différent du capteur. Par exemple, le profil Entrée/Sortie correspond à une utilisation du capteur sur une porte ou une fenêtre pour détecter l'ouverture et la fermeture de celle-ci, mais aussi dans quel sens la personne l'emprunte. Ce profil permet aussi de détecter des ouvertures de meubles et de tiroirs coulissant grâce à la détection d'un mouvement de translation.

3.2.2 Capteurs

Capteur universel

Le capteur universel (illustration figure 3.5) est une unité communicante équipée d'une centrale inertielle et en option d'un capteur de mouvement ou d'un bouton.

L'identifiant unique de chaque capteur permet de récupérer sur la base de données les valeurs qu'il a enregistrées.

Une unité peut être configurée suivant plusieurs profils de suivi qui détermineront les capteurs et les fonctionnalités qui seront activés lors de son fonctionnement. Tous les capteurs et les fonctionnalités ne peuvent pas être actifs de façon concurrente pour des raisons de durée de vie et de consommation d'énergie. Un profil est caractérisé par des propriétés : grandeurs physiques perçues par les différents capteurs de l'unité.

Exemple des propriétés de profil d'une chambre générique :

— mouvement
— température
— luminosité
— humidité

Chaque unité embarque un capteur de mouvement *PIR*, un capteur de température, de luminosité, d'humidité, un accéléromètre, un gyroscope et un bouton-poussoir.

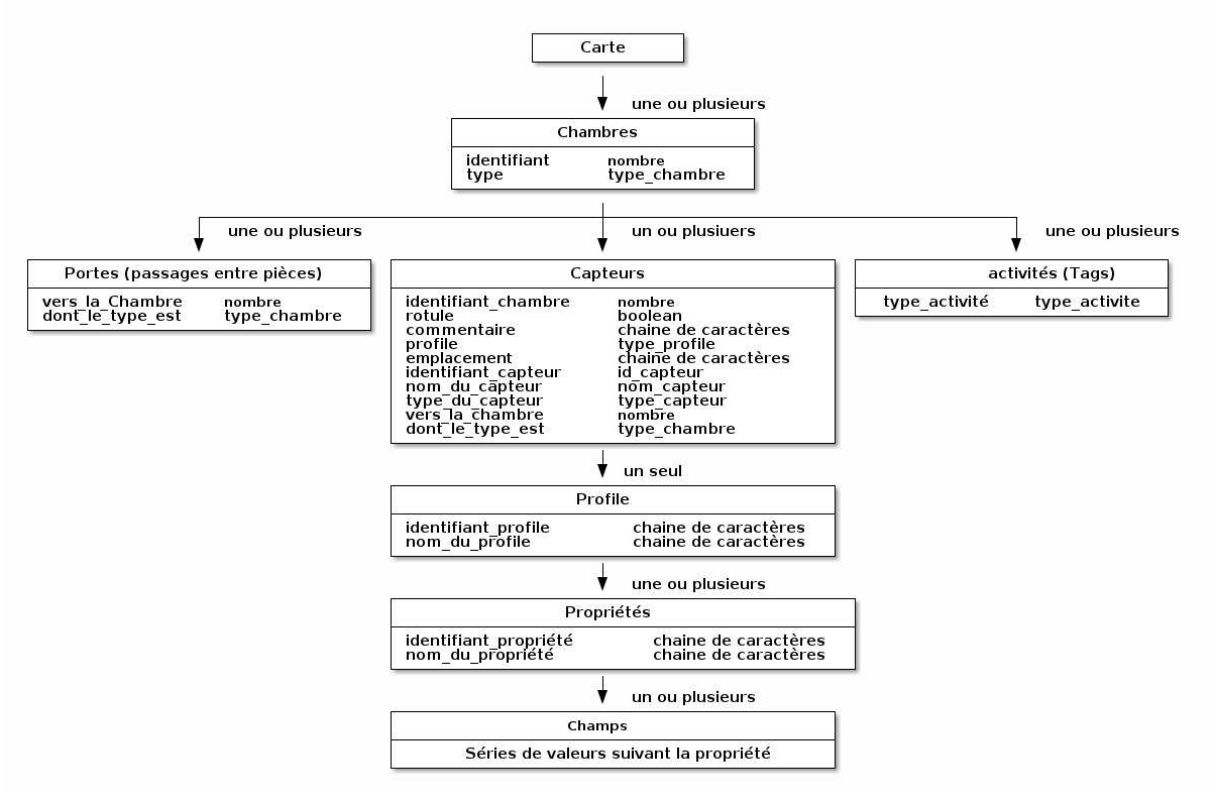


FIGURE 3.3 – Représentation structurelle des données

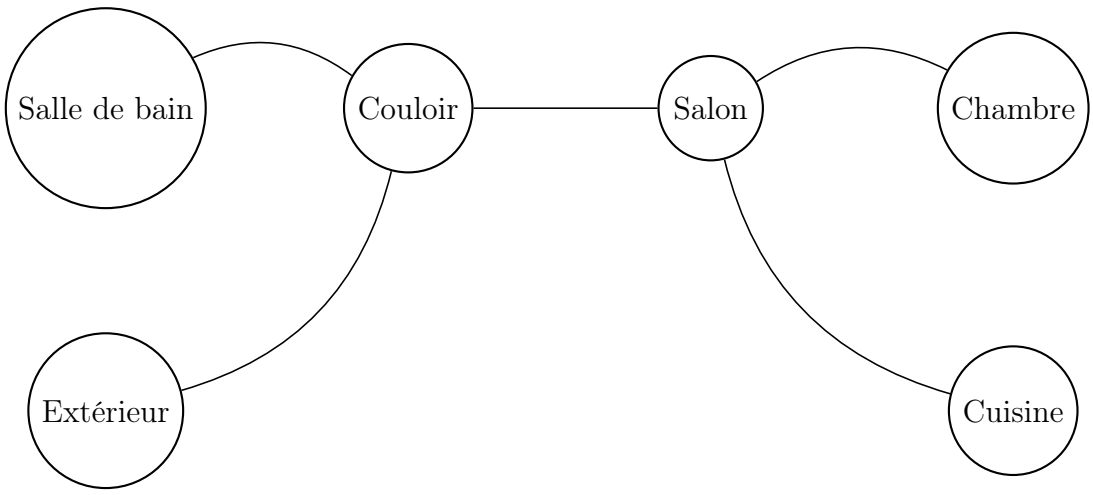


FIGURE 3.4 – Représentation de la carte d'un logement sous forme de graphe de transition entre ces différentes pièces

FIGURE 3.5 – Illustrations de la solution de capteur et box

On distingue 2 cas majeurs d'utilisation de ces capteurs :
- Quand le capteur est placé sur une porte
  - Profil de rotation/translation : La principale utilité du capteur sera, de par son gyromètre et accéléromètre, de détecter l'ouverture et la fermeture de la porte que ce soit une porte coulissante ou battante
  - Profil entré/sorti : En plus de la détection d'ouverture et fermeture le capteur est aussi capable de détecter s'il s'agit d'une entrée ou sortie de la pièce
- Quand le capteur est placé sur un mur pour détecter le mouvement de l'occupant, les variations de température et d'humidité de la pièce au fil du temps.

Les capteurs de mouvement émettent des données suivant un modèle événementiel illustré par la figure 3.6. À chaque nouveau mouvement, s'il est sporadique, un Dirac est émis à l'instant de l'observation. Mais si ce mouvement est conséquent, le capteur attend 10 minutes d'inactivité pour confirmer la fin du mouvement. La figure 3.6 représente l'automate décrivant ce comportement.

Capteur d'activité

Le capteur d'activité (illustré figure 3.7) est une caméra de profondeur permettant de suivre la silhouette des patients et d'identifier leurs différentes postures et propriétés physiques. Ces informations sont stockées sous la forme d'une série temporelle à deux dimensions : le type d'activité et sa durée.

Tous les états inférieurs à une certaine durée sont filtrés par le capteur et ne remontent pas sur le serveur.

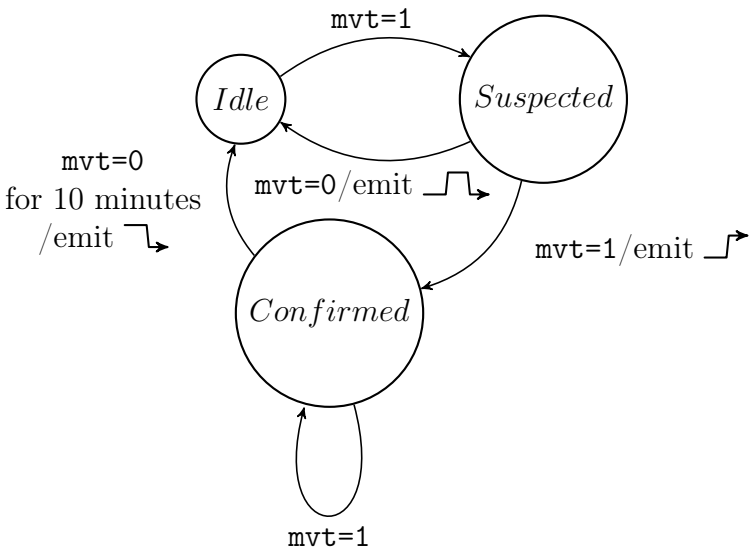


FIGURE 3.6 – Modèle de l'automate du capteur universel de mouvement

FIGURE 3.7 – Illustration du capteur d'activité en action

3.3 Données

Toutes les données sont sous la forme de séries temporelles (tableau 3.1) d'une ou plusieurs dimensions (tableau 3.2 et 3.3). Les données issues des capteurs universels sont des séries temporelles unidimensionnelles d'événements horodatés. Due à la nature du fonctionnement du capteur universel décrite plus haut, la quantité de données collectées sur le terrain reste relativement modeste par rapport à ce que des études en environnement contrôlé et cadré peuvent se permettre de collecter sans contraintes industrielles.

(tableau 3.2 et 3.3).

horodatage	température
2018-12-18 08:13:03+00:00	26.25
2018-12-18 08:28:03+00:00	26.75
2018-12-18 19:14:41+00:00	24.75

TABLE 3.1 – Exemple de série temporelle à une dimension ici mesurant la température

horodatage	paramètres des pas			vitesse de marche
	durée	longueur	nombre	
2016-07-06 13:25:32+00:00	0.880000	0.627651	6	0.032770
2016-07-06 13:28:18+00:00	1.366750	0.434064	2	0.013281
2016-07-06 14:42:54+00:00	1.500000	0.669095	2	0.018427

TABLE 3.2 – Exemple de série temporelle à plusieurs dimensions ici mesurant la démarche d'une personne

horodatage	activité	durée
2016-07-06 13:23:06+00:00	DEBOUT	3
2016-07-06 13:24:16+00:00	ASSIS	2
2016-07-06 13:24:52+00:00	MARCHE	3
2016-07-06 13:25:07+00:00	MARCHE	2

TABLE 3.3 – Exemple de série temporelle à plusieurs dimensions du capteur d'activité

3.4 Connaissances à priori

Il est important de s'aider et de rassembler toutes les connaissances *a priori* susceptibles d'aider à la construction des modèles de reconnaissance d'activités. Les activités principales qu'on cherche à rapporter avec nos capteurs ambiants sont :

1. WC : visite des toilettes,
2. Hygiène : douche, brossage des dents, utilisation du sèche-cheveux
3. Sommeil
4. Nutrition
5. Moments de repos
6. Domestiques : Activités de ménage, ...
7. Sociales : visite, discussions, jeux de sociétés, ...

On distingue trois caractéristiques sur lesquels on peut projeter ces activités :
— Topologie : qui concerne la relation de l'activité avec son lieu d'intérêt dans le milieu de vie
— Physique : qui lie l'activité à des grandeurs physiques et qui la discrimine
— Temporel : qui ancre l'activité dans des heures probables d'occurrence

Le tableau 3.5 montre l'*a priori* sur le lieu de ces types d'activités. Un lieu pondère plus ou moins l'activité examinée (très positivement ++, positivement +, neutrement ·, négativement - ou très négativement --). Typiquement, pour une activité de visite de toilettes, toute information provenant du lieu "WC" sera très riche et informative et aura un impact positif sur l'inférence de celle-ci. D'un autre côté, les informations provenant d'un autre lieu, par exemple le Salon, seront inversement proportionnelles sur la probabilité d'inférence de l'activité.

Suivant ce principe, on met en place le tableau 3.4 montrant les propriétés les plus intéressantes à regarder pour chaque activité.

	Salon	Cuisine	Chambre à coucher	WC	Salle de bain	Salle à manger
`WC`	-	-	-	++	+	+
`Hygiène`	-	-	+	+	++	-
`Sommeil`	-	-	+	+	+	-
`Nutrition`	-	++	-	-	-	++

TABLE 3.4 – Relation entre lieux d'intérêt et la nature physique des activités

	Mouvement	Température	Luminosité	Humidité	Rotation/Translation	Entrée/Sortie	Posture
WC	++	+	-	-	+	++	·
Hygiène	++	+	·	++	+	++	·
Sommeil	-	·	+	·	·	·	++
Nutrition	-	++	·	+	+	++	+

TABLE 3.5 – Relation entre lieux d'intérêt et les types d'activités principales que l'on cherche à trouver

3.5 Conclusion

Nous avons présenté dans ce chapitre le projet «36 mois de plus» qui est le cadre dans lequel se place cette thèse.

Ce projet implique différents acteurs et système dont le but est de construire et de faire subsister un écosystème complet pour l'assistance à l'autonomie à domicile des personnes âgées.

Les objectifs clés et concrets vont de la prévention de chutes, les risques de décompensation et la détection des signes de perte d'autonomie.

L'offre commerciale est composée de plusieurs sous-systèmes qui interagissent entre eux. Par exemple, des capteurs qui détectent et remontent l'activité du logement, des serveurs qui analysent les résultats et les communiquent au personnel adéquat, et même de l'intelligence locale capable de détecter les situations à haut risque et appeler une ligne d'urgence.

Les capteurs utilisés imposent ainsi des contraintes sur la nature des données que nos algorithmes doivent prendre en compte et savoir traiter : valeurs de données numériques ou booléennes, le temps et moment d'activation et sa durée, la topologie du logement, l'emplacement du capteur, etc.

La topologie du logement est potentiellement une source d'information sémantique sur l'activité rapportée par chaque capteur.

En perspective, un potentiel axe de recherche à explorer pourrait être de classifier automatiquement les activités suivant les lieux où ils ont été inférés.

4

Analyse de séries temporelles

Sommaire

4.1 Introduction

Une série temporelle est une réalisation d'un processus stochastique. En d'autres termes une suite de variables aléatoires (Y_t) qui évoluent dans le temps. Par exemple, le cours des actions en bourse, la consommation en électricité ou plus généralement des données issues de capteurs. Les problèmes impliquant les séries temporelles s'intéressent le plus souvent à l'analyse ou à la prédiction des tendances et des valeurs à venir.

L'une des grandes problématiques de l'analyse des séries temporelles est la segmentation en zones homogènes. La détection de points de rupture (*change point detection*) est une méthode de segmentation qui présente beaucoup d'intérêt, car de nombreuses applications cherchent à signaler, alerter au plus vite tout changement du processus mesuré. Utilisé à la base par **(author?)** [64] pour des applications d'analyse d'accidents industriels, en l'occurrence miniers, cette méthode et les problématiques associées représentent un domaine de recherche très actif [56]. Elle couvre des applications diverses telles que l'analyse d'anomalie sur le marché de l'électricité **(author?)** [40], la détection de brèches et d'attaques sur les réseaux informatiques **(author?)** [90], ou bien encore l'analyse EEG pour la détection d'activité dans le cerveau comme décrit par **(author?)** [52].

La finalité est la détection automatique des points de rupture et leur localisation. Le principe de base reste le même, quelle que soit l'application : la distribution de probabilités

(dont les variables aléatoires sont issues) change et notre objectif est d'automatiquement identifier ce changement et le moment de son occurrence. Néanmoins, plus les bases de données grandissent en taille et en volume, plus le nombre de solutions possibles augmente de façon combinatoire.

D'après **(author?)** [93] la détection de points de rupture (DPR) étant un problème de partitionnement d'une série temporelle, ceci en fait un problème hautement similaire à la segmentation d'une image. Celle-ci serait partitionnée en différents segments distincts suivant les textures et objets qu'elle représente. **(author?)** [93] argumentent que la DPR souffre du même problème de non-disponibilité de base de données de grande taille, extensive et de haute qualité pour l'évaluation de tels algorithmes à l'opposé des bases de données disponibles pour les problèmes de classification. Il propose ainsi une méthodologie d'évaluation mettant en jeu : une base de données, des métriques et les résultats de différents algorithmes de la littérature.

Il existe plusieurs algorithmes de DPR, parmi les principales on trouve la méthode de segmentation binaire [35] [85] [86], de segments adjacents [7] [10] ou l'algorithme PELT [55]. Le champ d'application de telles méthodes est très vaste et inclue la climatologie[79], la bio-informatique[36], la finance[97], l'océanographie[54], et l'imagerie médicale[71].

4.2 État de l'art

L'un des exemples les plus repris dans la littérature est un exemple qui exploite un jeu de données répertoriant les accidents se produisant dans les mines de charbon en Angleterre [47] [77].

```
[4, 5, 4, 0, 1, 4, 3, 4, 0, 6, 3, 3, 4, 0, 2, 6,
3, 3, 5, 4, 5, 3, 1, 4, 4, 1, 5, 5, 3, 4, 2, 5,
2, 2, 3, 4, 2, 1, 3, NaN, 2, 1, 1, 1, 1, 3, 0, 0,
1, 0, 1, 1, 0, 0, 3, 1, 0, 3, 2, 2, 0, 1, 1, 1,
0, 1, 0, 1, 0, 0, 0, 2, 1, 0, 0, 0, 1, 1, 0, 2,
3, 3, 1, NaN, 2, 1, 1, 1, 1, 2, 4, 2, 0, 0, 1, 4,
0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1]
```

La série temporelle, ci-dessus, dénombre année après année le nombre d'incidents impliquant plus de dix mineurs. On voit clairement dans ce jeu de donnée, figure 4.1 un changement qui apparait après la mise en place d'une réglementation rendant obligatoire de nouvelles règles de sécurité dans l'exploitation des mines.

Dans ce cas d'école nous pouvons visualiser qu'il y a bien un changement de distribution dans la répartition des points juste à l'aide d'une moyenne glissante comme le montre la figure4.1.

L'exercice de modélisation consiste à poser une variable aléatoire représentant le nombre d'incidents observés pour chaque année et supposer qu'elle suit une loi de poisson paramétré par un λ_1 ou λ_2 suivant que l'on se trouve avant ou après le point de changement de distribution :

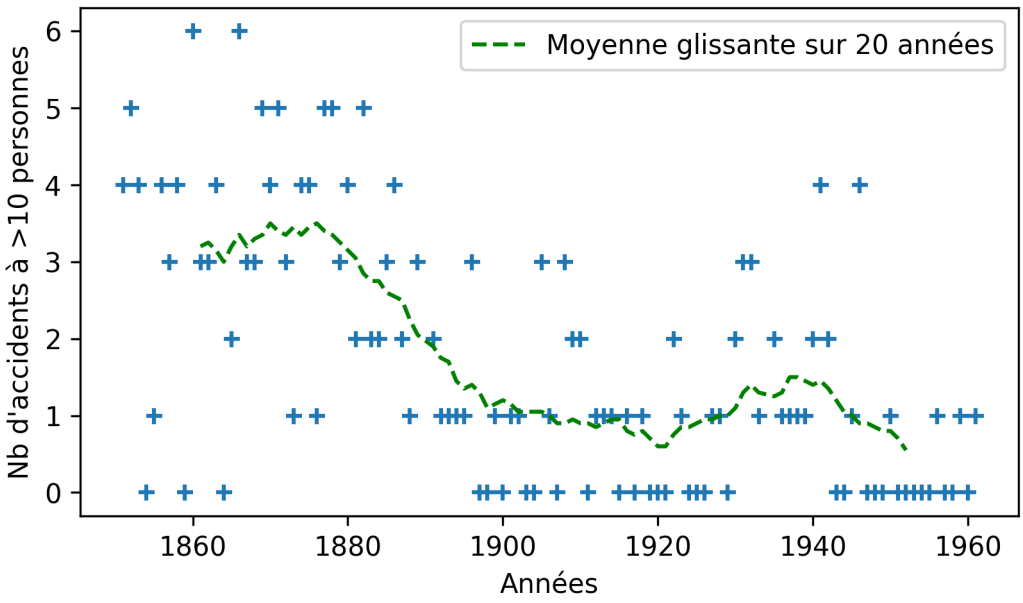


FIGURE 4.1 – Représentation de la série temporelle du nombre d'incidents ayant impliqué plus de dix mineurs avec une moyenne glissante sur 20 années

$$O_t \sim \mathcal{P}(\lambda_t) \tag{4.1}$$

$$\text{avec : } \lambda_t = \begin{cases} \lambda_1 & \text{si } \ t < T_{\text{switch}} \\ \lambda_2 & \text{sinon} \end{cases} \tag{4.2}$$

4.2.1 Formalisme

La formulation la plus simple et la plus répandue [72] [53] est basée sur les définitions suivantes :

Soit $y_{1:n} = (y_1, \ldots, y_n)$ une séquence d'observations ordonnée. Si $k, l \in \{1, \ldots, n\}$, avec $k < l$, on pose $y_{k:l} = (y_k, \ldots, y_l)$.

Définition 4.2.1. Une propriété statistique de $y_{k:l}$ peut être le résultat de toute fonction sur l'espace des séquences d'observations Y à valeur dans $\mathbb{R}$ (par exemple $\mathbb{R}^n$ pour une séquence de valeurs réelles ou $\mathbb{R}^{n \times d}$ pour une séquence de vecteurs réelles) :

$$f : Y \to \mathbb{R} \tag{4.3}$$

Par exemple, une moyenne de la séquence : $f(y_{k:l}) = \frac{\sum_{i=k}^{l} y_i}{l-k+1}$ est une propriété statistique.

Définition 4.2.2. On dit que $\tau \in \{2, \ldots, n-1\}$ est un point de rupture de $y_{1:n}$ si et seulement si *les propriétés statistiques* des séquences

$$y_{1:\tau-1} = (y_1, \ldots, y_\tau - 1) \quad \text{et} \quad y_{\tau:n} = (y_\tau, \ldots, y_n)$$

diffèrent d'une manière significative : c'est-à-dire que pour une propriété statistique f, étant donné une fonction de distance D, il existe un λ tel que :

$$D\left(f\left(y_{1:\tau-1}\right), f\left(y_{\tau:n}\right)\right) \geq \lambda \tag{4.4}$$

Cette définition peut bien sûr être étendue pour plusieurs points de rupture :

Définition 4.2.3. On dit que $y_{1:n} = (y_1, \ldots, y_n)$ admet m points de rupture

$$\tau_1, \tau_2, \ldots, \tau_m \in \{2, 3, \ldots, n-1\}$$

si

— $1 \leq \tau_1 < \tau_2 \cdots < \tau_m < n$,

— Pour tout $j = 1, \ldots, m+1$, les propriétés statistiques des séquences

$$y_{\tau_{j-1}+1:\tau_j}, \quad j = 1, \ldots m+1$$

diffèrent d'une manière significative.

On pose par convention $\tau_0 = 0$ et $\tau_{m+1} = n$.

En conséquence les $y_{\tau_{j-1}+1:\tau_j}$ vont subdiviser les observations $y_{1:n}$ en $m+1$ segments.

À chaque segment sont associés les paramètres $\{\theta_i, \phi_i\}$. ϕ_i est un ensemble possiblement nul de paramètres auxiliaires. θ_i et l'ensemble des paramètres dont on cherche les changements indiquant une rupture.

Typiquement, le but est de tester combien de segments sont nécessaires pour représenter les données. C.-à-d. une estimation du nombre et de la valeur des points de rupture présents dans une séquence d'observation.

Les méthodes pour la DPR peuvent principalement être caractérisées suivant plusieurs critères.

Un critère important est la quantité de données dont a besoin la méthode pour identifier des points de rupture. Ceci dicte sa capacité à inférer des résultats en temps réel ou bien «quasiment en temps réel». Les algorithmes sont décrits comme *hors ligne* quand ils ne sont applicables que sur l'ensemble des données à analyser. Un algorithme est dit ϵ-temps réel quand il nécessite d'observer une fenêtre de taille ϵ dans le futur après chaque candidat pour un point de rupture. Les algorithmes dits *en ligne* traitent généralement l'information en entrée sur une fenêtre glissante de taille n. De manière générale, n doit être assez large pour contenir assez d'information sur l'état de la série temporelle et assez serrée pour rester compatible avec la contrainte ϵ.

Un autre critère concerne plus particulièrement la nature des données en entrée de l'algorithme. En effet, les séries temporelles sur lesquelles la détection se fait peuvent être *univariées*, c'est-à-dire n'étant constituées que d'une seule variable qui évolue dans le temps. Quand la série temporelle est constituée de plusieurs variables, elle est dite

multivariées. Vu sous un autre angle c'est une série temporelle à une variable stochastique de dimension supérieure à 1.

Ces méthodes de DPR peuvent plus spécifiquement être subdivisées en plusieurs catégories :

— Les méthodes basées sur le ratio de vraisemblance, dont l'idée de base est d'analyser directement les propriétés statistiques des distributions à gauche et à droite de chaque candidat pour un point de rupture.

— Les méthodes à modèles probabilistes, souvent bayésiens qui répondent au problème en générant une distribution sur les positions possibles pour chaque candidat à un point de rupture.

— Les méthodes de regroupement, qui traitent le problème sous un différent angle où il s'agit de retrouver un nombre connu ou inconnu de regroupements de telle sorte que les observations dans chaque regroupement soient homogènes. C.-à-d. identiquement distribués. Si les observations entre un groupe et un autre sont différemment distribuées, alors il existe un point de rupture entre eux.

4.3 Approche par segmentation

4.3.1 Modélisation fréquentiste

Inférence avec l'estimateur du maximum de vraisemblance

Étant donné une séquence d'observations $y_{1:n} = (y_1, \ldots, y_n)$, nous supposons la présence d'un point de rupture τ au milieu de cette séquence.

Nous supposons également que les observations y_t suivent une loi de probabilité de paramètre θ_t et que ce paramètre change suivant la période à laquelle appartient y_t. C'est-à-dire :

$$\theta_t = \begin{cases} \theta_1 & \text{si } t < \tau \\ \theta_2 & \text{si } t \geq \tau \end{cases} \tag{4.5}$$

La figure 4.2 représente graphiquement ce modèle par un réseau bayésien.

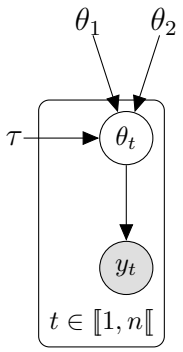


FIGURE 4.2 – Représentation graphique du réseau bayésien du modèle de segment de rupture

Pour inférer le meilleur découpage, il faut donc trouver les valeurs des paramètres du modèle τ, θ_1 et θ_2 qui maximisent la vraisemblance pour que la séquence d'observations $y = \{y_0, \ldots, y_n\}$ soit issue de ce modèle : $P(y_0, \ldots, y_n \mid \tau, \theta_1, \theta_2)$.

Ainsi, vu que nous cherchons $\hat{\tau}$:

$$\hat{\tau} = \underset{\tau}{\operatorname{argmax}} \max_{\theta_1, \theta_2} P(y_0, \ldots, y_n \mid \tau, \theta_1, \theta_2) \tag{4.6}$$

Nous supposons l'indépendance des variables y_t :

$$P(y_0, \ldots, y_n \mid \tau, \theta_1, \theta_2) = \prod_{t=1}^{n} p(y_t \mid \tau, \theta_1, \theta_2) \tag{4.7}$$

Nous pouvons ainsi décomposer ce produit pour trier les observations effectuées pendant la période $y_{1:\tau-1}$ et $y_{\tau:n}$:

$$\prod_{t=1}^{n} P(y_t \mid \tau, \theta_1, \theta_2) = \prod_{t \in [\![1,\ \tau[\![} p(y_t \mid \theta_1) \prod_{t \in [\![\tau,\ n]\!]} p(y_t \mid \theta_2) \tag{4.8}$$

Pour calculer l'équation 4.6 il faut maximiser ces deux termes correspondant à la vraisemblance pour que les séquences d'observations proviennent de la phase de rupture ou bien la phase normale :

$$\max_{\theta} p(y_t \mid \theta) \tag{4.9}$$

Dans la suite de cette thèse, nous supposerons que les observations y_t suivent une loi de Bernoulli de paramètre θ_t. Cette hypothèse se justifie à la fois pour simplifier les explications (l'extension à d'autres lois de probabilité est triviale) mais surtout car elle correspond à la mise en application de capteurs tout ou rien tel que ceux que nous utilisons (PIR, capteurs de contact, boutons poussoir et interrupteurs...).

Sachant que la valeur de θ_x qui maximise la vraisemblance d'une variable de loi de Bernoulli est la fréquence des observations, alors, la solution au problème d'optimisation 4.9 est :

$$\hat{\theta}_1 = \frac{\sum\limits_{t \in [\![1,\ \tau[\![} y_t}{\tau - 1}, \qquad \hat{\theta}_2 = \frac{\sum\limits_{t \in [\![\tau,\ n]\!]} y_t}{n - \tau + 1} \tag{4.10}$$

Démonstration. Soit une suite de variables aléatoires $(Y_i)_{i \in [\![1,n]\!]}$ issus d'une distribution de Bernoulli de paramètre θ :

$$\forall i \in [\![1, n]\!],\ Y_i \sim \mathcal{B}(\theta) \tag{4.11}$$

La fonction de vraisemblance est :

$$f(y_1, \ldots, y_n|\theta) = P(Y_1 = y_1, \ldots, Y_n = y_n \mid \theta) \tag{4.12}$$
$$= \theta^{y_1}(1-\theta)^{1-y_1} \ldots \theta^{y_n}(1-\theta)^{1-y_n} \tag{4.13}$$
$$= \theta^{\sum y_i}(1-\theta)^{\sum(1-y_i)} \tag{4.14}$$
$$= \theta^{\sum y_i}(1-\theta)^{n-\sum y_i} \tag{4.15}$$

On passe au logarithme pour faciliter les calculs tout en préservant la monotonie de la fonction :

$$\ln f = (\sum y_i) \ln \theta + (n - \sum y_i) \ln (1-\theta) \tag{4.16}$$

Et enfin, dériver pour trouver la valeur de θ pour laquelle la fonction est maximum (sa monotonie fait que ce point d'inflexion correspond à un maximum) :

$$\frac{\mathrm{d} \ln f}{\mathrm{d}\theta} = 0 \iff \frac{\sum y_i}{\hat{\theta}} - \frac{(n - \sum y_i)}{1-\hat{\theta}} = 0 \tag{4.17}$$
$$\iff \sum y_i - \hat{\theta} \sum y_i = n\hat{\theta} - \hat{\theta} \sum y_i \tag{4.18}$$
$$\iff \hat{\theta} = \frac{\sum y_i}{n} \tag{4.19}$$

□

Ainsi, en reprenant 4.6 et 4.8 avec ce dernier résultat nous pouvons calculer la vraisemblance qu'un τ donné corresponde à une rupture entre les deux périodes $[\![1, \tau[\![$ et $[\![\tau, n]\!]$ et extraire le $\hat{\tau}$ qui maximise cette vraisemblance :

$$\hat{\tau} = \underset{\tau}{\operatorname{argmax}} \max_{\theta_1, \theta_2} P(y_1, \ldots, y_n \mid \tau, \theta_1, \theta_2) \tag{4.20}$$
$$= \underset{\tau}{\operatorname{argmax}} \prod_{t \in [\![1,\ \tau[\![} p(y_t \mid \hat{\theta}_1) \prod_{t \in [\![\tau,\ n]\!]} p(y_t \mid \hat{\theta}_2) \tag{4.21}$$

Décision sur un test statistique

Nous proposons maintenant d'améliorer l'approche précédente en encadrant ce problème comme un problème de test et de décision statistique comme le décrit **(author?)** [13]. L'énoncé reste toujours le suivant : Étant donné une séquence d'observations $y_{1:n} = (y_1, \ldots, y_n)$ nous cherchons la présence d'un point de rupture τ correspondant à un changement dans les paramètres statistiques des observations. Par contre, à chaque pas de temps t, l'observation y_t est issue d'une distribution de Bernoulli de paramètre θ_t.

Nous posons ainsi les deux hypothèses suivantes :

$$H_0 : \quad \theta_1 = \theta_2 = \theta_3 = \cdots = \theta_n \tag{4.22}$$
$$H_1 : \quad \theta_1 = \theta_2 = \theta_3 = \cdots = \theta_{\tau-1} \neq \theta_\tau = \theta_{\tau+1} = \cdots = \theta_n \tag{4.23}$$

H_0 est l'hypothèse nulle qui suppose qu'il n'y a pas de point de rupture sur la séquence étudiée. H_1 est l'hypothèse alternative qui suppose qu'il y a un point de rupture quand $t = \tau$.

L'algorithme décrit par **(author?)** [53] sera basé sur le ratio du logarithme de la vraisemblance $\mathcal{L}$. La vraisemblance d'une hypothèse $\mathcal{L}(H_x)$ se calcule de la même façon que la vraisemblance d'une séquence d'observation :

$$\mathcal{L}(H_0) = P(Y \mid H_0) = \prod_{t=1}^{n} p(y_t \mid \theta_1) \tag{4.24}$$

Le paramètre θ_1 reste le même pour chaque pas de temps, car H_0 est l'hypothèse qui suppose que les paramètres pour toutes les tranches sont égaux (4.22).

De même pour l'hypothèse alternative :

$$\mathcal{L}(H_1) = P(Y \mid H_1) = \prod_{t=1}^{\tau-1} p(y_t \mid \theta_1) \prod_{t=\tau}^{n} p(y_t \mid \theta_\tau) \tag{4.25}$$

Le ratio du logarithme est donné pour tout $\tau \in [\![1, n]\!]$:

$$\mathcal{R}(\tau) = \log\left(\frac{\mathcal{L}_{H_1}}{\mathcal{L}_{H_0}}\right) = \sum_{t=1}^{\tau-1} \log p(y_t \mid \theta_1) + \sum_{t=\tau}^{n} \log p(y_t \mid \theta_\tau) - \sum_{t=1}^{n} \log p(y_t \mid \theta_1) \tag{4.26}$$

Celui-ci permet de calculer le maximum de log vraisemblance :

$$G = \max_{\tau \in [\![1,n]\!]} \mathcal{R}(\tau) \tag{4.27}$$

et l'estimation du maximum de vraisemblance de la valeur du point de rupture $\hat{\tau}$ dans le cas où l'hypothèse nulle est rejetée :

$$\hat{\tau} = \operatorname*{argmax}_{\tau \in [\![1,n]\!]} \mathcal{R}(\tau) \tag{4.28}$$

On note que la méthode suivie à la section précédente est quasiment la même que celle-ci. La seule différence étant l'hypothèse de présence d'un point de rupture faisant rejet systématique de l'hypothèse nulle.

Pour accepter ou rejeter l'hypothèse nulle, il faut se fier à la valeur de G. L'hypothèse est rejetée pour des valeurs assez grandes. C.-à-d. que la meilleure vraisemblance de l'hypothèse alternative est assez grande comparé à l'hypothèse nulle. En d'autres termes, il existe une valeur critique λ^* pour laquelle l'hypothèse nulle est rejetée :

$$G = \mathcal{R}(\hat{\tau}) > \lambda^* \tag{4.29}$$

Il existe plusieurs critères pour fixer cette valeur critique :

BIC (*Bayesian Information Criterion*) $\lambda^* = k \log n$

MBIC (*Modified Bayesian Information Criterion*) $\lambda^* = (k+1)\log n + \log(\tau) + \log(n - \tau + 1)$

AIC (*Akaike Information Criterion*) $\lambda^* = 2k$

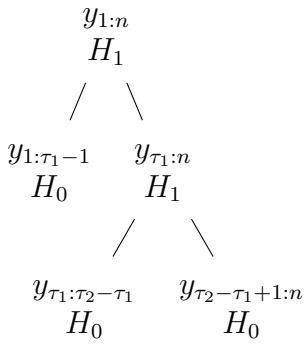
Hannan-Quinn $\lambda^* = 2k \log(\log n)$

k est le nombre de paramètres ajouté au modèle suite à l'insertion d'un point de rupture. Par exemple $k = 1$ s'il n'y a qu'un point de changement entre θ_0 et θ_τ. $k = 2$ si l'on ajoute un second point de rupture $\tau\prime$ qui fait qu'on a trois segments $y_{1:\tau-1}$, $y_{\tau:\tau\prime-1}$ et $y_{\tau\prime:n}$ avec les paramètres θ_1, θ_2 et θ_3.

Inférence de plusieurs points de rupture avec de la segmentation binaire

Jusqu'ici nous avons vu une méthode permettant d'évaluer si une seule segmentation est possible sur une série d'observations et séparer celle-ci en deux séries distinctes en trouvant un point de rupture. La généralisation à plusieurs points de rupture peut se faire à l'aide de plusieurs méthodes, parmi les plus populaires, on retrouve la méthode proposée par **(author?)** [85] qui se base sur un algorithme dichotomique appelé «segmentation binaire».

Le principe consiste à appliquer la méthode de détection d'un point de rupture sur un segment d'observations et d'ensuite, récursivement, de la réappliquer sur les deux segments résultants. L'arrêt est conditionné sur la validation de l'hypothèse nulle sur toutes les feuilles de l'arbre binaire généré :



Dans cet exemple, la série d'observations $y_{1:n}$ se découpe en trois séries $y_{1:\tau_1-1}$, $y_{\tau_1:\tau_2-\tau_1}$ et $y_{\tau_2-\tau_1+1:n}$, car elle admet deux points de rupture τ_1 et τ_2.

Le pseudo-code 1 décrit cet algorithme récursif. La fonction `segmentation` est une fonction qui effectue une segmentation à un seul point de rupture. L'avantage de cet algorithme est sa rapidité, son inconvénient est qu'il n'est pas optimal. Cependant, de nombreuses études montrent que d'un de vue opérationnel, cet algorithme donne de très bons résultats.

4.3.2 Modélisation bayésienne

L'approche fréquentiste se focalise sur l'estimation des valeurs des paramètres, par exemple ici les probabilités d'observation d'un mouvement et l'instant de rupture. Elle considère que ces paramètres sont fixés et que les données observées sont aléatoires.

Algorithm 1 Estimation des points de rupture par segmentation binaire

```
1: procedure SEGMENTATION BINAIRE(y_{i:j})
2:     τ ← Segmentation(y_{i:j})
3:     if τ = ∅ then
4:         return ∅
5:     else
6:         T_1 ← Segmentation Binaire(y_{i:τ−1})
7:         T_2 ← Segmentation Binaire(y_{τ:j})
8:         return {τ} ∪ T_1 ∪ T_2
9:     end if
10: end procedure
```

Mais si, par exemple, nous voulons en tirer des tendances ou effectuer des comparaisons, plutôt qu'étudier les processus sous-jacents, il nous faudra adopter une approche bayésienne.

D'un point de vue bayésien, nous admettons que les paramètres à estimer sont issus de lois de probabilités *a priori*. En reprenant l'exemple précédent, nous admettons maintenant que les observations y_i suivent une loi de Bernoulli $\mathcal{B}(\theta)$ dont le paramètre θ dépend de τ :

$$y_i = \begin{cases} \mathcal{B}(\theta_1) & \text{pour } t < \tau \\ \mathcal{B}(\theta_2) & \text{pour } t \geq \tau \end{cases} \tag{4.30}$$

Les seules quantités connues étant les données observées y_i, sachant que nous voulons inférer les inconnus θ_1, θ_2 et τ, il faut calculer la fonction de vraisemblance $P(Y \mid \theta_1, \theta_2, \tau)$ et la distribution *a priori* sur les paramètres à estimer $P(\theta_1, \theta_2, \tau)$, car :

$$P(\theta_1, \theta_2, \tau \mid Y) = \frac{P(Y \mid \theta_1, \theta_2, \tau) P(\theta_1, \theta_2, \tau)}{P(Y)} \tag{4.31}$$

$P(\theta_1, \theta_2, \tau \mid Y)$ étant la distribution *a posteriori*, la quantité à laquelle nous nous intéressons ici.

$P(Y \mid \theta_1, \theta_2, \tau)$ est la fonction de vraisemblance. C'est la vraisemblance d'observation des données Y étant donné le choix des paramètres du modèle.

$P(\theta_1, \theta_2, \tau)$ est la distribution *a priori* sur les paramètres du modèle.

$P(Y)$ est l'évidence sur les observations. Elle mesure la qualité du modèle, c.-à-d. si les choix de distributions supposés pour les observations sont bons ou bien douteux. Elle est typiquement difficile à évaluer analytiquement, mais peut être ignoré lors d'une maximisation en tant que constante.

Nous utilisons des *a priori* non informatifs (non intégrable à 1) pour les paramètres θ_1, θ_2 et τ. En particulier pour θ_1 et θ_2 nous utilisons l'*a priori* de Jeffery pour des expériences de Bernoulli. Celui-ci est une loi Beta de paramètres $\alpha = \beta = \frac{1}{2}$. Et pour τ une constante k.

$$P(\theta_1) = \frac{\theta_1^{\alpha-1}(1-\theta_1)^{\beta-1}}{B(\alpha,\beta)} \tag{4.32}$$

$$P(\theta_2) = \frac{\theta_2^{\alpha-1}(1-\theta_2)^{\beta-1}}{B(\alpha,\beta)} \tag{4.33}$$

$$P(\tau) = k \tag{4.34}$$

L'objectif est de retrouver une expression analytique de la distribution *a posteriori* du point de rupture τ : $P(\tau \mid d)$. Pour se faire, il faut marginaliser sur les paramètres θ_1 et θ_2 :

$$P(\tau \mid Y) = \int_0^1 \mathrm{d}\theta_1 \int_0^1 \mathrm{d}\theta_2 \; P(\theta_1, \theta_2, \tau \mid Y) \tag{4.35}$$

$$= \int_0^1 \mathrm{d}\theta_1 \int_0^1 \mathrm{d}\theta_2 \; \frac{P(Y \mid \theta_1, \theta_2, \tau)P(\theta_1)P(\theta_2)P(\tau)}{P(Y)} \tag{4.36}$$

$$\propto \int_0^1 \mathrm{d}\theta_1 \int_0^1 \mathrm{d}\theta_2 \; P(Y \mid \theta_1, \theta_2, \tau)P(\theta_1)P(\theta_2) \tag{4.37}$$

Il ne reste plus qu'à donner une expression pour la fonction de vraisemblance :

$$P(Y \mid \theta_1, \theta_2, \tau) = \prod_{i=1}^{\tau} P(y_i \mid \theta_1) \prod_{i=\tau+1}^{N} P(y_i \mid \theta_2) \tag{4.38}$$

$$= \prod_{i=1}^{\tau} \theta_1^{y_i}(1-\theta_1)^{1-y_i} \prod_{j=\tau+1}^{N} \theta_2^{y_j}(1-\theta_2)^{1-y_j} \tag{4.39}$$

$$= \theta_1^{\sum_{i=1}^{\tau} y_i}(1-\theta_1)^{\tau-\sum_{i=1}^{\tau} y_i} \theta_2^{\sum_{j=\tau+1}^{N} y_j}(1-\theta_2)^{N-\tau-\sum_{j=\tau+1}^{N} y_j} \tag{4.40}$$

Ainsi en injectant 4.32, 4.33 et 4.40 dans 4.37 on obtient :

$$P(\tau \mid Y) \propto \int \mathrm{d}\theta_1 \int \mathrm{d}\theta_2 \; \theta_1^{\alpha-1+\sum_{i=1}^{\tau} y_i}(1-\theta_1)^{\beta-1+\tau-\sum_{i=1}^{\tau} y_i} \theta_2^{\alpha-1+\sum_{j=\tau+1}^{N} y_j}(1-\theta_2)^{\beta-1+N-\tau-\sum_{j=\tau+1}^{N} y_j} \tag{4.41}$$

On reconnait ici un binôme différentiel dont la valeur intégrale est la fonction Beta incomplète en vertu du théorème de Chebyshev :

$$\int \lambda^p(1-\lambda)^q = \mathcal{B}_\lambda(p+1, q+1) \tag{4.42}$$

Nous pouvons ainsi donner une expression analytique pour la distribution *a posteriori* de τ :

$$\boxed{P(\tau \mid Y) \propto \mathcal{B}_{\theta_1}\left(\alpha + \sum_{i=1}^{\tau} y_i, \beta + \tau - \sum_{i=1}^{\tau} y_i\right) \times \mathcal{B}_{\theta_2}\left(\alpha + \sum_{j=\tau+1}^{N} y_j, \beta + N - \tau - \sum_{j=\tau+1}^{N} y_j\right)} \tag{4.43}$$

Et nous pouvons aussi mettre à jour notre *a priori* sur les paramètres θ_1 et θ_2 en posant la formule pour leurs distributions *a posteriori* :

$$P(\theta_1 \mid Y) \propto \int \mathrm{d}\theta_2 \int \mathrm{d}\tau \; \theta_2^{\alpha-1+\sum_{j=\tau+1}^{N} y_j} (1-\theta_2)^{\beta-1+N-\tau-\sum_{j=\tau+1}^{N} y_j} \tag{4.44}$$

$$\propto \mathcal{B}_{\theta_2}\left(\alpha + \sum_{j=\tau+1}^{N} y_j, \beta + N - \tau - \sum_{j=\tau+1}^{N} y_j\right) \tag{4.45}$$

de même :

$$P(\theta_2 \mid Y) \propto \mathcal{B}_{\theta_1}\left(\alpha + \sum_{j=1}^{\tau} y_j, \beta + N - \tau - \sum_{j=1}^{\tau} y_j\right) \tag{4.46}$$

Ceci, nous permet de remarquer la simplicité avec laquelle nous pouvons mettre à jour l'*a priori* d'une distribution Beta, avec son *a posteriori* en n'employant qu'une accumulation des données observées.

Le résultat de l'inférence 4.43 est donc une distribution des τ. Si notre application nécessite de retourner une seule valeur numérique, nous pouvons, par exemple, la déduire de la moyenne, la médiane ou l'argmax de la distribution.

4.4 Approche par classification

4.4.1 Espérance-Maximisation

L'algorithme d'espérance-maximisation est une méthode qui permet d'estimer les valeurs de paramètres de modèles statistiques qui maximisent la vraisemblance ou d'estimer le maximum *a posteriori* quand ces modèles dépendent de variables cachées dites latentes.

Théorie

L'espérance-maximisation se base sur un modèle d'observation illustré par la figure 4.3. La variable observée $\mathcal{Y}$ générée par un modèle p avec un paramètre θ^* suit une variable d'état caché X.

La finalité reste toujours d'arriver à une estimation du maximum de vraisemblance des paramètres étant donné les observations.

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}}\; l(\theta \mid Y) = \underset{\theta}{\operatorname{argmax}}\; \log P(Y \mid \theta) \tag{4.47}$$

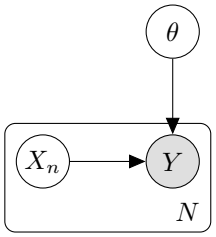


FIGURE 4.3 – Représentation graphique du réseau bayésien du modèle d'Espérance-Maximisation

Dans le cas de ce modèle, nous pouvons aussi étudier les valeurs de la variable latente X. Travailler avec la quantité $l(\theta \mid Y) = \log P(Y \mid \theta)$ peut souvent s'avérer difficile à cause de la marginalisation sur la variable latente X qui ne peut pas sortir du logarithme :

$$l(\theta \mid Y) = \log P(Y \mid \theta) = \log \sum_x P(Y, x \mid \theta) \tag{4.48}$$

L'idée principale de l'algorithme EM est d'itérer sur des estimations de paramètres θ en calculant des bornes inférieures pour $l(\theta \mid Y)$ de plus en plus croissantes et convergente vers un maximum local. Si l'on considère une distribution proxy q, on a avec l'inégalité de Jensen une borne inférieure $\mathcal{L}(q, \theta)$:

$$\begin{aligned} l(\theta \mid Y) &= \log \sum_x P(Y, x \mid \theta) && (4.49)\\ &= \log \sum_x q(x) \frac{P(Y, x \mid \theta)}{q(x)} && (4.50)\\ &\geq \sum_x q(x) \log \frac{P(Y, x \mid \theta)}{q(x)} \equiv \mathcal{L}(q, \theta) && (4.51) \end{aligned}$$

Qui peut être développée ainsi :

$$\begin{aligned} l(\theta \mid Y) \geq \mathcal{L}(q, \theta) &= \sum_x q(x) \log \frac{P(Y, x \mid \theta)}{q(x)} && (4.52)\\ &= \sum_x q(z) \log P(Y, x \mid \theta) - \sum_x q(x) \log q(x) && (4.53)\\ &= E_q[\log P(Y, X \mid \theta)] - E_q[\log q(x)] && (4.54) \end{aligned}$$

Pour EM, on ne considère que les distributions proxys de la forme $q_\vartheta(X) = P(X \mid Y, \vartheta)$ qui correspondent à l'*a posteriori* de la variable cachée.

Il ne reste plus qu'a itérer sur les étapes **E** et **M** de l'algorithme pour converger vers une estimation des paramètres $\hat{\theta}$ du modèle :

E-step Calcul d'une nouvelle borne inférieure :

$$\vartheta_{t+1} = \underset{\vartheta}{\operatorname{argmax}}\ \mathcal{L}(\vartheta, \theta_t) = \theta_t \tag{4.55}$$

M-step Calcul d'une nouvelle estimation des paramètres $\hat{\theta}$ du modèle, en utilisant la nouvelle borne inférieure calculée à l'étape E précédente :

$$\theta_{t+1} = \underset{\theta}{\operatorname{argmax}}\ \mathcal{L}(\vartheta_{t+1}, \theta) = \underset{\theta}{\operatorname{argmax}}\ E_q[\log p(Y, X \mid \theta)] \tag{4.56}$$

le second terme de $\mathcal{L}(\vartheta_{t+1}, \theta)$ disparait de l'argmax, car il ne dépend pas de θ. Le terme restant est communément appelé la fonction auxiliaire ou *Q-function* :

$$Q(\theta_t, \theta) = E_q[\log p(Y, X \mid \theta)] \tag{4.57}$$

Exemple sur des jets de pièces

Supposons une expérience de jet de deux pièces pipées et marquées où l'on effectue 10 jets à la suite pour chaque pièce :

Pièce	Jets
A	PFPFPPPFPF
A	FFPFPFFPPP
B	FFFPFFPFFP
B	PPFFFFPFFP
A	PPFPPPPFPF

TABLE 4.1 – Suites de jet de deux pièces A et B pipés

Dans le scénario détaillé au tableau 4.1, il est trivial d'estimer le biais de chaque pièce par simple décomptage. La pièce A donne 18 Piles, donc un biais de $18/30 = 0.6$ et la pièce B donne 7 Piles donc un biais de $7/20 = 0.35$.

Le problème se complique si les pièces ne sont plus marquées, en d'autres termes que l'on tire au hasard une pièce d'un sac, que l'on effectue 10 jets avec et qu'on la remette dans le sac :

Pièce	Jets
?	PFPFPPPFPF
?	FFPFPFFPPP
?	FFFPFFPFFP
?	PPFFFFPFFP
?	PPFPPPPFPF

TABLE 4.2 – Suites de jet de deux pièces A et B pipés

Dans ce cas, vu que nous n'avons pas d'information sur la pièce que l'on tire du sac et vu qu'elle n'est plus marquée, cette information peut être considérée comme cachée ou

latente. Par conséquent, on ne peut pas simplement calculer le biais pour chaque pièce car on ignore ses jets correspondants.

L'algorithme d'espérance-maximisation commence par une estimation initiale des valeurs de paramètres du modèle (les probabilités d'avoir Pile). Il permet d'en déduire l'espérance du nombre de piles pour chacune des classes. Celle-ci se calcule par l'intermédiaire d'une estimation de la vraisemblance que chaque séquence de jet provienne d'une pièce donnée. Ensuite, cette espérance du nombre de piles permet de recalculer de nouvelles valeurs en estimant les paramètres du modèle.

Par exemple, si nous reprenons l'exemple illustré au tableau 4.2, nous posons :

— $\theta_A = 0.7$ et $\theta_B = 0.4$ les valeurs initiales pour les paramètres du modèle

— $Y_1 = PFPFPPPFPF$ la première séquence observée

Nous pouvons alors calculer la probabilité d'observer la séquence Y_1 sachant que la pièce qui l'a produite est A ou B via la formule de la loi binomiale :

$$P(Y_1 \mid A, \theta_A) = \frac{10!}{4!6!} 0.7^6 0.3^4 \tag{4.58}$$

$$P(Y_1 \mid B, \theta_B) = \frac{10!}{4!6!} 0.4^6 0.6^4 \tag{4.59}$$

Qui est aussi en d'autres termes la vraisemblance de la séquence Y_1 étant donné la pièce A ou B.

Ainsi, nous pouvons calculer la probabilité *a posteriori* que la pièce jetée soit A ou B sachant les observations Y_1 grâce au théorème de Bayes :

$$P(A \mid Y_1) = \frac{P(Y_1 \mid A)P(A)}{P(Y_1 \mid A)P(A) + P(Y_1 \mid B)P(B)} \tag{4.60}$$

Sachant que $P(A) = P(B) = 0.5$, vu que l'on choisit aléatoirement dans un sac avec quelle pièce effectuer les 10 jets, nous pouvons simplifier et calculer les probabilités *a posteriori* d'avoir effectué le jet avec la pièce A ou B :

$$P(A \mid Y_1) = \frac{P(Y_1 \mid A)}{P(Y_1 \mid A) + P(Y_1 \mid B)} = \frac{0.7^6 0.3^4}{0.7^6 0.3^4 + 0.4^6 0.6^4} = 0.64 \tag{4.61}$$

$$P(B \mid Y_1) = \frac{P(Y_1 \mid B)}{P(Y_1 \mid A) + P(Y_1 \mid B)} = \frac{0.4^6 0.6^4}{0.7^6 0.3^4 + 0.4^6 0.6^4} = 0.36 \tag{4.62}$$

Le tableau 4.3 contient le calcul pour le reste des jets. Ceci constitue l'étape E dite d'estimation (ou *expectation* en anglais) de l'algorithme.

Maintenant, pour l'étape de maximisation, nous cherchons à mettre à jour les valeurs des paramètres qui maximisent l'évidence que l'on a sur la vraisemblance des observations. On commence d'abord par calculer l'espérance attendue des Piles qui est le produit du nombre de Piles de chaque jet par la probabilité que ce fût un jet de la pièce choisie. En d'autres termes, c'est l'espérance de la valeur de la pièce (pièce A : $X_n = A$ ou pièce B : $X_n = B$) étant donné les observations de Piles Y_i $E[X_n|Y_i]$:

Jets	$P(A\|Y_i)$	$P(B\|Y_i)$
PFPFPPPFPF	0.64	0.36
FFPFPFFPPP	0.34	0.66
FFFPFFPFFP	0.04	0.96
PPFFFFPFFP	0.13	0.87
PPFPPPPFPF	0.86	0.14

TABLE 4.3 – Probabilité *a posteriori* que les jets proviennent de la pièce A ou B

Jets	$P(A\|Y_i)$	$P(B\|Y_i)$	$E[A\|Y_i]$	$E[B\|Y_i]$
PFPFPPPFPF	0.64	0.36	3.84	2.16
FFPFPFFPPP	0.34	0.66	1.7	3.3
FFFPFFPFFP	0.04	0.96	0.12	2.88
PPFFFFPFFP	0.13	0.87	0.52	3.48
PPFPPPPFPF	0.86	0.14	6.02	0.98

TABLE 4.4 – Espérance du nombre de piles si la pièce était A ou B

Il ne reste plus qu'à calculer les nouvelles valeurs des paramètres. On divise l'espérance du nombre de Piles attendu pour chaque pièce par l'espérance du nombre de jets de celle-ci :

$$\theta_A^1 = \frac{3.84 + 1.7 + 0.12 + 0.52 + 6.02}{10 \times (0.64 + 0.34 + 0.04 + 0.13 + 0.86)} = 0.61 \tag{4.63}$$

$$\theta_B^1 = \frac{2.16 + 3.3 + 2.88 + 3.48 + 0.98}{10 \times (0.36 + 0.66 + 0.96 + 0.87 + 0.14)} = 0.43 \tag{4.64}$$

Il suffit ensuite d'enchaîner avec ces nouvelles valeurs les étapes E et M jusqu'à convergence de l'algorithme.

Cette approche permet ainsi de classifier des séquences d'observation. L'application à une série temporelle nécessite d'établir une certaine taille de fenêtre pour les observations est donc d'appliquer l'algorithme sur celles-ci. D'autres techniques de comparaison entre différentes tailles de fenêtres et chevauchement peuvent être appliquées pour améliorer et affiner les résultats. Cette méthode présente potentiel de pouvoir segmenter et de classifier en même temps une suite d'observations.

4.5 Conclusion

L'une des principales problématiques en analyse de séries temporelles est la détection de points de rupture. Les points de rupture représentent des moments clés lors desquels les données de la série changent de forme. Cela peut se traduire par un changement de

la moyenne des observations, leurs variances, ou toute autre grandeur statistique mesurable. Étant donnée la variété de domaines d'études, la littérature regorge de méthodes et d'applications sur l'analyse de point de rupture suivant divers paramètres et modèles. Nous avons commencé par présenter le formalisme le plus répandu. Ensuite, nous avons examiné les principales catégories des méthodes : les algorithmes qui peuvent analyser en temps réel ou non, ceux capables d'analyser des données multivariées ou limité à des données univariées, ainsi que ceux basés sur des mesures statistiques, modèles probabilistes ou sur des algorithmes de regroupement.

Nous avons présenté une approche statistique basée sur un calcul de l'estimateur du maximum de vraisemblance et une décision statistique. Nous l'avons appliqué à un problème simple où l'on cherche un seul point de rupture.

Ensuite, nous avons développé une approche bayésienne où nous calculons analytiquement l'expression de la distribution *a posteriori* du point de rupture.

Et enfin nous avons exploré une autre approche différente par une méthode de classification basée sur l'algorithme d'espérance-maximisation.

5

Inférence bayésienne pour la reconnaissance des périodes de sommeil

Sommaire

5.1 Introduction

Les périodes de sommeil d'une personne sont de bons indicateurs sur l'état de santé de celle-ci. Selon une étude de l'Inserm[45], une personne sur trois serait concernée par un trouble du sommeil en France. Un mauvais sommeil augmente le risque de maladies cardiovasculaires, d'obésité, de diabète, de cancer et d'accidents. Une bonne nuit de sommeil est composée d'entre trois et cinq cycles circadiens de 90 minutes ce qui correspond à une durée de sommeil entre 4h30 et 7h30 en moyenne, d'où l'importance du suivi de cet indicateur.

La méthode proposée dans ce chapitre repose sur le placement de capteurs ambiants à domicile, en particulier des détecteurs de mouvement ou capteurs pouvant générer des données sémantiquement équivalentes à de l'information sur de la présence. Ces capteurs télétransmettent des données révélant l'activité de la personne suivie vers un centre de données certifié. Ces données sont ensuite interprétées pour en extraire, entre autres, les activités de la vie quotidienne. Ici, nous nous concentrons sur l'évaluation du sommeil.

Nous présentons une méthode d'inférence peu coûteuse sur le plan computationnel pouvant s'appliquer à toutes sortes de configurations de capteurs ambiants. Cependant, les mêmes techniques peuvent être utilisées pour d'autres types d'activités.

5.2 État de l'art

5.2.1 Somnologie et capteurs

Dans le domaine médical, l'étude du sommeil est souvent associée à un examen médical dans un environnement contrôlé suivant un protocole assez précis connu sous le nom de *polysomnographie*. D'un autre côté, le marché de l'électronique grand public voit un afflux de dispositifs et de solutions de suivi de la santé physique et de *coaching* [39; 75]. Ces produits se présentent sous différentes formes que l'on peut catégoriser en trois axes : le type de capteur, la nature des données collectées et les types d'algorithmes utilisés. Suivant le type d'application visée, le suivi du sommeil peut être implémenté en utilisant différents types de capteurs que l'on peut catégoriser en deux groupes : les capteurs ambiants et les capteurs portés. Les capteurs portés sont les plus répandus de par la démocratisation des montres connectées, des bracelets de suivi d'activités sportives et aussi des «smartphones» [39]. Néanmoins, les solutions fondées sur des capteurs ambiants connectés constituent une approche qui prend de plus en plus d'ampleur [89]. Nous considérons que les capteurs ambiants constituent un bon compromis entre approches cliniques (médicalisées) du sommeil et les approches nécessitant des capteurs portés dont nous voudrions éviter l'usage sur les séniors afin de limiter la gêne, l'encombrement, la recharge et l'oubli du port.

Les solutions à base de capteurs ambiants offrent beaucoup de possibilités en termes de types de données collectées. Les capteurs de mouvement (PIR) sont une référence sûre, car les plus répandus pour le suivi d'activité en intérieur [63]. Ces capteurs embarquent souvent aussi des capteurs de température, luminosité, humidité, etc. [37]. D'autres permettent la capture d'images de profondeur, du squelette, de la silhouette de la personne [33] ou bien même la capture d'images thermique à l'aide d'ondes radio [43].

5.2.2 Base de données

Ce travail exploite la base de données publique «WSU CASAS Datasets» de l'université de l'état de Washington" [26] ainsi que des données réelles recueillies dans le cadre d'expérimentations Diatelic. Le projet CASAS dispose en effet d'une plate-forme expérimentale d'habitat instrumenté dont le but est de collecter des données annotées. Elles sont à la disposition de la communauté de recherche qui s'intéresse aux environnements instrumentés intelligents et à l'assistance médicale à domicile. Nous utilisons dans notre

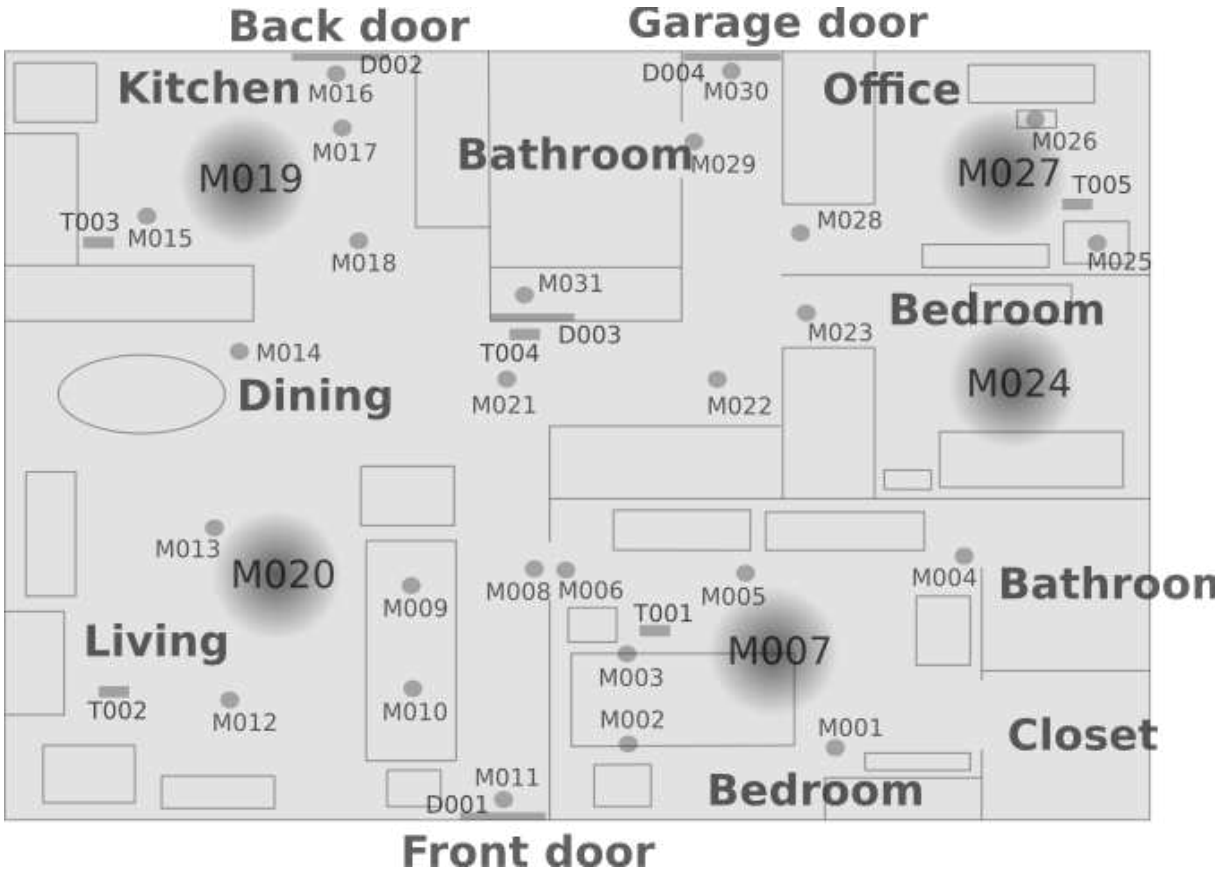


FIGURE 5.1 – Illustration du plan de l'appartement expérimental avec l'emplacement des différents capteurs

projet les données résultant d'une expérimentation dans laquelle un adulte volontaire a vécu pendant une période de 219 jours dans la maison expérimentale du projet CASA dont le plan est illustré sur la figure 5.1. Cette maison est équipée de détecteurs de mouvement binaires (`ON` et `OFF`) avec un horodatage à chaque début et fin des mouvements perçus. L'ensemble des données que nous utilisons contient donc 219 nuits couvrant la période du 2010-11-04 au 2011-06-11 avec plus de 1 713 128 activations.

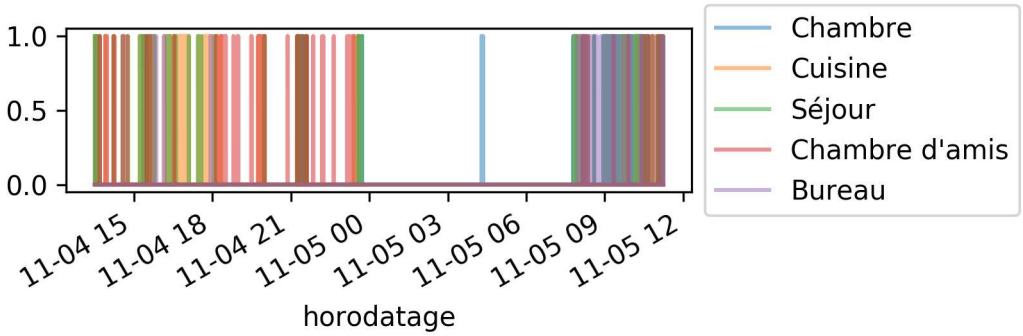


FIGURE 5.2 – Illustration de données de capteurs centrées autour de minuit montrant un clair changement dans le nombre d'activations.

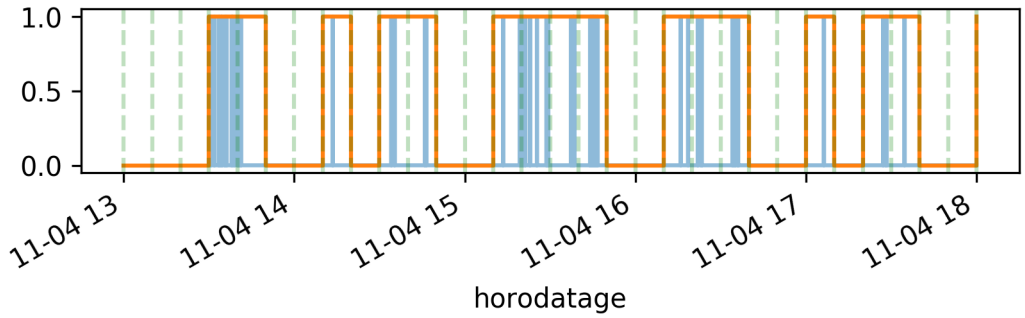


FIGURE 5.3 – Illustration de la discrétisation des données capteurs sur des tranches de 10 minutes. Le signal en bleu étant les données brutes et le signal en orange les données discrétisées.

5.3 Inférence MLE pour deux points de rupture

Comme nous l'avons vu lors du chapitre précédent, l'estimation du maximum de vraisemblance (MLE) est une méthode couramment utilisée pour l'inférence d'un point de rupture. Nous avons aussi vu une méthode permettant d'étendre l'inférence à plusieurs points de rupture quand cela est voulu. Ici, nous nous plaçons dans le cas particulier où nous ne cherchons que deux points de rupture caractérisant une période de sommeil. Nous pouvons simplifier l'approche en découpant nos données en journées débutant de midi d'un jour j et finissant à midi d'un jour j+1. Par conséquent, les journées étudiées sont centrées en minuit. Pour chaque journée, nous supposons la présence d'une période de sommeil entre un τ_s et τ_w qui marquent respectivement le début et la fin du sommeil. Cette période se caractérise par un paramètre θ_s et le reste de la journée par un paramètre θ_w.

Le calcul se base sur la formule 4.21 avec comme seule différence la séparation entre la période dite de sommeil et de réveil :

$$\hat{\tau}_s,\ \hat{\tau}_w = \underset{\tau_s,\ \tau_w}{\operatorname{argmax}} \max_{\theta_s,\theta_w} P(y_1,\dots,y_n \mid \tau_s,\tau_w,\theta_1,\theta_2) \tag{5.1}$$

$$= \underset{\tau}{\operatorname{argmax}} \prod_{t\in[\![\tau_s,\ \tau_w[\![} p(y_t \mid \hat{\theta}_s) \prod_{t\notin[\![\tau_s,\ \tau_w[\![} p(y_t \mid \hat{\theta}_w) \tag{5.2}$$

Ce modèle peut être amélioré en maximisant la distribution *a posteriori*. Ceci revient, grâce à une connaissance experte du domaine d'application, à y ajouter des termes représentant des distributions *a priori* sur les paramètres $\tau_s, \tau_w, \theta_1, \theta_2$. On s'attend, par exemple, à ce que l'heure du lever tourne autour de 8 h 00 du matin avec une durée de sommeil de 8 heures. On choisit un *a priori* sur la durée du sommeil au lieu de l'heure du coucher, car on postule que d'une manière générale les gens auront plus tendance à se réveiller à la même heure que de se coucher à une heure précise. Ceci est dû aux multiples facteurs externes qui arrivent généralement à l'heure du lever, tel qu'une alarme matinale ou tout simplement les rayons de soleil. Ainsi, un *a priori* sur la durée du sommeil permet à l'algorithme d'être plus flexible sur l'inférence de la valeur correcte pour le début du sommeil. Ceci donne le nouveau calcul suivant :

$$\hat{\tau}_s,\ \hat{\tau}_w = \underset{\tau_s,\ \tau_w}{\operatorname{argmax}} \max_{\theta_s,\theta_w} P(\tau_s,\tau_w,\theta_1,\theta_2 \mid y_1,\dots,y_n) \tag{5.3}$$

$$= \underset{\tau_s,\ \tau_w}{\operatorname{argmax}} \max_{\theta_s,\theta_w} P(\tau_s,\tau_w,\theta_1,\theta_2) P(y_1,\dots,y_n \mid \tau_s,\tau_w,\theta_1,\theta_2) \tag{5.4}$$

$$= \underset{\tau_s,\ \tau_w}{\operatorname{argmax}}\ P(\tau_w) P(\tau_w - \tau_s) \prod_{t\in[\![\tau_s,\ \tau_w[\![} p(y_t \mid \hat{\theta}_s) \prod_{t\notin[\![\tau_s,\ \tau_w[\![} p(y_t \mid \hat{\theta}_w) \tag{5.5}$$

5.3.1 Méthode d'évaluation

Évaluation de la distribution de l'erreur des inférences

Nous évaluons notre algorithme à l'aide des annotations de sommeil présentes dans la base de données CASAS. Nous exécutons l'inférence pour chaque nuit puis calculons

l'écart entre les valeurs inférées de T_s et T_w et la vérité terrain :

$$e_{T_x} = T_x - T_x^{annotated}$$

On obtient ainsi la distribution de l'erreur relative sur l'ensemble des 219 nuits disponibles. Nous présenterons les résultats sous forme d'un diagramme "boîtes à moustaches" afin de visualiser la quantité de valeurs aberrantes qui résulteraient de l'algorithme. Ce sont des points en dehors de la fourchette $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$.

Évaluation par test du χ^2 d'homogénéité

La comparaison de deux distributions de variables aléatoires discrètes peut se faire à l'aide de plusieurs méthodes de tests statistiques. Le test du χ^2 de Pearlson par exemple est souvent employé dans la littérature. Il est généralement utilisé soit pour comparer une série d'observations avec une certaine loi de probabilité, soit pour comparer deux séries d'observations entre elles. Le test est un calcul de distance entre les fréquences des réalisations des variables aléatoires et les fréquences attendues suivant la loi des probabilités.

Il est généralement admis que les effectifs des classes des distributions que l'on veut comparer doivent satisfaire une règle où 80% des classes contiennent au moins 5 éléments et le reste doit être non vide.

Il faut donc choisir un nombre adéquat de classes pour respecter ces règles.

Le test se déroule comme suit :

1. Déterminer k le nombre de degrés de liberté du test. Celui-ci est égal au nombre de classes des distributions moins 1 : $k = c - 1$
2. Calculer la statistique T du χ^2 qui correspond à une distance entre les deux distributions :
$$T = \sum_{i=1}^{k} \frac{(n_i - n'_i)^2}{n'_i} \tag{5.6}$$
Avec n_i et n'_i étant respectivement les effectifs de chaque classe des deux distributions comparées
3. Fixer un risque d'erreur correspondant au rejet de l'hypothèse alors qu'elle est vraie. $\alpha = 0.05$ étant souvent une valeur prise par défaut
4. Déterminer le seuil critique que T ne dois pas dépasser relatif à α et k qui est donné par soit une table du χ^2 soit la fonction quantile de celui-ci pour k et $1 - \alpha$
5. En déduire l'homogénéité selon la valeur de T inférieure au seuil critique ou non.

Évaluation par indices de performance d'une classification binaire

Pour quantifier et pouvoir plus facilement comparer les performances de l'algorithme nous proposons de l'évaluer comme un problème de classification binaire, étant donné chaque nuit analysée. Intrinsèquement nous cherchons à classifier chaque tranche $t : S_t = 0$ (Sommeil) ou $S_t = 1$ (Éveil).

Cette classification binaire donne lieu à la matrice de confusion de l'inférence de la période de sommeil (Positif = tranche dans la période, Négatif = tranche en dehors de la période) :

		Classe inférée	
		Éveil ($S_t = 1$)	Sommeil ($S_t = 0$)
Classe annotée	Éveil ($S_t = 1$)	Vrai Négatif (VN)	Faux Positif (FP)
	Sommeil ($S_t = 0$)	Faux Négatif (FN)	Vrai Positif (VP)

Les indicateurs de performance se basent sur quatre critères : justesse (*accuracy*), précision (*precision*), sensibilité et mesure F1 :

$$justesse = \frac{VP + VN}{VP + VN + FP + FN} \tag{5.7}$$

La justesse est la métrique la plus simple à calculer et à comprendre pour un algorithme de classification : Il s'agit de la fréquence des éléments correctement classifiés.

$$precision = \frac{VP}{VP + FP} \tag{5.8}$$

La précision indique la proportion des tranches de sommeil correctement inféré. Typiquement, si cette valeur est faible, l'algorithme a tendance à surestimer la période de sommeil. Réciproquement, si elle est proche de 1, l'algorithme estime la période correctement ou bien la sous-estime selon les valeurs des autres indicateurs.

$$sensibilite = \frac{VP}{VP + FN} \tag{5.9}$$

La sensibilité mesure l'inférence correcte d'une tranche comme étant une tranche de sommeil. Elle exprime à quel point l'algorithme est capable de couvrir toute la période de sommeil.

$$F_1 score = 2 \times \frac{precision \times sensibilite}{precision + sensibilite} \tag{5.10}$$

Le $F_1 score$ est une moyenne harmonique de la précision et de la sensibilité. C'est une autre mesure plus nuancée que la justesse pour valider les performances d'un algorithme. En effet, la valeur de la justesse peut être très haute à cause de l'influence des vrais négatifs, qui ici sont les tranches de réveils correctement inférés. Étant donné que nous étudions les périodes de sommeils qui ont une durée d'un peu moins d'un tiers de la période totale et que nous voulons évaluer l'algorithme sur sa capacité à précisément les trouver, alors, cette mesure se présente comme la plus judicieuse.

Nous disposons, pour chaque nuit, de deux tuples d'indices de la tranche du début et de fin de la période de sommeil : tuple inféré (I_1, I_2) et annoté (A_1, A_2). La figure 5.4 schématise sur une période donnée la classification de la suite des tranches disponibles suivant soit une période de sommeil inféré $[\![I_1, I_2[\![$ soit annotées $[\![A_1, A_2[\![$.

Le calcul de chaque élément de la matrice de confusion se fait ainsi :

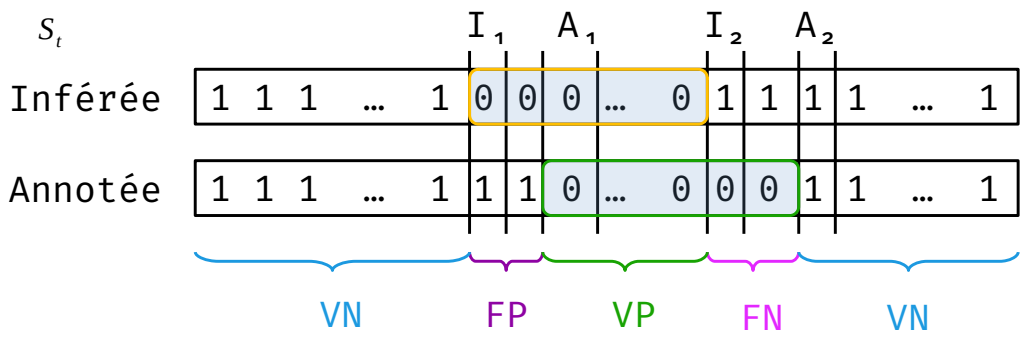


FIGURE 5.4 – Classification binaire des tranches pour une période d'observations

$$VN = \left|[\![0, \min(I_1, A_1)[\![\right| + \left|[\![\max(I_2, A_2), N[\![\right| = \min(I_1, A_1) + N - \max(I_2, A_2) + 2 \tag{5.11}$$

$$VP = \left|[\![\max(I_1, A_1), \min(I_2, A_2)[\![\right| = \min(I_2, A_2) - \max(I_1, A_1) + 1 \tag{5.12}$$

$$FP = (A_1 > I_1) \times \left|A_1 - I_1\right| + (I_2 > A_2) \times \left|A_2 - I_2\right| \tag{5.13}$$

$$FN = (I_1 > A_1) \times \left|A_1 - I_1\right| + (A_2 > I_2) \times \left|A_2 - I_2\right| \tag{5.14}$$

Ainsi, nous pouvons pour chaque nuit calculer les scores de justesse, précision, sensibilité et F1 pour en visualiser la distribution sur les nuits disponibles dans la base de données. Nous pouvons aussi en faire la somme pour en déduire des scores globaux sur toutes les nuits disponibles.

5.3.2 Résultats

La figure 5.5 représente les valeurs de la vraisemblance de l'inférence pour une nuit choisie dans la base CASAS. Les valeurs sont distribuées sur un spectre allant des plus basses en bleu vers les plus hautes en rouge en passant par des niveaux intermédiaires de vert puis de jaune et d'orange. La figure permet de bien visualiser comment s'effectue le choix du segment $[\tau_s, \tau_w]$ selon la valeur maximale de la vraisemblance. Le bandeau de couleur représente la probabilité que chaque segment appartienne à la période de sommeil.

Les résultats de l'inférence sur toute la base de données sont catalogués dans les figures 5.6 et 5.8.

La figure 5.6a montre la distribution de l'erreur sur les heures de coucher qui s'étendent entre - 3 h 50 et - 0 h 50 avec une médiane à 0 h 05 et une moyenne à - 0 h 17. La distribution de l'erreur sur les heures de lever s'étend entre - 0 h 20 et 0 h 10 avec une médiane à - 0 h 10 et une moyenne à - 0 h 01 comme on peut le constater sur la figure 5.6b. Nous observons ainsi que cette approche a plutôt tendance à sous-estimer les heures de coucher. Ceci peut se lire sur la figure 5.8 et le tableau 5.1 où l'on observe une précision

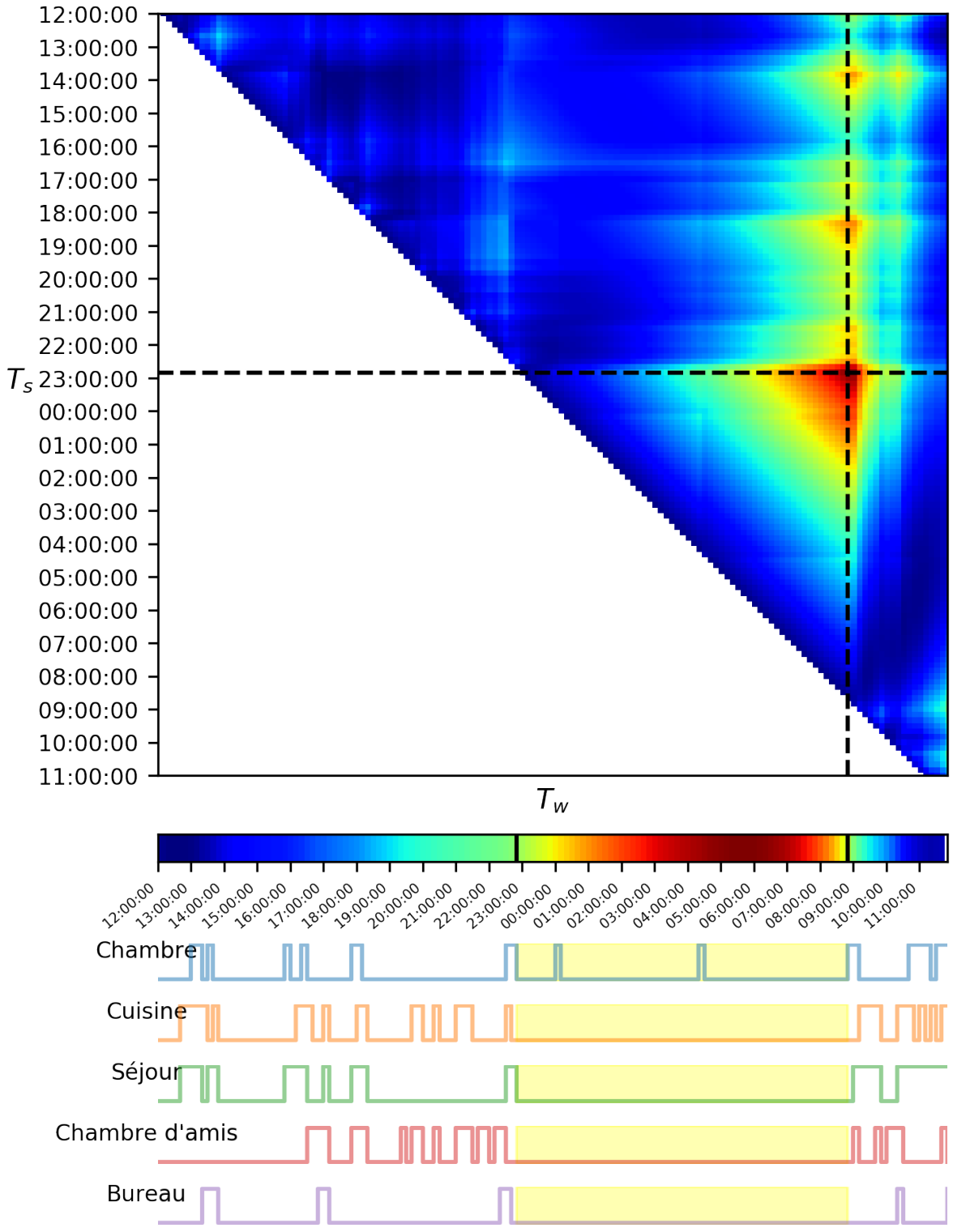
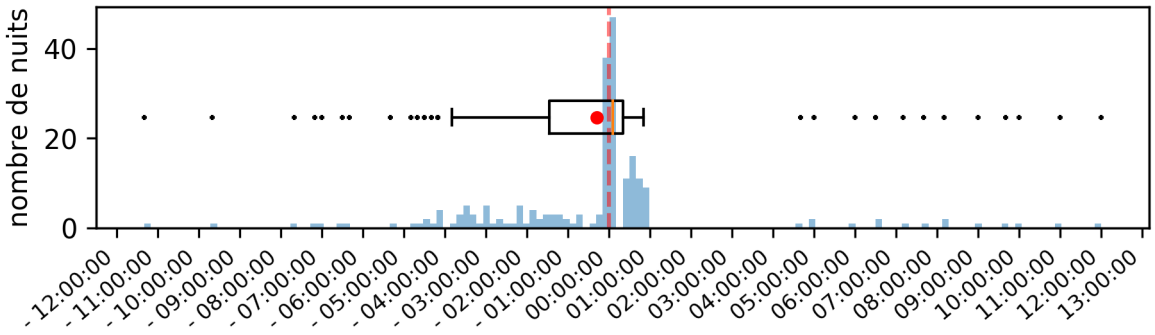
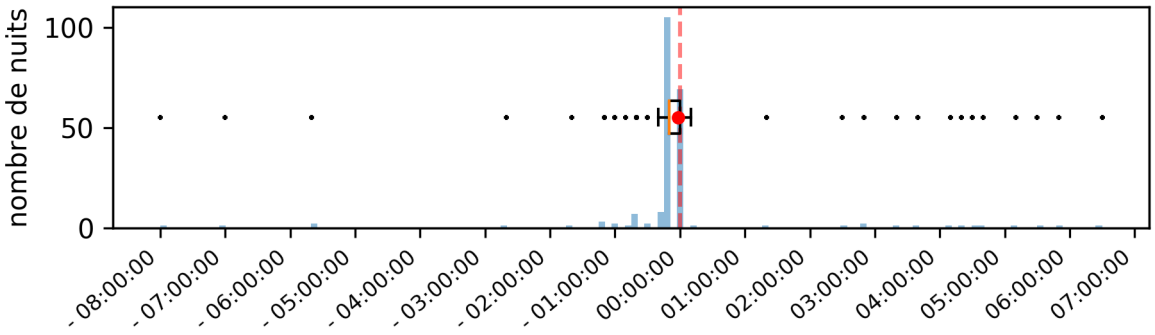


FIGURE 5.5 – Représentation des valeurs de la vraisemblance pour toutes les périodes de sommeil potentiels $[\tau_s, \tau_w]$



(a) L'erreur e_{T_s} sur les heures de coucher inférées



(b) L'erreur e_{T_w} sur les heures de lever inférés

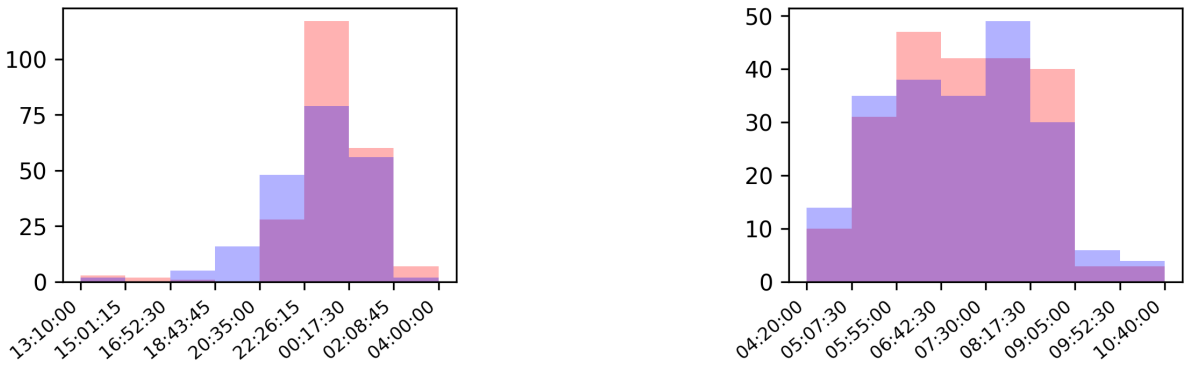
FIGURE 5.6 – Illustration des distributions d'erreurs sur les heure de coucher et de lever e_{T_s} et e_{T_w}

en moyenne plus basse que la sensibilité ($0.85 < 0.89$). On en déduit que l'algorithme a tendance à surestimer les heures de coucher, mais couvre toute la période de sommeil. Nous observons aussi l'utilité du score F1 (moyenne : 0.86, médiane : 0.95) qui permet de mieux refléter ce phénomène par rapport à la justesse (moyenne : 0.91, médiane : 0.97).

Les figures 5.7 représentent une comparaison des distributions des temps de coucher et de lever (en bleu) inférées par le modèle et celle des annotations (en rouge). La superposition des deux est donc en violet.

La statistique du χ^2 des heures de lever est de $T = 12.01$. La loi du χ^2 à 8 degrés de liberté donne la valeur critique pour un risque $\alpha = 0.05$: $P(T < 14.07) = 0.95$. Il suffit donc que la valeur de T soit inférieur à celle-ci pour en conclure que les deux distributions sont homogènes. Ce qui est bien le cas ici : $12.01 < 14.07$. Le test pour les heures de coucher n'est malheureusement pas tout aussi concluant avec un $T = 295.00$. En effet, on remarque une différence d'effectifs principalement dans les trois classes qui se situent entre `18:43:45` et `00:17:30`. Comme il a été montré plus haut, cette différence vient du fait que la méthode sous-estime les heures de coucher. L'allure de la distribution reste tout de même assez cohérente avec celle des annotations.

Les scores globaux (justesse, précision, rappel et F1) ainsi que leurs histogrammes montrent que cette méthode est également conforme et même légèrement meilleure que



(a) Comparaison des heures de coucher

(b) Comparaison des heures de lever

FIGURE 5.7 – Comparaison des distributions des heures de coucher et lever inférées (en bleu) aux annotations (en rouge)

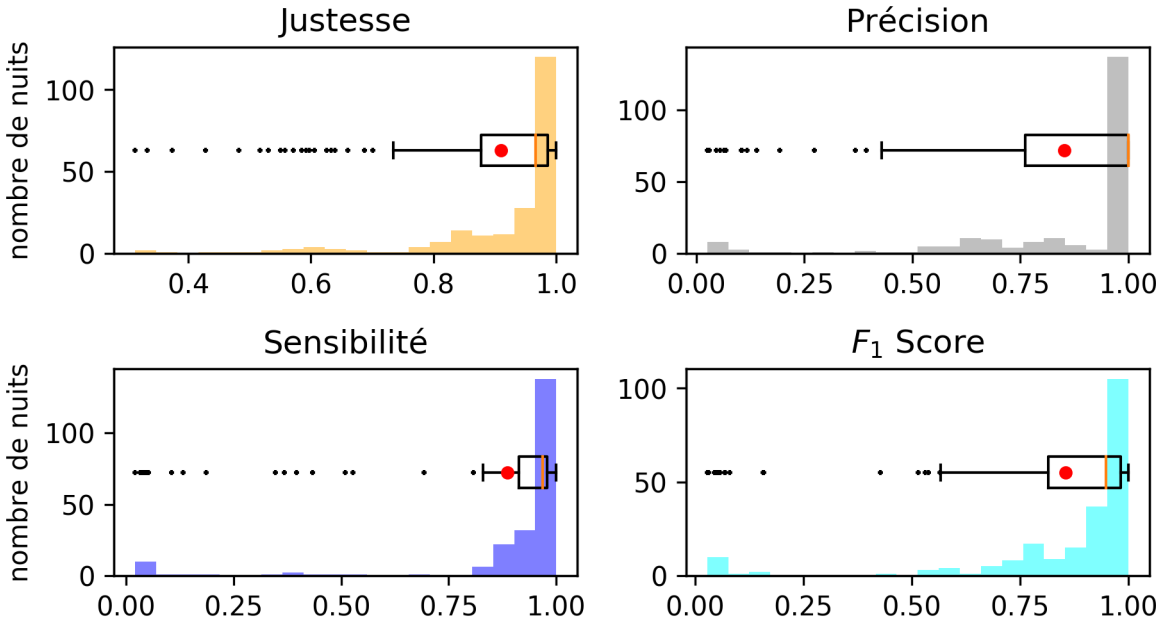


FIGURE 5.8 – Distributions de la justesse, précision, rappel et score F1 sur les nuits de la base de données

	Moyenne	Médiane
Justesse	0.91	0.97
Précision	0.75	1.00
Sensibilité	0.89	0.97
Score F1	0.86	0.95

TABLE 5.1 – Valeurs des indicateurs de performance

les résultats attendus par rapport à une méthode non supervisée[27].

SensibleSleep[27] est aussi une méthode basée sur des événements discrets pour extraire la période du sommeil. La méthode se base sur des événements de *smartphones*. L'inférence est basée sur la méthode de Monte-Carlo par chaîne de Markov (MCMC)[5] qui adapte le modèle aux données observées et sort la distribution *a posteriori* des paramètres. Par contraste, nous montrons qu'en utilisant notre modèle, le calcul direct d'une solution exacte permet d'obtenir des résultats comparables tout en étant peu coûteux sur le plan computationnel, même avec des pas de temps plus petits (notre méthode discrétise la journée par pas de 10 minutes au lieu de 15 minutes).

Un autre avantage de cette méthode est le fait que nous ne supposons pas que $P_s \leq P_w$ dans l'algorithme d'inférence, car ce n'est pas nécessairement vrai tout le temps. Par exemple, il y a de fortes chances que nous observions plus de mouvements dans la chambre à coucher pendant la nuit que pendant la journée. À cause de l'agitation du sommeil ou des visites aux toilettes, soit à cause de capteurs sensibles. Toutefois, cela explique les cas aberrants où l'algorithme trouve un segment de probabilité plus élevée avec plus d'activations de capteurs que la période de sommeil réelle.

5.3.3 Étude pratique de complexité

D'un point de vue pratique, les temps de calcul de nos algorithmes doivent pouvoir passer à l'échelle. En effet le nombre de personnes suivies peut devenir assez grand. Une implémentation naïve et itérative de l'algorithme de référence tournant sur un processeur applicatif classique cadencé à $2.99Ghz$ en plein régime nécessite environ $4.54ms \pm 284\mu s$ pour calculer la période de sommeil sur une journée. En d'autres termes, le temps de calcul est de l'ordre de la milliseconde par journée et par personne suivies.

Comme la complexité de l'algorithme est en $\mathcal{O}(n^3)$ si la fréquence d'échantillonnage est multipliée par 10 pour passer de 10 minutes à 1 minute, le temps de calcul sera multiplié par 10^3 c.-à-d. qu'il augmente de trois ordres de grandeur. En reprenant l'exemple précédent, le calcul passe ainsi de l'ordre de la milliseconde à la seconde, ce que confirment les expérimentations où on trouve un temps de calcul de $1.18s \pm 18.2ms$.

Néanmoins, ce temps peut être considérablement réduit, car le calcul peut être parallélisé. La complexité peut alors être réduite en $\mathcal{O}(\log(n))$ (complexité d'une somme et un max en parallèle) si on dispose de $\mathcal{O}(n^3)$ cœurs ou bien en $\mathcal{O}(\frac{n^3 \log(n)}{P})$ si on ne dispose que de P cœurs.

La limite principale de cet algorithme est qu'il n'est pas possible d'inférer plus d'une seule période de sommeil. L'algorithme est tout de même facilement modifiable pour changer le nombre de segments de sommeil à inférer. Il suffit d'ajouter autant de variables T_{sx}, T_{wx} et P_{sx}, P_{wx} que de nouveaux segments voulus.

Néanmoins, le nombre de segments recherchés doit être donné *a priori* et la complexité de l'algorithme augmente fortement en fonction du nombre de périodes recherchées. Si p est le nombre de segments recherchés, la complexité en en $\mathcal{O}(n^{3+2(p-1)}) = \mathcal{O}(n^{2p+1})$. Par exemple, en termes de temps d'exécution sur une même entrée de période de données capteurs, l'algorithme pour 1 segment de sommeil met un temps de l'ordre de la milliseconde comme vue plus haut, alors qu'un algorithme à 2 segments mettra un temps de

l'ordre de la dixième de seconde. Ce qui fait que la complexité des complexités des algorithmes à p segments de sommeil est exponentielle : $\mathcal{O}(2^p)$. On aurait donc pu envisager d'explorer l'espace des d'algorithmes à p segments en même temps que leurs débuts et fin. La complexité exponentielle de cette méthode décourage à continuer dans cet axe de recherche.

Il faudra se tourner vers d'autres méthodes pour l'inférence de plusieurs segments sur la journée.

5.4 Inférence par segmentation binaire

Comme vu au chapitre précédent, la condition d'arrêt de la méthode de segmentation binaire dépend du rejet de l'hypothèse nulle suivant un certain critère décrivant la quantité d'information contenue dans chaque segment analysé. Vu qu'ici nous ne cherchons qu'à trouver qu'une seule période de sommeil, on pourrait être tenté de simplifier cette condition d'arrêt en s'arrêtant dès la seconde récursion. Ce n'est pas une bonne idée car il n'y a aucune garantie que les trois premiers segments trouvés soient ceux recherchés. Ce n'est cependant pas très grave, la méthode donne le plus souvent plus que 3 segments. Il suffit d'utiliser les a priori sur la période de sommeil pour l'identifier.

5.4.1 Illustration

Il est donc fort possible que l'algorithme trouve d'abord des ruptures plus saillantes, mais correspondantes à d'autres activités que le sommeil.

Pour illustrer notre propos, prenons le jeu de donnée figure 5.9. On y voit clairement 7 segments et 6 points de rupture très marqués (traits rouges sur le chronogramme). L'algorithme 1 peut être facilement modifié pour fournir l'arbre binaire du découpage :

$$[A_0, [A_1, A_2, \dots], [A_k, A_{k+1}, \dots]] \text{ avec } \forall i \in [\![1, k[\![,\ A_i < A_0 < A_k \tag{5.15}$$

Donc avec l'exemple précédent si les points de rupture sont dans cet ordre

$$[\tau_{55}, \tau_{11}, \tau_{112}, \tau_{74}, \tau_{71}, \tau_{130}]$$

correspondants aux heures

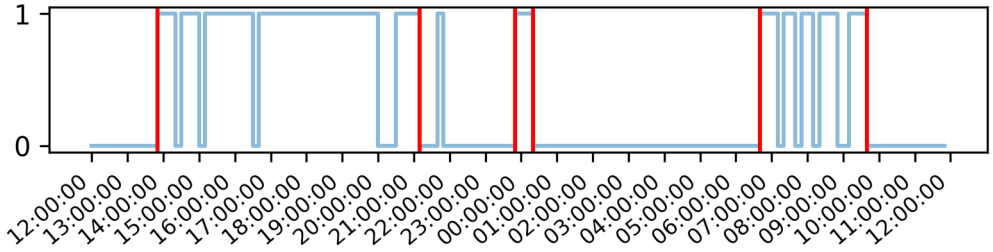


FIGURE 5.9 – Segmentation binaire sur une nuit d'observations

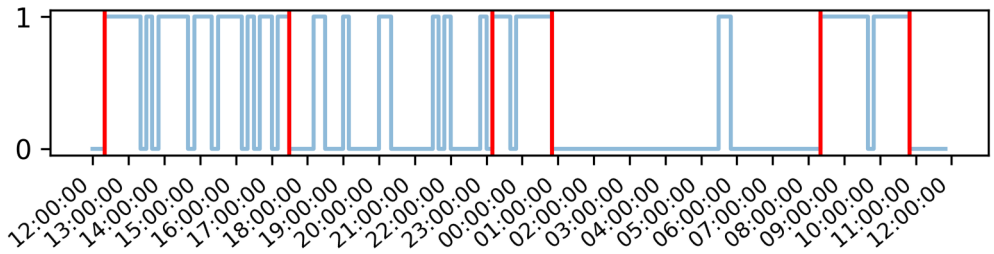
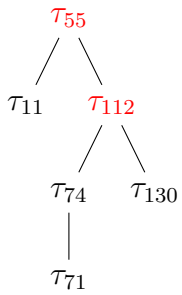


FIGURE 5.10 – Segmentation binaire sur une nuit d'observations

```
['21:10:00', '13:50:00', '06:40:00', '00:20:00', '23:50:00', '09:40:00']
```

Alors, l'arbre binaire de segmentation est ainsi (avec en rouge les points de rupture de ce qu'on pourrait considérer comme de bon début et fin de période de sommeil) :

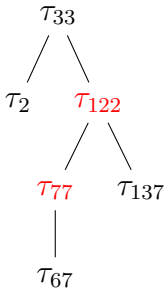


Ici, nous avons un exemple d'un cas ou les points de rupture qui nous intéressent sont bien restreints à deux récursions. Mais il existe aussi d'autres exemples où ce n'est pas le cas comme l'illustre la figure 5.10 où les points de rupture sont :

$$[\tau_{33}, \tau_2, \tau_{122}, \tau_{77}, \tau_{67}, \tau_{137}]$$

```
['17:30:00', '12:20:00', '08:20:00', '00:50:00', '23:10:00', '10:50:00']
```

et l'arbre binaire de segmentation est comme suit :



5.4.2 Méthode d'évaluation

Nous pouvons néanmoins évaluer les performances de cette méthode en définissant l'erreur sur l'inférence du temps de coucher ou de lever comme la plus petite différence entre ce temps annoté et les points de rupture trouvés :

$$e_{T_x} = \text{sign}(\tau_i - T_x^{annotated}) \min_i \left(\left|\tau_i - T_x^{annotated}\right|\right)$$

5.4.3 Résultats et comparaison

Les résultats de l'inférence sont illustrés dans les figures 5.11 et 5.13 juxtaposés avec les résultats précédents.

Nous pouvons voir sur la figure 5.11a que les résultats de la segmentation binaire s'étendent entre - 2 h 10 et - 1 h 20 avec une médiane à 0 h 00 et une moyenne à - 0 h 13. De même que sur la figure 5.11b, ceux des heures de lever s'étendent entre - 0 h 20 et 0 h 10 avec une médiane à - 0 h 10 et une moyenne à - 0 h 09. Nous observons ainsi que l'erreur sur les heures de lever reste autant resserrée autour de 0 que l'approche précédente, tout en ayant une sous-estimation des heures de coucher moins marquée. Les valeurs des indicateurs de performance observée sur la figure 5.13 et tableau 5.2 sont quasiment les mêmes. Ce qui montre que cette méthode permet d'avoir des résultats aussi robustes que ceux de la méthode par modélisation tout en offrant une segmentation plus riche sur toutes les activités possibles de la journée.

Les figures 5.12 représentent une comparaison des distributions des temps de coucher et de lever (en bleu) inférées par cette méthode et celle des annotations (en rouge).

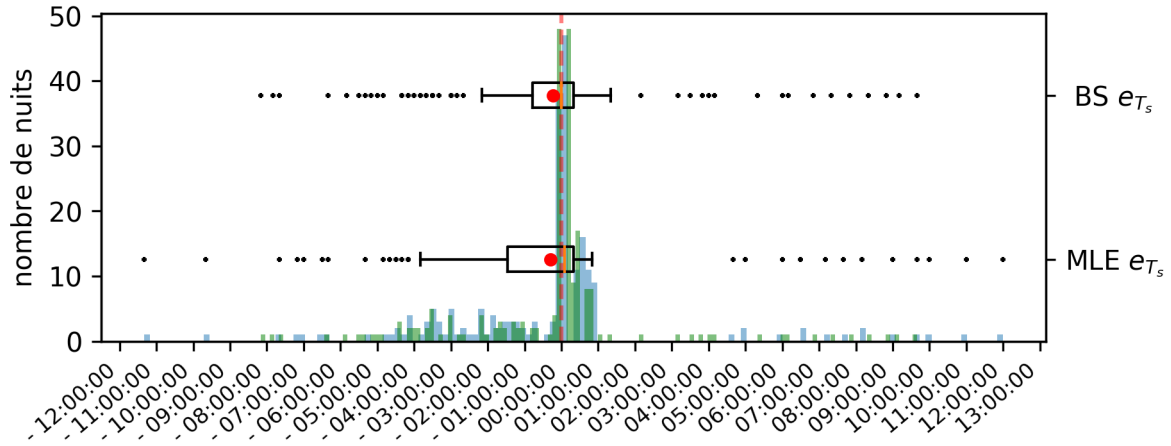
Tout comme les résultats de la méthode MLE, les heures de lever inférées sont bien homogènes aux annotations avec une statistique de $T = 8.97 < 14.07$ et celle des heures de coucher restent tout aussi inconcluantes malgré la ressemblance de l'allure des deux distributions.

	MLE		Segmentation Binaire	
	Moyenne	Médiane	Moyenne	Médiane
Justesse	0.91	0.97	0.92	0.97
Précision	0.75	1.00	0.90	1.00
Sensibilité	0.89	0.97	0.89	0.97
Score F1	0.86	0.95	0.87	0.96

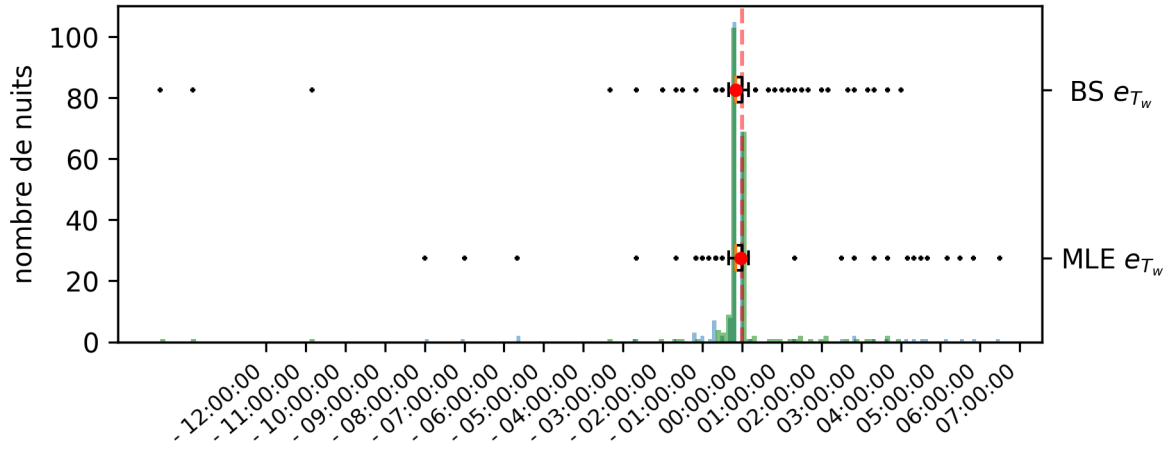
TABLE 5.2 – Valeurs des indicateurs de performance

Les deux algorithmes se basent sur le même principe de mesure de la vraisemblance. L'homogénéité des résultats de l'un et l'autre peut aussi être vérifiée avec un test du χ^2.

Les distributions des figures 5.14 représentent les temps de coucher et de lever des deux méthodes.

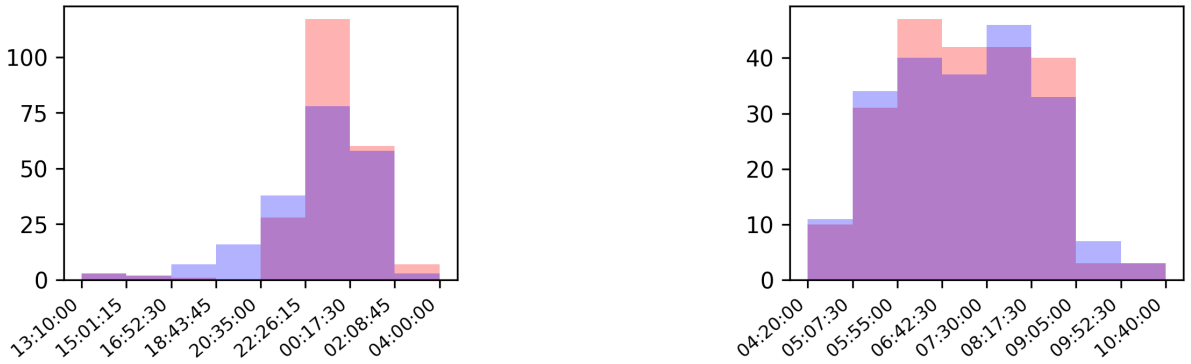


(a) Comparaison de l'erreur e_{T_s} de la segmentation binaire avec la méthode MLE



(b) Comparaison de l'erreur e_{T_w} de la segmentation binaire avec la méthode MLE

FIGURE 5.11 – Illustration des distributions d'erreurs sur les heures de coucher et de lever e_{T_s} et e_{T_w} pour la méthode d'inférence en segmentation binaire et comparaison avec les résultats précédents



(a) Comparaison des heures de coucher

(b) Comparaison des heures de lever

FIGURE 5.12 – Comparaison des distributions des heures de coucher et lever inférées par la méthode de la segmentation binaire (en bleu) aux annotations (en rouge)

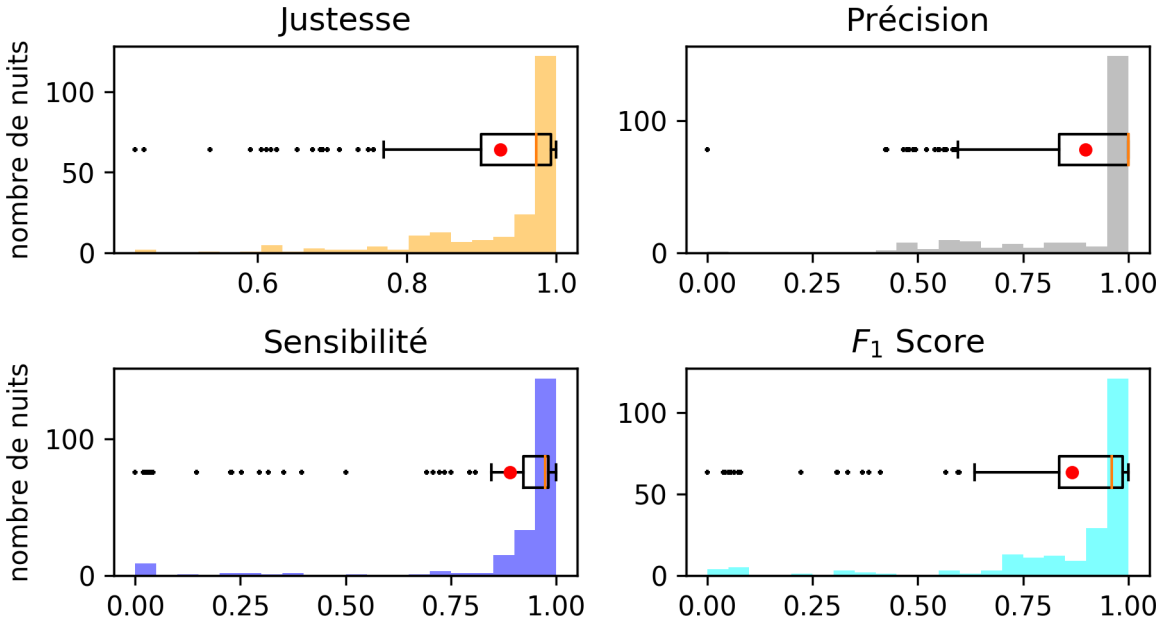
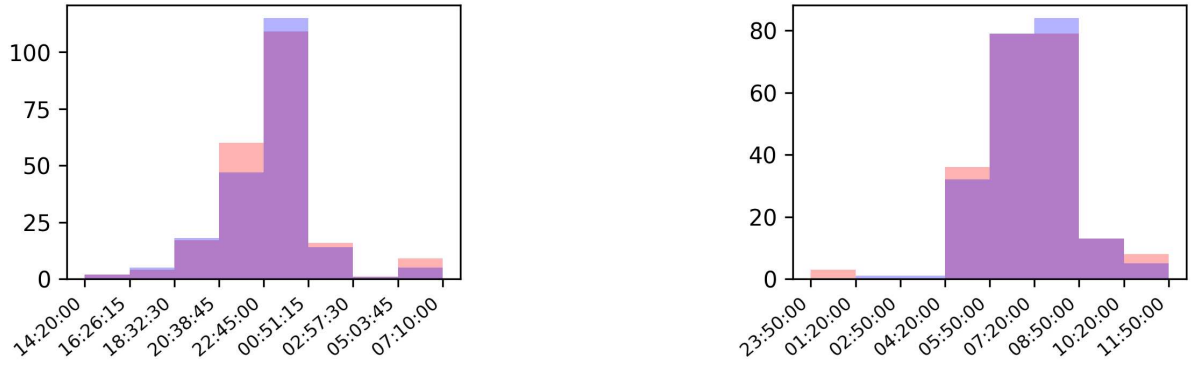


FIGURE 5.13 – Distributions de la justesse, précision, rappel et score F1 sur les nuits de la base de données pour la méthode d'inférence en segmentation binaire



(a) Comparaison des heures de coucher

(b) Comparaison des heures de lever

FIGURE 5.14 – Comparaison des distributions des heures de coucher et lever inférées par les deux méthodes. La méthode MLE en rouge et la méthode de segmentation binaire

Nous pouvons nous intéresser à l'homogénéité de ces deux distributions. Comme nous pouvons le constater sur les figures 5.14, les valeurs des effectifs de chaque classe sont très proches.

Concernant les heures de coucher, la statistique T du χ^2 prend la valeur $T = 7.65$. La loi du χ^2 à 8 degrés de liberté donne la valeur critique pour un risque $\alpha = 0.05$: $P(T < 14.07) = 0.95$. Comme $T = 7.65 < 14.07$ nous pouvons dire que les deux distributions sont bien homogènes. De même pour les heures de lever où $T = 12.28$.

La méthode par segmentation binaire offre donc un très bon outil permettant la segmentation d'une journée d'observation en plusieurs segments qui peuvent par la suite facilement se faire catégoriser par des méthodes de classifications. Nous avons exploré et validé, que la méthode marche efficacement sur un exemple d'activité, ici l'inférence des débuts et fin des activités de sommeil. Une limitation de cette méthode est qu'elle ne peut être appliquée que sur une séquence d'observation complète, car elle n'est pas adaptée pour pouvoir segmenter des séries temporelles en temps réel ou quasi-temps réel. En d'autres termes, comme cela a été expliqué lors du chapitre précédent, c'est une méthode dite *hors ligne*.

5.5 Inférence par programmation dynamique

Si nous voulons explorer des méthodes dites *en ligne* pour une inférence en temps réel, il faudra se tourner vers d'autres types d'algorithmes. **(author?)** [46] présentent un exemple d'algorithme qui se base sur le principe de la programmation dynamique. Cet algorithme permet d'explorer le vaste espace de solutions de partitionnements possibles en une complexité en $\mathcal{O}(N^2)$. Celui-ci se base sur la supposition qu'il existe une fonction g additive permettant d'évaluer une fonction objectif d'un partitionnement P. Ce partitionnement P est un partitionnement d'un interval I qui est un ensemble de M sous

intervalles qui ne se chevauchent pas $P(I) = \{B_m, m \in [\![1, 2, \ldots, M]\!]\}$:

$$V(P) = \sum_{m=1}^{M} g(B_m) \tag{5.16}$$

où $g(B_m)$ est la fonction qui évalue la conformité de chaque sous-intervalle d'observations. L'algorithme étant basé sur le principe de la programmation dynamique son implémentation la plus simple est récursive :

1. Définir $opt(n)$ comme la valeur de la fonction objectif d'un partitionnement optimal P_n^{max} des n premières observations de l'interval d'observations I
2. Définir $opt(0) = 0$
3. Étant donné $opt(j)$ déterminé pour tout $j = 0, 1, \ldots, n$
   — Définir $end(j, n+1) = g(B_{j,n+1})$ avec $B_{j,n+1}$ étant l'union des observations $j, j+1, \ldots, n+1$
   — Calculer $opt(n+1) = \max_{j=1,2,\ldots,n+1} (opt(j-1) + end(j, n+1))$
   — Garder en mémoire la valeur pour laquelle ce maximum est atteint comme $lastchange(n+1)$
4. Répéter l'étape précédente jusqu'à $n + 1 = N$ ou ainsi $opt(N)$ la valeur objectif optimale pour toutes les données observées est calculée
5. Remonter le vecteur *lastchange* pour identifier le début de chaque block du partitionnement optimal comme ceci : $n_1 = lastchange(N)$, $n_2 = lastchange(n_1 - 1)$, ... et ainsi de suite. Le dernier block dans P^{max} sera celui contenant les observations $n_1, n_1 + 1, \ldots, N$ et l'avant-dernier celui contenant $n_2, n_2 + 1, \ldots, n_1 - 1$ etc.

L'implémentation est trivialement immédiate suivant la description ci-dessus et est décrite par le pseudo-code 2.

Le vecteur *opt* évolue comme illustré sur la figure 5.15. À chaque étape i, la meilleure évaluation du découpage est comparée et choisie. L'indice de ce meilleur découpage est enregistré à l'emplacement correspondant dans le vecteur *lastchange* pour pouvoir à la fin de la récursion le remonter pour former la solution.

La seule question qui se pose pour l'implémentation de cet algorithme est celle du choix de la fonction $g(B_m)$ qui effectue l'évaluation de chaque sous-intervalle.

Si on choisit une fonction g basée sur le rapport de vraisemblance $\mathcal{R}$ avec un apriori sur la durée (par exemple $\mathcal{N}(8\ heures, 5\ heures)$), on trouve que l'algorithme retourne des résultats très similaires à l'algorithme de segmentation binaire présenté précédemment :

$$g(B_{j,k}) = \mathcal{R}_j(B_{1,k}) + \log \mathrm{pdf}_{\mathcal{N}}(|B_{1,k}|) \tag{5.17}$$

Les figures 5.16 et 5.17 montrent le découpage de l'algorithme de programmation dynamique sur les mêmes données des nuits montrées sur les figures 5.9 et 5.10.

5.5.1 Résultats et comparaison

Les résultats de l'inférence sont catalogués dans les figures 5.18 et 5.19.

Algorithm 2 Estimation des points de rupture par programmation dynamique

$memoisation \leftarrow \{\}$
$lastchange \leftarrow \{\}$

procedure OPT$(n+1)$
    if `not` $memoisation[n+1]$ **then**
        $candidats \leftarrow \{opt(j-1) + end(j, n+1),\ j \in [\![1, n+1]\!]\}$
        $memoisation[n+1] \leftarrow \max(candidats)$
        $lastchange[n+1] \leftarrow \text{argmax}(candidats)$
    end if
    return $memoisation[n+1]$
end procedure

procedure REMONTER VECTEUR$(lastchange)$
    $n \leftarrow \{\}$
    $courent \leftarrow N+1$
    while $courent - 1 \geq 0$ **do**
        $n \leftarrow n \cup \{lastchange[courent-1]\}$
        $courent \leftarrow lastchange[courent-1]$
    end while
    return n
end procedure

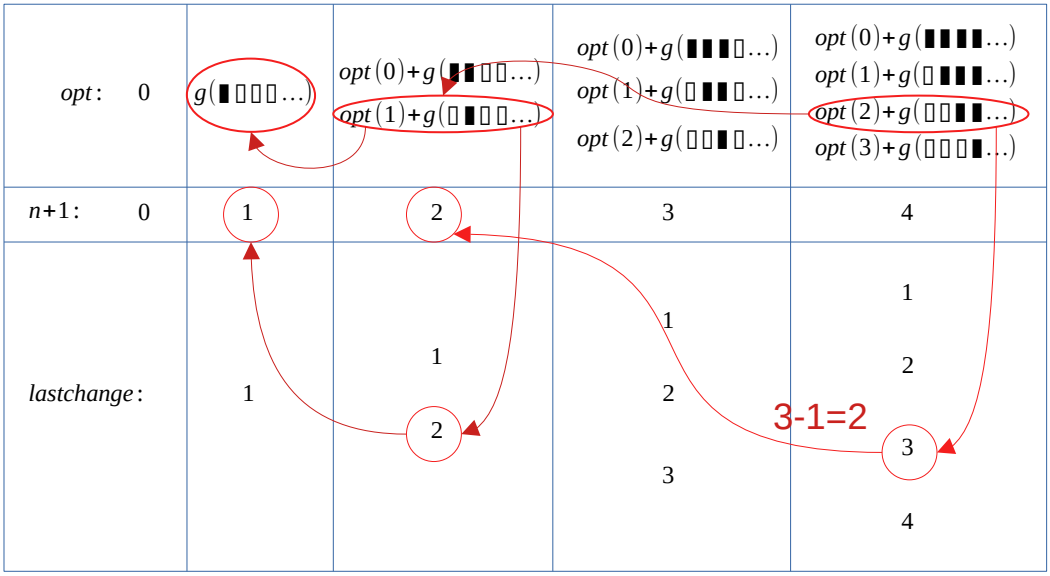


FIGURE 5.15 – Illustration de l'évolution de l'algorithme de programmation dynamique et la remontée de la solution

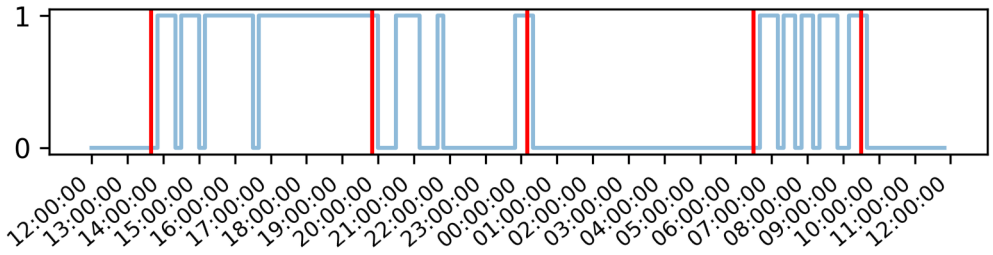


FIGURE 5.16 – Segmentation par programmation dynamique sur une nuit d'observations

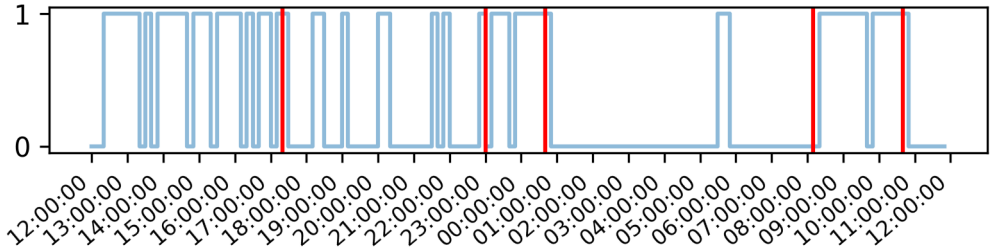


FIGURE 5.17 – Segmentation par programmation dynamique sur une nuit d'observations

La figure 5.18a montre la distribution de l'erreur sur les heures de coucher qui s'étendent entre - 4 h 20 et 2 h 20 avec une médiane à 0 h 10 et une moyenne à - 0 h 45. La distribution de l'erreur sur les heures de lever s'étend entre - 0 h 30 et 0 h 00 avec une médiane à - 0 h 20 et une moyenne à - 0 h 36 comme on peut le constater sur la figure 5.18b. Nous observons ainsi une dégradation de l'inférence des heures de coucher vis-à-vis de l'algorithme de segmentation binaire. L'erreur reste tout de même au même niveau que celle que l'on retrouve sur l'algorithme d'inférence de base. On remarque également que les heures de levers sont aussi légèrement sous-estimées.

Les indicateurs de performance observée sur la figure 5.19 et tableau 5.3 quant à eux restent quasiment inchangés, si ce n'est qu'une légère amélioration des moyennes de la sensibilité et du score F1.

Les figures 5.20 représentent une comparaison des résultats de cette méthode (en bleu) avec les annotations (en rouge). Tout comme les méthodes précédentes, le test du χ^2 reste non concluant pour les heures de lever. Le test ne l'est pas non plus pour les heures de coucher avec une marge assez fine : $T = 27.47 \not< 14.07$. Les allures des deux distributions restent tout de même assez proches.

	MLE		Segmentation binaire		Programmation dynamique	
	Moyenne	Médiane	Moyenne	Médiane	Moyenne	Médiane
Justesse	0.91	0.97	0.92	0.97	0.92	0.97
Précision	0.75	1.00	0.90	1.00	0.89	0.98
Sensibilité	0.89	0.97	0.89	0.97	0.90	0.96
Score F1	0.86	0.95	0.87	0.96	0.88	0.95

TABLE 5.3 – Valeurs des indicateurs de performance

Nous pouvons analyser l'homogénéité des résultats de cette méthode avec les précédentes, comme nous l'avons déjà effectuée entre l'inférence MLE et la segmentation binaire. Les figures 5.21 représentent cette comparaison. Nous constatons que le test est concluant pour la distribution du coucher entre la méthode MLE et de la programmation dynamique. Il l'est également pour la distribution du lever entre la méthode de segmentation binaire et programmation dynamique.

Cette méthode permet donc d'améliorer les précédentes en ajoutant la possibilité de segmenter les données en temps réel.

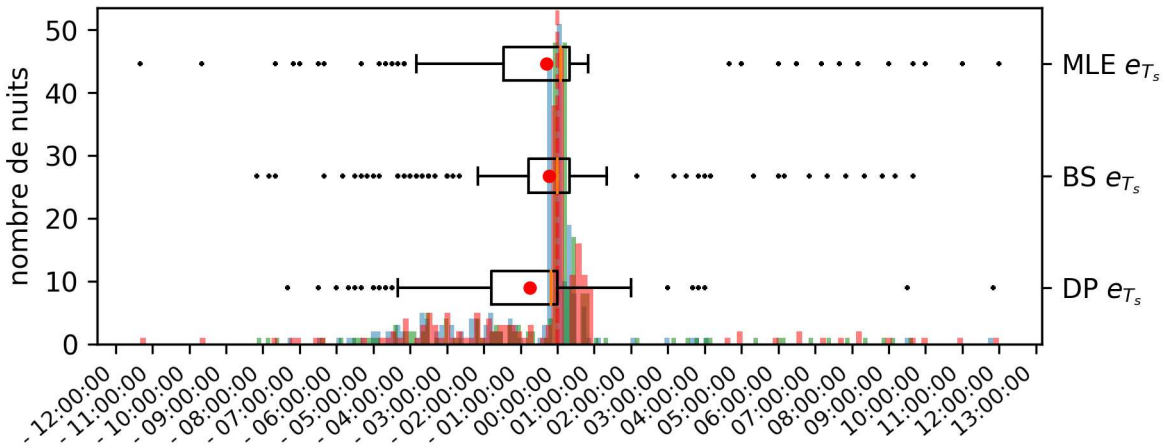
Nous pouvons résumer les caractéristiques principales des trois approches dans le tableau 5.4.

Temps réel La capacité de l'algorithme à traiter des données qui se mettent à jour en temps réel

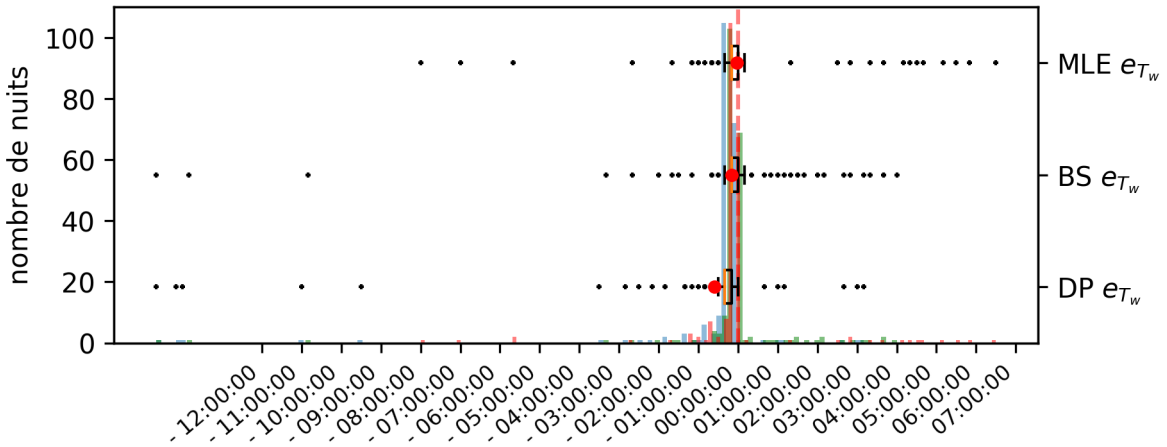
Complexité La complexité de l'algorithme par rapport à la taille des données en entrée

Passage à l'échelle La complexité de l'algorithme pour l'inférence de différents types d'activités

Optimalité La capacité de l'algorithme à trouver la solution optimale.



(a) L'erreur e_{T_s} sur les heures de coucher inférées



(b) L'erreur e_{T_w} sur les heures de lever inférés

FIGURE 5.18 – Illustration des distributions d'erreurs sur les heures de coucher et de lever e_{T_s} et e_{T_w} pour les trois méthodes

	Temps réel	Complexité	Passage à l'échelle	Optimalité
Modèle MLE	Non	$\mathcal{O}(n^3)$	$\mathcal{O}(n^{2p+1})$	Oui
Segmentation Binaire	Non	$\mathcal{O}(n\log(n))$	$\mathcal{O}(n\log(n))$	Non
Programmation dynamique	Oui	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$	Oui

TABLE 5.4 – Comparaison des caractéristiques principales des trois approches étudiés

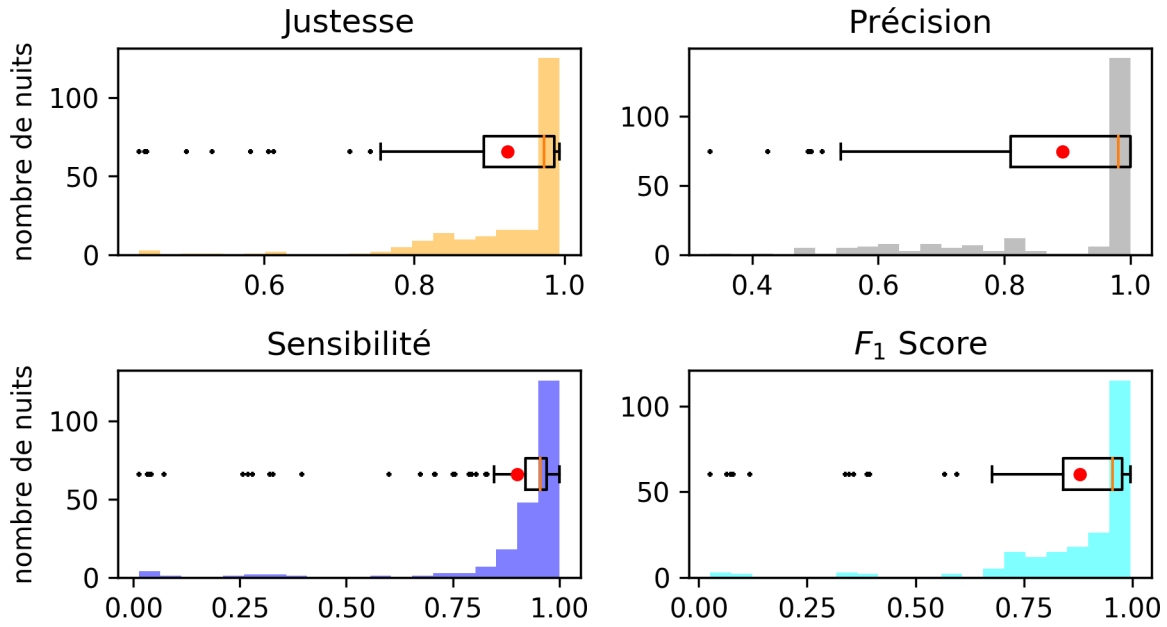
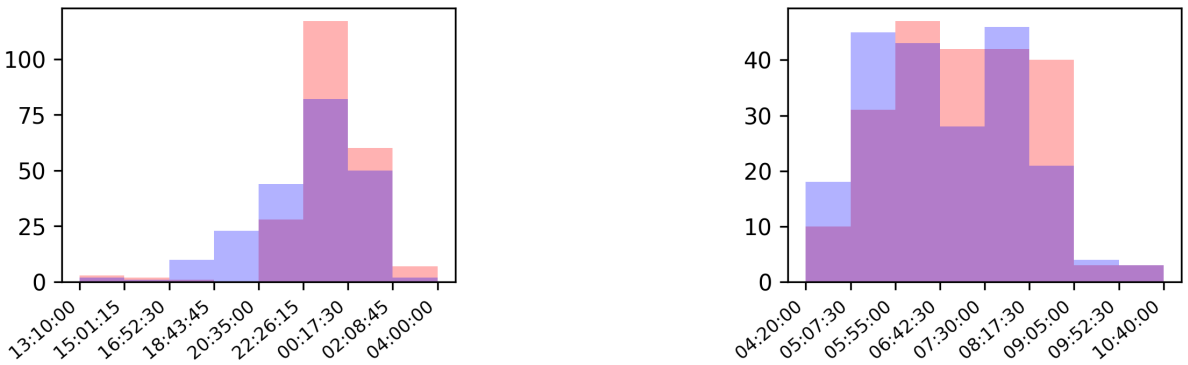


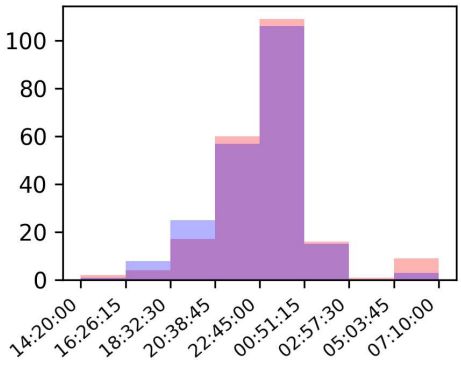
FIGURE 5.19 – Distributions de la justesse, précision, rappel et score F1 sur les nuits de la base de données pour la méthode d'inférence en programmation dynamique



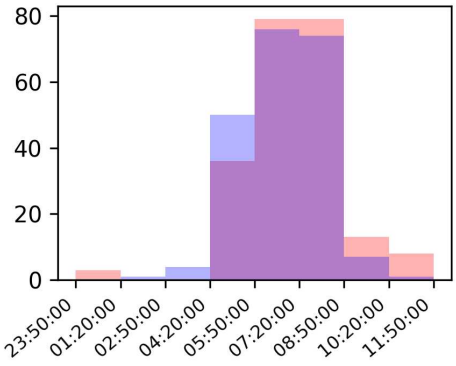
(a) Comparaison des heures de coucher

(b) Comparaison des heures de lever

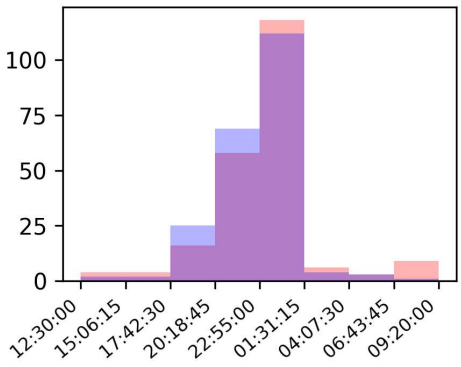
FIGURE 5.20 – Comparaison des distributions des heures de coucher et lever inférées par la méthode de programmation dynamique (en bleu) aux annotations (en rouge)



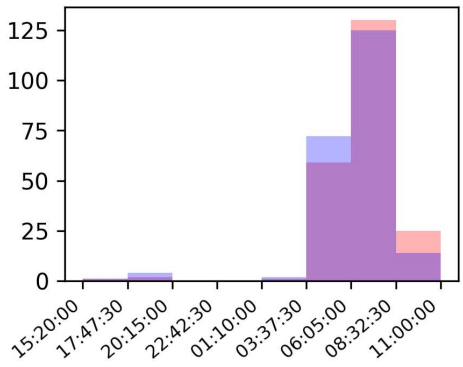
(a) (MLE vs DP T_s)
$T = 13.56 < 14.07$



(b) (MLE vs DP T_w)
$T = 32.99 \not< 14.07$



(c) (BS vs DP T_s)
$T = 17.23 \not< 14.07$



(d) (BS vs DP T_w)
$T = 9.49 < 14.07$

FIGURE 5.21 – Comparaison des distributions des heures de coucher et lever inférées par la méthode de programmation dynamique (en bleu) à la méthode MLE en première ligne et segmentation binaire en deuxième ligne (les deux en rouge).

5.6 Perspectives

L'algorithme d'inférence du sommeil avec la méthode du maximum *a priori* a été déployé dans plus de 50 installations, dont des maisons de retraite et des appartements privés. L'inférence de l'heure du lever est utilisée par le personnel infirmier pour estimer l'heure approximative à laquelle les résidents se réveillent afin qu'ils puissent aller les aider dans leur routine matinale. À partir de la période de sommeil présumée, nous pouvons également déduire l'activité nocturne pour détecter de potentiels troubles du sommeil ou des problèmes liés à des visites fréquentes aux toilettes. Et ceci, grâce aux algorithmes de segmentation binaire ou l'approche en programmation dynamique.

Des trois méthodes, l'algorithme de segmentation binaire semble être le mieux adapté pour cette application. Contrairement aux deux autres, il permet d'inférer les segments de sommeil sans `a priori` tout en étant le plus rapide. Il permet aussi de directement en déduire l'activité nocturne sans avoir à imbriquer les modèles.

5.7 Conclusion

Nous avons proposé une méthode non supervisée de détection des périodes de sommeil à l'aide de séries temporelles binaires de données actigraphiques. Cette méthode est très simple et peu coûteuse en calcul avec le bon paramétrage, ce qui permet de l'intégrer dans les dispositifs de surveillance médicale existants. Elle est également économique, car elle ne nécessite aucune configuration spéciale (calibrage, phase d'apprentissage ou placement spécial des capteurs) qui dépend par exemple de la géométrie ou de la nature des différentes pièces de la résidence. Cet algorithme est actuellement utilisé dans des studios et des appartements dans le cadre de la phase d'expérimentation du projet à l'origine de cette thèse. Ensuite, nous avons proposé des améliorations à cette méthode permettant de segmenter plus en détail les observations. Permettant ainsi l'inférence des levers nocturnes et même d'autres types d'activités sur la journée.

6

Analyse fréquentielle sur les tendances et rythmes d'activités

Sommaire

6.1 Introduction

Un système d'assistance à l'autonomie à domicile (AAD) est à destination de différents intervenants dont les besoins peuvent être spécifiques et très différents pour chacun. Étant données les configurations de logements les installations peuvent être très différentes d'un individu à l'autre. Les capteurs n'ont de sens que dans leur contexte. La présentation des données revêt donc une importance primordiale dont dépend l'usage effectif des services que l'on met en place. Les soignants sont souvent débordés, pas toujours familiarisés avec la technologie. La lecture des données doit être simple et rapide et permettre de tirer le plus d'information utiles sur le patient afin de prendre les meilleurs décisions qui permettront de conserver le plus longtemps possible l'autonomie du résident.

Dans ce chapitre, contrairement au chapitre précédant qui était focalisé sur l'analyse d'une journée d'activité, nous proposons une présentation des données adaptée à une analyse multi-échelle afin de pouvoir visualiser sur une semaine, un mois, un trimestre, une année les données actimétriques et d'en suivre les évolutions sur un temps long.

Les besoins sont différents en fonction des intervenants. [69] proposent huit types d'utilisateurs d'un système AAD. Parmi les principaux ils identifient :

— Le résident,
— Le personnel soignant sur-place,
— Le personnel soignant à distance,
— L'équipe de maintenance du système de télésurveillance.

Nous pouvons ajouter à cette liste l'équipe de recherche et développement du système et des sous-systèmes d'AAD qui elle aussi a besoin de visualiser les données collectées et cela sous différentes représentations pour améliorer les dispositifs en place. Comme nous l'avons évoqué plus haut, chacun de ces acteurs a des besoins très spécifiques et à des niveaux d'abstraction différents :

— au niveau des données brut,
— sur une représentation abstraite de ces données.

Le besoin peut être plus élaboré sous une forme de conseils, de rapports ou d'alertes.

Les besoins

Le bénéficiaire L'installation de capteurs à domicile nécessite un dialogue avec l'utilisateur qui doit conduire à son adhésion. Celle-ci ne pourra être entretenue si l'usager ne perçoit pas l'utilité des dispositifs installés. C'est pourquoi il est fondamental de penser lors de la conception d'un système de capteurs à domicile au moyens d'entretenir la motivation et la confiance dans le système. En particulier, cela passe par un affichage ergonomique, voire ludique des données collectées. Typiquement, un compte rendu simplifié à base d'iconographie résumant son état de santé peut lui être proposé et cela à un rythme régulier que l'on peut adapter en fonction de la demande. Il est également intéressant de lui prodiguer des conseils sur ses habitudes de vie. On peut envisager aussi le rappel des moments importants de la journée (hydratation, prise de médicaments, ..) et fournir des moyens de communication facilitant l'interaction avec ses proches et le personnel soignant.

Le personnel soignant Le personnel soignant sur place s'intéressera aux rythmes de vie de chacun des patients dont il a la charge alors que celui à distance aura plutôt besoin d'une synthèse regroupant les patients par catégories. En termes de visualisation, ils demandent des rapports graphiques journaliers des activités et des rapports de tendances hebdomadaires et mensuelles qu'ils pourront mettre à profit pour pouvoir évaluer la qualité du soin et le niveau d'autonomie du patient. Parmi les information très utiles on peut citer l'activité nocturne du résident ou sa propension à déambuler la journée, l'heure de la prise des repas, l'heure de coucher et de lever, le suivi de l'hygiène ou encore la fréquence d'utilisation des toilettes. Ces éléments peuvent être détectés à domicile avec des capteurs simple et peu coûteux.

Le médecin Le médecin est intéressé par des rapports plus globaux indiquant des tendances sur le rythme de vie afin d'identifier les anomalies ou les progrès réalisés. Il est intéressé aussi à analyser l'impact d'un traitement médicamenteux sur l'activité de son patient. Pour cela, il analyse les données sur une période de temps à plusieurs échelles :

semaine, mois, trimestre ... Il cherche soit à détecter des évolution lentes, insidieuses, ou au contraire des changements brutaux.

le développeur d'applications L'équipe de maintenance et de recherche et développement de son côté sera plus intéressée par une vue globale du fonctionnement des installations pour intervenir en cas de pannes ou imprévus, mais aussi pour ajuster les modèles d'inférence automatiques suivant l'évolution des données et des nouvelles connaissances expertes nouvellement acquises au fil du temps.

Ainsi dans ce chapitre nous nous intéressons à la représentation graphique des données capteurs et son potentiel pour l'étude, l'évaluation et l'inférence des tendances et habitudes de vie d'un patient. Nous présentons une technique de visualisation qui permet une lecture rapide de l'information, le but étant de ne pas surcharger les équipes soignantes. Nous montrons à travers une étude de cas que l'on peut visualiser les habitudes et activités de la vie quotidienne, et en inférer des situations de vie particulières.

6.2 État de l'art

La grande majorité des études de l'état de l'art proposent une représentation des actvités de la vie quotidienne sous forme matricielle dans lesquelles chaque ligne est un chronogramme. D'autres études proposent des représentations spécialisées pour le suivi médical [81] (Figure 6.2) ou autre [95] (Figure 6.1). Le point commun entre ces outils est une représentation du temps par journée sur un axe et des données actimétriques sur un autre axe. Les coefficients ou données représentés dans ces graphes ou matrices peuvent soit être directement des données capteurs bruts ou bien des informations plus élaborées.

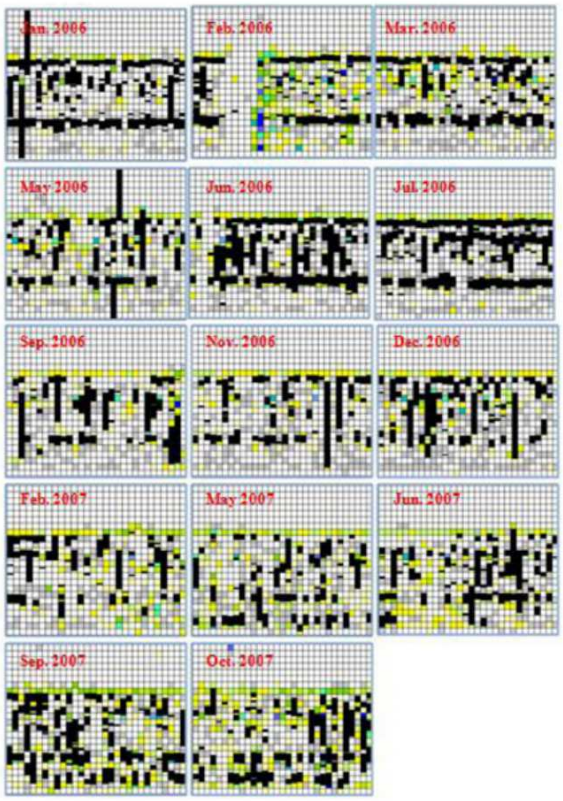


FIGURE 6.1 – Exemple de l'outil *activity density maps* appliqué à des données issues d'une installation à un résident

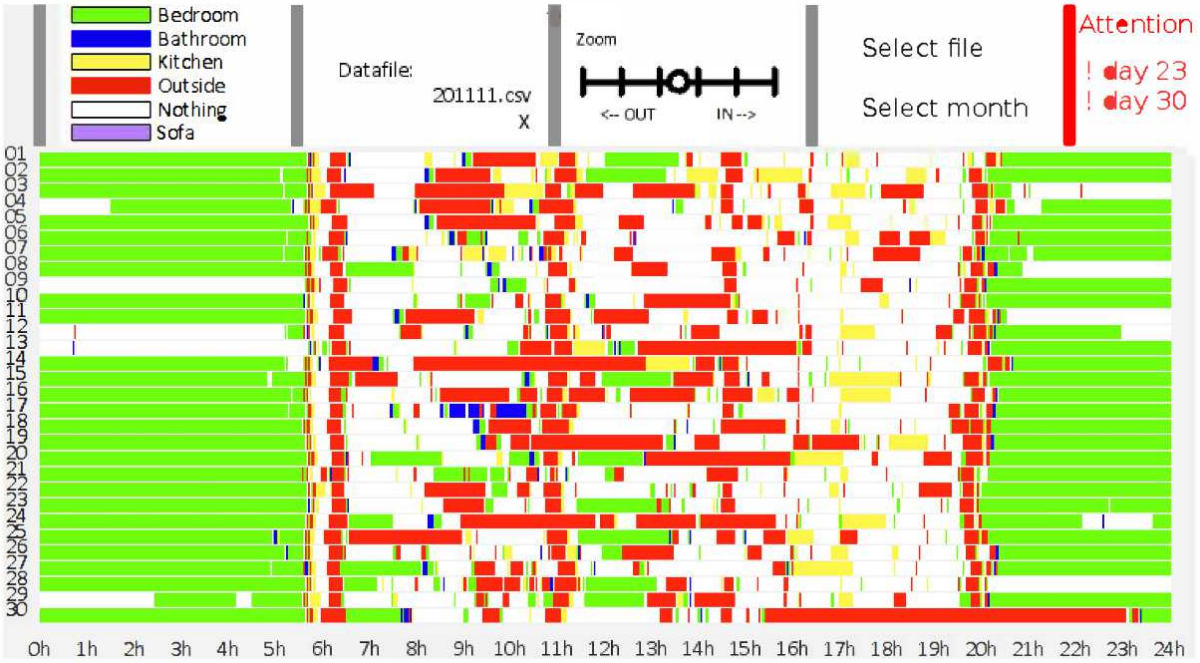


FIGURE 6.2 – Exemple de représentation des AVQ par plages horaires

6.3 Représentation Spectrale de l'activité d'une personne

Pour optimiser la lecture des données capteur, nous nous référons à la théorie de Gestalt qui définit les notions de continuité et de proximité comme principe perceptif à prendre en compte dans la représentation des données. [22].

6.3.1 Application sur une Série Temporelle Booléenne

Un outil répandu en analyse spectrale est le fenêtrage. Celui-ci permet l'observation de signaux à durée limitée dans le temps et permet dans le cas d'une moyenne glissante d'observer les tendances globales du signal c.-à-d. les variations à basse fréquence.

Étant donné un signal sous forme d'une série temporelle discrète $s = s_1, s_2, \ldots, s_n$, la transformée sur le domaine temporel avec application d'une fonction de fenêtrage F de taille de fenêtre w sera le produit de convolution suivant :

$$\forall t \in [\![w, n]\!] \quad S_t = \sum_{i=t-w+1}^{t} s_i \times F_w(i) = s_{[t-w+1:t]} \cdot F_w \tag{6.1}$$

ou bien :

$$\forall t \in [\![w, n]\!] \quad S_t = \sum_{i=t-\lceil \frac{t-w+1}{2} \rceil}^{t+\lceil \frac{t-w+1}{2} \rceil} s_i \times F_w(i) = s_{[t-\lceil \frac{t-w+1}{2} \rceil : t+\lceil \frac{t-w+1}{2} \rceil]} \cdot F_w \tag{6.2}$$

avec des fenêtres centrées.

Pour la fonction fenêtre, nous choisissons la fenêtre de *Blackman* :

$$F_w(n) = a_0 - a_1 \cos\left(\frac{2\pi n}{w}\right) + a_2 \cos\left(\frac{4\pi n}{w}\right) \tag{6.3}$$

avec et d'après **(author?)** [29] pour l'«*Exact Blackman*» :

$$a_0 = \frac{7938}{18608}; \quad a_1 = \frac{9240}{18608}; \quad a_2 = \frac{1430}{18608} \tag{6.4}$$

En appliquant cette transformée sur des données issues de capteurs de mouvement, et cela pour plusieurs tailles de fenêtres, nous obtenons une représentation en ondelettes dont la lecture nous permet de facilement distinguer certaines périodes clefs de la journée. La figure 6.3 représente une décomposition en ondelettes d'un signal issu d'un capteur de mouvement.

Les abscisses représentent le temps et les ordonnées représentent la taille de la fenêtre utilisée dans la décomposition. L'échelle de couleurs correspond à la valeur de la moyenne calculée sur la fenêtre glissante. Les valeurs bruts fournies par le capteur sont en bleu en haut de la figure. Les traits verticaux en vert donnent des repères sur l'heure : en trait plein "minuit" et en pointillés "midi". La courbe bleue en bas représente la décomposition sur une fenêtre de deux heures.

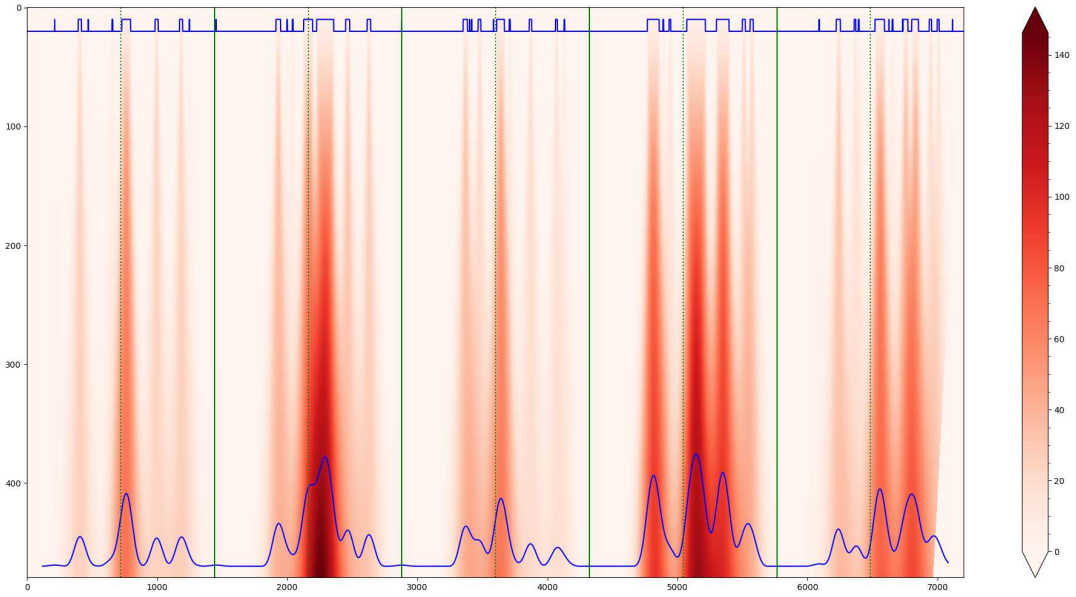


FIGURE 6.3 – Analyse par ondelettes d'un signal du capteur de mouvement

Nous pouvons constater qu'une telle représentation permet de bien mettre en évidence certaines périodes de la journée où l'activité est plus intense. Ici les données présentées sont issues d'un capteur placé dans la cuisine d'un logement de nos résidents. On voit bien un surcroit d'activité autour de midi et aussi mais moins prononcé le matin apres le levé et le soir au moment du petit déjeuner et du dîner.

6.4 Représentation matricielle et extraction des habitudes

La représentation spectrale vue plus haut ne passe pas à l'échelle pour une lecture des habitudes de vies sur plusieurs jours voir plusieurs mois. C'est ainsi qu'est venue l'idée de combiner la décomposition en ondelettes avec une représentation en deux dimensions où chaque ligne représente une journée comme présentées dans [95].

Plus formellement, après avoir calculé la transformée d'un signal pour une valeur de fenêtre particulière $S = S_0, S_1, \ldots, S_{N-1}$, il suffit de séparer les données correspondantes jour après jour i : $S_i = S_{iT}, S_{iT+1}, \ldots, S_{(i+1)T-1}$ (T étant le nombre de mesures journalières) et de les empiler pour construire la matrice 6.5 illustrée figure 6.4. :

$$\begin{pmatrix} S_0 & S_1 & \ldots & S_{T-1} \\ S_T & S_{T+1} & \ldots & S_{2T-1} \\ \vdots & \ddots & & \\ S_{iT} & S_{iT+1} & \ldots & S_{(i+1)T-1} \\ \vdots & & & \end{pmatrix} \tag{6.5}$$

La représentation spectrale vue plus haut nous permet de faire ressortir les points de

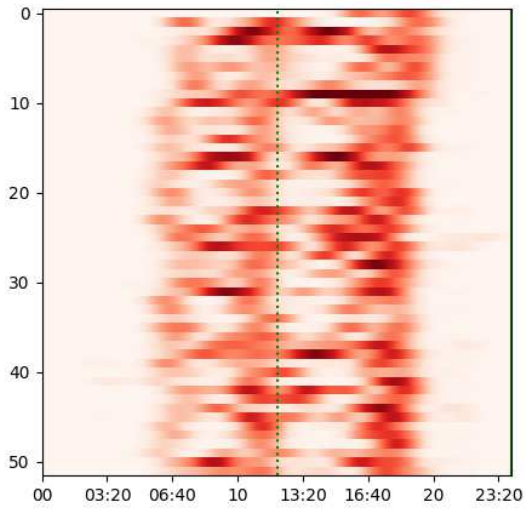


FIGURE 6.4 – Représentation matricielle du mouvement décomposé en ondelettes dans la cuisine d'un logement

forte activité sur une journée à l'aide de la décomposition en ondelettes.

Afin de lisser les activités dont on sait qu'elles sont répétitives d'une journée à l'autre, nous proposons de moyenner aussi les activités sur plusieurs journées. Pour cela nous réappliquons une décomposition en ondelettes sur l'axe des journées pour chaque tranche de temps $j \in [0, T[$:

$$S_j = \begin{pmatrix} S_0 \\ S_T \\ S_{2T} \\ \vdots \\ S_{nT} \end{pmatrix} \tag{6.6}$$

avec $n = \frac{N}{T}$ le nombre de jours disponibles dans les données pour suivre les principe de Gestalt dont nous avons prlé plus haut.

La figure 6.5 présente le résultat d'une telle opération sur différentes tailles de fenêtres.

Nous remarquons que cette opération permet de faire ressortir les habitudes autour des heures de petit déjeuner, déjeuner et le soir correspondant potentiellement au dîner. Nous attirons l'attention sur l'importance de la taille de fenêtre choisie. Une fenêtre trop courte ne lissera pas assez les segments journaliers alors qu'une fenêtre trop grande générera une trop grande diffusion des bandes de forte activité.

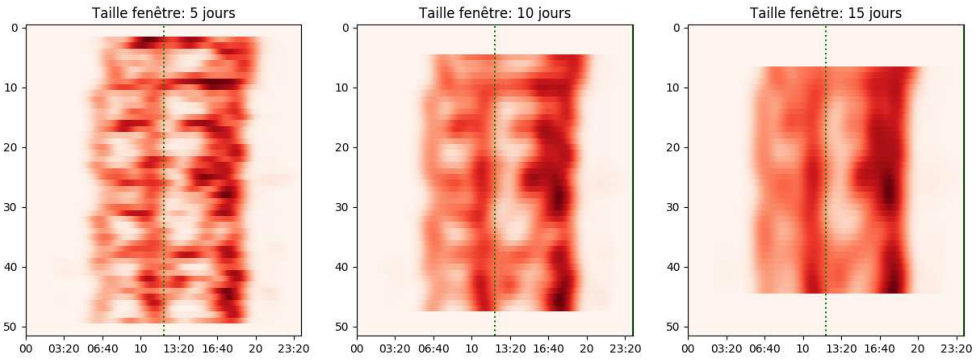


FIGURE 6.5 – Comparaison de représentations matricielles du mouvement décomposé en ondelettes dans la cuisine d'un logement avec différentes tailles de fenêtres de lissage sur l'axe des jours

6.5 Étude de cas

Dans cette section nous allons monter que la représentation que nous venons de proposer permet réellement d'inférer des situations de vie à partir des seuls actigrammes que nous avons construits à partir des quelques capteur tout ou rien placés dans l'appartement choisi. Celui-ci a été installé par la société Diatelic, qui a équipé aujourd'hui une cinquantaine d'appartement. Il illustre un cas reel. Tous les logements, suivant leur configuration (c.-à-d. types de pièces disponibles), sont équipés par certains des capteurs de mouvements référencés dans le tableau 6.1.

TABLE 6.1 – Référencement des capteurs de mouvement

Emplacement	référence
Salon	`MSA`
Chambre à couché	`MCH`
WC	`MWC`
Salle de bain	`MSDB`
Cuisine	`MCU`
Salle à manger	`MSA`
Bureau	`MBU`

Pour chacun de ces logements, nous dressons la matrice actigraphique des capteurs disponibles et rapportons les observations qui en découlent. Ce qui nous permet de montrer comment un tel outil peut être mis à contribution pour la lecture et l'analyse de différents phénomènes liés aux habitudes de la vie quotidienne.

Certains capteurs présentent des dysfonctionnements sur certaines périodes de temps, cela explique que l'on retrouve certaines plages vides de données sur certaines figures.

Les données sont représentées dans le fuseau horaire local, nous discuterons dans la prochaine section des implications de ce choix et des effets que cela a sur la lecture des

habitudes lors des changements d'heure en mars et octobre de chaque année.

6.5.1 Le Logement

Le logement que nous considérons ici dispose d'un Salon, d'une chambre principale, d'une cuisine, d'une salle de bain et d'un WC. Il est équipé des capteurs suivants : `MSA`, `MCH`, `MCU`, `MSDB`, `MWC`.

Salon

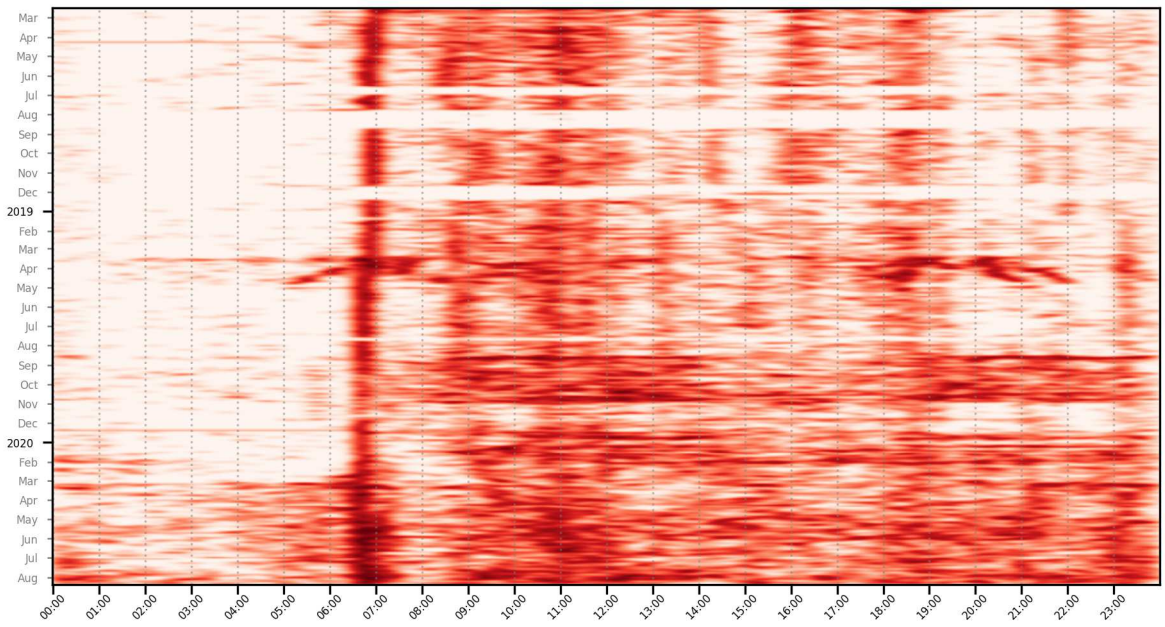


FIGURE 6.6 – Données du capteur `MSA` installé dans le Salon du Logement 1

La figure 6.6 illustre la représentation que nous avons construite comme indiquée dans le paragraphe précédant. Il s'agit de l'activité perçue par le capteur de mouvement `MSA` sur une période s'étendant de mars 2018 à août 2020. Il est installé dans le Salon du logement que nous avons choisi.

On remarque la présence de plages blanches [3] qui correspondent soit à des périodes d'absence du résidant soit un dysfonctionnement du capteur.

Le premier pic d'activité qui ressort de la figure 6.6 est situé vers 6 h 30 et 7 h. Il correspond au lever de la personne le matin. On relève une perturbation de cette habitude de vie entre mars et avril 2019 puis une reprise de la tendance habituelle jusqu'à août 2020. On note alors que la personne passe plus de temps dans son salon le matin jusqu'à mi-décembre. À partir de mars 2020, le capteur se met à relever beaucoup d'activité la nuit ce qui peut indiquer un problème de déambulation nocturne. Entre mars 2018 et août 2019, on relève une présence matinale dans le salon jusque vers midi et qui reprend de 16 h

3. Dans la suite, on se concentrera uniquement sur les périodes ayant des données complètes.

à 19 h. A partir d'août 2019, les données deviennent trop bruitées sur toute la journée pour qu'on puisse y relever une quelconque habitude marquée. La personne semble plus active ou agitée le soir après 19 h.

Cuisine

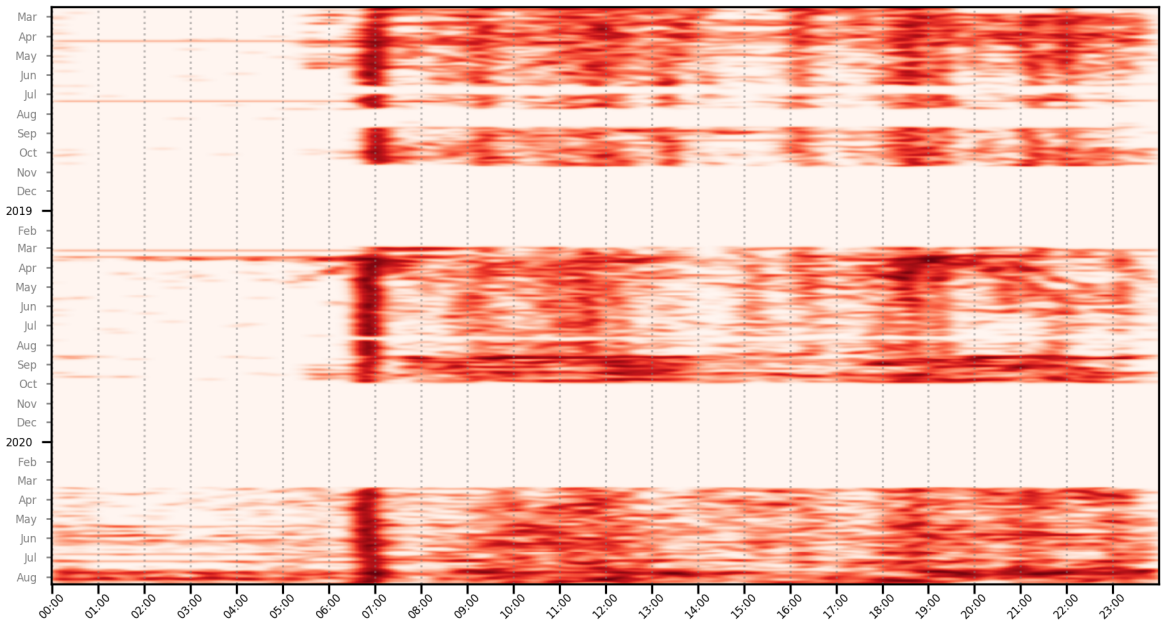


FIGURE 6.7 – Données du capteur `MCU` installé dans la cuisine du Logement 1

La figure 6.7 est un relevé du capteur de mouvement `MCU` installé dans la cuisine du logement 1 sur la même période de mars 2018 à août 2020 que le capteur du Salon vu au paragraphe précédent.

On peut noter ici des comportements qui nécessitent une analyse fine : entre octobre 2018 et mars 2019 et octobre 2019 et mars 2020 (correspondant au passage à l'heure d'hiver). On observe une activité matinale soutenue vers 7 h 00 du matin. On constate aussi une perturbation des habitude de vie sur la période située entre mars et avril 2019.

On relève aussi de manière constante trois principales périodes d'activité dans la cuisine entre 09 h 00 et 13 00, 18 h 00 et 20 h 00 et 21 h 00 et 23 h 00.

Chambre à coucher

La figure 6.8 illustre les données du capteur de mouvement `MCH` installé dans la chambre à coucher du logement 1. Contrairement à ce qu'on avait pu observer dans les autres pièces, on observe une quantité de mouvement assez soutenue pendant la nuit dans cette pièce. Une explication est que le capteur de mouvement peut être très sensible aux mouvements de la personne dans son lit pendant son sommeil, si celle-ci remue beaucoup. Notons également que le pic d'activité lors du réveil de la personne apparaît clairement.

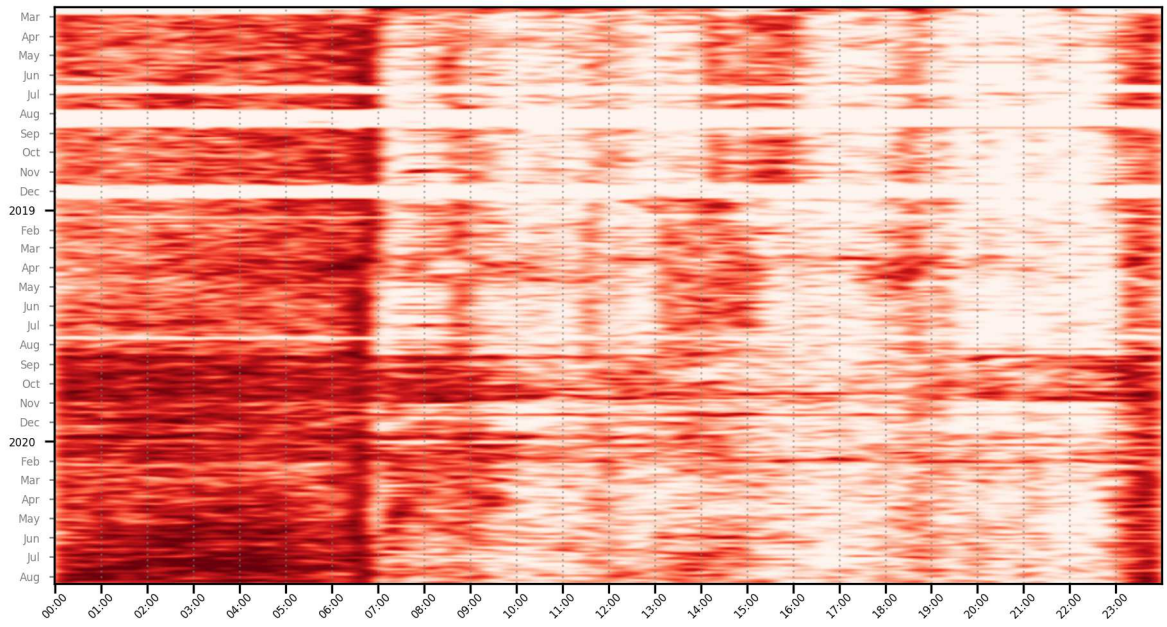


FIGURE 6.8 – Données du capteur `MCH` installé dans la chambre à coucher du Logement 1

Sur les périodes d'août à octobre 2019 et à partir de mai 2020 l'activité dans la chambre pendant la nuit se révèle être plus dense que d'habitude, cela pourrait être un signe de sommeil agité.

Notons également, entre septembre et novembre 2019, puis pendant la semaine de Noël ainsi que certains jours de janvier 2020 une rupture du rythme régulier de l'heure de réveil.

À partir de janvier 2020, la personne passe plus de temps dans sa chambre.

Pendant la journée, l'habitude la plus marquée serait l'activité dans la chambre entre 14 h 00 et 16 h 00 qui se décale d'une heure à partir de décembre 2018 entre 13 h 00 et 15 h 00. Cette activité pourrait correspondre à une sieste après le repas.

Salle de bain

La figure 6.9 exhibe le capteur de mouvement `MSDB` situé dans la salle de bain du logement 1.

Ce capteur relève des mouvements pendant la journée auxquels on ne s'attendrait pas. L'explication est probablement lié au fait que la porte de la salle de bain reste ouverte.

On remarque aussi le pic classique d'activité au réveil ainsi que deux visites matinales régulières sauf pendant la période entre septembre et novembre 2019.

WC

La figure 6.10 présente le compte rendu émis par le capteur de mouvement `MWC` installé dans les WC du logement 1.

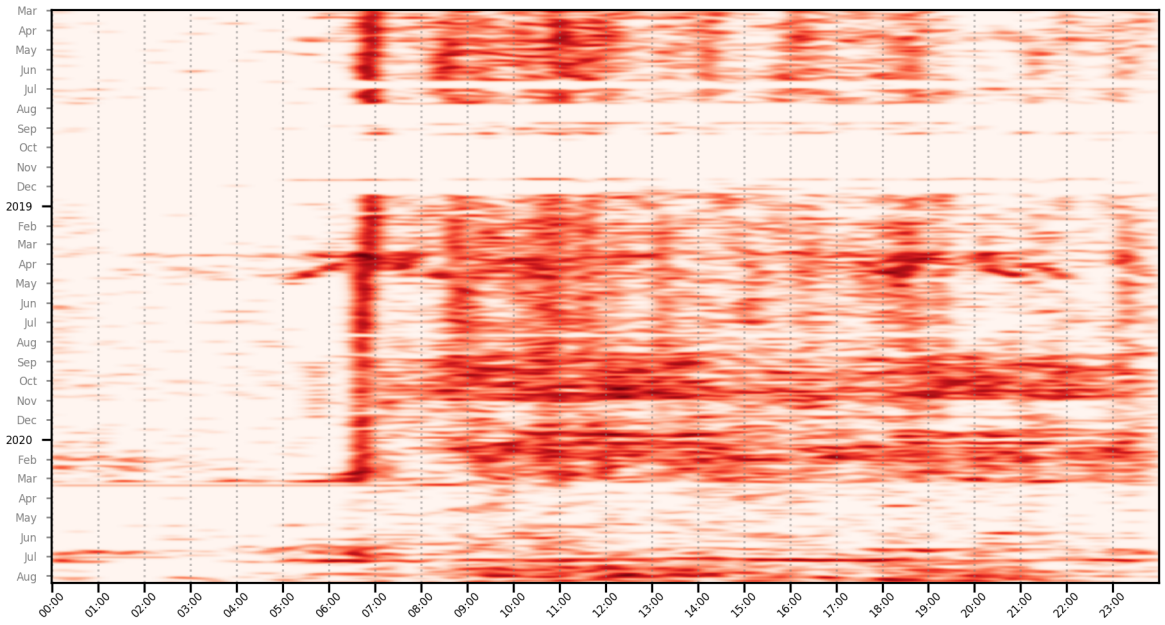


FIGURE 6.9 – Données du capteur `MSDB` installé dans la salle de bain du Logement 1

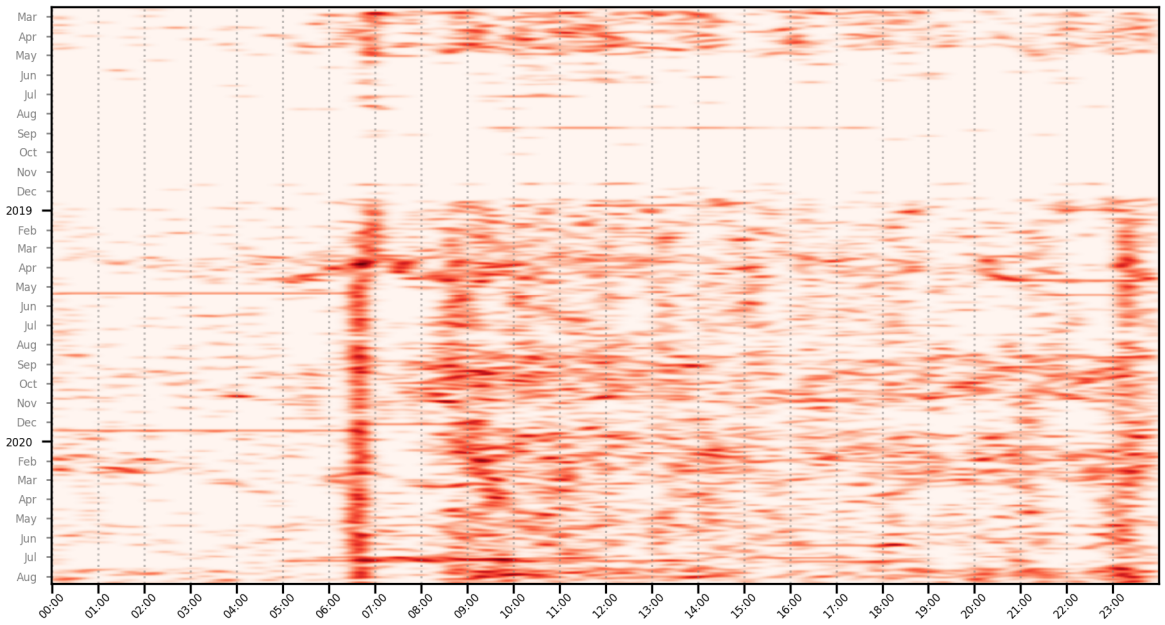


FIGURE 6.10 – Données du capteur `MWC` installé dans les WC du Logement 1

Le pic d'activité matinale ressort particulièrement bien. À l'instar des capteurs dans la salle de bain, le capteur relève une forte activité qu'on ne peut expliquer a priori. Une explication potentielle serait que la porte des WC reste ouverte et que le capteur détecte des activités hors de cette pièce.

6.5.2 Lecture et détection des habitudes de vie

Hygiène

L'hygiène est une composante importante dans l'évaluation de la qualité de vie et l'indépendance des patients suivis. Ainsi il est important que l'outil de visualisation permette la bonne lecture des activités liées aux visites de salle de bains et aux WC.

Comme nous l'avons vu dans l'étude de cas les données recueillies depuis les capteurs de la salle de bain et WC des différents logements il est clair que le placement des capteurs et le paramétrage de sa ligne de vue sont primordiaux pour en tirer des informations utiles. Typiquement, les rapports des capteurs pour les logements 1 et 2 laissent envisager que les cônes de vision des capteurs se chevauchent et ainsi invalident la possibilité d'avoir une granularité par pièce pour les événements d'activation. D'un autre côté, les logements 3 et 4 montrent qu'avec une bonne installation il est largement possible de visualiser des activités de prise de douche ou de visite des WC à travers que des données de mouvements.

Prise de repas

Tout comme l'hygiène, la prise de repas est aussi un facteur important dans l'évaluation de l'indépendance des patients suivis, la bonne prédiction de la dégradation de cette activité est primordiale pour l'optique du maintien à domicile des personnes âgées. En effet, une détection précoce des problèmes d'alimentation sera plus simple à prendre en charge pour permettre à la personne de continuer à vivre à son domicile. Plus le problème persistera, plus les conséquences seront importantes (perte de masse musculaire, augmentation du risque de chute, dégradation de la qualité de vie) et nécessiteront la mise en place de solutions plus radicales telles qu'embaucher une aide-soignante à domicile ou le déménagement vers une chambre en Ehpad.

Évaluation des heures de sommeil

L'activité de sommeil est parmi les plus faciles à lire à travers les matrices actigraphiques. En effet sur la grande majorité des diagrammes vue plus haut il ressort facilement une longue période de basse activité durant la nuit relative au sommeil des résidents de leurs logements respectifs. Ces périodes sont plus visibles sur les logements 3 et 4 que les logements 1 et 2.

Sur la figure 6.11 malgré la présence d'activité tout au long du jour et la nuit on voit bien que l'algorithme d'inférence du sommeil permet de faire ressortir les périodes de sommeil non triviales contenant des coupures. Bien sûr, l'algorithme d'inférence des périodes de sommeil trouve tout aussi facilement celles trivialement lisibles à l'oeil comme le montre la figure 6.12. Mais on voit bien ici de par sa limitation qui fait qu'il ne pourra pas inférer plusieurs périodes de sommeils pour une seule journée, l'importance de l'outil

de visualisation qui permet de faire ressortir l'activité de 3 h qui coupe la période de sommeil en deux.

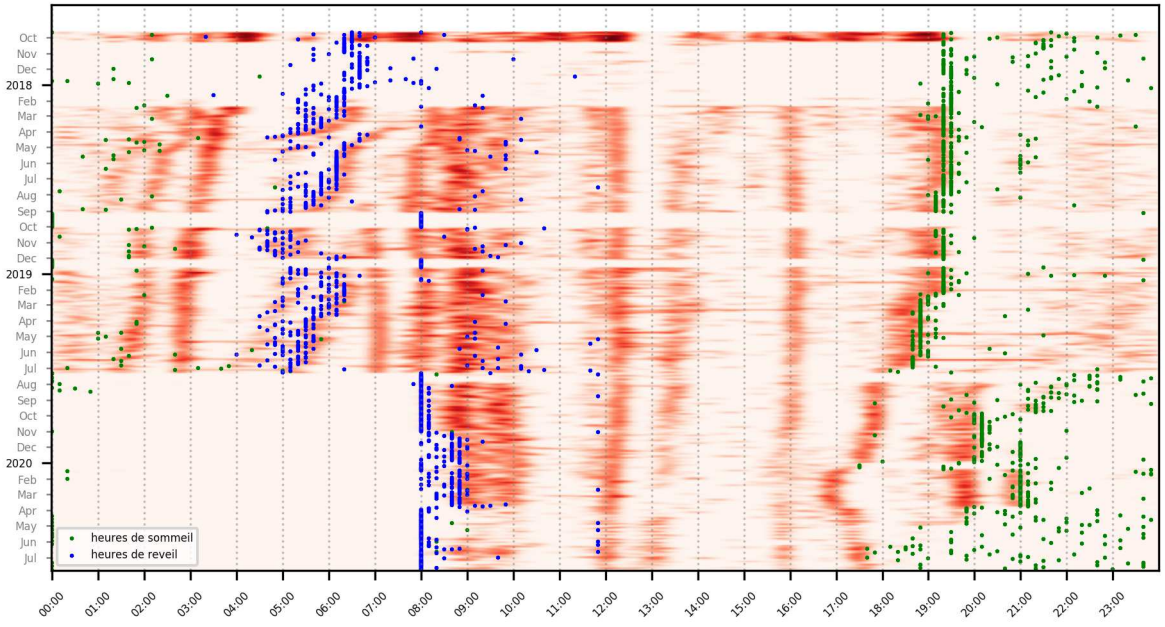


FIGURE 6.11 – Actigramme du capteur `MSA` du logement 2 avec les annotations des inférences des heures de sommeil et réveil en vert et bleu respectivement

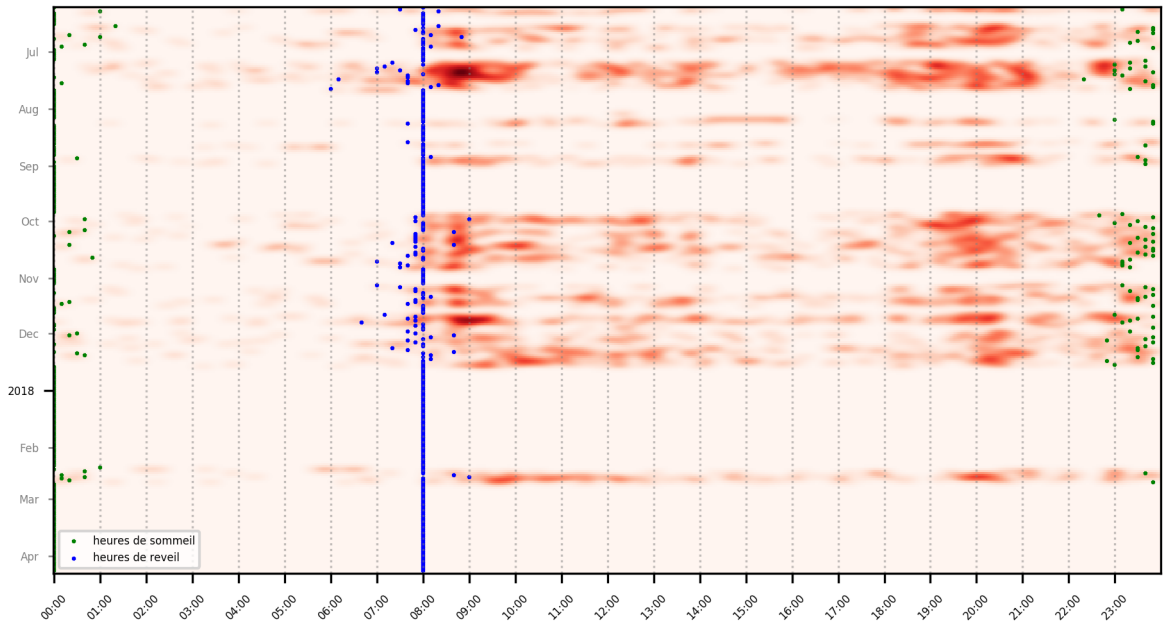


FIGURE 6.12 – Actigramme du capteur `MSDB` du logement 3 avec les annotations des inférences des heures de sommeil et réveil en vert et bleu respectivement

6.6 Conclusion

Nous avons présenté dans ce chapitre une technique de représentation graphique des données capteurs et son application à l'étude et l'évaluation des tendances et habitudes de vie d'une personne. Nous avons exploré les différents cas d'utilisations de cet outil, détaillé la méthode à suivre pour obtenir une telle représentation, appliqué l'outil à une installation expérimentale pour ensuite analyser des situations de vie.

Parmi les perspectives de ce travail nous pouvons citer, la reconnaissance automatique des blobs d'activités directement sur la matrice actigraphique qui est une contribution de cette thèse. Cela permettrait d'automatiser ou faciliter l'exploitation de ces données mais aussi d'identifier les logements qui restituent des données peu exploitables. Cela permettrait d'améliorer certaines installations.

Une limitation de l'outil est qu'il ne peut représenter qu'un seul capteur à la fois par matrice actigraphique. Pour des logements équipés de très nombreux capteurs, il devient difficile de lire les évolutions au fil de la journée et des pièces. La solution évidente serait d'afficher sur une seule matrice les données agglomérées des capteurs avec différentes couleurs, mais cela nécessiterait non seulement des données propres avec des capteurs dont les zones de couvertures ne se chevauchent pas, mais aussi un travail sur le choix des couleurs et le paramétrage de la convolution entre ces différents capteurs pour mieux visuellement faire ressortir les enchaînements de pièces.

7

Conclusion générale

L'augmentation de la population âgée et le désir du maintien à domicile de cette population est un enjeu majeur de nos sociétés modernes. Le choix de la maison de retraite est une grande préoccupation pour les personnes âgées et leurs familles et ne se fait que par défaut d'une autre solution. Cela est dû à des facteurs socio-économiques, mais aussi à des facteurs purement médicaux comme on a pu le constater lors la pandémie du Covid-19. L'alternative étant le soin à domicile, beaucoup de recherches et de développement sont menées dans le monde pour développer les technologies pouvant faciliter et assister le personnel et les résidents qui se tournent vers le maintien à domicile. Cette thèse s'inscrit dans le cadre d'un de ces travaux de recherche qui visent à proposer des produits et services pour le maintien et le suivi d'activités des personnes âgées.

Les contributions de la thèse sont des algorithmes pour la détection automatique d'activité et la génération de rapports transmis au personnel soignant et au bénéficiaires de nos installations.

Nous avons d'abord passé en revue les différents produits et systèmes de suivis d'activité pour contextualiser le positionnement et les contraintes du travail effectué. Puis, nous avons présenté les différents types de capteurs utilisés par la littérature, leurs avantages, inconvénients et limites. Avec cela nous avons présenté quelques principes, méthodes, algorithmes et outils répandus pour l'inférence d'activités. Nous avons également vu que certaines méthodes d'inférence nécessitaient des données étiquetées pour l'apprentissage de leurs modèles alors que d'autres se contentaient d'une fonction de critère pour cela.

La thèse s'inscrit dans le cadre d'un projet visant à commercialiser toute une solution d'aide au maintien à domicile des personnes âgées. Cette offre demande une coordination entre différents acteurs, systèmes, dispositifs et services. Il est ainsi nécessaire de guider notre travail de recherche en respectant les contraintes imposées par une telle coopération. Par exemple, ces contraintes sont le mode opératoire des capteurs domotiques développés par l'un de ces acteurs, ou le cahier des charges stipulant que la solution doit être immédiatement opérationnelle après son installation ce qui impose certains choix sur les algorithmes utilisés. Nous avons donc présenté ce contexte applicatif et la nomenclature des données sur lesquels l'inférence doit être effectuée pour générer les rapports ainsi qu'une analyse de certaines hypothèses et certains postulats sur l'environnement étudié.

La nature des données dont on dispose fait que l'approche naturelle est l'exploration des méthodes de segmentation de séries temporelles représentant des données issues de

capteurs de mouvement. Nous avons vu que c'est un problème bien formalisé dans la littérature, mais dont les recherches restent en constante évolution de par la diversité des domaines d'application et le grand intérêt qu'il lui est porté. Les méthodes de détection de point de rupture peuvent se subdiviser en plusieurs catégories : les méthodes basées sur le ratio de vraisemblance dite statistique, les méthodes basées sur des modèles probabilistes souvent bayésiens et les méthodes de regroupement souvent basé sur des mesures sur l'espace vectoriel des données analysées.

Nous nous sommes concentrés sur des méthodes à mis chemin entre les trois catégories présentées plus haut. Nous avons ainsi proposé trois algorithmes d'inférence d'activités. Le premier basé sur un modèle bayésien dont l'inférence se fait par la méthode du maximum de vraisemblance. Le second est un algorithme de segmentation basé sur le même principe que le premier, mais permettant l'inférence de plusieurs segments c.-à-d. qui passe à l'échelle pour plusieurs activités. Finalement, le troisième permet d'effectuer cette inférence à la volée sur des données capturées en temps réel. Pour chacun des algorithmes, nous avons analysé leur efficacité à l'aide d'un exemple sur des données étiquetées et présenté leurs avantages et inconvénients.

Lors de ce travail de recherche sur les algorithmes nous avons dû développer quelques outils pour manipuler et visualiser les données de travail. Nous présentons un de ces outils qui permet de visualiser les données de mouvement dans le logement sur d'assez longues périodes pour pouvoir en inférer visuellement un changement d'habitude ou de rythme de vie. En effet, l'un des principaux objectifs du projet est de non seulement développer des modèles d'analyse des données, mais aussi de les présenter sur des rapports journaliers au personnel soignant de façon claire et lisible pour les éclairer du mieux possible dans leurs décisions.

Ce travail a été déployé avec succès en phase de test sur une cinquantaine de résidences avec des rapports journaliers envoyé aux aides-soignantes sur les résidences d'assistance à domicile. Par exemple, ceci permet au personnel d'avoir une idée de l'état du sommeil de certains résidents, la nuit étant la seule période où ils ne peuvent pas interagir avec eux.

Bibliographie

[1] Abdel-Hamid, O., rahman Mohamed, A., Jiang, H., and Penn, G. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Mar. 2012), IEEE.

[2] Alsabti, K., Ranka, S., and Singh, V. An efficient k-means clustering algorithm. - (1997).

[3] Alsinglawi, B., Nguyen, Q. V., Gunawardana, U., Maeder, A., and Simoff, S. RFID systems in healthcare settings and activity of daily living in smart homes : A review. *E-Health Telecommunication Systems and Networks 06*, 01 (2017), 1–17.

[4] Andries, M., Simonin, O., and Charpillet, F. Localisation of humans, objects and robots interacting on load-sensing floors. *IEEE Sensors Journal 16*, 4 (Feb. 2016), 1026–1037.

[5] Andrieu, C., De Freitas, N., Doucet, A., and Jordan, M. I. An introduction to mcmc for machine learning. *Machine learning 50*, 1-2 (2003), 5–43.

[6] Arcelus*, A., Herry, C. L., Goubran, R. A., Knoefel, F., Sveistrup, H., and Bilodeau, M. Determination of sit-to-stand transfer duration using bed and floor pressure sequences. *IEEE Transactions on Biomedical Engineering 56*, 10 (2009), 2485–2492.

[7] Auger, I. E., and Lawrence, C. E. Algorithms for the optimal identification of segment neighborhoods. *Bulletin of mathematical biology 51*, 1 (1989), 39–54.

[8] Bachelet, M., and Anguis, M. Les médecins d'ici à 2040 : une population plus jeune, plus féminisée et plus souvent salariée. Tech. Rep. 1011, Direction de la Recherche, des Études, de l'Évaluation et des Statistiques, 2017. Disponible sur : `https://drees.solidarites-sante.gouv.fr/etudes-et-statistiques/publications/etudes-et-resultats/article/les-medecins-d-ici-a-2040-une-population-plus-jeune-plus-feminisee-et-plus` (Visité le 2020-12-15).

[9] Bae, I.-H. An ontology-based approach to ADL recognition in smart homes. *Future Generation Computer Systems 33* (Apr. 2014), 32–41.

[10] Bai, J., and Perron, P. Estimating and testing linear models with multiple structural changes. *Econometrica* (1998), 47–78.

[11] Bai, Y., and Ku, Y. Automatic room light intensity detection and control using a microprocessor and light sensors. *IEEE Transactions on Consumer Electronics 54*, 3 (2008), 1173–1176.

[12] Bank, W. Age dependency ratio, old (% of working-age population). Disponible sur : `https://data.worldbank.org/indicator/SP.POP.DPND.OL` (Visité le 2018-05-14), 2017.

[13] Bellei, C. Changepoint detection. part i - a frequentist approach. Available from : `http://www.claudiobellei.com/2016/11/15/changepoint-frequentist/`, 11 2016.

[14] Bellot, D., Boyer, A., and Charpillet, F. A new definition of qualified gain in a data fusion. In *The Fifth International Conference on Information Fusion - FUSION'2002* (Annapolis, Maryland, USA, 2002), p. 8 p. Colloque avec actes et comité de lecture. internationale.

[15] Betancourt, M. A conceptual introduction to hamiltonian monte carlo. *arXiv : Methodology* (2017).

[16] Bhagat, A., Kshirsagar, N., Khodke, P., Dongre, K., and Ali, S. Penalty parameter selection for hierarchical data stream clustering. *Procedia Computer Science 79* (2016), 24–31.

[17] Breiman, L. Random forests. *Machine Learning 45*, 1 (2001), 5–32.

[18] Bujnowski, A., Palinski, A., and Wtorek, J. An intelligent bathroom. In *2011 Federated Conference on Computer Science and Information Systems (FedCSIS)* (2011), pp. 381–386.

[19] Bureau, U. S. C. The majority of children live with two parents, census bureau reports. Tech. Rep. CB16-192, United States Census Bureau, 2016. Disponible sur : `https://www.census.gov/newsroom/press-releases/2016/cb16-192.html` (Visité le 2018-05-03).

[20] Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., and Sheikh, Y. A. Openpose : Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

[21] Casale, P., Pujol, O., and Radeva, P. Human activity recognition from accelerometer data using a wearable device. In *Pattern Recognition and Image Analysis*. Springer Berlin Heidelberg, 2011, pp. 289–296.

[22] Chang, D., and Nesbitt, K. V. Developing gestalt-based design guidelines for multi-sensory displays. In *Proceedings of the 2005 NICTA-HCSNet Multimodal User Interaction Workshop - Volume 57* (AUS, 2006), MMUI '05, Australian Computer Society, Inc., p. 9–16.

[23] Chen, R., Li, D., and Liu, Y. An activity of daily living primitive-based recognition framework for smart homes with discrete sensor data. *International Journal of Distributed Sensor Networks 13*, 12 (Dec. 2017), 155014771774949.

[24] Clark, R. A., Mentiplay, B. F., Hough, E., and Pua, Y. H. Three-dimensional cameras and skeleton pose tracking for physical function assessment : A

review of uses, validity, current developments and kinect alternatives. *Gait & Posture 68* (Feb. 2019), 193–200.

[25] Colombel, J., Bonnet, V., Daney, D., Dumas, R., Seilles, A., and Charpillet, F. Physically Consistent Whole-Body Kinematics Assessment Based on an RGB-D Sensor. Application to Simple Rehabilitation Exercises. *Sensors 20*, 10 (Jan. 2020), 18p.

[26] Cook, D. J. Learning setting-generalized activity models for smart spaces. *IEEE intelligent systems 27*, 1 (2012), 32–38.

[27] Cuttone, A., Bækgaard, P., Sekara, V., Jonsson, H., Larsen, J. E., and Lehmann, S. Sensiblesleep : A bayesian model for learning sleep patterns from smartphone events. *PLOS ONE 12*, 1 (01 2017), 1–20.

[28] Daher, M., Diab, A., El Badaoui El Najjar, M., Khalil, M., and Charpillet, F. Elder Tracking and Fall Detection System using Smart Tiles. *IEEE Sensors Journal 17*, 2 (Jan. 2017).

[29] Datar, A., Jain, A., and Sharma, P. Performance of blackman window family in m-channel cosine modulated filter bank for ECG signals. In *2009 International Multimedia, Signal Processing and Communication Technologies* (Mar. 2009), IEEE.

[30] de Lorraine, U. Une plateforme innovante figurant l'habitat du future. Disponible sur : `https://factuel.univ-lorraine.fr/node/2858` (Visité le 2020-12-16), 2015.

[31] Dib, A., and Charpillet, F. Pose Estimation For A Partially Observable Human Body From RGB-D Cameras. In *IEEE/RJS International Conference on Intelligent Robots and Systems (IROS)* (Hamburg, Germany, Sept. 2015), p. 8.

[32] Dubois, A., and Charpillet, F. Human activities recognition with RGB-depth camera using HMM. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (July 2013), IEEE.

[33] Dubois, A., and Charpillet, F. A gait analysis method based on a depth camera for fall prevention. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Aug 2014), pp. 4515–4518.

[34] Dubois, A., and Charpillet, F. Measuring frailty and detecting falls for elderly home care using depth camera. *JAISE - Journal of Ambient Intelligence and Smart Environments 9*, 4 (June 2017), 469 – 481.

[35] Edwards, A. W., and Cavalli-Sforza, L. L. A method for cluster analysis. *Biometrics* (1965), 362–375.

[36] Erdman, C., and Emerson, J. W. A fast bayesian change point analysis for the segmentation of microarray data. *Bioinformatics 24*, 19 (2008), 2143–2148.

[37] Fibaro. Fibaro motion sensors. `https://www.fibaro.com/en/products/motion-sensor/` (Visited on 07/05/2018), 2014.

[38] Genworth. Cost of care survey 2017 - summary, 2017. Disponible sur : `https://www.genworth.com/dam/Americas/US/PDFs/Consumer/corporate/cost-of-care/131168_081417.pdf` (Visité le 2018-05-03).

[39] GfK. A third of people track their health or fitness. Tech. rep., GfK, 2016. Available from : `https://www.gfk.com/insights/press-release/a-third-of-people-track-their-health-or-fitness-who-are-they-and-why-are-they-doing` (Visited on 2018-05-03).

[40] Gu, W., Choi, J., Gu, M., Simon, H., and Wu, K. Fast change point detection for electricity market analysis. In *2013 IEEE International Conference on Big Data* (2013), IEEE, pp. 50–57.

[41] Hevesi, P., Wille, S., Pirkl, G., Wehn, N., and Lukowicz, P. Monitoring household activities and user location with a cheap, unobtrusive thermal sensor array. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp'14 Adjunct* (2014), ACM Press.

[42] Hewson, D., Duchene, J., Charpillet, F., Saboune, J., Michel-Pellegrino, V., Amoud, H., Doussot, M., Paysant, J., Boyer, A., and Horgel, J.-Y. The PARAChute Project : Remote Monitoring of Posture and Gait for Fall Prevention. *EURASIP Journal on Advances in Signal Processing 2007*, ID : 27421 (2007), 15 pages. Editeurs scientifiques : Francesco G. B. De Natale, Aggelos K. Katsaggelos, Oscar Mayora, and Ying Wu ; ISSN : 1687-6172 ; e-ISSN : 1687-6180.

[43] Hsu, C.-Y., Ahuja, A., Yue, S., Hristov, R., Kabelac, Z., and Katabi, D. Zero-effort in-home sleep and insomnia monitoring using radio signals. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 1*, 3 (Sept. 2017), 59 :1–59 :18.

[44] Hu, Y.-N., Hu, G.-C., Hsu, C.-Y., Hsieh, S.-F., and Li, C.-C. Assessment of individual activities of daily living and its association with self-rated health in elderly people of taiwan. *International Journal of Gerontology 6*, 2 (June 2012), 117–121.

[45] Inserm. Sommeil : Faire la lumière sur notre activité nocturne. `https://www.inserm.fr/information-en-sante/dossiers-information/sommeil`, 2017.

[46] Jackson, B., Scargle, J. D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L., and Tsai, T. T. An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters 12*, 2 (2005), 105–108.

[47] JARRETT, R. G. A note on the intervals between coal-mining disasters. *Biometrika 66*, 1 (1979), 191–193.

[48] Jeanpierre, L., and Charpillet, F. Automated Medical Diagnosis with Fuzzy Stochastic Models : Monitoring Chronic Diseases. *Acta Biotheoretica 52*, 4 (2004), 291–311. Article dans revue scientifique avec comité de lecture. internationale.

[49] Jiang, L., Li, C., Wang, S., and Zhang, L. Deep feature weighting for naive bayes and its application to text classification. *Engineering Applications of Artificial Intelligence 52* (June 2016), 26–39.

[50] Ju Han, and Bhanu, B. Human activity recognition in thermal infrared imagery. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops* (2005), pp. 17–17.

[51] Juang, B.-H. On the hidden markov model and dynamic time warping for speech recognition-a unified view. *AT&T Bell Laboratories Technical Journal 63*, 7 (Sept. 1984), 1213–1243.

[52] Kaplan, A. Y., and Shishkin, S. L. Application of the change-point analysis to the investigation of the brain's electrical activity. In *Non-parametric statistical diagnosis*. Springer, 2000, pp. 333–388.

[53] Killick, R., and Eckley, I. changepoint : An r package for changepoint analysis. *Journal of statistical software 58*, 3 (2014), 1–19.

[54] Killick, R., Eckley, I. A., Ewans, K., and Jonathan, P. Detection of changes in variance of oceanographic time-series using changepoint analysis. *Ocean Engineering 37*, 13 (2010), 1120–1126.

[55] Killick, R., Fearnhead, P., and Eckley, I. A. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association 107*, 500 (2012), 1590–1598.

[56] Killick, R., Nam, C., Eckley, I., and Aston, J. changepoint.info : The changepoint repository. `http://www.changepoint.info` (Visited on 2021-01-27), 2021.

[57] Kovacshazy, T., and Fodor, G. New approach to passive infrared motion sensors signal processing for ambient assisted living applications. In *2012 IEEE International Instrumentation and Measurement Technology Conference Proceedings* (May 2012), IEEE.

[58] Kutaj, M., and Boroš, M. Development of a new generation of magnetic contact based on hall-effect sensor. *CBU International Conference Proceedings 5* (Sept. 2017), 1154–1158.

[59] Lee, M.-T., Jang, Y., and Chang, W.-Y. How do impairments in cognitive functions affect activities of daily living functions in older adults ? *PLOS ONE 14*, 6 (June 2019), e0218112.

[60] LIPS. LIPSedge™ AT Standalone FPGA-based 3D ToF Camera. Page web du produit : `https://www.lips-hci.com/product-page/lipsedge-at-standalone-fpga-based-3d-tof-camera` (Visité le 2020-07-23), 2019.

[61] Liu, J., Teng, G., and Hong, F. Human activity sensing with wireless signals : A survey. *Sensors 20*, 4 (Feb. 2020), 1210.

[62] Luo, X., Guan, Q., Tan, H., Gao, L., Wang, Z., and Luo, X. Simultaneous indoor tracking and activity recognition using pyroelectric infrared sensors. *Sensors 17*, 8 (July 2017), 1738.

[63] Luo, X., Guan, Q., Tan, H., Gao, L., Wang, Z., and Luo, X. Simultaneous indoor tracking and activity recognition using pyroelectric infrared sensors. *Sensors 17*, 8 (2017).

[64] Maguire, B. A., Pearson, E., and Wynn, A. The time intervals between industrial accidents. *Biometrika 39*, 1/2 (1952), 168–180.

[65] Malaisé, A. *Human movement learning with wearable sensors : towards ergonomic assessment automation*. Theses, Université de lorraine, July 2020.

[66] Mehrang, S., Pietilä, J., and Korhonen, I. An activity recognition framework deploying the random forest classifier and a single optical heart rate monitoring and triaxial accelerometer wrist-band. *Sensors 18*, 3 (Feb. 2018), 613.

[67] Meng, L., Miao, C., and Leung, C. Towards online and personalized daily activity recognition, habit modeling, and anomaly detection for the solitary elderly through unobtrusive sensing. *Multimedia Tools and Applications 76*, 8 (Jan. 2016), 10779–10799.

[68] Ministère de la santé et des sports, F. Décret $n^o2010-1229$ du 19 octobre 2010 relatif à la télémédecine. Disponible sur : `https://beta.legifrance.gouv.fr/jorf/id/JORFTEXT000022932449/` (Visité le 2020-07-21), 2010.

[69] Mulvenna, M., Carswell, W., Mccullagh, P., Augusto, J., Zheng, H., Jeffers, P., Wang, H., and Martin, S. Visualization of data for ambient assisted living services. *IEEE Communications Magazine 49*, 1 (Jan. 2011), 110–117.

[70] Murphy, K. P., et al. Naive bayes classifiers. *University of British Columbia 18* (2006), 60.

[71] Nam, C. F., Aston, J. A., and Johansen, A. M. Quantifying the uncertainty in change points. *Journal of Time Series Analysis 33*, 5 (2012), 807–823.

[72] Pawitan, Y. Change-point problem. *Encyclopedia of Biostatistics* (2005).

[73] Pharmagest. Projet 36 mois de plus. Disponible sur : `https://esante.pharmagest.com/logement-intelligent/projet-36-mois-de-plus/` (Visité le 2020-12-16), 2015.

[74] Piltaver, R., Gjoreski, H., and Gams, M. Identifying a person with door-mounted accelerometer. *Journal of Ambient Intelligence and Smart Environments 10*, 5 (Sept. 2018), 361–375.

[75] Pogue, D. What fitbit's 6 billion nights of sleep data reveals about us. Available from : `https://finance.yahoo.com/news/exclusive-fitbits-6-billion-nights-sleep-data-reveals-us-110058417.html` (Visited on 2018-05-03), 2018.

[76] Quinlan, J. R. Induction of decision trees. *Machine learning 1*, 1 (1986), 81–106.

[77] RAFTERY, A. E., and AKMAN, V. E. Bayesian analysis of a poisson process with a change-point. *Biometrika 73*, 1 (1986), 85–89.

[78] Rashidi, P., and Mihailidis, A. A survey on ambient-assisted living tools for older adults. *IEEE Journal of Biomedical and Health Informatics 17*, 3 (May 2013), 579–590.

[79] Reeves, J., Chen, J., Wang, X. L., Lund, R., and Lu, Q. Q. A review and comparison of changepoint detection techniques for climate data. *Journal of applied meteorology and climatology 46*, 6 (2007), 900–915.

[80] REZZOUG, I. Grille aggir : comment évaluer la perte d'autonomie ? Disponible sur : `https://fmh-association.org/grille-aggir-comment-evaluer-la-perte-dautonomie` (Visité le 2020-12-16), 2019.

[81] Robben, S., Boot, M., Kanis, M., and Krose, B. Identifying and visualizing relevant deviations in longitudinal sensor patterns for care professionals. In *Proceedings of the ICTs for improving Patients Rehabilitation Research Techniques* (2013), IEEE.

[82] Robert, C., and Casella, G. A short history of markov chain monte carlo : Subjective recollections from incomplete data. *Statistical Science 26*, 1 (Feb. 2011), 102–115.

[83] Rose, C., Saboune, J., and Charpillet, F. Reducing Particle Filtering Complexity for 3D Motion Capture using Dynamic Bayesian Networks. In *Twenty-Third Conference on Artificial Intelligence - AAAI-08* (Chicago, United States, July 2008).

[84] Saboune, J., and Charpillet, F. Markerless human motion capture for Gait analyis. In *3rd European Medical and Biological Engineering Conference - EMBEC'05* (Prague, République Tchèque, Nov. 2005).

[85] Scott, A. J., and Knott, M. A cluster analysis method for grouping means in the analysis of variance. *Biometrics* (1974), 507–512.

[86] Sen, A., and Srivastava, M. S. On tests for detecting change in mean. *The Annals of statistics* (1975), 98–108.

[87] Shoaib, M., Bosch, S., Incel, O., Scholten, H., and Havinga, P. Complex human activity recognition using smartphone and wrist-worn motion sensors. *Sensors 16*, 4 (Mar. 2016), 426.

[88] Singh, S., and Gupta, P. Comparative study id3, cart and c4. 5 decision tree algorithm : a survey. *International Journal of Advanced Information Science and Technology (IJAIST) 27*, 27 (2014), 97–103.

[89] Sleeptracker. Sleeptracker. `https://sleeptracker.com` (Visited on 2018-05-03), 2017.

[90] Tartakovsky, A. G., Rozovskii, B. L., Blazek, R. B., and Hongjoong Kim. A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods. *IEEE Transactions on Signal Processing 54*, 9 (2006), 3372–3382.

[91] Tegou, T., Kalamaras, I., Tsipouras, M., Giannakeas, N., Votis, K., and Tzovaras, D. A low-cost indoor activity monitoring system for detecting frailty in older adults. *Sensors 19*, 3 (Jan. 2019), 452.

[92] Uddin, M., Khaksar, W., and Torresen, J. Ambient sensors for elderly care and independent living : A survey. *Sensors 18*, 7 (June 2018), 2027.

[93] Van den Burg, G. J., and Williams, C. K. An evaluation of change point detection algorithms. *arXiv preprint arXiv :2003.06222* (2020).

[94] Venkatnarayan, R. H., and Shahzad, M. Enhancing indoor inertial odometry with WiFi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 3*, 2 (June 2019), 1–27.

[95] Wang, S., Skubic, M., and Zhu, Y. Activity density map visualization and dissimilarity comparison for eldercare monitoring. *IEEE Transactions on Information Technology in Biomedicine 16*, 4 (July 2012), 607–614.

[96] Wang, Y., Cang, S., and Yu, H. A survey on wearable sensor modality centred human activity recognition in health care. *Expert Systems with Applications 137* (Dec. 2019), 167–190.

[97] Zeileis, A., Shah, A., and Patnaik, I. Testing, monitoring, and dating structural changes in exchange rate regimes. *Computational Statistics & Data Analysis 54*, 6 (2010), 1696–1706.

Résumé

L'augmentation de la population sénior se révèle être un enjeu de taille et une grande question de santé publique. La part démographique des personnes âgées s'agrandit de plus en plus grâce au progrès et avancés de la médecine et nos systèmes de santé. Néanmoins le vieillissent de cette population implique naturellement une pléthore de problèmes de dépendance qui y sont associés, ceci bien sûr exponentiellement.

Les maisons de retraite sont des solutions généralement coûteuses et très peu appréciées. En conséquence, des solutions plus adaptées basées sur l'aide au maintien à domicile ce développement de plus en plus ces dernières années.

Cette problématique se retrouve dans la croisée des chemins entre les technologies de capteurs, la télétransmission de données, l'assistance aux personnes âgées à mobilité réduite et le suivi d'activités.

Cette thèse explore l'application d'algorithmes d'analyse de données pour le suivi d'activités des personnes âgées à domicile. L'idée étant qu'un suivi régulier des résidents permet d'inférer leur état de dépendance ou d'autonomie et permet aux personnels soignants d'intervenir en cas de détection d'un début de dégradation.

Nous avons exploré et adapté certaines méthodes d'inférence bayésienne et segmentation de séries temporelles pour la reconnaissance d'activités. Et ensuite, nous avons proposé un outil de visualisation permettant de faciliter la détection d'anomalies ou changements de rythme de vie.

Ce travail s'inscrit dans le cadre d'une thèse CIFRE. Ainsi tous les méthodes et algorithmes explorés ont été mis en production et sont exploités par la solution d'aide à domicile commercialisé par la société Diatelic.

Mots-clés: Intelligence artificielle, Décision, Reconnaissance des activités de la vie quotidienne, Habitat intelligent

Abstract

The increase of the senior population constitutes a major public health issue. The demographic share of the elderly is ever more growing thanks to the progress and advances in medicine and our health care systems. However, with the aging of this population comes a plethora of dependency problems, and this, of course, exponentially.

Retirement homes are an expensive and not very popular solution. As a result, we are seeing a surge in home assisted living solutions in the recent years.

This topic is in the crossroads between sensor technologies, data transmission, assistance to elderly people and activity monitoring.

This thesis explores the application of data analysis algorithms for activity monitoring of elderly people at home. The idea is that with day-to-day monitoring of residents it is

possible to infer their autonomy and capacity to perform day-to-day tasks. It also allows caregivers to intervene in cases where the start of some degradation is detected.

We explored and adapted some Bayesian inference and time series segmentation methods for activity recognition. And then, we proposed a visualization tool to facilitate the detection of anomalies or changes in everyday habits.

This work is part of a CIFRE thesis. The methods and algorithms presented have been put into production and are packaged into Diatelic's the assisted living commercial solution.

Keywords: Artificial intelligence, Decision, Activities of Daily Living recognition, Smart home