Fusion de Données Multicapteurs pour un Système de Télésurveillance Médicale de Personnes à Domicile

Florence Duchêne

UNIVERSITÉ JOSEPH FOURIER
Sciences, Technologie, Médecine – Grenoble 1

Institut National Polytechnique de Grenoble
École Doctorale en Électronique, Électrotechnique, Automatique et Traitement du signal (EEATS)

THÈSE

pour l'obtention du grade de

DOCTEUR DE L'UNIVERSITÉ JOSEPH FOURIER

Secteur de Recherche : Traitement du Signal et des Images

présentée et soutenue publiquement
le vendredi 15 octobre 2004

par

Florence DUCHÊNE

Fusion de données multicapteurs pour un système de télésurveillance médicale de personnes à domicile

Co-Directeurs de Thèse : Catherine GARBAY et Vincent RIALLE

Composition du jury

Rapporteurs :	Marie-Odile CORDIER	Professeur des Universités, Rennes 1
	Jean-Pierre THOMESSE	Professeur des Universités, ENSEM–Nancy
Examinateurs :	Michèle ROMBAUT	Professeur des Universités, UJF–Grenoble
	Bernard LEFEBVRE	Professeur des Universités, UQAM–Montréal (Canada)
Directeurs :	Catherine GARBAY	Directrice de Recherche, CNRS–Grenoble
	Vincent RIALLE	Maître de Conférences-Praticien Hospitalier, UJF–Grenoble

Laboratoire des Techniques de l'Imagerie, de la Modélisation et de la Cognition, TIMC - UMR CNRS 5525
Institut d'Ingénierie de l'Information de Santé, Faculté de médecine de Grenoble, 38 706 La Tronche cedex, France

Remerciements

Je tiens tout d'abord à remercier les co-directeurs de ce travail de thèse. Un grand merci à Madame Catherine Garbay qui m'a soutenue chaleureusement et m'a guidée dans l'analyse et le traitement rigoureux des étapes successives de recherche et des problèmes rencontrés. Je remercie aussi beaucoup Monsieur Vincent Rialle pour son aide plus spécifique dans l'exploration de la problématique complexe de télésurveillance médicale à domicile.

J'aimerais aussi remercier Monsieur Jacques Demongeot qui m'a accueillie au sein du laboratoire TIMC à Grenoble dans le cadre de cette thèse de doctorat, ainsi que Monsieur Norbert Noury qui m'a intégrée dans le projet de recherche de son équipe autour du concept d'habitat intelligent pour la santé. Je remercie aussi tous les membres des équipes AFIRM et SIC pour leur accueil et leur soutien.

De l'autre côté de l'atlantique, je remercie vivement Monsieur Bernard Lefebvre de m'avoir accueillie à l'Université du Québec à Montréal pour plusieurs séjours d'une durée totale de près d'un an durant ma thèse. Les échanges entretenus avec l'équipe du projet DOMUS de l'Université de Sherbrooke (Québec) ont également été très constructifs pour l'évolution de mon travail. Je remercie notamment sincèrement Madame Hélène Pigot pour le temps qu'elle m'a consacré et ses nombreux conseils liés à son expertise de la vie quotidienne des personnes âgées.

Un grand merci aussi à toutes les personnes qui m'ont guidée aux différentes étapes de la construction et de la validation de ce travail. Je voudrais remercier en particulier Madame Sylvie Charbonnier qui m'a permis d'accéder à un ensemble de données expérimentales qui se sont avérées indispensables à la réalisation de cette étude. Un grand merci aussi au Docteur Pierre Rumeau pour la transmission d'informations essentielles à la compréhension du contexte de recherche étroitement lié à sa spécialité de praticien-hospitalier en gériatrie. Merci aussi à Monsieur Olivier Gaudoin pour son expertise en statistiques et à Mademoiselle Céline Chevrier pour la transmission de ses connaissances et de son expérience en physiologie et chronobiologie.

Je suis également reconnaissante envers Madame Marie-Odile Cordier et Monsieur Jean-Pierre Thomesse d'avoir accepté d'être les rapporteurs de ma thèse, ainsi qu'envers Madame Michèle Rombaut et Monsieur Bernard Lefebvre pour leur participation au jury de soutenance.

Merci enfin à tout mon entourage, famille et amis, qui m'ont soutenue durant ces dernières années. Un gros merci tout particulier à Florence qui m'a donné de si précieux conseils pour le bon déroulement de ma thèse.

Sommaire

Table des figures **xi**

Liste des tableaux **xv**

Introduction générale **1**

Partie I Présentation de la problématique **3**

1 Contexte de recherche **5**
1.1 Télémédecine 5
1.1.1 Les apports et enjeux de la télémédecine 6
1.1.2 Les freins au développement 7
1.2 Télésurveillance médicale à domicile 8
1.2.1 Objectifs 8
1.2.2 Principe 9
1.2.3 Enjeux 9

2 État de l'art de la télésurveillance médicale à domicile **11**
2.1 Des projets variés dans le monde 11
2.1.1 Architecture et expérimentation globale de systèmes d'information . . 11
2.1.2 Systèmes de surveillance au domicile 12
2.1.3 Système de gestion et de stockage des données 12

2.1.4 Assistants intelligents pour l'analyse de données 12
2.1.5 Détection des situations critiques d'une personne à domicile 13
2.2 Description du projet HIS 14
2.2.1 Construction d'un appartement prototype 14
2.2.2 Collaborations nationales et internationales 15

3 La problématique de décision sur la situation d'une personne à domicile 17
3.1 Proposition d'une architecture de décision 17
3.1.1 Objectifs de décision 17
3.1.2 Système d'apprentissage et de décision 18
3.1.3 Définition de plusieurs niveaux de décision 18
3.1.4 Fusion de plusieurs types de données 20
3.1.5 Architecture du système : une approche granulaire 21
3.2 Formulation de la problématique
Étude des habitudes de la vie quotidienne 22
3.3 Principe de résolution
Vers l'apprentissage d'un profil comportemental 23
3.3.1 Caractéristiques des régularités observées dans les données 23
3.3.2 Résolution de l'identification des régularités comme un cycle de décision 24
3.4 Contraintes de résolution
Contexte et objectifs de la télésurveillance médicale 26
3.5 Synthèse de la résolution du problème de décision 27

Partie II Processus de simulation 29

**1 Introduction :
Pourquoi un processus de simulation ? 31**

2 État de l'art 33

3 Méthodologie pour la simulation **37**
3.1 Démarche incrémentale . . . 37
3.1.1 La simulation : une partie intégrante du cycle de résolution d'un problème 37
3.1.2 La simulation : un cycle de raffinement à part entière . . . 38
3.2 Approche hybride . . . 39
3.3 Techniques de simulation . . . 40
3.3.1 Généralités . . . 40
3.3.2 Principe de modélisation . . . 41
3.3.3 Principe de validation opérationnelle . . . 41
3.4 Synthèse . . . 42

4 Contexte de la télésurveillance médicale à domicile **43**
4.1 Paramètres observables . . . 43
4.2 Niveau de simulation . . . 45
4.3 Connaissances et données utiles à la simulation . . . 46
4.3.1 Connaissances *a priori* . . . 46
4.3.2 Données expérimentales . . . 51
4.4 Synthèse . . . 53

5 Modélisation **55**
5.1 Principe de construction du modèle . . . 55
5.2 Sous-modèles de simulation . . . 57
5.2.1 Déplacements . . . 57
5.2.2 Postures . . . 57
5.2.3 Niveau d'activité . . . 58
5.2.4 Fréquence cardiaque . . . 61
5.3 Synthèse . . . 71

6 Expérimentation et validation opérationnelle **73**
6.1 Contexte d'expérimentation et de validation . . . 73
6.1.1 Validation par les experts . . . 74
6.1.2 Validation par analyses mathématiques et statistiques . . . 75
6.2 Discussion sur la qualité des résultats . . . 77
6.3 Introduction d'une contrainte temporelle supplémentaire . . . 79
6.3.1 Principe de modélisation . . . 79
6.3.2 Validation conceptuelle . . . 80
6.3.3 Implémentation . . . 81
6.3.4 Validation opérationnelle . . . 82

6.4 Discussion sur le cycle de raffinement de la simulation 82
6.5 Simulation de modifications de comportement 87
6.5.1 Quelles modifications possibles du comportement ? 87
6.5.2 Modifications “normales” 88
6.5.3 Modifications inquiétantes 93
6.6 Synthèse 96

7 Discussion et Conclusion **97**

Partie III Système de décision **101**

1 Introduction **103**

2 État de l’art **107**
2.1 Fouille de données temporelles 107
2.1.1 Fouille de caractères dans les séquences temporelles 108
2.1.2 Classification des caractères 109
2.2 Représentation des séquences temporelles 110
2.2.1 Représentation à valeurs réelles 110
2.2.2 Discrétisation 111
2.2.3 Abstraction 111
2.2.4 Application de l’abstraction dans notre contexte 112
2.3 Mesure de similarité 113
2.3.1 Distance Euclidienne 113
2.3.2 Distance DTW 113
2.3.3 Distance $LCSS$ 113
2.3.4 Distance minimum discrète 114
2.4 Synthèse du positionnement par rapport à l’état de l’art 115

3 Méthodologie pour l'extraction de motifs temporels **117**
3.1 Problème considéré
Extraction de motifs pour la télésurveillance médicale . . . 117
3.2 Contexte expérimental
Critères guidant la construction d'une méthode d'extraction . . . 118
3.3 Système de décision
Méthodologie pour l'extraction de motifs . . . 119
3.4 Synthèse . . . 123

4 Mesure de similarité **125**
4.1 Distance homogène pour des composantes hétérogènes . . . 125
4.2 Distance réelle entre séquences temporelles . . . 126
4.2.1 Notion de plus longue sous-séquence commune, *LCSS* . . . 126
4.2.2 Extension d'une distance *LCSS* . . . 128
4.2.3 Calcul effectif de la distance *LCSS* . . . 129
4.3 Distance minimum . . . 130
4.3.1 Définition de la distance minimum . . . 130
4.3.2 Extension aux contraintes multidimensionnelles et hétérogènes . . . 131
4.4 Synthèse . . . 133

5 Approche proposée pour l'extraction de motifs temporels **135**
5.1 Abstraction des séquences temporelles . . . 136
5.1.1 Prétraitement . . . 136
5.1.2 Discrétisation . . . 136
5.1.3 Agrégation temporelle . . . 137
5.2 Fouille de caractères pour l'extraction de tentatives de motifs . . . 138
5.2.1 Objectifs et contraintes de la fouille de caractères . . . 138
5.2.2 Projections aléatoires . . . 142
5.2.3 Examen de la matrice de collisions . . . 144
5.2.4 Synthèse des tentatives de motifs . . . 148
5.3 Classification pour l'identification des motifs . . . 155
5.4 Synthèse . . . 158

6 Expérimentation et Validation **161**
6.1 Processus expérimental . . . 161
6.1.1 Contexte expérimental : la télésurveillance médicale . . . 161
6.1.2 Données expérimentales . . . 162
6.1.3 Méthode d'expérimentation . . . 164

6.1.4 Mesures de performance 164
6.1.5 Synthèse du processus expérimental 170
6.2 Qualité de la méthode 172
6.2.1 Mesure de similarité 172
6.2.2 Abstraction 182
6.2.3 Fouille de données 186
6.2.4 Synthèse sur la qualité de la méthode 192
6.3 Qualité des résultats 194
6.3.1 Exigences et paramètres clés du système 194
6.3.2 Performances et paramètres clés du système 195
6.3.3 Test de Sensibilité : modifications “normales” de comportement 207
6.3.4 Test de Spécificité : modifications inquiétantes de comportement . . . 209
6.4 Validation de la simulation par la décision 211
6.5 Synthèse 211

7 Discussion et Conclusion **213**

Conclusion et Perspectives **217**

Bibliographie **221**

Annexes **229**

A Journée type d’une personne âgée **229**

B Situations inquiétantes à domicile **231**
B.1 Infection urinaire 231
B.2 Insuffisance cardiaque 231
B.3 Dépression 232

C Tests statistiques sur la nature d’un échantillon **233**
C.1 Test d’ajustement graphique 233
C.2 Test de Kolmogorov-Smirnov 233
C.3 Test de Lilliefors 234
C.4 Application 234

D Données expérimentales **235**

E Analyse des données de modélisation **239**
E.1 Niveau d'activité . . . 239
E.1.1 Caractéristiques relatives des distributions . . . 239
E.1.2 Caractéristiques intrinsèques des distributions . . . 241
E.2 Fréquence cardiaque de repos . . . 244
E.3 Coût cardiaque d'une activité . . . 245
E.4 Influence de l'activité végétative sur la fréquence cardiaque . . . 247

F Implémentation du modèle de simulation **249**
F.1 Principe d'implémentation . . . 249
F.2 Interface et implémentation de la simulation . . . 249
F.2.1 Simulation d'un individu donné dans une certaine situation . . . 249
F.2.2 Simulation de modifications "normales" de comportement . . . 250
F.3 Structure des fichiers de paramétrisation . . . 253
F.4 Structure globale du programme de simulation . . . 257

G Implémentation de l'extraction de motifs **259**
G.1 Interface de l'extraction de motifs . . . 259
G.1.1 Définition des paramètres de l'extraction . . . 259
G.1.2 Analyse des résultats et performances du système . . . 259
G.1.3 Évaluation à grande échelle des performances du système . . . 260
G.2 Structure globale du programme d'extraction de motifs . . . 260

H Plus Longue Sous-séquence Commune **265**
H.1 Notion de plus longue sous-séquence commune . . . 265
H.2 Concepts de base . . . 266
H.3 Algorithme de résolution par programmation dynamique . . . 266
H.4 Algorithme de résolution par contours . . . 268
H.5 Algorithme implémenté dans notre application . . . 269
H.5.1 Présentation de l'algorithme . . . 269
H.5.2 Application sur un exemple . . . 271
H.5.3 Contexte de notre application . . . 272

I Mesure de distance selon le principe *DTW* **275**
I.1 Principe . . . 275
I.2 Calcul de la distance . . . 275
I.3 Calcul effectif . . . 276

J Liste des publications **277**

J.1 Revues 277

J.2 Congrès internationaux avec actes et comité de lecture 277

J.3 Congrès nationaux avec actes et comité de lecture 277

Table des figures

Partie I Présentation de la problématique **5**

1.1 Système d'information de la télésurveillance médicale à domicile. 9

2.1 L'HIS : appartement prototype mis en place à la faculté de médecine de Grenoble. 14

3.1 Une approche granulaire pour la détection de situations critiques. 21
3.2 Les étapes de résolution d'un problème de décision. 25

Partie II Processus de simulation **31**

3.1 Les étapes de mise en place d'un processus de simulation [100]. 38
3.2 Intégration des différents types de connaissances disponibles dans le cycle de construction et de validation d'un processus de simulation. 39
3.3 Synthèse de la méthodologie proposée pour la simulation. 42

4.1 Coût énergétique et accroissement de la fréquence cardiaque de repos pour diverses postures classées subjectivement par ordre de pénibilité. [31, 74] 47
4.2 Variation de la fréquence cardiaque pendant et après une marche ou course de 1600 mètres, à différentes vitesses. [74] 50
4.3 Données enregistrées d'une personne dans sa vie quotidienne. 52
4.4 Synthèse de la méthodologie proposée pour la simulation. 53

5.1 Structure en cascade pour la simulation. 56
5.2 Automate à états finis pour la simulation des postures. 58
5.3 Distribution des valeurs de niveau d'activité pour une personne effectuant des mouvements en posture allongée. 59
5.4 Test d'ajustement graphique à une loi normale de la distribution des valeurs de niveau d'activité enregistrées en posture allongée. 60
5.5 Coût cardiaque d'un travail musculaire : ΔFc. 62
5.6 Paramètres obtenus de l'analyse d'un rythme sinusoïdal par la méthode du cosinor. 63
5.7 Approximation sinusoïdale du rythme circadien de la fréquence cardiaque de repos. 65
5.8 Coefficients de corrélation linéaire entre fréquence cardiaque et niveau d'activité. 66
5.9 Moyenne et écart-type du coût cardiaque d'une activité. 67
5.10 Nombre de données expérimentales disponibles sur l'échelle du logarithme népérien des niveaux d'activité observés. 68

5.11 Écart-type du coût cardiaque des niveaux d'activité quelle que soit la posture. . . 69
5.12 Distribution des valeurs de coût cardiaque. 70
5.13 Principe de génération des valeurs de fréquence cardiaque. 71
5.14 Synthèse des techniques de construction et de validation du modèle de simulation. 72

6.1 Séquence temporelle à 4 dimensions générée par le processus de simulation. . . . 74
6.2 Séquence temporelle à 4 dimensions reconstituée à partir des données enregistrées par un système réel en fonction du contexte expérimental de la simulation. 76
6.3 Séquences temporelles représentatives d'une personne télésurveillée et mettant en évidence la nécessité de prise en compte d'une contrainte temporelle supplémentaire. 78
6.4 Relation moyenne entre la différence absolue entre deux valeurs successives dans le temps et la première de ces deux valeurs. 80
6.5 Séquences temporelles représentatives d'une personne télésurveillée et mettant en évidence les effets d'une réorganisation temporelles. 83
6.6 Effets de la réorganisation temporelle. 84
6.7 Distribution des couples de fréquence cardiaque et de niveau d'activité moyen pendant les deux minutes précédent la mesure de fréquence cardiaque. 85
6.8 Simulation de plusieurs profils de personnes et plusieurs types de situation. . . . 86
6.9 Observation expérimentale de la réalisation d'une même activité pour un sujet donné. 89
6.10 Simulation de modifications normales de comportement. 90
6.11 Simulation de modifications inquiétantes de comportement. 94

Partie III Système de décision **103**

2.1 Comparaison de deux séquences de même forme mais non alignées sur l'axe du temps. 114
2.2 Comparaison de deux séquences globalement identiques mis à part pour quelques points présents dans l'une mais pas dans l'autre. 115

3.1 Construction d'un système de reconnaissance. 120
3.2 Synthèse des étapes de la fouille de données temporelles. 123

4.1 La notion de similarité basée sur $LCSS$ et contrainte par ϵ. 127
4.2 La notion de similarité basée sur $LCSS$ et contrainte par δ. 127
4.3 Mise en évidence des points similaires identifiés dans la comparaison de deux séquences A et B. 129

5.1 Principe d'utilisation de la méthode des projections aléatoires pour l'identification des sous-séquences récurrentes d'une série temporelle à une dimension [28]. 139
5.2 Illustration de l'extension nécessaire des sous-séquences de base similaires identifiées par l'examen des fortes valeurs de la matrice de collisions. 141
5.3 Principe de la méthode des projections aléatoires proposée. 143
5.4 Voisinages de collisions pour l'identification et l'extension de sous-séquences récurrentes. 145
5.5 Principe d'examen de la matrice de collisions. 147
5.6 Illustration des objectifs de classification des sous-séquences identifiées à l'issue de l'examen de la matrice de collisions 149

5.7 Illustration des cas possibles de constitution d'une classe. 151
5.8 Illustration des cas possibles de division d'une classe non valide d'au moins trois éléments. 152
5.9 Illustration du cas d'une classe non valide dont la meilleure division est obtenue en conservant les deux sous-séquences k_1 et k_2 qui ne se superposent pas. 153
5.10 Synthèse des tentatives de motifs. 154
5.11 Classification hiérarchique ascendante. 156
5.12 Calcul du représentant d'une classe. 157
5.13 Synthèse des étapes successives de l'extraction de motifs à partir de données enregistrées par un ensemble de capteurs. 159

6.1 Synthèse du processus d'expérimentation de l'extraction de motifs. 171
6.2 Présentation de deux ensembles de séquences expérimentales. 173
6.3 Distances $LCSS$ observées en moyenne entre les séquences de la classe 0 et la séquence de référence n°0, en fonction du choix des paramètres ϵ_{LCSS} et δ_{LCSS}. . 175
6.4 Représentation des distances DTW et $LCSS$ entre chaque séquence des classes *a priori* 0 et 1 et la séquence de référence n°0. 177
6.5 Représentation des couples de points considérés similaires dans le calcul des distances $LCSS$ et DTW pour la comparaison des séquences 0 et 16. 178
6.6 Distances $LCSS$ observées en fonction du choix des paramètres ϵ_{LCSS} et δ_{LCSS}. . 179
6.7 Influence relative du choix des paramètres ϵ_{LCSS} et δ_{LCSS}. 181
6.8 Illustration du processus d'abstraction. 183
6.9 Sélection du "meilleur" seuil minimum de collisions en fonction des collisions observées pour les séquences des classes 0 et 1 avec la séquence de référence, et de la dimension de projection. 190
6.10 Réduction de la dimension de l'espace de recherche de sous-séquences récurrentes par la fouille de caractères. 191
6.11 Performances de l'identification des tentatives de motifs et de leur classification en motifs dans un contexte non bruité. 197
6.12 Performances de l'identification des tentatives de motifs et de leur classification en motifs dans un contexte faiblement bruité. 200
6.13 Indices de fractionnement des instances de motifs dans un contexte faiblement bruité, en fonction des paramètres clés du système. 201
6.14 Performances de l'identification des tentatives de motifs et de leur classification en motifs relativement pour différentes valeurs des seuils minimum de collisions (c_{min}) et maximum de distance (d_{max}) et de l'écart maximum de similarité sur les valeurs (ϵ_{LCSS}). 204
6.15 Illustration de l'extraction des instances d'un motif insérés initialement dans une séquence de données simulée à partir de déplacements aléatoires. 206
6.16 Performances de l'identification des tentatives de motifs et de leur classification en motifs en fonction des taux de bruit insérés entre les instances de motifs. . . . 208
6.17 Illustration de l'introduction d'une modification inquiétante de comportement entre les instances d'un motif. 210
6.18 Caractéristiques des motifs identifiés à partir de l'analyse d'une séquence simulée pour un individu "moyen" pendant sept journées consécutives. 212

Annexes **229**

D.1 Enregistrements du sujet 1. 236
D.2 Enregistrements du sujet 2. 236
D.3 Enregistrements du sujet 3. 236
D.4 Enregistrements du sujet 4. 237
D.5 Enregistrements du sujet 5. 237
D.6 Enregistrements du sujet 6. 237
D.7 Enregistrements du sujet 7. 238
D.8 Enregistrements du sujet 8. 238

E.1 Distribution des valeurs de niveau d'activité observées alors que la personne effectue des mouvements en posture allongée. 240
E.2 Distribution des valeurs de niveau d'activité observées alors que la personne effectue des mouvements en posture assise. 240
E.3 Distribution des valeurs de niveau d'activité observées alors que la personne effectue des mouvements en posture debout. 240
E.4 Distribution des valeurs de niveau d'activité observées alors que la personne marche. 240
E.5 Distribution des valeurs de niveau d'activité observées alors que la personne effectue des mouvements en posture allongée. 242
E.6 Distribution des valeurs de niveau d'activité observées alors que la personne effectue des mouvements en posture assise. 242
E.7 Distribution des valeurs de niveau d'activité observées alors que la personne effectue des mouvements en posture debout. 243
E.8 Distribution des valeurs de niveau d'activité observées alors que la personne marche. 243
E.9 Moyenne du coût cardiaque en fonction des niveaux d'activité observés, pour chaque type de posture. 246
E.10 Observation des tendances de variation du coût cardiaque en fonction des périodes d'alimentation sur la deuxième journée de surveillance du premier sujet de l'expérimentation. 248

F.1 Principe de l'implémentation du processus de simulation. 250
F.2 Interface du processus de simulation. 255
F.3 Interface d'introduction de motifs bruités dans une séquence de données. 256

G.1 Interface du processus d'extraction de motifs. 262
G.2 Interface de l'analyse des résultats et des performances de l'extraction. 263

H.1 Construction du tableau L pour la comparaison des deux mots $x = ATCGTT$ et $y = CTACTAATA$. 267
H.2 Tableau $L[1 \dots m, 1 \dots n]$ résultant de la programmation dynamique. 269
H.3 Fonctionnement de l'algorithme de recherche de la plus longue sous-séquence commune. [4] 273

I.1 Illustration sur un exemple de la recherche d'un chemin DTW. 276

Liste des tableaux

Partie I Présentation de la problématique **5**

Partie II Processus de simulation **31**

4.1 Caractéristiques des activités de la vie quotidienne (AVQ). 48
4.2 Classification des travaux physiques d'après la fréquence cardiaque. [30, 74] . . . 51

5.1 Moyenne, médiane et mode du niveau moyen d'activité en fonction de la posture associée. 60
5.2 Paramètres des variations sinusoïdales de repos observés en moyenne pour les sujets de l'expérimentation, en référence aux valeurs citées dans [46, 74]. 65

6.1 Coefficients de corrélation linéaire entre le niveau d'activité et la fréquence cardiaque. 87

Partie III Système de décision **103**

4.1 Distances entre symboles utilisées pour le calcul de distance minimum entre deux séquences discrètes. 131
4.2 Distances entre modalités utilisées pour le calcul de distance minimum entre deux séquences discrètes. 132
4.3 Distances entre modalités utilisées pour le calcul de distance minimum entre deux séquences discrètes. 132
4.4 Synthèse des mesures de similarité définies. 133

6.1 Intervalles de discrétisation du niveau d'activité et de la fréquence cardiaque. . . 185
6.2 Pourcentages de collisions observés entre les séquences des classes 0 et 1 et la séquence de référence selon la dimension de projection. 188
6.3 Pourcentages moyens de collisions observés entre les séquences des classes 0 et 1 et la séquence de référence, selon le nombre de projections réalisées. 189
6.4 Paramètres de la méthode proposée pour l'extraction de motifs. 193
6.5 Valeurs par défaut et variation des paramètres de réglage pour l'observation de leur influence sur les performances du système. 196
6.6 Variation des paramètres de réglage pour l'étude de leur influence relative sur les performances du système. 203
6.7 Valeurs par défaut des paramètres de réglage du système. 205

6.8 Indices moyens de performance de l'approche proposée pour l'extraction de motifs dans la configuration par défaut du système et dans un contexte "raisonnablement" bruité. 207
6.9 Résultats de l'extraction de motifs à partir de séquences contenant des instances normalement modifiées puis dégradées d'un même motif. 209

Annexes **229**

E.1 Paramètres des variations sinusoïdales de repos des sujets de l'expérimentation. . 245

Introduction générale

La recherche dans le domaine de la télésurveillance médicale à domicile a pris une grande ampleur ces dernières années face au vieillissement de la population et au manque d'infrastructures d'accueil de personnes exposées à des risques d'accident dans leur vie quotidienne ou de dégradation de leur état de santé. L'objectif est de permettre une prise en charge médicale et sociale des personnes isolées ou en perte d'autonomie. Les enjeux de le mise en place de tels systèmes sont nombreux, tant pour les patients, le personnel médical et la société en général.

Plusieurs axes de recherche sont impliqués dans le développement des systèmes de télésurveillance médicale. Ils concernent notamment le développement d'architectures de communication entre les acteurs de ces systèmes, d'équipements appropriés à la surveillance et à l'amélioration de la qualité de vie des personnes, de bases de stockages des données collectées au domicile et d'outils d'analyse et de traitement de ces grandes quantités de données. Il s'agit alors de détecter et de prévenir l'occurrence de situations critiques d'une personne à domicile, impliquant la transmission de messages et d'alarmes aux acteurs concernés et prêts à intervenir en cas de nécessité.

La travaux de cette thèse de doctorat se sont déroulés dans le contexte du projet HIS (Habitat Intelligent pour la Santé) de l'équipe AFIRM du laboratoire TIMC–IMAG à Grenoble. Ils se situent dans le cadre de la conception d'outils de fusion de tous types de données collectées au domicile, pour la détection des situations inquiétantes, voire critiques, d'une personne. On se place dans le contexte plus particulier de l'étude à long terme de l'évolution de l'état de santé d'une personne pour identifier l'installation de certaines pathologies plus ou moins progressives telles qu'une infection urinaire, une dépression ou encore une insuffisance cardiaque. Ce problème de décision sur la situation d'une personne à domicile concerne ainsi l'*analyse multidimensionnelle* de *données temporelles* qui peuvent être *hétérogènes*.

Les contraintes de résolution concernent en particulier la nécessité d'une approche centrée sur le comportement spécifique de chaque personne. Les habitudes de vie quotidienne aussi bien que les caractéristiques physiologiques (fréquence cardiaque par exemple) varient en fonction des individus. La complexité du problème réside ainsi dans une grande variabilité inter-individuelle des données enregistrées, mais également dans de larges modifications intra-individuelles possibles étant donné l'aspect souvent peu prévisible des comportements humains. De nombreux facteurs d'influence agissent également sur les différents paramètres observés et leurs variations relatives sont complexes. La jeunesse des projets dans le domaine de la télésurveillance médicale à domicile a par ailleurs pour conséquence le manque de données expérimentales issus de systèmes réels, et de connaissances *a priori* sur les relations conjointes des paramètres étudiés.

Dans ce contexte, cette étude a été menée en trois étapes, définissant les trois parties de ce document. Elles concernent successivement (I) la présentation de la problématique, (II) un processus de simulation de données collectées à domicile et (III) un système de décision sur la situation d'une personne.

La première partie présente la **problématique de décision** sur la situation d'une personne à domicile en la situant en particulier par rapport au contexte général de la télésurveillance. Étant données les spécificités individuelles de comportement, sa résolution est présentée comme la *construction d'un profil comportemental* d'une personne dans ses activités de la vie quotidienne, tout écart par rapport à ce profil étant considéré inquiétant. On présente également la nécessité de mise en place d'un *processus de simulation* de données collectées à domicile d'une part pour faire face au manque de données expérimentales et d'autre part pour une expérimentation complète et fiable d'un *système de décision* sur la situation d'une personne.

La deuxième partie concerne ainsi la mise en place d'un **processus de simulation** de séquences enregistrées par des capteurs à domicile. Il s'agit de générer des données correspondant à un ensemble possiblement hétérogène de paramètres interdépendants, tous liés dans notre contexte à la situation de la personne dans son environnement de vie. On montre en particulier la difficulté de définition d'une méthodologie générale pour la simulation. Quelle que soit la méthode utilisée, une démarche systématique consiste cependant en la *validation* progressive des différentes étapes de construction du processus. Le manque de données expérimentales issues d'un système réel de télésurveillance rend cependant difficile cette démarche. Compte tenu également du manque de connaissances *a priori* sur les paramètres observés à domicile, la construction et la validation du processus de simulation s'appuient finalement sur un ensemble de *connaissances hétérogènes* issues de la diversité des sources d'information disponibles. Le modèle proposé s'appuie également sur *plusieurs techniques de simulation* pour s'adapter à l'hétérogénéité des paramètres observés. On montre alors en particulier la diversité des profils de personnes et des types de situations générées, incluant la simulation de modifications "normales" et inquiétantes de comportement.

La troisième partie utilise les résultats du processus de simulation pour l'expérimentation d'un **système de décision** sur la situation d'une personne à domicile. Il s'agit d'abord de construire un processus d'apprentissage des comportements récurrents d'une personne dans sa vie quotidienne, la modification ou l'absence de ces activités habituelles étant considérée inquiétante. Compte tenu du large ensemble de données analysées, on rapproche en particulier ce problème de la *fouille de données temporelles* pour l'extraction de motifs. Les séquences analysées correspondent aux données collectées à domicile et constituent ainsi un ensemble multidimensionnel de données hétérogènes. Pour faire face aux contraintes liées au contexte de résolution, on propose alors une *approche générique et complètement non supervisée* pour l'extraction de motifs temporels multidimensionnels et hétérogènes. On intègre en particulier une étape de *fouille de caractères* pour l'identification des sous-séquences qui correspondent le plus certainement aux instances de motifs. On expose les résultats de l'expérimentation de la méthode proposée sur les séquences de données générées dans le contexte de la télésurveillance médicale par le processus de simulation. Les performances du système sont en particulier évaluées par des indices de sensibilité (tolérance aux modifications normales de comportement) et de spécificité (rejet des modifications inquiétantes).

Enfin, une conclusion générale et les perspectives de ce travail sont présentées. Un ensemble d'annexes liées aux différentes parties du document est également proposé. On présente en particulier en annexe J la liste des publications relatives à ces travaux.

Première partie

Présentation de la problématique

1

Contexte de recherche

Les travaux effectués au cours de cette thèse de doctorat se situent dans le cadre de la télésurveillance médicale à domicile, qui est une dimension de la télémédecine. Ce chapitre a ainsi pour objectif principal de situer ce contexte de recherche et ses enjeux.

1.1 Télémédecine

La télémédecine représente l'utilisation des Nouvelles Technologies de l'Information et de la Communication (NTIC) dans le secteur médical [20]. Elle médiatise l'acte médical en interposant un outil de communication entre les médecins ou entre un médecin et son patient. La télémédecine ne remplacera jamais le contact immédiat médecin–malade mais vient s'ajouter aux outils du médecin au service du patient [39]. Elle remet ainsi en cause une partie de la pratique médicale, mais représente un enjeu considérable pour l'amélioration des conditions de soin et de vie de beaucoup de personnes.

À l'origine, dans les années soixante à soixante-dix, les premiers programmes de télémédecine ont été adoptés par les pays les plus vastes où la densité de population est faible, pour répondre au problème d'isolement géographique de certaines populations [39]. Ce type d'organisation propose en effet une solution à la difficulté d'accès aux centres de soins spécialisés. D'après [39], les premières expérimentations ont ainsi été développées par exemple en Australie (suivi psychothérapique à distance), en Écosse (dermatologie et médecine à distance pour les plates-formes pétrolières) et dans les zones rurales des États-Unis (télésoin).

La télémédecine a aujourd'hui trouvé de nombreux champs d'applications, et se décline en différents termes dont il est difficile de déterminer une typologie unanime [5, 103]. On présente finalement cinq catégories d'applications en télémédecine :

- **Télésurveillance** – Enregistrement télémétrique, généralement au domicile, de paramètres physiologiques ou ciblant l'environnement ou le comportement d'un patient, transmis ensuite aux praticiens concernés.
- **Téléconsultation** – Examen d'un patient ou analyse des données le concernant sans interaction physique directe. On distingue deux types de téléconsultations : (1) soit le patient consulte de sa propre initiative un médecin par un réseau de communication interposé ; (2) soit le médecin consulté sollicite un avis diagnostic (télédiagnostic) ou thérapeutique (télé-expertise) auprès d'un confrère situé à distance. On peut également citer dans ce cadre l'envoi et la consultation d'images médicales à distance (télé-imagerie, téléradiologie).
- **Télé-assistance** – Aide thérapeutique directe apportée à distance au patient, conséquence possible de la téléconsultation.

- **Téléchirurgie** – Manipulation de matériel médical (instruments chirurgicaux) contrôlée à distance par le praticien sur le patient (appelée aussi télémanipulation).
- **Téléformation** – Utilisation de l'outil informatique en particulier pour l'aide à la formation continue des médecins : contacts professionnels via le réseau, consultation des informations médicales (banque de données, imagerie, suivi d'études épidémiologiques et d'essais cliniques), consultation de cours de formation et visioconférences dans les universités (téléenseignement) et réunions.

En particulier, l'application dénommée couramment **télésurveillance médicale à domicile** est fondamentale pour l'amélioration de la qualité de soins et de vie des personnes nécessitant des soins ou une attention particulière. Elle vise à mettre en place dans l'habitat d'une personne un dispositif qui permet de capturer des informations sur son état de santé, afin de rendre possible pour le praticien un diagnostic, voire une aide au patient à distance. Par rapport aux catégories précédentes, cet outil de médiation entre un médecin et son patient doit alors prendre en compte plusieurs éléments :

- *Télésurveillance*, pour l'enregistrement d'un ensemble de paramètres liés au patient ;
- *Téléconsultation*, à l'initiative soit du patient, soit du personnel médical pour l'analyse des données télésurveillées ;
- *Télé-assistance*, pour fournir quand c'est possible à distance une aide directe au patient.

En plus de la souscription à certaines règles techniques, cliniques, juridiques, économiques ou éthiques liées à la télémédecine en général [39], la considération conjointe de ces trois éléments dans le contexte de la télésurveillance médicale à domicile impose en particulier de définir un compromis entre la nécessité de *minimiser les contraintes de la télésurveillance pour le patient* (capteurs peu nombreux, dispositifs non invasifs) et de *maximiser l'information disponible pour le praticien* dans sa démarche de diagnostic et d'assistance.

1.1.1 Les apports et enjeux de la télémédecine

La télémédecine s'avère être une réalité médicale : elle s'impose déjà à travers l'usage d'outils comme le téléphone et la télécopie par exemple. Les progrès actuels des NTIC appliquées au domaine médical (imagerie médicale, débits de transmission, convivialité des systèmes, etc.), la miniaturisation des dispositifs, ouvrent des perspectives pour le développement de la télémédecine en termes d'accroissement de l'efficacité et de la qualité des soins, de partage des connaissances, ou encore de réduction des coûts de santé publique. Pour chaque acteur de la télémédecine, les avantages de ce type d'organisation sont nombreux [78, 103].

Pour les praticiens, il s'agit de développer une plus grande coopération entre les différents réseaux du milieu médical : ville-hôpital, généraliste-spécialiste, public-privé. L'idée est de créer des passerelles de communication, d'information et de transmission de savoir. Un des enjeux du développement de la télémédecine concerne ainsi les aspects de *partage de données et de connaissances* : nécessité de l'interopérabilité des systèmes, définition de protocoles de communication, d'ontologies, création d'un dossier médical électronique partagé, etc.

Pour les patients, la télémédecine permet d'améliorer la qualité des soins grâce à l'expertise possible à distance et, par conséquent, à la réduction des délais de prise en charge diagnostique et thérapeutique. Elle permet également de répondre au problème d'isolement géographique en assurant l'égalité d'accès aux soins. Les petits centres hospitaliers souffrent en effet du manque d'équipements et d'une pénurie de médecins. Si on considère le cas particulier de la surveillance à distance, la télémédecine répond au besoin d'autonomie, de sécurité et d'intégration sociale de patients souhaitant rester à leur domicile, et s'inscrit alors dans la dynamique des alternatives à l'hospitalisation.

L'intérêt des pouvoirs publics pour la télémédecine est directement lié à sa contribution dans la maîtrise des dépenses de santé publique, tout en améliorant l'accès à des soins de meilleure qualité. Des économies sont réalisées dans les transports – limitation des déplacements des patients et du personnel médical – mais aussi dans la mise en œuvre des soins – par exemple, diminution de la redondance des soins grâce à l'accès distant au dossier médical, utilisation plus efficace des professionnels de santé, diminution des durées moyennes de séjour en centre hospitalier. Le marché mondial de la télémédecine est globalement un secteur économique à fort potentiel de développement. La santé devrait être amenée à représenter une bonne part du chiffre d'affaire mondial des télécommunications. Les prix des équipements en télémédecine ont déjà beaucoup baissé. La télémédecine impose cependant beaucoup d'investissements, la formation du personnel médical ainsi que celle des patients dans certains cas à l'usage des NTIC.

Pour les chercheurs, une conséquence du développement de la télémédecine, et plus particulièrement de celui de la télésurveillance, est la collecte de grandes masses de données liées à différentes applications et à différents patients. Un des enjeux est ainsi la *conception d'outils "intelligents"* facilitant l'exploitation personnalisée de grandes quantités de données disponibles, dans le contexte de chaque patient. Ces ensembles expérimentaux peuvent alors être à la base de nombreux projets de recherche.

À terme, la télémédecine pourrait également agir en faveur du transfert mondial de connaissances médicales, et améliorer par exemple l'aide aux pays en voie de développement ou émergents. Le développement de la télémédecine intéresse également beaucoup certains secteurs médicaux pour lesquels elle serait parfois l'unique solution d'intervention pour l'apport de soins. Il s'agit par exemple de la médecine maritime, de la médecine sportive, de l'armée, qui considèrent la télémédecine comme un moyen d'assister à distance les marins, sportifs en zone isolée, soldats, spationautes, etc.

Malgré des potentialités certaines et reconnues, les enjeux restant à relever sont nombreux. Les développements actuels de la télémédecine n'en sont encore souvent qu'au stade expérimental, à l'étude de faisabilité des NTIC appliquées à la santé, et relèvent principalement d'initiatives locales. D'après [103], on a assisté ces dernières années à l'explosion de projets isolés, concernant des spécialités médicales différentes, dans des pays au système de santé parfois spécifique. La France en particulier est confrontée à des difficultés pour la mise en place d'une politique de développement et de régularisation des projets de télémédecine. La conséquence est le développement d'une multitude d'applications, expérimentées notamment en secteur hospitalier. Face au problème de désordre des normes informatiques et télématiques, à l'émergence de projets disparates, parfois concurrentiels, la nécessité d'un projet global pour le système de santé semblerait pourtant devoir s'imposer.

D'après [103], le bénéfice économique de la télémédecine reste ainsi encore incertain. L'analyse des coûts par rapport à l'efficacité des applications est complexe et nécessite de nouveaux outils d'évaluation. Le problème d'évaluation économique provient également du caractère encore expérimental des applications en télémédecine, qui rend difficile la mise en œuvre d'analyses à grande échelle.

1.1.2 Les freins au développement

Le développement de la télémédecine est confronté à des problèmes d'ordre culturel, juridique ou éthique, et à des réticences de la part des différents acteurs. Pelletier-Fleury a par exemple mis en évidence dans [86] plusieurs facteurs de frein de la diffusion de la télémédecine. Par ailleurs, le développement et l'efficacité des applications de télémédecine fait face à plusieurs contraintes méthodologiques importantes.

De manière générale, le scepticisme est encore présent quant à l'intérêt de la télémédecine. Médecins et patients craignent notamment qu'elle porte atteinte à la liberté d'exercice, au secret médical, et conduise finalement à une déshumanisation de la relation entre le médecin et son patient. De nombreuses réticences sont dues à la nécessité de changement de la structure organisationnelle du monde hospitalier et médical (modification des habitudes de travail, intégration de l'outil informatique, manque de temps, etc.), souvent considéré comme une charge de travail supplémentaire par les personnels soignants. En France, la complexité de la structuration et de l'organisation du système sanitaire fait de la télémédecine un domaine qui ne peut évoluer que lentement.

Ces nouvelles pratiques médicales soulèvent également de nombreux problèmes éthiques et juridiques [13, 95]. L'utilisation de l'outil informatique pour la consultation, le transfert et la sauvegarde des informations concernant les patients ne doit pas nuire à leur confidentialité et à leur fiabilité. D'autres questions concernent la responsabilité et la rémunération des praticiens. Les télépratiques médicales ne sont en effet pas encore reconnues comme des actes à part entière. Dans le cadre de la télé-expertise par exemple, l'avis consultatif d'un confrère, même s'il s'agit d'un acte de prudence, ne dégage cependant en rien le médecin traitant de sa responsabilité vis-à-vis de son patient, ce qui le rend plus vulnérable en terme de responsabilité civile. Le choix de la politique tarifaire de la télémédecine est également un problème important à résoudre.

Une autre crainte est celle de la fuite des compétences médicales des centres de soins les plus isolés. La délocalisation d'opérations médicales est en effet accompagnée du risque de regroupement des meilleurs spécialistes dans quelques grandes unités.

Au niveau méthodologique, l'hétérogénéité des besoins de chaque praticien et des normes informatiques impose de développer des applications à un degré de compatibilité et d'interopérabilité important. Leur efficacité dépend également d'une bonne gestion de la grande quantité d'informations générées par ces applications. Elles permettent en effet d'accéder à un large univers de données, de connaissances et d'informations, qui ne cesse de s'enrichir. Il devient alors difficile de constituer des corpus de connaissances et de rendre efficace leurs consultation et utilisation. La télémédecine nécessite en particulier un traitement personnalisé des informations, dans le contexte d'un patient, et prend ainsi en compte bien peu de règles d'interprétation générales issues de connaissances médicales.

1.2 Télésurveillance médicale à domicile

Les travaux de recherche effectués au cours de cette thèse de doctorat se situent dans le cadre de la télésurveillance médicale des personnes à domicile, qui représente une des dimensions de la télémédecine. Cette application prend en particulier en compte des éléments de *télésurveillance*, de *téléconsultation* et de *télé-assistance*.

1.2.1 Objectifs

L'objectif de tels systèmes est de permettre aux personnes de vivre chez elles le plus longtemps et le plus indépendamment possible, dans un environnement de confort et de sécurité. Il s'agit de détecter et de prévenir l'occurrence de situations critiques à domicile ou une dégradation de l'état de santé d'une personne. Ces systèmes représentent ainsi une alternative momentanée ou durable à l'hospitalisation ou au recours aux établissements d'hébergement de longue durée – maisons de retraite ou centres spécialisés. Le patient n'est alors plus contraint de renoncer à son domicile et à la vie en société. Il conserve une large autonomie dans son environnement social et privatif, tout en bénéficiant de services préventifs de santé. Ces systèmes concernent particulièrement les

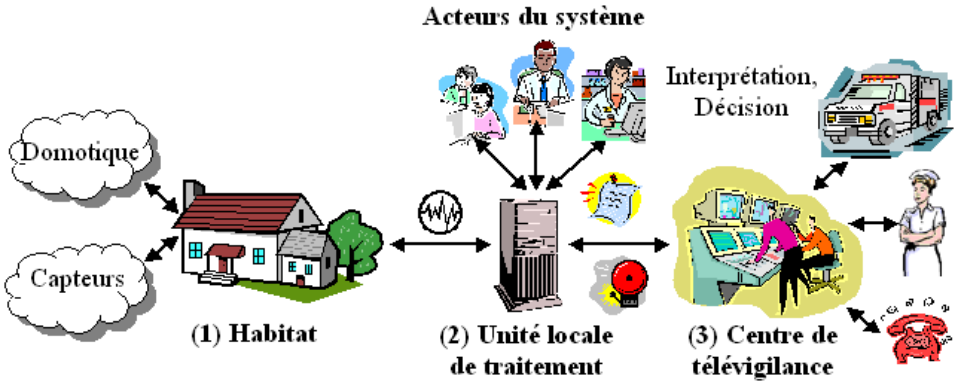


FIG. 1.1 – Système d'information de la télésurveillance médicale à domicile.

personnes âgées, mais plus généralement les personnes présentant des risques d'affection motrice (chute par exemple) ou cognitive (dépression, démence sénile, etc.), ou nécessitant des soins ou une attention particulière (diabétiques, asthmatiques, etc.).

1.2.2 Principe

La télésurveillance médicale d'une personne à domicile s'appuie sur un système d'information global comprenant les éléments suivants (voir Fig. 1.1) :

(1) **Un ensemble de capteurs** de différents types (physiologie, environnement, activité) installés dans l'habitat ou portés par la personne, reliés en réseaux pour la collecte en temps réel de données, **et d'appareillages automatiques** (domotique) pour adapter l'environnement de vie de la personne à ses capacités personnelles, motrices et cognitives ;

(2) **Une unité locale de traitement**, au niveau de chaque habitat, responsable du stockage et du traitement des signaux reçus des capteurs, de la gestion d'une base de connaissances relative à la personne télésurveillée, et de l'émission de messages et d'alarmes ;

(3) **Un centre de télévigilance** pour le traitement des messages et alarmes reçus des habitats.

Un ensemble d'acteurs (personnel médical, personne télésurveillée et membres de sa famille) peuvent accéder à tout moment, après authentification et selon leurs privilèges, aux données du système, au niveau de l'unité locale de traitement.

1.2.3 Enjeux

Les principales fonctionnalités nécessaires à la mise en place de systèmes de télésurveillance médicale à domicile sont la perception, l'analyse, le stockage et la transmission de données et d'informations relatives à la personne télésurveillée.

On identifie alors d'après [106] cinq sous-systèmes clés du développement des systèmes d'information pour les services de soin à domicile :

1. **Système de surveillance local** – Il s'agit d'un réseau local au domicile pour l'enregistrement télémétrique de données relatives à une personne par l'intermédiaire de capteurs physiologiques, d'environnement et d'activité.
2. **Système d'analyse de données** – La grande quantité de données collectées nécessite la conception d'assistants intelligents pour l'extraction d'informations pertinentes permettant la génération de messages et d'alarmes, l'aide au diagnostic et à la décision.

3. **Système de base de données** – Les données collectées ou les informations extraites doivent être stockées et accessibles pour leur consultation ou leur mise à jour.
4. **Système d'interfaces** – Les données et informations issues de la télésurveillance et de l'analyse des données collectées doivent être facilement accessibles aux différents acteurs du système.
5. **Système de communication** – Il s'agit de permettre l'interopérabilité des quatre sous-systèmes précédents à travers un réseau médical qui relie les habitats de patients, les centres hospitaliers, les centres de télévigilance et plus généralement les différents acteurs du système.

La complexité de ces systèmes réside dans le nombre d'acteurs impliqués, la diversité des techniques informatiques utilisées aux différents niveaux d'enregistrement, de stockage, d'analyse et de transmission des données, la quantité croissante des données collectées, la nécessaire personnalisation de leur traitement dans le contexte de chaque patient, la difficulté de modélisation de l'état de santé d'une personne. Une des spécificités de la télésurveillance médicale est la contrainte de traitement rapide de larges ensemble de données évoluant au cours du temps, afin de répondre à l'objectif de détection "au plus vite" des situations inquiétantes à domicile. Les difficultés de ces analyses sont en particulier liées à l'hétérogénéité des données collectées, aux facteurs d'influence agissant parfois fortement sur les paramètres observés, ainsi qu'aux dépendances mutuelles de ces paramètres.

2

État de l'art de la télésurveillance médicale à domicile

De nombreux projets isolés sont menés dans le monde sur le thème de la télésurveillance médicale des personnes à domicile. L'objectif de ce chapitre est ainsi de mettre en évidence la diversité des concepts et objectifs concernés par les différents projets sur la base d'exemples de travaux de recherche engagés. On situe alors notre problématique de détection des situations critiques d'une personne à partir des données collectées à domicile, et on présente enfin le projet HIS (Habitat Intelligent pour la Santé) du laboratoire TIMC (Grenoble) dans le cadre duquel ont été réalisées ces recherches.

2.1 Des projets variés dans le monde

Des projets pilotes variés dans les concepts et objectifs sont menés à travers le monde. Ils visent par exemple à définir une architecture générique pour de tels systèmes de surveillance, à expérimenter un système de télésurveillance sur une catégorie spécifique de patients (insuffisants cardiaques et pulmonaires, asthmatiques, diabétiques, patients souffrant de la maladie d'Alzheimer, etc.), ou encore à concevoir des appartements, des capteurs, des systèmes d'alarmes adaptés aux exigences de la télésurveillance médicale. Un ensemble de projets et de concepts relatifs au domaine de la télésurveillance médicale à domicile sont présentés par exemple dans [45, 94].

2.1.1 Architecture et expérimentation globale de systèmes d'information

De nombreux projets visent à concevoir une architecture appropriée aux objectifs des systèmes de télésurveillance médicale à domicile. Il s'agit particulièrement de répondre aux exigences d'interopérabilité et de fiabilité des sous-systèmes impliqués, dans le respect de la vie privée des patients, et pour assurer leur sécurité et le suivi à distance de leur état de santé.
Au Royaume-Uni, Williams *et al.* [113, 114] ont par exemple conçu une architecture générique d'un système de télésurveillance médicale (CarerNet) mise en œuvre sous la forme du prototype MIDAS [36].
Rodriguez *et al.* [98], en Espagne, ont développé une architecture du même type dans le cadre du projet EPIC (*European Prototype for Integrated Care*) de l'Union Européenne.
En France, Thomesse *et al.* [106] ont développé le projet TISSAD ayant pour objectif principal la définition d'une architecture générique, modulaire et ouverte pour les systèmes de télésur-

veillance, adaptable à diverses pathologies traitées à domicile (suivi de personnes âgées, d'insuffisants cardiaques et rénaux).

En terme du développement et de l'expérimentation de ces systèmes, le projet Shahal [97], en Israël, est probablement le plus abouti puisque plus de 40 000 patients y ont déjà souscrit. Il fournit un service d'urgence et de prévention de risques cardiaques et pulmonaires.
D'autres prototypes sont expérimentés à moindre échelle sur une catégorie spécifique de personnes, tels HAT [38] développé aux Etats-Unis pour des asthmatiques, le projet suédois de Lind *et al.* [71] pour des diabétiques, ou encore le projet PROSAFE [24] en France expérimenté sur des patients souffrant de la maladie d'Alzheimer.

Les projets liés à la conception globale de systèmes d'information pour la télésurveillance médicale à domicile s'intéressent ainsi soit à la mise en place des architectures de tels systèmes, soit à leur expérimentation sur une population bien ciblée, présentant une pathologie ou un risque particulier.

2.1.2 Systèmes de surveillance au domicile

D'autres expérimentations sont menées dans l'objectif de développer plus précisément les infrastructures de surveillance situées au domicile des personnes. Il s'agit en particulier de concevoir des habitats et des capteurs adaptés aux objectifs et contraintes de la télésurveillance médicale en termes à la fois technique et éthique, pour répondre à la finalité de la surveillance dans le respect de la vie privée des personnes.

Au Japon par exemple, Ogawa *et al.* [81] ont évalué la faisabilité de l'enregistrement de mesures physiologiques à distance à long terme. Celler *et al.* [21, 22, 23] ont étudié la possibilité d'évaluation de l'état de santé d'une personne à partir d'un ensemble de capteurs. Najafi *et al.* [76] ont proposé un capteur cinématique porté sur la poitrine pour la surveillance de l'activité physique d'une personne par la détection des postures du corps (allongée, assise et debout) et des périodes de marche.

Quelques projets se concentrent plutôt sur la conception d'appartements adaptés aux exigences de la télésurveillance médicale : SmartBo en Suède [37], AID HOUSE au Royaume-Uni [12], "Smart House in Tokushima" au Japon [104], ou le projet de Van Berlo *et al.* aux Pays-Bas [108].

Noury, Rialle *et al.* [79, 92, 93] ont également conçu et mis en œuvre à des fins expérimentales un habitat intelligent pour la santé (HIS) connecté à un réseau pour permettre la gestion d'utilisateurs, d'informations sur les personnes télésurveillées et d'alarmes.

2.1.3 Système de gestion et de stockage des données

Une des difficultés liées à la gestion et au stockage des données collectées réside dans la définition de modèles qui permettent facilement le partage et l'échange. Certaines projets se concentrent ainsi sur la définition de protocoles de gestion de données et de connaissances [1], d'ontologies ou encore d'un dossier médical électronique partagé pour les patients [29].

2.1.4 Assistants intelligents pour l'analyse de données

Vers la conception de systèmes intelligents, autonomes et non intrusifs
En terme d'exploitation des données enregistrées, la première génération de télé-alarmes consistait en de simples "boutons portables" liés par une communication sans fil à un système central de surveillance. L'inconvénient majeur est la nécessité de participation du patient à la génération

manuelle de ces alarmes. Il doit par conséquent d'une part réaliser que quelque chose ne va pas, et d'autre part être encore éveillé et conscient.

Les générations suivantes de systèmes visent ainsi un niveau supérieur d'alarmes avec la conception de **systèmes autonomes**, qui ne nécessitent pas l'intervention de la personne télésurveillée. Noury *et al.* [80] ont par exemple étudié la mise au point d'un capteur de chute qui intègre sur un même support trois accéléromètres disposés orthogonalement et un microcontrôleur qui détermine l'inclinaison du corps puis élabore de manière autonome l'information de chute.

D'autres systèmes se concentrent par ailleurs sur des **approches non intrusives**, pour lesquelles le patient n'est équipé d'aucun instrument. Chan *et al.* [24] se sont par exemple intéressés à l'activité d'une personne dans une pièce, pour la détection de situations critiques, à partir de l'étude des données enregistrées par un ensemble de capteurs infrarouges permettant de détecter les mouvements dans différentes zones de la pièce.

Les données collectées à domicile sont également exploitées pour la **télé-assistance à domicile**. Au Canada, Pigot *et al.* [88] développent par exemple un projet d'assistance cognitive aux personnes dans la réalisation des activités de la vie quotidienne. Les interventions du système sont personnalisées en fonction des capacités de la personne, de son état de santé, de ses habitudes de vies et préférences.

Problèmes spécifiques à la conception d'assistants intelligents

Cette génération d'*assistants intelligents* pour l'analyse de données collectées à domicile est ainsi caractérisée par sa capacité d'autonomie et ses facultés de perception, de raisonnement et de prise de décision [106].

La complexité de leur conception réside d'abord dans les contraintes de robustesse, de fiabilité et de pertinence des informations disponibles. Cette question soulève en particulier le problème du choix d'un ensemble cohérent de capteurs à considérer pour la télésurveillance. L'objectif est de trouver un compromis entre la nécessité de disposer des informations optimales en terme de surveillance et la contrainte du respect de la vie privée des patients.

Une difficulté supplémentaire vient de la nécessité de conception de systèmes qui permettent le traitement personnalisé des données de différents patients. Une solution consiste alors à exploiter des algorithmes issus de l'intelligence artificielle. La grande quantité de données à analyser peut également justifier l'utilisation de techniques de fouille de données.

Un autre problème concerne la manière de combiner des données de différents types pour fournir des informations pertinentes aux praticiens [106]. Les systèmes proposés initialement concernent en effet principalement la génération d'alarmes ciblées, à "bas niveau", ou déclenchées à partir de l'observation d'un seul type de paramètre. Cette caractéristique d'analyse multidimensionnelle et hétérogène nécessite de s'intéresser particulièrement aux méthodes de fusion de données.

2.1.5 Détection des situations critiques d'une personne à domicile

Le problème de détection des situations critiques d'une personne à partir des données collectées à domicile concerne en particulier la conception d'**assistants intelligents**. De grandes quantités de données temporelles, hétérogènes, sont analysées en temps réel pour l'identification des situations inquiétantes ou critiques. Les projets développés et les plus avancés jusqu'à présent dans ce contexte s'intéressent souvent à une pathologie particulière, ou bien à un ensemble restreint ou spécifique de paramètres.

Ainsi, on n'a pas identifié de recherches avancées vers la conception d'un assistant intelligent

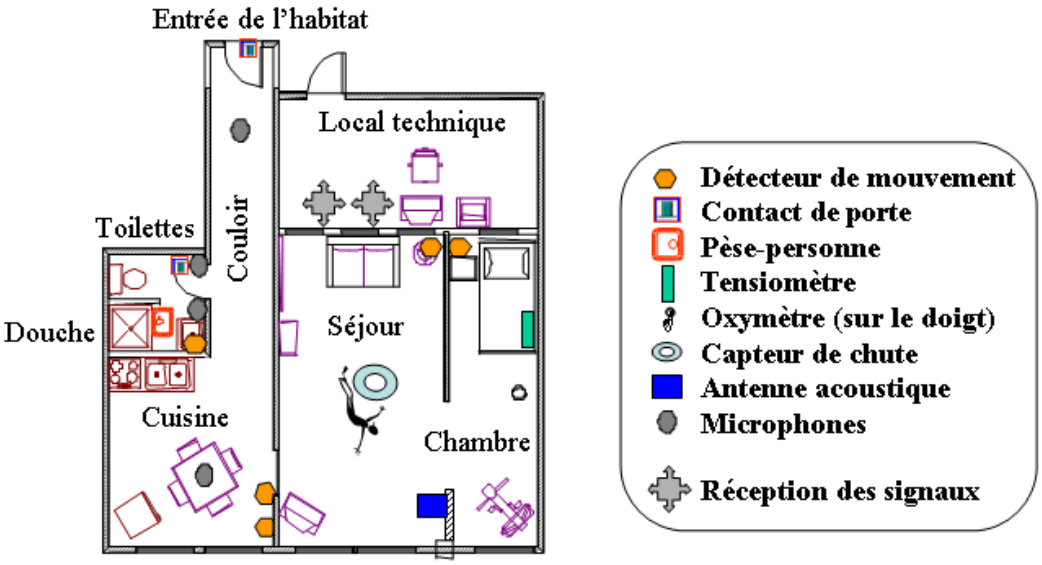


FIG. 2.1 – L'HIS : appartement prototype mis en place à la faculté de médecine de Grenoble.

"générique", qui permette d'analyser les données relatives à n'importe quels paramètres, quels que soient le patient et les pathologies particulières qui le concernent, et dans un objectif de détection de situations inhabituelles rencontrées. On décide alors de mener des travaux dans ce sens étant données la quantité croissante de capteurs développés qui peuvent être installés au domicile, la diversité des pathologies qui peuvent nécessiter ce type de surveillance, et la spécificité du comportement de chaque personne. L'état de l'art des différents thèmes concernés pour la résolution de ce problème est détaillé dans les parties suivantes. Il est en effet nécessaire de formuler d'abord précisément le problème pour identifier les axes de recherche qui nécessitent effectivement d'être abordés.

2.2 Description du projet HIS

Les travaux de recherche de cette thèse de doctorat se situent dans le cadre du projet HIS de mise en place d'un *Habitat Intelligent pour la Santé*, développé au sein de l'équipe AFIRM (Acquisition, Fusion d'Informations et Réseaux pour la Médecine) du laboratoire TIMC–IMAG à Grenoble. Ce projet concerne la problématique de surveillance de personnes à domicile et investigue en particulier le concept d'*habitat intelligent* [94]. Il s'inscrit cependant dans un projet plus vaste définissant le système d'information complet associé à un habitat télésurveillé dans son environnement médico-social : SIC–HIS (Système d'Information et de Communication de l'Habitat Intelligent pour la Santé). L'objectif est d'intégrer sur un même réseau médical plusieurs habitats de patients, des centres de télévigilance, des postes de médecins et d'autres membres du corps médical et des postes de consultants occasionnels. Dans un premier temps, l'étude est cependant limitée à la surveillance d'un seul habitat.

2.2.1 Construction d'un appartement prototype

Dans le cadre de ce projet, un appartement prototype (HIS) a été mis en place dans les bâtiments de la Faculté de médecine de Grenoble [92] (voir Fig. 2.1). C'est un appartement complet de type

T1 (40m^2 environ) qui comprend les pièces principales d'un habitat à savoir la chambre, le séjour, la cuisine, les toilettes, la douche et le couloir. Un local technique attenant à l'appartement a été ajouté afin de recevoir le système informatique d'expérimentation, c'est-à-dire en particulier l'unité locale de traitement. L'HIS met en oeuvre la chaîne complète de traitement d'informations, allant de leur perception à leur analyse :

- **Perception de la personne et de son environnement** par différents types de capteurs installés dans l'habitat ou portés par la personne à domicile. Les capteurs sont des capteurs de présence infrarouges, des capteurs magnétiques de contact de porte, des microphones, un tensiomètre, un pèse-personne, un oxymètre. Des travaux sont par ailleurs menés au sein de l'équipe sur la mise en place de capteurs dits "intelligents" qui possèdent une intelligence locale pour délivrer une information pertinente à partir des données brutes enregistrées et une évaluation de sa qualité, tel le capteur de chute [80].
- **Transmission des informations** par un bus domotique vers l'unité locale de traitement située dans l'habitat. On peut ainsi visualiser l'état des capteurs à tout instant.
- **Gestion du dossier médical** afin d'avoir connaissance des données cliniques nécessaires au suivi médical de la personne. On dispose notamment de fonctionnalités de rédaction et de gestion d'ordonnances et de comptes-rendus de visite, de droits d'accès et de niveau de confidentialité, de gestion de scénarios d'événements à risques.
- **Analyse et interprétation des données du système** afin de détecter en temps réel toute situation critique pour la personne et de prévenir l'occurrence de ces situations en identifiant les détériorations de son état de santé.
- **Présentation de courbes de tendances, de données et d'informations relatives à la personne**, afin que le personnel médical dispose en local ou à distance d'un ensemble d'informations pertinentes pour l'évaluation de la situation de la personne à domicile et de l'évolution de son état de santé.
- **Déclenchement de différents types d'alarmes** suivant les situations détectées vers un poste de télévigilance qui en assure le traitement adapté.

2.2.2 Collaborations nationales et internationales

L'équipe AFIRM développe ces différents aspects de la télésurveillance médicale à domicile en collaboration avec des laboratoires universitaires, des équipes hospitalo-universitaires et des partenaires industriels. La plateforme HIS a en particulier permis une collaboration avec l'équipe GEOD du CLIPS pour la reconnaissance de signaux sonores de détresse dans un Habitat Intelligent pour la Santé. D'autres projets visent à modéliser le système d'information pour l'organisation des soins à domicile dans le cadre des personnes âgées, des insuffisants cardiaques et rénaux (projet TIISSAD [106]), avec la collaboration de plusieurs autres centres de recherche – LORIA (Nancy), LAAS (Toulouse), INSERM U558 (Toulouse), INSERM ERM107 (Lyon) ; ou encore à la conception et au développement d'équipements (biocapteurs) de télé-assistance médicale intégrés dans un vêtement (projet VTAMN), en collaboration avec le MEDES à Toulouse, l'ITECH et l'INSA à Lyon, et plusieurs sociétés.

L'HIS a également été mis à disposition du projet EPICT pour l'évaluation d'une nouvelle approche de la prise en charge d'une pathologie chronique, l'insuffisance cardiaque, en privilégiant la participation active du patient grâce à la télémédecine, en collaboration avec les sociétés ADDS à Paris, TAM Télé-Santé à Aix en Provence, CORONIS à Montpellier, et les services hospitaliers du CHU de Grenoble (AGL-HTA) et de l'Hôpital Européen Georges Pompidou. Une première expérimentation en vraie grandeur du dispositif HIS est mise en place dans le

service d'Oncologie Médicale du CHU de Grenoble (projet OncologHIS) pour évaluer l'activité pré et post-thérapeutique de patients subissant une chimiothérapie. Une chambre hospitalière du service d'Oncologie a été équipée d'un réseau de capteurs de détection de présence et d'un dispositif d'analyse de ces données afin d'évaluer la mobilité et la fréquence des déplacements du patient tout au long de son séjour. Un projet de plus grande envergure (projet AILISA), propose de mettre en place des plateformes de validation médico-technologiques dans des environnements contrôlés situés dans un service de gériatrie de l'Hôpital Charles Foix à Paris, dans un service de gériatrie du CHU de Toulouse, puis dans un domicile de centre d'accueil dépendant du CCAS de Grenoble. Ces plateformes sont destinées à servir ensuite de lieux d'expérimentation et de validation d'usages pour d'autres équipes et projets.

Au niveau international, l'équipe AFIRM collabore depuis plusieurs années avec des équipes canadiennes, notamment avec le département d'informatique de l'Université du Québec à Montréal (UQAM), et plus récemment avec la laboratoire DOMUS de l'Université de Sherbrooke, en particulier sur le sujet de l'informatique diffuse et distribuée, et le Centre de Recherche sur le Vieillissement de l'Université de Sherbrooke sur les aspects de la physiologie de la chute. Dans le cadre de ces collaborations, j'ai eu l'occasion de plusieurs séjours à l'UQAM pendant ma thèse, d'une durée totale de près d'un an.

3

La problématique de décision sur la situation d'une personne à domicile

Dans le cadre de ces travaux de thèse, on s'intéresse particulièrement aux opérations effectuées au niveau de l'unité locale de traitement d'un système de télésurveillance médicale de personnes à domicile, c'est-à-dire à l'analyse de l'ensemble des signaux reçus des capteurs installés dans l'habitat pour la génération de messages et d'alarmes. Cette étape est fondamentale pour une exploitation efficace des potentialités de collecte de grandes masses de données dans l'objectif d'améliorer le suivi et la sécurité des patients à domicile et de prévenir une dégradation de leur état de santé. Les informations extraites sur la situation du patient doivent être pertinentes pour l'aide au diagnostic et à la décision des praticiens submergés par la masse de données disponibles. La complexité de cette démarche réside en particulier dans la quantité et la diversité des données, ainsi que dans la nécessité d'un traitement personnalisé pour chaque patient.

Ce paragraphe présente une proposition d'architecture de décision pour un tel système de surveillance (3.1), et détaille la formulation (3.2), le principe (3.3) et les contraintes (3.4) de résolution du problème qui nous intéresse particulièrement dans ces travaux : l'étude des habitudes de vie d'une personne à domicile dans l'objectif d'évaluer l'évolution de son état de santé à long terme.

3.1 Proposition d'une architecture de décision

3.1.1 Objectifs de décision

La problématique de décision sur la situation d'une personne dans le cadre de la télésurveillance médicale à domicile se situe au niveau de l'unité locale de traitement d'un habitat (voir Fig. 1.1). Les objectifs de décision se situent principalement à deux niveaux :

1. **Détecter l'occurrence d'une situation critique,** à plus ou moins long terme : de la détection d'une chute, d'une infection, d'une crise cardiaque, à l'apparition des premiers symptômes de dépression ou de démence sénile par exemple. Une situation critique se définit ainsi par un écart à une certaine "normalité" de comportement.
2. **Prévenir l'occurrence de ces situations,** en fournissant au personnel soignant des informations parfois difficilement observables même par une visite quotidienne de leur part – sur les habitudes de la personne, son comportement à domicile, etc. – et importantes pour le suivi médical : de l'évolution des valeurs de paramètres directement enregistrés par

les capteurs installés dans l'habitat à l'abstraction d'informations plus complexes de ces séquences de valeurs.

Le système de détection et de prévention des situations critiques à domicile est ainsi un système de décision complexe qui prend en compte **plusieurs niveaux d'appréhension de la situation de la personne** et de gravité des situations observées. Ce système est construit sur la base de l'observation de **plusieurs types de données relatives au patient**, issues des capteurs de l'habitat ou d'indications subjectives du patient lui-même ou du personnel médical. Aux différents niveaux d'analyse, il repose également sur l'**apprentissage** du comportement habituel d'une personne télésurveillée à partir des paramètres observés à domicile. Cet apprentissage permet d'extraire de nouvelles informations et d'enrichir la base des connaissances utiles à la décision.

Les paragraphes suivants décrivent ainsi les principes de décision selon ces trois caractéristiques : construction d'un système d'apprentissage et en décision (paragraphe 3.1.2), impliquant plusieurs niveaux de décision (paragraphe 3.1.3), à partir de plusieurs types de données (paragraphe 3.1.4). On aboutit finalement à la proposition d'une approche granulaire pour la construction de tels systèmes (paragraphe 3.1.5).

3.1.2 Système d'apprentissage et de décision

Le développement des projets en télésurveillance médicale et la collecte de données en environnement réaliste n'en sont encore qu'au stade expérimental. Ainsi, on n'a pas particulièrement acquis de connaissances sur les évolutions conjointes attendues des paramètres contrôlés par les capteurs (physiologie, environnement, activité) en fonction des caractéristiques d'une personne (données cliniques par exemple). On ne peut pas non plus envisager la description exhaustive de l'ensemble des situations critiques possibles d'une personne, et on ne dispose pas de moyens d'apprentissage de telles situations (observation de sujets évoluant vers ces situations critiques). Cette faible connaissance du domaine oblige à se baser sur peu ou pas de connaissances *a priori* (connaissances déclaratives, seuils de décision, etc.) et à considérer spécifiquement chaque personne télésurveillée.

Le système de détection de situations critiques proposé fonctionne ainsi à la fois en **apprentissage** et en **décision**. Il repose sur les informations d'une base de connaissances qui s'enrichit progressivement au cours du temps. On espère ainsi apprendre les caractéristiques du comportement d'une personne pour détecter ensuite les situations critiques comme des écarts à ce profil comportemental. Ce problème de décision peut en cela se rapprocher de la détection d'anomalies dans les systèmes informatiques [19, 44, 63, 66] – on construit alors le profil d'un utilisateur ou d'un programme. Un exemple de résolution de ce problème d'apprentissage et de décision concerne l'étude de l'activité de patients souffrant de la maladie d'Alzheimer. En se fondant uniquement sur des capteurs de mouvement infrarouges répartis dans une chambre, Chan *et al.* [24] ont utilisé un automate à états finis modélisant les déplacements du patient et du personnel médical dans la chambre pour caractériser les activités d'un patient et détecter tout comportement suspect.

3.1.3 Définition de plusieurs niveaux de décision

Pour réaliser des systèmes de décision complexes, il est intéressant de décomposer la problématique en plusieurs niveaux de compréhension. Ce type d'approche granulaire est couramment utilisé par exemple pour la reconnaissance, à partir de signaux vidéos, de mouvements [11] ou de

gestes complexes [41], ou encore pour la détection d'intrusions dans les systèmes informatiques [6].

Dans le domaine de la télésurveillance médicale, le comportement d'une personne à domicile – lié directement à son état de santé – est déjà intuitivement décrit à plusieurs niveaux – en terme d'un ensemble d'actions élémentaires effectuées (se lever, marcher, etc.) puis d'activités (dormir, prendre un repas faire sa toilette, etc.) [24, 85, 112] – ce qui rend facilement compréhensible ce type d'approche. Une dégradation de l'état de santé d'un patient entraîne généralement des troubles du comportement visibles à plusieurs niveaux : augmentation des risques de chute, lenteur à effectuer des actions élémentaires, ou diminution globale de l'activité par exemple.

Ce lien entre comportement et autonomie d'un patient d'une part et état de santé d'autre part est déjà largement utilisé dans la pratique médicale. L'évolution de l'état de santé d'un patient, liée à son autonomie, est fréquemment évaluée en terme de capacité à effectuer les activités de la vie quotidienne (AVQ ou ADL – *Activity of Daily Living*) telles que faire sa toilette ou se nourrir par exemple. Dans les AVQ on distingue parfois les activités instrumentales de la vie quotidienne (IAVQ ou IADL – *Instrumental Activity of Daily Living*) regroupant des activités plus complexes (faire le ménage ou faire ses courses par exemple). Plusieurs projets [85, 112, 115] ont intégré l'évaluation de l'ADL dans des systèmes de surveillance de l'état de santé d'une personne. D'après Ramos *et al.* [91], une estimation de la dépendance d'une personne dans la réalisation des activités de la vie quotidienne devrait constituer une part essentielle de n'importe quelle évaluation de l'état de santé d'une personne âgée. Dans le contexte spécifique de réduction de l'incidence des chutes, Cameron *et al.* [18] soulignent aussi l'utilité de la surveillance en temps réel de paramètres liés à l'activité de la personne, tels que les données enregistrés par des capteurs de mouvement, de vitesse de déplacement, ou encore de surveillance de l'occupation des pièces. Le critère d'évaluation de l'AVQ correspond cependant à une notion globale et les méthodes d'évaluation et les échelles d'appréciation sont variées [9, 35, 53, 64, 87, 105].

Afin de réduire la complexité du système et d'en optimiser les performances, il est donc intéressant de considérer plusieurs **niveaux de décision**, correspondant à des niveaux croissants de complexité et de compréhension de la situation du patient. Deux échelles de décision sont ainsi mises en évidence :

(1) Échelle de temps correspondant à des décisions à plus ou moins long terme (augmentation de la largeur de la fenêtre d'observation) ;

(2) Échelle de gravité des situations à chaque niveau.

On définit ainsi plusieurs types de messages et d'alarmes, chaque type présentant plusieurs niveaux de gravité : de situations critiques ponctuelles (trébucher, chuter, etc.) à la détection de symptômes sur une plus longue durée (oublis, diminution globale d'activité, etc.).
Les messages ou alarmes générés par le système sont le résultat de plusieurs types d'analyse :

(1) Détection d'incohérences dans les données reçues ou les paramètres évalués, pouvant signifier par exemple le mauvais fonctionnement d'un capteur ou la présence de plusieurs personnes dans l'appartement.

(2) Détection du dépassement d'une valeur critique d'un paramètre ou d'une combinaison de paramètres – un paramètre pouvant être défini à un haut niveau de complexité (paramètre d'activité, d'adéquation au profil, etc.).

(3) Détection d'un scénario d'événements, connu a priori comme étant critique pour le patient.

La proximité d'un seuil ou la vraisemblance de l'occurrence d'un scénario sont autant de paramètres qui définissent le **niveau de gravité des messages** ou alarmes. On prend ensuite à tout

niveau une décision sur la génération éventuelle de messages et d'alarmes liés à la situation de la personne.

3.1.4 Fusion de plusieurs types de données

Le système de décision repose sur les données et connaissances disponibles dans le cadre de la télésurveillance médicale de personnes à domicile sur la situation de la personne, dans l'objectif d'en extraire des informations pertinentes pour la détection de situations critiques et l'aide au diagnostic et à la décision pour les praticiens. D'après [102], "les données sont le résultat d'observations d'une expérience ; les informations sont le résultat de l'interprétation de ces données ; les connaissances définissent la manière dont les données et les informations vont être manipulées."

Une grande variété de capteurs permettent de collecter des données liées à l'état de santé et peuvent être installés dans un habitat. Ils concernent l'activité, l'environnement et la physiologie de la personne télésurveillée. Ces capteurs fournissent des données **différentes**, voire **complémentaires**, ou même **redondantes**, et permettent d'inférer des informations de plus haut niveau sur la situation de la personne à domicile. On dispose également d'un ensemble de connaissances *a priori* relatives à la personne télésurveillée et à son comportement à domicile. Il s'agit de données cliniques et de connaissances extraites de l'apprentissage au cours du fonctionnement du système.

Une appréhension complète et fiable de la situation d'une personne à domicile et de son évolution est ainsi obtenue à tout niveau de décision par l'analyse d'un ensemble de données, connaissances et informations relatives au patient impliquant :

(1) les données issues des capteurs,

(2) un ensemble de données cliniques (connaissances *a priori*),

(3) une base de connaissances constituée d'informations *a priori* – selon le contexte, il peut s'agir de seuils de normalité sur certains paramètres par exemple – puis enrichie au cours du fonctionnement du système par les résultats d'apprentissages réalisés à partir des données disponibles.

La question fondamentale d'analyse de ces grands ensembles de données hétérogènes pour prendre une décision à tout moment sur la situation d'un patient peut se définir comme un problème de **fusion de données**. D'après [42], "un processus de fusion de données permet, grâce à la combinaison d'informations hétérogènes provenant de différents capteurs pouvant être géographiquement répartis, de fournir une représentation synthétique de l'univers d'intérêt." Bloch [10] propose une définition plus générale relative à la fusion d'informations, qui consiste à "combiner des informations issues de plusieurs sources afin d'améliorer la prise de décision." Dans tous les cas, la masse de données nécessaire au cours du traitement est extrêmement importante. Elle comprend en effet des connaissances *a priori*, des connaissances issus des capteurs et des traitements précédents. Les auteurs de [42] précisent encore la nécessité d'une définition large donnée au terme "capteur", pouvant inclure des capteurs traditionnels (caméras, radars, etc.) mais aussi des renseignements en provenance d'observateurs humains. Ces définitions du contexte et des objectifs de la fusion de données correspondent ainsi à ceux définis pour l'analyse de données dans le cadre la télésurveillance médicale à domicile. Il s'agit bien de fournir une représentation pertinente et fiable de la situation d'une personne à domicile à partir de divers types de capteurs et de connaissances dans un objectif d'aide à la décision.

Les problèmes liés à la fusion de données sont variés, tels que la gestion du temps réel, de la masse de données nécessaires, des incertitudes et imprécision des informations, du choix des capteurs, de leur synchronisation, etc. L'étendue des techniques de fusion de données utilisées

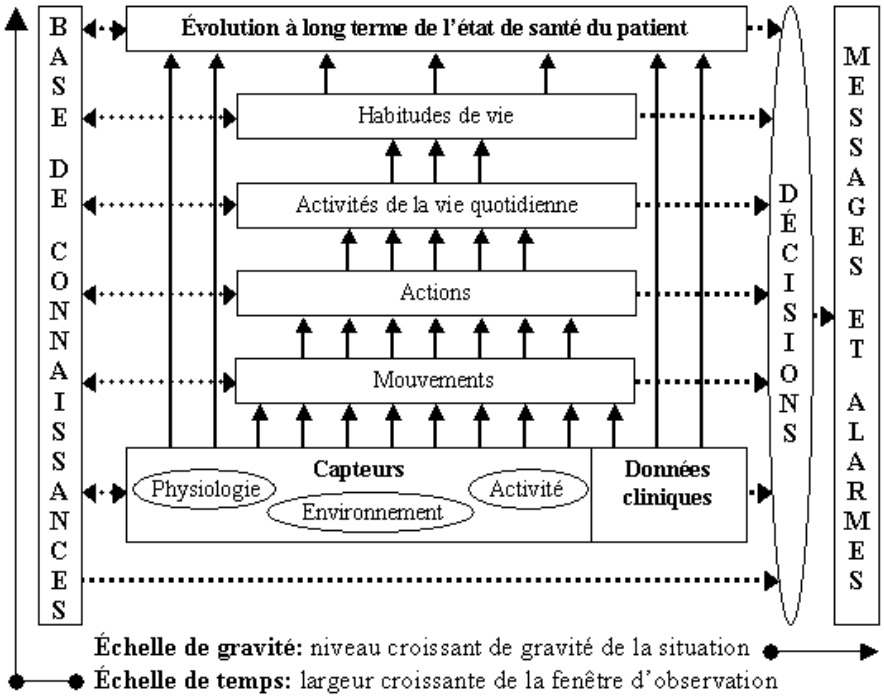


FIG. 3.1 – Une approche granulaire pour la détection de situations critiques.

pour répondre au contexte et aux objectifs d'une applications est vaste et nécessite un travail interdisciplinaire [42] (intelligence artificielle, imagerie, traitement du signal, parallélisme, reconnaissance de formes, aide à la décision, etc.). Dans le cadre de la télésurveillance médicale à domicile, une caractéristique importante est le traitement de *données temporelles* issues de l'observation de l'évolution dans le temps de la situation d'une personne à domicile. La définition d'une méthode appropriée de fusion de données dans ce contexte nécessite cependant une formulation précise la problématique d'intérêt (voir section 3.2).

3.1.5 Architecture du système : une approche granulaire

Un système de décision sur la situation d'une personne à domicile est finalement décomposé en plusieurs étapes correspondant à des niveaux croissants de compréhension de la situation du patient (**échelle de temps**) (voir Fig. 3.1). À chaque niveau peut être généré un ensemble de messages et d'alarmes (**échelle de gravité**), et ainsi chaque étape constitue à elle seule un sous-système de décision intégré au système de décision global. La base de l'architecture est constituée de l'ensemble des capteurs et des données cliniques du patient. Les décisions dépendent des informations contenues dans une base de connaissance : seuils de décision, historique d'événements, profil comportemental dans des conditions habituelles, etc. Les niveaux croissants de complexité sont décrits ci-dessous :

- **Capteurs :** données brutes ou prétraitées (cas de capteurs "intelligents"), correspondant à plusieurs classes de capteurs : (a) activité du patient (position, déplacements, chute, etc.), (b) environnement (température, utilisation des portes, des fenêtres, de l'éclairage, etc.) et (c) physiologie (pressions artérielles, poids, etc.).
- **Mouvements :** données brutes des capteurs filtrées, échantillonnées et organisées. On

peut également associer un niveau d'incertitude aux données reçues selon une décision binaire (le patient est dans la chambre et dans aucune des autres pièces par exemple), une décision floue (possibilité pour le patient d'être dans chacune des pièces), ou une décision probabiliste (probabilité pour le patient d'être dans chacune des pièces).

- **Actions :** séquences de mouvements analysées pour la reconnaissance (a) des postures du patient (être couché, s'asseoir, se lever, marcher, etc.), (b) de ses déplacements dans les pièces de l'habitat et (c) d'un ensemble de sons usuels (vaisselle, téléphone, etc.).
- **Activités de la vie quotidienne (AVQ) :** fusion de séquences temporelles de différents types d'actions et d'informations sur l'environnement (utilisation des portes, des fenêtres, du matériel ménager, etc.) et en tenant compte des connaissances acquises sur le comportement du patient (dormir, faire sa toilette, préparer un repas, se reposer, etc.).
- **Habitudes de vie :** observation journalière de séquences d'activités et caractérisation de ces activités en terme de (a) fréquence, (b) intensité, (c) durée, (d) horaire et (e) ordre d'occurrence par exemple. Il s'agit ensuite notamment de comparer ces observations avec un modèle de comportement de la personne - obtenu par apprentissage - pour évaluer leur "normalité".
- **Evolution à long terme :** évaluation globale de la situation du patient sur le long terme (plusieurs semaines voire plusieurs mois) à partir d'informations de différentes origines : (a) des paramètres représentatifs des habitudes de vie de la personne et de leur "normalité", (b) des caractéristiques de son environnement telles que les températures intérieures et extérieures (susceptibles d'influencer le rythme cardiaque par exemple), (c) des données physiologiques, (d) des données cliniques telles que l'âge du patient ou son genre et (e) des critères globaux de la base de connaissances réévalués périodiquement (cognition, vision, équilibre, nombre de chutes antérieures, etc.).

3.2 Formulation de la problématique *Étude des habitudes de la vie quotidienne*

Les travaux de recherche réalisés au cours de cette thèse concernent l'**étude à long terme de l'évolution de l'état de santé d'une personne**. Cela correspond au niveau supérieur de compréhension de la situation d'une personne (voir Fig. 3.1). Dans cet objectif, on s'intéresse ainsi particulièrement à l'**observation des habitudes de vie quotidienne** de la personne à domicile. Toute dégradation de l'état de santé a en effet des répercussions immédiates sur le comportement dans les activités de la vie quotidienne : observation d'une lenteur à effectuer certaines activités ou diminution globale de l'activité par exemple.

D'après l'architecture proposée pour le système de décision sur la situation d'une personne à domicile, l'observation des habitudes de vie quotidienne peut être formulée comme la construction d'un **système d'apprentissage** dont l'objectif est d'identifier un profil ou modèle comportemental de la personne **à haut niveau** à partir de la **fusion de différents types de données**. L'évolution vers une situation inquiétante correspond alors à un écart observé par rapport à ce profil.

Des informations sur les habitudes de vie sont disponibles par l'intermédiaire des capteurs et par les connaissances *a priori* de la personne et de son fonctionnement (données cliniques, base de connaissances). On se place ainsi dans une perspective de **fusion de données**, permettant l'analyse conjointe de données hétérogènes (quantitatives ou qualitatives) en provenance à la fois de capteurs, de la base de données relative à la personne télésurveillée (données cliniques) et

d'une base de connaissances dynamique (historique de données et d'événements, connaissances déclaratives, seuils de décision, etc.). Cela rejoint la vision de Bracio *et al.* [15] qui considèrent la problématique de fusion de données en biomédical comme la combinaison d'un ensemble de mesures disponibles (issues de capteurs ou non) pour obtenir la meilleure estimation possible de l'état de santé d'une personne.

À chaque instant, la décision sur l'état d'une personne dépend des valeurs des paramètres à cet instant mais aussi aux instants précédents, ainsi que d'un ensemble de connaissances obtenues par apprentissage dans le temps et par enregistrement d'événements antérieurs. La chronologie et la répartition des valeurs de paramètres et événements jouent un rôle fondamental dans l'évaluation d'une situation, et la **composante temporelle** est donc essentielle à considérer. Par ailleurs, même si *a priori* certains événements tels qu'une chute semblent indépendants du temps car difficiles à prévoir, leur observation temporelle permettra probablement une meilleure connaissance a posteriori de leur contexte d'occurrence voire finalement leur prévention. On sait déjà qu'un grand nombre de critères peuvent en effet influencer l'apparition de ces situations critiques et aider à les prévenir : Cameron *et al.* [18] ont par exemple étudié les facteurs agissant sur le risque de chute chez les personnes âgées pour aboutir à une meilleur prévention des accidents.

Finalement, la résolution du problème de décision sur l'évolution de la situation générale d'une personne à domicile correspond à la **fusion de séquences temporelles de données multidimensionnelles et hétérogènes – qualitatives ou qualitatives**. L'objectif est de repérer des **régularités dans le temps** pour définir un modèle ou profil de comportement d'une personne, les situations critiques correspondant alors à des écarts par rapport à ce modèle. On s'intéresse ainsi à l'**apprentissage du profil comportemental d'une personne à domicile à partir de l'étude de séquences temporelles de données multidimensionnelles et hétérogènes** le caractérisant.

3.3 Principe de résolution
Vers l'apprentissage d'un profil comportemental

Le principe de résolution du problème d'étude des habitudes de vie d'une personne pour la détection de situations critiques est ainsi défini comme nécessitant la construction de son profil comportemental, caractérisé par un ensemble d'activités habituellement réalisées à domicile. Il s'agit alors de concevoir un système d'apprentissage de ces activités régulières, à partir de l'observation de séquences temporelles de données multidimensionnelles et hétérogènes représentatives d'une situation "normale" de la personne – les données enregistrées par les capteurs, appelées dans ce contexte les *données d'apprentissage*. L'objectif est d'**extraire les régularités temporelles**, c'est-à-dire les sous-séquences récurrentes dans l'ensemble de ces données. La caractérisation et la classification des sous-séquences récurrentes permet d'identifier celles qui sont représentatives des activités types réalisées par la personne dans sa vie quotidienne.

Ce paragraphe décrit d'abord dans ce contexte en 3.3.1 les caractéristiques des régularités observées, puis en 3.3.2 le principe de résolution de leur identification par assimilation de ce problème d'apprentissage à un cycle de décision.

3.3.1 Caractéristiques des régularités observées dans les données

Étant donné que l'on considère un ensemble multidimensionnel de données hétérogènes comme ensemble de données d'apprentissage, l'extraction des régularités comportementales d'une per-

sonne peut s'envisager selon deux types d'analyse :

(a) Analyse monodimensionnelle de chaque paramètre, permettant d'extraire les régularités temporelles pour chaque paramètre et de les fusionner ensuite pour obtenir une vision globale des régularités pour l'ensemble des paramètres observés ;

(b) Analyse multidimensionnelle des régularités temporelles aboutissant directement à l'identification des comportements fréquents de la personne en terme de variation conjointe des différents paramètres.

Dans notre contexte d'étude, on suppose que les paramètres considérés sont sélectionnés comme un compromis entre deux types de critères fondamentaux :

(1) Critères individuels, éthiques et sociaux : être facilement observables, de façon non-invasive, et en respectant la vie privée des personnes télésurveillées ;

(2) Critères d'utilité et d'efficacité : donner une appréciation la plus complète possible de la situation de la personne à domicile, sensible à toute détérioration de son état de santé.

Par conséquent, tous les paramètres observés sont fortement liés les uns aux autres, et il convient ainsi de préserver au maximum les variations conjointes des paramètres observés par la recherche de régularités temporelles – appelées aussi *motifs* – sur la base d'une **analyse multidimensionnelle**.

Par ailleurs, les relations entre les paramètres varient dans le temps, en fonction par exemple du moment de la journée, de l'activité réalisée, mais aussi de la situation générale et de l'état de santé de la personne télésurveillée. Toute réduction de dimension sur la base des variations observées entre les différents paramètres *en général* ou *dans une situation habituelle de la personne* correspond alors à une sur-simplification du système qui n'est plus assez sensible et ne détecte pas certaines évolutions critiques de la situation de la personne. Par exemple, si sur la base de l'observation d'une forte corrélation entre le niveau d'activité d'une personne et sa fréquence cardiaque on décide de simplifier le système en ne considérant qu'un de ces deux paramètres ou en les projetant sur une seule dimension, on prend le risque de ne plus observer une augmentation globale de la fréquence cardiaque avec un rythme d'activité moins soutenu. Cette situation peut pourtant être considérée comme inquiétante.

Notre objectif est ainsi de réaliser l'extraction de **motifs multidimensionnels** dans des séquences de données temporelles multidimensionnelles et hétérogènes, représentatives des habitudes de vie de la personne.

3.3.2 Résolution de l'identification des régularités comme un cycle de décision

Cet apprentissage des régularités temporelles peut être assimilé à un **système de décision** sur l'appartenance ou non de sous-séquences des données d'apprentissage à un comportement régulier de la personne, ou plus généralement à un **système de reconnaissance** qui concerne l'identification des habitudes de vie quotidienne. Dans toute problématique de décision, il est indispensable de bien spécifier les objectifs et le contexte de la prise de décision pour assurer une bonne adéquation des systèmes implémentés à la résolution du problème. Cette étape est particulièrement critique dans le cadre de problématiques complexes mettant en jeu plusieurs niveaux de détail entre les connaissances disponibles, les données enregistrées et les objectifs de la décision. C'est le cas de notre contexte d'études puisqu'on dispose de données très "bas niveau" – les données enregistrées par les capteurs – pour une prise de décision "haut niveau" – l'identification des comportements quotidiens réguliers de la personne. Il est alors nécessaire de

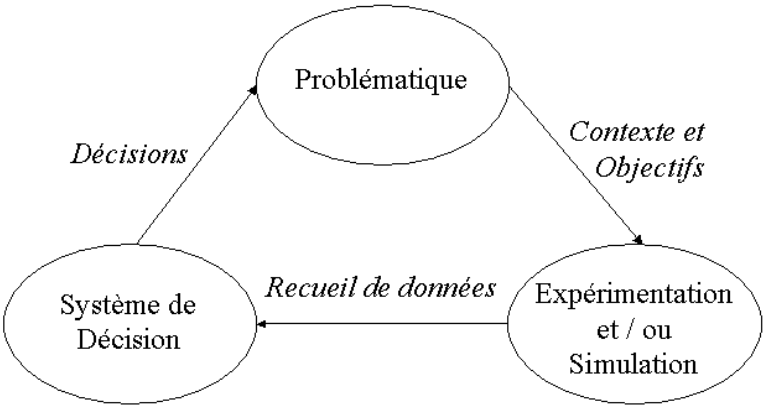


FIG. 3.2 – Les étapes de résolution d'un problème de décision.

bien poser les besoins, les exigences et les contraintes du système afin que la prise de décision corresponde au mieux aux objectifs.

Finalement, la mise en place d'un processus de décision s'inscrit dans un schéma récursif dont les grandes étapes sont les suivantes (voir Fig. 3.2) :

1. **Définition du contexte et des objectifs généraux de la décision.** L'objectif est de spécifier l'espace des données, des connaissances et des informations qu'il est pertinent de considérer en rapport aux objectifs du processus de décision. Les questions importantes que l'on se pose à cette étape sont par exemple : Quelles sont les observations pertinentes à réaliser dans un contexte expérimental ? À quel niveau de détail doit-on considérer ces observations ? Quelles sont les performances attendues du système de décision ?
2. **Enregistrement ou génération de données appropriées au contexte de la décision.** La mise en place d'un processus expérimental pour la collecte de données adaptées au problème de décision est guidée par les objectifs et le contexte de la décision, définis à l'étape précédente de résolution du problème. Cependant, l'acquisition d'ensembles de données suffisamment complets et représentatifs pour la mise au point d'un système de décision performant peut s'avérer difficile pour certaines applications. La mise en place d'un processus de simulation est alors bien utile comme une première étape de résolution du problème pour faire face au manque de données expérimentales. En faisant varier les paramètres de la simulation, on peut aussi obtenir des données représentatives de tout un ensemble de situations possibles, mais qu'il n'est pas toujours facile d'observer exhaustivement dans la réalité d'un processus expérimental. Les performances du système de décision peuvent ainsi être évaluées plus complètement grâce à la simulation.
3. **Test de méthodes adaptées pour la résolution du problème de prise de décision.** Les données acquises par l'expérimentation ou générées par la simulation sont utilisées comme entrées du processus de décision implémenté pour résoudre le problème. La sensibilité et la spécificité des algorithmes expérimentés doivent correspondre aux exigences de résolution du problème définies lors de la première étape.

Une fois un algorithme de décision implémenté et expérimenté, l'adéquation entre les performances de la décision et les objectifs de résolution permet de déterminer si une prochaine étape de raffinement du processus de décision est nécessaire ou non. La résolution du problème de décision doit aussi prendre en compte des étapes de validation des choix effectués et des résul-

tats obtenus à chaque phase du processus afin de s'assurer continuellement de leur adéquation au contexte et aux objectifs de la décision. Ces étapes de validation sont réalisées soit avec des experts des différents domaines impliqués, soit à l'aide d'outils mathématiques et statistiques lorsque des données de référence sont disponibles.

3.4 Contraintes de résolution *Contexte et objectifs de la télésurveillance médicale*

D'après le schéma de résolution d'un problème de décision (voir Fig. 3.2), la définition rigoureuse de la problématique considérée – c'est-à-dire du contexte et des objectifs de résolution – a des conséquences d'une part sur la mise en place d'un processus expérimental et/ ou de simulation adapté, et d'autre part sur la construction d'un système de décision efficace à partir des données acquises par l'expérimentation et/ ou de la simulation.
Dans ce paragraphe, on définit ces caractéristiques fondamentales de résolution d'un problème en **trois types de contraintes** sur la mise en place de l'expérimentation (ou de la simulation) et la décision :

- **Niveau de performance exigé.** Les performances de la décision sont définies et évaluées en terme de taux de bonnes détections et de fausse alarmes (sensibilité et spécificité), associés éventuellement à un temps de décision "acceptable".
- **Niveau de détail nécessaire.** La définition du niveau de détail nécessaire pour la résolution d'un problème est fondamentale pour sélectionner au mieux : (a) les connaissances les plus appropriées, et par conséquent le niveau de détail de l'acquisition de données et de leur représentation ; et (b) les algorithmes les plus adaptés à la résolution du problème. Il y a un compromis à trouver entre la nécessité de préserver la complexité du problème considéré, et la restriction à un niveau de détail pertinent, c'est-à-dire qui correspond aux objectifs et aux performances exigées de la décision.
- **Niveau de connaissance disponible.** La définition de ce critère permet d'identifier l'ensemble des connaissances qu'il est possible de prendre en compte pour la décision, en tenant compte du niveau de détail nécessaire. Le manque de connaissances liées au problème étudié peut nécessiter de s'appuyer sur une variété de sources d'information, incluant : (a) des connaissances *a priori* – intuitives ou académiques – et (b) des connaissances extraites d'ensembles de données expérimentaux. La fusion de plusieurs types de connaissances est particulièrement adaptée á la résolution de problèmes complexes.

Dans le contexte de la télésurveillance médicale à domicile, la recherche de l'extraction de motifs représentatifs du comportement habituel d'une personne est une **problématique "haut niveau"**. L'idée fondamentale de cette étude n'est en effet pas d'identifier précisément les problèmes qui ont lieu à domicile, mais de relater le contexte de toute modification globale de comportement. Par conséquent, l'extraction de motifs a pour objectif d'identifier les comportements récurrents à l'échelle d'intervalles de temps relativement longs, c'est-à-dire d'environ trente minutes à plusieurs heures. Par ailleurs, ces comportements récurrents ne sont pas identifiés précisément : il existe en effet une grande variabilité dans la réalisation d'une même activité par une personne.
Les "bons" niveaux de paramètres à considérer pour la résolution de ce problème sont ainsi les suivants :

- **Niveau de performance.** Les résultats du système de décision doivent correspondre aux objectifs généraux de la surveillance : détecter tous les comportements réguliers (sensibi-

lité), avec un faible taux de fausses détections (spécificité), et en un temps de détection "acceptable". On cherche à ne détecter que les motifs "haut niveau", représentatifs d'un comportement habituel d'une personne, et sans la nécessité d'un trop grand nombre de données d'apprentissage.

- **Niveau de détail.** Pour la résolution de problèmes à haut niveau de décision, comme c'est le cas dans notre contexte, le système de décision ne nécessite pas une grande précision des connaissances et des données mises en jeu dans le processus. Les données expérimentales de bas niveau, telles que les données enregistrées par les capteurs de l'habitat, vont ainsi nécessiter un haut niveau de représentation pour mettre en évidence les tendances de variation sur le long terme, et supprimer les variations locales non significatives à cette échelle de décision.
- **Niveau de connaissance.** L'apprentissage d'un profil de comportement est spécifique à la personne télésurveillée. Par conséquent, il y a très peu de connaissances *a priori* qui puissent être intégrées dans le processus de décision. Ce dernier doit donc être construit essentiellement à partir des données enregistrées au domicile par l'ensemble des capteurs considérés, et de façon non supervisée. Afin d'obtenir une appréciation complète de la situation de la personne, sensible à toute détérioration de son état de santé, les données sont acquises de capteurs de différents types : (1) activité (localisation dans le domicile, posture de la personne, etc.), (2) environnement (température, utilisation des portes, des fenêtres, de l'éclairage, etc.), et (3) physiologie (fréquence cardiaque, pressions artérielles, poids, etc.). Des contraintes éthiques, sociales et individuelles doivent être également considérées pour la sélection des paramètres les plus pertinents : respect de la vie privée, confidentialité des données, discrétion des installations qui équipent les habitats, et amélioration de la qualité de vie des patients (confort, sécurité). Par ailleurs, l'adoption de tels systèmes de surveillance repose sur des contraintes économiques fortes : pas de coûts excessifs engagés par les patients et réduction des coûts de santé publique. Le respect de l'ensemble de ces contraintes est une des clés de l'acceptabilité sociale et individuelle des systèmes de télésurveillance médicale à domicile.

3.5 Synthèse de la résolution du problème de décision

La résolution du problème de décision sur le situation d'une personne à domicile est finalement formulée comme la **construction d'un profil comportemental** de cette personne dans ses activités de la vie quotidienne. Les situations inquiétantes ou critiques correspondent alors à l'observation d'un écart de comportement par rapport à ce profil.

On s'intéresse ainsi dans ce travail à l'apprentissage du profil comportemental d'une personne, qui s'exprime comme l'**extraction non supervisée de motifs temporels multidimensionnels de "haut niveau"** à partir de **séquences temporelles multidimensionnelles et hétérogènes observées à "bas niveau"**, par l'intermédiaire de capteurs fonctionnant en permanence au domicile, et représentatives du comportement habituel que l'on cherche à modéliser.

Une fois le contexte et les objectifs de la problématique bien identifiés, et selon le schéma de résolution d'un problème de décision proposé précédemment (voir Fig. 3.2), la mise en place du processus de résolution met en jeu deux grandes étapes successives :

(1) d'une part l'acquisition de données appropriées á la résolution du problème, par l'expérimentation et / ou la simulation ; et

(2) d'autre part l'expérimentation d'algorithmes adaptés pour la résolution effective du problème posé.

Le développement des projets en télésurveillance médicale et la collecte de données en environnement réaliste n'en étant encore qu'au stade de leur mise en oeuvre expérimentale (voir chapitre 2 sur l'état de l'art), la première de ces deux étapes consiste exclusivement dans notre contexte en la mise au point d'un processus de simulation de données qui sont potentiellement enregistrées à partir de capteurs installés dans l'habitat de la personne télésurveillée.

Ainsi, les deux grandes étapes de ce travail de thèse, présentées dans les deux prochaines parties, sont les suivantes :

Partie II. **Processus de simulation** – Construction d'un processus de simulation de séquences temporelles de données multidimensionnelles et hétérogènes correspondant à des données "bas niveau" obtenues par un ensemble de capteurs installés à domicile et représentatives du comportement habituel de la personne.

Partie III. **Système de décision** – Mise au point d'une méthode générique d'extraction non supervisée de motifs temporels multidimensionnels à partir de séquences de données multidimensionnelles et hétérogènes. L'expérimentation de la méthode proposée est réalisée dans le cadre de l'extraction de régularités temporelles "haut niveau", qui soient représentatives des activités typiques d'une personne dans sa vie quotidienne.

Chacune de ces parties du travail implique des phases d'expérimentation et de validation des méthodes implémentées et des résultats obtenus. Le dernier chapitre présente les conclusions et perspectives de ce travail.

Deuxième partie

Processus de simulation

1

Introduction : Pourquoi un processus de simulation ?

La mise en place d'un processus de simulation a pour objectif la génération d'un grand nombre de séquences temporelles de données multidimensionnelles représentatives du comportement habituel d'une personne à domicile. Ce processus se situe comme une part intégrante du schéma de résolution de la problématique d'étude du comportement habituel d'une personne à domicile (voir Fig. I.3.2) : la mise en place de la simulation est conditionnée par le contexte et les objectifs de résolution de la problématique d'une part, afin de produire des données appropriées à l'expérimentation d'algorithmes de décision d'autre part. L'implémentation et la mise au point de tout système de décision nécessite en effet des ensembles de données réalistes et adaptés à la problématique étudiée. Ces ensembles peuvent être constitués à partir d'expérimentations sur le terrain et / ou générées par la simulation.

Dans le domaine de la télésurveillance médicale à domicile, les projets de recherche en sont encore aux premiers stades de leur développement, et les expérimentations en environnement réaliste ont ainsi à peine débuté. On ne peut donc pas disposer d'ensembles de données suffisamment réalistes et complets pour constituer la seule base de mise au point de systèmes de décision sur les habitudes de vie d'une personne à domicile. Par ailleurs, une étude complète et fiable nécessite de prendre en compte plusieurs profils de personnes face à plusieurs types de situations. Cela permet d'expérimenter au mieux le système de décision en testant idéalement tous les cas que doit savoir traiter le système. La constitution par l'expérimentation de ces ensembles de test suffisamment complets et représentatifs s'avère de toutes façons être une tâche plutôt difficile dans notre contexte : observer dans des conditions "normales" de vie à domicile un grand nombre de personnes, faisant face à tout un ensemble de situations. C'est pour ces raisons que beaucoup de chercheurs se tournent vers la simulation, comme un moyen de contourner les difficultés de constitution de larges ensembles de données expérimentales complets et représentatifs. L'utilisation de données expérimentales enregistrées en environnement réaliste reste cependant absolument indispensable comme seconde étape de validation des systèmes de décision, l'utilisation de la simulation ne constituant qu'une première étape de test.

Par rapport à l'expérimentation, l'utilisation de la simulation permet ainsi aux chercheurs de disposer d'un univers de données complet et mieux contrôlé pour l'expérimentation de systèmes de décision. Les avantages se situent à plusieurs niveaux :

- Production de larges ensembles de données pour l'expérimentation d'algorithmes de décision.
- Génération de données représentatives d'un maximum de situations que le système de

décision peut être amené à traiter : plusieurs profils de personnes et de situations à domicile dans notre contexte.
- Construction d'un processus crédible et facilement compréhensible par n'importe quel acteur du système, par opposition aux approches de modélisation analytiques bien plus complexes.
- Meilleure connaissance *a posteriori* des paramètres étudiés : tendances et variations conjointes de différents paramètres observés simultanément dans le cadre de la télésurveillance médicale à domicile.
- Test de l'efficacité et de la robustesse des algorithmes de décision en faisant varier les paramètres de la simulation pour la génération de données couvrant un maximum de cas possibles.

Le manque de données expérimentales dans notre contexte de recherche d'une part, et les avantages à disposer de données simulées d'autre part, motivent la mise en place d'un processus de simulation. L'objectif est alors de construire un modèle dont la modification des paramètres permet de simuler différents profils de comportement, c'est-à-dire de générer des données associées à la surveillance de différents individus. Pour chacun d'eux, d'autres modifications progressives des paramètres permettent alors de simuler l'évolution d'une personne vers différentes situations, en particulier des conditions habituelles à des changements inquiétants dans la vie quotidienne.

La démarche de simulation a été réalisée avec la collaboration d'un ensemble de chercheurs de plusieurs domaines, tant pour la construction d'un modèle que pour sa validation. Cette étude nécessite en effet des connaissances variées qui concernent par exemple les habitudes de vie quotidienne d'une personne, les caractéristiques de paramètres physiologiques que l'on peut enregistrer à domicile, ou encore la démarche d'analyse statistique de données expérimentales.

2

État de l'art

La simulation peut être définie comme un processus de conception d'un modèle d'un système réel, et d'expérimentation de ce modèle à des fins de compréhension du comportement de ce système et / ou d'évaluation de ses différents modes de fonctionnement [101]. Cette définition met en évidence deux étapes dans un processus de simulation : d'une part la construction du modèle et d'autre part son expérimentation.

Dans le domaine médical, la simulation est actuellement largement utilisée mais principalement afin d'améliorer l'enseignement et la formation des étudiants et des praticiens, sans faire prendre de risques supplémentaires aux patients. Plus généralement, Kelton [54] souligne le succès croissant connu dans tous les domaines par la recherche et la pratique de la simulation dans les 25 dernières années. La simulation a été l'un des bénéficiaires de l'augmentation significative de la puissance de calculs des machines dans les dernières décades. L'idée même de la pratique de la simulation est ainsi devenue populaire et son utilité a été démontrée dans plusieurs contextes. Dans le contexte de la prise de décision, Shannon [101] considère que la simulation est l'un des outils les plus puissants disponible pour la conception et la mise en œuvre de processus et de systèmes complexes. Les avantages de la simulation par rapport aux modèles analytiques et mathématiques pour l'analyse de systèmes complexes se situent à différents niveaux :

(1) Les concepts de la simulation sont faciles à appréhender et à comprendre ;

(2) Un modèle de simulation est plus crédible qu'un modèle analytique car son comportement est fondé et comparé directement au fonctionnement du système réel et nécessite peu d'hypothèses ;

(3) La simulation permet d'expérimenter facilement de nouvelles situations peu familières.

Malgré ces avantages, la simulation n'est cependant pas toujours possible ou simple à mettre en œuvre. Toujours selon Shannon [101], l'utilité d'un processus de simulation dépend de plusieurs critères :

(1) Qualité du modèle de simulation ;

(2) Adéquation et qualité des données utilisées pour la construction du processus de simulation au contexte de celles-ci ;

(3) Bonne spécification du contexte de simulation pour la génération de données cohérentes avec les objectifs de résolution de la problématique étudiée.

Il est ainsi nécessaire d'être précis dans la définition du système et du modèle conceptuel afin de déterminer les niveaux d'abstraction et de simplification qui ne correspondent ni à une "sur-simplification" du système, ni à la génération de données à un niveau de détail trop fin [101]. En ce qui concerne la complexité d'un modèle de simulation, Chwif [32] conseille de construire

un modèle simple, et d'ajouter ensuite de la complexité si cela s'avère finalement absolument nécessaire. L'objectif est alors de déterminer le meilleur niveau de complexité d'un modèle donné qui lui permet de rester valide et approprié aux objectifs de son utilisation. Cependant, Chwif mentionne aussi le manque de méthodologies disponibles afin de guider le concepteur dans la réalisation du modèle le plus simple et adapté à sa problématique.

Plus généralement, Kelton [54] souligne les problèmes méthodologiques auxquels fait face la simulation pour la définition des techniques de modélisation, de conduite des expérimentations et d'interprétation des résultats. La recherche sur les méthodologies adaptées à la simulation est organisée autour de plusieurs thèmes principaux tels que la modélisation, la génération de nombres et de processus aléatoires, la conception et l'analyse statistique. Dans chacun de ces axes, les recherches peuvent être encore orientées autour de plusieurs domaines d'intérêt tels que les méthodes orientées objet, les méthodes graphiques, les architectures parallèles. Les méthodologies adaptées à la simulation varient ainsi largement en fonction du contexte de l'application concernée. Selon Kelton, il semble même que les recherches en simulation se sont dispersées dans des directions où finalement il n'y a plus forcément d'applications possibles des résultats obtenus. Dans [82], Ören met aussi en évidence la diversité des méthodologies relatives à la simulation. Il présente une taxonomie qui recense de nombreux types de simulation dont le choix dépend de critères variés tels que la prise en compte d'une composante temporelle, la relation fonctionnelle entre des variables descriptives, l'organisation des composantes du modèle, les objectifs de modélisation, la nature du modèle. Cela conduit à la définition d'une centaine de types de simulations possibles. On comprend alors mieux qu'il soit quasiment impossible d'extraire des lignes de conduite générales pour la conception d'un processus de simulation.

Il existe cependant une étape incontournable à toute conception d'un processus de simulation : la validation et la vérification du modèle conçu et implémenté et de son comportement. Cette problématique est abordée dans plusieurs articles tels que [100, 101]. Les étapes de vérification et de validation ont respectivement pour objectif d'évaluer si le processus de simulation (1) s'exécute conformément aux attentes de l'analyste, et (2) a un comportement en adéquation avec celui du système réel qu'il modélise. Dans [100], Sargent traite exclusivement de la vérification et de la validation des modèles de simulation. Il considère en particulier la question de l'intégration des étapes de vérification et de validation dans le processus de conception, d'implémentation et d'expérimentation d'un processus de simulation. Il présente les grandes étapes d'un tel processus et les techniques appropriées pour leur validation.

Dans notre contexte de recherche, on considère une approche incrémentale et hybride pour la simulation. La méthodologie utilisée est **incrémentale** pour permettre progressivement un raffinement du processus à deux niveaux :

1. **Au niveau de sa bonne adéquation à la résolution de la problématique** – vision globale de la simulation comme une étape du cycle de résolution d'une problématique donnée,
2. **Au niveau de la validation du modèle et de son implémentation** pour répondre aux exigences et objectifs d'un contexte donné – vision plus limitée au cycle de mise en place d'un processus de simulation, dans un contexte et avec des objectifs bien définis.

On décrira par ailleurs la nécessité d'une approche **hybride** dans notre contexte pour répondre à plusieurs objectifs :

1. Prendre en compte des données hétérogènes ;
2. S'adapter à la simulation de données multidimensionnelles et corrélées ;
3. Intégrer plusieurs types de connaissances dans le processus : des connaissances *a priori* –

connaissances de sens commun et académiques – aussi bien que des connaissances extraites de données expérimentales.

3

Méthodologie pour la simulation

La mise en place d'un processus de simulation est une démarche d'autant plus complexe qu'elle concerne des paramètres hétérogènes, corrélés, dans un contexte où différents types de connaissances sont disponibles. Dans les paragraphes de ce chapitre nous décrivons une méthodologie pour la simulation à ce niveau de complexité.

3.1 Démarche incrémentale

3.1.1 La simulation : une partie intégrante du cycle de résolution d'un problème

La conception d'un processus de simulation n'a de sens que dans la perspective plus globale du contexte et des objectifs de son utilisation. La simulation doit en effet être considérée comme une étape d'une démarche plus générale de résolution d'un problème de décision (voir Fig. 3.2). La génération par la simulation de données appropriées au contexte de la décision – de même que la collecte de données dans un contexte expérimental – est guidée par le contexte et les objectifs généraux de la problématique de décision. Cette définition du cadre de la simulation permet de bien spécifier et de limiter au minimum requis l'espace des données, informations et connaissances à prendre en compte dans la simulation. L'objectif est de construire un modèle qui ne soit ni plus complexe que nécessaire, ni trop simplifié par rapport aux objectifs d'utilisation des données qu'il génère. Les questions importantes à considérer pour concevoir un modèle bien adapté sont alors du type : quels sont les paramètres pertinents à intégrer dans le processus de simulation ? ou encore quel est le niveau de détail nécessaire dans les données générées ? Une mauvaise adéquation des données simulées – ou collectées expérimentalement – aux objectifs de la décision induit de mauvaises performances associées à la décision quel que soit le système de décision considéré : entre autres, la sensibilité et la spécificité ne correspondent pas aux taux escomptés. Un cycle supplémentaire dans la démarche de résolution du problème est alors nécessaire, afin de mieux répondre aux objectifs fixés.

La conception d'un processus de simulation doit ainsi prendre en compte les paramètres clés de la résolution de tout problème de décision : (1) le niveau de performance exigé, (2) le niveau de détail nécessaire et (3) le niveau de connaissance disponible (voir paragraphe I.3.4).

(1) Le **niveau de performance exigé** intervient au moment de la validation des résultats du système de décision. Des résultats en adéquation avec les objectifs attestent d'une part de l'efficacité du système de décision et d'autre part de la pertinence des données utilisées en entrée du système et, par conséquent, de la validité des données produites par la simulation.

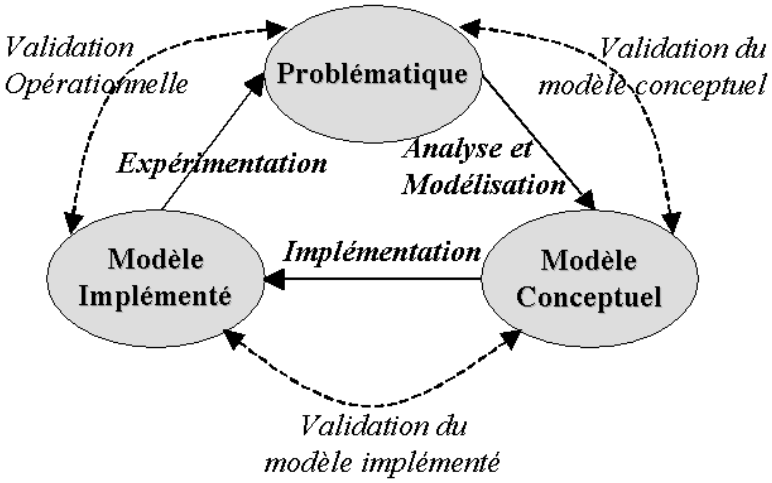


FIG. 3.1 – Les étapes de mise en place d'un processus de simulation [100].

Au contraire, de mauvais résultats peuvent remettre en cause l'un et l'autre du processus de simulation – ou d'expérimentation – et du système de décision.

(2) Une question particulièrement critique pour la mise en place d'un processus de simulation adapté aux performances exigées de la décision concerne le **niveau de détail nécessaire**. Plus la décision est "bas niveau", plus les données simulées devront être détaillées et précises. Au contraire, une décision plus "haut niveau" nécessite surtout de respecter les tendances globales et les variations conjointes dans les valeurs des paramètres simulés.

(3) **Le niveau des connaissances disponibles** et intégrées dans la conception du modèle de simulation doit être approprié au niveau de détail nécessaire. La simulation de processus complexes nécessite d'appuyer la modélisation et la validation des données simulées sur une diversité des sources d'informations relatives aux différents concepts à intégrer dans le processus de simulation. Les sources d'informations disponibles incluent un ensemble de connaissances *a priori* – **connaissances de sens commun** et **connaissances académiques** – et des ensembles de données expérimentales. D'autres connaissances – nommées **connaissances extraites** – sont inférées à partir d'études mathématiques et statistiques sur les données expérimentales.

3.1.2 La simulation : un cycle de raffinement à part entière

De même que le cycle de résolution d'une problématique de décision dans lequel il est intégré, un processus de simulation est un cycle de raffinement à part entière, pour progressivement aboutir à la génération de données appropriées au contexte de décision considéré (voir Fig. 3.1). La mise en place d'un processus de simulation comprend les étapes suivantes [100] :

1. **Construction du modèle conceptuel** en fonction de l'analyse des besoins et des objectifs de la problématique de décision dans laquelle la simulation s'intègre.
2. **Implémentation de ce modèle** en un modèle opérationnel.
3. **Expérimentation** pour la génération de larges ensembles de données.

Une démarche de vérification et validation est nécessaire lors de chacune des étapes de construction du processus de simulation. Les objectifs successifs sont les suivants :

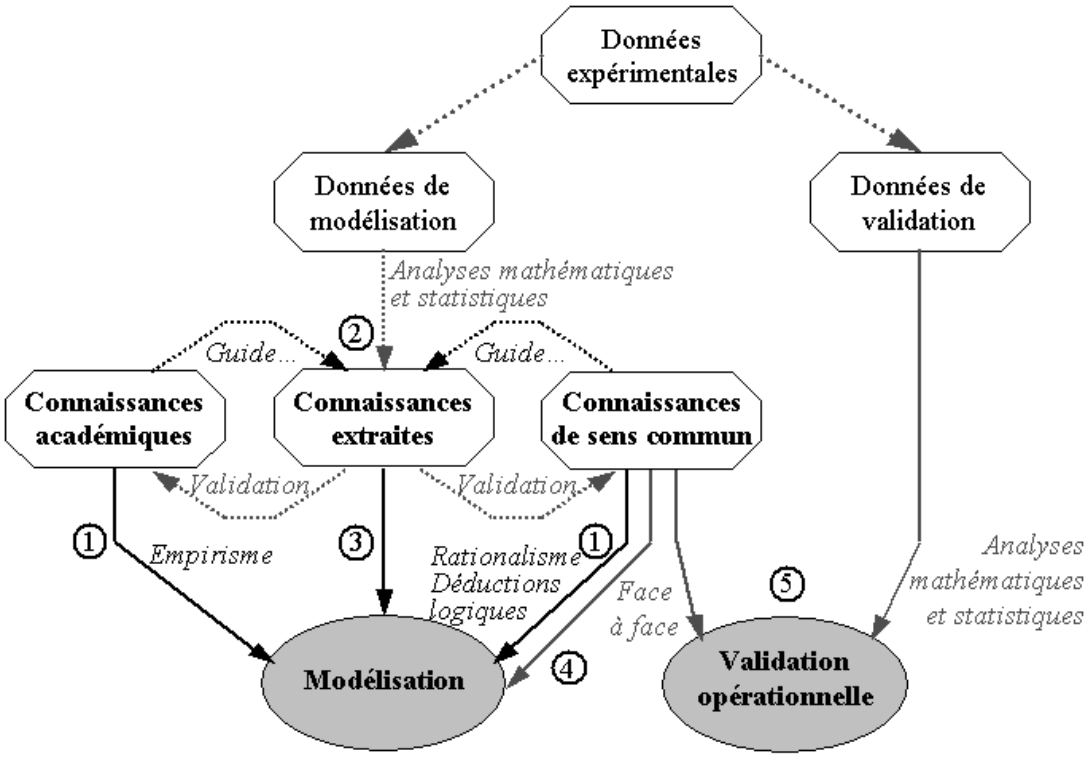


FIG. 3.2 – Intégration des différents types de connaissances disponibles dans le cycle de construction et de validation d'un processus de simulation.

1. Validation des théories et des hypothèses sous-jacentes à la définition du modèle conceptuel, et vérification de la structure du modèle et de sa logique par rapport aux objectifs fixés.
2. Vérification de l'implémentation correcte du modèle et du bon fonctionnement des programmes informatiques.
3. Validation du comportement de la simulation par rapport au système réel et aux objectifs fixés.

3.2 Approche hybride

Une approche hybride est nécessaire pour la simulation afin de répondre à la complexité du problème considéré. Pour aborder des problématiques complexes, on a souvent recours à la fusion de données. On propose ainsi, pour faire face à la complexité, d'intégrer plusieurs types de connaissances aux différentes étapes du processus de simulation et de sa validation, incluant principalement d'une part la construction et la validation du modèle conceptuel (modélisation), et d'autre part la validation des données produites par la simulation (validation opérationnelle) (voir Fig. 3.2) :

- **Connaissances de sens commun** – connaissances intuitives d'un expert sur le fonctionnement d'un système ;
- **Connaissances académiques** – connaissances avérées et validées au terme de plusieurs travaux de recherche, dans différents contextes expérimentaux ;

- **Connaissances extraites de données expérimentales** – nouvelles connaissances sur le fonctionnement d'un système, dont la définition peut être guidée par des connaissances déjà acquises (connaissances de sens commun et/ ou académiques), et qui sont ensuite validées par l'observation de données expérimentales.

Les connaissances de sens commun sont exclusivement qualitatives, alors que les deux autres types de connaissances peuvent être soit qualitatives, soit quantitatives. Les **connaissances qualitatives** sont particulièrement intéressantes à deux niveaux :

1. Ces connaissances donnent une idée des concepts fondamentaux sous-jacents au modèle de simulation ;
2. Les connaissances des sens commun des experts peuvent servir aux étapes de validation, à la fois (a) pour la validation du modèle conceptuel, et (b) pour celle des données générées par la simulation lorsqu'on ne dispose pas de données expérimentales pour comparer les résultats obtenus aux données enregistrées par un système réel.

Les **connaissances quantitatives** permettent la validation et/ ou la quantification des concepts afin de disposer d'un modèle prêt pour l'implémentation. Certaines de ces connaissances sont obtenues par l'analyse de données expérimentales. Cependant, seule une partie des données expérimentales disponibles est utilisée à cette fin d'extraction de connaissances – les *données de modélisation*, l'autre partie étant dédiée à la validation des séquences de données produites par la simulation – les *données de validation*.

On propose finalement les étapes successives suivantes pour l'intégration des connaissances disponibles dans le processus de simulation, selon la numérotation de la figure 3.2 :

1. Définition des concepts sous-jacents au modèle de simulation ;
2. Validation de certains concepts par l'extraction de connaissances à partir de données expérimentales ;
3. Intégration de ces connaissances extraites dans le modèle – en particulier, quantification des concepts ;
4. Validation des autres concepts de modélisation par les experts ;
5. Validation des données de simulation produites à partir du modèle conceptuel implémenté.

3.3 Techniques de simulation

3.3.1 Généralités

Les types de connaissances intégrées à chaque étape du processus de simulation et de sa validation déterminent le choix des techniques de modélisation et de validation appropriées. Quelques exemples de techniques et leur contexte d'utilisation, extraites de [100], sont détaillés ci-dessous :

- **Rationalisme et déductions logiques :** cette technique repose sur des connaissances de sens commun et suppose que n'importe qui est capable de savoir si les hypothèses sous-jacentes sont vraies.
- **Empirisme :** cette technique met en jeu des connaissances académiques et exige que chaque hypothèse soit validée de façon empirique.
- **Analyses mathématiques et méthodes statistiques :** ces analyses sont utiles au test de théories et d'hypothèses sous-jacentes au modèle conceptuel, à partir de données expérimentales appropriées. Elles permettent d'extraire des informations sur le domaine étudié et d'en inférer de nouvelles connaissances.

- **Face à face avec un expert :** cette technique implique de demander à des personnes expertes dans les domaines concernés si le modèle envisagé ou le comportement du système leur paraît "raisonnable", en faisant appel à leurs connaissances de sens commun.

Par ailleurs, le type des paramètres simulés et le niveau de détail requis pour la génération de leurs valeurs successives conditionnent le type et le niveau des connaissances nécessaires, et par conséquent le choix d'une technique de construction et de validation de la simulation. Une simulation "bas niveau" – simulation précise des valeurs de paramètres – met probablement en jeu un grand nombre de connaissances de type quantitatif, c'est-à-dire des connaissances académiques ou extraites de l'analyse de données expérimentales, même si les concepts intégrés dans le modèle peuvent être issus de connaissances qualitatives. Une simulation "haut niveau" – simulation de tendances de variation – intègre par contre plus certainement des connaissances de type qualitatif, c'est-à-dire des connaissances de sens commun ou académiques, ou bien des connaissances quantitatives à faible niveau de précision sur les variations des paramètres étudiés – connaissances académiques ou extraites de données expérimentales.

3.3.2 Principe de modélisation

Les principes de modélisation utilisés pour la simulation dépendent principalement du type des paramètres correspondant aux données simulées.

- **Paramètres qualitatifs.** Dans le cadre de la simulation, on suppose qu'un paramètre qualitatif a un ensemble fini de valeurs qualitatives possibles, appelées *modalités*. La modalité la plus probable ou la plus réaliste à chaque instant est alors déterminée par l'état le plus probable d'un **automate à états finis**. La structure de cet automate est guidée par le *rationalisme* ou l'expérience. Les probabilités initiales de chaque état et de transition entre les états peuvent dépendre d'un ensemble de paramètres d'influence du paramètre modélisé. Les paramètres de l'automate sont déterminés par l'*analyse de données expérimentales* si le système réel est observable et qu'on dispose des données expérimentales correspondantes. Sinon ils sont déterminés *intuitivement* et validés avec l'aide d'une personne ayant une bonne connaissance du domaine appréhendé.
- **Paramètres quantitatifs.** Un paramètre quantitatif prend ses valeurs dans un intervalle continu de valeurs réelles. Les valeurs les plus probables à chaque instant sont déterminées à partir de **distributions** de valeurs appropriées, dépendantes du contexte observé à l'instant considéré. Cette approche est guidée par le *rationalisme* – intuition du type de distribution de valeurs observé dans un contexte donné – ou l'*empirisme* – connaissances académiques relatives à ces distributions. Ces distributions sont estimées à partir d'*analyses mathématiques et statistiques* sur des ensembles appropriés de données expérimentales, et peuvent elles aussi dépendre d'un ensemble d'autres paramètres.

La définition de ces modèles doit ainsi prendre en compte les dépendances identifiées entre les différents paramètres considérés. Les valeurs de certains paramètres, et ainsi les paramètres des automates à états finis ou distributions qui modélisent leurs variations, peuvent en effet dépendre d'un ensemble d'autres paramètres. Le niveau de détail nécessaire pour la simulation détermine la précision requise dans la définition des modèles et leur validation.

3.3.3 Principe de validation opérationnelle

Pour ce qui concerne la validation des séquences de données produites par la simulation – *validation opérationnelle*, la question fondamentale pour le choix d'une technique de validation est de

savoir si le système réel correspondant est observable ou pas. S'il est possible de disposer de données expérimentales enregistrées à partir du système réel, les techniques de validation consistent en la comparaison des comportements entre d'une part les entrées et les sorties du système réel, et d'autre part celles du processus de simulation. La comparaison peut se faire par exemple au moyen de graphes (comparaison subjective) ou de tests statistiques (comparaison objective). Si par contre le système réel n'est pas observable, les données produites par la simulation sont validées subjectivement par des experts, ou bien par comparaison aux résultats produits par d'autres processus de simulation.

3.4 Synthèse

Finalement, la construction d'un processus de simulation correspond à une *démarche incrémentale* et s'inscrit dans un double cycle de raffinement pour atteindre progressivement les objectifs de résolution du problème posé. La nécessité d'une *approche hybride* implique que plusieurs types de données et connaissances disponibles et appropriées au contexte et objectifs de décision sont intégrés aux différentes étapes de résolution. On prend également en compte différents modèles de simulation selon le type de paramètre : *qualitatif* (automates à états finis) ou *quantitatif* (distributions).
Cette méthodologie pour la simulation est synthétisée sur la figure 3.3.

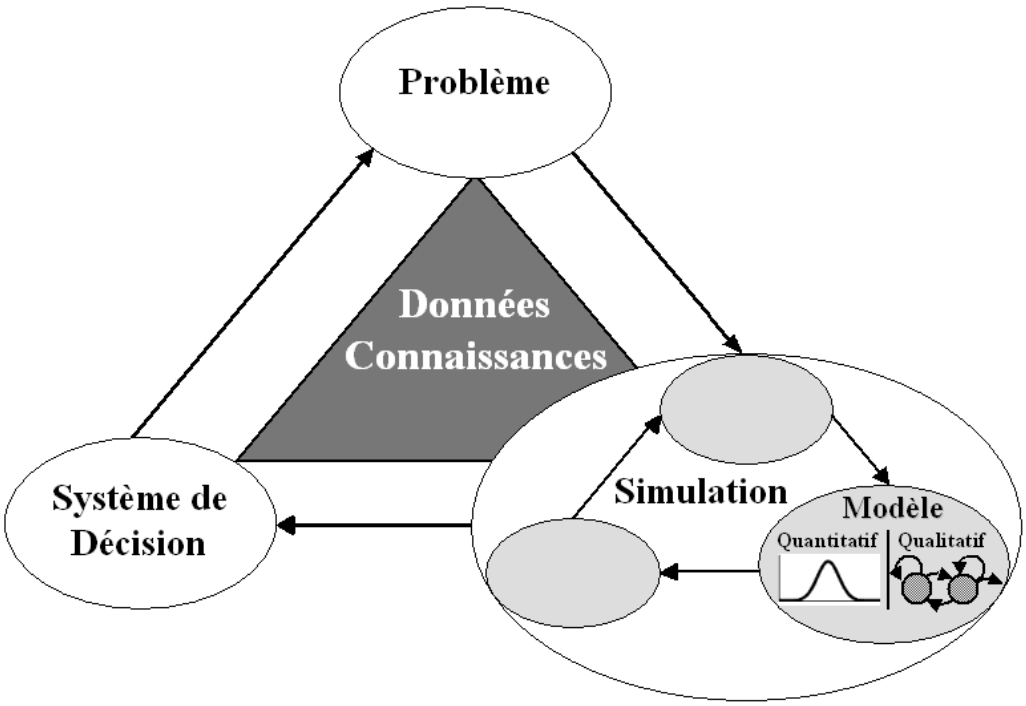


FIG. 3.3 – Synthèse de la méthodologie proposée pour la simulation.

4

Contexte de la télésurveillance médicale à domicile

L'objectif de ce chapitre est de présenter le contexte de la télésurveillance médicale à domicile en terme (1) des paramètres observables à domicile, (2) du niveau de détail et de connaissances requis pour la simulation des valeurs de ces paramètres, et (3) des connaissances et données utiles à la construction du processus de simulation.

4.1 Paramètres observables

Les paramètres observés doivent permettre de répondre au problème de l'étude des habitudes de vie d'une personne à domicile, dans l'objectif de détecter toute évolution critique de l'état de santé de la personne comme une déviation de son profil comportemental. Comme on ne dispose d'aucune connaissance *a priori* des habitudes de vie, des activités de la vie quotidienne, des actions réalisées, etc. (voir Fig. I.3.1), on s'intéresse aux **paramètres observés au niveau bas de l'échelle de décision** – les paramètres enregistrés en temps réel et en continu à domicile par un ensemble de capteurs. Les paramètres cliniques sont quant à eux observés à bien plus basse fréquence et ne varient pas significativement à l'échelle d'une journée. On pourrait étudier leur simulation et leur intégration à un plus haut niveau de décision, en conjonction avec un ensemble d'informations de même niveau extraites des données brutes des capteurs. Ces paramètres ne sont cependant pas pris en compte dans une première étude pour la mise en place d'un processus de simulation.

La définition de l'ensemble des paramètres dont on souhaite simuler l'évolution dans les conditions habituelles de vie d'une personne doit ainsi satisfaire les contraintes suivantes :

(1) Les valeurs des paramètres doivent pouvoir être obtenues simplement à partir des données enregistrées par un ou plusieurs capteurs installés dans l'habitat ou portés par la personne.

(2) L'ensemble des paramètres disponibles est limité par le choix des capteurs, qui est contraint par des critères éthiques, sociaux et individuels : respect de la vie privée, discrétion, facilité d'utilisation, faible coût.

(3) Les paramètres considérés doivent être suffisamment représentatifs des activités réalisées par la personne et sensibles à une détérioration de son état de santé pour être pleinement en adéquation avec les objectifs de décision.

Ce dernier point réfère à la considération de la simulation comme une étape du cycle de résolution du problème de construction du profil comportemental d'une personne. Les séquences de données doivent en effet être simulées dans l'objectif d'être utilisées en entrée du système de décision correspondant. Dans notre contexte, le choix des paramètres à simuler est donc primordial pour qu'on puisse effectivement extraire des informations sur l'évolution de l'état de santé de la personne télésurveillée. L'annexe B présente les scénarios de modifications observées de la situation et du comportement d'une personne âgée lors de l'installation progressive de certaines pathologies très spécifiques, mais sensibles, à moyen terme. Ces scénarios nous ont été décrits par le Docteur Pierre Rumeau, praticien hospitalier en gériatrie au CHU La Grave-Casselardit à Toulouse. Ils mettent en évidence la relation étroite entre l'installation d'une pathologie particulière et les habitudes de la vie quotidienne, associée souvent à une modification des variations habituelles de certains paramètres physiologiques "simples" tels que les fréquences cardiaque et respiratoire.

La sélection des paramètres considérés pour la simulation n'est cependant pas critique dans un premier temps puisqu'on a par ailleurs comme objectif pour l'étape suivante de construire un système de décision générique, adapté au traitement de n'importe quel type de paramètres fournis en entrée du système. On pourra ainsi facilement modifier l'ensemble des paramètres considérés s'il s'avère que le choix des paramètres n'est pas optimal pour répondre aux objectifs de décision. L'idée fondamentale de ce premier cycle de résolution est ainsi de constituer un ensemble de paramètres pertinent et représentatif de l'évolution de l'état de santé d'une personne, mais pas de définir dès ce niveau d'étude le meilleur ensemble exhaustif de paramètres à utiliser dans ce contexte de télésurveillance médicale. Par ailleurs, afin d'expérimenter la généricité de la méthode de décision proposée dans la suite des travaux, on sélectionne un ensemble de paramètres hétérogènes, aussi bien quantitatifs que qualitatifs, à modalités ordonnées ou non.

On fait finalement reposer le choix des paramètres de la simulation sur deux hypothèses importantes dans le cadre de l'étude du comportement d'une personne et de son état de santé :

(1) Toute détérioration de l'état de santé d'une personne entraîne très certainement des modifications de son comportement habituel ;

(2) La fréquence cardiaque est une mesure physiologique importante, facilement observable, et représentative à la fois de l'activité d'une personne et de son état de santé [74].

À partir de ces hypothèses et des contraintes à considérer dans le choix des paramètres, on décide de s'intéresser à un ensemble des paramètres permettant de décrire partiellement **l'activité de la personne à domicile et son état cardiaque** :

1. **Déplacements de la personne.** Ce paramètre est qualitatif et prend ses valeurs dans l'ensemble des pièces de l'habitat, auxquelles peut s'ajouter la considération d'une pièce "extérieur" pour modéliser les sorties du domicile (non considérée dans le cas présent). La présence dans chacune des pièces est détectée par un ensemble de capteurs de mouvements infrarouges.
2. **Postures.** Ce paramètre est également qualitatif, mais ses modalités peuvent être ordonnées en fonction de l'effort requis par la posture. On considère trois classes de postures, ordonnées subjectivement selon une dépense énergétique associée croissante : "allongé", "assis", puis "debout". Les postures peuvent être inférées à partir des données d'un ensemble de capteurs portés par la personne (accéléromètres et/ ou magnétomètres). Najafi *et al.* [76] ont par exemple proposé un capteur porté sur le torse qui détecte ces trois postures principales d'une personne.

3. **Niveau d'activité moyen.** C'est un paramètre quantitatif dont la valeur est définie comme la norme de l'accélération dans la direction de l'axe antérieur-postérieur. Il a été montré que ce paramètre fournit en moyenne une bonne approximation du niveau d'activité de la personne [26], et des dépenses énergétiques associées à l'activité réalisée. Les valeurs correspondantes sont obtenues par l'intermédiaire des données enregistrées par un accéléromètre porté sur le haut du torse, moyennées toutes les minutes.

4. **Fréquence cardiaque moyenne.** C'est un paramètre quantitatif dont la valeur est définie comme la moyenne de la fréquence cardiaque instantanée sur des périodes d'une minute. Les valeurs peuvent être calculées à partir des données enregistrées par un dispositif portatif d'enregistrement de l'électrocardiogramme (ECG).

4.2 Niveau de simulation

La définition du niveau de simulation est fondamentale et concerne les **niveaux de détail nécessaire et de connaissances disponibles**, en fonction des objectifs et de la complexité du problème étudié. Dans le cadre de la télésurveillance médicale à domicile, la complexité réside particulièrement dans l'**hétérogénéité** et les **variations conjointes** des paramètres à simuler. Le processus de simulation doit en effet générer les variations dans le temps d'un ensemble de paramètres tous représentatifs de la situation de la personne à domicile et de son état de santé, donc interdépendants. Les valeurs de ces paramètres, obtenues expérimentalement par des capteurs, peuvent également correspondre à des paramètres aussi bien quantitatifs (paramètres physiologiques par exemple) que qualitatifs (observation des déplacements de la personne dans son habitat par exemple).

Le **niveau de détail nécessaire** à la simulation doit permettre de répondre aux objectifs de la simulation et aux performances exigées de la décision. Dans notre contexte d'étude, on s'intéresse au comportement d'une personne à domicile dans l'objectif de détecter ensuite d'éventuelles modifications des habitudes comportementales qui peuvent être représentatives d'une évolution critique de son état de santé. Cette problématique a été présentée en introduction comme une problématique "haut niveau", dont l'objectif est l'observation de tendances à long terme (voir paragraphe I.3.3). Les séquences de données simulées ne nécessitent donc pas un grand niveau de détail et de précision, tout en restant réalistes. Ce sont les variations conjointes des différents paramètres observés, mettant en jeu des phénomènes et relations complexes, qu'il est fondamental de préserver dans le processus de simulation, et donc d'intégrer dans le modèle. Les étapes de validation du processus de simulation doivent en particulier être adaptées à ces objectifs.

Les **connaissances disponibles** et utiles à la conception du modèle de simulation doivent également être adaptées au niveau de détail nécessaire. Une difficulté réside dans le manque de connaissances *a priori* sur le fonctionnement et les observations enregistrées par les systèmes de télésurveillance médicale, puisque les dispositifs expérimentaux permettant d'observer conjointement les évolutions de ces paramètres ne sont pas encore complètement opérationnels. La construction du processus de simulation s'appuie ainsi sur une diversité des sources d'information et de connaissances relatives aux paramètres étudiés et à leurs variations conjointes, c'est-à-dire à l'activité d'une personne à domicile d'une part, à son état cardiaque d'autre part, et enfin aux variations relatives de ces caractères. On intègre dans la construction du modèle de simulation des connaissances *a priori* – **connaissances de sens commun** et **connaissances académiques** – et des **connaissances extraites** de données expérimentales.

4.3 Connaissances et données utiles à la simulation

4.3.1 Connaissances *a priori*

Les connaissances *a priori* correspondent à des connaissances bien reconnues et validées, c'est-à-dire d'une part aux connaissances de sens commun, et d'autre part aux connaissances académiques. Dans notre contexte elles concernent :

(1) l'activité d'une personne à domicile, caractérisée par les déplacements observés, les postures et le niveau d'activité moyen ;

(2) son état cardiaque, caractérisé par la fréquence cardiaque moyenne ;

(3) les variations relatives de ces caractères.

Activités de la vie quotidienne

Le déroulement des activités quotidiennes d'une personne à domicile est principalement décrit par les connaissances de sens commun d'experts dans les domaines liés à l'observation des habitudes de vie d'une personne (gériatrie, ergothérapie, etc.). Deux grandes catégories d'activités sont communément définies :

- **Activités dites "de base",** tournées vers la personne, et qui correspondent à ce qu'on appelle les AVQ (Activités de la Vie Quotidienne) ou *ADL* en anglais *(Activities of Daily Living)* : (1) dormir (bien que cette activité ne soit finalement jamais citée dans les articles sur le sujet), (2) faire sa toilette, (3) s'alimenter, (4) aller au cabinet. Ces activités sont présentes quotidiennement dans le rythme de vie d'une personne.
- **Activités plus complexes**, communément appelées IAVQ (Activités Instrumentales de la Vie Quotidienne) ou *IADL* en anglais *(Instrumental Activities of Daily Living)*, bien plus tournées vers la communauté. On en dénombre habituellement huit [64] : (1) préparer un repas, (2) faire les courses, (3) faire le ménage, (4) s'occuper de la lessive, (5) utiliser le téléphone, (6) utiliser les transports, (7) prendre ses médicaments, (8) gérer son budget. Ces activités sont présentes occasionnellement dans la vie quotidienne.

L'annexe A détaille par exemple la journée type d'une dame âgée fragile de 87 ans, veuve, proposée par le Docteur Pierre Rumeau. Cette description met bien en évidence la présence quotidienne des activités de base de la vie quotidienne, associée à l'observation de la réalisation habituelle d'activités plus complexes. Il existe ainsi un **rythme** que l'on peut observer **dans les activités quotidiennes** d'une personne à domicile, caractérisé par un ensemble d'activités régulières. On identifie alors deux types d'activités régulières :

- D'une part, les **activités de base** (AVQ) sont observées quotidiennement : le sommeil, la toilette, l'alimentation et les passages aux toilettes.
- D'autre part, des **activités plus complexes** peuvent également être habituelles pour une personne dans sa vie quotidienne. Elles correspondent soit à des IAVQ – par exemple, "faire le ménage pendant environ deux heures tous les vendredi matin", soit à tout autre activité que la personne à l'habitude d'effectuer – par exemple, "lire pendant environ une heure au salon tous les jours en fin d'après-midi".

Les activités régulières sont caractérisées par plusieurs paramètres qui concernent (1) d'une part leurs caractéristiques intrinsèques – fréquence, moment et durée – et (2) d'autre part leurs caractéristiques extrinsèques – ordre de réalisation.

Caractéristiques intrinsèques

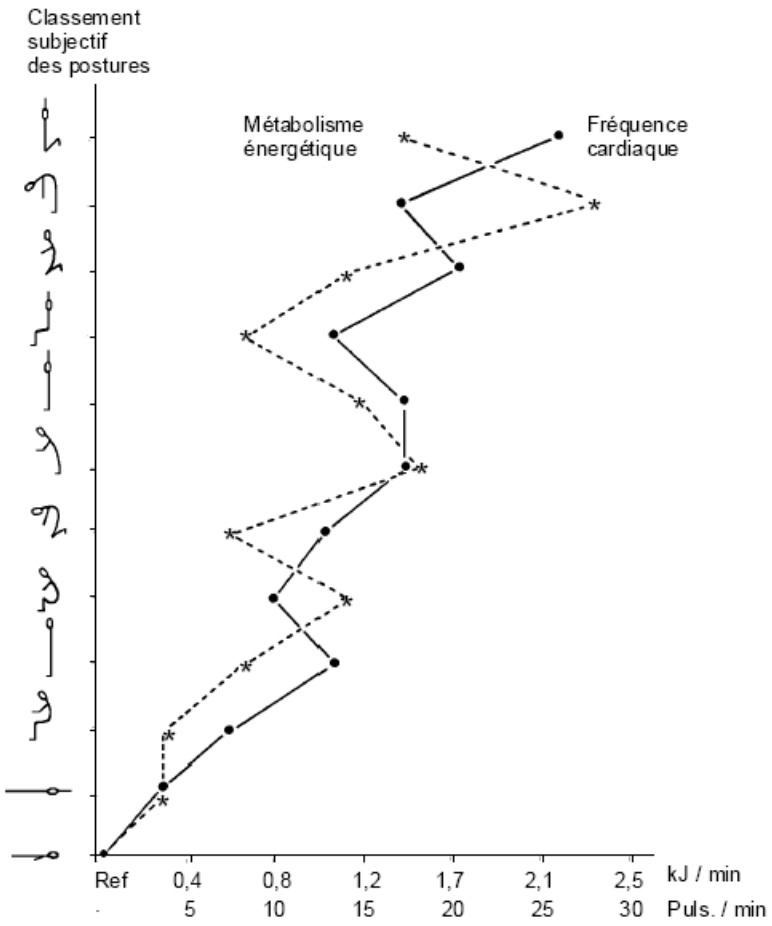


FIG. 4.1 – Coût énergétique et accroissement de la fréquence cardiaque de repos pour diverses postures classées subjectivement par ordre de pénibilité. [31, 74]

À partir des connaissances d'experts et d'une synthèse sur les niveaux d'énergie requis pour chaque tâche de la vie quotidienne réalisée dans [84], on décrit la réalisation des activités de base de la vie quotidienne en terme des caractéristiques intrinsèques d'une activité (fréquence, moment et durée de réalisation) et des valeurs des paramètres qui nous intéressent particulièrement pour décrire une activité : lieu de l'activité, posture requise et dépenses énergétiques associées. Les résultats de cette synthèse sont présentés dans le tableau 4.1.

Caractéristiques extrinsèques

Qu'elles soient habituelles ou non, les activités réalisées par une personne à domicile présentent une certaine cohérence dans leur réalisation et leur enchaînement. Par conséquent, les paramètres directement représentatifs de l'activité d'une personne sont liés dans leurs variations et contraints au cours du temps. Ces contraintes dans la réalisation des activités sont issues de connaissances de sens commun, telles que :

- "Une personne ne passe pas toute une après-midi dans la salle de bain ou les toilettes."
- "La posture debout n'est jamais maintenue sur une longue période, sauf lors de certaines activités telles que le repassage ou le ménage, pour lesquelles rester debout pendant une heure par exemple n'a rien d'étonnant."

Activités de la Vie Quotidienne	
SOMMEIL	
Fréquence	1 à 2 fois par jour.
Moment	La nuit, et éventuellement une autre fois pour une sieste plus certainement l'après-midi.
Durée	Plusieurs heures pendant la nuit, mais plutôt 1 à 2 heures pendant la sieste.
Lieu	Dans la chambre, éventuellement dans le salon pour la sieste.
Posture	Allongée, avec des passages en postures assise et debout pendant le coucher et le lever.
Dépenses énergétiques	Dépenses énergétiques faibles pendant le sommeil, et plus élevées au coucher et au lever.
TOILETTE	
Fréquence	1 fois par jour.
Moment	Plutôt le matin.
Durée	Probablement moins d'une heure.
Lieu	Dans la salle de bain.
Posture	Debout essentiellement.
Dépenses énergétiques	Activité relativement coûteuse en énergie.
ALIMENTATION	
Fréquence	3 fois par jour.
Moment	Le matin, le midi et le soir, dans la cuisine.
Durée	L'alimentation peut durer d'un quart d'heure à une heure environ, voire plus pour de longues préparations. La préparation des repas concerne principalement les repas du midi et du soir, mais il est possible d'observer qu'une personne prépare en une fois plusieurs repas qui seront alors ensuite pris plus rapidement.
Lieu	Dans la cuisine, éventuellement dans le salon pour l'alimentation en elle-même.
Posture	Alternances fréquentes entre les postures assise et debout.
Dépenses énergétiques	L'alimentation en elle-même est peu coûteuse en énergie, mais par contre toutes les autres tâches liées à l'alimentation, et notamment la préparation des repas, le sont beaucoup plus.
PASSAGE AU CABINET	
Fréquence	Au moins une fois par jour.
Moment	–
Durée	Durée relativement courte.
Lieu	Dans les toilettes.
Posture	Assis.
Dépenses énergétiques	Activité plutôt coûteuse en énergie.

Tab. 4.1 – Caractéristiques des activités de la vie quotidienne (AVQ).

– "Une personne âgée particulièrement ne réalise pas successivement plusieurs activités longues et coûteuses en énergie."
– etc.

Il existe notamment un lien étroit et reconnu entre la posture de la personne et les dépenses énergétiques associées [31, 74] : les dépenses énergétiques augmentent avec l'effort requis par la posture (voir Fig. 4.1).

Finalement, par rapport aux paramètres que l'on souhaite simuler pour décrire l'activité d'une personne à domicile (déplacements, postures et niveau d'activité), les connaissances disponibles sont les suivantes :

(1) On a identifié les caractéristiques des activités de base de la vie quotidienne, présentes tous les jours : répartition au cours d'une journée (fréquence, moment, durée) et déplacements, postures et dépenses énergétiques associés ;

(2) On sait la présence possible d'autres activités régulières d'une journée ou d'une semaine à l'autre par exemple ;

(3) Plus généralement, on a la connaissance ou l'intuition d'un ensemble de critères définissant des relations cohérentes entre les différents paramètres observés quelle que soit l'activité réalisée.

Fréquence cardiaque

On dispose de connaissances académiques générales sur les variations de la fréquence cardiaque d'une personne, extraites de [74]. La fréquence cardiaque varie en permanence autour d'une valeur moyenne, et on définit ainsi deux grandeurs :

- **La fréquence cardiaque instantanée** – Elle est obtenue à partir de l'intervalle de temps séparant deux battements consécutifs, et ramenée à la minute. Chez le sujet au repos, les variations peuvent atteindre $\pm 20\%$ par rapport à la moyenne.
- **La fréquence cardiaque moyenne** – Elle est obtenue en comptant les battements cardiaques pendant un temps suffisamment long (30 secondes ou au mieux 1 minute), en raison des variations cycliques et aléatoires de la fréquence instantanée. Les variations rapides de l'activité cardiaque présentent peu d'intérêt pour des études de longue durée. Les fluctuations légères de la fréquence cardiaque moyenne, observées au cours des minutes successives d'un état stable, témoignent d'un effort permanent de régulation.

Dans notre contexte d'étude à long terme de l'évolution de la situation d'une personne à domicile, on s'intéresse ainsi aux variations de la **fréquence cardiaque moyenne**. Elle est habituellement calculée à partir des données d'un ECG moyennées toutes les 30 secondes à 1 minute.

Chez l'homme adulte sain au repos couché, la fréquence cardiaque est d'environ 65 battements par minute (bpm) en moyenne. Au repos, elle varie cependant suivant les conditions de repos et d'observation, et avec les individus. Il existe plusieurs causes de variation affectant les stabilité de la **fréquence cardiaque au repos** :

- **Variations inter-individuelles :** la fréquence cardiaque moyenne au repos varie largement en fonction des individus et avec l'âge du sujet, l'entraînement sportif, le nycthémère (rythme de variations approximativement circadien).
- **Variations intra-individuelles :** pour un sujet donné, la fréquence cardiaque de repos subit des variations en fonction de l'activité végétative (légère augmentation après les

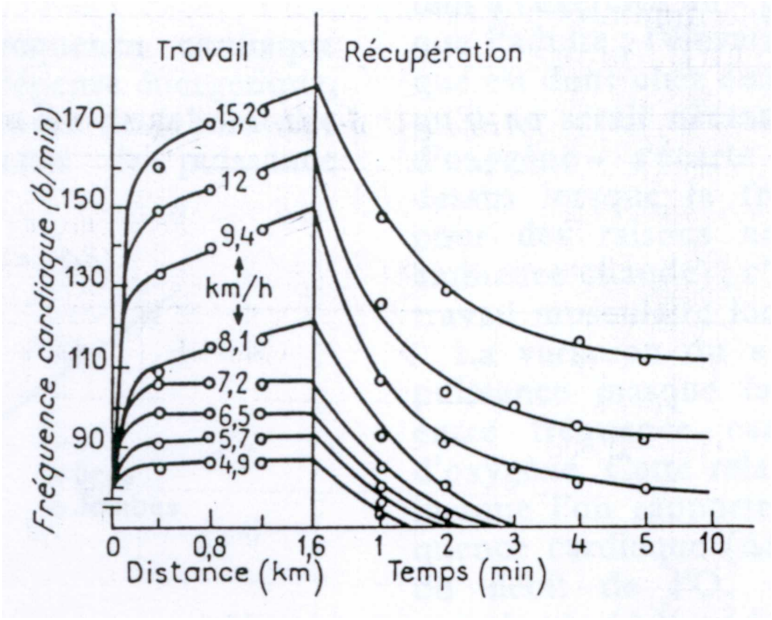


FIG. 4.2 – Variation de la fréquence cardiaque pendant et après une marche ou course de 1600 mètres, à différentes vitesses. [74]

repas), de la posture (augmentation avec l'effort requis par la posture), de l'ambiance thermique (accroissement comme une fonction linéaire de la température ambiante), etc.

Aux variations de repos de la fréquence cardiaque viennent se superposer d'importantes **variations dues au travail musculaire**, mais aussi au niveau de vigilance, à l'importance de la contrainte mentale, etc. Il existe également une variation permanente du rythme cardiaque – l'*arythmie sinusale* – qui est maximale au repos. Les travaux musculaires s'accompagnent d'une suppression de cette arythmie et d'une accélération de la fréquence cardiaque. Au cours d'un effort mental, on observe également la suppression de l'arythmie sinusale, mais sans augmentation de la fréquence cardiaque.

Relation entre l'activité et la fréquence cardiaque

La fréquence cardiaque est un témoin global du niveau d'activité, reflétant non seulement le coût énergétique de la contraction musculaire, mais également les réactions de l'organisme aux variations du contexte physique, aux émotions, et aux agressions de diverses natures. Les activités réalisées ont notamment une forte influence sur les valeurs de la fréquence cardiaque moyenne, en particulier dues aux différents **postures** requises et à l'**intensité du travail musculaire** nécessaire.

Influence de la posture

Monod et Pottier [74] précisent l'influence de la posture sur la fréquence cardiaque au repos :

(a) En position assise, la fréquence cardiaque moyenne est environ 10% supérieure à celle constatée en position allongée.

(b) En position debout, elle est de 20 à 30% supérieure à celle enregistrée en position allongée.

La figure 4.1 présente la relation entre la posture et la fréquence cardiaque d'une part, et les dépenses énergétiques d'autre part. Le non parallélisme des deux courbes s'explique par une prise en compte différente des caractéristiques des postures par chacun des indicateurs.

Influence du travail musculaire
Les effets d'un travail musculaire d'intensité faible à modérée sur les valeurs de la fréquence cardiaque moyenne sont décrits de la façon suivante (voir Fig. 4.2, jusqu'à une vitesse de 7.2 Km/h) :

(1) Augmentation linéaire des valeurs de la fréquence cardiaque moyenne avec la puissance du travail ;

(2) Rapide stabilisation des valeurs jusqu'à un plateau caractérisant l'état constant ;

(3) Nécessité de prendre en compte de 1 à 3 minutes de récupération pour revenir à la fréquence cardiaque de repos après une activité d'intensité modérée.

Au-dessus d'une certaine puissance de travail (voir Fig. 4.2, pour les vitesses supérieures ou égales à 8.1 km/h), la fréquence cardiaque ne se maintient plus en plateau et continue à augmenter jusqu'à observer un effet de saturation aux valeurs maximales, spécifiques de chaque individu. La période de récupération s'accroît considérablement.
Finalement, le tableau 4.2 présente une classification des travaux physiques d'après la fréquence cardiaque.

Niveau d'activité	**Fréquence cardiaque (bpm)**
Très léger	< 75
Léger	75 à 100
Modéré	100 à 125
Dur	125 à 150
Très dur	150 à 175
Extrêmement dur	> 175

Tab. 4.2 – Classification des travaux physiques d'après la fréquence cardiaque. [30, 74]

4.3.2 Données expérimentales

Dans le contexte de la télésurveillance médicale à domicile, on ne dispose pas pour le moment d'enregistrements de capteurs installés dans un environnement réaliste de télésurveillance. Cependant, des données expérimentales correspondant à la surveillance de 8 sujets dans leur vie quotidienne ont été collectées pendant deux périodes non consécutives de 24 heures – ce qui donne 16 séquences d'enregistrements sur 24 heures. Les sujets de l'expérimentation sont des hommes et des femmes, entre 20 et 30 ans, et en bonne santé. Ces enregistrements ont été réalisés par l'équipe de Sylvie Charbonnier, du Laboratoire d'Automatique de Grenoble (LAG). Les données enregistrées sont les suivantes (voir Fig. 4.3) :

- **Date et heure.**
- **Annotation des activités réalisées** (toutes les 15 minutes), parmi un ensemble de 14 activités exigeant des niveaux d'énergie différents : dormir, être allongé, être assis calmement, être assis et discuter, être assis et travailler, manger, être debout, être debout et travailler, marcher doucement, marcher vite, faire de la bicyclette, courir, monter et descendre les escaliers.

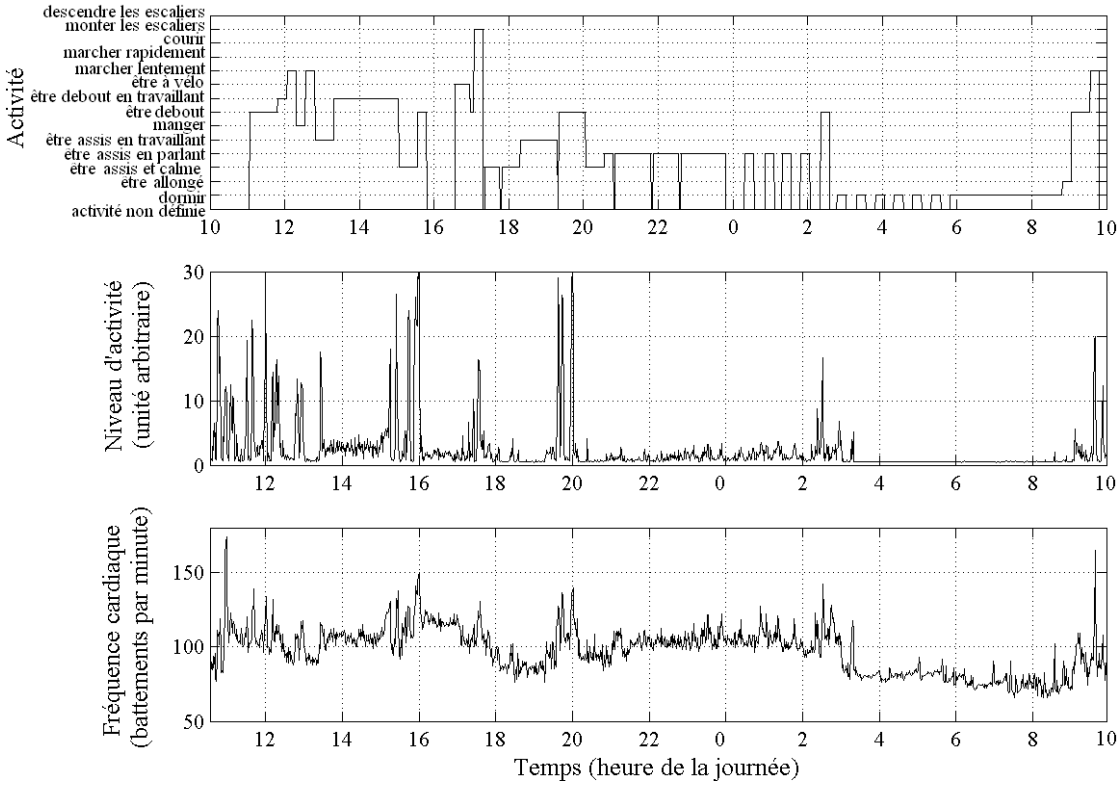


FIG. 4.3 – Données enregistrées d'une personne dans sa vie quotidienne.
Le niveau d'activité et la fréquence cardiaque sont moyennés toutes les minutes.

- **Niveau d'activité**, défini sur une échelle arbitraire, et correspondant à la moyenne toutes les minutes de l'accélération enregistrée suivant l'axe antérieur-postérieur.
- **Fréquence cardiaque moyenne**, calculée toutes les minutes à partir des données enregistrées par un appareil portatif de mesure de l'ECG.

Il est à noter cependant que 2 des 8 sujets n'ont pas annoté leurs activités – sujets 3 et 6 – et les données expérimentales correspondantes ne sont donc pas complètement exploitables. Les données enregistrées pour l'ensemble des sujets de l'expérimentation sont présentées en annexe D. L'*a priori* de la spécificité du comportement de chaque sujet s'y trouve en particulier confirmé. Les enregistrements de niveau d'activité par exemple sont très variés – certains ne dépassent qu'à peine une valeur de 10, alors que d'autres atteignent régulièrement 20 ou 30. Les variations de fréquence cardiaque sont également diverses, et on remarque par ailleurs un alignement fréquent de leurs pics de valeurs sur ceux du niveau d'activité.

Pour se rapprocher du cadre de notre étude, on restreint l'ensemble des données expérimentales disponibles afin de ne considérer que des enregistrements significatifs dans le contexte de la télésurveillance médicale à domicile. Cette application concerne en effet principalement des personnes âgées ou souffrant d'affections motrices ou cognitives. On ne considère ainsi que les données correspondant à des activités d'intensité faible à modérée, et en supposant que la personne habite dans une maison de plein pied, ce qui correspond aux activités suivantes : dormir, être allongé, être assis calmement, être assis et discuter, être assis et travailler, manger, être debout, être debout et travailler et marcher doucement. Par ailleurs, on ne prend pas non plus en compte les données enregistrées pendant les 4 minutes suivant la réalisation d'une activité

exigeant un important travail musculaire afin de s'affranchir d'une grande partie des effets sur la fréquence cardiaque moyenne du temps de récupération nécessaire pour stabiliser les valeurs autour des variations au repos.

Les données expérimentales sont ensuite séparées en deux sous-ensembles utilisés respectivement pour la construction et la validation du modèle conceptuel – données de modélisation (sujets 1 à 4) – et pour la validation opérationnelle – données de validation (sujets 5 à 8). Bien qu'on dispose de deux enregistrements pour chaque sujet, il est préférable de constituer les deux sous-ensembles de modélisation et de validation de façon à ce que les données relatives à chaque sujet ne figurent pas à la fois pour une part dans l'ensemble de modélisation, et pour l'autre part dans celui de validation. L'ensemble de validation des résultats de la simulation est ainsi complètement "nouveau" au regard de celui qui a été impliqué dans la construction du modèle.

4.4 Synthèse

Finalement, pour un individu donné, on décide de simuler à "*haut niveau*" les variations de quatre paramètres qui peuvent être enregistrés à domicile : les *déplacements*, les *postures*, le *niveau d'activité* et la *fréquence cardiaque*. Dans ce contexte, les connaissances utiles à la simulation sont constituées d'une part de *connaissances a priori* – connaissances de sens commun et académiques – et d'autre part de connaissances que l'on peut extraire de *données expérimentales*. Afin de répondre au souci de préserver la complexité du système, et en particulier les variations conjointes des différents paramètres, on s'intéresse bien sûr aux connaissances sur leurs caractéristiques intrinsèques, mais aussi à leurs influences relatives.
Les connaissances et données disponibles et utiles à la construction du processus de simulation sont synthétisées sur la figure 4.4.

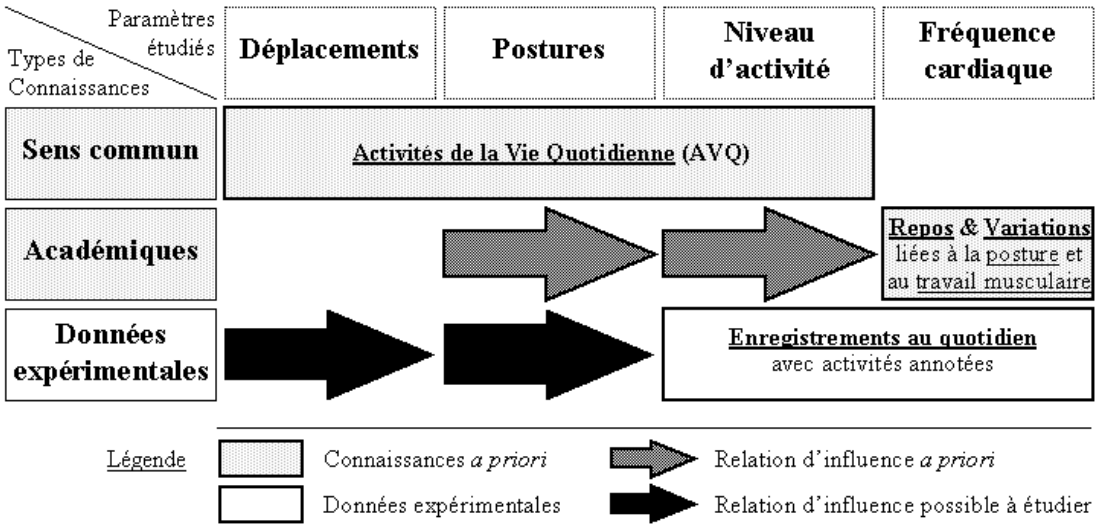


FIG. 4.4 – Synthèse de la méthodologie proposée pour la simulation.

5

Modélisation

Dans le contexte de décision visant à la construction d'un profil comportemental d'une personne à domicile, l'objectif de la simulation est de générer des séquences de données représentatives des conditions habituelles de vie d'une personne télésurveillée. Dans l'objectif de décision, on recherche non pas la simulation d'un "individu moyen", mais la construction d'un modèle qui permet, en agissant sur paramètres, de générer des données représentatives de plusieurs profils de comportement. Plus tard, l'étude de la détection d'éventuelles tendances critiques dans l'évolution des paramètres observés nécessite la simulation de perturbations dans ces séquences.

L'étape de modélisation comprend à la fois la construction du modèle conceptuel utilisé pour la simulation et sa validation (voir Fig. 3.1).

5.1 Principe de construction du modèle

La construction du modèle conceptuel pour la simulation est guidée par les connaissances *a priori* dont on dispose sur les dépendances entre les paramètres considérés. Il est en effet fondamental de préserver la complexité du système simulé, c'est-à-dire en particulier dans notre contexte les variations conjointes des paramètres. Une grande précision dans leurs valeurs n'est pas forcément nécessaire, et même impossible à obtenir puisque les connaissances que l'on a des paramètres à simuler démontrent qu'il existe une grande variabilité inter-individuelle dans les valeurs enregistrées sur le comportement et l'évolution de l'état cardiaque d'une personne dans ses activités de la vie quotidienne. Dans le contexte d'observation d'une personne à domicile au travers de ses déplacements et postures, des variations du niveau d'activité et de la fréquence cardiaque, on sait *a priori* que :

(1) Les déplacements dans l'habitat suivent le rythme des activités quotidiennes ;

(2) Les postures successives dépendent de l'activité réalisée, et par conséquent de la pièce occupée et du moment de la journée ;

(3) Le niveau d'activité observé est également fonction de l'activité réalisée, et donc des déplacements observés et des postures ;

(4) Les principales variations de la fréquence cardiaque sont liées à la posture et au niveau d'activité.

Le modèle de simulation est ainsi défini par une ***structure en cascade*** (voir Fig. 5.1), impliquant quatre sous-modèles de simulation. L'expérimentation consiste alors en quatre étapes successives de simulation, chacune correspondant à un sous-modèle, permettant de générer successivement les données correspondant aux paramètres suivants :

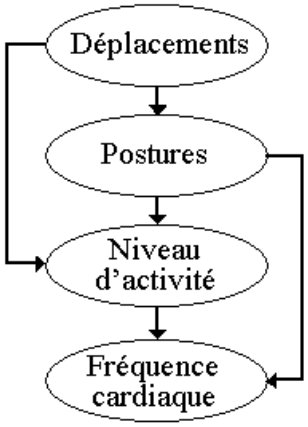


FIG. 5.1 – Structure en cascade pour la simulation.

1. **Déplacements** de la personne pendant l'intervalle de temps considéré pour la simulation.
2. **Postures** successives cohérentes avec les déplacements observés, en fonction du moment de la journée.
3. **Valeurs cohérentes de niveau d'activité**, en fonction des déplacements et des postures observées, et du moment de la journée,
4. **Valeurs de fréquence cardiaque** réalistes, en fonction de l'activité réalisée, c'est-à-dire principalement en fonction de la posture et du niveau d'activité observés.

Chaque sous-modèle utilisé pour la génération des valeurs successives d'un paramètre prend ainsi en entrée les résultats de simulation correspondant à un ou plusieurs des sous-modèles précédents dans la structure en cascade.

La conception des sous-modèles de simulation est ainsi guidée par les **connaissances *a priori*** sur les variations du paramètre correspondant et ses relations avec les paramètres précédemment simulés. Si des données expérimentales sont disponibles, des analyses mathématiques et statistiques sont réalisées en fonction de ces *a priori* afin de valider et de quantifier les concepts sous-jacent aux sous-modèles. On génère ainsi un ensemble de **connaissances extraites**, quantitatives, à intégrer dans le modèle de simulation. Ces connaissances extraites d'analyses statistiques donnent une idée de valeurs pertinentes pour le réglage des paramètres du modèle : elles correspondent à une distribution de valeurs réalistes pour ces paramètres plutôt qu'à des valeurs précises, définies avec un intervalle de confiance limité. Il existe en effet une grande variabilité dans les caractéristiques des paramètres observés en fonction du sujet considéré, et donc dans les paramètres des modèles qui génèrent les séquences de valeurs observées. Les connaissances extraites interviennent donc comme des paramètres du modèle de simulation, permettant de produire des séquences de données représentatives de plusieurs profils de personnes.

Les techniques de modélisation – conception et validation du modèle – sont spécifiques à chaque sous-modèle, et définies en fonction du type de paramètre à simuler et des données et connaissances disponibles (voir paragraphe 3.3.2). En particulier, les paramètres qualitatifs sont simulés à partir d'automates à états finis, et les paramètres quantitatifs à l'aide de distributions. Le paragraphe suivant détaille les sous-modèles utilisés pour chaque paramètre considéré dans la simulation.

5.2 Sous-modèles de simulation

Ce paragraphe présente les quatre sous-modèles associés à la simulation successive des déplacements, postures, niveaux d'activité, puis fréquence cardiaque. Pour chacun, en accord avec la méthode de "conception-validation" du modèle (voir Fig. 3.1), nous présentons d'abord le principe de modélisation, puis sa validation.

Selon notre schéma d'intégration des différents types de connaissance (voir Fig. 3.2), cette dernière démarche peut inclure l'analyse de données expérimentales pour la validation – par *analyses mathématiques et statistiques* – des concepts de modélisation issus du *rationalisme* ou de l'*empirisme*. Si par contre aucun ensemble expérimental approprié n'est disponible, cette validation ne peut être réalisée qu'*avec l'aide d'experts* du domaine.
Dans le cadre de ce travail, on a ainsi en particulier collaboré avec le Professeur Hélène Pigot, ergothérapeute et professeur en informatique au laboratoire DOMUS de l'Université de Sherbrooke (Québec, Canada) et le Docteur Pierre Rumeau, praticien hospitalier en gériatrie au CHU La Grave-Casselardit à Toulouse.

5.2.1 Déplacements

Principe de modélisation

Le principe de modélisation des déplacements d'une personne à domicile est fondé sur le *rationalisme*. Le modèle de simulation pour la génération des déplacements de la personne a été construit et implémenté par G. Virone [109]. Il repose sur un ensemble de réseaux de Pétri définis à partir des connaissances de sens commun sur les activités de la vie quotidienne d'une personne. Les réseaux de Pétri représentent ainsi les déplacements attendus d'une personne en fonction du moment de la journée. Une journée est divisée en sept périodes de temps : (1) nuit, (2) lever, (3) matinée, (4) déjeuner, (5) après-midi, (6) soirée, (7) coucher ; et l'habitat est supposé comprendre six pièces sur un seul étage : (1) cuisine, (2) séjour, (3) chambre, (4) salle de bain, (5) toilettes et (6) couloir. Par ailleurs, la présence dans une pièce n'est enregistrée que si la durée d'occupation de la pièce dépasse 10 secondes, ce qui a pour conséquence de n'observer en pratique quasiment jamais la présence dans le couloir.

Validation

Comme on ne dispose pas d'enregistrements expérimentaux représentatifs des déplacements, la validité de l'utilisation d'un réseau de Pétri pour modéliser les déplacements d'une personne dans ses activités de la vie quotidienne est confirmée par des personnes ayant une bonne connaissance du domaine appréhendé. Les paramètres du réseau sont déterminés *intuitivement avec l'aide d'experts*. Ils pourront être ajustés lors de la validation des séquences de déplacements générées par le modèle.

5.2.2 Postures

Principe de modélisation

Le principe de modélisation des postures successives d'une personne dans ses activités de la vie quotidienne est fondé sur le *rationalisme*. Les postures de la personne représentent une information *statique* sur son activité à domicile. Le modèle de génération des postures est défini à partir des connaissances de sens commun sur les postures successives d'une personne en fonction de ses activités à domicile. L'observation des déplacements au cours de la journée donne par ailleurs

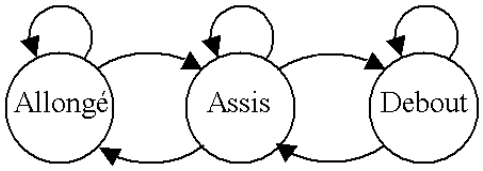


FIG. 5.2 – Automate à états finis pour la simulation des postures.

l'intuition de ces activités. Le modèle proposé est ainsi fondé sur un automate à états finis fonction de la sortie du sous-modèle de simulation précédent – les déplacements au cours du temps – et comprenant trois états représentatifs des trois classes de postures définies intuitivement pour une personne : (1) debout, (2) assise et (3) allongée. Dans le contexte de la télésurveillance médicale à domicile, on suppose par ailleurs que les postures "allongé" et "debout" ne peuvent être atteintes que par passage en posture "assis" (voir Fig. 5.2).

Validation

Comme on ne dispose pas d'enregistrements expérimentaux représentatifs des postures, le modèle proposé est validé par les connaissances de sens commun, et *avec l'aide d'experts*. Les probabilités de transition de l'automate sont définies intuitivement et dépendent de la pièce occupée (déplacements) et du moment de la journée – une des sept périodes définies pour la génération des déplacements. Ces paramètres permettent en effet d'avoir une idée de l'activité réalisée par la personne, et ainsi d'ajuster les postures générées en conséquence. Par exemple si la personne est dans la cuisine vers midi, on peut supposer qu'elle prépare le repas ou est en train de manger, ce qui entraîne une forte alternance entre les postures "debout" et "assis" (voir Tab. 4.1). On définit par ailleurs en fonction de ces mêmes paramètres la moyenne et l'écart-type d'une distribution normale permettant de déterminer une durée de transition entre la génération de deux postures successives. Il est cependant possible que la nouvelle posture générée soit identique à la précédente.

Les probabilités de transition et les paramètres de la distribution normale associée aux durées entre deux transitions pourront être ajustés lors de la validation des séquences de postures générées par le modèle.

5.2.3 Niveau d'activité

Principe de modélisation

Le principe de modélisation des niveaux d'activité d'une personne dans ses activités de la vie quotidienne est fondé sur le *rationalisme*. Le niveau d'activité représente une information *dynamique* sur l'activité de la personne. Dans le contexte des activités réalisées à domicile par une personne télésurveillée, on suppose que le niveau d'activité – norme de l'accélération selon l'axe antérieur-postérieur du torse – représente assez précisément l'effort réalisé et donc les dépenses énergétiques de la personne [26].

Les connaissances de sens commun guident ainsi la construction du modèle (voir paragraphe 4.3.1) : "Il existe une distribution des niveaux d'activité caractéristique du type d'activité réalisée". En particulier, le type d'activité peut être supposé à partir de l'observation des déplacements et des postures de la personne. Puisque le niveau d'activité est une approche "dynamique" de

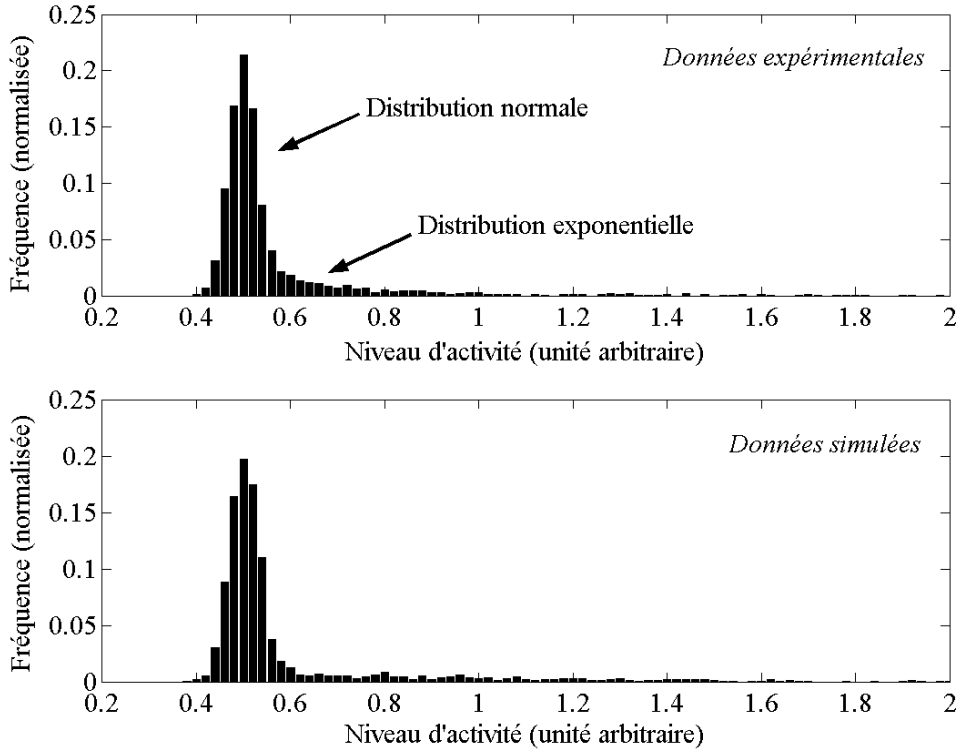


FIG. 5.3 – Distribution des valeurs de niveau d'activité pour une personne effectuant des mouvements en posture allongée.
Le graphe du haut représente la distribution des données expérimentales, et le graphe du bas celle de données simulées après estimation des paramètres des distributions normale et exponentielle. Ces deux distributions sont identiques d'après le test de Kolmogorov-Smirnov.

l'activité, on lui associe la notion de *mouvement*. On peut notamment définir le type de mouvement observé, selon quatre classes : les mouvements associés aux différentes postures (allongé, assis, debout), et même plus spécifiquement les mouvements associés à la posture "debout" qui correspondent à de la marche – la personne est par exemple forcément en train de marcher lorsqu'elle se déplace d'une pièce à une autre. On s'attend à ce que le niveau d'activité moyen observé augmente avec l'effort requis par la posture associée : les activités réalisées en posture allongée sont en effet intuitivement moins coûteuses en énergie que celles associées à la posture assise, puis à la posture debout.

Les niveaux moyens d'activité de la personne sont alors générés aléatoirement (toutes les minutes) en fonction d'une distribution caractéristique du type de mouvement observé au même moment. Afin d'obtenir une génération plus subtile de niveaux d'activité, prenant en compte le fait que différentes activités peuvent être réalisées lors de l'observation d'un même type de mouvement, le processus de simulation sélectionne aléatoirement une distribution de valeurs parmi un ensemble de distributions possibles compte tenu du moment de la journée, de la pièce occupée et de la posture.

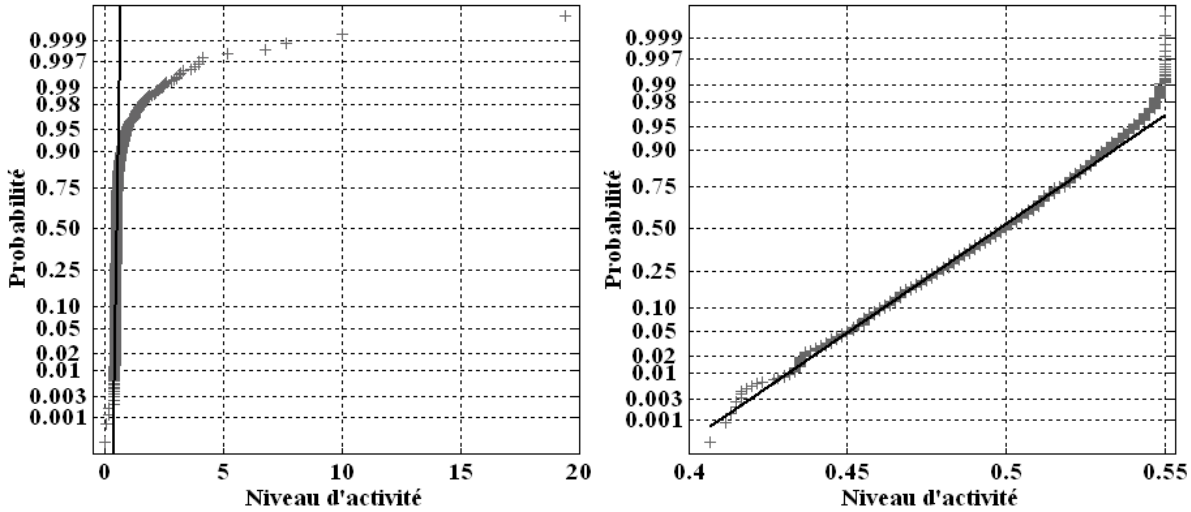


FIG. 5.4 – Test d'ajustement graphique à une loi normale de la distribution des valeurs de niveau d'activité enregistrées en posture allongée.
Le graphe de gauche représente les résultats du test sur l'ensemble des valeurs de expérimentales, mettant en évidence que seule la partie correspondant aux faibles valeurs de niveau d'activité semble normale ; celui de droite a ainsi été réalisé uniquement sur les faibles niveaux d'activité, confirmant cette présomption de normalité.

Validation

La validation de cette approche est réalisée par des *analyses sur les données expérimentales dédiées à la modélisation*. On dispose de valeurs de niveau d'activité correspondant à plusieurs situations – dormir, être allongé, être assis calmement, être assis et discuter, être assis et travailler, manger, être debout, être debout et travailler et marcher doucement – et on peut ainsi observer les distributions de valeurs correspondant aux quatre types de mouvements sélectionnés – mouvements en posture allongée, assise, debout, et pendant la marche.

Cette étude confirme l'*a priori* d'un niveau moyen d'activité croissant avec l'effort requis par la posture associée (voir tableau 5.1). L'observation des niveaux d'activité associés à la marche – cas particulier d'un mouvement réalisé en posture debout – montre une moyenne supérieure à celle associée à chaque type de posture. Cependant, le mode et la médiane de la distribution sont légèrement en dessous des valeurs observées pour l'ensemble des mouvements réalisés en posture

Posture	**Niveau d'activité**		
	Moyenne	**Médiane**	**Mode**
Allongé	0.60	0.51	0.51
Assis	1.58	0.86	0.56
Debout	3.46	1.77	0.74
Marche	4.82	1.51	0.58

TAB. 5.1 – Moyenne, médiane et mode du niveau moyen d'activité en fonction de la posture associée.
Les niveaux d'activité sont observés sur une échelle arbitraire.

debout. Une interprétation possible de ces remarques est que beaucoup des niveaux d'activité les plus élevés sont observés en posture debout lorsqu'une personne marche – ce qui augmente la moyenne de la distribution correspondante. Néanmoins, la marche n'est pas associée fréquemment aux niveaux d'activité les plus élevés dans l'ensemble des activités réalisées en posture debout.

L'estimation des paramètres des distributions de niveau d'activité n'est cependant pas si évidente, malgré l'intuition de leur ressemblance avec une distribution bien connue, correspondant à une *loi gamma*. Des tests d'adéquation montrent qu'en fait les distributions expérimentales n'ont rien de commun avec une quelconque distribution gamma. En collaboration avec Olivier Gaudoin, professeur et statisticien au Laboratoire de Modélisation et de Calcul (LMC) à Grenoble, nous avons finalement découvert que les distributions observées pour chaque type d'activité correspondent plutôt bien à un mélange entre :

(1) **une distribution normale** pour les faibles valeurs de niveau d'activité, et

(2) **une distribution exponentielle** pour les valeurs plus élevées.

La figure 5.3 illustre cette constitution sur l'exemple de la distribution des niveaux d'activité en posture allongée. Les distributions expérimentales relatives aux différents types de mouvements sont présentées en annexe E.1. Une interprétation possible de ces résultats est qu'une forte concentration des niveaux d'activité associés à un type de mouvement oscille autour d'une valeur moyenne, avec cependant la possibilité d'observer quelques fois des valeurs plus élevées. On détermine alors les paramètres des distributions expérimentales correspondant à des activités suffisamment longues – la taille de l'ensemble des valeurs doit être suffisante pour permettre l'estimation des paramètres de la distribution – et durant lesquelles la posture et le type de mouvement observés restent les mêmes. La méthode utilisée consiste à déterminer le niveau d'activité "de coupure" entre la génération des valeurs par une distribution normale ou exponentielle et les paramètres de chaque distribution :

(1) en réalisant d'une part un test d'ajustement graphique à une loi normale des données expérimentales du niveau d'activité (voir Fig. 5.4) et

(2) en vérifiant d'autre part que les distributions expérimentales et simulées sont identiques – test de Kolmogorov-Smirnov (voir Fig. 5.3).

Des détails sur ces tests statistiques relatifs à la nature d'un échantillon sont présentés en annexe C.

Ainsi, le type de mouvement observé est inféré à chaque instant – c'est-à-dire toutes les minutes – de l'observation des déplacements et postures préalablement simulés. Une valeur de niveau d'activité est alors générée aléatoirement à partir d'une distribution elle-même sélectionnée aléatoirement dans l'ensemble des distributions estimées et appropriées au contexte observé – type de mouvement et moment de la journée.

5.2.4 Fréquence cardiaque

Principe de modélisation

L'approche utilisée pour la simulation des valeurs de fréquence cardiaque est guidée par l'*empirisme* (voir paragraphe 4.3.1) – "Il est nécessaire de considérer une distribution spécifique des valeurs de fréquence cardiaque en fonction de la posture et du niveau d'activité. La valeur moyenne de fréquence observée dans un contexte donné est une fonction linéaire du niveau d'activité, cette fonction étant également spécifique à la posture." La construction du modèle prend ainsi en compte une sensibilité dominante de la fréquence cardiaque à la posture de la personne et à l'intensité des activités réalisées. En plus des variations intra-individuelles, il existe une grande

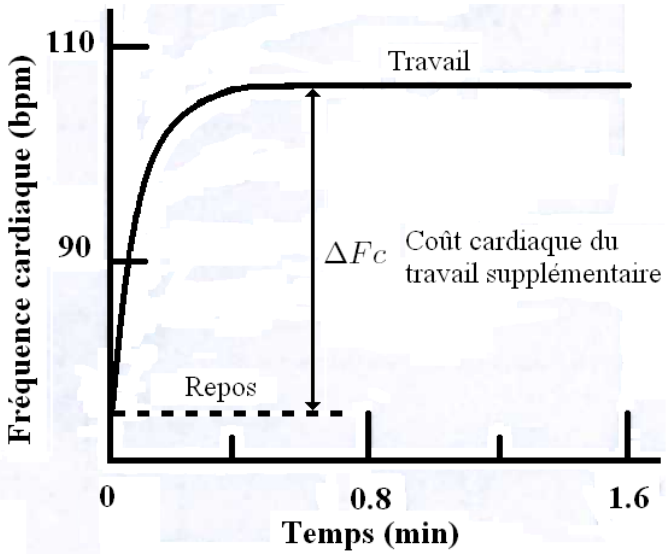


FIG. 5.5 – Coût cardiaque d'un travail musculaire : ΔFc.

variabilité inter-individuelle, en particulier dans les variations de la fréquence cardiaque moyenne au repos.
Ainsi, la simulation des variations de la fréquence cardiaque comprend les étapes suivantes :

(1) Variations de repos ($Fc_{repos}(t)$) – Génération des séquences temporelles correspondant aux variations de la fréquence cardiaque au repos.

(2) Coût cardiaque de l'activité ($\Delta Fc(t)$) – Introduction de variations par rapport aux valeurs de repos en fonction de la posture et du niveau d'activité observés au même moment.

Les variations de la fréquence cardiaque, $Fc(t)$, sont alors obtenues en ajoutant les variations de repos à celles du coût cardiaque, selon (5.1).

$$Fc(t) = Fc_{repos}(t) + \Delta Fc(t). \tag{5.1}$$

Le *coût cardiaque* est un terme proposé par Brouha [16] par analogie avec celui de coût énergétique. Il correspond au nombre de pulsations comptées au-dessus du niveau de repos entre le début et la fin d'un travail musculaire, et rapportées à la minute. On considère que le coût cardiaque, ΔFc, est à peu près équivalent à la différence entre la valeur de repos et le niveau constant atteint par la fréquence cardiaque rapidement après le début de tout travail musculaire d'intensité faible à modérée (voir Fig. 5.5). Une valeur de coût cardiaque est déterminée aléatoirement à chaque instant à partir d'une distribution appropriée en fonction du travail effectué – estimé par le niveau d'activité pour chaque type de posture.

Validation

La validation de ce modèle est réalisée par des *analyses statistiques sur l'ensemble des données expérimentales dédiées à la modélisation.* L'objectif est de vérifier et de quantifier les connaissances académiques prises en compte dans la définition du principe de modélisation. Les données expérimentales disponibles concernant plusieurs sujets, il convient de prendre en compte la variabilité inter-individuelle de la fréquence cardiaque pour réaliser une analyse conjointe des toutes

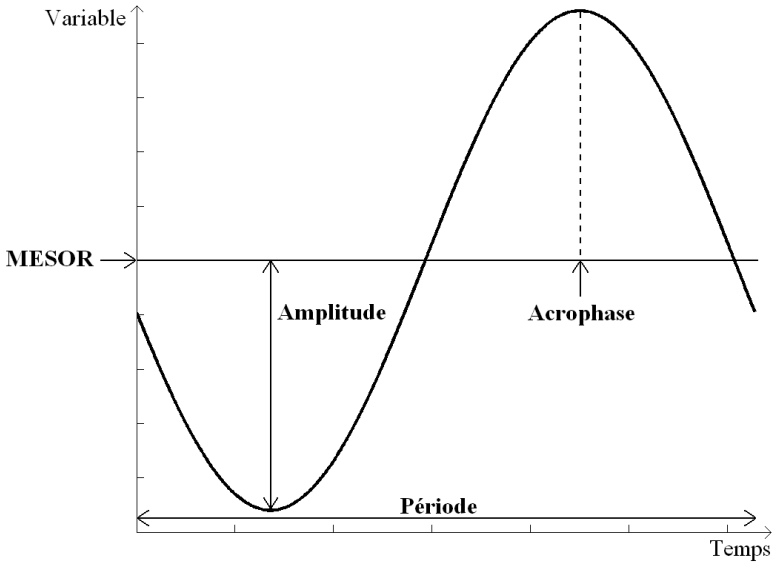


FIG. 5.6 – Paramètres obtenus de l'analyse d'un rythme sinusoïdal par la méthode du cosinor.

les données enregistrées. Comme les variations de repos portent une grande partie de cette variabilité, on propose de normaliser les valeurs de fréquence cardiaque observées pour chaque sujet en fonction d'une estimation des variations de sa fréquence cardiaque de repos.

Les étapes de validation comprennent l'analyse (1) des caractéristiques de variation de la fréquence cardiaque de repos et (2) du coût cardiaque du niveau d'activité en fonction de la posture. En particulier l'analyse des caractéristiques de repos a été guidée par la collaboration avec Céline Chevrier, alors doctorante au Centre de Recherches du Service de Santé des Armées (CRSSA) à Grenoble sur des recherches intégrant des notions de chronobiologie. Sylvie Charbonnier, maître de conférences au Laboratoire d'Automatique de Grenoble (LAG), nous a également apporté son expérience sur l'analyse de ces données qu'elle avait collectées dans son équipe.

Variations au repos

Principe de l'estimation

Un paramètre physiologique tel que la fréquence cardiaque moyenne a des variations qui suivent quasiment un rythme circadien (période de 24 heures). D'après [14], la méthode la plus utilisée par les chronobiologistes pour l'analyse des rythmes circadiens sur des données expérimentales recueillies sur 24 heures a longtemps été la *technique du cosinor* [46], pour laquelle les données collectées sur une période de 24 heures sont représentées par la meilleure fonction sinusoïdale en terme des moindres carrés. Plus tard, des méthodes plus puissantes ont été développées pour estimer un rythme circadien sans l'hypothèse d'une approximation sinusoïdale. Cependant, une grande précision dans l'estimation des paramètres du rythme observé n'est pas toujours nécessaire, et l'utilisation de la seule technique du cosinor permet alors d'obtenir des informations utiles sur les rythmes biologiques [14].

Dans notre contexte expérimental, on dispose de peu de données correspondant à chaque sujet au repos – deux journées non consécutives d'enregistrements lors de diverses activités – et les résultats d'analyse des rythmes circadiens seront de toutes façons peu précis, quelle que soit la méthode. Compte tenu de ce manque de précision, on utilise la **méthode du cosinor**

pour l'estimation du rythme circadien de la fréquence cardiaque de repos pour chaque sujet. Les caractéristiques de variation extraites par cette analyse sont les suivantes (voir Fig. 5.6) :

(1) ***MESOR,*** *M – Midline Estimating Statistic Of Rythm*, c'est-à-dire le niveau moyen autour duquel les valeurs oscillent ;

(2) **Amplitude,** A – Mesure de l'étendue des variations possibles dans les valeurs ;

(3) **Acrophase,** $A\phi$ – Moment de la journée où l'approximation sinusoïdale du rythme atteint la valeur maximale, où ϕ est la phase des variations, en unité trigonométrique.

Les variations temporelles des valeurs de fréquence cardiaque de repos, $Fc_{repos}(t)$, sont alors exprimées en fonction de ces paramètres selon l'équation (5.2), où t est le temps exprimé en heures [14].

$$Fc_{repos}(t) = M + A \cdot sin\left(\frac{2\pi}{24}t + \phi\right). \tag{5.2}$$

Estimation expérimentale

La méthode du cosinor pour l'estimation du rythme circadien de repos doit être théoriquement appliquée pour chaque sujet sur des données expérimentales correspondant à l'observation de la personne au repos couché – la situation de référence – et rapportées ensuite à une période de 24 heures. Les enregistrements expérimentaux dont on dispose ne contiennent cependant que peu de données de ce type, puisque le repos couché est observé principalement pendant la nuit. Afin d'augmenter la quantité de données appropriées à l'estimation, on sélectionne ainsi pour chaque sujet les valeurs proches de la fréquence cardiaque de repos, c'est-à-dire associées à un faible niveau d'activité – en pratique, inférieures à 0.6.

Cependant, d'après les connaissances académiques sur les variations de la fréquence cardiaque (voir paragraphe 4.3.1), on sait que la fréquence cardiaque varie également suivant la posture, même lorsque le sujet a une activité très légère. Il est ainsi nécessaire d'ajuster les valeurs expérimentales sélectionnées qui sont associées à une autre posture qu'allongée afin de tenir compte de cette variation de la fréquence cardiaque. D'après [74], la variation observée est de 10% en posture assise et de 20 à 30% en posture debout par rapport à la fréquence cardiaque en posture allongée. En prenant pour référence la fréquence cardiaque de repos en posture allongée, on divise alors les valeurs observées en posture assise par 1.1, et en posture debout par 1.25.

Les analyses du cosinor alors effectuées pour chaque sujet donnent pour les différents paramètres des valeurs moyennes d'environ : (1) $M \approx 70$ bpm pour le niveau moyen autour desquelles les valeurs oscillent, (2) $A \approx 6$ bpm pour l'amplitude des variations et (3) $A\phi \approx 16$ heures pour l'heure moyenne d'occurrence de la plus forte valeur sur une journée. La figure 5.7 montre le rythme circadien sinusoïdal estimé pour l'un des sujets. Les paramètres estimés pour les différents sujets sont détaillés en annexe E.2.

Même si cette analyse n'est ni très juste – les données utilisées ne correspondent pas exactement au repos couché mais à un faible niveau d'activité rapporté à cette posture, conditions pour lesquelles on dispose de peu de valeurs expérimentales – ni très précise – la méthode du cosinor fait l'hypothèse de variations circadiennes sinusoïdales – les résultats obtenus donnent une estimation plus ou moins réaliste mais significative du rythme de variations de la fréquence cardiaque au repos. Par exemple, une valeur d'acrophase atypique est obtenue pour le premier sujet – valeur maximale vers 3h le matin – et, en vérifiant l'annotation de ses activités, on constate que la personne travaille en fait toute la nuit, ce qui a par conséquent pu décaler progressivement son rythme biologique si ce travail nocturne est habituel. Par contre, le même type de décalage obtenu pour le cinquième sujet ne s'explique pas *a priori* d'après les informations disponibles : il faudrait les enregistrements de plusieurs journées consécutives pour l'expliquer, ou bien les

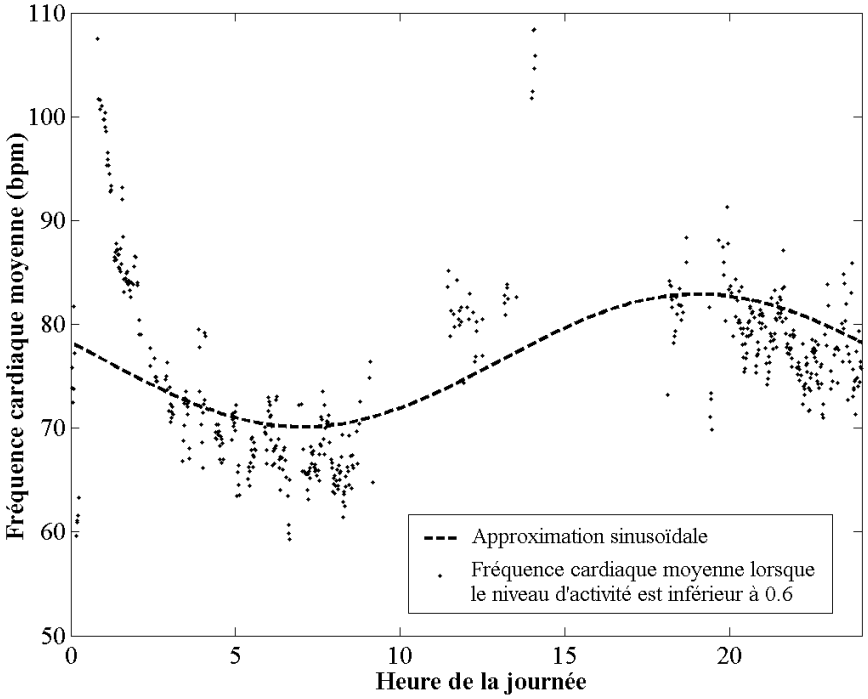


FIG. 5.7 – Approximation sinusoïdale du rythme circadien de la fréquence cardiaque de repos.

données expérimentales sont trop peu nombreuses ou pas assez significatives pour réaliser une estimation correcte du rythme biologique pour cette journée. De façon générale, les résultats moyens obtenus sont néanmoins conformes aux valeurs attendues pour ces paramètres, issues de [46, 74] (voir Tab. 5.2).

	MESOR (bpm)	**Amplitude (bpm)**	**Acrophase (heure)**
Moyenne	69.6	6.1	16h00
Référence [46, 74]	65	[0.7,13.9]	[10h40,19h30]

TAB. 5.2 – Paramètres des variations sinusoïdales de repos observés en moyenne pour les sujets de l'expérimentation, en référence aux valeurs citées dans [46, 74].

Les paramètres décrivant les variations sinusoïdales de repos de la fréquence cardiaque pourront ainsi être modifiés au cours des simulations afin de générer des données correspondant à plusieurs profils physiologiques et comportementaux. Dans la suite des analyses sur les données expérimentales, on normalise les valeurs de fréquence cardiaque observées, $Fc_{obs}(t)$, par rapport aux variations estimées de la fréquence cardiaque de repos, $Fc_{repos}(t)$, selon (5.3), afin de s'affranchir des spécificités individuelles et de considérer conjointement les données expérimentales de la fréquence cardiaque de tous les sujets en terme du coût cardiaque des activités réalisées, $\Delta Fc(t)$.

$$\Delta Fc(t) = Fc_{obs}(t) - Fc_{repos}(t). \tag{5.3}$$

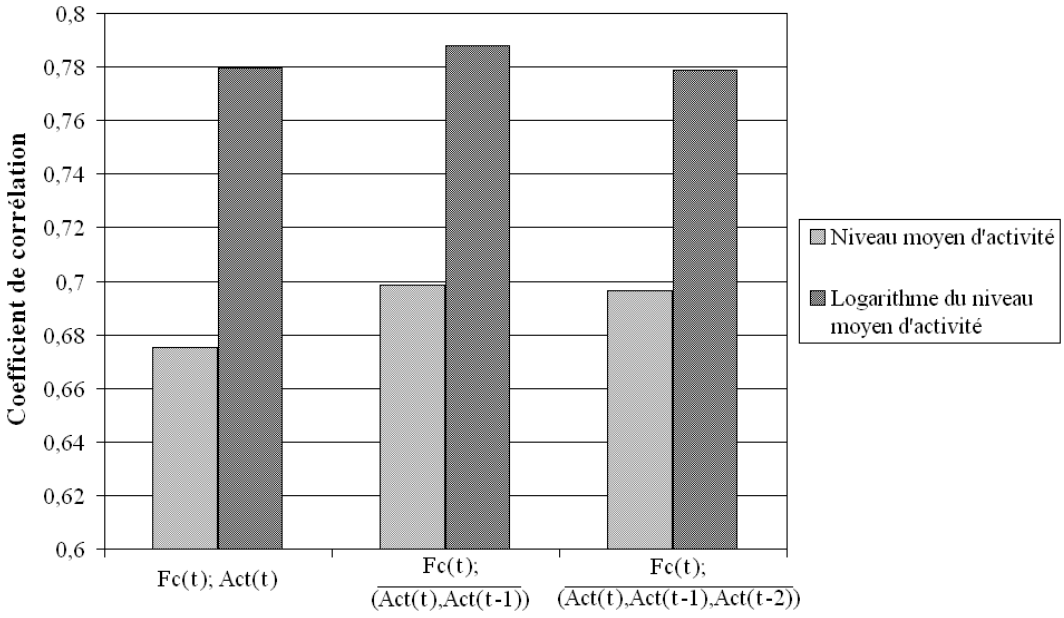


FIG. 5.8 – Coefficients de corrélation linéaire entre fréquence cardiaque et niveau d'activité. On considère la fréquence cardiaque moyenne sur 1 minute et le niveau d'activité moyen sur 1, 2 ou 3 minutes précédent le calcul de la fréquence cardiaque moyenne. Les coefficients de corrélation présentés correspondent aux valeurs moyennes observées pour les sujets de l'expérimentation, d'une part en considérant les niveaux d'activité moyens, et d'autre part en considérant le logarithme de ces valeurs.

Coût cardiaque de l'activité

L'analyse du coût cardiaque des activités réalisées nécessite de prendre en compte les éléments suivants :

1. **Corrélation linéaire entre le niveau d'activité et la fréquence cardiaque**, pour évaluer la portée dans le temps de l'influence d'une certaine activité sur les valeurs enregistrées pour la fréquence cardiaque ;
2. **Variations du coût cardiaque** en fonction des niveaux d'activité observés et pour chaque type de posture ;
3. **Distribution des valeurs du coût cardiaque** d'une posture et d'un niveau d'activité donnés.

Corrélation linéaire entre niveau d'activité et fréquence cardiaque

Dans [74], Monod *et al.* décrivent *a priori* comme une fonction linéaire l'influence de l'effort requis par une activité sur la fréquence cardiaque. Mais compte tenu du besoin de récupération après n'importe quelle activité, une valeur de fréquence cardiaque observée à un instant donné ne dépend pas simplement de l'activité réalisée au même moment. Par conséquent, on a besoin de déterminer l'intervalle de temps précédent un calcul de fréquence cardiaque moyenne pendant lequel le niveau d'activité a une influence sur la valeur observée. Une analyse d'intercorrélation entre ces deux paramètres met en évidence un pic de valeurs quand les deux séquences sont en phase. Ce résultat indique que la relation temporelle peut être décrite par une équation du type (5.4), où $Fc(t)$ représente la valeur moyenne de la fréquence cardiaque pendant la minute précédent l'instant t, et $Act(t)$ la valeur moyenne du niveau d'activité pendant ce même intervalle

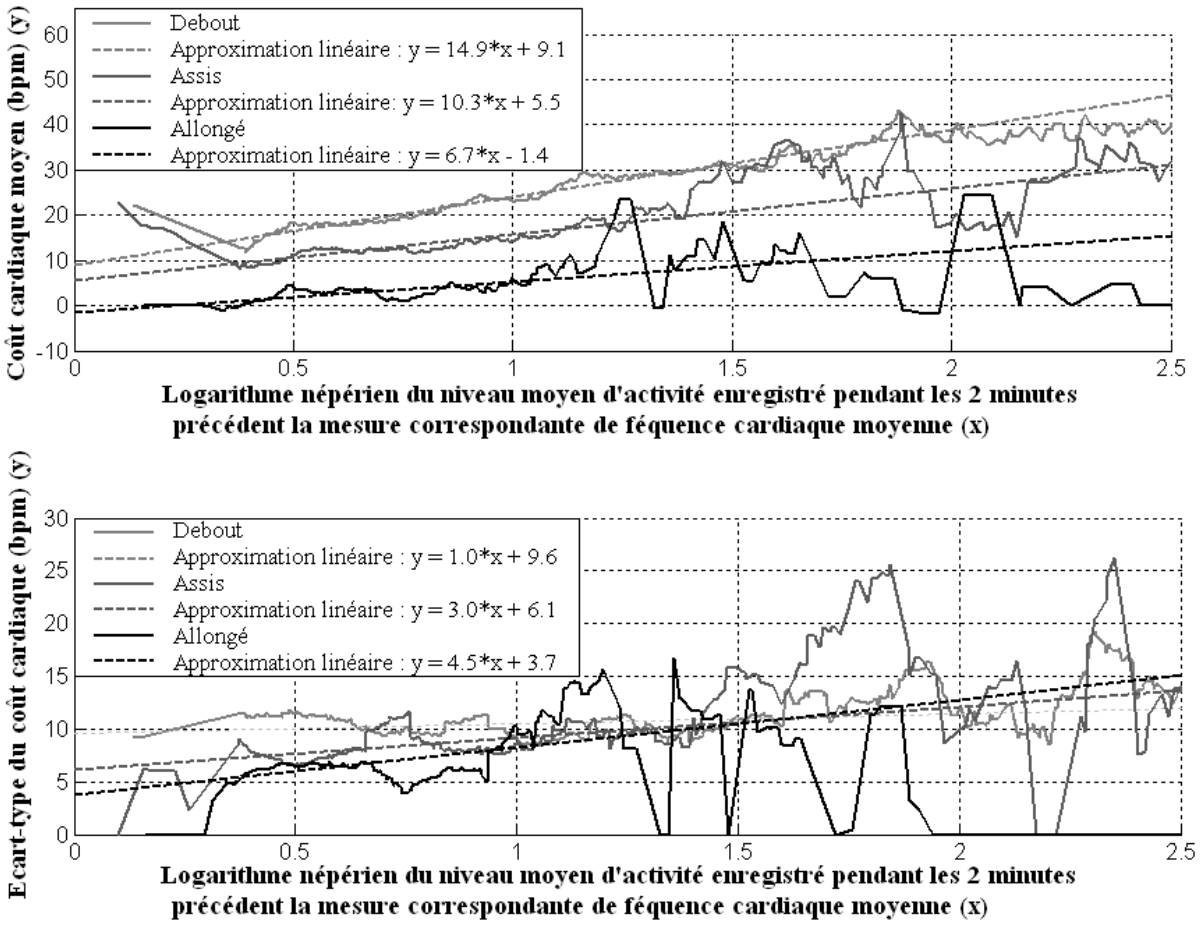


FIG. 5.9 – Moyenne et écart-type du coût cardiaque d'une activité.
Les graphes du haut et du bas présentent respectivement la **moyenne** et l'**écart-type** du coût cardiaque des niveaux d'activité observés en moyenne sur une période de deux minutes précédent la mesure de fréquence cardiaque correspondante, pour les différentes postures considérées.

de temps :

$$Fc(t) = f(Act(t), Act(t-1), Act(t-2), \ldots). \tag{5.4}$$

Par ailleurs, le meilleur coefficient de corrélation linéaire – le plus proche de 1 – est obtenu en moyenne lorsque les niveaux d'activité sont moyennés sur les deux minutes précédent la détermination de la fréquence cardiaque (voir Fig. 5.8). La différence observée n'est pas très significative, mais par ailleurs la connaissance de la nécessité d'en moyenne 2 minutes de récupération après une activité modérée incite à en tenir compte. L'équation décrivant la relation temporelle entre la fréquence cardiaque moyenne et l'activité – posture et niveau d'activité – est alors décrite par l'équation (5.5).

$$Fc(t) = f(\overline{(Act(t), Act(t-1))}). \tag{5.5}$$

Enfin, les valeurs de fréquence cardiaque dépendent également de la posture du sujet, même au repos. L'équation (5.5) se réécrit ainsi comme (5.6), où $Pos(t)$ représente la posture observée pendant la minute précédent l'instant t :

$$Fc(t, Pos(t)) = f(\overline{(Act(t), Act(t-1))}). \tag{5.6}$$

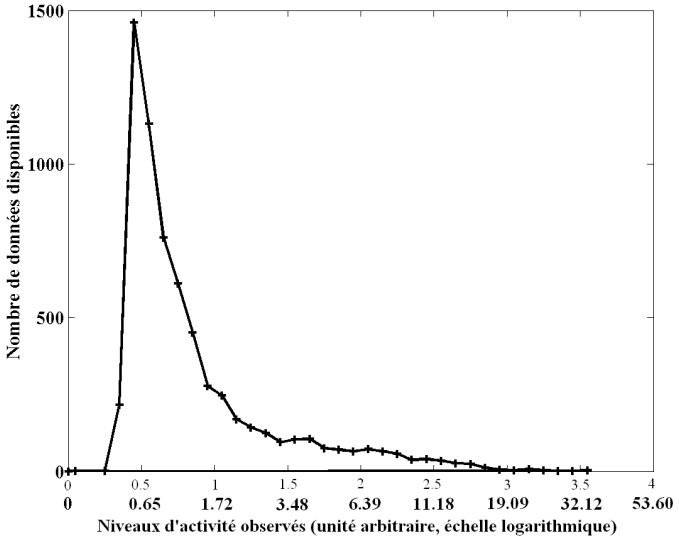


FIG. 5.10 – Nombre de données expérimentales disponibles sur l'échelle du logarithme népérien des niveaux d'activité observés.

Variations du coût cardiaque

On sait par ailleurs qu'il existe une relation le plus souvent quasiment linéaire entre les valeurs de fréquence cardiaque et de niveau d'activité [74]. La relation (5.6) est détaillée à partir d'analyses mathématiques et statistiques sur les données expérimentales normalisées (voir équation (5.3)), indépendamment pour les trois types de posture : allongé, assis et debout. Par ailleurs, pour faire face à l'effet de saturation des valeurs de la fréquence cardiaque avec des niveaux d'activité croissants, on réalise l'étude de la relation entre le coût cardiaque et le logarithme népérien du niveau d'activité moyen pendant les deux minutes précédent la détermination d'une valeur moyenne de fréquence cardiaque. Les coefficients de corrélation linéaire observés entre coût cardiaque et activité dans ce contexte sont d'ailleurs significativement supérieurs (voir Fig. 5.8).

La figure 5.9 représente la moyenne et l'écart-type des coûts cardiaques observés en fonction du niveau d'activité, et pour chaque type de posture. Ces résultats sont en accord avec [74] (voir 4.3.1) et confirment une relation quasiment linéaire entre ces deux paramètres, avec pour chaque niveau d'activité un coût cardiaque d'autant plus élevé que l'effort requis par la posture est important. Il faut cependant noter que peu de données sont disponibles pour les niveaux d'activité modérés à élevés, en particulier lorsque la personne est assise ou allongée (voir Fig. 5.10) – enregistrer d'importants niveaux d'activité n'est par ailleurs pas vraiment habituel dans ces situations. De même on dispose de peu de données correspondant à de faibles niveaux d'activité en posture assise ou debout. Une régression linéaire sur les valeurs moyennes significatives du coût cardiaque donne une estimation des paramètres qui le lient aux valeurs moyennes du logarithme du niveau d'activité. On constate que la pente et l'ordonnée à l'origine de l'approximation linéaire augmentent avec l'effort requis par la posture. Par ailleurs, l'examen individuel des données enregistrées pour chaque sujet montre que cette relation quasiment linéaire est toujours observée, avec cependant de légères différences dans l'estimation des paramètres (voir Annexe E.3).

Pour ce qui concerne la variabilité des valeurs, la relation avec le niveau d'activité pour chaque type de posture est beaucoup moins évidente d'après le graphe du bas de la figure 5.9, représentant pour chaque type de posture l'écart-type des valeurs du coût cardiaque en fonction du niveau d'activité. Le graphe de la figure 5.11 montre la relation moyenne observée entre le

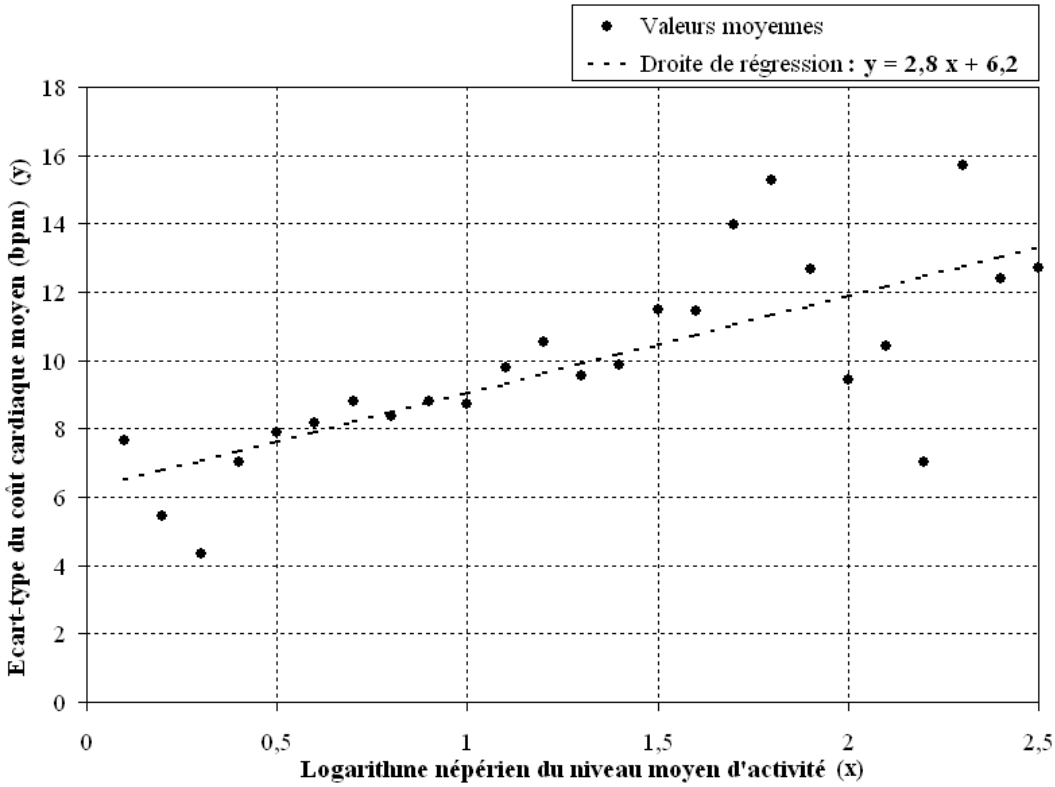


FIG. 5.11 – Écart-type du coût cardiaque des niveaux d'activité quelle que soit la posture. Le graphe présente l'écart-type du coût cardiaque en fonction des niveaux d'activité observés en moyenne sur une période de deux minutes précédent la mesure de fréquence cardiaque correspondante, indifféremment pour l'ensemble des postures.

niveau d'activité et l'écart-type des valeurs de coût cardiaque quelle que soit la posture. L'écart-type sur les valeurs de coût cardiaque est ainsi approximé grossièrement par une fonction linéaire des niveaux moyens d'activité indépendamment de la posture. D'après [74], il existe en effet une variation permanente du rythme cardiaque – l'arythmie sinusale – qui est maximale au repos, et supprimée au cours des travaux musculaires ou d'un effort mental. Cette faible variabilité de la fréquence cardiaque avec l'effort est donc vraie si l'on observe l'évolution des valeurs au cours de la réalisation d'une même activité. Dans notre contexte expérimental, les données de plusieurs sujets réalisant plusieurs types d'activités sont analysées conjointement. La relation linéaire croissante observée en moyenne sur les données expérimentales s'interprète alors surtout (1) d'une part par la diversité des activités réalisées dans chaque type de posture lorsque le niveau d'activité augmente, et (2) d'autre part par l'approximation du coût énergétique d'une activité par le niveau d'activité observé – hypothèse d'autant moins exacte que le coût énergétique d'une activité est élevé. Les tendances de variation du coût cardiaque sont alors largement variables avec des niveaux croissants d'activité. L'écart-type du coût cardiaque ne traduit ainsi pas une variabilité ponctuelle des valeurs du coût cardiaque associé à un certain niveau d'activité et une certaine posture – d'une valeur à la suivante – mais une variabilité sur une certaine durée ou entre différentes activités possiblement réalisées.

Compte tenu de la variabilité inter-individuelle observée, les paramètres définissant la valeur moyenne et l'écart-type du coût cardiaque pourront être modifiés au cours des simulations afin

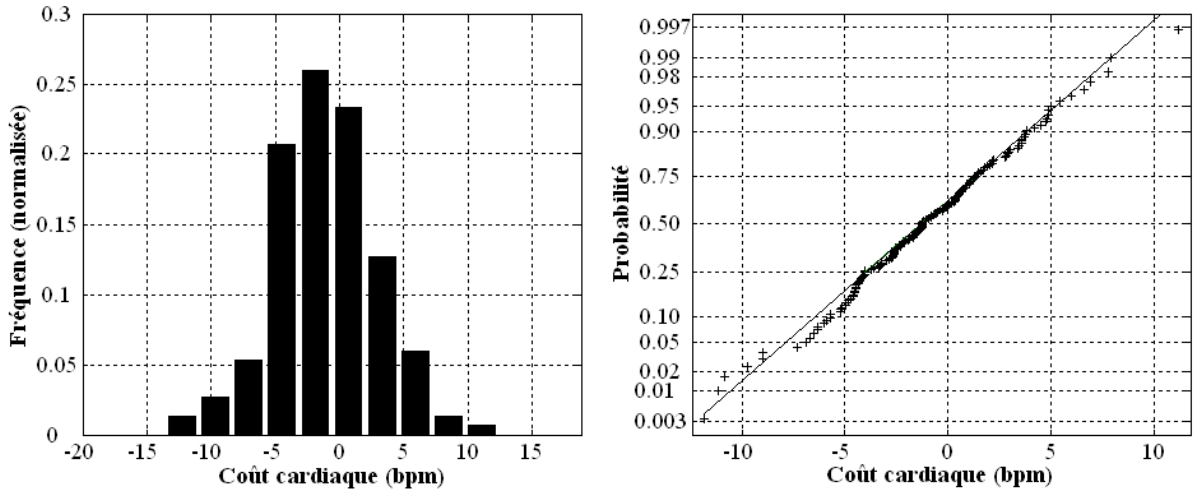


FIG. 5.12 – Distribution des valeurs de coût cardiaque.
Distribution correspondant à des niveaux d'activité compris entre 0.495 et 0.500, lorsque la personne est allongée : **Histogramme** (graphe de gauche) et **Test d'ajustement graphique** à une loi normale (graphe de droite).

de générer des données correspondant à plusieurs profils de personnes réalisant plusieurs types d'activités. La variabilité intra-individuelle dans les valeurs de coût cardiaque observées permet par ailleurs de simuler implicitement la réalisation de plusieurs types d'activité au cours du temps.

Distribution des valeurs de coût cardiaque
Enfin, en considérant une posture et un niveau d'activité moyen donné, on approxime la distribution des valeurs de coût cardiaque observées par une distribution normale, dont la moyenne et l'écart-type sont déterminés d'après les relations précédemment identifiées. Cette hypothèse de normalité ne peut être vérifiée que sur de petits intervalles de variation du niveau d'activité pour lesquels on dispose de beaucoup de données expérimentales correspondant à une même posture. Ainsi, des tests de normalité sont en particulier réalisés en posture allongée, et pour des niveaux d'activité proches du mode des valeurs observées dans cette posture – 0.51 – pour lequel on dispose d'un grand nombre de données.

Un exemple de l'histogramme d'une distribution expérimentale et du test d'ajustement graphique à une loi normale est présenté sur la figure 5.12. Dans le cas présenté, le test de Lilliefors confirme la présomption de normalité de la distribution issue de l'observation du tracé graphique – courbe voisine d'une droite. Des détails sur les tests statistiques relatifs à la nature d'un échantillon sont présentés en annexe C.

Finalement, les valeurs de fréquence cardiaque moyenne sont générées suivant le processus suivant (voir Fig. 5.13) :

1. **Variations circadiennes sinusoïdales** des valeurs de fréquence cardiaque de repos pour le sujet "simulé",
2. **Variations du coût cardiaque en fonction de la posture et du niveau d'activité :** les valeurs de coût cardiaque sont déterminées aléatoirement à partir d'une distribution normale dont la moyenne et l'écart-type correspondent aux caractéristiques linéaires observées entre le coût cardiaque et le niveau d'activité, pour chaque type de posture.

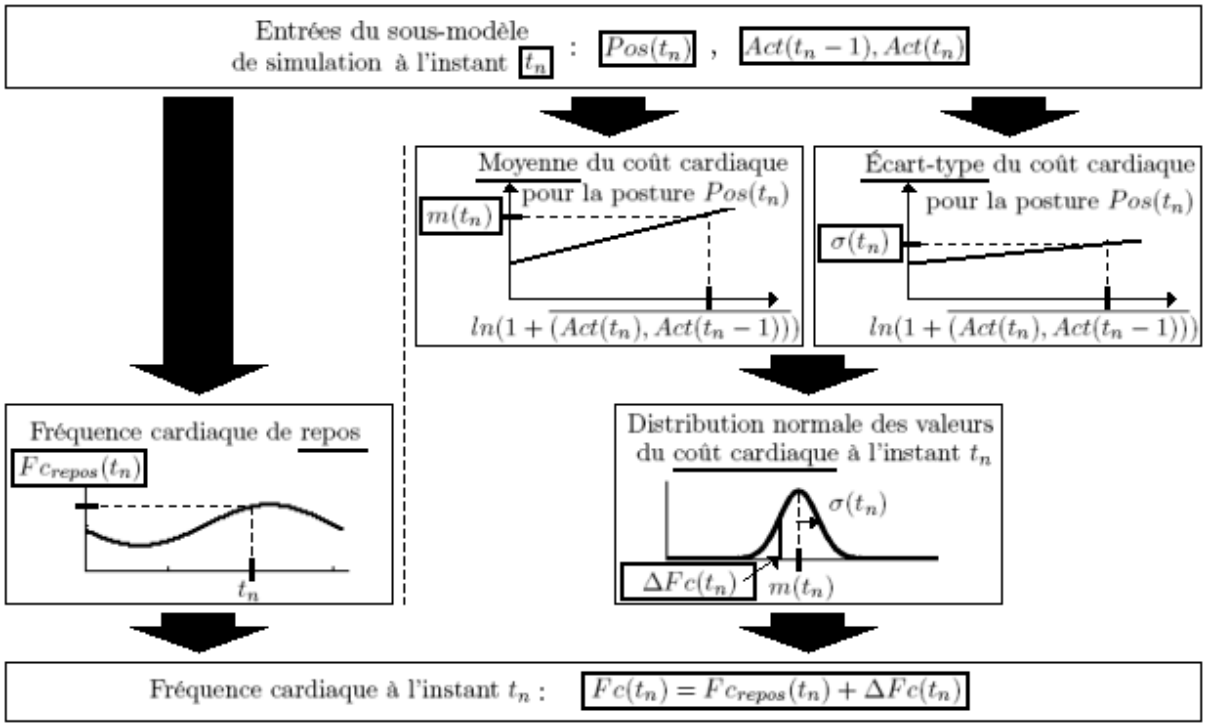


FIG. 5.13 – Principe de génération des valeurs de fréquence cardiaque.
Les valeurs de fréquence cardiaque sont fonction de la posture et des niveaux d'activité enregistrés dans les minutes précédentes.

3. **Addition des variations de repos et du coût cardiaque**, à chaque instant.

On considère ainsi que les facteurs principaux d'influence sur la fréquence cardiaque sont liés à l'activité physique de la personne : posture et niveau d'activité. Il existe cependant d'autres facteurs d'influence tels que le stress, la prise de médicaments, l'activité végétative. Ces événements de la vie quotidienne étaient d'ailleurs probablement plus ou moins présents chez les sujets de l'expérimentation, et, par conséquent, on en a tenu implicitement compte en terme de variabilité des valeurs de fréquence cardiaque observées. Pour ce qui concerne l'activité végétative, une analyse expérimentale de son influence est possible puisque les périodes de repas sont parfois annotées par les sujets lors des expérimentations. Cependant, aucune influence flagrante sur la fréquence cardiaque pendant les 3 heures suivant le repas [74] n'a pu être observée (voir Annexe E.4). Il s'agit probablement d'une influence trop légère pour qu'elle puisse être observée dans notre contexte, compte tenu de l'ensemble des autres facteurs d'influence non explicités et de l'imprécision de l'estimation des variations de repos et du coût cardiaque des activités et postures.

5.3 Synthèse

Le schéma de la figure 5.14 synthétise le principe de simulation de l'ensemble des paramètres – structure en cascade – et les techniques de modélisation et de simulation utilisées à chaque étape. Ce schéma présente dans des élipses les paramètres simulés, et dans des rectangles les sous-modèles utilisés pour la simulation, en détaillant de haut en bas : (1) la technique de

modélisation utilisée, (2) le sous-modèle de simulation construit, et (3) la technique utilisée pour la validation de ce modèle. Les flèches pleines vont d'un modèle vers les données qu'il produit ; celles en pointillés vont des données vers le(s) modèle(s) qui les utilise(nt).

La démarche de simulation intègre ainsi plusieurs types de connaissances et de techniques de modélisation et de validation. La conception de chaque sous-modèle est guidée pas des connaissances *a priori* – de sens commun (*rationalisme*) ou académiques (*empirisme*). Les modèles sont fondés soit sur des automates à états finis dans le cas de paramètres qualitatifs, soit sur un ensemble de distributions dans le cas de paramètres quantitatifs. Leur validation est ensuite réalisée par des analyses statistiques sur des données expérimentales si elles sont disponibles, sinon avec l'aide d'un expert du domaine.

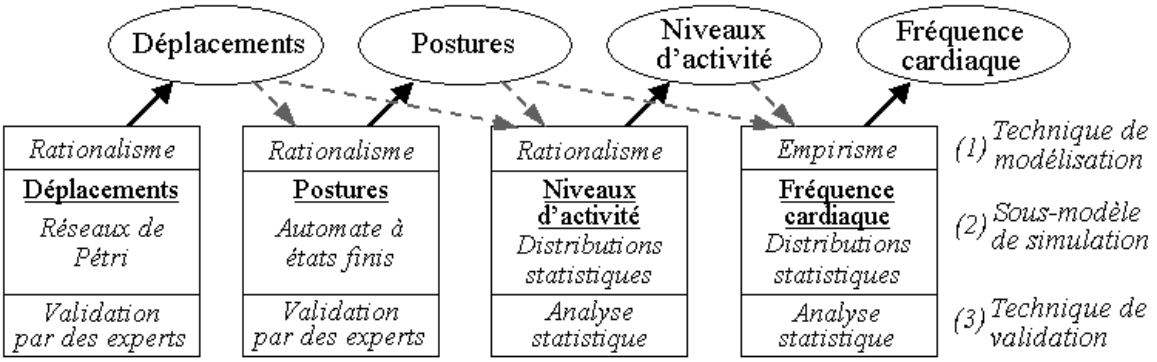


FIG. 5.14 – Synthèse des techniques de construction et de validation du modèle de simulation.

6

Expérimentation et validation opérationnelle

Ce chapitre propose l'analyse et la validation des séquences de données générées par le processus de simulation construit selon le modèle défini au chapitre précédent.

6.1 Contexte d'expérimentation et de validation

Le modèle de simulation est implémenté sous Matlab et expérimenté pour la génération de séquences de données correspondant à plusieurs profils de personnes et types de situations. La complexité du modèle, particulièrement en terme de la grande quantité de paramètres à définir *a priori*, offre l'avantage de la diversité des profils et des situations simulés, mais rend aussi difficile le choix des paramètres qui permettent la génération de données représentatives d'un type de comportement ou d'une certaine situation. La plupart nécessitent d'être définis exclusivement intuitivement d'après leur signification dans le modèle et le sens de la simulation, particulièrement lorsqu'ils concernent les sous-modèles de simulation des paramètres qualitatifs. La détermination des paramètres des autres sous-modèles est plus contrainte par les résultats des analyses sur les données expérimentales. On a malgré tout observé qu'il existe une grande variabilité, inter-individuelle notamment, dans leur définition.
La démarche de simulation peut ainsi nécessiter des tâtonnements et n'est pas évidente à mettre en oeuvre. Les détails de l'implémentation proposés en annexe F montrent l'interface graphique développée pour la définition de l'ensemble des paramètres d'une simulation.

Le graphe de la figure 6.1 présente un exemple de séquences de données générées par le processus de simulation lors de premières expérimentation. Conformément au schéma de mise en place d'un processus de simulation (voir Fig. 3.1), l'objectif est alors de valider la pertinence de ces séquences, non pas en terme d'un réalisme "parfait", mais en fonction des caractéristiques importantes à préserver au regard de la décision. Ainsi, il s'agit en particulier d'évaluer à "haut niveau" la pertinence et l'interprétation possible des séquences générées en terme de la réalisation des activités quotidiennes d'une personne, dans le respect des caractéristiques de variation conjointe des différents paramètres simulés.

Deux techniques de validation des séquences simulées sont utilisées, selon que l'on dispose ou non de données expérimentales pour les paramètres considérés :

1. Soit une **validation intuitive** avec l'aide des connaissances d'un expert si on ne dispose pas de données expérimentales associées à un ou plusieurs paramètres ;

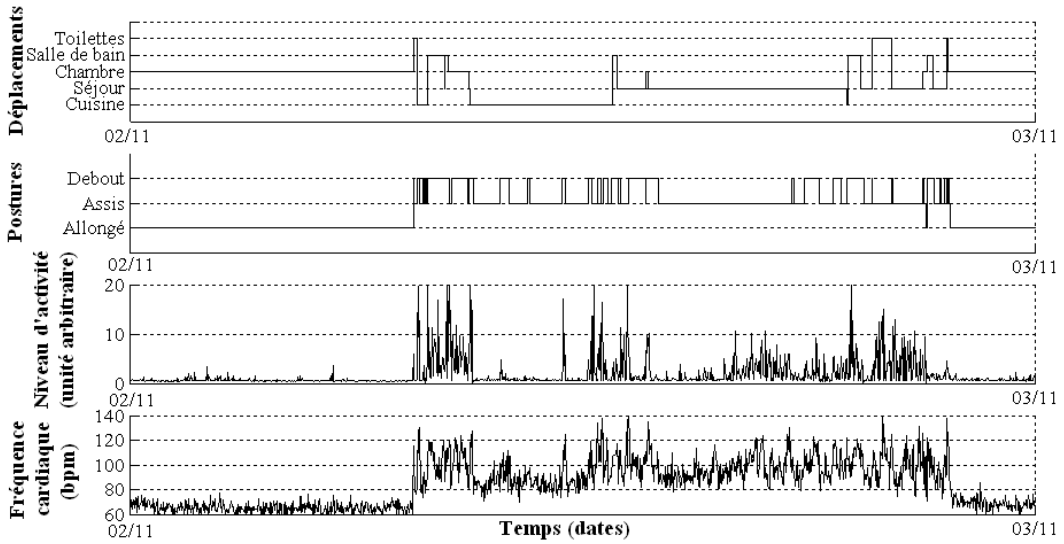


FIG. 6.1 – Séquence temporelle à 4 dimensions générée par le processus de simulation.
La séquence présentée est représentative d'une personne télésurveillée pendant une journée, incluant de haut en base les paramètres suivants : (1) déplacements, (2) postures, (3) niveau d'activité moyen et (4) fréquence cardiaque moyenne.

2. Soit une **validation par analyses mathématiques et statistiques** comparativement sur les données expérimentales et simulées.

Dans notre contexte, on ne dispose pas de données expérimentales correspondant aux *paramètres qualitatifs* – déplacements et postures – dont les séquences simulées sont donc validées intuitivement. Par contre, on dispose d'enregistrements dans un environnement de surveillance médicale des *paramètres quantitatifs* – niveau d'activité et fréquence cardiaque – qui permettent de comparer les données simulées à celles collectées à partir de l'observation d'un système réel.

6.1.1 Validation par les experts

Les séquences de données générées pour les ***paramètres qualitatifs*** – déplacements et postures – sont validées par interaction avec des experts. Ces échanges permettent d'affiner le réglage des paramètres afin de générer des séquences de déplacements et postures réalistes, c'est-à-dire qui peuvent être interprétées de façon cohérente par rapport au rythme attendu d'une personne dans ses activités quotidiennes à domicile. L'observation des variations des déplacements et postures au cours d'une journée donne en effet une bonne idée des activités réalisées à chaque instant, confirmées par les variations du niveau d'activité et de la fréquence cardiaque.

D'après les connaissances *a priori* sur le déroulement des activités de la vie quotidienne (voir paragraphe 4.3.1), la validation des séquences de données par des experts se fait à deux niveaux :

- **Vérifier la réalisation des activités de base de la vie quotidienne.** On s'attend à retrouver quotidiennement des sous-séquences de données représentatives des activités de base – sommeil, toilette, alimentation, passage aux cabinets – avec des horaires et durées plus ou moins variables, et même si on n'est pas forcément capable de décomposer précisément leur réalisation : faire sa toilette par exemple doit inclure de se laver et de s'habiller, de même que s'alimenter peut inclure la préparation du repas.

- **S'assurer de la cohérence des séquences correspondant aux autres périodes de temps.** Les données simulées doivent pouvoir être interprétées en terme de la réalisation d'une activité réaliste pendant les périodes correspondantes, avec une certaine cohérence dans leur réalisation et leur enchaînement.

L'avis d'experts sur les séquences simulées par rapport aux habitudes de vie d'une personne à domicile guide alors en particulier le réglage des paramètres des automates à états finis (probabilités de transition) modélisant les variations des déplacements et postures. Cela permet aussi de vérifier la cohérence des tendances de variation des niveaux d'activité et de la fréquence cardiaque observés en fonction des déplacements et postures. On assure ainsi la génération de séquences de données multidimensionnelles globalement représentatives d'un rythme de vie réaliste pour une personne.

Cette démarche a en particulier été réalisée avec la collaboration du Professeur Hélène Pigot. Des incohérences entre les données simulées et les variations attendues dans des conditions habituelles de vie ont ainsi été détectées, induisant de nouveaux réglages pour certains paramètres. Les premières séquences simulées n'étaient par exemple pas toujours réalistes en terme de la fréquence des changements de postures. En particulier, lorsqu'une personne est le midi dans la cuisine, son activité s'interprète comme la préparation et la prise d'un repas. On s'attend alors à de fréquents changements entre les postures debout et assis, notamment au début (préparation) et à la fin (rangement) de cette activité de base, associés à des niveaux d'activités plutôt élevés. Le soir par contre la période du repas peut être plus courte et moins active : il est en effet relativement fréquent, en particulier pour une personne âgée à domicile, de ne réchauffer pour le soir qu'un plat préparé d'avance le midi.

6.1.2 Validation par analyses mathématiques et statistiques

Les séquences de données générées pour les ***paramètres quantitatifs*** – niveau d'activité et fréquence cardiaque – sont validées par comparaison avec des données expérimentales enregistrées dans un contexte de surveillance médicale (voir paragraphe 4.3.2). Une partie des données expérimentales disponibles est réservée à cette étape de validation, l'autre partie ayant été utilisée pour la modélisation. Compte tenu de l'influence des déplacements et postures sur le niveau d'activité et la fréquence cardiaque, la comparaison entre les séquences réelles et simulées n'a de sens que si les données considérées correspondent à un même contexte expérimental. À partir d'une séquence de données simulées à quatre dimensions – déplacements, postures, niveau d'activité et fréquence cardiaque – l'idée est alors d'arranger sur la même durée des sous-séquences expérimentales – niveau d'activité et fréquence cardiaque – caractéristiques de chaque type d'activité réalisée, c'est-à-dire cohérentes avec les séquences de déplacements et postures générées par la simulation. Afin d'être au plus proche de séquences de données qui auraient été enregistrées réellement dans le contexte considéré, on favorise l'insertion des plus longues sous-séquences possibles issues des enregistrements expérimentaux. Il est ainsi pertinent de comparer les séquences réelles "reconstruites" et simulées, par l'intermédiaire de graphes ou de tests statistiques.

La méthode de construction de ces séquences réelles pour la validation opérationnelle repose sur l'annotation lors des expérimentations des couples successifs de niveau d'activité et fréquence cardiaque en fonction du type de mouvement observé au même moment – mouvements en posture allongée, assise, debout, ou marche. Pour s'affranchir des spécificités de chaque sujet et considérer conjointement toutes les données expérimentales disponibles, on normalise les valeurs de fréquence cardiaque en fonction des variations de repos sinusoïdales estimées pour chaque sujet. On dispose ainsi de séquences de niveau d'activité et du coût cardiaque associé en fonction du type de mouvement réalisé au même moment. A chaque instant – toutes les minutes – et

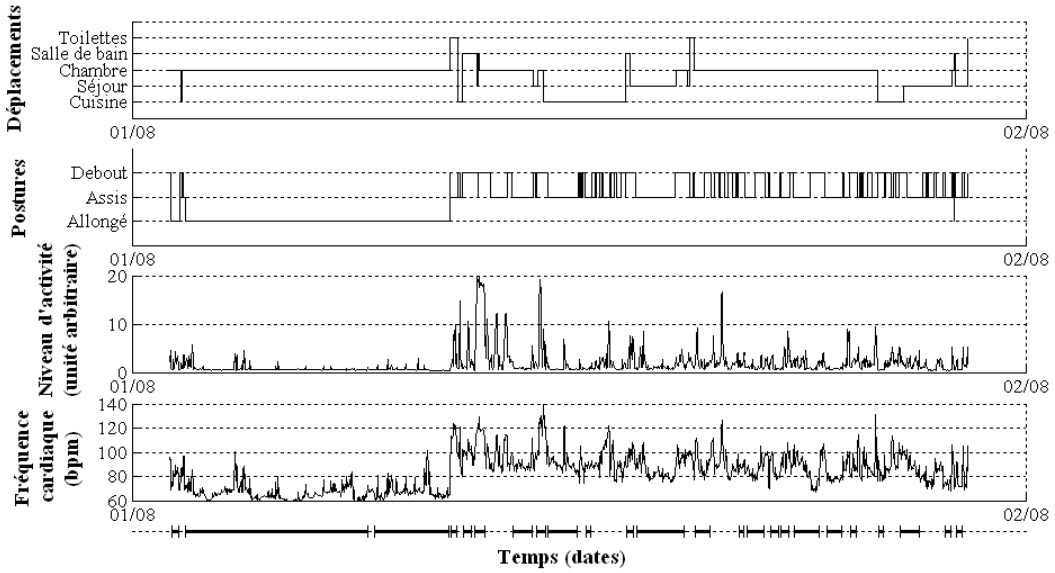


FIG. 6.2 – Séquence temporelle à 4 dimensions reconstituée à partir des données enregistrées par un système réel en fonction du contexte expérimental de la simulation.
La séquence comprend, de haut en bas, les paramètres suivants : (1) déplacements, (2) postures, (3) niveau d'activité moyen et (4) fréquence cardiaque moyenne. Les déplacements et postures sont issus de la simulation. Les intervalles tracés en bas du graphe mettent en évidence les sous-séquences significatives – de plus de 10 minutes – correspondant exactement à des sous-séquences expérimentales pour le niveau d'activité et la fréquence cardiaque.

sur toute la durée de la simulation, on sélectionne alors aléatoirement un couple approprié de valeurs réelles parmi l'ensemble des données disponibles, selon les critères suivants permettant de s'assurer d'un même contexte expérimental :

(1) Le type de mouvement observé est le même que celui simulé au même instant ;

(2) Le moment d'enregistrement des données réelles est assez proche de l'instant expérimental considéré ;

(3) Si elles peuvent être définies, les moyennes du niveau d'activité sur les deux minutes précédent l'instant considéré, d'une part pour la séquence réelle reconstruite pour la validation, et d'autre part pour la séquence réelle contenant le couple de valeurs sélectionnées, doivent être similaires.

Une fois qu'un couple de valeurs approprié est sélectionné à un instant donné, le couple suivant dans la séquence réelle enregistrée, s'il est pertinent, est sélectionné de préférence pour continuer la construction de la séquence réelle de validation à l'instant d'après. On s'assure ainsi de construire une séquence de validation qui contient un grand nombre de sous-séquences correspondant à des successions réelles de couples de valeurs niveau d'activité, coût cardiaque associé. Enfin, les valeurs de fréquence cardiaque sont calculées en ajoutant les variations circadiennes du sujet "simulé" aux valeurs de coût cardiaque sélectionnées sur la durée expérimentale considérée.

Le graphe de la figure 6.2 présente un exemple de séquence temporelle dédiée à la validation, reconstruite à partir de séquences réelles de niveau d'activité et fréquence cardiaque, dans un contexte expérimental posé par les résultats de la simulation des déplacements et postures.

6.2 Discussion sur la qualité des résultats

Dans l'objectif de validation des résultats du processus de simulation, on compare des séquences de données simulées à celles reconstituées dans le même contexte de simulation à partir des enregistrements de niveau d'activité et de fréquence cardiaque d'un système réel. Les trois premiers graphes de la figure 6.3 présentent un exemple de ces séquences : le graphe (1) présente le contexte de simulation – déplacements et postures ; le graphe (2) la séquence simulée de niveaux d'activité et fréquence cardiaque ; et le graphe (3) la séquence correspondant au même contexte mais reconstituée à partir d'enregistrements d'un système réel. Avant de chercher à effectuer une validation objective des résultats obtenus par comparaison aux données enregistrées par un système réel, une simple observation subjective de l'allure des séquences simulées met en évidence une trop grande variabilité haute fréquence dans les valeurs simulées par rapport aux séquences de données expérimentales. Par contre, les tendances globales de variation apparaissent cohérentes avec les déplacements et postures observés : on observe de faibles niveaux d'activité et des valeurs de fréquence cardiaque proches du repos pendant les périodes de sommeil, la nuit – posture allongée, la nuit, dans la chambre – et de plus importantes valeurs pendant la journée, notamment en posture debout où lorsqu'on note de fréquents changements de posture. Par ailleurs, la construction du modèle de simulation assure des distributions de valeurs effectivement similaires entre les données simulées et les enregistrements réels.

Ces considérations signifient que des caractéristiques autres que statistiques doivent être intégrées dans la conception du modèle de simulation afin de rendre plus réaliste l'allure des séquences générées, en accord avec le type de variations observé sur les enregistrements réels. Les observations précédentes renforcent l'idée intuitive de la nécessité de prendre en compte des informations sur l'organisation temporelle des valeurs de niveau d'activité et de fréquence cardiaque au cours de chaque activité. D'ailleurs, les séquences de validation sont construites à partir d'enregistrements expérimentaux en intégrant le plus possible de longues sous-séquences de données collectées réellement de façon continue dans le temps, plutôt qu'en sélectionnant indépendamment à chaque instant une paire de valeurs de niveau d'activité et fréquence cardiaque appropriée au contexte de simulation. Le graphe (4) de la figure 6.3 présente une séquence de validation cohérente avec le contexte de simulation du graphe (1) mais dont les valeurs ont été sélectionnées indépendamment à chaque instant, sans chercher à conserver les successions de valeurs réellement observées. L'allure de cette séquence de validation apparaît alors plus proche de celle des données simulées. Ceci confirme la nécessité de prendre en compte une contrainte temporelle supplémentaire dans la construction du modèle de simulation des deux paramètres quantitatifs, de même que l'utilisation d'un graphe à états finis pour la simulation des paramètres qualitatifs définit implicitement des relations temporelles entre les modalités des paramètres. Dans le cas des paramètres quantitatifs, seule la sélection aléatoire d'une distribution de valeurs de niveau d'activité appropriée au contexte de l'activité réalisée et au moment de la journée prend en compte le temps. Chaque distribution sélectionnée est représentative des valeurs expérimentale observées dans la réalisation d'une même activité sur une certaine durée. Par conséquent, on respecte bien globalement la succession de différents types d'activité au cours d'une journée.

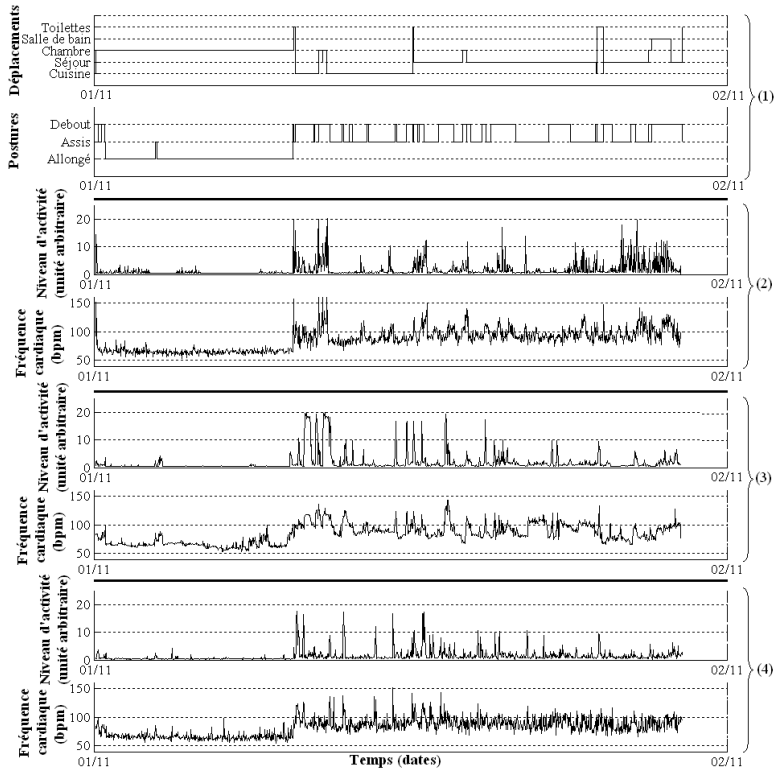


FIG. 6.3 – Séquences temporelles représentatives d'une personne télésurveillée et mettant en évidence la nécessité de prise en compte d'une contrainte temporelle supplémentaire.
Les séquences présentées correspondent à une journée et incluent, de haut en bas :

(1) Une séquence de déplacements et postures générés par le processus de simulation, posant le contexte expérimental ;

(2) Une séquence à deux dimensions de niveaux d'activité et fréquence cardiaque simulés dans ce contexte ;

(3) Une séquence reconstituée à partir d'enregistrements de niveau d'activité et fréquence cardiaque d'un système réel, en insérant le plus possible de longues séquences réelles caractéristiques du contexte de simulation ;

(4) Une séquence reconstituée à partir d'enregistrements d'un système réel, mais en sélectionnant point à point des vecteurs adaptés au contexte observé, sans souci de conserver une cohérence temporelle entre les couples de niveau d'activité et fréquence cardiaque sélectionnés.

6.3 Introduction d'une contrainte temporelle supplémentaire

Ce paragraphe explique et discute une étape de raffinement du processus de simulation afin de mieux prendre en compte la composante temporelle au niveau de chaque activité réalisée. Selon le cycle de mise en place d'un processus de simulation (voir Fig. 3.1), on présente alors successivement (1) le principe de modélisation, (2) la validation conceptuelle du modèle proposé, (3) son implémentation et (4) la validation opérationnelle des nouvelles données générées par le processus de simulation.

6.3.1 Principe de modélisation

L'analyse préliminaire des résultats précédents met en évidence la nécessité d'ajouter une contrainte temporelle dans la construction du modèle de simulation du niveau d'activité et de la fréquence cardiaque. La prise en compte de la composante temporelle est complexe et difficile à formaliser et à valider car on dispose de peu connaissances *a priori* sur la constitution des séquences de niveau d'activité et de fréquence cardiaque. Elle est cependant cruciale dans le cadre d'une étude ayant pour objectif la détection des évolutions critiques de la situation d'une personne dans le temps.

Dans notre contexte, les hypothèses réalisées pour mieux prendre en compte la composante temporelle sont ainsi guidées par les connaissances de sens commun sur les caractéristiques d'évolution dans le temps des valeurs de niveau d'activité et de fréquence cardiaque lorsqu'une certaine activité est réalisée. On suppose que les valeurs de niveau d'activité et de fréquence cardiaque ne sont pas distribuées aléatoirement au cours de la réalisation d'une certaine activité, mais que la succession des valeurs respecte un ***principe général de continuité physique***. Dans notre contexte, cela signifie que la différence absolue entre deux valeurs enregistrées de façon rapprochée dans le temps reste le plus souvent assez faible.

Jusque là, une distribution des valeurs de niveau d'activité estimée à partir des données expérimentales correspondant à une certaine activité et appropriée au contexte de simulation – déplacements, postures et moment de la journée – est utilisée pour la génération aléatoire de valeurs. De même, une distribution des valeurs de fréquence cardiaque est estimée à chaque instant à partir de la connaissance de la posture et du niveau moyen d'activité dans les deux minutes précédentes. La prise en compte de cette contrainte de continuité sur la succession de valeurs possible entraîne la nécessité de réorganiser dans le temps les valeurs de niveau d'activité et de fréquence cardiaque, une fois générées conformément aux distributions de valeurs attendues dans un certain contexte de simulation. Cette réorganisation s'effectue sur les intervalles de temps où l'on sait que les valeurs successives ont été générées selon des distributions similaires, afin de préserver la complexité de la simulation et la cohérence entre les valeurs de l'ensemble des paramètres simulés. Cela signifie que : (1) les niveaux d'activité sont réorganisés sur les périodes où une même activité est réalisée, et (2) les fréquences cardiaques le sont lorsque la personne maintient une même posture avec des niveaux d'activité proches. On respecte par ailleurs la structure en cascade du modèle de simulation en réorganisant d'abord les valeurs de niveau d'activité sur les intervalles de temps appropriés, correspondant à une même activité réalisée, puis selon le même principe les valeurs alors générées pour la fréquence cardiaque sur les périodes de niveaux d'activité successifs proches avec une même posture maintenue.
Les conséquences attendues sur les résultats de la simulation sont alors les suivantes :

- **Conservation des distributions de valeurs** – La cohérence des variations conjointes des paramètres est conservée puisqu'on ne réorganise les valeurs que sur les intervalles de temps où une distribution similaire est utilisée pour leur génération.

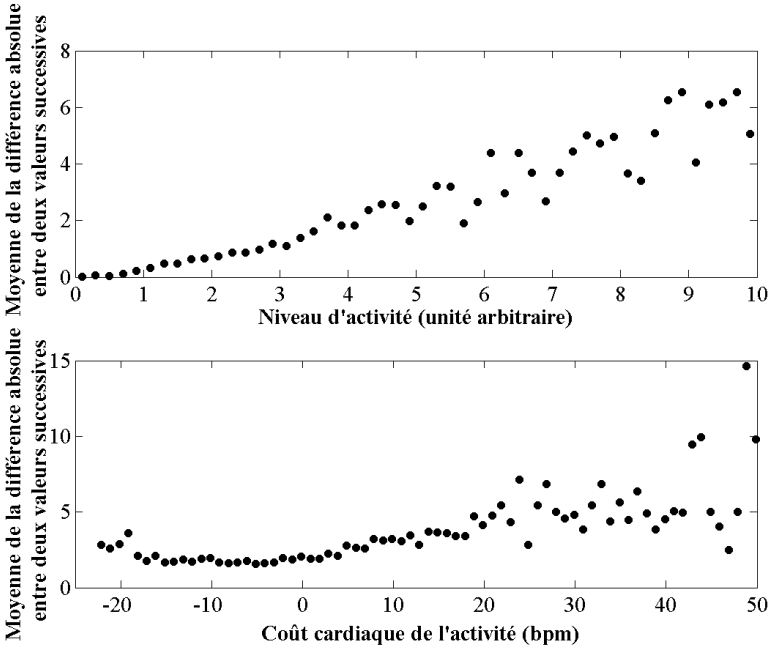


FIG. 6.4 – Relation moyenne entre la différence absolue entre deux valeurs successives dans le temps et la première de ces deux valeurs.
Les deux graphes présentés concernent, (1) pour celui du haut, le niveau d'activité, et (2) pour celui du bas, le coût cardiaque d'une activité.

- **Réduction de la variabilité haute fréquence** – Les "perturbations" – pics de valeurs – observées au cours d'une même activité pour les paramètres quantitatifs sont rassemblées de façon à ce qu'il n'y ait pas de variations trop brusques dans les valeurs successives des différents paramètres.
- **Conservation des caractéristiques basses fréquences** – Les tendances de variation des différents paramètres, spécifiques à chaque activité, sont conservées puis qu'on ne réorganise les valeurs quantitatives que sur des intervalles de temps correspondant au plus à la durée de réalisation d'une activité. Il n'y a donc aucune influence sur la répartition des activités au cours de la journée.

La validation de ces hypothèses fait partie intégrante d'un cycle d'évolution du processus de simulation (voir Fig. 3.1), c'est-à-dire principalement à deux niveaux : (1) **Validation conceptuelle** des nouveaux concepts intégrés dans la construction du modèle, et (2) **Validation opérationnelle**, des séquences de données générées après l'**implémentation** de ce nouveau modèle.

6.3.2 Validation conceptuelle

Des analyses mathématiques et statistiques sont réalisées sur l'ensemble des données expérimentales dédiées à la modélisation afin de valider les nouveaux concepts introduits dans le modèle de simulation. On s'intéresse particulièrement à l'organisation temporelle des valeurs de niveau d'activité et de fréquence cardiaque, c'est-à-dire aux valeurs probables de ces paramètres à un instant donné sachant leur valeurs aux instants précédents.

Une étude possible est de s'intéresser à l'estimation des distributions *a priori* des valeurs de ces paramètres, en fonction des valeurs observées aux instants précédents. Une estimation fiable

nécessite cependant un grand nombre de données expérimentales représentatives de chaque valeurs possibles des paramètres. Dans le contexte d'observation des tendances globales de variation des différents paramètres et d'imprécision de la définition de leurs variations conjointe, une étude aussi précise n'a par ailleurs pas vraiment de sens.

On décide alors d'étudier plus simplement la relation moyenne entre une valeur observée à un instant t et la différence absolue avec la valeur observée à l'instant suivant, $t+1$. La figure 6.4 représente l'observation de cette relation pour les valeurs de niveau d'activité et de fréquence cardiaque. La partie droite des graphes présentés, de même que la partie la plus à gauche du graphe des coûts cardiaques, ne sont pas très significatives car on dispose de peu de données expérimentales associées aux valeurs correspondantes de niveau d'activité et de coût cardiaque. Les graphes présentés montrent que la différence absolue entre deux valeurs successives dans le temps reste effectivement relativement faible, et augmente à peu près linéairement avec les valeurs observées. L'interprétation intuitive de ces résultats peut être que : (1) les faibles valeurs correspondent aux conditions physiologiques de repos, avec par conséquent une variabilité relativement faible dans les valeurs successives observées puisqu'elles doivent rester assez faibles ; (2) les valeurs plus élevées correspondent à des perturbations de ces conditions de repos – par le début d'une activité par exemple – si bien qu'une plus grande variabilité est alors possible dans les valeurs observées, selon le type d'activité réalisée et son déroulement dans le temps.

On définit ainsi un intervalle de tolérance dans lequel peut être définie une valeur en fonction de la valeur précédente. La largeur de cet intervalle de tolérance définissant les valeurs possibles à un instant donné t est approximée par une relation linéaire fonction de la valeur observée à l'instant précédent $t-1$. La valeur observée à l'instant t, $val(t)$, doit satisfaire les contraintes de l'équation (6.1), où $val(t-1)$ représente la valeur observée à l'instant $t-1$, et a et b sont respectivement la pente et l'ordonnée à l'origine de la relation linéaire caractérisant la largeur moyenne de l'intervalle de tolérance.

$$|val(t) - val(t-1)| \leq a \times val(t-1) + b. \tag{6.1}$$

En appliquant cette contrainte temporelle sur les intervalles de temps durant lesquels une même activité est réalisée, les perturbations dans les valeurs observées sont alors regroupées dans le temps. Ceci réduit la variabilité haute fréquence dans les séquences temporelles tout en conservant les caractéristiques basse fréquence de ces séquences.

6.3.3 Implémentation

L'implémentation de cette hypothèse de continuité physique dans le modèle de simulation consiste en la réorganisation temporelle des valeurs de niveau d'activité et de fréquence cardiaque le long des intervalles de temps pendant lesquelles ces valeurs sont générées aléatoirement suivant des distributions similaires.

Considérons une séquence de valeurs de longueur N vérifiant ces hypothèses, et une fonction linéaire d'une valeur à un instant donné qui caractérise l'intervalle de tolérance pour la valeur suivante observée. L'objectif est de permuter les N valeurs de l'intervalle afin de vérifier le plus possible les contraintes sur les valeurs successives admises. Si on considère la $k^{ème}$ valeur de la séquence, $val(k)$, la valeur suivante est sélectionnée dans les $(N-k)$ valeurs suivantes de la séquence. Ces valeurs sont parcourues afin d'identifier la première valeur, la $p^{ème}(p > k)$, qui respecte l'intervalle de tolérance autorisé pour la $(k+1)^{ème}$ valeur : $|val(k) - val(p)| \leq a \times val(k) + b$. Si aucune valeur n'est trouvée qui vérifie cette contrainte, l'intervalle de tolérance est augmenté jusqu'à ce qu'une valeur convienne. La $p^{ème}$ valeur est alors déplacée vers la $(k+1)^{ème}$

position dans la séquence, et les valeurs comprises jusqu'alors entre la $(k+1)^{ème}$ et la $(p-1)^{ème}$ positions sont déplacées entre la $(k+2)^{ème}$ et la $(p)^{ème}$ positions.

Cet algorithme est répété pour toutes les valeurs de la séquence temporelle, de la $1^{ère}$ à la $(N-2)^{ème}$. Chaque réorganisation temporelle est sensible à la première valeur de la séquence considérée. On décide par conséquent de sélectionner la valeur à placer au début de la séquence en fonction de la dernière valeur de l'intervalle précédent si elle existe, selon le même principe de continuité. Cela signifie que seule la toute première valeur de la séquence complète générée par le processus de simulation est conservée selon une génération aléatoire en fonction d'une distribution donnée.

6.3.4 Validation opérationnelle

Il s'agit alors de valider la pertinence des séquences de données produites par la simulation en intégrant les nouvelles contraintes temporelles, par comparaison aux séquences de validation composées de séquences de données expérimentales. La figure 6.5 présente les résultats de simulation obtenus en appliquant les hypothèses de continuité physique – graphe (4) – par comparaison aux séquences générées sans ces hypothèses – graphe (2) – et aux séquences de validation – graphe (3) – dans un contexte de simulation donné – graphe (1). La réorganisation temporelle des valeurs conduit effectivement à une réduction de la variabilité haute fréquence, tout en préservant l'allure générale des séquences. L'effet obtenu est ainsi proche de celui d'un filtre passe-bas, avec cependant les avantages suivants : (1) les valeurs initialement générées par les distributions appropriées ne sont pas modifiées – ce qui préserve les propriétés statistiques des séquences – et (2) la réorganisation des valeurs est réalisée successivement sur des intervalles de différentes longueurs, en fonction des activités observées.

Les séquences de données obtenues ont alors subjectivement un aspect plus proche des séquences de validation (voir Fig. 6.5). La figure 6.6 permet de visualiser plus précisément l'effet de la réorganisation temporelle des valeurs de niveau d'activité et de fréquence cardiaque moyenne, par comparaison avec l'aspect d'enregistrements expérimentaux de même durée qui s'étendent environ sur le même intervalle de valeurs. La figure 6.7 montre par ailleurs que malgré ces réorganisations successives de valeurs, les couples de niveau d'activité – fréquence cardiaque apparaissent globalement identiquement distribués entre les valeurs simulées et celles enregistrées d'une système réel.

Une analyse plus objective des coefficients de corrélation linéaire entre niveau d'activité et fréquence cardiaque montre par ailleurs que la prise en compte d'une contrainte temporelle dans le modèle de simulation permet d'obtenir des résultats plus proches de ceux observés sur les enregistrements d'un système réel (voir Tab. 6.1). Les coefficients de corrélation augmentent en moyenne de 0.05 à 0.15 après la réorganisation temporelle des valeurs, ce qui les rend plus proches des coefficients d'environ 0.6 calculés sur les séquences reconstituées dans le même contexte de simulation – déplacements et postures – à partir des séquences expérimentales.

6.4 Discussion sur le cycle de raffinement de la simulation

L'introduction d'une contrainte temporelle supplémentaire dans la construction du modèle de simulation réduit notablement la variabilité très haute fréquence dans les séquences de données générées, ce qui accroît leur similarité avec les enregistrements réels. Le principal avantage de l'algorithme proposé est qu'il préserve les propriétés statistiques des séquences initiales, puisqu'on ne modifie pas les valeurs et qu'on respecte ainsi les distributions qui les ont générées.

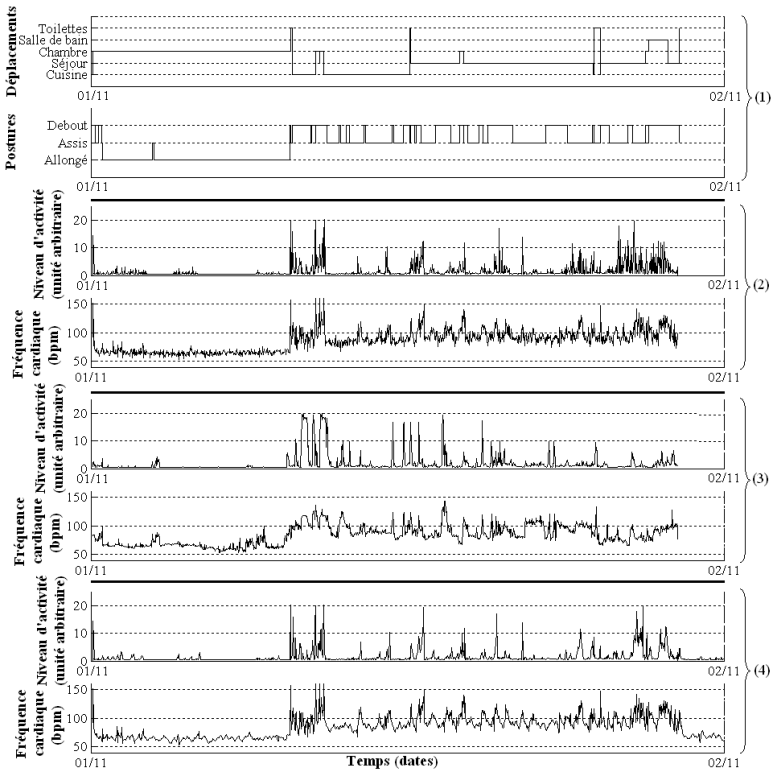


FIG. 6.5 – Séquences temporelles représentatives d'une personne télésurveillée et mettant en évidence les effets d'une réorganisation temporelles.
Les séquences correspondent à une journée d'une personne télésurveillée et présentent, de haut en bas :

(1) Une séquence de déplacements et postures associées générés par le processus de simulation, posant le contexte expérimental ;

(2) Une séquence à deux dimensions de niveaux d'activité et fréquence cardiaque simulés dans ce contexte ;

(3) Une séquence reconstituée à partir d'enregistrements de niveau d'activité et fréquence cardiaque d'un système réel, en insérant le plus possible de longues séquences réelles caractéristiques du contexte de simulation ;

(4) Les données simulées pour le niveau d'activité et la fréquence cardiaque en prenant en compte une contrainte sur les valeurs successives possibles.

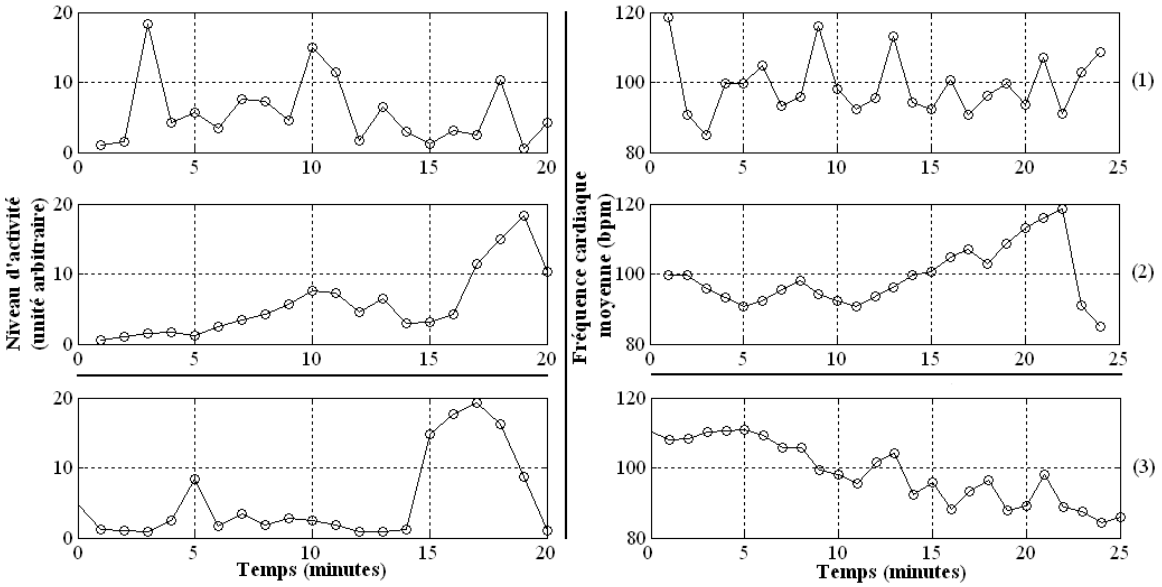


FIG. 6.6 – Effets de la réorganisation temporelle.
Les graphes de gauche et de droite présentent respectivement la réorganisation temporelle de valeurs de niveau d'activité et de fréquence cardiaque, sur une durée de 20 minutes.
De haut en bas, les graphes représentent :

(1) Une séquence de valeurs générée par la simulation ;

(2) Ces mêmes valeurs réorganisées selon le principe de continuité physique ;

(3) Une séquence de valeurs observée environ sur la même échelle de variations et extraite d'une séquence expérimentale.

Un des inconvénients est qu'il peut être considéré comme trop intuitif et quelque peu restrictif, puisqu'il exige que l'intervalle entre deux valeurs successives excède rarement un certain seuil de tolérance. Ce problème est mis en évidence sur les graphes de la figure 6.6 qui concernent le niveau d'activité : un faible pic de valeurs – le premier des deux pics – qui ne pourrait pas être présent dans des séquences simulées à cause de la contrainte temporelle – graphe (2) – est par ailleurs observé sur les séquences de données expérimentales – graphe (3). Plusieurs paramètres tels que les seuils définissant la largeur des intervalles de tolérance peuvent cependant être ajustés afin d'obtenir si nécessaires des résultats de simulation plus précis en terme de variabilité des valeurs observées, tout en préservant les propriétés statistiques.

La nécessité d'amélioration du processus de simulation afin de générer des séquences de données plus précises ou plus réalistes est cependant liée au problème fondamental de simulation de séquences appropriées au contexte de résolution du problème posé. Dans le contexte "haut niveau" défini *a priori* pour la simulation, il s'agit plus de préserver dans la simulation les tendances globales de variation des paramètres et leurs relations conjointes. Dans un contexte où on ne dispose pas de données expérimentales pour une validation plus précise du respect de ces objectifs, c'est l'utilisation des données simulées en entrée d'un système de décision qui permettra de mettre en évidence l'éventuelle nécessité d'un cycle supplémentaire dans la mise en place d'un processus de simulation approprié.

Dans l'objectif de décision, l'ajustement des paramètres des différents sous-modèles de simulation permet de générer des séquences de données représentatives de différents profils de

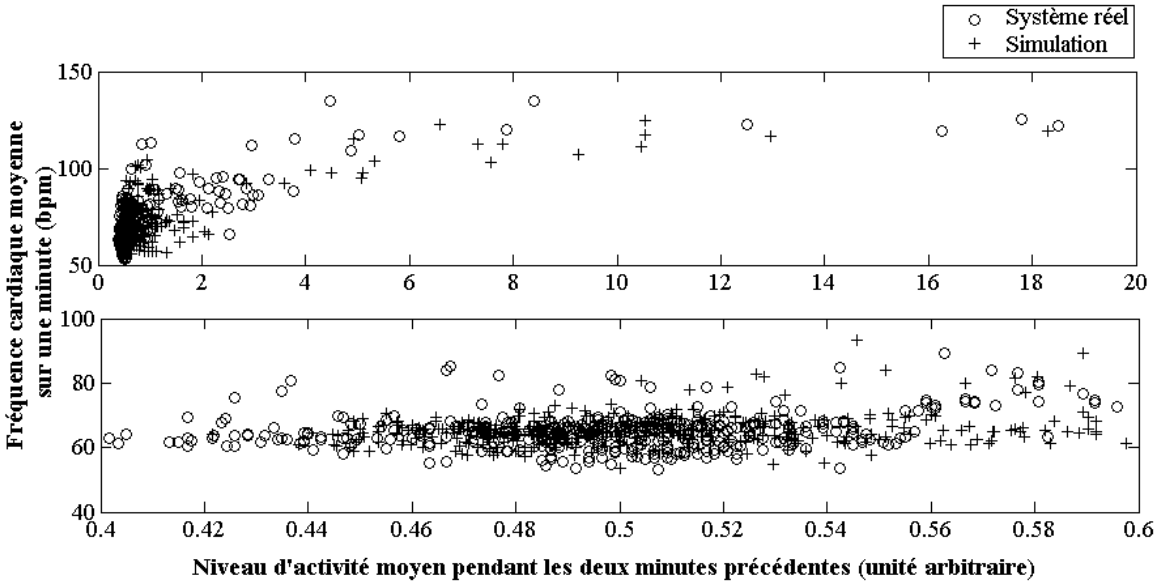


FIG. 6.7 – Distribution des couples de fréquence cardiaque et de niveau d'activité moyen pendant les deux minutes précédent la mesure de fréquence cardiaque.
On présente pour comparaison les couples issus d'un système réel (+) et du processus de simulation (O). Le graphe du bas est un zoom du graphe du haut pour les faibles niveaux d'activité.

personnes et de différents types de situations, tel que présenté sur les graphes de la figure 6.8. D'abord, le graphe (1) présente une séquence à quatre dimensions correspondant à une situation dite "habituelle". Puis, le graphe (2) correspond à une séquence générée à partir des mêmes déplacements mais en modifiant les caractéristiques d'alternance des postures – paramètres de l'automate à états finis définissant les relations entre les postures successives. L'objectif de cette expérimentation est de générer une certaine lenteur dans la réalisation des activités de la vie quotidienne : on le constate effectivement sur la succession des postures, mais on peut également observer les conséquences sur les valeurs générées de niveau d'activité et fréquence cardiaque. Enfin, le graphe (3) présente une séquence générée à partir des mêmes déplacements mais en modifiant les caractéristiques individuelles de variation de la fréquence cardiaque : la tendance générale des successions de valeurs est alors différente, avec en particulier des valeurs de fréquence cardiaque globalement plus faibles et une plus importante arythmie sinusale – variabilité permanente du rythme cardiaque.

Ces graphes mettent bien en évidence la variabilité des données qui correspondent pourtant *a priori* à des profils réels de personnes et de situations. Dans le contexte de la surveillance d'une personne en particulier, il est cependant nécessaire de bien préciser l'interprétation de ces modifications possibles du comportement. Il est en effet fondamental pour la construction d'un système de décision de distinguer en particulier les modifications "normales" de comportement de celles qui sont représentatives d'une situation inquiétante à domicile.

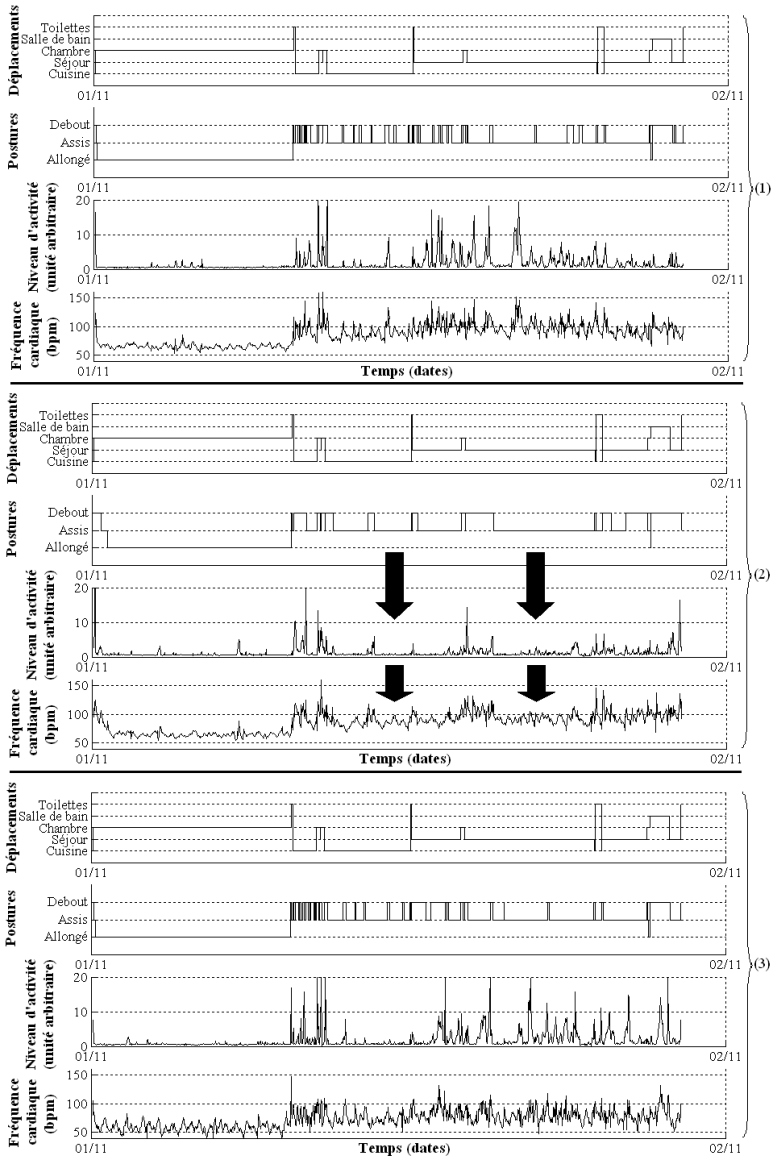


FIG. 6.8 – Simulation de plusieurs profils de personnes et plusieurs types de situation.
Chaque séquence présentée correspondent à une personne télésurveillée pendant une journée et inclut, de haut en bas : (1) déplacements, (2) postures, (3) niveau d'activité moyen et (4) fréquence cardiaque moyenne.
La figure présente, de haut en bas :

(1) Une séquence représentative d'une situation dite "habituelle" ;

(2) Une séquence générée à partir des mêmes déplacements mais en modifiant les caractéristiques d'alternance des postures ;

(3) Une séquence générée à partir des mêmes déplacements mais en modifiant les caractéristiques individuelles de variation de la fréquence cardiaque.

Sur le graphe (2), les flèches mettent en évidence les conséquences de peu d'alternances de postures sur les valeurs de niveau d'activité et de fréquence cardiaque.

Coefficient de corrélation	Simulation	Simulation avec contraintes temporelles	Système réel
k = 0	0.4257	0.5730	0.5829
k = 1	0.5384	0.6142	0.5894
k = 2	0.5839	0.6278	0.5839
k = 0	0.4219	0.5906	0.6094
k = 1	0.5371	0.6413	0.6192
k = 2	0.5756	0.6552	0.6054
k = 0	0.4349	0.5771	0.6199
k = 1	0.5312	0.6201	0.6261
k = 2	0.5797	0.6257	0.6200

TAB. 6.1 – Coefficients de corrélation linéaire entre le niveau d'activité et la fréquence cardiaque. Les coefficients de corrélation linéaire sont calculés entre les séquences correspondant à une journée des valeurs de fréquence cardiaque moyenne et de niveaux moyen d'activité pendant les k minutes précédent chaque calcul de fréquence cardiaque moyenne ($0 \leq k \leq 2$). Le tableau présente les coefficients de corrélation des séquences simulées, avec et sans contrainte temporelle, par comparaison aux séquences reconstituées dans le contexte de simulation à partir d'enregistrements d'un système réel, pour trois expérimentations.

6.5 Simulation de modifications de comportement

6.5.1 Quelles modifications possibles du comportement ?

Un processus de simulation fait partie intégrante du cycle de résolution d'un problème (voir Fig. I.3.2), et doit en particulier générer des données appropriées à l'expérimentation d'un système de décision permettant de répondre à ce problème. Dans notre contexte, on s'intéresse à la construction d'un système de décision pour l'identification du profil comportemental d'une personne à domicile, dans l'objectif de détection des déviations de comportement par rapport à ce profil, significatives d'une situation inquiétante de la personne. Cela impose de savoir discriminer et simuler distinctement des séquences représentatives :

- d'une part des **habitudes de vie** d'une personne, et
- d'autre part de **modifications de comportement**.

Les habitudes de vie d'une personne peuvent en effet être modifiées de façon inquiétante, et les activités quotidiennes réalisées dans de mauvaises conditions : par exemple, faiblesse générale et donc lenteur dans l'exécution des activités, interruptions intempestives par des passages aux toilettes dans le cas de certaines pathologies, etc. Cependant, la modification du comportement d'une personne n'est pas systématiquement significative d'une situation critique. Une activité donnée n'est par exemple jamais réalisée exactement dans les mêmes conditions, et il existe ainsi de larges différences possibles dans les sous-séquences de données qui la représentent.

Le système de décision doit ainsi être capable de différencier les sous-séquences représentatives d'une activité donnée, et celles représentatives de cette même activité réalisée dans une situation inquiétante de la personne. Pour l'expérimentation complète et fiable du système, il est par conséquent nécessaire de savoir simuler explicitement ces deux types de modification possible du comportement, "normale" et inquiétante.

1. **Modification "normale".**

   Les activités de la vie quotidienne ne sont pas toujours réalisées exactement de la même manière, par exemple : une personne va aux toilettes pendant la préparation d'un repas, ou bien elle a oublié de prendre ses vêtements dans la chambre avant d'aller dans la salle de bain, ou encore elle a l'habitude de feuilleter quelques revues dans le séjour en début d'après-midi, mais pas toujours le même nombre de revues ni pendant la même durée.

   Ces comportements variables induisent de larges différences dans les sous-séquences de données représentatives de ces activités. Puisqu'elle est "normale", cette variabilité doit être considérée comme du *bruit* par le système de décision.

2. **Modification inquiétante.**

   Toute dégradation de l'état de santé d'une personne a des répercussions sur la réalisation de ses activités quotidiennes. Par exemple une personne d'habitude plutôt active dès le lever se retrouve alors très peu active et principalement assise au cours de la journée, ou encore elle interrompt souvent ses activités par de longs passages aux toilettes dans le cas de certaines pathologies, etc.

   Les différences alors visibles sur les sous-séquences représentatives des activités quotidiennes doivent induire un changement dans les résultats du système de décision, qui ne doit plus associer ces sous-séquences à la réalisation de l'activité habituelle.

Au niveau du processus de simulation, une modification "normale" de comportement correspond à la variabilité des séquences générées pour la représentation d'une certaine activité selon une configuration donnée des paramètres de simulation. Au contraire, une modification inquiétante intervient sur la définition des paramètres du modèle de simulation. Dans certains cas il est cependant possible qu'il soit difficile de discriminer ces deux types de modifications. Par exemple, une activité peut être normalement interrompue par un passage aux toilettes, mais une "trop grande" fréquence ou durée de ces interruptions peut par contre devenir inquiétante.

6.5.2 Modifications "normales"

La simulation des modifications "normales" du comportement d'une personne nécessite d'abord de bien les définir, puis d'identifier pour chacune les conséquences sur les types de bruit présents dans les sous-séquences représentatives d'une même activité quotidienne. Ces modifications sont définies principalement intuitivement, et avec l'aide d'experts. En particulier, elles ont été discutées avec le Professeur Hélène Pigot. Pour les paramètres quantitatifs uniquement, on dispose par ailleurs d'enregistrements expérimentaux, mais pour lesquels les activités annotées associées sont très peu spécifiques – du type "marcher lentement", "être debout et travailler", etc. Ces annotations peuvent ainsi s'interpréter selon plusieurs types d'activités de la vie quotidienne, ou plusieurs tâches successives d'une même activité, et il n'est par conséquent pas facile d'évaluer à partir de ces données les caractéristiques des modifications "normales" de comportement d'une personne dans la réalisation d'une certaine activité.

La figure 6.9 présente cependant les données expérimentales relatives à l'interprétation de la réalisation d'une *activité de base* pour l'un des sujets de l'expérimentation, lors de chacune des deux journées de surveillance. Il s'agit du *repas du soir*, comprenant *a priori* la préparation en posture "debout" suivie de l'alimentation. On constate par comparaison des deux enregistrements que cette activité n'est *ni toujours réalisée exactement au même moment* – un peu avant 20h ou après 21h – *ni pendant la même durée* – l'alimentation est en particulier plus longue lors de la deuxième journée. Par ailleurs, concernant les variations du niveau d'activité et de la fréquence cardiaque, on peut faire les remarques suivantes sur la variabilité dans les valeurs observées :

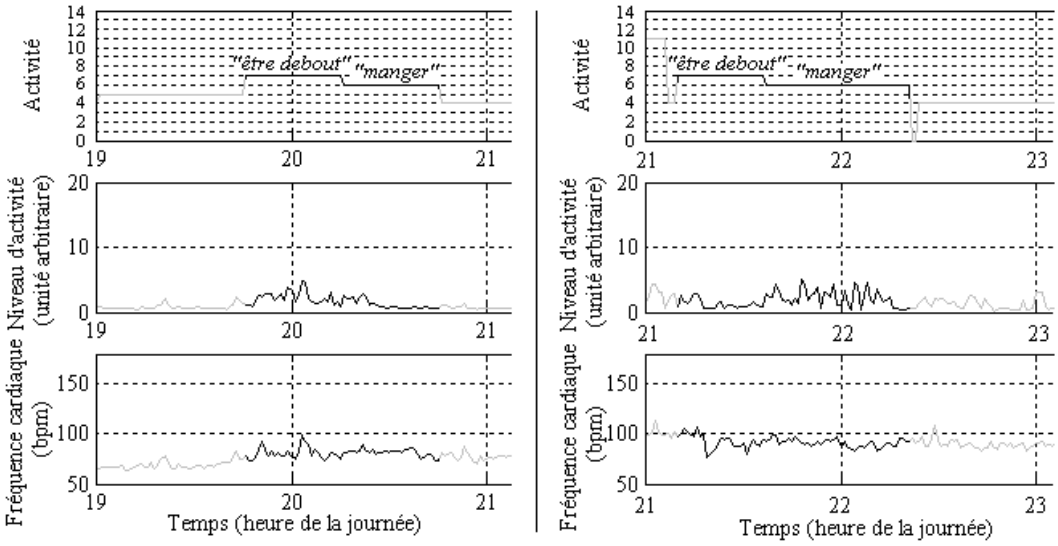


FIG. 6.9 – Observation expérimentale de la réalisation d'une même activité pour un sujet donné. Les deux séquences proposées sont issues respectivement des deux journées d'enregistrement de données expérimentales pour le sujet 5.

- Les paramètres correspondent globalement aux *mêmes tendances de variation à l'échelle de l'activité*, c'est-à-dire à *basse fréquence* – variation stationnaire du niveau d'activité entre 0 et 5 et de la fréquence cardiaque autour de 80 à 90 battements par minute ;
- Les successions de valeurs ont les mêmes caractéristiques à *haute fréquence* – relativement continues, sans changements brusques ou répétés dans les valeurs ;
- Par contre, les tendances locales de variation, à *moyenne fréquence*, diffèrent.

D'après ces constatations et l'intuition des caractéristiques des activités quotidiennes, on identifie alors les modifications "normales" suivantes dans la réalisation récurrente d'une même activité :

1. Instant de réalisation,
2. Interruption,
3. Déformation temporelle,
4. Variabilité dans les valeurs.

Ces modifications sont réalisées *en agissant sur les séquences générées dans une configuration donnée du processus de simulation*, pour l'activité considérée. Étant donné que l'instant de réalisation n'a pas de forte influence sur les séquences observées, on n'en tient pas compte dans la modification "normale" de l'instance d'un motif. C'est le moment d'insertion des instances dans une séquence de données représentative de la personne qui varie alors pour prendre en compte cette caractéristique.

Dans les paragraphes qui suivent, on présente successivement pour chaque type de modification : sa définition, les paramètres nécessaires à sa prise en compte et un exemple d'application sur une séquence représentative de la réalisation d'une activité de référence (Fig. 6.10).

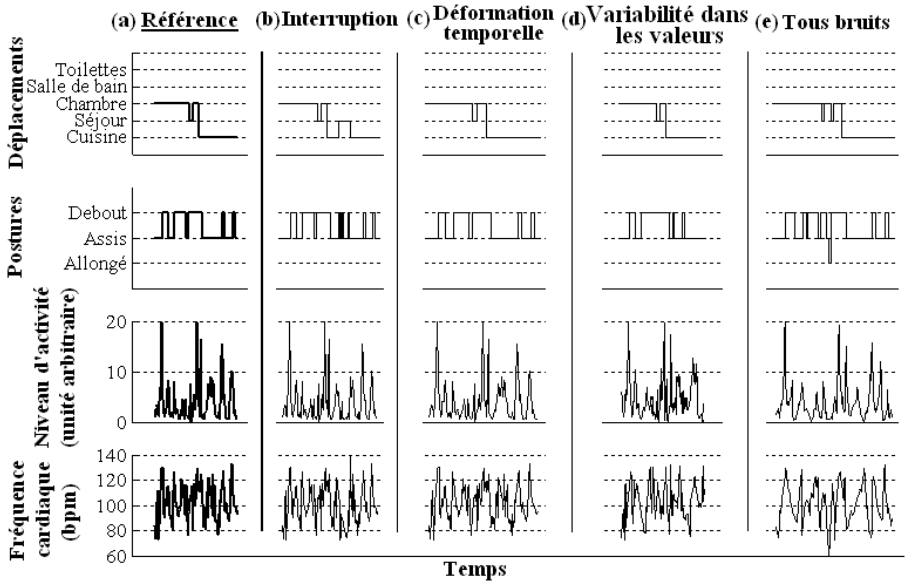


FIG. 6.10 – Simulation de modifications normales de comportement.
Simulation des différents types de bruit normalement présents dans les séquences de données représentatives d'une même activité. Chaque séquence est représentée selon une même échelle de temps et de valeurs, et sur 4 dimensions incluant de haut en bas les paramètres suivants : (1) Déplacements, (2) Postures, (3) Niveau d'activité et (4) Fréquence cardiaque.
Les graphes proposés représentent, de gauche à droite, les séquences suivantes :

- **(a) Séquence de référence,** représentative d'une activité habituelle : *la période du lever le matin* ;
- **(b) Interruption :** l'activité est interrompue par un coup de téléphone dans le séjour ;
- **(c) Déformation temporelle :** l'activité s'étend plus longuement dans le temps ;
- **(d) Variabilité dans les valeurs :** l'activité n'est pas réalisée exactement de la même manière, mais les caractéristiques générales sont préservées ;
- **(e) Tous types de bruits** présents dans la séquence.

Interruption

Définition

Une personne va parfois être amenée à interrompre une tâche planifiée sur une certaine durée pour en effectuer une autre plus urgente comme répondre au téléphone ou aller aux toilettes. Tant que la durée de la *tâche interruptrice* reste bien inférieure à celle de la *tâche dite principale*, on doit pouvoir continuer à reconnaître cette dernière comme étant celle effectuée par la personne à ce moment-là. Il est donc important de tester la robustesse de la caractérisation des activités quotidiennes à la présence d'interruptions.

On remarque alors que l'identification des activités de la vie quotidienne exclut la considération des passages aux toilettes comme une activité principale, puisque cette tâche est le plus souvent interruptrice d'une autre activité. Cette activité étant par ailleurs importante puisque considérée au niveau clinique comme une *activité de base*, sa reconnaissance sera réalisée spécifiquement, à un autre niveau de détail de l'analyse. Par contre, si un passage aux toilettes devient anormalement long, il commence à occulter l'activité principale qui ne sera alors plus forcément reconnue, entraînant indirectement la reconnaissance d'une situation inhabituelle de la personne.

Paramétrisation
Une interruption se traduit par l'insertion, dans une séquence représentative d'une certaine activité, d'une sous-séquence correspondant à une tâche sans rapport direct avec l'activité principale. On la définit ainsi comme se déroulant dans une autre pièce que celle occupée pour l'activité principale au moment de l'interruption.
La simulation d'interruptions nécessite la définition des paramètres suivants :

- **Durée maximum d'une interruption,** comme un rapport de la durée totale de la sous-séquence correspondant à la réalisation de la tâche principale (réel compris entre 0 et 1).
- **Nombre d'interruptions,** si on considère qu'une activité principale peut-être interrompue plusieurs fois (entier positif ou nul).

Un exemple de l'introduction d'une interruption dans la réalisation d'une activité principale est présenté sur le graphe (b) de la figure 6.10, en comparaison au graphe (a) représentant l'activité de référence.

Déformation temporelle

Définition
Une déformation temporelle correspond à la modification de la durée de réalisation d'une certaine activité. Étant donné qu'il s'agit de comportements humains, ce type de modification est obligatoirement et normalement présent dans la vie quotidienne d'une personne. Par conséquent, les séquences représentatives d'une même activité doivent pouvoir correspondre à différentes durées.

Paramétrisation
Une déformation temporelle se traduit par l'étirement ou la compression dans le temps d'une séquence représentative d'une certaine activité. Il s'agit ainsi respectivement d'ajouter ou de supprimer des points de cette séquence. Étant donné qu'une personne ne modifie pas forcément de la même façon la durée de réalisation de chacune des tâches élémentaires composant une activité principale, l'étirement ou la compression n'est pas uniforme dans le temps.

La simulation de déformations temporelles nécessite ainsi la définition des paramètres suivants :

- **Taux de variation de la durée de réalisation de l'activité principale,** défini par un réel positif ou négatif selon qu'il s'agit respectivement d'étirement ou de compression dans le temps de la durée de réalisation de l'activité principale.
- **Durée minimale d'une activité dite principale,** définie par un entier positif ou nul, afin de ne pas compresser jusqu'à un extrême non significatif la réalisation d'une activité.

La durée maximale d'une activité n'est par contre pas paramétrée directement car elle est fortement liée à l'activité réalisée. Elle est cependant implicitement définie par l'intervalle de valeurs considéré pour le taux de variation de la durée. Les instants d'étirement ou de compression sont ensuite déterminés aléatoirement.
Un exemple de déformation temporelles dans la réalisation d'une activité principale est présenté sur le graphe (c) de la figure 6.10, en comparaison au graphe (a) représentant l'activité de référence.

Variabilité dans les valeurs

Définition

En plus de n'être pas toujours réalisée sur une même durée, une activité quotidienne ne sera jamais non plus réalisée exactement de la même manière. Ainsi, même si les tendances globales observées sur les variations des différents paramètres surveillés sont conservées, il existe une grande variabilité possible dans les valeurs.

Ce bruit fondamental à considérer peut être appelé dans notre contexte "bruit moyenne fréquence" car il modifie les tendances de variation à l'échelle des moyennes fréquences – échelle de la réalisation de différentes tâches élémentaires – tout en conservant d'une part la tendance globale de variation basse fréquence représentative de l'activité et d'autre part les caractéristiques haute fréquence de ces variations. Ce type de bruit ne semble cependant pas décrit dans la littérature, et on tente ainsi d'en définir plutôt intuitivement une implémentation réaliste.

Paramétrisation

La variabilité possible dans les valeurs est définie par un seul paramètre qui précise le taux de variabilité autorisé dans les valeurs observées, compris entre 0 et 1. Le taux de variabilité est interprété différemment pour chaque type de paramètre, selon qu'il est *qualitatif* ou *quantitatif*, et toujours de manière à préserver les caractéristiques fondamentales de variation "basse" et "haute" fréquences.

- **Paramètre quantitatif.**
  L'introduction de variabilité se traduit dans le cas de paramètres quantitatifs par une modification des valeurs enregistrées à chaque instant. Les principes de préservation des deux caractéristiques fondamentales de variation sont les suivants :
  - *Variations basse fréquence.*
    Pour préserver les tendances globales de variation, les valeurs sont modifiées selon une distribution normale d'un écart-type donné, mais de moyenne nulle. L'écart-type est défini de façon à générer des valeurs cohérentes par rapport aux données caractérisant l'instance initiale du motif. On décide ainsi d'une valeur d'écart-type définie comme une proportion de l'écart-type observé sur les données initiales, selon le taux de variabilité considéré. Ainsi, dans le cas d'un taux de variabilité nul, l'écart-type associé aux modifications de valeurs est nul et aucune valeur n'est modifiée ; plus le taux de variabilité augmente, plus l'écart-type se rapproche de l'écart-type des données initiales et les valeurs sont largement modifiées.
  - *Variations haute fréquence.*
    Les nouvelles valeurs ainsi générées sont réorganisées selon le principe général de continuité physique, présenté au paragraphe 6.3. Il permet de conserver les propriétés statistiques des données tout en contraignant la succession de valeurs réalistes.
- **Paramètre qualitatif.**
  L'introduction de variabilité se traduit dans le cas de paramètres qualitatifs par une modification des changements successifs de modalité. Dans notre contexte, seules les successions de postures sont concernées à ce niveau, un changement dans les déplacements observés au cours d'une activité est en effet considéré comme une interruption. Par ailleurs, des modifications "normales" de posture sont pertinentes si la personne est debout ou assise ; si la personne est allongée, une modification de posture correspond plutôt à une interruption – interruption de la période de repos ou de sommeil. Enfin, la préservation des caractéristiques de variation "haute fréquence" n'a pas vraiment de sens dans ce contexte, seule la conservation des caractéristiques générales d'alternance des postures est significative.

Concernant les postures d'une personne, il semble intuitivement que la variabilité possible dans les modalités correspond :

– *soit à l'introduction d'alternances de posture supplémentaires*, mais de courte durée, pendant les périodes où la personne maintient longtemps une même posture par rapport aux caractéristiques moyennes d'alternance observées durant l'activité – par exemple, si une personne prend le thé en regardant la télévision, il semble réaliste qu'elle se lève à un moment donné si elle est assise depuis longtemps pour aller chercher le sucrier ou encore le programme télévisé ;

– *soit à la suppression de changements de posture* de courte durée par rapport à la posture principale maintenue pendant la même période – dans le même exemple que précédemment, si la personne se lève pour une courte durée à un moment donné, on peut interpréter cette tâche élémentaire comme secondaire par rapport à l'activité principale et la supprimer à ce moment-là, sachant qu'une alternance de ce type peut être ajoutée à un autre moment d'après le cas précédent.

Le taux de variabilité définit la quantité et la durée des alternances de posture ajoutées ou supprimées. Si ce taux est nul, on est certain de ne modifier aucune alternance de posture.

À l'extrême, si le taux de variabilité est égal à 1, on inverse toutes les alternances de posture dont la durée est inférieure à la durée moyenne d'une posture ; on ajoute également une alternance supplémentaire pour chaque période correspondant à cette durée moyenne et pendant laquelle une même posture est effectivement observée. La durée de l'alternance est sélectionnée aléatoirement au maximum égale à la durée moyenne d'une posture.

Pour des valeurs du taux de variabilité strictement comprises entre 0 et 1, les alternances possibles sont toujours définies selon les critères précédents, mais les alternances effectives sont déterminées en proportion du taux considéré, en terme de leur nombre et de leur durée.

Un exemple d'insertion de variabilité dans les valeurs lors de la réalisation d'une activité principale est présenté sur le graphe (d) de la figure 6.10, en comparaison au graphe (a) représentant l'activité de référence.

Finalement, le dernier graphe de la figure 6.10 illustre la prise en compte de tous ces types de bruit possibles dans la réalisation de la même activité que celle représentée par ailleurs sur la graphe du haut de cette même figure. Une interface pour l'introduction de ces différents types de bruit dans des *motifs*, c'est-à-dire des sous-séquences représentatives d'activités "type" d'une personne dans sa vie quotidienne, est présentée en annexe F.2.2.

6.5.3 Modifications inquiétantes

Les modifications inquiétantes du comportement d'une personne correspondent, au niveau de la simulation, à l'observation de séquences de données qui ne vérifient plus les propriétés de variation des différents paramètres observés. Leur application nécessite alors de définir les dégradations possibles en terme de la définition des paramètres du modèle de simulation.

D'après la description de situations inquiétantes présentée en annexe B et obtenue du Docteur Pierre Rumeau, on constate que l'installation d'une pathologie a plusieurs types de conséquences sur la situation de la personne à domicile, en terme d'activité et de physiologie.

- *La structure des activités habituelles change* – par exemple, beaucoup plus de passages aux toilettes, et éventuellement dans la chambre et la salle de bains dans le cas d'une infection urinaire ;

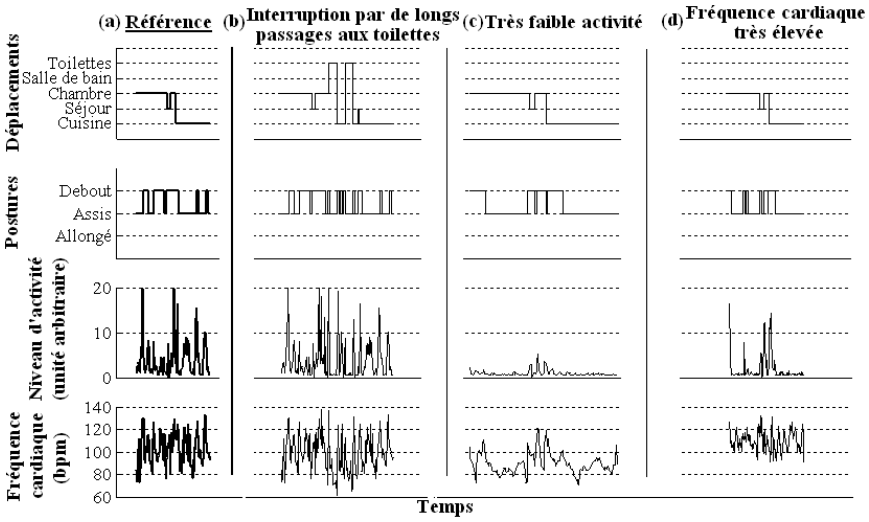


FIG. 6.11 – Simulation de modifications inquiétantes de comportement.
Chaque séquence est représentée selon une même échelle de temps et de valeurs, et sur 4 dimensions incluant de haut en bas les paramètres suivants : (1) Déplacements, (2) Postures, (3) Niveau d'activité et (4) Fréquence cardiaque.
Les graphes proposés représentent de gauche à droite :

(a) Une **séquence de référence** représentative d'une activité habituelle : *la période du lever le matin* ; puis, les autres séquences représentent la même activité avec :

(b) Deux longues interruptions par des passages aux toilettes ;

(c) Un trop faible niveau d'activité ;

(d) Une fréquence cardiaque très élevée indépendamment de l'activité.

- *Les caractéristiques de variation de certains paramètres sont modifiées* – par exemple, dans le cas de l'installation d'une insuffisance cardiaque, accélération de la fréquence cardiaque de repos, jusqu'à 90 ou 100 battements par minute en moyenne au lieu des 70 habituels ;
- *Les relations habituelles entre plusieurs paramètres peuvent également être affectées* – par exemple, dans le cas d'une dépression, les niveaux d'activités enregistrés au cours de la journée sont probablement très faibles quelle que soit la pièce occupée et le moment de la journée.

On identifie finalement trois principales classes de perturbations inquiétantes du comportement, *simulées en agissant sur les paramètres du modèle* :

1. Modification inhérente à la réalisation d'une activité, affectant particulièrement la succession des déplacements,
2. Modification intrinsèque des variations d'un ou plusieurs paramètres,
3. Modification de la relation conjointe de variation de plusieurs paramètres.

Dans le cas de ces modifications inquiétantes, il n'est pas nécessaire de définir des paramètres spécifiques à leur considération. Ce sont les paramètres du processus de simulation qui sont directement modifiés vers des valeurs inhabituelles pour la génération de séquences de données "dégradées" relatives à la personne télésurveillée.

Les graphes de la figure 6.11 illustrent la simulation de certaines des modifications inquiétantes possibles dans la réalisation d'une activité, en référence à la séquence décrivant une réalisation habituelle de cette même activité. On considère de nouveau, pour référence, la période du lever, le matin.

Modification inhérente à la réalisation d'une activité.

La structure même de la réalisation d'une activité peut être modifiée. Au niveau de la simulation, une conséquence est la modification des déplacements observés habituellement au cours de l'activité concernée. En particulier, elle peut alors contenir une ou plusieurs longues interruptions inquiétantes.

Le graphe (b) de la figure 6.11 montre l'exemple de l'interruption d'une activité par deux longs passages aux toilettes.

Modification intrinsèque des variations d'un ou plusieurs paramètres.

Les valeurs observées pour un paramètre ne sont plus en accord avec ses caractéristiques générales de variation. Au niveau de la simulation, il s'agit de définir des valeurs extrêmes ou inattendues pour les paramètres du sous-modèle de simulation correspondant.

Le graphe (c) de la figure 6.11 montre l'exemple de la définition de valeurs extrêmes pour les paramètres du sous-modèle de simulation des postures, correspondant à une faible alternance entre les positions assise et debout au regard de la durée de correspondante. La conséquence est une grande lenteur dans la réalisation de l'activité, mise en évidence sur les variations de la posture, mais également sur les valeurs de niveau d'activité et de fréquence cardiaque qui en dépendent.

Modification de la relation conjointe de variation de plusieurs paramètres.

Les variations relatives de plusieurs paramètres ne sont plus en accord avec les caractéristiques générales observées dans des situations habituelles. Au niveau de la simulation, la structure en cascade du modèle implique une dépendance séquentielle entre les différents paramètres, et donc entre les sous-modèles correspondant à chacun d'eux (voir Fig. 5.1). Il s'agit alors de modifier en des valeurs extrêmes ou inquiétantes les paramètres d'un sous-modèle qui décrivent sa relation avec les résultats de simulation issus de l'un des sous-modèles précédents.

Le graphe (d) de la figure 6.11 montre l'exemple de la rupture d'une relation quasiment linéaire entre la fréquence cardiaque et le niveau d'activité : les valeurs de fréquence cardiaque restent anormalement élevées indépendamment du niveau d'activité.

Pratiquement, l'application de ces modifications inquiétantes est réalisée à partir de l'*interface de base* du processus de simulation définie en annexe F.2.1, en choisissant des valeurs limites ou critiques pour certains paramètres du modèle.

Finalement, la définition et l'application de modifications "normales" et inquiétantes de comportement ont été validées exclusivement intuitivement et à partir de constatations d'experts ou de l'observation de données expérimentales proches du contexte considéré. Elles restent donc à valider plus précisément par comparaison des résultats de la simulation de ces modifications aux enregistrements expérimentaux d'une activité réalisée dans différentes conditions issus d'un système réel. On ne dispose cependant pas pour le moment de tels enregistrements pour ce type de validation.

6.6 Synthèse

La validation opérationnelle du processus de simulation – dans des conditions habituelles ou inquiétantes – a été réalisée avec l'aide d'experts et par des analyses mathématiques et statistiques pour les paramètres qu'il a été possible d'observer en environnement réel. Dans le cadre de la simulation de séquences représentatives d'une situation habituelle d'un individu donné, une première démarche de validation a montré la nécessité de prise en compte, pour la modélisation, d'une contrainte temporelle supplémentaire sur la succession des valeurs des paramètres quantitatifs, au cour des périodes durant lesquelles une même distribution est utilisée pour leur génération. Un cycle de raffinement dans la mise en place du processus de simulation est ainsi réalisé.

Le modèle de simulation proposé et implémenté permet alors la génération de séquences de données qui semblent pouvoir être représentatives des activités d'une personne dans sa vie quotidienne à domicile. Le processus de simulation s'avère complexe, particulièrement au niveau du nombre de paramètres à définir. Les valeurs appropriées des paramètres sont déterminées intuitivement ou guidées par les résultats d'analyse de données expérimentales. Cette complexité s'associe cependant à une grande richesse des profils de personnes et des types de situations qu'il est possible de simuler. En modifiant de façon inhabituelle certains de ces paramètres, on génère en particulier des séquences représentatives de situations inquiétantes d'une personne à domicile.

7

Discussion et Conclusion

Le processus de simulation proposé a été construit selon une **approche hybride et incrémentale** afin de répondre aux objectifs suivants :

- d'une part s'adapter à la complexité d'un contexte hétérogène et multidimensionnel ;
- d'autre part assurer la validation du système au niveau du respect des objectifs prédéfinis dans un contexte donné, et de sa pertinence plus globale pour la résolution d'un problème.

La discussion s'articule ainsi autour de deux points fondamentaux concernant d'abord la **démarche de simulation** dans le respect de la complexité et des objectifs du contexte de la télésurveillance médicale, puis sa **vision plus globale dans le cycle de résolution du problème** de construction du profil de comportement d'une personne pour assurer la détection de situations critiques.

La démarche de simulation dans le contexte de la télésurveillance médicale à domicile

La discussion sur la construction d'un processus de simulation concerne successivement : (1) la validation des ensembles de données expérimentaux, (2) la construction du modèle de simulation, et (3) les résultats de son expérimentation.

Données expérimentales

Les données expérimentales utilisées pour la construction et la validation du processus de simulation ne sont pas complètement appropriées puisqu'elles ont été enregistrées sur des sujets d'une classe restreinte de la population (hommes et femmes jeunes, entre 20 et 30 ans, et en bonne santé), très peu représentative de la population ciblée par les applications de télésurveillance médicale à domicile (personnes âgées ou handicapées, présentant des risques moteurs et cognitifs). Par ailleurs, les informations concernant les activités réalisées par le sujet ont été fournies par le sujet lui-même qui en a pris note au cours du déroulement de la journée. Ces annotations sont donc plutôt subjectives, et il est par exemple difficile de distinguer objectivement, de la même façon pour chaque sujet, une marche "lente" d'une marche "rapide". Enfin, les données enregistrées à la suite de la réalisation d'une activité de forte intensité risquent d'être perturbées par la nécessité d'un temps de récupération, même si les données suivantes les plus proches de la fin de l'activité intense ont été supprimées de l'ensemble des données expérimentales considérées. Toutes ces raisons font que le modèle de simulation risque déjà d'être biaisé.

Construction du modèle

D'autres remarques peuvent être avancées concernant la construction du modèle de simulation, particulièrement pour ce qui concerne la simulation des valeurs de la fréquence cardiaque

moyenne. L'intensité des activités réalisées et les postures de la personne ont été considérées comme les facteurs d'influence principaux des valeurs de fréquence cardiaque moyenne. Cette hypothèse est probablement vraie en général, mais beaucoup d'autres facteurs peuvent cependant également avoir une influence. C'est le cas par exemple de facteurs externes tels que la température extérieure, la prise de médicaments, le stress, la sonnerie du téléphone, le claquement d'une porte, et de facteurs internes tels que la productivité de l'organisme (activité végétative par exemple). Les contextes des données expérimentales et de la simulation ne permettent pas de prendre en compte ce niveau de précision dans la définition des facteurs d'influence des différents paramètres considérés.

Par ailleurs, certaines méthodes utilisées lors de l'analyse des données expérimentales, telles que l'estimation d'un rythme circadien sinusoïdal et la normalisation des valeurs expérimentales à partir de cette estimation des valeurs de repos, sont également loin d'être parfaites puisqu'elles reposent sur plusieurs hypothèses simplificatrices. La définition d'une contrainte temporelle dans la succession des valeurs de niveau d'activité et de fréquence cardiaque est également très intuitive, même si confirmée par les données expérimentales, et ne permet pas une modélisation très précise, au niveau de chaque activité réalisée, des contraintes de variation des paramètres quantitatifs dans le temps.

Validation opérationnelle

Les imprécisions de la modélisation, discutées précédemment, résultent en la génération de séquences de données reflétant surtout les tendances de variation des différents paramètres, ce qui semble finalement plutôt approprié au contexte de simulation "haut niveau" défini *a priori*. Comme cela avait été précisé au chapitre 2, il ne s'agit ni de sur-simplifier le système, ni de construire un modèle particulièrement complexe qui génère des données qu'on espère le parfait reflet d'un comportement réel, alors que ce niveau de précision n'est pas obligatoirement requis dans un contexte donné. On préfère ainsi une démarche incrémentale qui apporte progressivement de la complexité supplémentaire au modèle, et seulement si nécessaire, tel que suggéré par Chwif [32].

Dans notre contexte, un nouveau cycle de raffinement du processus de simulation aurait pour objectif de prendre en compte des connaissances et une modélisation plus précises des paramètres pour inférer une plus grande précision dans les valeurs simulées. Avant de procéder à un raffinement éventuel de la simulation, il est ainsi nécessaire d'avancer dans la définition et l'expérimentation d'un processus de décision afin de déterminer si un niveau plus fin de détail est effectivement nécessaire pour répondre aux objectifs de résolution du problème. La conséquence éventuelle est la redéfinition du contexte et des objectifs de la simulation pour la génération de séquences à un plus grand niveau de détail et de précision.

Cependant, étant donné le manque de données expérimentales appropriées au contexte de la télésurveillance médicale à domicile, la validation des résultats de la simulation est en grande partie intuitive et, par conséquent, incomplète même au faible niveau de précision exigé. En particulier, on ne dispose pas d'enregistrements réels associés aux déplacements et postures d'une personne dans sa vie quotidienne. Il est par conséquent difficile d'avoir une idée intuitive, même approximative, et même "haut niveau", du réalisme des séquences générées.

Malgré cela, la démarche de simulation a eu pour objectif l'intégration des caractéristiques de réalisation des activités quotidiennes d'une personne, dans le respect des variations conjointes des différents paramètres simulé. Dans ce contexte, on a en particulier proposé pour la modélisation et la validation :

– d'une part une **approche hybride** intégrant (a) **différents types de connaissances** liées à la diversité des sources d'information disponibles (voir Fig. 3.2), mais aussi (b) **dif-**

férentes techniques de modélisation selon le type des paramètres simulés – quantitatifs ou qualitatifs ;
– d'autre part un principe de **simulation en cascade** pour préserver la complexité de la relation entre les paramètres étudiés (voir Fig. 5.1).

La conséquence attendue est la génération, pour un individu donné, de séquences multidimensionnelles cohérentes, corrélées ou dépendantes, et représentatives d'un certain rythme – c'est-à-dire contenant un ensemble de sous-séquences fréquentes représentatives de certaines activités habituelles. La modification des paramètres du modèle de simulation permet de faire varier ces caractéristiques de variation et de générer des données représentatives d'une part de différents profils de personnes, et d'autre part de différents types de situations. On espère ainsi se rapprocher des conditions réelles d'enregistrement de séquences de données représentatives de la vie quotidienne à domicile, appropriées en particulier à l'expérimentation d'un système de décision pour la construction du profil comportemental d'une personne et la détection des situations critiques.

La simulation dans le cycle d'apprentissage d'un profil comportemental

La simulation de données appropriées au contexte de la télésurveillance médicale à domicile rend possible l'expérimentation d'un système d'apprentissage des habitudes de vie d'une personne dans l'objectif de détecter des déviations inquiétantes de comportement. Un des intérêts de la simulation est en effet de disposer de larges ensembles de données représentatives d'une grande variété de situations pas toujours faciles à observer dans la réalité. Mais la simulation permet également d'avoir une meilleure connaissance *a posteriori* des paramètres étudiés, ce qui a des conséquences sur la définition d'un système de décision approprié sur ces données.

Dans ce paragraphe, nous discutons ainsi les apports de la simulation pour la construction du système de décision sur les habitudes de vie d'une personne à domicile. On explique d'abord pour quelles raisons le système de décision, présenté dans la partie suivante, ne doit pas *a priori* être basé directement sur le modèle de simulation. Puis, on décrit les connaissances *a posteriori* liées à la construction du modèle, qui confirment la nécessité d'un système de décision sur les habitudes de vie d'une personne qui soit d'une part **non basé directement sur le modèle de simulation**, et d'autre part complètement **non supervisé**.

Un système de décision a priori non basé directement sur le modèle de simulation

Dans notre contexte, le processus de simulation et les résultats de son expérimentation ne sont pas complètement validés puisqu'ils ne l'ont pas été dans un contexte réel de télésurveillance médical à domicile. Il n'est ainsi pas véritablement possible de fonder la décision concernant les habitudes de vie d'une personne à domicile directement sur le modèle de simulation proposé.

Par ailleurs, le meilleur ensemble exhaustif de paramètres à considérer pour cette étude ne peut pas non plus être déterminé sans l'étude des enregistrements d'un système réel. Il s'agit d'identifier l'ensemble des paramètres les plus représentatifs du comportement d'une personne, mais aussi les plus discriminants vis à vis de la détection de situations critiques. Cette démarche implique d'expérimenter le système de décision sur plusieurs ensembles de paramètres, en affinant progressivement leur choix en fonction des performances avérées du système. On souhaite ainsi avoir la possibilité d'étendre et de modifier facilement l'ensemble de paramètres considérés sans influence sur la définition du système de décision. Il est par conséquent préférable de disposer d'une méthode générique de décision, indépendante d'un modèle de variation des paramètres analysés.

Connaissances a posteriori utiles à la décision

Malgré cette nécessité d'indépendance de la simulation et de la décision, la modélisation des paramètres peut néanmoins guider en partie la décision. La mise en place du processus de simulation apporte en effet des connaissances *a posteriori* sur le contexte de la télésurveillance médicale qu'il est utile de prendre en compte dans la définition d'une méthode de décision.

Tout d'abord, la complexité des relations entre les paramètres qui peuvent être observés à domicile et des facteurs d'influence à prendre en compte montre qu'il est très difficile de construire un modèle précis et exhaustif de leurs variations. Cela confirme la pertinence de disposer d'**un système de décision qui ne soit pas basé sur un modèle** de comportement explicite décrivant les variations habituelles des différents paramètres surveillés.

La validation du modèle conceptuel a également renforcé l'*a priori* de la **spécificité de comportement de chaque individu** dans la vie quotidienne, au niveau des activités réalisées (variations diverses des niveaux d'activité par exemple), mais aussi au niveau physiologique (variations individuelles de la fréquence cardiaque, au repos d'une part, et en réaction à l'effort d'autre part). La conséquence pour le système de décision est la nécessité d'un **apprentissage complètement non supervisé** du profil comportemental d'une personne.

La partie suivante de ce travail concerne alors la définition et l'expérimentation d'une méthode générique et non supervisée d'apprentissage du profil comportemental d'une personne à domicile, dans l'objectif de détecter l'occurrence de situations inquiétantes.

Troisième partie

Système de décision

1

Introduction

Dans le cadre de la télésurveillance médicale et dans l'objectif de détecter les situations critiques de personnes à domicile, on souhaite mettre en place un système d'**apprentissage du profil comportemental d'une personne** dans la réalisation de ses activités de la vie quotidienne. Toute modification du comportement habituel, tant en terme des activités réalisées que du comportement physiologique, porte en effet la forte présomption d'une dégradation de l'état de santé de la personne. Il s'agit ainsi d'**identifier** et de **classer** les **activités habituelles** d'une personne à partir des données collectées d'un ensemble de capteurs installés au domicile.

Nous présentons dans cette introduction d'abord les données et connaissances disponibles, puis la formulation du problème d'apprentissage à partir de ces données, et enfin les axes de recherche appropriés à sa résolution et l'originalité de l'approche proposée.

Données et connaissances disponibles pour l'apprentissage

L'apprentissage des activités régulières d'une personne est réalisé à partir des données enregistrées en continu par un ensemble de capteurs à domicile. Ces données sont ainsi de différents types en terme :

- **de l'aspect du comportement qu'elles décrivent** : les capteurs enregistrent des informations qui concernent la physiologie, l'activité et/ ou l'environnement de la personne ;
- **de leurs caractéristiques intrinsèques** : les données collectées peuvent être quantitatives (fréquence cardiaque, etc.) ou qualitatives (posture, etc.).

Il s'agit ainsi d'analyser des **séquences multidimensionnelles de données hétérogènes** relatives au comportement de la personne télésurveillée.

Étant donné le manque de données expérimentales, on utilise dans une première phase d'étude comme données d'apprentissage les séquences générées par le processus de simulation dans un contexte "normal" de vie d'une personne. L'avantage de l'exploitation d'un processus de simulation réside également dans la possibilité de génération de données représentatives de différents profils de personnes et types de décision pour l'expérimentation du système. On étudie ainsi des enregistrements à quatre dimensions composés des séquences de déplacements et postures (paramètres qualitatifs), et des séries de niveaux d'activité et fréquence cardiaque (paramètres quantitatifs). Cette sélection ne correspond *a priori* pas aux paramètres à considérer pour une étude optimale, mais constitue un ensemble pertinent en terme de l'observation des activités quotidiennes pour la détection de situations critiques (voir paragraphe II.4.1).

Les connaissances *a priori* impliquées dans l'apprentissage d'un profil comportemental sont et même doivent être peu nombreuses, compte tenu d'une part de la variabilité des comporte-

ments observés en fonction des personnes télésurveillées, et d'autre part du manque d'études sur l'ensemble optimal de paramètres à considérer pour caractériser le comportement d'une personne et détecter les situations inquiétantes à domicile. Les conséquences sur les caractéristiques du système d'apprentissage sont les suivantes :

- **Apprentissage non supervisé**, pour tenir compte des spécificités individuelles d'un profil comportemental ;
- **Système générique d'apprentissage**, indépendant du modèle des données étudiées, afin de pouvoir ensuite étendre et/ ou modifier l'ensemble des paramètres considérés dans cette première étude, pour l'optimiser et améliorer la détection des situations critiques d'une personne. Idéalement, les paramètres considérés sont ceux les plus caractéristiques du comportement d'une personne, mais aussi les plus sensibles à toute dégradation de son état de santé.

Formulation du problème d'apprentissage

L'apprentissage des activités régulières d'une personne peut être assimilé à un système de décision sur l'appartenance ou non des données successives de l'apprentissage, enregistrées par les capteurs au domicile, à la réalisation d'une activité habituelle. Les sous-séquences de ces données effectivement associées à un comportement régulier sont alors considérées comme les instances d'un **motif** : chaque motif représente ainsi une activité type réalisée habituellement par la personne et est supposé avoir des instances récurrentes dans la séquence d'apprentissage.

Les caractéristiques de ces motifs dépendent du contexte de l'apprentissage, tel qu'explicité sur le schéma de résolution de tout problème de décision (voir Fig. I.3.2). Notre contexte d'apprentissage des habitudes de vie quotidienne d'une personne, détaillé dans le chapitre I.3, impose en particulier les contraintes suivantes :

- **Extraction de motifs multidimensionnels** à partir des séquences d'apprentissage, afin de préserver la complexité du problème dans la considération conjointe de l'ensemble des paramètres surveillés par les capteurs installés dans l'habitat ;
- **Recherche de motifs "haut niveau"**, pour s'adapter à la grande variabilité possible, au "bas niveau" des données issues des capteurs, des séquences de données correspondant à la réalisation d'une même activité de la vie quotidienne. Il existe ainsi une *grande quantité de bruit* possible entre des séquences temporelles qui sont pourtant les instances d'un même motif. Cette caractéristique des motifs impose au système d'apprentissage de prendre particulièrement en compte les *imprécisions* possibles entre les instances d'un même motif, la présence éventuelle d'*interruptions*, ou encore de *déformations* et de *translations dans le temps*.

Axes de recherche et originalité de l'approche proposée pour la résolution

L'apprentissage du profil comportemental d'une personne à domicile se rapporte tout à fait à une problématique de **fouille de données temporelles** puisqu'il est réalisé dans un *contexte non supervisé* et à partir d'un large ensemble de données temporelles. La forte augmentation des possibilités de collecte et de stockage de données de différents types, dans le contexte de toutes sortes d'applications, a particulièrement encouragé depuis ces dernières années les chercheurs à s'intéresser à la découverte de motifs, de tendances caractéristiques, ou des particularités de constitution de larges ensembles de données. C'est ce qu'on appelle la *découverte de connaissances – "knowledge discovery"* – ou encore la *fouille de données – "data mining"*. L'objectif est d'extraire

des informations implicitement contenues dans les données analysées, qui ne sont pas évidentes à identifier *a priori*, et jusqu'alors inconnues [40]. Par ailleurs, de nombreuses applications ayant pour objectif la prédiction de comportements ou l'aide au diagnostic concernent particulièrement des ensembles de *données temporelles* [96]. C'est ce qui a motivé le développement du domaine de recherche correspondant : la *fouille de données temporelles – "time-series mining"*. L'objectif de la fouille de données temporelles dans notre contexte est l'apprentissage de motifs, c'est-à-dire l'identification et la classification des sous-séquences récurrentes dans les séquences temporelles étudiées. Les motifs extraits des séquences d'apprentissage doivent être représentatifs des activités habituelles de la personne dans sa vie quotidienne.

Un système d'apprentissage approprié doit également prendre en compte une grand *écart entre le niveau des données analysées et celui de la décision*. On recherche en effet l'extraction de motifs "haut niveau" représentatifs des activités régulières à partir des données "bas niveau" issues des capteurs. La méthode proposée doit ainsi comprendre plusieurs niveaux de **représentation** et de **fouille des données** des capteurs, pour aboutir à des informations dont le niveau de détail est approprié à l'évaluation de la situation de la personne.

Une originalité de ce travail par rapport aux recherches déjà effectuées dans le domaine de l'extraction de motifs temporels est la prise en compte de séquences de *données multidimensionnelles et hétérogènes*, dans l'objectif d'extraction de *motifs multidimensionnels*. Par conséquent, il est en particulier nécessaire de définir une **mesure de similarité** adaptée à la comparaison de séquences de ce type.

Dans cette partie, nous décrivons ainsi d'abord, dans le chapitre 2, un état de l'art des méthodes de fouille de données temporelles, de représentation et de mesure de similarité. Le chapitre 3 est consacré à la présentation de la méthodologie de recherche de motifs temporels. Les chapitres 4 et 5 détaillent ensuite la mesure de similarité puis l'approche proposées pour l'extraction de motifs temporels, multidimensionnels et hétérogènes. Enfin, le chapitre 6 présente les résultats de leur expérimentation sur les séquences de données générées dans le contexte de la télésurveillance médicale par le processus de simulation décrit dans la partie II.

2

État de l'art

La recherche non supervisée de motifs "haut niveau" à partir de données enregistrées en continu à "bas niveau" au regard de la décision concerne la fouille de données temporelles, les méthodes de représentation et les mesures de similarité appropriées à la considération de séquences multidimensionnelles et hétérogènes. Nous décrivons ainsi successivement l'état de l'art de ces trois grands axes recherche.

2.1 Fouille de données temporelles

La **fouille de données temporelles** se rapporte à l'analyse non supervisée des variations dans le temps d'un ou plusieurs paramètres. L'objectif est d'extraire de nouvelles informations implicites contenues dans ces données qui concernent les caractères des paramètres étudiés. Ces dernières années, depuis l'essor des capacités de collecte et de stockage de données, la fouille de données temporelles a été le sujet de nombreuses recherches, dans des domaines d'application variés qui concernent particulièrement la prédiction et l'aide au diagnostic. Les données temporelles sont souvent désignées sous les noms **séquences temporelles** – *"temporal sequences"* – ou **séries temporelles** – *"time-series"*. D'après Antunes *et al.* [3], les séquences temporelles sont relatives à la succession de symboles d'un alphabet, alors que les séries temporelles concernent des éléments continus, à valeurs réelles. Étant donné qu'on s'intéresse ici à des séquences multidimensionnelles et hétérogènes de données variant au cours du temps, on les nommera donc indifféremment séquences ou séries temporelles.

Une vue d'ensemble de la fouille de données temporelles est présentée dans des articles tels que [3] et [96]. Les algorithmes développés visent à extraire les caractéristiques temporelles importantes de ces séquences telles que similarités, tendances, périodicité, dans un objectif de description ou de prédiction [77]. La découverte de caractères particuliers des séquences temporelles est particulièrement utile à la représentation de ces séquences [48], aussi bien qu'à des tâches d'apprentissage telles que la découverte de règles d'association [34, 50], la classification supervisée [55], ou non supervisée [110]. Dans notre contexte, la fouille de données temporelle est utile à l'identification des sous-séquences récurrentes dont la classification indique finalement les *motifs* (classes de sous-séquences récurrentes) présents dans la séquence analysée.

Étant donnée la quantité exponentielle de sous-séquences qu'il est possible de constituer à partir d'une séquence initiale, une première étape de fouille de données temporelles consiste souvent en une **fouille de caractères** permettant l'identification des sous-séquences les plus susceptibles de correspondre effectivement aux instances de motifs. Ces sous-séquences sont dénommées les *caractères* des données temporelles.

Enfin, la fouille de données temporelles, incluant la fouille de caractères, peut être réalisée dans un objectif d'apprentissage. Elle nécessite alors une méthode de classification non supervisée des caractères pour l'identification effective des motifs.

2.1.1 Fouille de caractères dans les séquences temporelles

La fouille de données temporelles a parfois un rôle de prétraitement d'une séquence temporelle initiale pour en extraire le meilleur ensemble de sous-séquences à placer alors en entrée d'algorithmes d'apprentissage [68] pour des tâches telles que la découverte de règles d'association ou la classification. Par analogie avec les domaines d'analyse de données non séquentielles, et en raison du nombre exponentiel de sous-séquences qu'il est possible de constituer à partir d'une séquence initiale, la fouille de séries temporelles dans un objectif d'apprentissage est alors souvent référencée comme *fouille de caractères – "feature mining"* [62, 67]. Les caractères sont alors toutes les sous-séquences extraites de la séquence temporelle analysée. Dans le cas d'une analyse de données non séquentielles, la sélection de caractères correspond en effet à l'identification d'un espace optimal de taille m à partir de l'espace complet des caractères observés, de dimension d, où idéalement on a $m \ll d$. Dans les domaines temporels, la sélection de caractères a ainsi pour objectif la sélection du meilleur sous-ensemble de caractères temporels, c'est-à-dire l'identification des sous-séquences de la séquence initiale qui portent le plus d'information sur les spécificités des données analysées [67].

Une telle étape est particulièrement intéressante pour améliorer les performances d'un système d'apprentissage quand les séries temporelles en entrée du système contiennent à la fois des sous-séquences représentatives de la séquence complète et d'autres qui ne le sont pas, comme dans le contexte de [48]. Selon [67], les critères qui guident cette sélection de caractères temporels doivent prendre en compte les éléments suivants : les caractères doivent être fréquents, distinctifs d'au moins une classe, et non redondants. Il existe généralement deux grandes classes de méthodes de découverte des caractères de séries temporelles :

- **Méthodes supervisées :** les caractères recherchés sont décrits par un ensemble de connaissances *a priori*, ou similaires à ceux d'une séquence connue (par exemple, dans [2, 55, 65]).
- **Méthodes non supervisées :** les caractères sont recherchés sans connaissances *a priori* sur les régularités contenues dans les séquences étudiées (par exemple, [28, 49, 48]) : c'est la *fouille de caractères*.

D'après ces définitions, un exemple de fouille de caractères est alors l'identification des sous-séquences récurrentes dans une séquence initiale, comme une première étape de fouille de données avant leur classification. Dans ce cas particulier d'une recherche non supervisée de sous-séquences similaires, fréquemment représentées dans une série temporelle, et sur lesquelles on n'a aucun *a priori*, Lin *et al.* [69] ont introduit la notion de *motifs de séries temporelles* – "*time-series motifs*". Ces motifs sont aussi nommés "formes de base" – *primitive shapes* [34] – ou "caractère temporel fréquent" – *frequent temporal patterns* [50] – selon les auteurs. L'identification de sous-séquences récurrentes est réalisée par des techniques variées selon les applications, en fonction :

- **des spécificités des séquences temporelles** étudiées d'une part : séries ou séquences temporelles, à une ou plusieurs dimensions, etc. ;
- **des caractéristiques des motifs recherchés** d'autre part : variabilité possible dans les valeurs, transformations autorisées entre différentes instances d'un même motif, etc.

Dans le cadre de l'extraction non supervisée de motifs temporels, à une ou plusieurs dimensions, à partir de séquences de données à valeurs réelles, des techniques telles que l'utilisation de réseaux de neurones récurrents, d'automates à états finis, de projections aléatoires, ont été expérimentées.

Par exemple, Hong *et al.* ont implémenté un processus d'apprentissage à partir d'un réseau de neurones récurrent pour l'extraction non supervisée de motifs temporels multidimensionnels [49]. Cette méthode s'est cependant avérée peu efficace en présence de bruit entre les instances d'un même motif. L'utilisation d'automates à états finis [48] donne de meilleurs résultats dans ce contexte, mais ces derniers s'effondrent avec l'augmentation de la dimension considérée pour les séries temporelles, et par conséquent du nombre d'états à prendre en compte dans l'automate.

Toujours dans le contexte d'extraction non supervisée de motifs temporels, Chiu *et al.* [28] ont expérimenté un algorithme efficace fondé sur la comparaison des projections aléatoires, sur une dimension inférieure, des sous-séquences d'une séquence étudiée. Cela signifie qu'on ne considère pour leur comparaison qu'une partie des points ou symboles définissant chaque sous-séquence. Cette méthode a été initialement proposée par Buhler et Tompa [17] pour la recherche de motifs dans des séquences de nucléotides. L'application de cet algorithme à la représentation discrète de séquences de données réelles est particulièrement intéressante car l'algorithme permet d'extraire rapidement des résultats approximatifs sur les motifs présents dans les séquences réelles étudiées, et en particulier sur leur localisation dans la séquence initiale. Il est par ailleurs efficace même en présence de bruit et de symboles "imprévus" ou non significatifs dans une sous-séquence pourtant représentative d'un motif. L'expérimentation de cet algorithme proposée dans [28] ne s'intéresse cependant qu'à la recherche de motifs dans une série temporelle à une dimension, et la problématique d'extraction de motifs multidimensionnels et hétérogènes n'est ainsi pas considérée. La méthode de projections aléatoires telle qu'elle a été implémentée dans [28] n'autorise pas non plus de déformation temporelle entre les instances d'un même motif.

Dans notre contexte, la fouille de caractères a pour objectif l'identification des sous-séquences récurrentes dans une séquence d'apprentissage, caractéristiques des habitudes de vie de la personne. On effectue ensuite leur classification afin de mettre en évidence les activités "types" réalisées dans la vie quotidienne. On appelle alors **tentatives de motifs** les sous-séquences récurrentes, et **motifs** les classes de ces sous-séquences.

L'utilisation d'un algorithme réalisant des *projections aléatoires* [17, 28] des sous-séquences de la séquence étudiée comme une étape d'identification des sous-séquences récurrentes est particulièrement intéressante. Cette technique est en effet complètement non supervisée, et permet l'extraction de sous-séquences pour lesquelles la présomption de récurrence est forte, avec une bonne résistance au bruit – valeurs inattendues ou variabilité dans les valeurs. Il est cependant nécessaire de s'adapter aux séquences multidimensionnelles et hétérogènes, et à la considération de différentes durées possibles pour les instances de motifs.

2.1.2 Classification des caractères

La fouille de caractères a été présentée comme une première étape de fouille de séries temporelles dans un objectif d'*apprentissage*. Il s'agit par exemple d'extraire les sous-séquences récurrentes dans une séquence initiale, puis de les classer afin d'identifier finalement les motifs – classes de sous-séquences récurrentes – caractérisant cette séquence. Finalement, si le système d'apprentissage est complètement non supervisé, il peut être assimilé dans sa globalité à la résolution d'un problème de fouille de données temporelles, constituée alors de deux étapes successives : (1) d'une part la fouille de caractères, et (2) d'autre part l'application d'un algorithme de classification non supervisée sur ces caractères.

Dans les travaux mentionnés sur la recherche non supervisée de motifs, la fouille de données temporelles est effectivement souvent réalisée en plusieurs étapes successives permettant une identification d'abord grossière, puis progressivement plus fine de motifs, dans l'objectif d'extraire puis de classer les sous-séquences récurrentes [28, 49, 48, 65]. Le principe est de limiter

le nombre de sous-séquences considérées afin de réduire par conséquent la dimension de l'espace de classification. Seules les sous-séquences correspondant le plus probablement aux instances de motifs sont prises en compte à l'étape de classification, réalisée selon un seuil sur une mesure de similarité par exemple.

Les méthodes de classification des caractères correspondant aux sous-séquences récurrentes sont cependant souvent rapidement abordées dans la littérature sur l'extraction de motifs à partir de données temporelles. Pourtant, selon la méthode de fouille de caractères utilisée, les sous-séquences récurrentes identifiées peuvent être redondantes et/ ou se recouvrir mutuellement. Il convient alors de définir une méthode d'identification des sous-séquences récurrentes les plus pertinentes afin de vérifier les critères définis par Lesh *et al.* [67] : les caractères doivent être fréquents, distinctifs d'au moins une classe, et non redondants.

Leur classification nécessite ensuite la définition d'une mesure de similarité appropriée aux séquences analysées, ainsi que le précisent Antunes *et al.* [3]. Ces mesures de distances utiles à la classification peuvent être calculées à partir de la représentation ou de l'abstraction des sous-séquences récurrentes identifiées. Cependant, elles sont aussi souvent appliquées sur les données initiales, simplement prétraitées, correspondant aux sous-séquences récurrentes, afin d'affiner à cette étape l'identification des motifs [28, 69]. Cette démarche impose alors de disposer d'une mesure de distance appropriée à la comparaison de séquences des données initiales, c'est-à-dire dans notre contexte à des séquences temporelles multidimensionnelles et hétérogènes.

2.2 Représentation des séquences temporelles

Pour être appliquées efficacement, les méthodes de fouille de données temporelles doivent être appliquées sur des séquences dont la représentation est adaptées au sens et au niveau de la recherche d'informations. Un algorithme aussi pertinent soit-il n'est malgré tout pas efficace si les données qu'il reçoit sont biaisées, ou mal appropriées [51]. Une grande attention doit ainsi être portée à leur prétraitement et à leur représentation significative par rapport au problème posé pour rendre efficace les tâches suivantes de fouille de données (fouille de caractères et classification). De manière générale, la représentation des séries temporelles est particulièrement importante pour dépasser la difficulté d'analyse directe et efficace de séquences de données continues, et plus spécialement encore lorsque les séquences étudiées sont de grande dimension, hétérogènes, et éventuellement bruitées. Il s'agit d'extraire les caractères importants de ces séquences au regard des objectifs de reconnaissance ou de décision.

Plusieurs modes de représentation sont utilisés en fonction principalement du type des données analysées et du niveau de représentation nécessaire. Dans ce paragraphe, on décrit ainsi les représentations à valeurs réelles, la discrétisation, puis l'abstraction. On présente enfin notre positionnement par rapport à ces méthodes de représentation : on utilise une abstraction pour s'adapter au "haut niveau" de décision, incluant une *discrétisation* pour rendre homogènes les séquences étudiées.

2.2.1 Représentation à valeurs réelles

De nombreuses méthodes de représentation ont été proposées pour des séries temporelles, c'est-à-dire pour des séquences de données à valeurs réelles. Certaines techniques définissent une description approximative des séquences en les représentant par leur tendance de variation sur des segments prédéfinis ou estimés au cours de la représentation, et en leur affectant éventuellement des poids différents en fonction de leur importance relative dans la séquence. Cette approximation par morceaux peut être linéaire (*PLA – Piecewise Linear Approximation*) [55, 83], ou constante.

Des exemples d'approximation par des constantes sur une partition de la séquence initiale sont la méthode *PCA – Piecewise Constant Approximation* [57] – dans le cas de segments de longueur fixe ; la méthode *PAA* – Piecewise Aggregate Approximation [59, 116] – qui remplace chacun des segments, tous de même longueur, par la valeur moyenne associée ; ou encore *APCA – Adaptative Piecewise Constant Approximation* [58] – qui permet d'obtenir une approximation constante sur des segments de longueur variable.

D'autres techniques proposent un changement de l'espace de représentation, telles que la transformée de Fourier discrète (*DFT – Discrete Fourier Transform*) [90], la transformée en ondelettes discrète (*DWT – Discrete Wavelet Tranform*) [25], la décomposition en valeurs singulières (*SVD – Singular Value Decomposition*) [60], ou encore la représentation de chaque sous-séquence monodimensionnelle de longueur n par un point dans un espace temporel de dimension n [43].

Toutes ces transformations s'appliquent sur des séries temporelles et sont intéressantes en terme de la mise en évidence des caractéristiques importantes des données analysées afin de rendre plus efficaces les étapes suivantes de traitement. Par contre, l'espace de représentation reste celui des valeurs réelles, et les méthodes de fouille de données qui peuvent être appliquées dans ce contexte sont assez limitées étant donné le nombre de valeurs, et par conséquent de sous-séquences possibles.

2.2.2 Discrétisation

Quelle que soit la transformation réalisée, la considération de séries de valeurs réelles limite le choix des algorithmes applicables de fouille de données, plus appropriés à la considération d'ensembles finis et dénombrables de valeurs pour les paramètres étudiés. Les chercheurs se sont ainsi tournés vers des représentations symboliques, et de nombreuses méthodes de discrétisation des séries temporelles ont été développées. Les intervalles de discrétisation peuvent être déterminés par classification non supervisée sur un ensemble de valeurs expérimentales, selon des méthodes telles que les k plus proches voisins [34] ou une variante proposée dans [48], appelée *DLK – Dynamic Local K-means*. L'algorithme *DLK* apprend le nombre de classes en fonction d'un seuil maximum défini *a priori* sur la variance de chaque classe. D'autres méthodes visent à l'équiprobabilité des symboles produits par la discrétisation. Dans [70] par exemple, la partition des valeurs est réalisée dans cet objectif et à partir de l'hypothèse d'une distribution fortement Gaussienne des valeurs normalisées.

Dans le cadre de l'analyse de séquences multidimensionnelles hétérogènes, une méthode de discrétisation est particulièrement intéressante pour rendre homogènes les séquences analysées.

2.2.3 Abstraction

Plus généralement, il existe pour la représentation de séquences temporelles des méthodes qui consistent en une simple transformation ou réduction de l'espace de représentation des données, alors que d'autres nécessitent des opérations plus complexes telles qu'une démarche de classification. C'est le cas de la recherche d'intervalles de discrétisation par une méthode des k plus proches voisins par exemple, ou encore de la recherche d'une partition de l'axe temps en segments de longueur variable pour une approximation significative d'une série temporelle. Ces démarches visent souvent à résumer les séries temporelles, en fournissant une représentation synthétique, interprétable, en fonction du problème à résoudre. C'est ce qu'on appelle l'*abstraction* de séries temporelles, qui résulte en la détermination d'une séquence d'intervalles, pouvant correspondre à différentes longueurs, et étiquetés par des attributs discrets ou numériques [51].

L'abstraction de séries temporelles soulève en particulier le problème de *partitionnement* d'une séquence dans le temps, qu'on propose de formaliser à l'inverse comme le *regroupement* d'une succession de vecteurs homogènes, correspondant à un même tendance de variation dans un certain intervalle de temps. Ainsi, le *partitionnement* d'une séquence peut se traduire comme l'*agrégation* de données successives dans le temps en symboles représentatifs chacun d'une même situation. D'après sa définition, l'agrégation est en effet l'"Action de réunir des éléments distincts pour former un tout homogène" (Dictionnaire Larousse, 2002). Dans la littérature, l'agrégation se rapporte souvent au regroupement dans l'espace de segments similaires, mais observés à des instants différents, de trajectoires d'objets mobiles au cours du temps, comme dans [72]. D'après [75], l'agrégation de sous-séquences est alors réalisée selon des segments de longueur de prédéfinies (agrégation appelée alors en anglais *span grouping*) ou de longueur variable en fonction des données enregistrées à chaque instant (*instant grouping*).

Dans le contexte du résumé des variations temporelles d'un ensemble de paramètres, ces deux méthodes correspondent bien au problème de *partitionnement supervisé* – longueurs prédéfinies – ou *non supervisé* – longueurs variables – d'une série temporelle. Cela confirme la proximité des deux problématiques de *partitionnement* et d'*agrégation* de données temporelles. Il semble ainsi légitime de parler pour l'abstraction d'un problème d'agrégation de données temporelles. Dans un contexte non supervisé, l'abstraction – et par conséquent l'agrégation temporelle – nécessite d'après [51] de disposer d'une **mesure de distance** pour évaluer la similarité ou dissimilarité des segments possibles.

2.2.4 Application de l'abstraction dans notre contexte

Une spécificité importante de notre contexte est l'apprentissage d'un profil comportemental "haut niveau" à partir de données "bas niveau". La différence entre les données initiales "bas niveau" – données issues des capteurs – et l'objectif de décision plutôt "haut niveau" – mettre en évidence les tendances caractéristiques de variation – nécessite que cette étape de représentation soit une **abstraction** des séquences étudiées. L'objectif est d'extraire des informations dont le niveau de détail est approprié à celui de la décision, et interprétables au regard de la résolution du problème : l'identification du profil comportemental d'une personne à domicile.

L'abstraction est finalement réalisée à deux niveaux, en agissant d'une part sur l'axe des valeurs (discrétisation) et d'autre part sur l'axe temporel (agrégation).

- **Axe des valeurs.** La nécessité de générer des séquences abstraites adaptées aux principales méthodes de fouille de données suggère de rendre leurs composantes discrètes. Puisqu'on considère des séquences de données hétérogènes, il s'agit de **discrétiser** les paramètres quantitatifs et, si possible et nécessaire, de réduire le cardinal des modalités des paramètres qualitatifs.
- **Axe temporel.** Dans notre contexte expérimental, l'abstraction a ensuite pour objectif l'obtention d'une représentation des séquences interprétable dans le temps en terme des activités effectuées à domicile. Si on décrit une activité comme la succession d'actions élémentaires ayant chacune une certaine durée, l'idée intuitive consiste alors en l'agrégation de données successives dans le temps en symboles représentatifs chacun d'une même situation, c'est-à-dire d'une même tâche élémentaire réalisée, pendant une durée variable selon l'instance de la tâche. Ainsi, le partitionnement des séquences pour l'agrégation doit permettre de définir de façon non supervisée des *segments de différentes longueurs*, correspondant chacun à un *état stationnaire* des paramètres surveillés à l'échelle de décision.

2.3 Mesure de similarité

D'après les paragraphes précédents, une mesure de similarité est nécessaire pour la fouille de données temporelles, aussi bien que pour certaines méthodes d'abstraction préalable de ces données afin qu'elles prennent du sens par rapport aux objectifs de résolution du problème. Il est en effet indispensable de disposer d'une méthode de comparaison de sous-séquences pour l'identification précise des sous-séquences récurrentes et leur classification : on compare alors les **sous-séquences initiales, multidimensionnelles et hétérogènes**.

Dans notre contexte, l'abstraction nécessite également une mesure de proximité entre sous-séquences discrètes. En effet, cette étape est décrite comme incluant l'agrégation des vecteurs discrets successifs décrivant un état stationnaire de la personne – interprété comme une même tâche élémentaire réalisée. Il s'agit ainsi de comparer des **sous-séquences multidimensionnelles discrètes** – une sous-séquence donnée et la séquence constituée uniquement du vecteur moyen correspondant – pour décider de l'agrégation des données correspondantes en leur vecteur moyen.

On expose ainsi dans ce paragraphes les méthodes utilisées pour la comparaison de séquences de données réelles – les distances Euclidienne, DTW et $LCSS$ – ainsi qu'une méthode proposée pour l'approximation de la similarité de séquences discrètes – la distance minimum.

2.3.1 Distance Euclidienne

L'approche la plus simple communément utilisée pour définir une fonction de similarité est fondée sur la distance Euclidienne, ou des extensions de cette distance qui autorisent certaines transformations entre des séquences similaires, telles que différentes échelles ou décalages. Chiu *et al.* [28] ont utilisé avec succès dans certaines applications une distance Euclidienne pour l'extraction de motifs dans des séries temporelles à une dimension. Cependant, cette méthode est particulièrement sensible à la présence de valeurs inattendues dans les instances de motifs et aux distorsions éventuelles de l'axe temporel.

2.3.2 Distance DTW

Une approche consiste alors à utiliser une autre distance – appelée en anglais *Dynamic Time Warping (DTW)* – qui autorise les déformations dans le temps, et par conséquent la comparaison de séquences de différentes longueurs [56, 61]. Le principe est d'identifier les associations des points des deux séquences qui minimisent leur distance globale. Cette distance est alors calculée sur le principe d'une distance euclidienne à partir des "meilleurs" couples de points constitués. Tous les points de chaque séquence doivent être au moins associés à une point de l'autre séquence, mais il est en plus possible de réaliser des associations multiples. On respecte bien sûr l'ordre des séquences de points dans le temps : un point d'une séquence ne peut être associé qu'à un ou plusieurs points successifs de l'autre séquence. Le schéma de la figure 2.1, issu de [56], illustre le principe de cette distance par comparaison à la distance euclidienne. La présence de valeurs inattendues entre deux séquences pourtant similaires résulte cependant toujours en une grande valeur de distance, même si les différences observées entre les deux séquences ne le sont que sur quelques points.

2.3.3 Distance $LCSS$

Des mesures de distance non métriques ont ainsi été proposées pour comparer efficacement des séquences de données bruitées [2, 33, 110]. L'idée est de saisir la notion intuitive de similarité

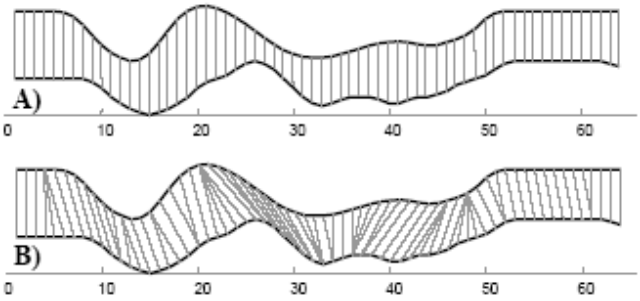


FIG. 2.1 – Comparaison de deux séquences de même forme mais non alignées sur l'axe du temps. Dans le cas **A)**, une distance euclidienne suppose que le $i^{ème}$ point d'une séquence est aligné avec le $i^{ème}$ point de l'autre, et la distance alors obtenue est dramatiquement pessimiste. Dans le cas **B)**, une distance DTW permet l'association de points non alignés dans le temps et la mesure obtenue est bien plus réaliste. [56]

présentée dans [2] : "deux séquences devraient être considérées similaires si elles contiennent suffisamment de couples de sous-séquences similaires, ne se recouvrant pas, et dans le même ordre." Cela revient à identifier la plus longue sous-séquence commune – *Longest Common Subsequence (LCSS)* – entre les deux séquences comparées. On recherche les paires de points effectivement similaires entre les deux séquences, selon un seuil d'écart maximum entre deux valeurs, et on calcule ensuite la proportion de ces points similaires dans les séquences pour estimer leur distance. Tous les points de chaque séquence ne sont alors pas obligatoirement associés à un point de l'autre séquence.

Cette approche permet alors la présence de valeurs inattendues entre deux sous-séquences similaires, de même que différents facteurs d'échelle et de possibles décalages dans les valeurs. Une distance non métrique de ce type est alors plus adaptée pour la comparaison de séquences dans un contexte bruité. La figure 2.2 illustre le principe de cette distance par comparaison à la distance DTW.

Les travaux cités sur les mesures de distance entre trajectoires concernent cependant principalement des séries temporelles de faible dimension – de une à trois composantes – et qui ne considèrent pas le problème de composantes hétérogènes – qualitatives et quantitatives – pour la description d'un objet mobile. Dans un contexte bruité et hétérogène, notre objectif est alors d'étendre l'approche fondée sur la détermination de la longueur de la plus longue sous-séquence commune – $LCSS$ [110] – à la prise en compte de séquences temporelles multidimensionnelles et hétérogènes.

2.3.4 Distance minimum discrète

Une mesure de distance permettant de comparer des séquences discrètes est nécessaire pour décider de l'agrégation ou non d'une succession de vecteurs discrets en leur vecteur moyen. Chiu *et al.* [28] ont proposé une *distance minimum* pour comparer sur une dimension des séquences temporelles discrètes de même longueur. Une distance minimum entre séquences temporelles est une borne inférieure de la mesure de distance entre les séquences des valeurs réelles correspondantes. Elle est intéressante pour avoir une idée de leur similarité et peut ainsi être utilisée comme mesure de distance approchée pour autoriser ou non l'agrégation de vecteurs discrets successifs en fonction d'un seuil maximum.

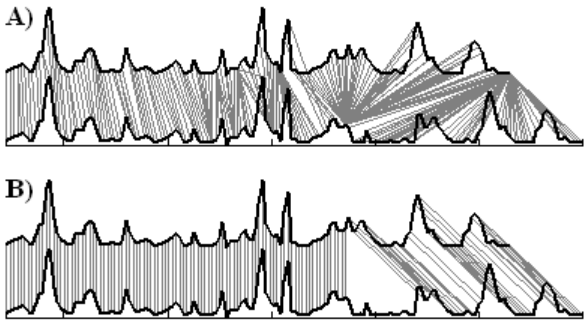


FIG. 2.2 – Comparaison de deux séquences globalement identiques mis à part pour quelques points présents dans l'une mais pas dans l'autre.
Dans le cas **A)**, une distance *DTW* doit associer tous les points de chaque séquence si bien que la présence d'une sous-séquence "différente" dans l'une des deux perturbe l'ensemble des associations de points, même sur les sous-séquences identiques. Dans le cas **B)**, une distance de type *LCSS* permet de ne pas tenir compte dans le calcul de distance des quelques points qui diffèrent d'une séquence à l'autre.

2.4 Synthèse du positionnement par rapport à l'état de l'art

Par rapport aux travaux déjà réalisés dans le domaine de la fouille de données temporelles, notre objectif est de proposer une méthode de fouille de séquences temporelles qui permet l'**extraction non supervisée de motifs temporels multidimensionnels et hétérogènes**, à haut niveau par rapport aux données initiales. Dans ce contexte, et selon l'état de l'art des différents axes de recherche abordés, on a ainsi identifié certaines caractéristiques des étapes d'analyse nécessaires.

(1) La représentation doit être une abstraction des séquences temporelles.

On agit à cette étape sur l'**axe des valeurs** en discrétisant les paramètres quantitatifs. L'objectif est d'une part de constituer des séquences homogènes plus faciles à analyser. D'autre part, bien plus de méthodes de fouille de données sont applicables sur des séquences discrètes.

On intervient également sur l'**axe temporel** en agrégeant dans le temps, selon des intervalles de longueur variable, les vecteurs successifs ne présentant pas de forte variation au regard du niveau de décision.

(2) La fouille de données se décompose en fouille de caractères et classification.

Une première étape de fouille de caractères est nécessaire pour identifier les sous-séquences les plus pertinentes pour l'identification des sous-séquences récurrentes et l'application de méthodes de classification pour la découverte des motifs.

Au niveau de la fouille de caractères, une méthode intéressante d'identification de sous-séquences a été expérimentée par Chiu *et al.* dans le contexte de séries temporelles à une dimension. Cette technique basée sur les projections aléatoires des sous-séquences possibles et leur comparaison deux à deux nécessite cependant d'être étendue à la considération de séquences multidimensionnelles. Un autre point fondamental est son extension à l'identification de sous-séquences récurrentes dont les instances peuvent être de différentes longueurs.

La classification peut être ensuite réalisée simplement sur la base d'une mesure de distance entre sous-séquences récurrentes. Il est cependant nécessaire de s'assurer au préalable de la pertinence des sous-séquences récurrentes considérées à l'issue de la fouille de carac-

tères. Les caractères identifiés doivent en effet être fréquents, distinctifs d'au moins une classe, et non redondants.

(3) Une mesure de similarité entre séquences temporelles est nécessaire.

D'une part notre contexte nécessite la définition d'une mesure de similarité entre séquences multidimensionnelles et hétérogènes, pour l'identification précise des sous-séquences récurrentes et leur classification en motifs.

D'autre part une mesure approchée de similarité, appliquée sur des séquences discrètes, est utilisée pour l'agrégation des vecteurs successifs correspondant à un état stationnaire du système.

3

Méthodologie pour l'extraction de motifs temporels

L'apprentissage du profil comportemental d'une personne correspond à l'extraction de motifs significatifs à partir de séquences temporelles de données multidimensionnelles et hétérogènes issues d'un ensemble de capteurs. Dans ce contexte, les motifs identifiés dans les séquences doivent correspondre aux comportements habituels d'une personne dans ses activités de la vie quotidienne. Les spécificités individuelles de comportement imposent par ailleurs une approche complètement non supervisée pour l'identification des motifs, sans *a priori* sur les sous-séquences correspondant aux activités régulières.

Ce chapitre a ainsi pour objectif de présenter une méthodologie pour l'extraction non supervisée de motifs temporels adaptée au problème considéré et au contexte expérimental de notre application. Nous rappelons d'abord dans la section 3.1 les contraintes liées au problème de la télésurveillance médicale à domicile, puis en 3.2 celles relatives aux données disponibles pour sa résolution, pour présenter enfin en 3.3 un principe général de fouille de données temporelles, en particulier approprié à ce contexte.

3.1 Problème considéré
Extraction de motifs pour la télésurveillance médicale

La conception d'un système de décision est réalisée dans la perspective de résolution d'un problème (voir Fig. I.3.2). Elle est ainsi guidée par le contexte d'étude et les objectifs fixés, et doit prendre en compte les paramètres clés de la résolution de tout problème de décision : (1) le niveau de performance exigé, (2) le niveau de détail nécessaire et (3) le niveau de connaissance disponible (voir paragraphe I.3.4).

(1) Niveau de performance. L'objectif est d'identifier les comportements récurrents d'une personne dans sa vie quotidienne (sensibilité), tout en n'identifiant et ne classifiant pas comme habituels ceux qui ne le sont pas (spécificité).

(2) Niveau de détail. L'identification des comportements habituels ne nécessite pas un grand niveau de détail, c'est une décision "haut niveau" en comparaison aux données "bas niveau" issues des capteurs. Par conséquent, il n'est en particulier pas nécessaire de localiser précisément les sous-séquences récurrentes dans le séquences initiales. L'essentiel est de les reconnaître effectivement et de les caractériser grossièrement en termes des tendances de variation des différents paramètres, du moment, de la fréquence et de l'ordre d'occurrence.

(3) Niveau de connaissance. L'apprentissage des habitudes de vie d'une personne à domicile doit être réalisé sur la base de chaque individu, compte tenu des spécificités individuelles de comportement en termes d'activité et de physiologie. Les études réalisées dans le cadre de la simulation montrent d'ailleurs bien cette variabilité inter-individuelle des différents paramètres considérés (voir par exemple les annexes E.2 et E.3). L'extraction des motifs est ainsi réalisée de façon non supervisée, et avec peu de connaissances *a priori* sur les activités de la vie quotidiennes. On prend par exemple en compte la durée minimale de la réalisation d'une *tâche* pour qu'elle puisse être considérée comme une *activité* de la vie quotidienne.

3.2 Contexte expérimental *Critères guidant la construction d'une méthode d'extraction*

La conception et l'implémentation d'un système de décision impose de définir le contexte expérimental approprié aux objectifs et aux exigences du problème considéré. Les séquences temporelles générées et/ ou collectées sont alors adaptées à l'expérimentation d'algorithmes de résolution du problème. Leurs caractéristiques – bien identifiées dans le contexte expérimental – sont des critères qui guident la mise en place du système de décision. Dans un contexte de recherche de motifs représentatifs de comportements humains – les activités de la vie quotidienne d'une personne à domicile – à partir de données hétérogènes enregistrées d'un ensemble de capteurs, le système de décision doit en particulier être adapté aux caractéristiques suivantes :

- **Séquences temporelles multidimensionnelles.** Système approprié à l'analyse de l'évolution de n'importe quel objet ou de n'importe quelle situation décrite par un ensemble de paramètres.
- **Composantes hétérogènes.** Cohérence pour l'extraction de motifs à partir de l'observation d'objets ou de situations décrits par un ensemble composé de paramètres quantitatifs et qualitatifs.
- **Séquences temporelles mixtes.** Apprentissage des motifs à partir de séquences composées à la fois de motifs et de "non motifs". La vie quotidienne se compose en effet de comportements très prévisibles et réguliers – les *activités de la vie quotidienne (AVQ)* par exemple – aussi bien que de comportements inattendus ou imprévisibles.
- **Imprécision des motifs.** Forte présence de bruit entre les instances d'un même motif, particulièrement parce que les motifs représentent des comportements humains. On s'intéresse alors aux tendances qui se trouvent dans les séquences temporelles plutôt qu'aux valeurs précises des paramètres à chaque instant. On considère ainsi la découverte de motifs de "haut niveau".
- **Valeurs imprévisibles.** Valeurs imprévisibles potentiellement contenues dans les instances des motifs, pouvant correspondre à des anomalies de fonctionnement des capteurs, mais plus sûrement à un comportement humain imprévisible bien que pas du tout anormal dans le cours d'une activité de la vie quotidienne – par exemple, un passage aux toilettes pendant la préparation d'un repas.
- **Translation dans le temps.** Translation possible dans le temps des instances des motifs représentant les activités de la vie quotidienne, d'une journée à l'autre. Des activités identiques peuvent en effet se produire à n'importe quel moment, et on souhaite ainsi être capable de les reconnaître sans *a priori* sur leur moment d'occurrence.

- **Déformation temporelle.** Déformation possible des instances de motifs dans le temps, avec des variations de durée pouvant être assez importantes. Une même activité n'est en effet pas forcément réalisée pendant la même durée.

3.3 Système de décision *Méthodologie pour l'extraction de motifs*

Le processus d'apprentissage non supervisé des motifs contenus dans une séquence temporelle peut être assimilé à un système de décision sur l'appartenance ou non de sous-séquences des données d'apprentissage à l'instance d'un motif. Plus généralement, c'est aussi un **système de reconnaissance** qui concerne l'identification de motifs dans des séquences temporelles issues de capteurs. Dans ce paragraphe nous déclinons ainsi le schéma général d'un système de reconnaissance pour aboutir à la définition d'un système de reconnaissance sur de larges ensembles de données temporelles, telles que disponibles dans notre contexte. Cette définition repose sur les étapes définies par Antunes *et al.* [3] pour la fouille de données temporelles. Les schémas **A)** à **C)** de la figure 3.1 illustrent cette démarche.

Système "classique" de reconnaissance

Un système de reconnaissance de formes réalise deux tâches principales successives [73] : (1) Extraction des caractères et (2) Classification. Une étape de prétraitement des données utilisées en entrée du système est par ailleurs souvent nécessaire afin de s'affranchir au moins des erreurs, imprécisions de mesure et variabilité très haute fréquence des paramètres étudiés. Le fonctionnement d'un processus de reconnaissance de formes est alors décrit par la succession des étapes suivantes (voir Fig. 3.1, schéma **A)**) :

(1) **Prétraitement** des données issues des capteurs considérés pour la résolution du problème ;

(2) **Extraction des caractères** appropriés pour répondre aux objectifs du système ;

(3) **Classification** pour l'identification des formes recherchées, telles que des tendances de variation ou motifs.

Dans le cas qui nous intéresse d'une reconnaissance complètement non supervisée sur de larges ensembles de données temporelles, un tel système d'apprentissage peut également être interprété comme répondant à un problème de *fouille de données temporelles*. La fouille de données concerne en effet l'extraction d'informations inconnues *a priori* contenues implicitement dans un ensemble de données analysées. On cherche alors à établir le lien entre le schéma "classique" d'un système d'apprentissage et celui que l'on peut proposer pour un système de fouille de données temporelles dédié à l'apprentissage, à partir des étapes définies par Antunes *et al.* [3] pour la fouille de données temporelles.

Système de fouille de données temporelles proposé par Antunes *et al.* [3]

Dans le cadre de la fouille de données temporelles, Antunes *et al.* [3] identifient entre autres deux éléments clés de la définition d'un processus d'apprentissage (voir Fig. 3.1, schéma **B)**) :

1. **Représentation des séquences temporelles.** Prétraitement, puis transformation des séquences de données pour fournir en entrée du système les données les plus pertinentes en fonction des objectifs fixés de la reconnaissance. Il s'agit bien d'**extraire les caractères** appropriés à la résolution du problème, tel que proposé après le prétraitement des données dans le schéma d'un système de reconnaissance de forme.

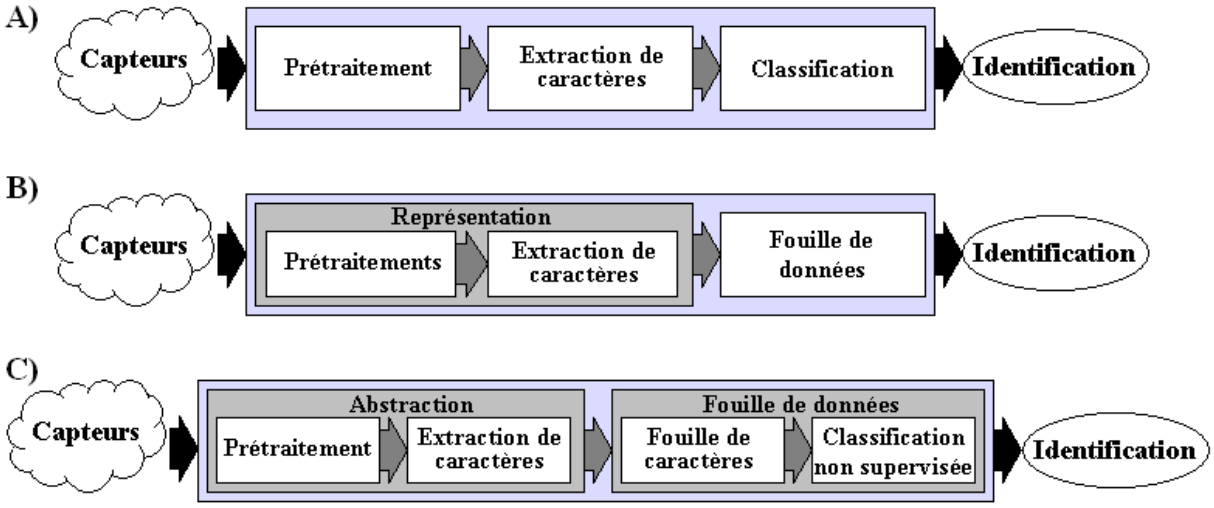


FIG. 3.1 – Construction d'un système de reconnaissance.
À partir de données enregistrées par un ensemble de capteurs, les différentes visions des étapes d'un système de reconnaissance sont les suivantes :

A) Schéma "classique" d'un système de reconnaissance, incluant (1) Prétraitement, (2) Extraction de caractères et (3) Classification.

B) Vision de Antunes *et al.* [3] d'un processus d'apprentissage dans le cadre de la fouille de données temporelles : (1) Représentation des séquences temporelles (prétraitement et transformation) et (2) Opérations de fouille de données.

C) Notre vision d'un processus de fouille de données temporelles dédié à l'apprentissage, incluant (1) Abstraction des séquences temporelles (représentation avec notion d'interprétation) : prétraitement et extraction de caractères ; (2) Opérations des fouille de données : fouille de caractères et classification non supervisée.

2. **Opérations de fouille de données.** Application des modèles et méthodes de représentation des données à la résolution de la véritable problématique de fouille de données (découverte de règles d'association, classification, prédiction). L'étape de **classification** du schéma d'un système de reconnaissance est ainsi opérée à ce niveau d'analyse.

Les étapes de **représentation** et de **fouille de données** sont alors redéfinies pour s'adapter en particulier à deux caractéristiques de notre contexte : (1) la grande différence entre les niveaux des données analysées et de l'objectif de reconnaissance et (2) le caractère mixte des données analysées, contenant à la fois des motifs et des "non motifs".

De la *Représentation* à l'*Abstraction*

Dans le cas d'un problème complexe intégrant une grande différence entre le niveau de détail des données initiales et le niveau de décision requis, il est nécessaire de redéfinir **la représentation** en une étape d'**abstraction**. L'objectif est alors d'extraire les *informations* "haut niveau" adaptées au niveau de la fouille de données, et qui permettent une représentation interprétable des données brutes en terme du problème posé. Plus qu'une simple transformation dans un certain formalisme, l'abstraction se situe à un niveau d'interprétation et génère une représentation synthétique qui met en évidence la structure fondamentale des séquences temporelles du point de vue des objectifs de reconnaissance.

L'**abstraction** nécessaire à la fouille de données temporelles correspond ainsi à l'étape de **représentation** considérée par Antunes *et al.*, c'est-à-dire aux étapes successives de **prétraitement** et d'**extraction de caractères** d'un système de reconnaissance.

Fouille de données

La fouille de données telle que définie par Antunes *et al.* est appliquée à la représentation des données initiales, et concerne par exemple une démarche de classification. Dans un système "classique" de reconnaissance, la classification suit également l'étape d'extraction de caractères. Dans un contexte non supervisé d'identification de motifs dans des séquences temporelles, à partir de séquences comprenant à la fois des sous-séquences représentatives de certains motifs et d'autres ne correspondant à aucun motif, les algorithmes de **classification** sont cependant difficilement applicables directement aux caractères extraits de données séquentielles, en terme des caractéristiques de leurs variations dans le temps.

Il est ainsi intéressant d'utiliser un mécanisme supplémentaire d'extraction de caractères, tel que cité dans [3], cette fois en terme de l'identification des sous-séquences porteuses d'informations pertinentes en fonction des objectifs de classification : c'est la **fouille de caractères**. Dans cette expression, les caractères ne sont plus les informations pertinentes extraites tout au long de la séquence analysée, mais les sous-séquences pertinentes extraites de cette séquence. Un exemple est l'identification des sous-séquences récurrentes dans la séquence initiale comme une première étape de fouille de données avant leur classification [68] pour l'identification effective des motifs, c'est-à-dire des classes de sous-séquences récurrentes significatives.

Les opérations de fouille de données sont ainsi divisées en deux étapes successives permettant d'atteindre progressivement le niveau de reconnaissance.

1. **Fouille de caractères.** Sélection des caractères significatifs dans les séquences temporelles initiales, c'est-à-dire des sous-séquences caractéristiques des objectifs de reconnaissance.
2. **Classification.** Application d'un l'algorithme de classification non supervisée à ces caractères.

La **fouille de données** temporelles peut être ainsi interprétée comme correspondant aux étapes successives d'**extraction de caractères** et de **classification** d'un système de reconnaissance.

Système de reconnaissance sur de larges ensembles de données temporelles

Finalement, par rapport au schéma de principe de tout système de reconnaissance présenté sur le schéma **A)** de la figure 3.1, la considération d'un problème d'analyse de séquences temporelles, à "bas niveau" par rapport à l'objectif de reconnaissance, divise l'**extraction de caractères** en deux étapes successives intégrées respectivement aux niveaux de l'abstraction, puis de la fouille de données.

1. **Abstraction.** L'extraction de caractères a pour objectif la représentation de la séquence temporelle initiale (après prétraitement) par une succession d'informations pertinentes au regard de la décision, sur une ou plusieurs dimensions. Le résultat de l'abstraction est ainsi une séquence temporelle décrivant le système observé sur la même durée que la séquence initiale.
2. **Fouille de données.** Étant donné le nombre exponentiel de sous-séquences constituées à partir de la séquence initiale, l'extraction de caractères consiste dans ce cas en la sélection des sous-séquences – les caractères – les plus significatives par rapport au problème posé afin de restreindre l'espace des caractères considérés. Le résultat de la fouille de caractères est un ensemble de sous-séquences de la séquence initiales appropriées à la classification.

On propose ainsi le schéma **C)** de la figure 3.1 comme représentatif des étapes successives d'un système de reconnaissance sur de larges ensembles de données temporelles. Le paragraphe suivant décrit les conséquences de ce principe de reconnaissance sur la définition d'une **mesure de similarité**. Les deux principales étapes considérées nécessitent en effet toutes deux la définition d'une mesure de similarité, appropriée d'abord à une estimation grossière de similarité (abstraction), puis à une évaluation plus fine de la proximité réelle de sous-séquences (fouille de données).

Mesure de similarité

La description proposée d'un système de reconnaissance utilisant des données temporelles à des conséquences sur la définition d'une mesure de similarité. Au niveau de complexité considéré, une mesure de similarité est en effet nécessaire, d'une part pour l'abstraction, et d'autre part pour la fouille de données.

1. **Abstraction.**
   Dans un contexte non supervisé, il est nécessaire de disposer d'une mesure de similarité appropriée pour l'étape d'abstraction [50]. L'objectif est de partitionner la séquence initiale en sous-séquences "homogènes", ou encore "stationnaires", c'est-à-dire présentant des variations peu significatives au regard de la reconnaissance. La pertinence de l'abstraction ou agrégation d'une succession de "points" ou vecteurs en un seul symbole décrivant la continuité d'une même situation pendant la durée correspondante est évaluée par leur proximité globale.
   A ce niveau de l'analyse, une approximation de la similarité réelle des vecteurs d'une sous-séquence est suffisante. On s'intéresse en effet aux tendances de variations des paramètres dans leur ensemble, pour une décision plutôt "haut niveau" par rapport aux données analysées. La considération conjointe des paramètres hétérogènes nécessite par ailleurs une étape préalable de discrétisation des paramètres quantitatifs.
   On propose ainsi l'utilisation d'une **distance discrète minimum** pour estimer la proximité globale d'une succession de vecteurs et du "vecteur moyen" correspondant, et faire éventuellement l'hypothèse que ce vecteur moyen représente bien la sous-séquence pendant la durée correspondante.
2. **Fouille de données.**
   Une fois localisées les sous-séquences les plus probablement représentatives d'instances de motifs, une mesure de similarité est nécessaire pour évaluer la proximité effective de ces sous-séquences, à deux niveaux :
   (a) pour la **fouille de caractères**, afin d'évaluer si une sous-séquence est suffisamment similaire à au moins une autre sous-séquence et peut par conséquent être considérée comme un caractère, appelé alors *tentative de motif* ;
   (b) pour la **classification** des tentatives de motifs afin d'effectuer des regroupements significatifs au regard de la décision et finalement identifier les *motifs* – classes de *tentatives de motifs* – présents dans la séquence initiale.
   Afin d'affiner progressivement l'identification des motifs, ces deux étapes nécessitent une mesure de **distance réelle**, entre les séquences temporelles initiales simplement prétraitées. Cette distance doit ainsi être appropriée à la comparaison de séquences multidimensionnelles et hétérogènes.

3.4 Synthèse

Le tableau de la figure 3.2 synthétise les étapes successives de la fouille de données temporelles pour l'extraction non supervisée de motifs, ainsi que la mesure de similarité nécessaire à chaque niveau de l'analyse.

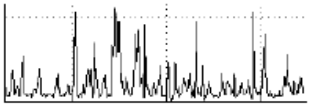
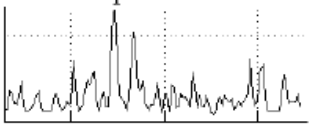
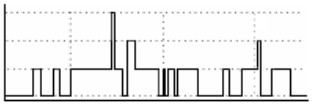
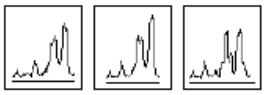
Étapes	Type de données	Mesure de distance
	Données brutes 	
1. Abstraction **1.1** Prétraitement	Données prétraitées 	Distance approchée
1.2 Extraction de caractères	Données abstraites 	
2. Fouille de données **2.1** Fouille de caractères	Tentatives de motif 	Distance réelle
2.2 Classification	Motif 	

FIG. 3.2 – Synthèse des étapes de la fouille de données temporelles.

Ce tableau met en évidence deux définitions importantes :

- **Tentatives de motifs :** Sous-séquences récurrentes, non redondantes, identifiées dans la séquence initiale à l'issue de la fouille de caractères ;
- **Motifs :** Classes significatives de tentatives de motifs.

Le chapitre 4 présente les mesures de similarité nécessaires pour l'abstraction et la fouille de données. Puis, le chapitre 5 propose une approche d'extraction non supervisée de motifs selon le schéma défini dans ce chapitre d'un système de reconnaissance dédié à l'analyse de données temporelles.

4

Mesure de similarité

Ce chapitre décrit les mesures de similarité nécessaires au cours du processus d'extraction non supervisée de motifs dans des séquences de données temporelles multidimensionnelles et hétérogènes. Dans un contexte bruité et hétérogène, les mesures de distance non métriques, fondées notamment sur la notion de plus longue sous-séquence commune aux séquences comparées ($LCSS$), semblent plus efficaces que les distances métriques telles que les distances Euclidienne et DTW (voir le paragraphe 2 sur l'état de l'art).

Notre objectif est alors d'étendre l'approche $LCSS$ à la comparaison de séquences temporelles multidimensionnelles et hétérogènes. Cela nécessite de définir d'abord, dans le section 4.1, une **distance entre deux valeurs** qui soit homogène pour l'ensemble des paramètres considérés, qualitatifs ou quantitatifs. Nous détaillons ensuite en 4.2 son utilisation pour le calcul d'une **distance réelle** entre séquences temporelles, utilisée pour la *fouille de données*. Enfin, nous utilisons en 4.3 ces définitions pour étendre l'approche proposée dans [28] pour le calcul d'une **distance minimum** entre séries temporelles : c'est une distance approchée nécessaire à l'étape d'*abstraction*.

4.1 Distance homogène pour des composantes hétérogènes

On considère la description possible d'un objet par plusieurs paramètres pouvant être de différents types :

- Quantitatif ;
- Qualitatif, à modalités ordonnées ;
- Qualitatif, à modalités non ordonnées.

La manière la plus simple d'assurer une certaine cohérence dans la comparaison des valeurs des différents paramètres est de générer des distances qui varient entre 0 et 1 pour chaque type de paramètre. Notons a et b deux valeurs d'un paramètre donné, et $d(a, b)$ la distance entre ces deux valeurs. Dans le cas d'un paramètre qualitatif, notons v le nombre de ses modalités, les valeurs possibles étant alors symbolisées par les entiers compris entre 1 et v. Selon le type de paramètre considéré, la distance $d(a, b)$ est définie selon l'une des équations (4.1) à (4.3).

$$d(a, b) = |a - b|\,, \tag{4.1}$$

$$d(a, b) = \frac{|a - b|}{v - 1}, \tag{4.2}$$

$$d(a, b) = min(|a - b|\,, 1). \tag{4.3}$$

Les équations (4.2) et (4.3) sont utilisées pour les paramètres qualitatifs, à modalités respectivement ordonnées ou non. L'équation (4.1) est appliquée aux valeurs des paramètres quantitatifs. L'obtention d'une distance variant entre 0 et 1 dans ce cas nécessite cependant une étape de normalisation des valeurs afin qu'elles varient elles aussi dans ce même intervalle. On utilise une *normalisation min-max*, fonction des bornes minimum et maximum de l'intervalle de variation des paramètres. Ces bornes sont définies soit par les experts, soit à partir d'analyses statistiques sur les données d'apprentissage. Toutes les valeurs sont alors limitées à ces bornes, des valeurs inférieures ou supérieurs étant interprétées comme bruitées ou erronées.

Notons X_{min} et X_{max} respectivement les bornes minimum et maximum de variation des valeurs x d'un certain paramètre X. On définit alors la valeur normalisée $norm(x)$ de toute valeur x selon l'équation (4.4).

$$norm(x) = \frac{max\,(0, min\,(x, X_{max}) - X_{min})}{X_{max} - X_{min}} \tag{4.4}$$

4.2 Distance réelle entre séquences temporelles

4.2.1 Notion de plus longue sous-séquence commune, *LCSS*

Une fonction efficace de similarité entre des trajectoires temporelles bruitées est basée sur leur plus longue sous-séquence commune (*LCSS*). Cette notion est déjà utilisée par Vlachos *et al.* [110] dans le contexte de l'analyse de séries temporelles multidimensionnelles – généralement deux ou trois dimensions – de paramètres quantitatifs. L'idée générale est de compter le nombre de couples de points considérés comme similaires lorsqu'on parcourt les deux séquences comparées, notées A et B. La similarité de deux points est définie selon un seuil maximum ϵ sur les valeurs correspondantes (voir Fig. 4.1). Un point ne peut jamais être associé deux fois à un point de l'autre séquence, si bien que le nombre maximum de points similaires rencontrés en comparant deux séquences correspond à la plus petite des deux longueurs. Une autre constante prédéfinie, δ, contrôle l'écart de temps maximum entre deux points de chacune des séquences pour qu'ils puissent être considérés similaires (voir Fig. 4.2).

La plus longue sous-séquence commune à A et B est alors une séquence que l'on peut obtenir en supprimant certains points de A, ou bien de B. Ce n'est ainsi pas une "sous-séquence" au sens où l'on utilise ce terme dans le reste du document : une sous-séquence de A est considérée par ailleurs plus restrictivement comme un ensemble de points *successifs* issus de A.

Supposons que les objets considérés sont des points évoluant dans un espace de dimension p : $(x_1, \ldots, x_p)$. Notons A et B les trajectoires de deux objets mobiles, de dimension respectivement n et m.

$$A = ((a_{x_1,1}, \ldots, a_{x_p,1}), \ldots, (a_{x_1,n}, \ldots, a_{x_p,n}))$$

$$B = ((b_{x_1,1}, \ldots, b_{x_p,1}), \ldots, (b_{x_1,m}, \ldots, b_{x_p,m}))$$

Pour une trajectoire donnée A, notons $Head(A)$ la sous-séquence correspondant aux positions successives de l'objet jusqu'à l'avant-dernière :

$$Head(A) = ((a_{x_1,1}, \ldots, a_{x_p,1}), \ldots, (a_{x_1,n-1}, \ldots, a_{x_p,n-1})).$$

Étant donnés un entier δ et un réel ϵ, $0 < \epsilon < 1$, la mesure de similarité telle que présentée dans [110] correspond au nombre de points similaires entre deux séquences comparées. La fonction de similarité entre les séquences A et B, notée $LCSS_{\delta,\epsilon}(A, B)$, est définie selon l'algorithme récursif (4.5). Cet algorithme décrit un processus récursif initialisé par la considération de deux

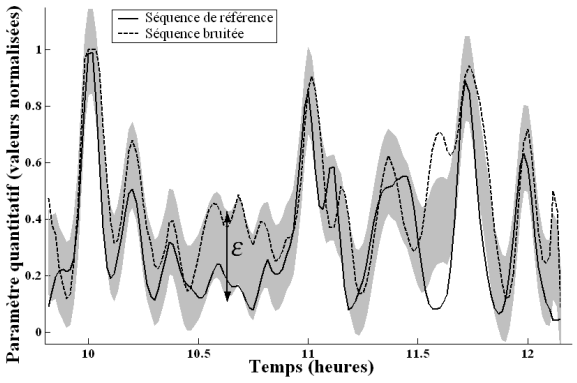


FIG. 4.1 – La notion de similarité basée sur *LCSS* et contrainte par ϵ.
En comparant les trajectoires point à point le long de l'axe du temps, les paires de points dont les deux sont dans la région grise peuvent être considérés comme similaires.

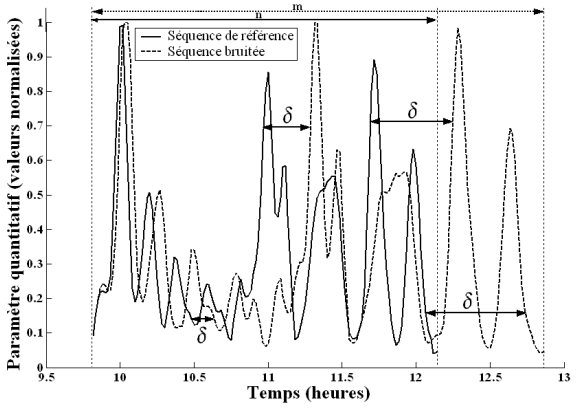


FIG. 4.2 – La notion de similarité basée sur *LCSS* et contrainte par δ.
Deux points de chacune des trajectoires peuvent être considérés similaires si l'intervalle de temps qui les sépare est inférieur à la valeur maximale autorisée pour δ.

trajectoires A et B dont on compare d'abord les derniers points. Si au moins l'une des deux séquences est vide, la mesure de similarité est nulle et le processus s'arrête. Sinon, si les deux derniers points des séquences sont similaires, la mesure de similarité vaut 1 plus la similarité calculée entre les deux trajectoires privées de leur dernier point, $Head(A)$ et $Head(B)$. Si enfin les deux derniers points des séquences A et B ne sont pas similaires, la mesure de similarité est égale au maximum de la similarité entre une séquence et l'autre privée de son dernier point, A et $Head(B)$, et réciproquement, $Head(A)$ et B.

$$LCSS_{\delta,\epsilon}(A,B) = \begin{cases} 0 \quad \text{si } A \text{ ou } B \text{ est vide,} \\ \\ 1 + LCSS_{\delta,\epsilon}(Head(A), Head(B)), \\ \quad \text{si } |a_{x_k,n} - b_{x_k,m}| < \epsilon,\ \forall 1 \leq k \leq p, \text{ et } |n-m| \leq \delta, \\ \\ max\left(LCSS_{\delta,\epsilon}(Head(A), B), LCSS_{\delta,\epsilon}(A, Head(B))\right) \\ \quad \text{sinon.} \end{cases} \tag{4.5}$$

Le nombre de points similaires entre A et B, $LCSS_{\delta,\epsilon}(A,B)$, est ensuite normalisé par la plus petite longueur des deux trajectoires comparées, de façon à ce que la mesure de similarité varie entre 0 et 1. La distance $D_{\delta,\epsilon}(A,B)$ entre les deux trajectoires A et B est alors définie comme suit [110] :

$$D_{\delta,\epsilon}(A,B) = 1 - \frac{LCSS_{\delta,\epsilon}(A,B)}{min(n,m)}.$$

$D_{\delta,\epsilon}(A,B)$ vérifie bien les propriétés de base d'une distance entre deux points, telles que décrites par (4.6).

$$\begin{cases} D_{\delta,\epsilon}(A,B) & \geq 0 \\ D_{\delta,\epsilon}(A,B) & = 0 \Leftrightarrow A = B \\ D_{\delta,\epsilon}(A,B) & = D_{\delta,\epsilon}(B,A) \end{cases} \tag{4.6}$$

Dans notre contexte, l'égalité entre A et B ($A = B$) n'a pas un sens d'égalité stricte mais de similarité, soit plutôt $A \approx B$.

4.2.2 Extension d'une distance *LCSS*

La mesure de distance que nous proposons pour comparer des séquences temporelles multidimensionnelles et hétérogènes diffère de celle de Vlachos *et al.* [110] à deux niveaux.

(1) Contrainte temporelle supplémentaire.

Une contrainte temporelle supplémentaire est nécessaire pour prendre en compte le cas où une grande partie des points de la plus courte des deux séquences est similaire à un ensemble de ceux de la plus longue, sans recouvrement, et dans le même ordre sur l'axe du temps. Un exemple typique est celui où la plus courte des séquences correspond exactement au début de la plus longue (voir Fig. 4.3). Selon la définition de Vlachos *et al.* [110] présentée dans le paragraphe précédent, leur similarité est égale à 1 (distance nulle), quelle que soit la longueur de la plus longue des séquences.

Notons N et M la longueur des séquences A et B respectivement à la première étape de l'algorithme récursif (4.5), n et m étant les longueurs considérées à l'étape courante de l'algorithme. Afin d'empêcher les fortes valeurs de similarité non justifiées, on contraint l'écart temporel pour la définition de deux points similaires à partir de la fin de chaque séquence autant qu'à partir du début. Une contrainte temporelle supplémentaire pour l'acceptation de la similarité effective de deux points $(a_{x_1,n}, \ldots, a_{x_p,n})$ et $(b_{x_1,m}, \ldots, b_{x_p,m})$ est alors définie par (4.7).

$$|n - m| \leq \delta \text{ et } |N - n - M + m| \leq \delta. \tag{4.7}$$

(2) Prise en compte des paramètres qualitatifs.

Nous avons par ailleurs besoin d'étendre la contrainte sur la similarité entre les valeurs de chaque composante des vecteurs définissant une séquence au cas où certains paramètres sont qualitatifs. Cette contrainte est basée sur les distances entre points définies au paragraphe 4.1, variant toutes entre 0 et 1. L'idée est de considérer que deux valeurs d'un paramètre qualitatif sont similaires si et seulement si elles sont égales. Par conséquent, on définit une valeur de la contrainte ϵ entre deux valeurs similaires appropriée pour chaque type de paramètre, de la façon suivante :

- Quantitatif : $0 < \epsilon < 1$,
- Qualitatif, à modalités ordonnées : $\epsilon = \frac{1}{v-1}$,
- Qualitatif, à modalités non ordonnées : $\epsilon = 1$.

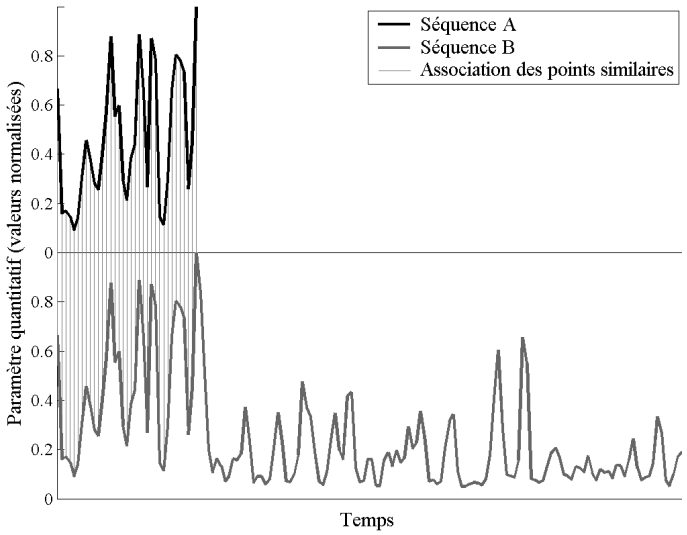


FIG. 4.3 – Mise en évidence des points similaires identifiés dans la comparaison de deux séquences A et B.
Selon la définition de Vlachos [110], quels que soient les contraintes de similarité ϵ et δ, la distance entre ces séquences pourtant largement différentes est nulle.

Considérons alors deux valeurs $a_{x_k,n}$ et $b_{x_k,m}$, correspondant à la $k^{ème}$ dimension, x_k, des $n^{ème}$ et $m^{ème}$ points des trajectoires A et B respectivement. La contrainte de similarité entre ces deux valeurs est alors définie par l'équation (4.8), où $d\left(a_{x_k,n}, b_{x_k,m}\right)$ représente la distance entre les valeurs $a_{x_k,n}$ et $b_{x_k,m}$ telle que définie par les équations (4.1) à (4.3) en fonction du type du paramètre correspondant.

$$d\left(a_{x_k,n}, b_{x_k,m}\right) < \epsilon, \ \forall 1 \leq k \leq p \tag{4.8}$$

Finalement, on propose de calculer le nombre de points similaires à deux séquences A et B, noté $LCSS_{\delta,\epsilon}(A,B)$, selon la formule récursive suivante (4.9), plutôt que selon la formule (4.5) proposée par Vlachos *et al.*.

$$LCSS_{\delta,\epsilon}(A,B) = \begin{cases} 0 \quad \text{si } A \text{ ou } B \text{ est vide,} \\ \\ 1 + LCSS_{\delta,\epsilon}(Head(A), Head(B)), \\ \quad \text{si } d\left(a_{x_k,n}, b_{x_k,m}\right) < \epsilon, \forall 1 \leq k \leq p, \\ \quad \text{et } |n-m| \leq \delta \text{ et } |N-n-M+m| \leq \delta, \\ \\ max\left(LCSS_{\delta,\epsilon}(Head(A), B), LCSS_{\delta,\epsilon}(A, Head(B))\right) \\ \quad \text{sinon.} \end{cases} \tag{4.9}$$

4.2.3 Calcul effectif de la distance *LCSS*

Le calcul d'une distance entre trajectoires nécessite ainsi d'identifier au moins la longueur de leur plus longue sous-séquence commune ($LCSS$). L'implémentation naïve de la fonction récursive proposée est de complexité exponentielle dans le cas où les deux sous-séquences n'ont pas

d'éléments en commun. De nombreux algorithmes ont ainsi été proposés pour résoudre ce problème, dont la plupart ont leur prédécesseur dans les travaux soit de Hunt et Szymanski [52], soit de Hirschberg [47].

Dans notre contexte, on implémente une variante d'un algorithme de Hirschberg [47] proposée par Apostolico [4]. Le complexité du problème est également diminuée en examinant uniquement les paires de points qui vérifient la contrainte temporelle décrite par (4.7). Ainsi, l'algorithme devient très efficace dès que la constante δ est petite, autorisant un court écart temporel entre deux valeurs similaires.

Des détails sur le problème de recherche de la plus longue sous-séquence commune à deux séquences et sur l'algorithme implémenté sont présentés en annexe H.

4.3 Distance minimum

Une distance minimum entre deux séquences temporelles de vecteurs discrétisés est définie comme mesure approchée de leur similarité. Cette distance est utile à l'abstraction d'une séquence initiale en une séquence de symboles représentatifs d'un "état stationnaire" des paramètres observés, à l'échelle de décision. En particulier, à partir d'une représentation discrète d'une séquence initiale, l'objectif est d'agréger les vecteurs successifs en un seul symbole qui les représente tant qu'ils ne correspondent à aucune variation significative dans le temps. L'évaluation de la possibilité d'agrégation d'une succession de vecteurs discrets est réalisée par un seuil maximum sur une mesure approchée de distance entre cette succession de vecteurs et une séquence de même longueur composée uniquement du vecteur moyen correspondant. Il s'agit donc d'une mesure de distance entre deux séquences de vecteurs discrets de même longueur.

Chiu *et al.* [28] ont proposé et utilisé efficacement une distance minimum pour comparer grossièrement sur une dimension des séquences temporelles discrètes de même longueur, dans un objectif de classification. Une distance minimum entre séquences temporelles est une approximation par les valeurs inférieures de la distance réelle, intéressante pour avoir une idée grossière de leur similarité. De même que la distance réelle, une distance minimum varie entre 0 et 1. Une valeur nulle signifie que les deux séquences peuvent être considérées comme assez similaires. Notre contexte nécessite cependant d'étendre la distance minimum proposée dans [28] pour la prise en compte de séquences multidimensionnelles et hétérogènes.

4.3.1 Définition de la distance minimum

La distance minimum telle que proposée dans [28] est calculée sur des séquences de valeurs réelles discrétisées, à une dimension. Sa définition utilise les bornes des intervalles de discrétisation. Notons $B = \beta_1, ..., \beta_{a-1}$ la liste ordonnées par valeurs croissantes des bornes de ces intervalles pour une paramètre quantitatif discrétisé en a symboles α_1,...,α_a, avec $a \geq 1$. β_0 et β_a sont définis comme $-\infty$ et $+\infty$ respectivement. Une séquence $C = c_1, ..., c_n$ de longueur n peut être ainsi transformée en sa représentation discrète $\hat{C} = \hat{c}_1, ..., \hat{c}_\omega$, où $\hat{c}_i = \alpha_j$ si et seulement si $\beta_{j-1} \leq c_i < \beta_j$. En utilisant le principe de la distance euclidienne, la distance minimum entre les séries temporelles initiales Q et C, correspondant à deux représentations discrètes $\hat{Q}$ et $\hat{C}$, appelée $mindist(\hat{Q}, \hat{C})$, est définie selon l'équation (4.10) [28].

$$mindist(\hat{Q}, \hat{C}) = \sqrt{\frac{n}{\omega}} \sqrt{\sum_{i=1}^{\omega} (d(\hat{q}_i, \hat{c}_i))^2} \tag{4.10}$$

Le rapport $\frac{n}{\omega}$ représente le taux de compression des séquences dans le temps au moment de la discrétisation, où n est la longueur des séquences initiales Q et C, et w celle de leur représentation symbolique $\hat{Q}$ et $\hat{C}$. On a $n \neq w$ s'il y a une réduction de dimension au moment de la discrétisation. La distance $d(\hat{q}_i, \hat{c}_i)$ représente la distance entre les $i^{èmes}$ valeurs discrètes des séquences $\hat{Q}$ et $\hat{C}$, respectivement $\hat{q}_i$ et $\hat{c}_i$. Supposons que ces valeurs discrètes correspondent aux symboles α_k et α_p de l'alphabet ordonné α_1,...,α_a, $1 \leq k, p \leq a$. La distance $d(\alpha_k, \alpha_p)$ entre ces deux symboles est définie à partir des bornes des intervalles de discrétisation correspondant à chaque symbole, d'après l'équation (4.11) [28].

$$d(\alpha_k, \alpha_p) = \begin{cases} 0 & \text{si } |k-p| \leq 1 \ , \\ \beta_{max(k,p)-1} - \beta_{min(k,p)} & \text{sinon.} \end{cases} \tag{4.11}$$

L'implémentation d'un tableau permettant d'avoir une vue d'ensemble des distances entre symboles est illustrée dans le tableau 4.1. La distance entre deux symboles est nulle tant que ces symboles correspondent à des intervalles de valeurs adjacents.

	α_1	α_2	α_3	α_4
α_1	0	0	0.25	0.57
α_2	0	0	0	0.32
α_3	0.25	0	0	0
α_4	0.57	0.32	0	0

Tab. 4.1 – Distances entre symboles utilisées pour le calcul de distance minimum entre deux séquences discrètes.
On considère un alphabet de cardinal 4, $\alpha_1, ...\alpha_4$ dont les symboles sont obtenus par discrétisation de l'intervalle $[0, 1]$ selon les limites $\beta_1 = 0.12$, $\beta_2 = 0.37$, et $\beta_3 = 0.69$. La distance entre deux symboles est lue à l'intersection des ligne et colonne correspondantes. Par exemple, $d(\alpha_1, \alpha_2) = 0$ et $d(\alpha_1, \alpha_3) = 0.25$.

4.3.2 Extension aux contraintes multidimensionnelles et hétérogènes

Pour étendre la définition d'une distance minimum aux séquences temporelles multidimensionnelles et hétérogènes, on prend en compte la distance minimum entre les symboles correspondant à chaque composante des vecteurs définissant la séquence. Notons dans un contexte multidimensionnel $C = ((c_{1,1}, ..., c_{1,p}), ..., (c_{n,1}, ..., c_{n,p}))$ et $Q = ((q_{1,1}, ..., q_{1,p}), ..., (q_{n,1}, ..., q_{n,p}))$ deux séquences temporelles hétérogènes de longueur n et de dimension p représentées respectivement par les séquence de symboles, de longueur ω, $\hat{C} = ((\hat{c}_{1,1}, ..., \hat{c}_{1,p}), ..., (\hat{c}_{\omega,1}, ..., \hat{c}_{\omega,p}))$ et $\hat{Q} = ((\hat{q}_{1,1}, ..., \hat{q}_{1,p}), ..., (\hat{q}_{\omega,1}, ..., \hat{q}_{\omega,p}))$. La distance minimum $mindist(\hat{Q}, \hat{C})$ entre les deux séquences initiales Q et C, représentées par $\hat{Q}$ et $\hat{C}$, est alors redéfinie selon l'équation (4.12).

$$mindist(\hat{Q}, \hat{C}) = \sqrt{\frac{n}{\omega}} \sqrt{\sum_{i=1}^{\omega} \left(\sum_{j=1}^{p} \left(d(\hat{q}_{i,j}, \hat{c}_{i,j}) \right)^2 \right)}. \tag{4.12}$$

Dans un contexte hétérogène, la fonction $d(\hat{q}_{i,j}, \hat{c}_{i,j})$ définissant la distance entre les valeurs sur la $j^{ème}$ composante des $i^{ème}$ vecteurs discrets, $\hat{q}_{i,j}$ et $\hat{c}_{i,j}$, dépend du type de la $j^{ème}$ composante.

- **Paramètres quantitatifs.** La distance entre deux valeurs discrétisées est définie selon l'équation (4.11) proposée par Chiu *et al.* [28], et illustrée par le tableau 4.1.

- **Paramètres qualitatifs.** Il n'y a pas de discrétisation et la distance entre deux modalités est celle définie au paragraphe 4.1. L'implémentation d'un tableau permettant d'avoir une vue d'ensemble des distances entre modalités est illustrée dans les tableaux 4.2 et 4.3, respectivement pour un paramètre à modalités non ordonnées et ordonnées.

	α_1	α_2	α_3	α_4
α_1	0	1	1	1
α_2	1	0	1	1
α_3	1	1	0	1
α_4	1	1	1	0

TAB. 4.2 – Distances entre modalités utilisées pour le calcul de distance minimum entre deux séquences discrètes.
On considère un ensemble fini de modalités non ordonnées, de cardinal 4, $\alpha_1, ...\alpha_4$.

	α_1	α_2	α_3
α_1	0	0.5	1
α_2	0.5	0	0.5
α_3	1	0.5	0

TAB. 4.3 – Distances entre modalités utilisées pour le calcul de distance minimum entre deux séquences discrètes.
On considère un ensemble fini de modalités ordonnées, de cardinal 3, $\alpha_1 < \alpha_2 < \alpha_3$.

Finalement, la définition de la fonction $d(\hat{q}_{i,j}, \hat{c}_{i,j})$ en fonction du type de la $j^{ème}$ composante des vecteurs est résumée ci-dessous. L'alphabet ordonné des a symboles de discrétisation d'un paramètre quantitatif est noté α_1,...,α_a, $a \geq 1$; et v est le nombre de modalités dans le cas d'un paramètre qualitatif. Pour assurer un calcul de distance homogène – toutes les distances sont comprises entre 0 et 1 – on utilise les valeurs normalisées des paramètres quantitatifs.

$$\begin{aligned}
\text{Quantitatif} \quad d(\hat{q}_{i,j}, \hat{c}_{i,j}) = & \begin{cases} 0 & \text{si } |k-p| \leq 1 \\ \beta_{max(k,p)-1} - \beta_{min(k,p)} & \text{sinon} \end{cases} \\
& \text{ssi } \hat{q}_{i,j} = \alpha_k \text{ et } \hat{c}_{i,j} = \alpha_p \\
\text{Qualitatif, ordonné} \quad d(\hat{q}_{i,j}, \hat{c}_{i,j}) = & \frac{|\hat{q}_{i,j} - \hat{c}_{i,j}|}{v-1} \\
\text{Qualitatif, non ordonné} \quad d(\hat{q}_{i,j}, \hat{c}_{i,j}) = & \; min(|\hat{q}_{i,j} - \hat{c}_{i,j}|, 1)
\end{aligned}$$

4.4 Synthèse

Dans ce chapitre on a ainsi défini deux mesures de similarité, dont le principe et le contexte d'utilisation sont résumés dans le tableau 4.4.

Distance approchée (Abstraction)	**Contexte :** Séquences multidimensionnelles (dimension p), de même longueur ω, où les paramètres quantitatifs sont discrétisés : $\hat{Q} = ((\hat{q}_{1,1}, ..., \hat{q}_{1,p}), ..., (\hat{q}_{\omega,1}, ..., \hat{q}_{\omega,p}))$, $\hat{C} = ((\hat{c}_{1,1}, ..., \hat{c}_{1,p}), ..., (\hat{c}_{\omega,1}, ..., \hat{c}_{\omega,p}))$
	Paramètres : –
	Définition : $mindist(\hat{Q}, \hat{C}) = \sqrt{\frac{n}{\omega}} \sqrt{\sum_{i=1}^{\omega} \left(\sum_{j=1}^{p} \left(d(\hat{q}_{i,j}, \hat{c}_{i,j}) \right)^2 \right)}$, où n est la longueur des séquences avant discrétisation. et w la longueur des séquences discrétisées.
Distance réelle (Fouille de données)	**Contexte :** Séquences multidimensionnelles (dimension p), hétérogènes, de longueur différente n et m : $A = ((a_{x_1,1}, \ldots, a_{x_p,1}), \ldots, (a_{x_1,n}, \ldots, a_{x_p,n}))$, $B = ((b_{x_1,1}, \ldots, b_{x_p,1}), \ldots, (b_{x_1,m}, \ldots, b_{x_p,m}))$
	Paramètres : Contraintes de similarité entre deux points, dans le temps (δ) et sur les valeurs (ϵ)
	Définition : $D_{\delta,\epsilon}(A, B) = 1 - \frac{LCSS_{\delta,\epsilon}(A,B)}{min(n,m)}$, où $LCSS_{\delta,\epsilon}(A, B)$ est le nombre de points similaires entre A et B au sens de la nouvelle définition proposée.

TAB. 4.4 – Synthèse des mesures de similarité définies.

5

Approche proposée pour l'extraction de motifs temporels

Dans ce chapitre nous décrivons successivement les grandes étapes de l'approche proposée pour l'extraction de motifs temporels, dans l'objectif ultime de surveillance des tendances critiques de variation des paramètres considérés. Il s'agit alors d'identifier des **motifs multidimensionnels** qui peuvent être de "**haut niveau**" par rapport au niveau des données brutes, avec la possibilité de considérer des **composantes hétérogènes**. On propose par ailleurs une approche complètement **non supervisée**, et appropriée à la **présence de bruit** entre les instances des motifs – longueurs différentes, variabilité dans les valeurs, présence de valeurs imprévisibles.

Les paragraphes suivants présentent, selon le schéma d'un système de reconnaissance approprié à l'analyse de données temporelles proposé sur la figure 3.1 (cas **C)**), une approche d'extraction en trois principales étapes.

- La section 5.1 présente la première étape d'**abstraction**, qui permet de représenter de façon significative pour la découverte de motifs les données issues des capteurs.
- La section 5.2 propose une méthode de **fouille de caractères** ayant pour objectif l'extraction des sous-séquences répétitives, qui sont celles les plus susceptibles de correspondre aux instances de motifs.
- La section 5.3 présente enfin la **classification** non supervisée de ces caractères – appelés *tentatives de motifs* – en classes de sous-séquences similaires – les *motifs*.

La nécessité d'une approche complètement non supervisée impose une méthode non supervisée d'une part pour la fouille de caractères définissant les tentatives de motifs et d'autre part pour la classification en motifs.

Dans notre contexte expérimental, les *tentatives de motifs* sont ainsi représentatives des comportements répétitifs d'une personne à domicile. Leur durée doit être représentative de la description d'une activité, quelle que soit la fréquence. Les *motifs* sont ensuite définis comme les classes significatives de *tentatives de motifs* – en particulier, une classe doit être suffisamment instanciée dans la séquence initiale. Chaque classe ou *motif* représente alors un comportement particulier fréquemment observé pour la personne télésurveillée.

5.1 Abstraction des séquences temporelles

L'abstraction est une représentation "haut niveau" qui, plus qu'une simple transformation de l'espace ou simple réduction de dimension, inclut l'interprétation des données pour leur donner un sens et un niveau de détail adapté au regard de l'objectif de leur analyse. Lorsque les données initiales sont très "bas niveau" – les données issues des capteurs – et la décision plutôt "haut niveau", l'objectif est alors de mettre en évidence les tendances à plus ou moins long terme dans les séquences temporelles initiales. L'abstraction permet ainsi de s'affranchir d'une partie du bruit présent dans les données analysées au regard de la décision. On considère ainsi les étapes suivantes d'abstraction :

1. **Prétraitement** des données brutes pour s'affranchir de la variabilité très haute fréquence ;
2. **Discrétisation** des données quantitatives pour obtenir une représentation homogène des données multidimensionnelles initialement hétérogènes ;
3. **Agrégation** des séquences de vecteurs discrets tant que les variations entre une séquence de vecteurs et le vecteur moyen correspondant ne sont pas trop importantes ;

Les paragraphes suivants détaillent le principe de ces différentes étapes. Leur validité est illustrée en commentée dans le paragraphe 6.2.2 du chapitre 6 présentant l'ensemble des résultats expérimentaux.

5.1.1 Prétraitement

Le prétraitement des données brutes inclut l'alignement temporel des différentes composantes, une éventuelle réduction temporelle, un filtrage pour le lissage des variations hautes fréquences, et la normalisation des valeurs des paramètres quantitatifs. Bien que ces opérations de prétraitement soient bien connues et habituelles, elles sont néanmoins véritablement importantes car elles gouvernent au moins en partie le niveau de détail de l'analyse. Étant donné que le prétraitement est suivi par l'extraction des caractères significatifs permettant une réduction des dimensions temporelle et spatiale, l'alignement temporel est réalisé sans réduction sur la fréquence maximum commune à l'ensemble des paramètres. Pour le filtrage, on choisit une méthode de moyenne mobile pondérée pour ne pas trop "aplatir" la courbe et conserver au maximum les "pics" de valeurs significatifs. La normalisation est ensuite réalisée par la méthode du min-max décrite au paragraphe 4.1.

5.1.2 Discrétisation

La discrétisation de séries temporelles permet d'étendre les possibilités d'utilisation d'algorithmes d'apprentissage et de fouille de données. Dans un contexte hétérogène, elle permet également de générer une représentation homogène des données multidimensionnelles initialement hétérogènes. Elle est ainsi nécessaire avant toute autre opération d'abstraction afin de pouvoir facilement considérer l'ensemble des paramètres pour construire une représentation significative en terme de leurs variations conjointes. Les séquences initiales s'expriment alors comme une succession de vecteurs discrets dans le temps.

Détermination des intervalles de discrétisation
La détermination d'intervalles de discrétisation peut être complètement intuitive ou empirique, reposer sur des contraintes spécifiques telles que l'équiprobabilité des symboles [28, 69] ou une variance maximum pour les valeurs associées à chaque symbole [48], ou encore sur une classification supervisée ou non des valeurs possibles [33, 48]. Dans un contexte de surveillance, les paramètres

quantitatifs sont normalisés par une méthode du min-max (voir section 4.1) et, contrairement aux constatations de [69], les distributions de valeurs n'ont aucune raison d'être normales, ou même plus simplement symétriques. D'après la construction d'un modèle de simulation dans le cadre de la télésurveillance médicale, on constate d'ailleurs qu'elles ne le sont pas pour la surveillance du niveau d'activité par exemple (voir section II.5.2.3).

Afin de fournir une abstraction significative et interprétable des données, il est ainsi intéressant de conserver les spécificités de distribution des valeurs des différents paramètres surveillés au niveau de leur représentation discrète. La génération de valeurs équiprobables n'est alors pas appropriée, et on préfère apprendre les classes ou intervalles caractéristiques de variation des valeurs observées à partir de l'algorithme des *k plus proches voisins*. Le nombre d'intervalles de discrétisation qu'il est pertinent de considérer est par contre déterminé empiriquement ou avec l'aide d'experts.

5.1.3 Agrégation temporelle

Dans le contexte d'une étude "haut niveau" par rapport aux données exploitées, une simple discrétisation ne suffit pas forcément à une abstraction des séquences temporelles appropriée aux objectifs de décision. Elle ne permet en effet pas de s'affranchir de l'ensemble des variations non significatives à l'échelle de la décision.

Ainsi, afin d'obtenir une représentation concise des séquences pour une étude "haut niveau", l'idée est d'identifier des segments temporels de longueur variable durant lesquels l'état des différents paramètres est stationnaire au regard du niveau de décision. La complexité des variations relatives des paramètres est par ailleurs préservée dans la considération de la stationnarité conjointe des différents paramètres. On propose alors d'agréger dans le temps les sous-séquences correspondantes de vecteurs discrets, c'est-à-dire les vecteurs successifs observés tant que l'écart avec leur vecteur moyen n'est pas trop important. Chacune de ces sous-séquences identifiant un état stationnaire des paramètres est ainsi synthétisée en un symbole. Les symboles sont les vecteurs moyens discrets sur les segments d'agrégation, de même dimension que le nombre de paramètres étudiés, et associés chacun à une durée spécifique.

Définition d'une séquence temporelle agrégée

L'agrégation est réalisée en fonction d'un seuil maximum σ sur une mesure de distance minimum discrète (voir 4.3) entre une sous-séquence discrète, notée $\hat{C}$, et son vecteur moyen discret, noté $\widehat{aggr(C)}$. Considérons un espace de dimension p, éventuellement hétérogène, et une sous-séquence C d'une trajectoire donnée, de longueur n, définie dans cette espace par :

$$C = \left(\begin{pmatrix} c_{1,1} \\ \vdots \\ c_{1,p} \end{pmatrix}, \cdots, \begin{pmatrix} c_{n,1} \\ \vdots \\ c_{n,p} \end{pmatrix} \right)$$

Les composantes j, $1 \leq j \leq p$, du vecteur moyen de C, noté $aggr(C)$, correspondent à la moyenne, notée $moy()$, des valeurs observée sur chaque composante : $aggr(C)_j = moy(c_{1,j}, \ldots, c_{n,j})$. Pour les composantes qualitatives, la valeur moyenne correspond à la modalité la plus représentée. Le vecteur moyen discret, $\widehat{aggr(C)}$, correspond alors à la discrétisation du vecteur moyen :

$$\widehat{aggr(C)} = \begin{pmatrix} \widehat{moy\,(c_{1,1}, ..., c_{n,1})} \\ \vdots \\ \widehat{moy\,(c_{1,p}, ..., c_{n,p})} \end{pmatrix}$$

La condition d'agrégation des vecteurs de la sous-séquence discrète $\hat{C}$ est alors définie par la relation 5.1,

$$mindist(\hat{C}, \widehat{AGGR}(C)) \leq \sigma, \tag{5.1}$$

où $\widehat{AGGR}(C)$ est une séquence de même longueur n que $\hat{C}$, composée uniquement de la répétition du vecteur moyen discrétisé $\widehat{aggr}(C)$:

$$\widehat{AGGR}(C) = (\widehat{aggr}(C), ..., \widehat{aggr}(C)),$$

et $mindist()$ est la fonction de distance minimum discrète (voir 4.3). À partir du premier point de la séquence, on cherche ainsi les plus longs intervalles de temps pour lesquelles l'agrégation temporelle est permise selon les définitions précédentes. Finalement, la séquence initiale est ainsi représentée par une succession de vecteurs discrets – appelés symboles – dont chacun correspond à une durée spécifique.

5.2 Fouille de caractères pour l'extraction de tentatives de motifs

5.2.1 Objectifs et contraintes de la fouille de caractères

La fouille de caractères a pour objectif l'extraction des sous-séquences récurrentes les plus significatives au regard de la classification, afin de réduire l'espace de recherche des motifs. D'après [67], les critères de sélection des caractères dépendent du domaine d'application et du classifieur utilisé. Il existe cependant trois heuristiques qui guident leur sélection quel que soit le contexte :

(1) les caractères doivent être fréquents,

(2) caractéristiques d'au moins une classe,

(3) non redondants.

D'après les résultats de l'abstraction (voir 5.1), la fouille de caractères pour l'identification de sous-séquences récurrentes est réalisée sur une succession de symboles – vecteurs discrets – estampillés en fonction de la durée qui leur est associée.

Dans le contexte de l'extraction des sous-séquences récurrentes présentes dans des séries temporelles à une dimension, Chiu *et al.* [28] ont expérimenté un algorithme utilisant les *projections aléatoires* de l'ensemble des sous-séquences discrètes issues de la série temporelle analysée, tel qu'illustré sur la figure 5.1. Des sous-séquences discrètes, dites *sous-séquences de base*, sont extraites par une fenêtre glissante de longueur fixe sur la série temporelle initiale. À chaque position de la fenêtre glissante (cas de C_1), la sous-séquence discrète associée ($\hat{C}_1$) est calculée à partir des valeurs moyennes observées sur chaque segment de discrétisation composant la fenêtre, selon un ensemble de symboles équiprobables (3 symboles **a**, **b** et **c** dans cet exemple). La projection des sous-séquences a alors le sens d'une réduction de dimension. Elle est définie comme le masquage d'un nombre fixé *a priori* de symboles pour la comparaison des sous-séquences discrètes deux à deux. La position de ce masque sur les sous-séquences est déterminée aléatoirement à chaque projection. Les projections sont ainsi réalisées sur une dimension inférieure à la longueur de chaque sous-séquence, à partir d'un masque aléatoire de projection dont la dimension est prédéfinie. Elles sont itérées un certain nombre de fois. À chaque projection, les collisions entre sous-séquences – égalité des sous-séquences projetées – sont enregistrées dans une matrice de collisions. Les valeurs élevées de cette matrice correspondent ainsi à une forte présomption de similarité entre les sous-séquences correspondantes.

Cette méthode de projections aléatoires, utilisée pour la première fois pour la recherche de motifs dans des séquences de nucléotides [17], est intéressante car elle permet l'extraction

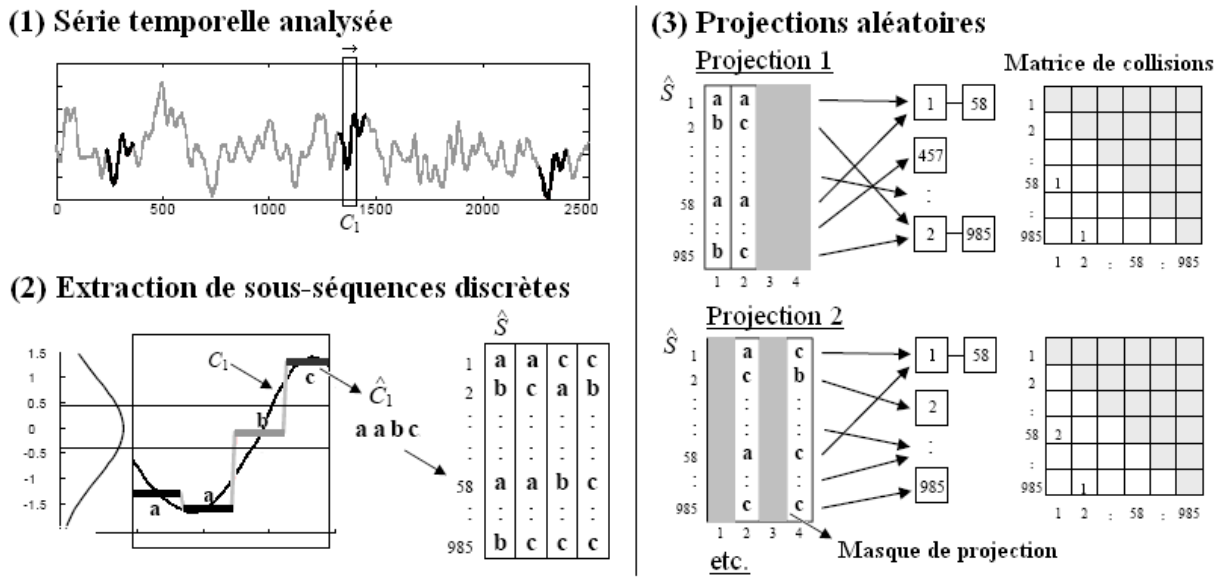


FIG. 5.1 – Principe d'utilisation de la méthode des projections aléatoires pour l'identification des sous-séquences récurrentes d'une série temporelle à une dimension [28].
Les étapes successives sont les suivantes :

(1) Définition d'une largeur de fenêtre glissante et d'un pas de déplacement sur la série temporelle analysée ;

(2) Extraction des sous-séquences discrètes pour chaque position de la fenêtre, selon les valeurs moyennes observées sur des segments temporels de longueur fixe. Selon le pas de déplacement de la fenêtre, les sous-séquences discrètes successives peuvent n'avoir aucun symbole en commun ;

(3) Projections aléatoires à partir de l'ensemble des sous-séquences discrètes, et selon un masque de projections de dimension prédéfinie et dont la position est sélectionnée aléatoirement à chaque projection. La matrice de collisions est mise à jour selon les sous-séquences projetées identiques.

de caractères en autorisant du bruit et de l'imprécision entre les sous-séquences récurrentes des séquences discrètes observées. Ces caractéristiques *de bruit* et d'*imprécision* sont justement particulièrement présentes entre les instances des motifs recherchés dans notre contexte. Il faut cependant chercher à étendre cet algorithme à deux niveaux :

– pour la considération de séquences de symboles multidimensionnels,
– et pour l'identification de sous-séquences récurrentes pouvant être de différentes longueurs.

Dans [28], chaque valeur discrète est en effet définie sur un intervalle de temps de longueur fixe. Dans notre contexte, l'abstraction génère par contre des séquences de symboles associés chacun à une durée spécifique. Ainsi, l'application de projections aléatoires à ces séquences de symboles répond déjà au moins en partie à ce dernier critère d'extraction de sous-séquences récurrentes de différentes longueurs.

Les critères sur la définition des caractères, proposés dans [67], ne sont cependant pas tous vérifiés par la seule application de projections aléatoires.

(1) Fréquence. Le critère de fréquence des caractères est vérifié puisqu'il est inhérent à l'algorithme de projections. Les sous-séquences extraites correspondent en effet à un nombre important ou au moins strictement positif de collisions. Cela signifie que chacune est similaire à au moins une autre sous-séquence.

(2) Signification. La signification des caractères en terme d'instances de motifs ne peut pas être assurée directement à l'issue des projections puisque cette méthode donne simplement une *idée* des récurrences dans la séquence initiale, à une *échelle de temps limitée* par le nombre prédéfini de symboles successifs composant les sous-séquences. Une opération supplémentaire est ainsi nécessaire au moment de l'examen de la matrice de collisions pour assurer la pertinence des tentatives de motifs, en trois étapes successives.

- **Distance réelle.** Puisque qu'une importante valeur de collisions n'est qu'un indicateur de possible similarité, on vérifie la proximité effective par un calcul de distance réelle entre les sous-séquences, à partir des données initiales prétraitées auxquelles elles correspondent.
- **Extension.** Étant données les durées variables des instances de motifs, et la présence éventuelle d'interruptions, il convient de fixer la longueur des sous-séquences de base considérées pour les projections aléatoires au nombre de symboles décrivant *a priori* la plus courte représentation d'une instance de motifs. Les tentatives de motifs peuvent ainsi correspondre en fait à plusieurs sous-séquences de base successives, et générer par conséquent plusieurs valeurs adjacentes importantes dans la matrice de collisions. On propose alors une méthode extensive d'examen de la matrice de collision pour identifier des sous-séquences récurrentes de longueur variable, en terme de leur durée effective et du nombre de symboles qui les représente. L'extension de sous-séquences *de base* à des sous-séquences significatives est illustrée sur la figure 5.2.
- **Seuil minimum de durée.** Un seuil minimum sur la durée des tentatives de motifs permet de ne sélectionner que les sous-séquences les plus significatives au niveau de leur durée par rapport aux objectifs de l'application.

(3) Redondance. Les sous-séquences de base considérées sont extraites à partir d'une fenêtre glissante sur les données initiales, si bien qu'elles se recouvrent forcément. Par ailleurs, une sous-séquence peut-être similaire à plusieurs autres sous-séquences et être ainsi identifiée plusieurs fois comme récurrente au moment de l'examen de la matrice de collisions, mais pas forcément exactement selon les mêmes délimitations. Le critère de non redondance impose alors de définir une méthode de synthèse de l'ensemble des sous-séquences récurrentes identifiées après examen de la matrice de collisions afin de générer finalement un groupe de sous-séquences disjointes – les *tentatives de motifs*.

La fouille de caractères est ainsi réalisée en trois étapes successives, agissant respectivement sur les critères de fréquence, signification et redondance, et détaillées dans les paragraphes suivants :

- **5.2.2. Projections aléatoires** à partir d'une séquence de symboles multidimensionnels, afin d'extraire des sous-séquences **récurrentes** ;
- **5.2.3. Examen de la matrice de collisions** pour l'identification d'un ensemble de sous-séquences récurrentes dans la séquences initiales, de différentes longueurs, et **significatives** au regard des objectifs de l'application ;
- **5.2.4. Synthèse des tentatives de motifs** à partir des sous-séquences récurrentes identifiées pour l'extraction de caractères **non redondants**.

A l'issue de ces étapes, les tentatives de motifs identifiés vérifient les critères de fréquence, pertinence et non redondance.

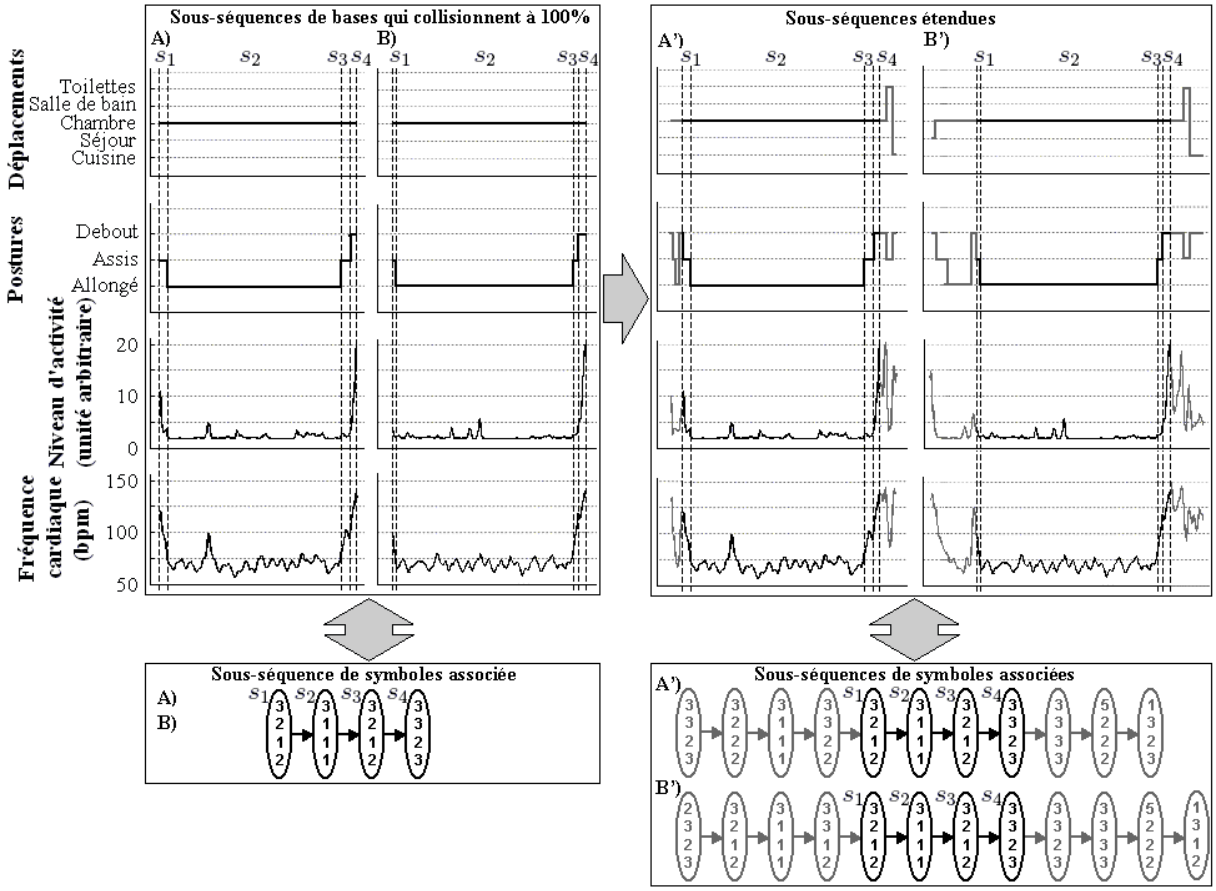


FIG. 5.2 – Illustration de l'extension nécessaire des sous-séquences de base similaires identifiées par l'examen des fortes valeurs de la matrice de collisions.
L'illustration proposée concerne des séquences hétérogènes à quatre dimensions, incluant (1) les déplacements, (2) les postures, (3) le niveau d'activité et (4) la fréquence cardiaque moyenne.
La figure de gauche présente deux sous-séquences de base correspondant à un nombre maximum de collisions : les séquences de symboles qui les représentent sont donc strictement identiques.
La figure de droite présente leur extension vers des sous-séquences significatives au regard de la décision. Les séquences discrètes associées diffèrent en partie en terme des symboles qu'elles contiennent et de leur nombre. Leur distance réelle reste cependant inférieure au seuil maximum de distance fixé *a priori*.

5.2.2 Projections aléatoires

L'avantage d'appliquer des projections aléatoires pour comparer des sous-séquences discrètes est d'autoriser le rapprochement de sous-séquences même si elles ne sont pas parfaitement identiques en terme de la succession de symboles qui les représente. On ne compare en effet, à chaque projection, qu'une partie des symboles composant les sous-séquences de base. D'une part on permet ainsi un certain taux de bruit entre des sous-séquences pourtant considérées similaires ; et d'autre part, on gagne en flexibilité sur la précision nécessaire pour l'abstraction. Les étapes suivantes de traitement permettent ensuite d'affiner la sélection des tentatives de motifs. L'idée est alors de ne pas en manquer dès les premières étapes d'analyse et de rejeter progressivement les sous-séquences qui ne sont finalement pas pertinentes lors des étapes suivantes.

Pour répondre aux contraintes spécifiques de notre contexte, l'algorithme de projections aléatoires proposé et implémenté dans [17, 28] est étendu à la considération de symboles multidimensionnels. Son contexte d'application sur une séquence de symboles dont chaque instance est associée à une durée spécifique rend par ailleurs possible la découverte de motifs dont les instances sont de différentes longueurs.

Les projections aléatoires d'un ensemble de *sous-séquences de base* produisent une *matrice de collisions*, carrée, symétrique, de dimension égale au nombre de sous-séquences de base définies à partir d'une fenêtre glissante sur la séquence initiale de symboles. Cette matrice comptabilise le nombre de "*collisions*" deux à deux entre les sous-séquences projetées, c'est-à-dire le nombre de fois où deux sous-séquences sont identiques en terme de la comparaison des symboles projetés uniquement. Dans un contexte multidimensionnel, on définit la projection d'une sous-séquence comme le masquage d'un nombre fixé *a priori* de symboles et de paramètres par symbole, pour sa comparaison avec les autres sous-séquences. La position de ce masque multidimensionnel sur chaque sous-séquence est déterminée aléatoirement à chaque itération des projections. À chaque projection, le nombre de collisions est incrémenté en fonction de l'égalité des sous-séquences projetées. Ainsi, si deux sous-séquences de symboles sont absolument identiques, le nombre de collisions est maximum, c'est-à-dire égal au nombre de projections.

Le schéma de la figure 5.3 résume le principe des projections aléatoires, selon les étapes suivantes :

(1) Séquence prétraitée. C est une séquence prétraitée de dimension p et de longueur n :

$$C = ((c_{1,1}, \ldots, c_{1,p}), \ldots, (c_{n,1}, \ldots, c_{n,p})).$$

(2) Abstraction. La séquence C est représentée par abstraction en une séquence de N symboles estampillés par t_i, $1 \leq i \leq N$, $N \leq n$, $(t_1, \ldots, t_N)$ étant les instants d'occurrence des symboles ordonnés dans le temps. Les symboles sont des vecteurs discrets de dimension p, $\hat{q}_{i,j}$, $1 \leq i \leq N$, $1 \leq j \leq p$:

$$\hat{C} = (((\hat{q}_{1,1}, \ldots, \hat{q}_{1,p}), t_1), \ldots, ((\hat{q}_{N,1}, \ldots, \hat{q}_{N,p})), t_N).$$

(3) Projections aléatoires.

(a) Sous-séquences. Les projections aléatoires sont réalisées sur les sous-séquences de base de w symboles extraites par une fenêtre glissante de longueur w sur la séquence $\hat{C}$ de N symboles représentant la séquence initiale C. On obtient ainsi une matrice $\hat{S}$ de dimension $(N - w + 1) \times w$.

(b) Masque de projections. On sélectionne aléatoirement un masque de dimension $w_{mask} \times p_{mask}$, où w_{mask} et p_{mask} sont des entiers tels que $0 \leq w_{mask} < w$ et $0 \leq$

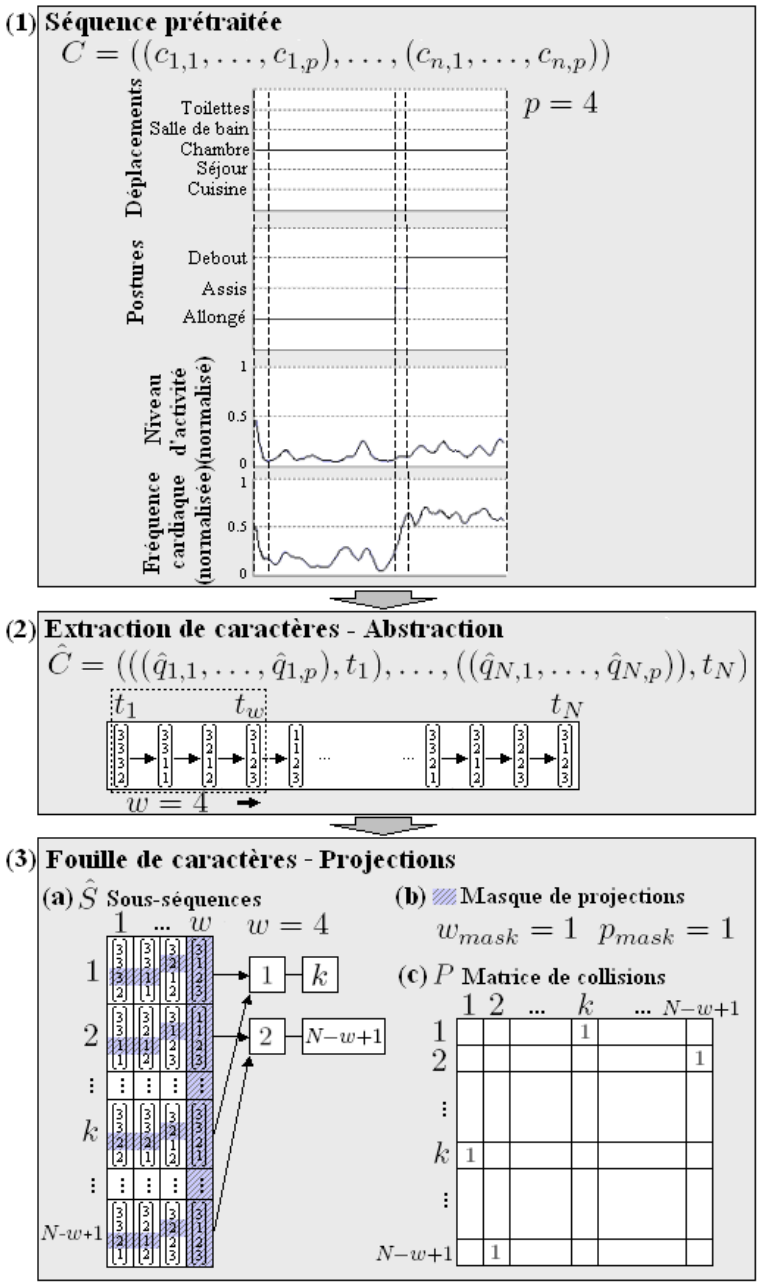


FIG. 5.3 – Principe de la méthode des projections aléatoires proposée.

(1) Les projections aléatoires sont réalisées à partir d'une séquence C de dimension p et longueur n.

(2) La séquence C est représentée par $\hat{C}$, une succession de N symboles estampillés. Chaque symbole multidimensionnel, $(\hat{c}_{k,1}, ..., \hat{c}_{k,p})$, $1 \leq k \leq n$, a une durée t_k.

(3) Les étapes successives et répétées des projections sont les suivantes :

- **(a)** Construction de la matrice $\hat{S}$ de toutes les $n-w+1$ sous-séquences possibles en faisant glisser une fenêtre de longueur w sur la séquence de N symboles ;
- **(b)** Construction du masque de projections de w_{mask} des w symboles, et de p_{mask} des p paramètres par symbole projeté.
- **(c)** Répétition de projections successives des sous-séquences de la matrice $\hat{S}$ pour construire la matrice des collisions P, de dimensions $(n-w+1) \times (n-w+1)$.

$p_{mask} < p$, correspondant respectivement au nombre de symboles et de paramètres par symbole masqués à chaque projection. Sur la figure 5.3 où $w = 4$ et $p = 4$, le masque correspond à $w_{mask} = 1$ symbole par sous-séquence, et $p_{mask} = 1$ paramètre par symbole. Par exemple le $4^{ème}$ symbole a été aléatoirement sélectionné pour être masqué, et pour les 3 autres symboles respectivement les $3^{ème}$, $3^{ème}$ et $2^{ème}$ paramètres sont masqués.

(c) Matrice de collisions. Les $(N - w + 1)$ sous-séquences de $\hat{S}$ sont ensuite associées en fonction de l'égalité ou non de leurs valeurs projetées – c'est-à-dire non masquées – et la matrice de collisions P, de dimension $(N - w + 1) \times (N - w + 1)$, est mise à jour selon le principe suivant : si deux sous-séquences i et j sont identiques, on incrémente la valeur correspondante de la matrice de collisions, $P(i, j) = P(i, j) + 1$, sachant que la matrice initiale est nulle. Par exemple, sur le schéma proposé, les sous-séquences projetées 1 et k sont identiques, de même que les sous-séquences 2 et $(N - w + 1)$, et les valeurs $P(k, 1)$ et $P(N - w + 1, 2)$ initialement nulles sont positionnées à 1.

Les étapes (b) et (c) sont répétées un nombre approprié de fois pour que les nombres de collisions soient significatifs. La pertinence des collisions comptabilisées dépend également du choix approprié des paramètres impliqués en fonction des objectifs de l'extraction de motifs.

Ainsi, les paramètres clés de l'application de projections aléatoires sont les suivants :

- w, nombre de symboles par sous-séquence,
- w_{proj}, nombre de symboles projetés,
- p_{proj}, nombre de paramètres projetés pour chaque symbole,
- n_{proj}, nombre de projections.

5.2.3 Examen de la matrice de collisions

Principe général

Le résultat des projections successives est la constitution d'une *matrice de collisions* dont les valeurs donnent une forte indication des sous-séquences potentiellement similaires deux à deux, en fonction d'un *seuil minimum de collisions* défini *a priori*. Pour affiner l'identification des tentatives de motifs, la vérification de ces hypothèses de proximité est alors réalisée en fonction d'un *seuil maximum de distance réelle* entre ces paires de sous-séquences, défini également *a priori* (voir paragraphe 4.2). Les valeurs de la matrice de collisions sont ainsi examinées successivement à partir de la plus élevée. L'algorithme s'arrête après l'examen de toutes les valeurs supérieures au seuil minimum de collisions.

Par ailleurs, afin de tenter d'identifier des instances complètes de motifs, indépendamment du nombre de symboles considérés pour la définition des sous-séquences dites "de base" lors des projections aléatoires (w), on propose leur extension possible dès l'examen de la matrice de collisions. C'est ce que Hong et al. [49] appellent dans un autre contexte méthodologique la "croissance des motifs" – en anglais, *pattern growing* – afin d'identifier des instances complètes de motifs. Cette démarche permet également de simplifier par avance l'étape suivante de synthèse des sous-séquences pour constituer un groupe de tentatives de motifs significatifs et non redondants. Tant que les critères de collisions et de distances sont vérifiés sur les extensions possibles de deux sous-séquences comparées, on les étend alors "à droite" et/ ou "à gauche".

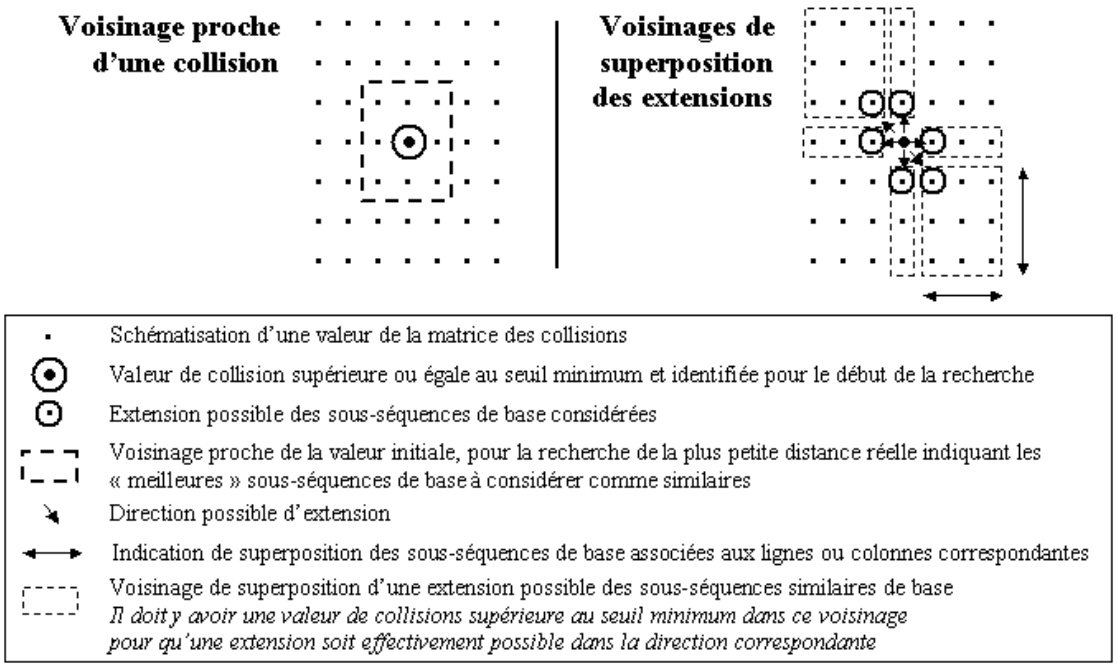


FIG. 5.4 – Voisinages de collisions pour l'identification et l'extension de sous-séquences récurrentes.

Adaptation aux bruits et imprécisions

Les principes de représentation d'une séquence initiale et de projections aléatoires des sous-séquences de base de longueur fixe identifiées dans cette séquence permettent une étude grossière des variations et récurrences. La présence de bruit entre des séquences similaires et la possible imprécision de leur représentation ne permet cependant pas d'être certain d'aboutir à une forte présomption de récurrence, selon le seuil minimum de collisions, pour l'ensemble des sous-séquences de base composant chaque séquence effectivement similaire à une autre. Afin d'affiner l'identification des sous-séquences récurrentes et de la rendre plus efficace en présence de bruit, on décide ainsi d'apporter un peu de souplesse dans l'examen de la matrice de collisions par rapport à l'indicateur de récurrence dû au nombre de collisions supérieur à un seuil prédéfini. L'identification des sous-séquences de base effectivement similaires et de leurs extensions possibles est ainsi fonction des valeurs supérieures à ce seuil observées dans un voisinage des points correspondants de la matrice de collisions.

Plus précisément, si on considère un nombre de collisions supérieur au seuil minimum et les deux sous-séquences de base correspondantes :

- L'identification plus précise des sous-séquences de base similaires est réalisée en recherchant la distance réelle la plus faible entre les sous-séquences de base correspondant à un nombre de collisions supérieur au seuil dans un **voisinage proche** de celles initialement identifiées.
- Dans le cas d'une extension, on autorise la vérification de la distance réelle entre deux sous-séquences étendues s'il existe un nombre de collisions supérieur au seuil dans les **voisinages de superposition** des extensions considérées, à gauche et à droite, pour chacune des deux sous-séquences.

Cet assouplissement de la démarche nécessite ainsi la définition de deux types de voisinages, illustrés sur la figure 5.4 :

- **Voisinage proche d'une collision – pour l'*identification des sous-séquences de base.***

  C'est l'ensemble des valeurs de collisions adjacentes à celle initialement identifiée, définissant neuf comparaisons possibles de sous-séquences de base. La définition de ce voisinage permet d'affiner l'identification des sous-séquences de base supposées similaires par la forte valeur de collisions observée. Pour chaque valeur de collisions supérieure au seuil minimum dans ce voisinage, on calcule la distance réelle entre les deux sous-séquences de base correspondantes, et on sélectionne finalement les sous-séquences de base correspondant à la plus petite distance réelle.

- **Voisinage de superposition d'une extension – pour l'*extension des sous-séquences de base similaires.***

  Les *directions possibles d'extension* correspondent dans la matrice de collisions aux valeurs en haut et/ ou à gauche, ainsi qu'en bas et/ ou à droite, par rapport aux sous-séquences de bases considérées. En se déplaçant à gauche et/ ou à droite dans cette matrice, on ajoute en effet un symbole respectivement à gauche et/ ou à droite à l'une des sous-séquences de base. De même, en se déplaçant en haut et/ ou en bas on ajoute un symbole respectivement à gauche et/ ou à droite à l'autre de ces sous-séquences. Toutes les combinaisons d'extensions sont autorisées. Le voisinage de superposition d'une extension est défini dans chaque direction d'extension possible. Il s'agit de l'ensemble des valeurs de collisions associées à des sous-séquences de base qui se superposent avec la (ou les) sous-séquence(s) étendue(s) dans une direction donnée. Ces sous-séquences "de superposition des extensions" contiennent alors :

  (a) soit le symbole de l'extension si on s'intéresse à l'extension d'une seule des deux sous-séquences – cas d'un *voisinage rectangulaire*, horizontal ou vertical selon la sous-séquence considérée pour l'extension ;

  (b) soit l'un et l'autre des symboles des extensions si on cherche à étendre les deux sous-séquences – cas d'une *voisinage carré* dans le sens d'extension oblique.

Le processus d'extension des sous-séquences est itéré sur les sous-séquences étendues jusqu'à ce que plus aucune extension ne soit possible en terme des seuils de collisions et de distance. Finalement, à partir d'une valeur supérieure au seuil dans la matrice des collisions, correspondant à la comparaison de deux sous-séquences "de base", on identifie les plus longues sous-séquences étendues vérifiant les critères suivants à chaque extension :

1. Au moins une des valeurs de collisions dans le voisinage de superposition des extensions est supérieure au *seuil minimum de collisions* ;
2. La distance réelle entre les sous-séquences étendues est inférieure au *seuil maximum de distance* ;
3. On favorise par ailleurs à chaque itération la sélection de l'extension qui ajoute le plus de symboles aux sous-séquences considérées.

Ainsi, les paramètres clés de l'examen de la matrice de collisions sont les suivants :

- c_{min}, seuil minimum de collisions,
- d_{max}, seuil maximum de distance.

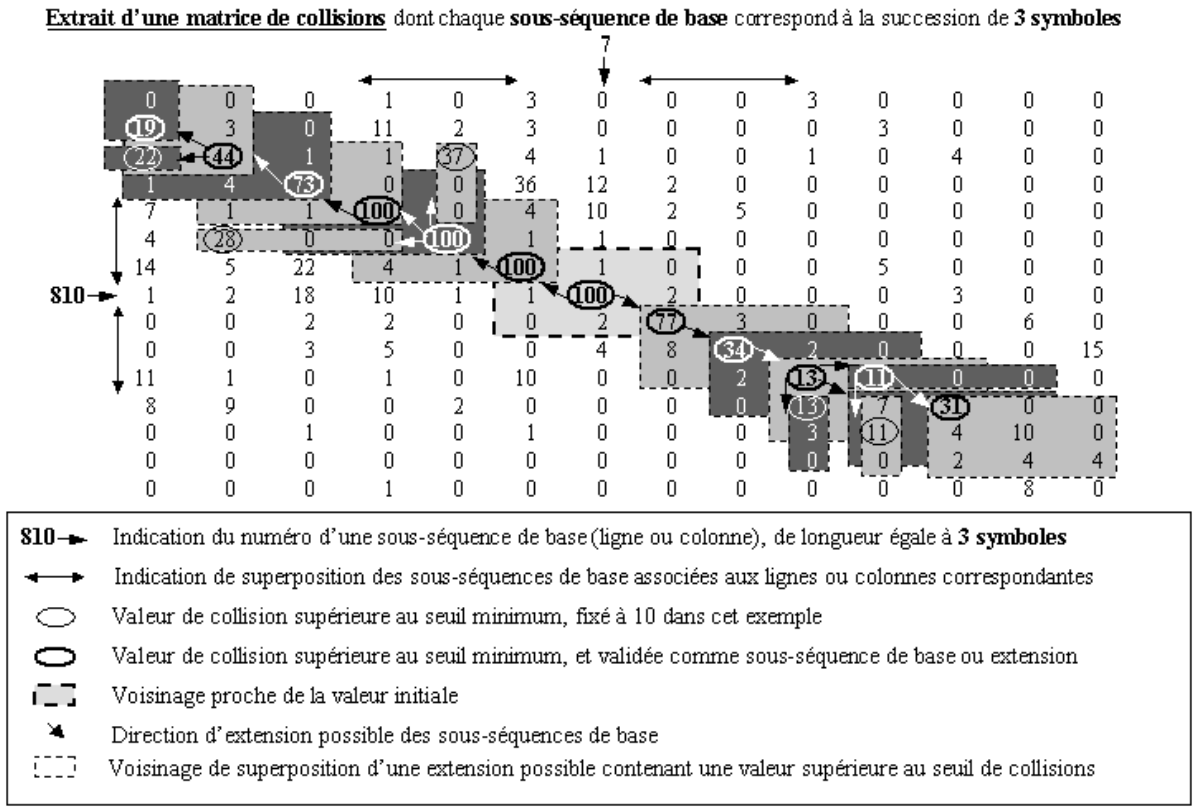


FIG. 5.5 – Principe d'examen de la matrice de collisions.
Recherche des sous-séquences correspondant aux instances complètes d'un motif à partir de l'identification d'une forte valeur de la matrice de collisions – 100 collisions / 100 projections aléatoires entre les sous-séquences numérotées 7 et 810 en référence à leur position dans l'ensemble des sous-séquences de base qui composent, dans l'ordre, la séquence initiale. La longueur des sous-séquences de base est fixée *a priori* à 3 symboles.

L'utilisation d'une distance réelle entre trajectoires nécessite par ailleurs implicitement la définition des deux paramètres liés à cette mesure. Ils fixent les écarts maximum sur les valeurs et dans le temps entre deux points similaires :

- ϵ_{LCSS}, écart maximum sur les valeurs,
- δ_{LCSS}, écart maximum dans le temps.

Illustration sur un exemple

La figure 5.5 illustre cette recherche de la sous-séquence la plus complète correspondant certainement à l'instance d'un motif à partir de la comparaison de deux sous-séquences de base identifiées par une valeur forte de la matrice de collisions. Dans cet exemple, chaque sous-séquence de base est composée de 3 symboles, 100 projections aléatoires de ces sous-séquences ont été réalisées, et le seuil minimum de collisions est fixé strictement à 10. Ainsi, les voisinages de collisions sont au plus de hauteur et/ ou de largeur égale à 3.

La première étape de recherche de l'instance supposée être la plus complète d'un motif consiste à identifier dans le voisinage proche de la valeur initiale examinée le couple de sous-séquences

associé à un nombre de collisions supérieur au seuil minimum, et dont la distance réelle est inférieure au seuil maximum. Dans l'exemple proposé, on sélectionne alors le couple initialement identifié, $(810, 7)$.

La seconde étape consiste à étendre de façon itérative ces sous-séquences de base jusqu'à ce que plus aucune extension ne soit possible. Pour exemple, la démarche de la première extension est la suivante :

1. **Définition des extensions possibles.** Les seuls voisinages de superposition des extensions contenant un nombre de collisions supérieur au seuil minimum de 10 strictement correspondent aux directions d'extension obliques, c'est-à-dire à l'ajout de la considération d'un symbole à droite et à gauche pour chaque sous-séquence de base.
2. **Vérification de la distance réelle.** La distance réelle entre les deux sous-séquences étendues est inférieure au seuil maximum, et on valide ainsi l'extension proposée des sous-séquences de base.
3. **Itération.** Étant donné le critère de distance validé, on itère le processus d'extension à partir cette fois des sous-séquences étendues couvrant pour l'une les sous-séquences de base 809 à 811, et pour l'autre celles numérotées de 6 à 8.

Dans le cas où plusieurs voisinages de superposition des extensions sont possibles, on favorise la plus grande extension. Par exemple à la troisième itération de l'extension, vers la gauche, on choisit l'extension oblique étant donné que le critère de distance est vérifié. Il n'y a cependant *a priori* aucune raison pour que la présence d'un nombre plus important de collisions dans un certain voisinage soit associé à une distance réelle plus faible des extensions correspondantes. Un contre-exemple est la quatrième extension vers la droite pour laquelle une distance réelle inférieure au seuil maximum n'est observée que dans le voisinage horizontal, pour lequel le nombre maximum de collisions est 11, alors qu'il est de 31 dans le voisinage oblique.

Finalement, dans cet exemple, la recherche des sous-séquences "complètes" est réalisée en six étapes d'extension pour identifier finalement comme similaires et potentiellement récurrentes les sous-séquences couvrant les sous-séquences de base 804 à 814 d'une part, et 1 à 12 d'autre part.

5.2.4 Synthèse des tentatives de motifs

L'examen de la matrice de collisions permet d'identifier un ensemble de sous-séquences récurrentes, ou *caractères*, présentant les caractéristiques suivantes :

(1) Fréquence. Chaque sous-séquence est similaire à au moins une autre selon les seuils minimum de collisions et maximum de distance.

(2) Signification. Les sous-séquences correspondent aux sous-séquences "de base" issues des projections, éventuellement étendues jusqu'aux plus longues sous-séquences significatives selon les critères de collisions et de distance. Elles ont ainsi différentes longueurs en terme du nombre de symboles qui les représente et/ ou de la durée à laquelle elles correspondent.

(3) Redondance. Certaines sous-séquences sont redondantes ou au moins se recouvrent partiellement.

L'une des caractéristiques de l'extraction de caractères étant de définir un ensemble de caractères non redondants [67], l'étape suivante consiste à les synthétiser en un ensemble de sous-séquences toutes disjointes – les *tentatives de motifs*.

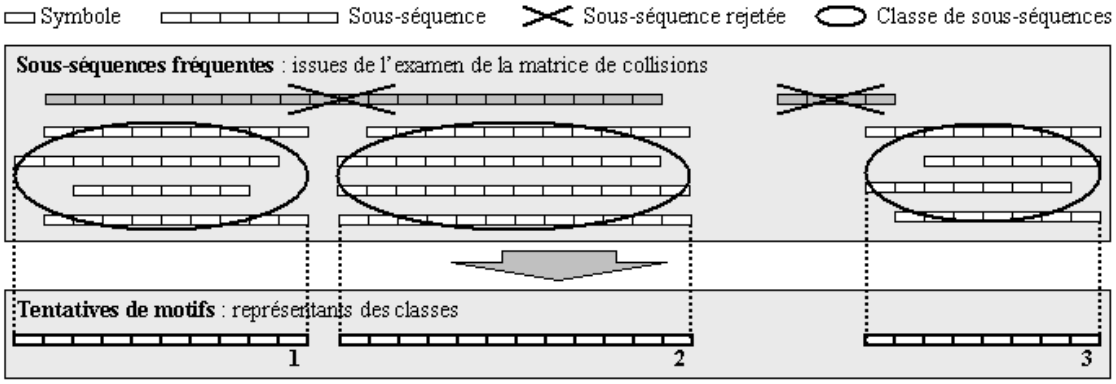


FIG. 5.6 – Illustration des objectifs de classification des sous-séquences identifiées à l'issue de l'examen de la matrice de collisions
Rejet des sous-séquences non significatives par rapport à l'ensemble des autres et constitution de classes de sous-séquences significatives, qui se recouvrent toutes mutuellement au moins partiellement.

Présentation du problème de synthèse des tentatives de motifs

La synthèse des tentatives de motifs consiste en la classification des sous-séquences fréquentes identifiées à l'issue de l'examen de la matrice de collisions, en fonction de leur pertinence à représenter une même instance d'un motif. On définit ensuite le représentant de chaque classe comme une *tentative de motif*, c'est-à-dire comme une sous-séquence correspondant potentiellement à une instance d'un des motifs présents dans la séquence initiale. Les sous-séquences d'une classe doivent toutes se superposer entre elles pour s'assurer qu'elles sont effectivement significatives de la présence d'une même instance d'un motif. Les représentants des différentes classes sont par contre définis de façon à ce qu'il n'y ait aucun recouvrement entre eux – les tentatives de motifs doivent être disjointes puisque les instances de motifs ne peuvent pas se superposer.

La synthèse des sous-séquences identifiées lors de l'examen de la matrice de collisions est ainsi une classification non supervisée de ces séquences. La figure 5.6 présente une illustration schématique des objectifs de la synthèse : construire des classes de sous-séquences qui se recouvrent toutes mutuellement au moins partiellement, en même temps qu'elles sont disjointes de tous les membres de chacune des autres classes. Compte tenu de la constitution relative possible des sous-séquences identifiées à ce niveau de l'analyse, cela impose de supprimer la considération de certaines sous-séquences non pertinentes, par exemple parce qu'elles se sont trouvées "par hasard" similaires à au moins une autre, ou parce qu'elles ont été trop étendues par rapport à la sous-séquence véritablement représentante d'un motif – cas de la proximité répétée mais pas systématique de plusieurs instances de motifs par exemple. Il est par contre également possible qu'une sous-séquence fréquente ne soit pas étendue suffisamment par rapport à l'instance de motif qu'elle représente en raison de bruit ou d'imprécision dans les valeurs. Globalement, dans un contexte particulièrement bruité, les sous-séquences seront souvent "sous-étendues" et rarement "sur-étendues". Par conséquent, on suppose dans ce contexte qu'une fois la classification réalisée les sous-séquences de chaque classe sont toutes vraiment représentatives d'une instance de motif, si bien que la plus longue des sous-séquences est probablement la plus significative. Ainsi, le représentant de la classe est estimé par la sous-séquence dont les indices de début et de fin dans la séquence initiale sont respectivement le plus faible et le plus élevé de ceux observés pour l'ensemble de ses membres. Considérons une classe "valide" de k sous-séquences, chacune se superposant avec toutes la autres de le classe. Si on note $t_{j,1}$ et t_{j,n_j}, où $1 \leq j \leq k$, $n_j > 1$ et

$t_{j,1} < t_{j,n_j}$, les indices de début et de fin dans la séquence initiale de chaque sous-séquence j du groupe, la *tentative de motif* représentant cette classe est alors définie par les indices de début et de fin, respectivement t^i et t^f, tels que :

$$t^i = min\left(\{t_{j,1}\}_{1\leq j\leq k}\right) \text{ et } t^f = max\left(\{t_{j,n_j}\}_{1\leq j\leq k}\right).$$

Résolution par une classification divisive

Pour répondre aux exigences de cette classification, on utilise une méthode divisive de classification non supervisée, qui est une méthode itérative généralement peu usitée. Les méthodes divisives partent de la considération d'une seule classe constituée de l'ensemble des éléments, et divisent progressivement la ou les classes jusqu'à un certain critère d'arrêt ; au contraire, les méthodes agglomératives, très communes, considèrent initialement chaque élément comme une classe constituée d'un seul membre.

L'intérêt d'une méthode divisive de classification dans notre contexte est qu'on est capable de définir facilement les critères de rejet d'une ou plusieurs sous-séquences d'un groupe alors que le contraire n'est pas si évident. Par ailleurs, si on considère un groupe de sous-séquences où chacune en recouvre partiellement au moins une autre, l'utilisation d'une méthode divisive pour leur classification nécessitera moins d'itérations qu'une méthode agglomérative, sachant que globalement peu de séquences doivent être exclues du groupe, et que le groupe lui même sera éventuellement divisé mais en un faible nombre de sous-groupes (voir l'exemple de la figure 5.6).

Trois critères fondamentaux définissent une méthode divisive de classification [27] : (1) le critère du choix de la classe à diviser, (2) le critère de division d'une classe et (3) le critère d'arrêt des divisions. Dans notre contexte de synthèse d'un ensemble de sous-séquences, ces critères sont définis selon les caractéristiques suivantes :

(1) Critère du choix de la classe à diviser. Le choix de la classe à diviser à chaque itération repose sur la présence ou non d'une sous-séquence qui ne se superpose pas avec toutes les sous-séquences de la classe à laquelle elle appartient alors.

(2) Critère de division d'une classe. Le choix de la méthode de division d'un groupe repose sur l'objectif de supprimer le moins de sous-séquences possible de l'ensemble initialement sélectionné, en partant du principe que chaque classe est globalement représentative d'une instance d'un motif. La "meilleure" division selon ce critère ne peut par conséquent être déterminée qu'*a posteriori* par comparaison de toutes les divisions possibles. Afin d'éviter un temps exponentiel d'exécution de toutes les divisions, on définit des critères permettant d'identifier *a priori* l'une des meilleures.

(3) Critère d'arrêt des divisions. Les divisions s'arrêtent lorsqu'on a constitué des classes disjointes de sous-séquences qui se superposent par ailleurs avec toutes les sous-séquences de la classe à laquelle elles appartiennent.

Algorithme de division

L'ensemble des sous-séquences sélectionnées peut être tout d'abord facilement divisé en un certain nombre de classes telles que chaque sous-séquence d'une classe (1) se superpose au moins avec une autre sous-séquence de sa classe et (2) ne se superpose à aucune sous-séquence des autres classes. L'objectif des divisions successives de chacune de ces classes afin de les rendre plus certainement significatives est alors de regrouper dans une classe uniquement des sous-séquences qui se superposent avec toutes les autres de la classe (et non avec au moins une comme c'est le cas initialement), tout en conservant le critère de ne se superposer avec aucune des autres classes.

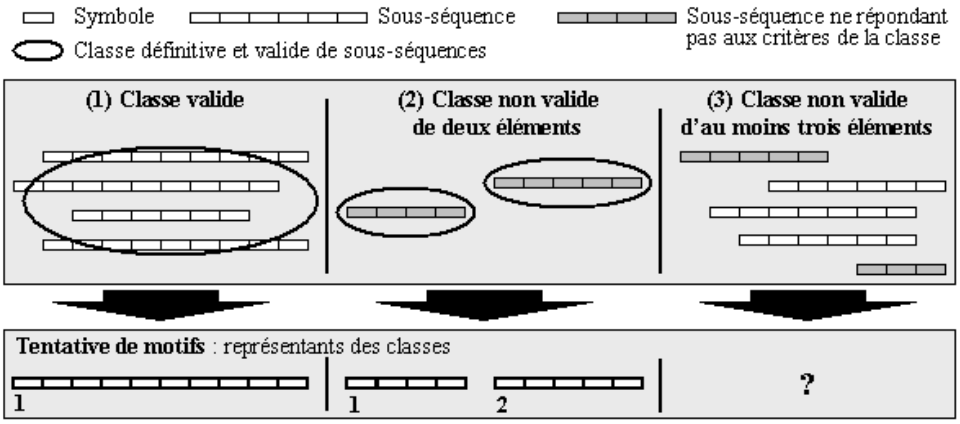


FIG. 5.7 – Illustration des cas possibles de constitution d'une classe.

À chaque itération de la division d'une classe, plusieurs configurations peuvent se présenter, tel qu'illustré sur la figure 5.7 :

(1) Chaque sous-séquence de la classe à diviser se superpose avec toutes les autres, si bien qu'aucune division supplémentaire de la classe n'est nécessaire ;

(2) La classe n'est constituée que de deux sous-séquences, et celles-ci ne se superposent pas : on forme alors deux classes contenant chacune l'une des deux sous-séquences ;

(3) La classe est constituée d'au moins trois sous-séquences dont au moins deux ne se superposent pas.

Dans ce dernier cas, la détermination du "bon" représentant de la classe n'est pas évidente. On fait l'hypothèse qu'au moins un des éléments de la classe n'est pas complètement significatif et doit être supprimé avant de réitérer la division. Notons k_1 et k_2 les sous-séquences qui ne se superposent pas. On définit k_1 ou k_2 en fonction de l'identification des sous-séquences qui se superposent avec le moins d'autres sous-séquences de la classe. La figure 5.8 illustre les quatre choix de suppression possibles pour définir une ou plusieurs nouvelles classes et réitérer l'algorithme de division :

(a) Supprimer k_1 et k_2 de la classe,

(b) Supprimer uniquement k_1,

(c) Supprimer uniquement k_2,

(d) Supprimer une ou plusieurs autres sous-séquences de la classe que k_1 ou k_2.

Le choix de l'une de ces solutions dépend du *critère de division d'une classe*, défini comme la conservation d'un nombre maximum de sous-séquences dans la classe. Dans les cas (a), (b) et (c), la nouvelle classe à considérer pour sa division éventuelle est simplement la classe initiale privée des éléments supprimés. Dans le cas (d), on commence par supprimer les sous-séquences qui se superposent à la fois avec k_1 et k_2 puisqu'on a fait le choix de les conserver tous les deux et qu'elles ne se superposent pas. On forme ensuite deux classes contenant pour l'une les éléments qui se superposent avec k_1, et pour l'autre ceux qui se superposent avec k_2. S'il reste ensuite des sous-séquences ne se superposant ni avec k_1, ni avec k_2, ni avec aucune des sous-séquences des deux classes précédemment constituées, on les regroupe alors en une troisième classe. Finalement on réitère l'algorithme de division sur ces deux ou trois nouvelles classes.

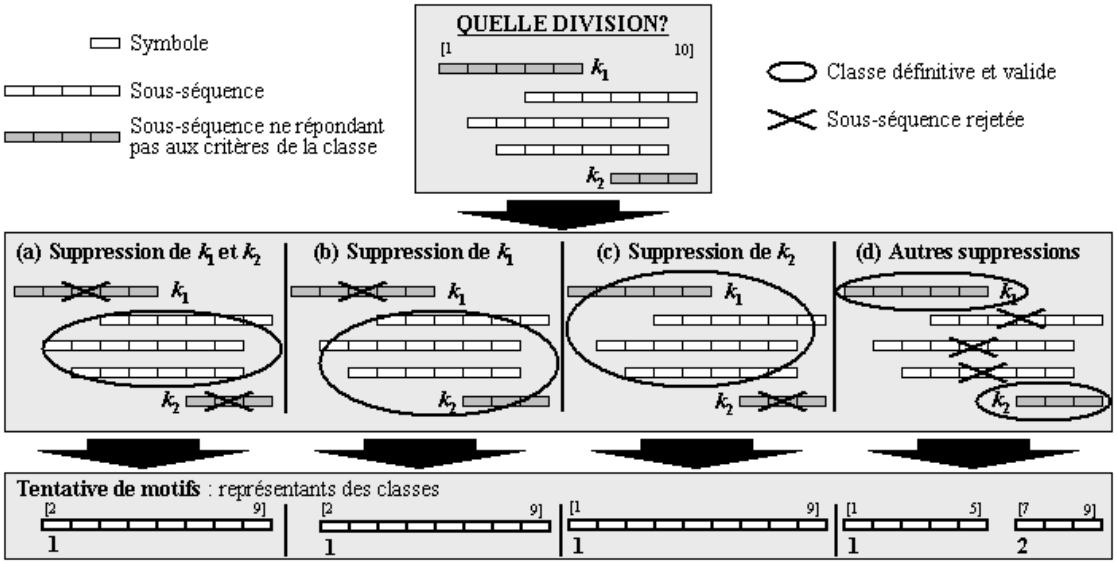


FIG. 5.8 – Illustration des cas possibles de division d'une classe non valide d'au moins trois éléments.
Une classe non valide signifie qu'au moins deux sous-séquences de la classe ne se superposent pas, notées k_1 et k_2

L'exemple de la figure 5.8 met en évidence qu'il n'est pas forcément nécessaire de tester tous les cas possibles de division d'une classe. Les cas (a), (b) et (c) peuvent en effet être définis comme "acceptables", d'autant plus qu'on ne cherche pas une délimitation précise des instances de motifs puisqu'on s'intéresse à une décision de "haut niveau". Afin d'améliorer les performances de la classification en terme du temps d'exécution, on définit alors d'une part un *ordre de priorité* pour le test des différents cas de division – (a), (b), (c) et enfin (d) – et d'autre part un critère de division moins exhaustif que "la suppression du moins de sous-séquences possibles", fonction d'un *seuil minimum* sur un taux "acceptable" du nombre de sous-séquences conservées par rapport au nombre total de sous-séquences initialement dans la classe. L'ordre de priorité est défini selon l'hypothèse qu'une sous-séquence est plus fréquemment "sous-étendue" par rapport à l'instance du motif qu'elle représente dans un contexte très bruité. Si par ailleurs le seuil n'est pas défini trop bas, la figure 5.9 illustre le bon fonctionnement de l'algorithme même dans le cas d'une sous-séquence "sur-étendue" : le cas (a) n'aboutit à la constitution d'une classe valide qu'en conservant une sous-séquence ; les cas (b) et (c) n'en conservent au mieux que quatre ; enfin le cas (d) en conserve six sur sept : c'est le "meilleur".

Finalement, la méthode de classification proposée permet d'identifier un ensemble de sous-séquences de la séquence initiale, toutes disjointes, et qui sont similaires à une ou plusieurs autres sous-séquences. Cet ensemble correspond aux *tentatives de motifs*. La méthode de synthèse proposée ne nécessite la définition d'aucun paramètre en particulier, mis à part le "taux acceptable" de sous-séquences conservées par rapport au nombre initial de sous-séquences mais qui a pour unique objectif l'exécution de l'algorithme dans un temps non exponentiel.

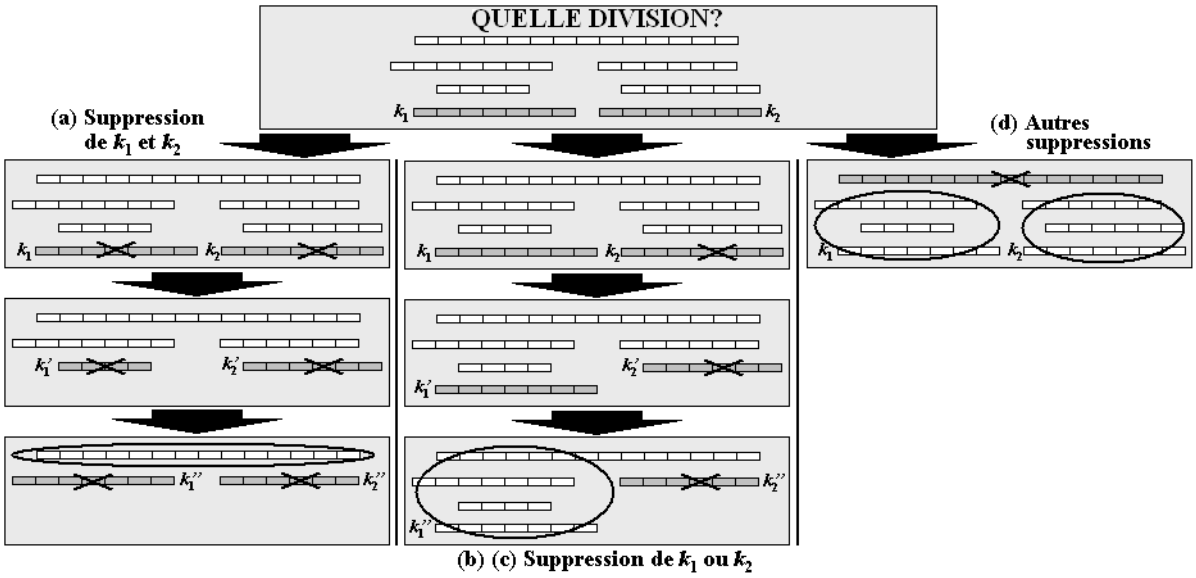


FIG. 5.9 – Illustration du cas d'une classe non valide dont la meilleure division est obtenue en conservant les deux sous-séquences k_1 et k_2 qui ne se superposent pas.

Illustration sur un exemple

Le schéma de la figure 5.10 présente un exemple des résultats de l'expérimentation de l'algorithme de classification sur une partie des sous-séquences issues de l'examen d'une matrice de collisions – sous-séquences de base 460 à 493, de longueur égale à 4 symboles. L'application de l'algorithme de division de l'ensemble des sous-séquences fréquentes identifiées résulte ne l'identification de trois tentatives de motifs, correspondant aux sous-séquences de base 461 à 466, 471 à 479, et 489 à 493.

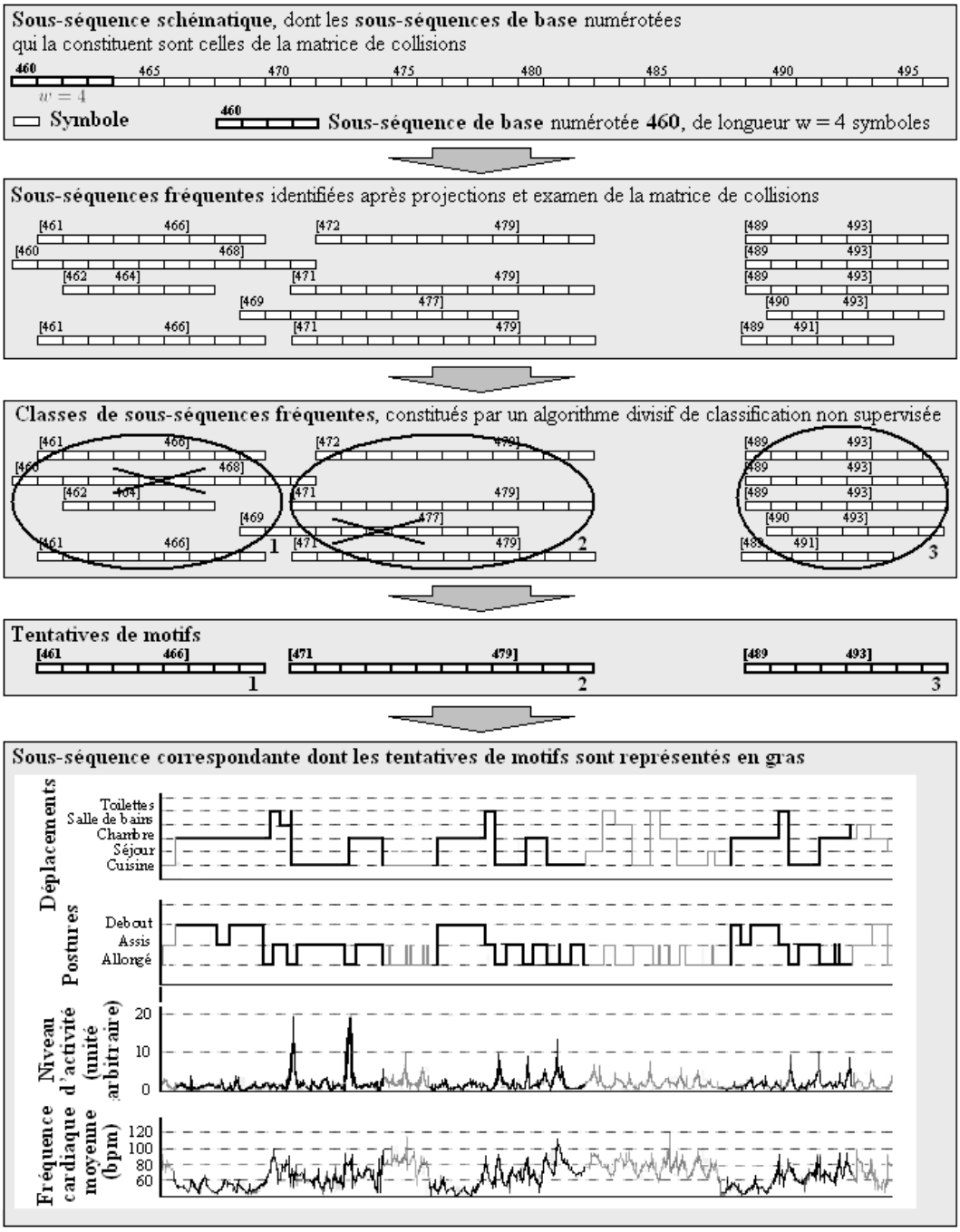


FIG. 5.10 – Synthèse des tentatives de motifs.

Illustration de la classification des sous-séquences identifiées à l'issue de l'examen de la matrice de collisions afin de constituer des classes significatives de sous-séquences dont chaque représentant est finalement une tentative de motif.

Les séquences correspondantes, issues du processus de simulation dans le cadre de la télésurveillance médicale à domicile, sont de dimension quatre incluant : (1) les déplacements, (2) les postures, (3) le niveau d'activité et (4) la fréquence cardiaque moyenne.

5.3 Classification pour l'identification des motifs

La dernière étape est alors l'identification des *motifs*, c'est-à-dire des classes significatives de *tentatives de motifs* en fonction de leurs similarités. Puisqu'on dispose d'une mesure de similarité pour comparer ces sous-séquences, une simple méthode de classification hiérarchique ascendante est utilisée pour l'identification des motifs, basée uniquement sur la table des distances réelles entre toutes les tentatives de motifs deux à deux. Cette méthode est une méthode agglomérative non supervisée, qui considère initialement autant de classes que d'éléments à classer pour les regrouper successivement selon un critère de similarité (seuil maximum de distance) : à chaque itération on regroupe les éléments "les plus similaires".

Classification hiérarchique ascendante
La mesure de similarité utilisée pour comparer deux à deux ces sous-séquences de longueurs différentes *a priori* est la *distance réelle* définie au paragraphe 4.2. Un regroupement est alors effectué entre deux classes d'éléments (pouvant ne contenir qu'un élément) si la distance entre ces deux éléments ou classes est inférieure au *seuil de distance* défini comme la distance maximum autorisée entre deux sous-séquences représentatives d'une même motif. Les regroupements sont effectués dans l'ordre croissant des distances observées, jusqu'au seuil de distance, le même que celui utilisé lors de l'examen de la matrice de collisions (voir paragraphe 5.2.3). Il a en effet la même signification – définir l'intervalle de similarité admissible entre deux sous-séquences – et la limitation du nombre de paramètres considérés est par ailleurs importante pour la robustesse du système. La distance entre une nouvelle classe ainsi constituée et une classe existante est définie comme la distance maximum observée entre les éléments de chacune des classes, comparés deux à deux. Cette définition permet de ne pas attribuer à une classe un élément si on n'est pas sûr que sa distance avec tous les éléments de la classe est inférieure au seuil de distance. On obtient ainsi un ensemble de classes dont les éléments sont, pour chaque classe, distants deux à deux d'au maximum le seuil de distance considéré.

La classification hiérarchique ascendante est ainsi réalisée à partir de la matrice carrée et symétrique des distances entre les classes constituées à chaque itération. Initialement, il y a autant de classes que de tentatives de motifs, et la matrice contient ainsi les distances entre toutes ces sous-séquences deux à deux. On itère le processus de classification tant que la plus petite distance contenue dans la matrice est inférieure au seuil de distance (d_{max}). À chaque itération, les deux éléments les plus proches – sous-séquences ou classes – sont regroupés en une seule classe dont on recalcule la distance par rapport à toutes les autres. Afin de s'assurer de n'avoir dans une classe que des sous-séquences séparées au plus d'une distance égale au seuil de distance, le principe de définition de la distance $d(c_{ij}, c_k)$ entre (1) la classe notée c_{ij} qui regroupe les deux classes notées c_i et c_j telles que $d(c_i, c_j) \leq d_{max}$, et (2) chacune des autres classes, notées c_k, est donc le suivant :

$$d(c_{ij}, c_k) = max(d(c_i, c_k), d(c_j, c_k)), \forall k \text{ tel que } 1 \leq k \leq n, k \neq i, k \neq j$$

où $1 \leq i, j \leq n$ et n est le nombre de classes à l'itération considérée.

Le schéma de la figure 5.11 illustre la classification hiérarchique ascendante de cinq tentatives de motifs notées (a) à (e) en deux classes de motifs {a,b,e} et {c,d}, selon le seuil de distance d_{max}. Les distances entre les nouvelles classes constituées à chaque itération sont mentionnées au niveau de chaque barre de regroupement sur le graphe de gauche, ainsi que dans les matrices de distances successives calculées, D_0 à D_3, présentées sur la droite de la figure.

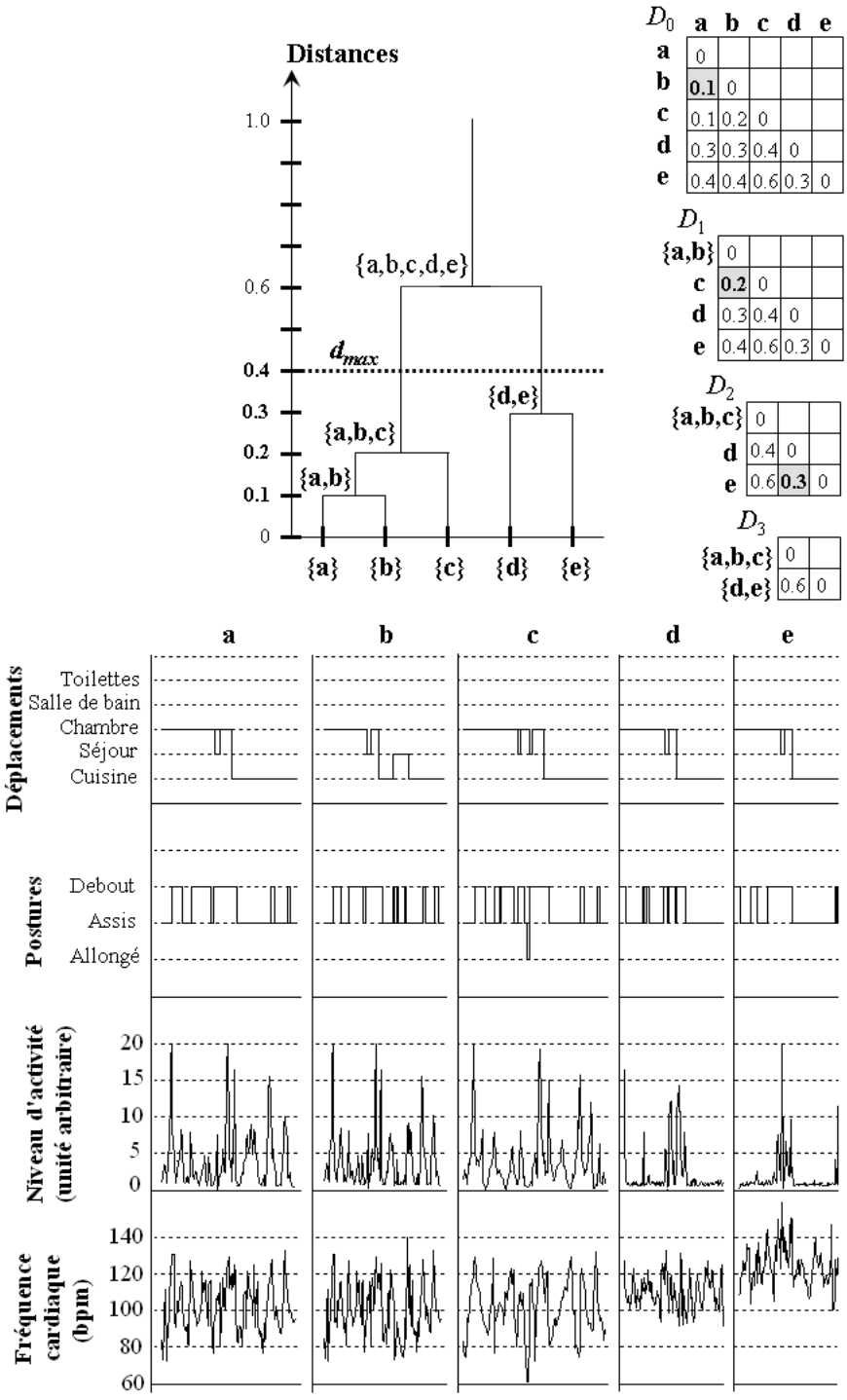


FIG. 5.11 – Classification hiérarchique ascendante.

On réalise la classification des tentatives de motifs notées (a) à (e) – représentées sur les graphes du bas de ka figure – selon la matrice initiale des distances D_0 et le seuil de distance $d_{max} = 0.4$. On forme alors deux classes regroupant {a,b,c} d'une part et {d,e} d'autre part.

Les sous-séquences concernées sont représentées en bas de la figure. Elles sont issues du processus de simulation dans le cadre de la télésurveillance médicale à domicile, et comportent quatre dimensions : (1) les déplacements de la personne télésurveillée, (2) les postures, (3) le niveau d'activité et (4) la fréquence cardiaque.

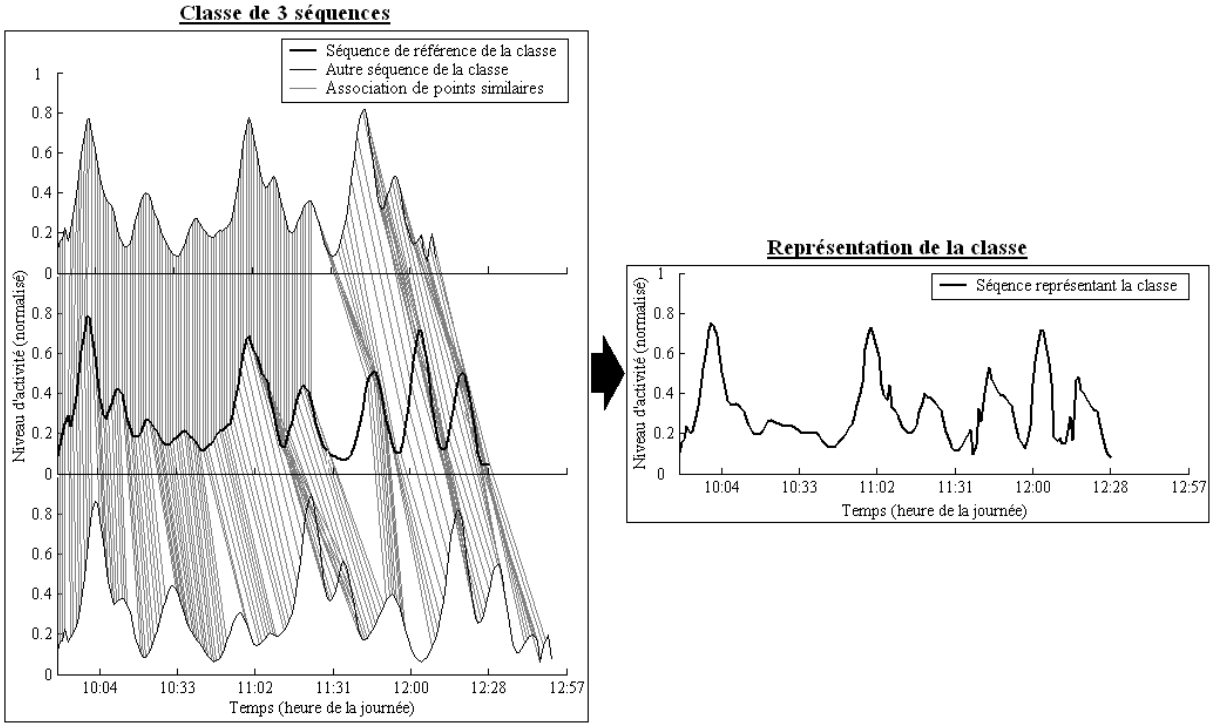


FIG. 5.12 – Calcul du représentant d'une classe.
Recherche du représentant d'une classe de trois séquences définies par un seul paramètre quantitatif dont les valeurs sont normalisées. À partir de chaque point de la séquence de référence – celle dont la longueur est la plus proche de la longueur moyenne des éléments de la classe – on calcule la valeur moyenne de ce point et de ceux des autres séquences auxquels il est associé.

Définition du représentant d'une classe

Le représentant de chaque classe est ensuite déterminé à partir de la sous-séquence de la classe dont la durée est la plus proche de la durée moyenne observée pour la classe, appelée *sous-séquence de référence*, et d'après les distances réelles observées avec les autres sous-séquences de la classe. Lors d'un calcul de distance basé sur la plus longue sous-séquence commune *LCSS*, des couples de points similaires entre les deux séquences comparées sont constitués selon les écarts maximum autorisés dans le temps et sur les valeurs, δ_{LCSS} et ϵ_{LCSS}. Le principe de calcul du représentant d'une classe est alors le suivant :

- La sous-séquence représentant la classe est de même longueur que la sous-séquence de référence, la plus proche en longueur de la moyenne des longueurs observées pour la classe ;
- Pour chaque point de cette sous-séquence de référence, et pour chaque composante, on calcule la moyenne des valeurs de l'ensemble des points considérés comme similaires à celui-ci lors des calculs de distance réelle entre la sous-séquence de référence et chacune des autres de la classe.

Le schéma de la figure 5.12 illustre le calcul du représentant moyen d'une classe dans le cas d'un effectif de trois sous-séquences. On note que cette représentation est un peu "hachée" mais met bien en évidence les trois pics de valeurs caractéristiques présents dans chacune des trois séquences de la classe. Par ailleurs, étant donné le "haut niveau" d'analyse et de décision, la représentation d'une classe est forcément imprécise et plutôt grossière. Cependant, dans le cas

où le représentant d'une classe est utilisé par exemple pour des tâches supervisées de recherche de caractères et de classification, il conviendrait sûrement d'assurer une meilleure continuité des valeurs successives qui le constituent. Dans le cadre de la télésurveillance médicale à domicile, une classification *supervisée* est par exemple nécessaire au processus de décision sur une situation habituelle ou non d'une personne : les caractères – ou sous-séquences récurrentes – constituant son profil comportemental, issus de l'apprentissage, sont recherchés dans les données collectées en temps réel pour vérifier leur adéquation au profil.

Ainsi, les paramètres clés de la classification des tentatives de motifs pour l'identification des motifs sont ceux déjà définis aux étapes précédentes de l'analyse :

- d_{max}, seuil maximum de distance,
- ϵ_{LCSS}, écart maximum dans les valeurs,
- δ_{LCSS}, écart maximum dans le temps.

5.4 Synthèse

Le schéma de la figure 5.13 résume les étapes principales de l'identification et de la classification des motifs détaillées dans ce chapitre :

(1) l'abstraction des données brutes issues des capteurs,

(2) la fouille de caractères, c'est-à-dire l'extraction non supervisée des sous-séquences récurrentes – les *tentatives de motifs*,

(3) la classification non supervisée de ces sous-séquences en classes de sous-séquences similaires – les *motifs*.

Pour chaque étape, on définit également l'ensemble des paramètres impliqués, incluant si nécessaires les paramètres liés au calcul de la distance réelle entre deux sous-séquences. On constate qu'il existe un grand nombre de paramètres à définir pour mettre en oeuvre l'approche proposée pour l'extraction de motifs. Une partie d'entre eux peut néanmoins être déterminée assez simplement dans un contexte donné. L'objectif du chapitre suivant est alors d'expérimenter et de valider l'approche proposée pour l'extraction de motifs. On présente en particulier la démarche de sélection de valeurs appropriées pour les différents paramètres.

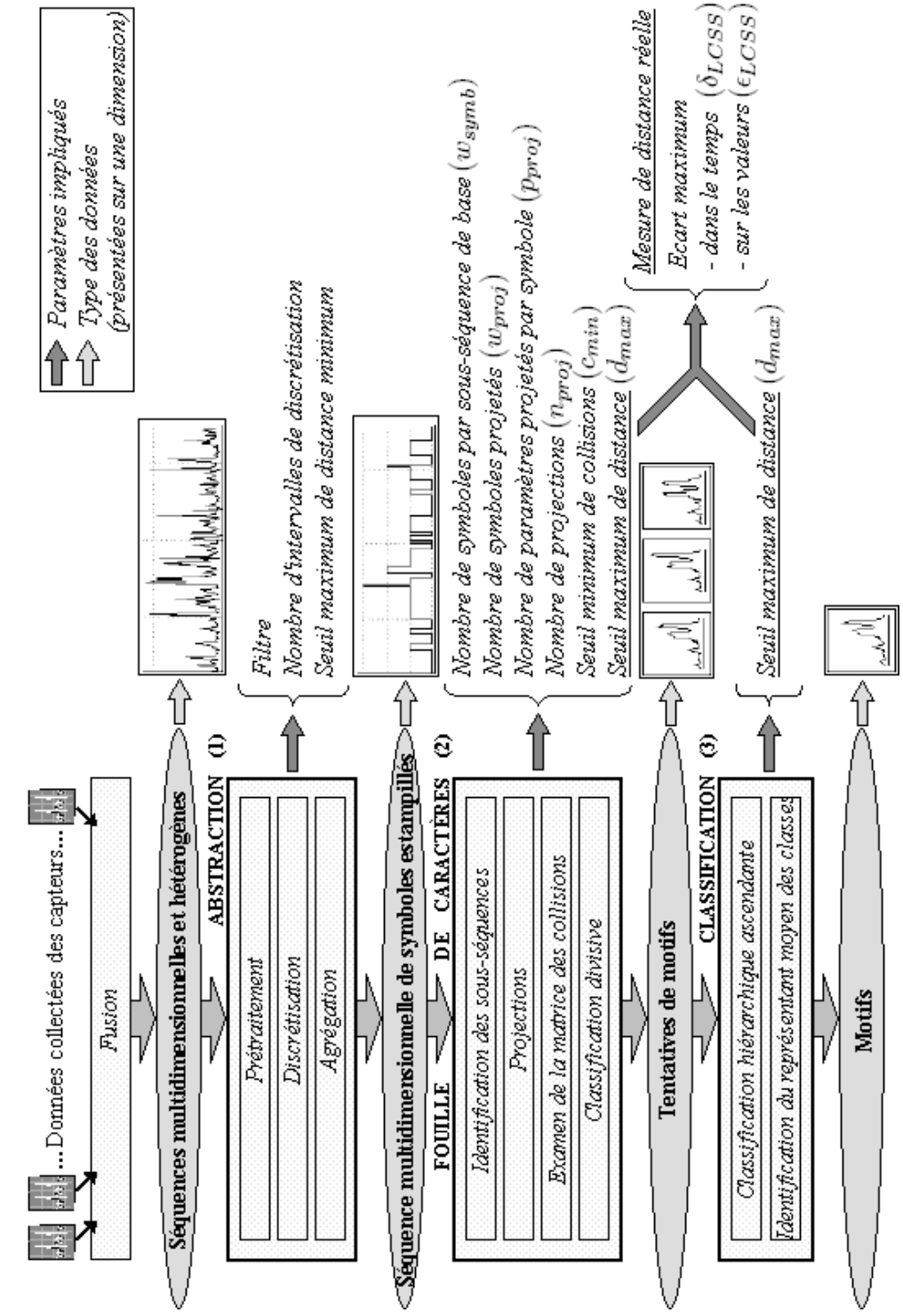


FIG. 5.13 – Synthèse des étapes successives de l'extraction de motifs à partir de données enregistrées par un ensemble de capteurs.

6

Expérimentation et Validation

L'approche proposée pour l'extraction non supervisée de motifs multidimensionnels et hétérogènes est expérimentée dans le cadre de la télésurveillance médicale à domicile. Des détails sur l'implémentation du processus sous `MATLAB` sont présentés en annexe G. Dans ce chapitre nous définissons d'abord en 6.1 le processus expérimental approprié à l'évaluation du système d'extraction de motifs. Puis, la qualité de la méthode et des résultats sont présentés et discutés dans les paragraphes 6.2 et 6.3.

6.1 Processus expérimental

La mise en place d'un processus expérimental nécessite la définition :
- 6.1.1. du contexte expérimental,
- 6.1.2. des ensembles de données nécessaires à l'expérimentation,
- 6.1.3. de la méthodologie d'expérimentation et
- 6.1.4. des mesures de performances nécessaires à l'évaluation du système.

6.1.1 Contexte expérimental : la télésurveillance médicale

L'expérimentation de l'approche proposée pour l'extraction de motifs est réalisée dans le cadre de la télésurveillance médicale à domicile. Dans ce contexte, le système a pour objectif l'apprentissage des habitudes de vie d'une personne. Ainsi :
- **les tentatives de motifs** doivent représenter les comportements récurrents d'une personne dans sa vie quotidienne ;
- **les motifs** sont alors les classes de ces comportements réguliers, représentatives des activités "types" réalisées habituellement dans la vie quotidienne.

Les caractéristiques intrinsèques de ces activités, associées à leurs contraintes de réalisation (ordre de réalisation par exemple), constituent les **habitudes de vie** d'une personne.

Étant donné qu'on ne dispose pas d'enregistrements réels suffisamment larges, complets et représentatifs, l'expérimentation du système d'apprentissage est réalisée à partir des données issues de la simulation (voir Partie II). Les séquences considérées sont ainsi hétérogènes, et à quatre dimensions correspondant aux paramètres suivants : (1) Déplacements, (2) Postures, (3) Niveau d'activité, et (4) Fréquence cardiaque. L'avantage de considérer des données issues de la simulation est la possibilité de générer des séquences représentatives de plusieurs types de situations, incluant plusieurs types de modifications du comportement d'une personne (voir paragraphe II.6.5) :

- **Modifications "normales"** dans la réalisation répétée d'une même activité : interruptions, déformations temporelles, variabilité dans les valeurs ;
- **Modifications inquiétantes** correspondant à une rupture dans les principes de variation des différents paramètres considérés dans la construction des sous-modèles de simulation.

La mise en place d'un processus expérimental pour évaluer les performances d'un système de décision dans un contexte donné nécessite ensuite de bien définir les éléments suivants :

- **6.1.2 Quels ensembles de données** sont appropriés au contexte expérimental et à l'évaluation des performances du système de décision ?
- **6.1.3 Quelle méthode d'expérimentation** permet une évaluation complète du système ?
- **6.1.4 Quelles mesures de performance** sont les plus appropriées pour une évaluation objective de la robustesse et de l'efficacité du système ?

6.1.2 Données expérimentales

Une validation objective des performances d'un système d'extraction de motifs nécessite d'avoir connaissance *a priori* des motifs présents dans les séquences analysées et de la localisation de leurs instances. Les séquences issues de la simulation portent, d'après leur construction, un certain rythme dans les activités qu'elles représentent implicitement. En particulier, la succession des déplacements est caractéristique des habitudes de vie, et les valeurs successives des autres paramètres en découlent. Cependant, les activités habituelles et le moment de leur réalisation ne sont pas connus explicitement d'après le modèle de simulation.

L'objectif d'une première expérimentation est alors d'analyser des séquences de données pour lesquelles on connaît précisément les motifs et les instances de ces motifs qu'elles contiennent. Le principe de construction de ces séquences expérimentales consiste (1) à générer des séquences ne contenant sûrement aucun motif, à partir de séquences de déplacements aléatoires, puis (2) à définir un ensemble de motifs dont on insère pour chacun plusieurs instances dans les séquences précédemment générées.

Génération d'une séquence de "non motifs"

Compte tenu de la construction du processus de simulation, on génère des séquences dont on sait qu'elles ne contiennent *a priori* aucune sous-séquence récurrente à partir de la génération de déplacements aléatoires. Les déplacements successifs ne correspondent ainsi à aucune habitude de vie. Le processus de simulation permet ensuite de générer à partir de ces déplacements des valeurs cohérentes pour les autres paramètres, si bien que les séquences générées sont malgré tout réalistes en terme de leur enregistrement possible à un moment donné dans un contexte expérimental.

Introduction d'instances de motifs

L'introduction d'instances de motifs dans une séquence de "non motifs" est réalisée en plusieurs étapes successives.

1. **Sélection des motifs.**

   Les motifs sont par hypothèse représentatifs de comportements habituels au domicile. Leur sélection est ainsi réalisée aléatoirement dans une séquence générée par le processus de simulation, correspondant à une journée dans des conditions habituelles de vie d'une personne. Les sous-séquences sélectionnées peuvent ainsi probablement être interprétées en terme de la réalisation d'une certaine activité à un moment donné.

Quelques contraintes sont par ailleurs prises en compte pour la sélection des sous-séquences définissant les motifs :

(a) *La longueur des sous-séquences* est déterminée aléatoirement dans un intervalle de limites significatives – par exemple, on impose que chacune corresponde à une durée comprise entre 30 minutes et 2 heures ;

(b) *Chaque sous-séquence doit être significative* en terme de la réalisation d'une activité et non pas seulement d'une ou deux tâches élémentaires. Les sous-séquences extraites doivent par conséquent être représentées par plusieurs symboles si on leur applique le processus d'abstraction.

Une estimation du nombre de symboles significatif est réalisée à partir des séquences générées de la simulation. Il s'agit d'identifier intuitivement des sous-séquences représentatives d'activités de base de la vie quotidienne. L'abstraction limitée à ces sous-séquences donne ensuite une idée du nombre minimum de symboles qui caractérise ces activités. On considère finalement que l'instance significative d'un motif est composée d'au moins **4 symboles**. La description d'une activité en comporte en général beaucoup plus, mais il est néanmoins possible qu'elle en comporte si peu, comme dans le cas particulier des périodes de sommeil.

2. **Instanciation de ces motifs.**

L'étape suivante consiste à créer des instances représentatives de ces motifs. L'instanciation doit en particulier prendre en compte les modifications "normales" possibles dans la réalisation d'une même activité. Différents types de bruits sont par conséquent introduits dans chaque instance de motif : interruptions, déformations temporelles et variabilité dans les valeurs (voir paragraphe II.6.5.2).

L'introduction de ces bruits est caractérisée par un ensemble de paramètres : durée et fréquence des interruptions, taux de déformation dans le temps et de variabilité dans les valeurs. Le réglage de ces paramètres permet d'expérimenter l'identification des motifs dans différentes conditions et d'évaluer ainsi la qualité des résultats obtenus tant que les valeurs de ces paramètres restent représentatives d'une modification "normale" de comportement.

3. **Insertion de ces instances.**

La dernière étape consiste à introduire les instances de chaque motif dans la séquence de "non motifs" générée. On tient en particulier compte de la translation possible dans le temps de la réalisation d'une même activité. L'instant d'insertion d'une instance de motif (t_{inst}) est ainsi déterminé aléatoirement selon une distribution normale (fonction $randn$). La moyenne de cette distribution correspond à l'instant d'occurrence de l'instance initiale du motif lors de sa sélection (t_0), et l'écart-type correspond à un paramètre de décalage typique autorisé autour de cette valeur (t_{std}).

$$t_{inst} = max(0, t_0 + t_{std} \times randn)$$

On s'assure également de ne pas superposer les instances des motifs, suivant l'hypothèse que les activités "typiques" d'une personne sont toujours réalisées les unes après les autres, et peuvent simplement être momentanément interrompues par un tâche secondaire. Une personne peut bien sûr réaliser plusieurs activités en même temps – par exemple si elle a l'habitude de préparer son sac de sport le matin en même temps qu'elle prend son petit déjeuner – mais l'activité "type" considérée alors sera la réalisation simultanée de ces deux tâches.

6.1.3 Méthode d'expérimentation

Le processus expérimental doit être approprié à l'évaluation des performances du système de décision à deux niveaux :

- **Qualité de la méthode**
  L'objectif est à ce niveau d'évaluer la pertinence de la méthode proposée et de préciser les critères de choix pour chaque paramètre des valeurs les plus adaptées au contexte considéré. L'analyse de la qualité de la méthode inclut ainsi l'évaluation de l'efficacité des étapes critiques impliquées dans le système de décision :
  1. **la mesure de similarité** utilisée,
  2. **l'abstraction** des séquences initiales,
  3. **la fouille de données** pour l'extraction de motifs.
- **Qualité des résultats**
  Une fois évaluée la qualité de la méthode, l'objectif suivant est d'évaluer la capacité du système à bien localiser et classer les instances de motifs prédéfinis insérés dans des séquences de "non motifs", particulièrement dans un contexte bruité. On prend en compte les différents types de bruits possibles entre les instances d'un même motif : interruptions, déformations temporelles et variabilité dans les valeurs.
  Cette démarche comprend plusieurs étapes :
  1. **Évaluer l'efficacité du système dans un contexte faiblement bruité.** On étudie en particulier l'influence du choix des paramètres du système pour atteindre les objectifs de décision.
  2. **Test de sensibilité : évaluer les performances dans un contexte bruité.** Le système doit reconnaître les instances bruitées d'un motif tant qu'elles correspondent à des modifications "normales" de comportement.
  3. **Test de spécificité : évaluer les performances dans un contexte critique.** Le système ne doit plus reconnaître comme instances d'un motif les représentants de la réalisation anormalement modifiée de l'activité correspondante.

Enfin, une fois le système configuré dans notre contexte expérimental, on propose d'évaluer la **qualité relative des étapes de simulation et de décision**. Il s'agit d'appliquer le système d'extraction des motifs sur les séquences produites par le processus de simulation dans des conditions habituelles de vie d'une personne. L'objectif est de vérifier si les données générées par le simulateur contiennent effectivement des régularités au regard du processus d'apprentissage. Des résultats non significatifs remettent alors en cause soit la validité du processus de simulation, soit la pertinence de la méthode d'extraction de motifs proposée dans le contexte expérimental considéré.

6.1.4 Mesures de performance

La mise en place d'un processus expérimental nécessite de disposer de critères appropriés à l'évaluation objective des performances du système, en comparaison aux résultats attendus. Dans notre contexte, il s'agit ainsi d'expérimenter le système d'extraction de motifs sur des séquences de données pour lesquelles on connaît *a priori* d'une part les motifs présents, et d'autre part la localisation de leurs instances.
Les performances du système sont évaluées à deux niveaux :

- **Identification des tentatives de motifs.** Évaluation de la bonne localisation et délimitation des sous-séquences récurrentes en fonction de la connaissance *a priori* des instances de motifs introduites dans la séquence initiale.
- **Classification en motifs.** Évaluation de la bonne classification des sous-séquences récurrentes selon les classes de motifs connues *a priori*.

On définit alors des méthodes d'évaluation de la *sensibilité* – *"bonnes détections"* – et de la *spécificité* – *"fausses alarmes"* – des résultats obtenus à ces deux niveaux de l'extraction de motifs.

Identification des tentatives de motifs

Une première étape est d'évaluer l'efficacité de la méthode dans la résolution des objectifs suivants :

(a) Sensibilité : Identifier comme *tentatives de motifs* des sous-séquences qui correspondent effectivement à des sous-séquences récurrentes dans la séquence initiale ;

(b) Spécificité : Ne pas identifier comme *tentatives de motifs* des sous-séquences issues d'intervalles de "non motifs".

Ces critères de sensibilité et de spécificité sont des réels compris entre 0 et 1, la valeur 1 correspondant à un indice "parfait". Ils sont calculés à partir des taux de vrais/ faux positifs/ négatifs – notés VP, FP, VN, FN – en considération des hypothèses suivantes pour chaque point de la séquence temporelle considérée :

- H_0 = "*le point ne fait pas partie de l'instance d'un motif*"
- H_1 = "*le point fait partie de l'instance d'un motif*"

On estime alors les taux suivants :

- VP = "*points considérés à raison comme appartenant à l'instance d'un motif*" ;
- VN = "*points considérés à raison comme n'appartenant pas à l'instance d'un motif*" ;
- FP = "*points considérés à tort comme appartenant à l'instance d'un motif*" ;
- FN = "*points considérés à tort comme n'appartenant pas à l'instance d'un motif*".

Les indices de sensibilité (Se) et de spécificité (Sp) sont alors estimés selon les équations suivantes :

$$Se = \frac{VP}{VP + FN} \text{ et } Sp = \frac{VN}{VN + FP}.$$

Étant donné qu'ils comparent les sous-séquences récurrentes identifiées et les instances de motifs insérés point à point le long de la séquence initiale, ces indices de performance sont très précis et sensibles. Le contexte d'extraction "haut niveau" des sous-séquences récurrentes n'exige ainsi pas forcément de parfaites valeurs pour ces indices.

Classification en motifs

L'objectif de la mesure de performance relative à la classification des tentatives de motifs est d'évaluer la capacité de l'algorithme à catégoriser correctement les sous-séquences récurrentes identifiées, en fonction des motifs insérés *a priori* dans la séquence initiale. Les critères de sensibilité et de spécificité peuvent alors être définis comme suit :

(a) Sensibilité : Regrouper toutes les sous-séquences représentatives d'un même motif dans une seule classe ;

(b) Spécificité : Ne regrouper dans un classe donnée que des sous-séquences représentatives d'un seul et même motif.

Dans cet objectif, on exploite le contenu de la *matrice de confusion* – dite parfois aussi *matrice de contingences* – qui met en correspondance :
- les catégorisations effectuées par le système, en n classes, et
- les connaissances *a priori* sur les classes d'appartenance des différentes sous-séquences récurrentes, parmi m motifs.

La matrice de confusion est alors de dimension $m \times n$, et notée $C = \{c_{ij}\}_{1 \leq i \leq m, 1 \leq j \leq n}$. Chaque élément c_{ij} représente le nombre d'éléments de la classe j qui correspondent effectivement à une instance du motif i. Par conséquent, la somme des éléments sur chaque ligne i doit correspondre au nombre d'instances du motif i, et la somme sur chaque colonne j au nombre d'éléments de la classe j.

Définitions spécifiques liées au contexte

Dès qu'un système de classification est supervisé, ou que le nombre de classe est défini *a priori*, l'évaluation des performances est réalisée à partir d'une matrice de confusion carrée. Ce contexte particulier est ainsi largement abordé dans la littérature. Cependant, un système de classification non supervisé ne 'découvre" pas toujours le même nombre de classes que celui initialement présent dans les données, et il existe ainsi des cas peu abordés où $m \neq n$.

Par ailleurs, notre contexte présente des spécificités qui compliquent encore le problème d'évaluation des performances de la classification. Les éléments à classer – les *tentatives de motifs* – ne sont en effet pas connus directement, mais identifiés dans une séquence initiale par un système d'apprentissage non supervisé. Par conséquent, on peut observer les particularités suivantes :
- **Des instances de motifs ne sont pas identifiées,** donc pas classées – les *faux négatifs* de l'évaluation précédente.
  La somme sur chaque ligne i de la matrice de confusion n'est ainsi pas forcément égale au nombre d'instances du motif i.
- **Des sous-séquences ne correspondant pas à une instance d'un motif sont identifiées,** donc classées – les *faux positifs* de l'évaluation précédente.
  La somme sur chaque colonne j n'est ainsi pas forcément égale à l'effectif de la classe j.
- **Une instance d'un motif peut être fractionnée,** c'est-à-dire reconnue comme plusieurs sous-séquences appartenant à une ou plusieurs classes.
  La somme sur chaque ligne i de la matrice de confusion peut ainsi être supérieure au nombre d'instances du motif i.

On propose alors de nouvelles définitions des indices de *sensibilité* et *spécificité* appropriées à ce contexte particulier.

(a) Sensibilité. *"Toutes les instances d'un même motif doivent être regroupées dans une seule classe, sans fractionnement – c'est-à-dire qu'une instance d'un motif ne doit être associée qu'à une seule sous-séquence d'une seule classe."*

(b) Spécificité. *"Tous les éléments d'une classe doivent être représentatifs d'un seul et même motif, sans fractionnement – c'est-à-dire qu'il n'y a qu'une seule sous-séquence de la classe associée à chaque instance du motif."*

Implémentation par la notion d'entropie

D'après les définitions précédentes, manquer la reconnaissance de certaines instances de motifs, les identifier comme plusieurs tentatives de motifs, ou ne pas regrouper correctement les tentatives

de motifs en motifs, sont autant de critères de *désordre* qui font diminuer l'indice de performance calculé à l'issue de la classification. Ces définitions de sensibilité et de spécificité sont ainsi implémentées à l'aide de la notion d'*entropie*, qui traduit bien une mesure de désordre issue de la thermodynamique. La fonction d'entropie est positive ou nulle, où une valeur nulle représente l'*"ordre parfait"* – c'est-à-dire, dans notre contexte, une parfaite cohérence des résultats de la classification, qui doit donc se traduire par des valeurs de sensibilité et spécificité maximum, égales à 1. Par ailleurs, le maximum de l'entropie correspond au "désordre maximum", c'est-à-dire dans notre contexte à l'équiprobabilité de la présence d'une instance d'un motif dans chacune des classes par exemple. On rapporte ainsi l'entropie aux notions de sensibilité (Se) et de spécificité (Sp) par la formule suivante :

$$Se \text{ ou } Sp = 1 - \frac{entropie}{max(entropie)}.$$

Le lien entre les indices de sensibilité et spécificité et le contenu de la matrice de confusion est alors le suivant :

- **Sensibilité (Se).** La sensibilité de la classification est liée à la bonne identification et classification des instances de chaque motif i dans une seule classe j. Elle correspond ainsi à une mesure d'entropie sur chaque ligne i de la matrice de confusion par rapport à l'ensemble des colonnes représentatives des classes constituées par le système ($1 \leq j \leq n$).

  Par exemple, si les instances d'un motif i se retrouvent en même proportion dans chacune des classes, l'entropie est maximum et la sensibilité est donc nulle. À l'inverse, si les instances d'un motif i sont toutes associées à une seule même classe j, l'entropie correspondante est nulle et la sensibilité maximum.

- **Spécificité (Sp).** La spécificité de la classification est liée à la composition homogène de chaque classe j constituée par le système, comme représentative des instances d'un seul motif i. Elle correspond ainsi à une mesure d'entropie sur chaque colonne j de la matrice de confusion par rapport à l'ensemble des lignes représentatives des motifs initialement présents dans les séquences ($1 \leq i \leq m$).

  Par exemple, si les sous-séquences d'une classe j sont associées dans une même proportion à chaque motif i, l'entropie est maximum et la spécificité est donc nulle. À l'inverse, si les éléments d'une classe j sont tous associés à un seul même motif i, l'entropie correspondante est nulle et la spécificité maximum.

Ces indices Se_i et Sp_j associés à chaque ligne i, $1 \leq i \leq m$, et chaque colonne j, $1 \leq j \leq n$, de la matrice de confusion $C = \{c_{ij}\}$ peuvent alors être définis comme suit :

$$Se_i = 1 + \frac{1}{log(n)} \cdot \sum_{j=1}^{n} \frac{c_{ij}}{m_i} \cdot log\left(\frac{c_{ij}}{m_i}\right) \tag{6.1}$$

$$Sp_j = 1 + \frac{1}{log(m)} \cdot \sum_{i=1}^{m} \frac{c_{ij}}{n_j} \cdot log\left(\frac{c_{ij}}{n_j}\right) \tag{6.2}$$

où m est le nombre de motifs initialement insérés,
n nombre de classes identifiées par le système,
m_i nombre d'instances du motif i,
n_j nombre d'éléments de la classe j.

Les valeurs $log(n)$ et $log(m)$ représentent les entropies maximum associées respectivement aux lignes (n classes possibles) et aux colonnes (m motifs possibles). L'entropie maximum d'un système à N états, $max(entropie(N))$, correspond en effet à l'équiprobabilité des états, $1/N$, si bien que :

$$max(entropie(N)) = -\sum_{i=1}^{N} \frac{1}{N} \cdot log\frac{1}{N} = log(N).$$

Pour appliquer les équations (6.1) et (6.2), les propriétés suivantes doivent cependant être vérifiées :

$$\sum_{j=1}^{n} \frac{c_{ij}}{m_i} = 1 \text{ et } \sum_{i=1}^{m} \frac{c_{ij}}{n_j} = 1$$

Dans notre contexte cependant, on a remarqué qu'il est possible de ne pas reconnaître des instances de motifs, ou de les reconnaître comme fractionnées en plusieurs sous-séquences associées alors à une ou plusieurs classes. Par conséquent, les sommes $\sum_{j=1}^{n} c_{ij}/m_i$ et $\sum_{j=1}^{n} c_{ij}/n_j$ peuvent être soit *inférieures* – quand on ne reconnaît pas certaines instances – soit *supérieures* – quand la reconnaissance fractionne certaines instances – à 1. Pour assurer des sommes égales à 1, le plus simple est alors de considérer, au lieu de (6.1) et (6.2), les équations suivantes (6.3) et (6.4) :

$$Se_i = 1 + \frac{1}{log(n)} \sum_{j=1}^{n} \frac{c_{ij}}{\sum_{j=1}^{n} c_{ij}} \cdot log\left(\frac{c_{ij}}{\sum_{j=1}^{n} c_{ij}}\right), \tag{6.3}$$

$$Sp_j = 1 + \frac{1}{log(m)} \sum_{i=1}^{m} \frac{c_{ij}}{\sum_{i=1}^{m} c_{ij}} \cdot log\left(\frac{c_{ij}}{\sum_{i=1}^{m} c_{ij}}\right). \tag{6.4}$$

Il faut cependant alors adapter ces équations pour prendre en compte les éventualités (1) d'une instance non reconnue d'une part, et (2) de la reconnaissance fractionnée d'une instance d'autre part.

(1) Instance non reconnue.

Afin de prendre en compte l'éventualité d'une instance non reconnue, on introduit la notion de *taux de bonne reconnaissance* pour pondérer les indices de sensibilité et de spécificité. Ainsi, on diminue la sensibilité dans les cas où l'instance d'un motif n'est pas reconnue ; et la spécificité dans le cas où des éléments d'une classe ne sont représentatifs d'aucun motif. Les taux de reconnaissance sont ainsi des réels compris entre 0 et 1, et définis pour chaque motif i, ρe_i, et chaque classe j, ρp_j, selon les équations suivantes :

$$\rho e_i = \frac{\sum_{j=1}^{n} c_{ij}}{m_i} \text{ et } \rho p_j = \frac{\sum_{i=1}^{m} c_{ij}}{n_j}. \tag{6.5}$$

(2) Instance fractionnée.

Afin de disposer d'une mesure de performance appropriée à la découverte éventuellement fractionnée de l'instance d'un motif, on introduit la notion de *taux de fractionnement* pour chaque instance k d'un motif i, noté $1/\eta_{ik}$, où η_{ik} est le nombre de tentatives de motifs associés à la même instance k d'un motif i. Si une instance d'un motif correspond à une seule sous-séquence d'une seule classe, ce taux est égal à 1.

La prise en compte des taux de fractionnement associés aux différentes instances de motifs intervient au moment de la construction de la matrice de confusion. La matrice de confusion

$C = \{c_{ij}\}$, initialement nulle, est construite en considérant successivement chaque sous-séquence de chaque classe j issue de la classification, et en ajoutant 1 à c_{ij} quand la sous-séquence considérée de la classe j correspond à l'instance k d'un motif i. Pour prendre en compte le cas où plusieurs sous-séquences sont associées à une même instance k d'un motif i, on propose alors d'utiliser une nouvelle **matrice de confusion** $C' = \{c'_{ij}\}$ qui est construite en ajoutant non plus 1 mais $1/\eta_{ik}$ à c'_{ij} chaque fois qu'une sous-séquence de la classe j correspond à l'instance k d'un motif i. Si dans le cas inverse une sous-séquence recouvre plusieurs instances de motifs, elle sera par défaut associée à l'une d'entre elles et les autres considérées comme non reconnues.

Ainsi :
- La somme $\sum_{j=1}^{n} c'_{ij}$ représente le nombre d'instances du motif i correctement identifiées, même si pas forcément bien classées, si bien que cette somme est un entier qui vérifie la propriété suivante : $\sum_{j=1}^{n} c'_{ij} \leq m_i$;
- La somme $\sum_{i=1}^{m} c'_{ij}$ représente le nombre d'instances distinctes d'un motif représentées dans la classe j. Cette valeur n'est pas systématiquement entière car une sous-séquence d'une classe peut correspondre à une fraction seulement de l'instance d'un motif, l'autre fraction ayant été intégrée dans une autre classe. On a également : $\sum_{i=1}^{m} c'_{ij} \leq n_j$.

En conséquence de la redéfinition de la matrice de confusion, on reforme également les équations (6.5) définissant les *taux de reconnaissance*, ρe_i et ρp_j, selon les nouvelles équations suivantes :

$$\rho e_i = \frac{\sum_{j=1}^{n} c'_{ij}}{\sum_{j=1}^{n} c'_{ij} + m'_i} \text{ et } \rho p_j = \frac{\sum_{i=1}^{m} c'_{ij}}{\sum_{i=1}^{m} c'_{ij} + n'_j}, \tag{6.6}$$

où m'_i est le nombre d'instances du motif i non reconnues, dans aucune classe,
n'_j nombre d'éléments de la classe j qui ne sont représentatifs d'aucune instance de motif.

On définit également un indice de fractionnement des instances de motifs, λ, selon l'équation 6.7. Un indice maximum égal à 1 signifie que la reconnaissance des tentatives de motifs n'est absolument pas fractionnée.

$$\lambda = \frac{1}{n \cdot m} \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{c'_{ij}}{c_{ij}} \tag{6.7}$$

Finalement, en utilisant les dernières définitions de ρe_i et ρp_j, et à partir des équations initiales (6.3) et (6.4), les indices de sensibilité et de spécificité respectivement pour chaque motif et chaque classe sont redéfinis selon les équations (6.8) et (6.9) :

$$Se_i = \rho e_i \cdot \left(1 + \frac{1}{log(n)} \sum_{j=1}^{n} \frac{c'_{ij}}{\sum_{j=1}^{n} c'_{ij}} \cdot log\left(\frac{c'_{ij}}{\sum_{j=1}^{n} c'_{ij}}\right)\right), \tag{6.8}$$

$$Sp_j = \rho p_j \cdot \left(1 + \frac{1}{log(m)} \sum_{i=1}^{m} \frac{c'_{ij}}{\sum_{i=1}^{m} c'_{ij}} \cdot log\left(\frac{c'_{ij}}{\sum_{i=1}^{m} c'_{ij}}\right)\right). \tag{6.9}$$

On peut finalement définir les valeurs moyennes de sensibilité et de spécificité en considérant conjointement les valeurs calculée respectivement pour toutes les lignes et toutes les colonnes, selon les équations suivantes :

$$Se = \frac{1}{m} \sum_{i=1}^{m} Se_i \text{ et } Sp = \frac{1}{n} \sum_{j=1}^{n} Sp_j.$$

Les exigences sur ces indices de performance sont particulièrement fortes puisqu'ils permettent d'attester finalement la bonne reconnaissance et la correcte classification des comportements récurrents, ce qui constitue la réponse aux objectifs du système. La localisation précise des sous-séquences récurrentes dans la séquence initiale est largement moins indispensable.

Interprétation en terme des notions classiques de sensibilité et de spécificité
Dans les cas particuliers où $m = 1$ ou $n = 1$, les équations ci-dessus ne peuvent pas être appliquées et n'auraient de toutes façons pas de sens. On définit alors les indices de sensibilité de chaque motif et de spécificité de chaque classe comme :

$$Se_i = \rho e_i = \frac{\sum_{j=1}^{n} c_{ij}^{'}}{\sum_{j=1}^{n} c_{ij}^{'} + m_i^{'}} \text{ et } Sp_j = \rho p_j = \frac{\sum_{i=1}^{m} c_{ij}^{'}}{\sum_{i=1}^{m} c_{ij}^{'} + n_j^{'}}.$$

Compte tenu des définitions des *taux de reconnaissance*, ρe_i et ρp_j définis par les équations (6.6), on retrouve bien des équations qui semblent homogènes aux définitions classiques de sensibilité et de spécificité. Les valeurs $m_i^{'}$ ou $n_j^{'}$ impliquées dans ces équations représentent en effet le nombre d'*erreurs* – respectivement, "faux négatifs" ou "faux positifs" – dans la considération des classes constituées par rapport aux motifs initiaux :

- $m_i^{'}$ est le nombre d'instances de motifs non reconnues – "faux négatifs", et
- $n_j^{'}$ le nombre d'éléments d'une classe qui ne correspondent en fait à aucun motif – "faux positifs".

Dans le cas de la sensibilité liée à la reconnaissance d'un motif i, l'équation définissant ρe_i s'interprète alors particulièrement bien :

$$Se_i = \rho e_i = \frac{\sum_{j=1}^{n} c_{ij}^{'}}{\sum_{j=1}^{n} c_{ij}^{'} + m_i^{'}}.$$

où $\sum_{j=1}^{n} c_{ij}^{'}$ représente le nombre d'instances du motif i correctement identifiées, c'est-à-dire une sorte de taux de "vrais positifs" (VP),
$m_i^{'}$ le nombre d'instances du motif i non reconnues, dans aucune classe, c'est-à-dire une sorte de taux de "faux négatifs" (FN),

si bien que l'équation définissant la sensibilité correspond bien à une définition classique :

$$Se = \frac{VP}{VP + FN}.$$

6.1.5 Synthèse du processus expérimental

Le schéma de la figure 6.1 résume le principe général d'expérimentation de l'extraction de motifs, et reprend en particulier la définition des étapes suivantes : (1) le contexte expérimental, (2) les données de l'expérimentation, (3) les méthodes d'évaluation du système et (4) les mesures de performances.

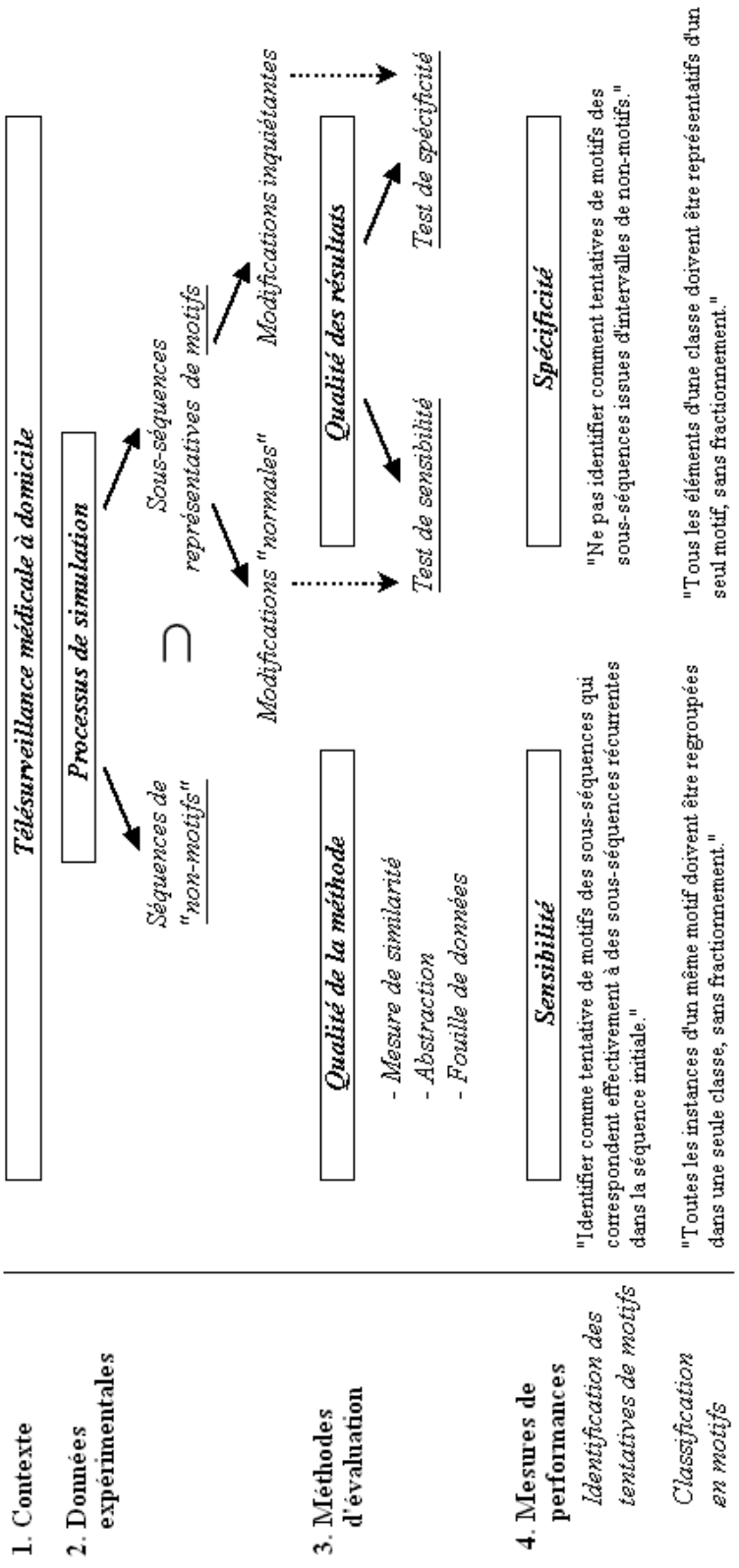


FIG. 6.1 – Synthèse du processus d'expérimentation de l'extraction de motifs.

6.2 Qualité de la méthode

La validation de la qualité de la méthode d'extraction de motifs proposée est réalisée à plusieurs niveaux, concernant :

- 6.2.1 L'efficacité de la **mesure de similarité** proposée entre séquences multidimensionnelles et hétérogènes ;
- 6.2.2 La pertinence de l'**abstraction** des séquences brutes en une succession de symboles estampillés ;
- 6.2.3 L'efficacité de la **fouille de données** pour l'identification et la classification des sous-séquences récurrentes.

6.2.1 Mesure de similarité

Dans ce paragraphe on s'intéresse à la validation de la mesure de distance réelle proposée pour la comparaison de séquences de données multidimensionnelles et hétérogènes, présentée au paragraphe 4.2. L'autre mesure de distance proposée – la distance minimum – est une distance approchée définie exclusivement pour l'agrégation de vecteurs successifs à l'étape d'abstraction. Elle est ainsi discutée dans le paragraphe 6.2.2 dédié à la validation de cette étape.

Processus expérimental

L'efficacité de la mesure de distance proposée est évaluée par sa capacité à considérer comme proches des séquences effectivement représentatives d'un même comportement, et à l'inverse comme éloignées celles qui ne le sont pas. Dans le contexte de la télésurveillance médicale à domicile, on s'attend ainsi à trouver les caractéristiques de distance suivantes :

- *La distance entre des séquences représentatives de la réalisation d'une même activité, dans des conditions habituelles, doit être faible.*
  On doit prendre en compte les modifications "normales" possibles entre les séquences correspondantes : interruptions, déformations temporelles, et variabilité dans les valeurs (voir paragraphe II.6.5.2).
- *La distance entre des séquences représentatives d'activités différentes doit être élevée.*
  On doit prendre en compte à ce niveau les modifications inquiétantes possibles d'un comportement, afin de ne plus considérer comme proches des sous-séquences correspondant à la réalisation d'une même activité mais dans de mauvaises conditions pour l'une des deux (voir paragraphe II.6.5.3).

Étant données des séquences représentatives d'une certaine activité et d'autres qui ne le sont pas, on s'attend alors à ce que la mesure de distance proposée permette de classer correctement ces séquences en deux classes, selon un seuil de distance prédéfini. Les graphes de la figure 6.2 présentent ainsi deux classes *a priori* de séquences expérimentales, définies comme suit :

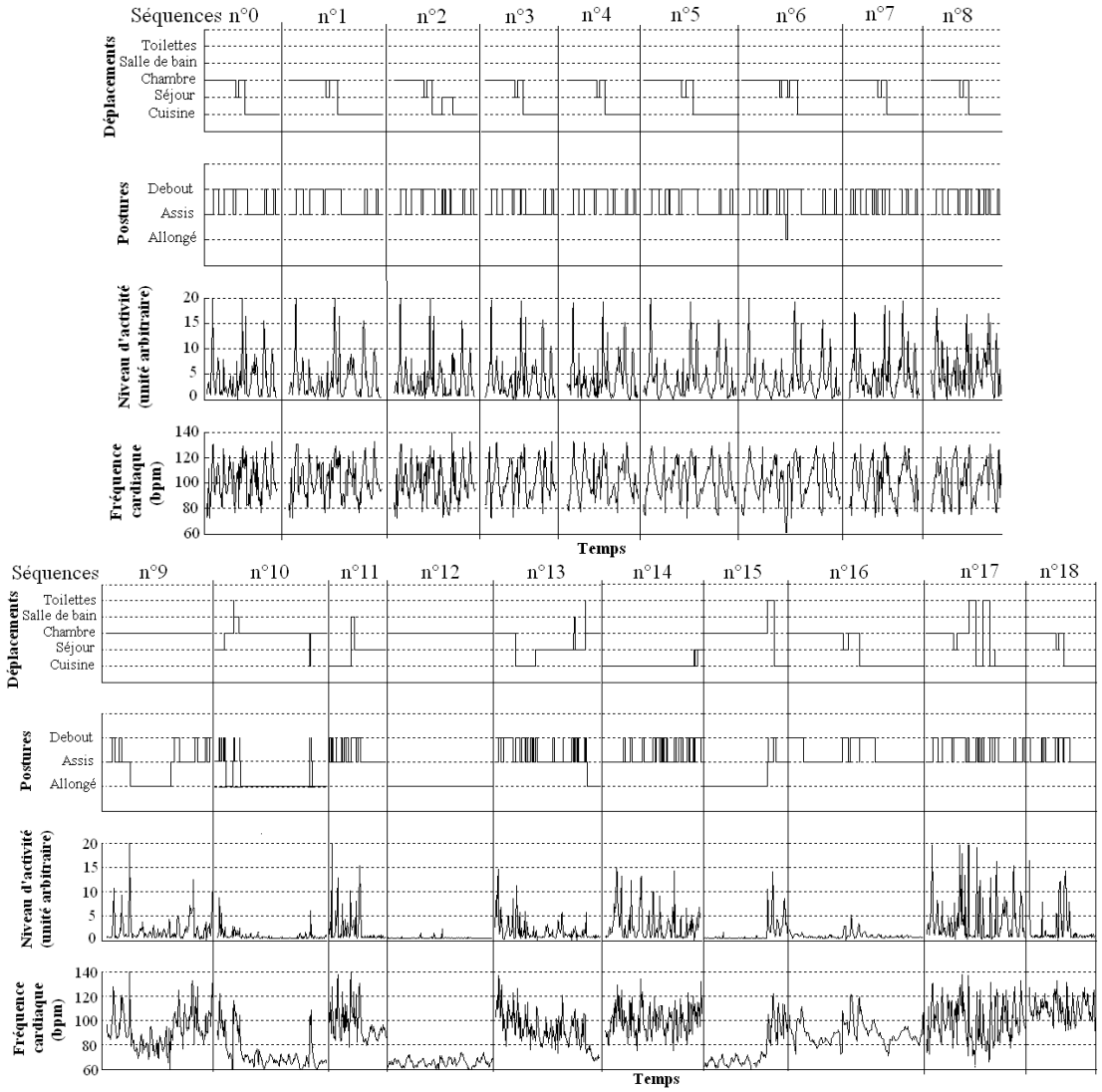


FIG. 6.2 – Présentation de deux ensembles de séquences expérimentales.

- **Classe 0.** Les séquences représentées sur la figure du haut (0 à 8) sont toutes représentatives d'une même activité ;
- **Classe 1.** Les séquences représentées sur la figure du bas (9 à 18) correspondent à des activités différentes.

Classe 0. **Séquences numérotées de 0 à 8, représentatives d'une certaine activité** – *le lever le matin.* Ces séquences sont toutes générées à partir d'une séquence de référence – séquence n°0 – en lui appliquant une ou plusieurs modifications "normales" de comportement, avec différents taux de bruit appliqués selon des valeurs croissantes avec le numéro de la séquence. En particulier, les modifications apportées successivement pour générer les séquences 1 à 8 comprennent :

(a) Déformation temporelle : séquences 1, 5 et 6 ;

(b) Interruption : séquences 2 et 6 ;

(c) Variabilité dans les valeurs, selon des **taux croissants** : 0.1 (séquence 3), 0.3 (séquences 4 à 6), 0.6 (séquence 7) et 1.0 (séquence 8).

Classe 1. **Séquences numérotées de 9 à 18, représentatives d'autres activités** – *dormir, prendre un repas, profiter d'une activité calme, etc.* On inclut également dans cet ensemble des séquences représentatives de l'activité de référence mais réalisées dans de mauvaises conditions – séquences 16 à 18 : respectivement, lenteur dans les activités, longues interruptions aux toilettes, fréquence cardiaque anormalement élevée.

La séquence n°0 servant de référence, le processus expérimental a pour objectif le classement des autres séquences, numérotées de 1 à 18, en comparaison à la séquence n°0, et selon un seuil de distance, noté th. Si la mesure de distance est appropriée, elle doit alors permettre de discriminer correctement les séquences en calculant des distances inférieures au seuil th entre la séquence 0 et les séquences 1 à 8 (classe 0), et supérieures au seuil entre la séquence 0 et les séquences 9 à 18 (classe 1). On compare en particulier les résultats obtenus avec deux méthodes :

- **Distance *LCSS*.** Méthode détaillée dans le paragraphe 4.2, basée sur la longueur de la plus longue sous-séquence commune : recherche de la plus longue sous-séquence de paires de points similaires, un point d'une séquence ne devant être associé qu'à au plus un point de l'autre séquence.
- **Distance *DTW*.** Méthode basée sur le principe de "*Dynamic Time Warping*" : recherche des paires de points qui minimisent la distance totale, chaque point d'une séquence devant être au moins associé à un point de l'autre séquence. Cette méthode est présentée de façon détaillée en annexe I.

Par ailleurs, on compare à chaque fois les résultats obtenus en calculant les distances à partir des séquences brutes d'une part, et des séquences prétraitées (filtrage moyen pondéré et normalisation) d'autre part. On évalue ainsi l'influence du prétraitement sur les distances calculées et par conséquent la pertinence de l'application de cette mesure aux séquences brutes et/ ou prétraitées.

Sélection des paramètres de la distance *LCSS*

La distance dite *LCSS* nécessite la définition *a priori* de deux paramètres, δ_{LCSS} et ϵ_{LCSS}, qui définissent l'écart maximum autorisé, respectivement dans le temps et sur les valeurs, entre deux points similaires. Les graphes de la figure 6.3 présentent un exemple de l'influence du choix de ces paramètres sur les valeurs de distance observées en moyenne entre les séquences de la classe 0 et la séquence de référence (séquences similaires).

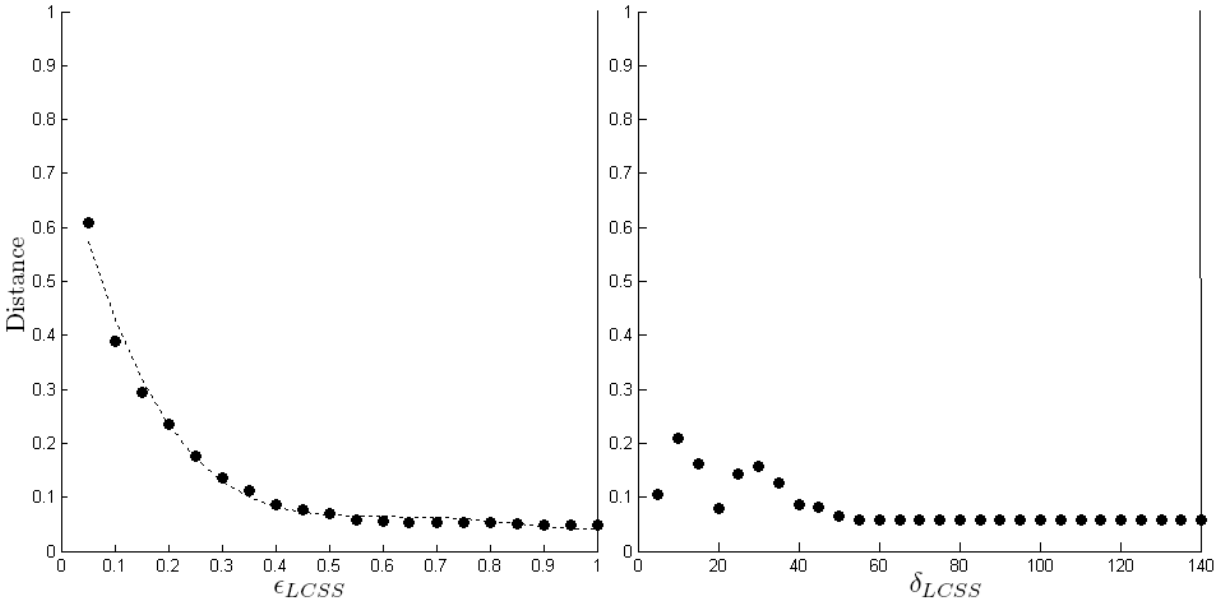


FIG. 6.3 – Distances $LCSS$ observées en moyenne entre les séquences de la classe 0 et la séquence de référence n°0, en fonction du choix des paramètres ϵ_{LCSS} et δ_{LCSS}.
Le graphe de gauche présente les variations de ces distances en fonction du choix de ϵ_{LCSS}, et avec δ_{LCSS} non restrictif ; celui de droite les variations en fonction du choix de δ_{LCSS}, et avec $\epsilon_{LCSS} = 0.3$.

L'écart temporel δ_{LCSS} est défini au plus grand dans cette expérimentation, de façon non restrictive – c'est-à-dire, δ_{LCSS} est égal à la longueur de la plus longue des deux sous-séquences comparées. L'idée est de ne pas contraindre dans le temps la détermination des paires de points similaires, et de comparer ainsi "au plus sévère" la distance $LCSS$ par rapport à la distance DTW qui, elle, ne donne pas la possibilité de cette limitation.
En choisissant une valeur de δ_{LCSS} plus restrictive, c'est-à-dire plus faible, on autorise potentiellement moins d'associations de points similaires. Ceci résulte globalement en une augmentation des distances observées (voir Fig. 6.3).

L'écart sur les valeurs ϵ_{LCSS} est défini par quelques expérimentations préliminaires sur des calculs de distance. L'objectif est de définir une valeur qui permette de suffisamment discriminer les classes de sous-séquences qui se ressemblent de celles qui sont très différentes, et par conséquent de générer des distances "suffisamment" faibles entre séquences similaires. L'évolution de la distance moyenne des séquences de la classe 0 avec la séquence de référence – séquences similaires – en fonction de ϵ_{LCSS} (voir Fig. 6.3) montre un début de stabilisation des valeurs de distance vers des mesures assez faibles (< 0.15) au-delà de $\epsilon_{LCSS} = 0.3$, valeur que l'on sélectionne ainsi *a priori* pour le paramètre.
En choisissant une valeur de ϵ_{LCSS} supérieure, on diminue encore un peu la distance entre des séquences similaires, mais on diminue également celle de ces séquences avec d'autres qui doivent être considérées éloignées. Par ailleurs, plus la valeur de ϵ_{LCSS} est faible, plus le nombre de points potentiellement similaires entre deux séquences diminue. Par conséquent, les distances observées sont globalement plus élevées.

Classification des séquences expérimentales : discussion sur la qualité des résultats

Remarques générales
Étant données les valeurs définies *a priori* pour les paramètres δ_{LCSS} – valeur non restrictive – et ϵ_{LCSS} – valeur fixée à 0.3 – on compare les performances de la distance non métrique *LCSS* et de la distance métrique *DTW*. L'expérimentation consiste à classer selon un seuil de distance par rapport à la séquence de référence les séquences expérimentales 1 à 18 dont on connaît les classes d'appartenance *a priori*. Les résultats sont présentés sur le graphe de gauche de la figure 6.4.

De façon générale, on remarque que les distances *DTW* calculées sont bien plus faibles que les distances *LCSS* qui s'étalent quant à elles complètement sur la plage de valeurs entre 0 et 1. Cette différence s'explique par deux raisons principales :

- l'ordre différent de calcul : 1 pour *LCSS* et 2 pour *DTW*,
- la possibilité d'associations multiples de chaque point des séquences comparées avec une distance *DTW*, en recherchant par ailleurs les paires de points qui induisent une distance minimum. Par conséquent, la distance globale – une distance euclidienne à partir de ces paires de points similaires – peut souvent rester assez faible.

Supériorité de la distance LCSS ?
Si on exclut pour le moment le cas de la séquence 17, la supériorité d'une distance *LCSS* par rapport à *DTW* dans notre contexte est mise en évidence par une meilleure discrimination des classes dans le cas d'utilisation de *LCSS*. En particulier, les distances *DTW* associées aux séquences 16 et 18, représentatives d'une modification inquiétante du comportement de référence, ne sont pas discriminées des séquences similaires à la séquence de référence, alors qu'elles le sont dans le cas d'une distance *LCSS*. Les raisons de cette mauvaise discrimination des classes dans le cas d'une distance *DTW* par rapport à *LCSS* est illustré sur la figure 6.5 dans le cas de la comparaison de la séquence 16 avec la séquence de référence. Cette figure présente le comportement de chacune des distances en terme de la constitution des associations de points similaires de chacune des séquences, nécessaires dans chaque cas au calcul de distance.

- **Distance *DTW*.** Les associations multiples sont possibles et tous les points de chaque séquence doivent être associés à au moins un point de l'autre séquence, selon un critère de distance minimum. Dans cet exemple, à cause de la proximité des séquences de déplacements et de postures, le peu de points correspondant à de faibles niveaux d'activité et fréquences cardiaques dans la séquence 0 est multiplement associé à la grande quantité de tels points dans la séquence 16. Réciproquement, le peu de fortes valeurs de ces paramètres sur la séquence 16 entraîne leur association multiple à la grande quantité de ces points sur la séquence 0. Finalement, les associations de points sont réalisées de telle façon qu'il y a peu de couples correspondant vraiment à une grande valeur de distance, si bien que la distance globale est faible.
- **Distance *LCSS*.** La force de *LCSS* dans ce contexte est de fonder la similarité entre les points sur un critère de seuil maximum, en permettant ainsi de ne pas apparier certains points, et en excluant par ailleurs la possibilité d'associations multiples. Les distances sont alors élevées dès qu'il y a peu de couples de points similaires par rapport aux dimensions des séquences. Si on contraint en plus la similarité des points dans le temps par le paramètre δ_{LCSS}, on obtient une distance encore plus élevée entre les séquences 0 et 16.

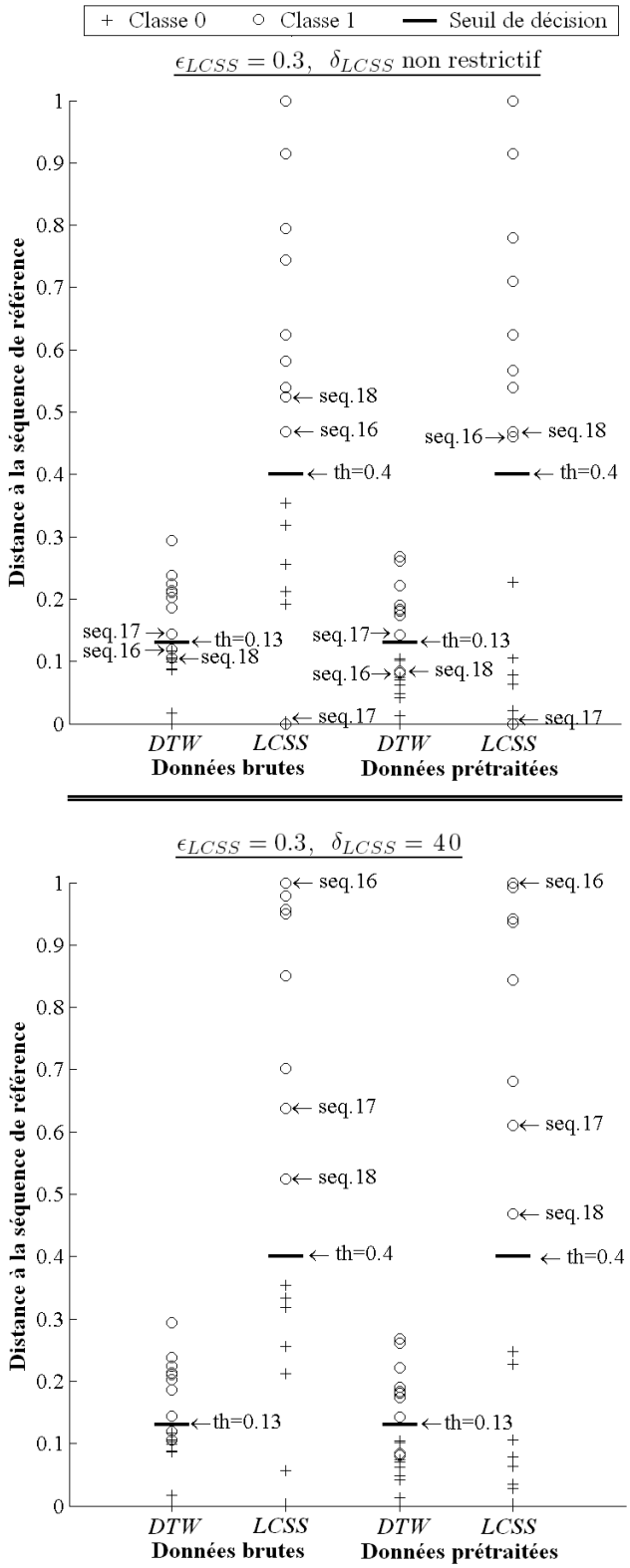


FIG. 6.4 – Représentation des distances *DTW* et *LCSS* entre chaque séquence des classes *a priori* 0 et 1 et la séquence de référence n°0.

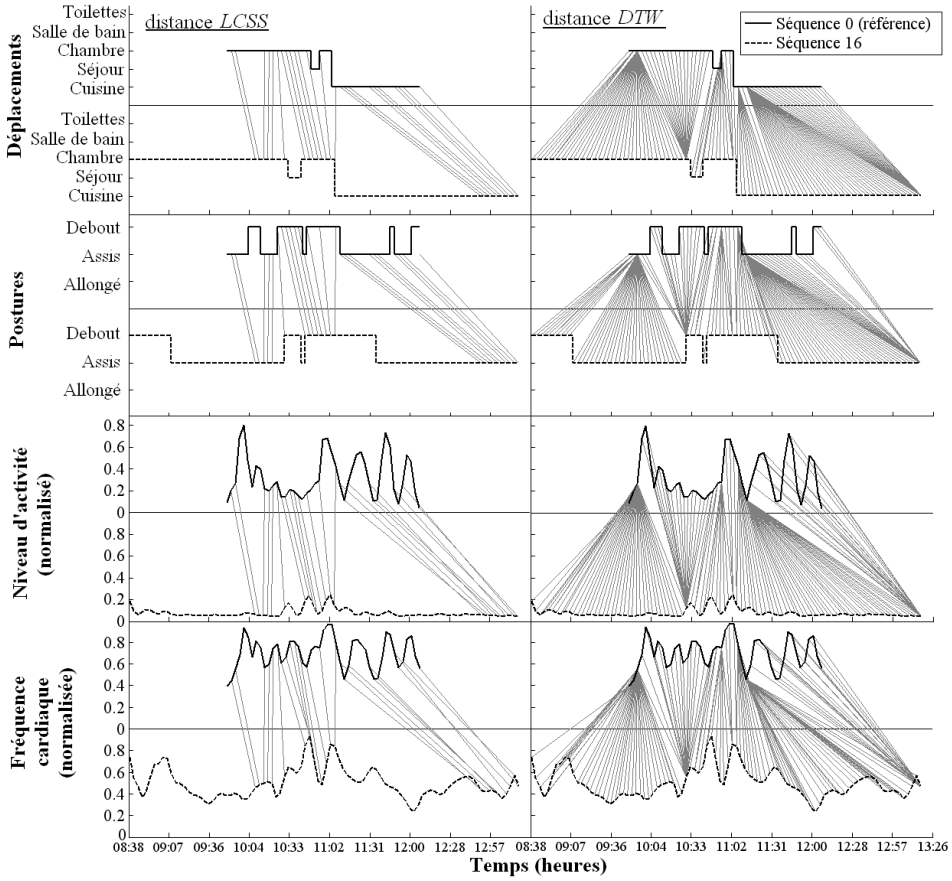


FIG. 6.5 – Représentation des couples de points considérés similaires dans le calcul des distances $LCSS$ et DTW pour la comparaison des séquences 0 et 16.

Cas de la séquence 17

Malgré cette forte présomption de supériorité de $LCSS$, le graphe de gauche de la figure 6.4 montre aussi l'échec de cette distance pour la bonne classification de la séquence 17, à laquelle est même associée une distance minimum, nulle, avec la séquence de référence. Cette distance s'explique cependant tout simplement du fait de la constitution de cette séquence : il s'agit exactement de la séquence de référence dans laquelle on a inséré deux longues interruptions inquiétantes correspondant à des passages aux toilettes (voir Fig. 6.2). La plus longue sous-séquence commune aux séquences 0 et 17 est ainsi exactement la séquence 0, quel que soit le seuil ϵ_{LCSS}, ce qui résulte en une similarité de 1 puis une distance nulle.

Le problème ainsi mis en évidence se résout en agissant sur la valeur du paramètre δ_{LCSS}, choisi initialement non restrictif. Ce choix correspondait à la volonté de comparer d'abord les distances $LCSS$ et DTW dans des conditions proches d'appariement des points de chaque séquence, c'est-à-dire indépendamment de la définition d'une contrainte d'écart temporel entre points similaires. Le graphe de droite de la figure 6.4 montre les résultats alors obtenus en choisissant un écart maximum de $\delta_{LCSS} = 40$ minutes entre deux points similaires. Un calcul de distance $LCSS$ permet d'aboutir à une classification parfaite des séquences expérimentales, et par ailleurs d'obtenir une bonne séparation des classes avec un écart de distance d'environ 0.1 dans le cas des séquences brutes, et de plus de 0.2 dans le cas des séquences prétraitées.

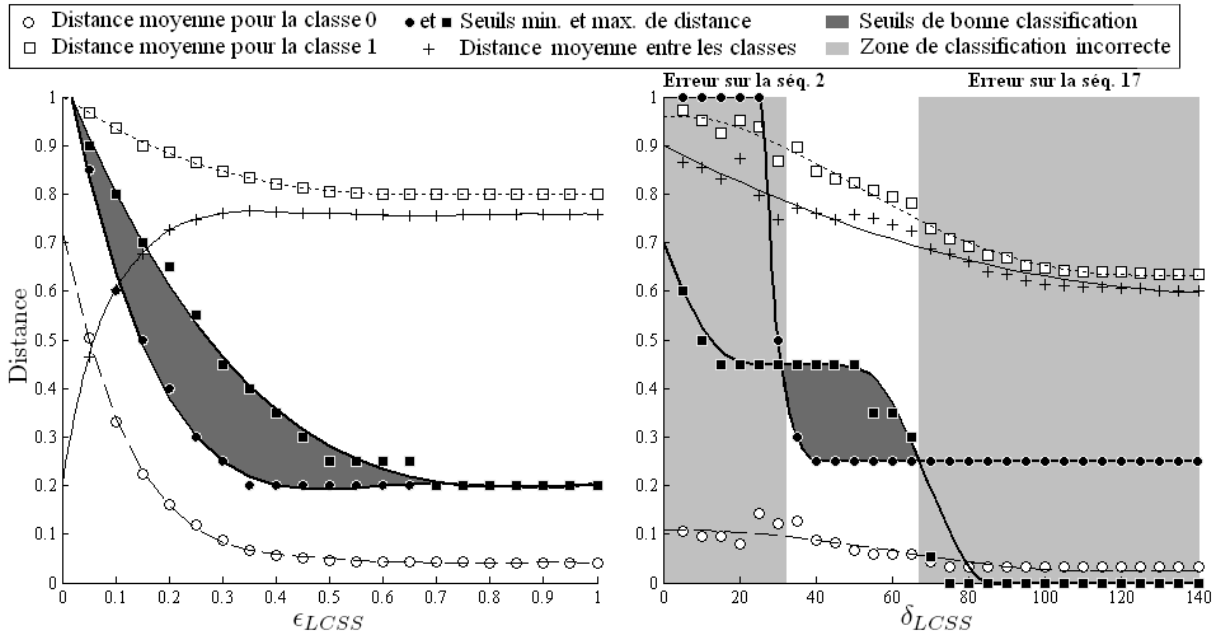


FIG. 6.6 – Distances *LCSS* observées en fonction du choix des paramètres ϵ_{LCSS} et δ_{LCSS}. On observe les distances *LCSS* pour la classification des séquences prétraitées 1 à 18 dans les classes 0 ou 1, selon leur proximité à la séquence de référence n°0, et en fonction d'un seuil de distance. Les graphes montrent aussi les seuils minimum et maximum de distance nécessaires pour une classification correcte, ainsi que l'écart moyen de distance entre les classes constituées.

- Le graphe de gauche présente les variations de ces distances lorsque le seuil de similarité sur les valeurs, ϵ_{LCSS}, varie entre 0 et 1, alors que $\delta_{LCSS} = 40$ minutes ;
- Le graphe de droite présente ces variations pour des valeurs croissantes du seuil de similarité dans le temps, δ_{LCSS}, exprimé en minutes ou nombre de points de chaque séquence (la période d'échantillonnage étant d'une minute), alors que $\epsilon_{LCSS} = 0.3$.

Facteurs d'influence des mesures de distance

Influence du prétraitement des séquences

Les remarques précédentes sont valables à la fois pour la considération des séquences brutes et prétraitées. Si on utilise une distance *LCSS*, on remarque par ailleurs que les classes sont mieux séparées avec les données prétraitées. Le prétraitement des séquences induit des distances largement plus faibles entre séquences similaires (baisse d'environ 0.1), alors qu'elles ne diminuent que faiblement entre séquences différentes. Ceci confirme l'efficacité de la distance *LCSS* dans notre contexte, de même que l'intérêt d'une étape de prétraitement des séquences étudiées.

Influence du bruit : contraintes sur le choix des paramètres ϵ_{LCSS} et δ_{LCSS}

Les résultats obtenus sur l'expérimentation des mesures de distance sont finalement conformes à la littérature qui tend à prouver qu'une distance métrique de type *DTW* est très peu résistante à la présence de bruit dans les données, par rapport à une distance non métrique de type *LCSS* (cas par exemple de la séquence 16 – voir Fig. 6.5). On a mis cependant aussi en évidence la sensibilité du réglage des deux paramètres fondamentaux décrivant les contraintes de similarité entre points :

- ϵ_{LCSS}**, l'écart de valeurs maximum** autorisé entre deux points pour qu'ils puissent être considérés similaires ;
- δ_{LCSS}**, l'écart temporel maximum** entre deux points similaires.

Les graphes de la figure 6.6 illustrent en particulier dans le contexte de l'expérimentation proposée l'évolution en fonction des valeurs de ces deux paramètres de la distance moyenne des séquences de chaque classe avec la séquence de référence, de l'écart entre les classes et des seuils minimum et maximum permettant d'aboutir à une bonne classification des sous-séquences. Le seuil minimum est la plus petite valeur du seuil de distance qui induit une sensibilité maximum de la classification : toutes les sous-séquences de la classe 0 *a priori* sont effectivement regroupées comme similaires à la séquence de référence. Le seuil maximum est la plus grande valeur de seuil qui préserve une spécificité maximum : le groupe des sous-séquences considérées comme similaires à la séquence de référence ne contient que des sous-séquences de la classe 0. D'après ces définitions, si le seuil minimum est supérieur au seuil maximum, la classification des sous-séquences ne peut pas être parfaite, quel que soit le seuil sélectionné (cas des zones grisées du graphe de droite de la figure 6.6). Sinon, toutes les valeurs de seuil comprises entre le minimum et le maximum conduisent à une classification exacte des sous-séquences. Dans l'exemple proposé, les valeurs de seuil sont données à 0.05 près.

Le choix de ϵ_{LCSS}, contraignant l'écart possible sur les valeurs de points similaires, est particulièrement important à régler en fonction de la **variabilité possible dans les valeurs** de séquences similaires. Sur l'exemple proposé, on observe que les valeurs de distance se stabilisent à partir d'une certaine valeur du paramètre. Cette limite non nulle des distances calculées pour des valeurs croissantes de ϵ_{LCSS} est due aux contraintes sur l'égalité des paramètres qualitatifs. La similarité de deux points nécessite en effet au minimum l'égalité des paramètres qualitatifs, ce qui impose une limite supérieure sur le nombre de points similaires entre deux séquences. Lorsque le choix de l'écart maximum sur les valeurs est très élevé – c'est-à-dire, n'est plus contraignant – ce sont ainsi les valeurs des paramètres qualitatifs exclusivement qui régissent la mesure de distance et imposent une limite qui peut être strictement positive pour les distances.
La valeur de ϵ_{LCSS} ne doit par ailleurs pas être trop faible afin de supporter la présence de bruit dans les valeurs et générer des distances faibles même entre des séquences bruitées. En augmentant ϵ_{LCSS}, on diminue ainsi les distances mais on augmente également *en moyenne* la distance entre les classes. Par contre on observe sur cet exemple que l'écart absolu entre les classes – environ la différence entre les seuils maximum et minimum – devient très faible, voire quasiment nul pour de grande valeurs de ϵ_{LCSS}. Cela est dû à la séquence 18, associée à la classe 1 car les valeurs de fréquence cardiaque sont trop élevées par rapport à la séquence de référence. Avec l'augmentation de ϵ_{LCSS}, sa distance devient de plus en plus faible avec la séquence de référence, jusqu'à la limite supérieure des distances observées pour la classe 0. La classification reste ainsi correcte mais les classes sont très mal discriminées. *On met ainsi en évidence la nécessité de trouver un compromis dans le choix de ϵ_{LCSS} pour la discrimination des taux "normaux" et inquiétants de bruit dans les valeurs.*
Ainsi, d'après cette expérimentation, un intervalle assez large de valeurs de ϵ_{LCSS} semble appropriées lorsque $\delta_{LCSS} = 40$ – environ entre 0.2 et 0.6 – du moment que le seuil de distance est adapté à la valeur choisie. On préfère cependant *a priori* les valeurs plus centrales de cet intervalle, comprises entre 0.3 et 0.4, qui permettent de mieux discriminer les classes.

Le choix de δ_{LCSS}, contraignant l'écart possible dans le temps entre deux points similaires, est particulièrement important à régler en fonction des **déformations temporelles** et des **interruptions** présentes entre des séquences similaires. Sur l'exemple proposé, on constate que

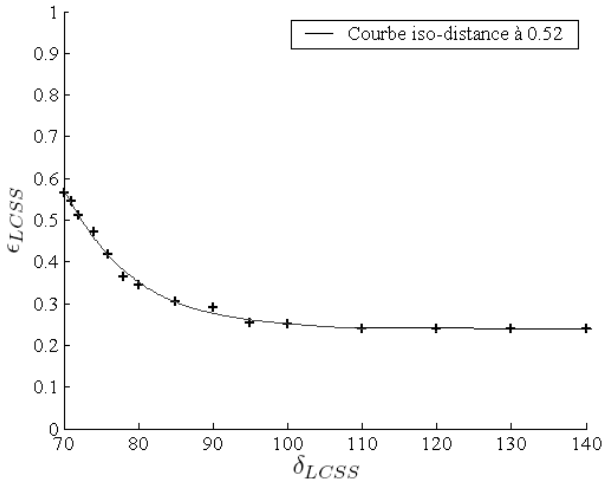


FIG. 6.7 – Influence relative du choix des paramètres ϵ_{LCSS} et δ_{LCSS}.
On trace les couples de valeurs $(\delta_{LCSS}, \epsilon_{LCSS})$ qui permettent d'obtenir une distance d'environ 0.52 entre la séquence 16 et la séquence de référence.

la classification des sous-séquences est erronée aux limites inférieure et supérieure de l'intervalle de variation de δ_{LCSS} (zones grisées du graphe). Vers la limite inférieure, ce sont progressivement les séquences 1, 2, 5 et 6 contenant interruptions et déformations temporelles par rapport à la séquence de référence qui ne sont pas correctement classées et considérées comme éloignées de celle-ci ; vers la limite supérieure, c'est la séquence 17 contenant une longue interruption inquiétante qui n'est pas correctement classée car considérée comme identique à la séquence de référence. *On met ainsi en évidence la nécessité de trouver un compromis dans le choix de δ_{LCSS} pour la discrimination des taux "normaux" et inquiétants de déformations et d'interruptions.*
En augmentant beaucoup la valeur de δ_{LCSS}, on diminue par ailleurs beaucoup plus les distances relatives aux séquences différentes de la séquence de référence (classe 1) que les autres, induisant une diminution de la distance entre les classes. En effet, avec plus de flexibilité dans le temps il est plus facile de trouver des points similaires entre des séquences qui ne le sont pourtant globalement pas, alors qu'il n'y a plus trop d'influence sur les associations de points similaires entre des séquences qui le sont effectivement.
D'après cette expérimentation dans le contexte bruité de notre application, une valeur de δ_{LCSS} comprise environ entre 35 et 65 semble appropriée lorsque $\epsilon_{LCSS} = 0.3$.

Dans ces conditions, les valeurs appropriées du seuil maximum de distance se situent environ entre 0.2 et 0.6, en fonction du choix des paramètres de similarité.

Influence relative des paramètres ϵ_{LCSS} et δ_{LCSS}
Le choix des paramètres ϵ_{LCSS} et δ_{LCSS} est lié à la quantité de bruit présent dans les séquences étudiées. Le paramètre ϵ_{LCSS} s'adapte particulièrement à la *variabilité possible dans les valeurs*, et δ_{LCSS} aux **déformations temporelles** et *interruptions*. Les valeurs de ces deux paramètres agissent ainsi sur la mesure de distance pour permettre le rapprochement de séquences bruitées.

Au-delà des spécificités de ces paramètres au regard du bruit dans les séquences étudiées, la figure 6.7 met en évidence les conséquences relatives de leur réglage. Le graphe proposé présente la contrainte nécessaire sur les valeurs, ϵ_{LCSS}, en fonction de celle sur l'axe du temps, δ_{LCSS}, afin de maintenir une même distance dans la comparaison des séquences 0 et 16 (activité de référence mais avec une certaine lenteur dans l'exécution). Plus la contrainte de similarité dans le temps

est forte – valeurs faibles de δ_{LCSS} – et plus la contrainte sur les valeurs doit être faible – valeurs élevées de ϵ_{LCSS} – si l'on veut maintenir une même mesure de distance entre les deux séquences, si bien qu'il faut trouver un "bon" compromis entre ces deux contraintes.

Ces deux paramètres n'ayant cependant pas la même spécificité, l'idéal est de leur définir *indépendamment* des valeurs appropriées. Chaque valeur doit être adaptée à la constitution des meilleures associations de points lors du calcul de distance entre sous-séquences, en fonction des types de bruits présents dans les données. Un critère est en particulier de générer de faibles distances entre séquences similaires, et réciproquement.

Finalement, une mesure de distance fondée sur *la plus longue sous-séquence commune* ($LCSS$) semble appropriée dans notre contexte. Le réglage des paramètres ϵ_{LCSS} et δ_{LCSS} permet par ailleurs de s'adapter aux différents types de bruits présents dans les séquences, respectivement, la variabilité dans les valeurs et les déformations temporelles ou interruptions.

6.2.2 Abstraction

L'abstraction des données brutes comporte plusieurs étapes successives (voir paragraphe 5.1) :

1. **Prétraitement** : filtrage, normalisation, réduction temporelle ;
2. **Discrétisation** des paramètres quantitatifs ; et
3. **Agrégation** des vecteurs le long des intervalles de temps sur lesquels on n'observe pas de variation significative des paramètres.

La validation de la qualité de l'abstraction est alors réalisée à deux niveaux :

1. Vérifier que l'abstraction préserve dans les séquences les tendances de variation qui semblent importantes pour l'identification du profil comportemental d'une personne ;
2. Analyser l'influence des paramètres et des étapes de l'abstraction sur la préservation de ces tendances fondamentales.

Cette étape de validation est largement intuitive, et peut être réalisée avec l'aide d'experts. L'idée générale est de vérifier la pertinence de l'abstraction au regard de la décision. Les paramètres impliqués dans cette tâche de représentation sont cependant fortement contraints par le contexte et les objectifs de la décision, si bien que le choix des valeurs possibles de ces paramètres est relativement limité. On compare alors intuitivement les séquences résultant de l'abstraction avec les valeurs possibles de ces paramètres en vérifiant les critères suivants :

- préserver les tendances globales de variation,
- supprimer les variations qui ne paraissent *a priori* pas significatives pour l'étude des habitudes de vie d'une personne.

Finalement, la validation des étapes d'abstraction et la définition des paramètres clés sont réalisées sur les séquences de données simulées représentatives d'un individu "moyen", afin de comparer les résultats obtenus, décrits ci-dessous, avec les connaissances empiriques "moyennes" mentionnées dans la littérature. Les paramètres de la simulation étudiés par des analyses mathématiques et statistiques sont ainsi définis selon les valeurs moyennes observées d'après les données expérimentales. Les résultats ont été validés intuitivement, avec parfois le support de connaissances empiriques (cas de la définition des intervalles de discrétisation par exemple). La figure 6.8 illustre par ailleurs les différentes étapes de l'abstraction.

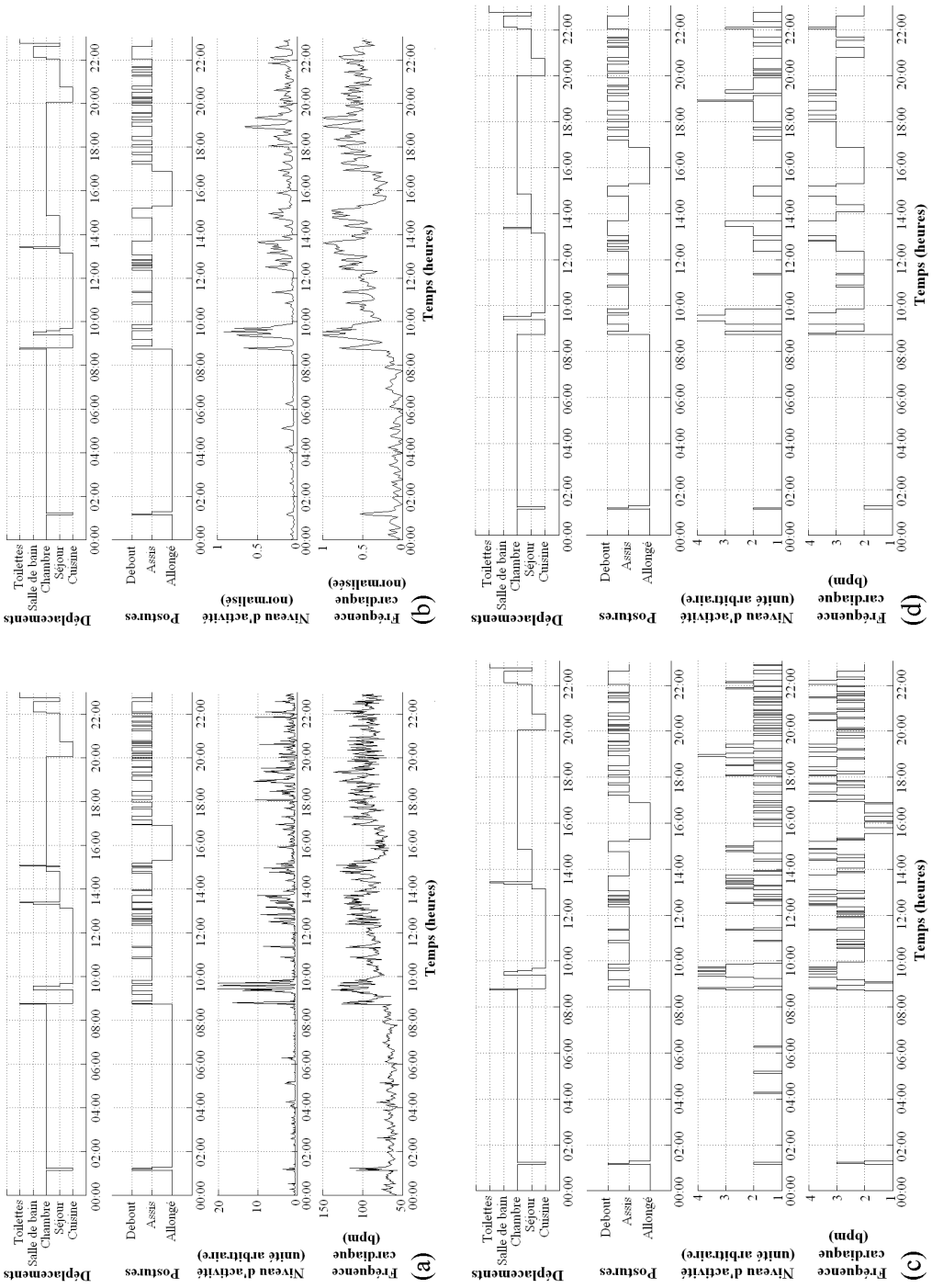


FIG. 6.8 – Illustration du processus d'abstraction.
Les graphes présentés correspondent aux séquences de données à quatre dimensions représentatives d'une personne télésurveillée à domicile, prenant en compte pour chaque graphe, de haut en bas : (1) les déplacements de la personne, (2) ses postures, (3) le niveau d'activité et (4) la fréquence cardiaque moyenne.
De gauche à droite, et de haut en bas, les graphes représentent successivement les séquences suivantes : (a) **séquence brute**, (b) **séquence prétraitée**, (c) **séquence discrétisée**, et (d) **séquence des vecteurs discrets agrégés**.

Prétraitement

Type de filtrage

Pour le filtrage des données brutes, on choisit une méthode de moyenne mobile pondérée pour ne pas trop "aplatir" la courbe et conserver au maximum les "pics" de valeurs significatifs (voir les graphes du haut de la Fig. 6.8). Après expérimentation de filtres moyens et médians, pondérés ou non, il s'avère en effet intuitivement qu'un filtre moyen pondéré met le mieux en évidence les tendances globales de variation, tout en préservant les pics de valeurs importants à considérer en particulier dans un contexte de surveillance. On utilise ainsi un filtre moyen pondéré, symétrique, avec le maximum au centre de la fenêtre. Les pondérations évoluent ensuite de façon homogène et décroissante vers "les bords". On utilise en pratique dans cette application un filtre de largeur correspondant à 11 minutes pondéré par 6 au centre en diminuant les régulièrement les pondérations jusqu'à 1 sur les bords.

Taux de réduction temporelle

Une analyse de la réduction temporelle montre qu'une réduction trop rapide de la fréquence des données fait prendre le risque de perdre des informations essentielles à la compréhension du comportement de la personne, et notamment les pics de valeurs. Par ailleurs, l'étape d'agrégation qui termine l'abstraction génère des séquences de symboles approximativement de même longueur quelle que soit la réduction temporelle effectuée à ce niveau. On décide par conséquent de reposer la réduction de la dimension des données dans le temps sur l'agrégation significative des vecteurs successifs.

Normalisation

La normalisation est ensuite réalisée par la méthode du *min-max* décrite au paragraphe 4.1. Dans le cas de la simulation d'un "individu moyen", on constate d'après les résultats de la simulation des intervalles de variation des valeurs, *pour les séquences filtrées*, d'en moyenne environ 0 à 12 et 60 à 120 respectivement pour le niveau d'activité et la fréquence cardiaque. Ces limites sont ainsi utilisées comme bornes minimum et maximum pour la normalisation.

Enfin, les comparaisons faites dans la section précédente 6.2.1 sur les mesures de distance *LCSS* obtenues à partir des séquences brutes ou prétraitées semblent donner raison à cette définition du prétraitement. En effet, l'écart de distance entre la classe contenant les séquences représentatives d'une même activité et l'autre classe contenant tous types de séquences augmente lorsqu'on utilise les données prétraitées. Cette constatation laisse supposer que le prétraitement met bien en évidence les caractéristiques fondamentales et discriminantes de variation dans ces séquences.

Discrétisation

Méthode de discrétisation

Dans un contexte de surveillance, on apprend les classes ou intervalles caractéristiques de variation des valeurs observées à partir de l'algorithme des *k plus proches voisins* sur les séquences de données générées par le processus de simulation. Cet algorithme nécessite de fixer *a priori* le nombre de classes. Dans notre contexte il s'agit donc de fixer empiriquement ou avec l'aide d'experts le nombre d'intervalles de discrétisation qu'il est pertinent de considérer pour les paramètres quantitatifs. Puisque l'abstraction doit fournir une représentation interprétable des séquences initiales, le nombre d'intervalles de discrétisation doit être significatif en terme des objectifs de décision.

Nombre et délimitation des intervalles

Dans le cadre de la télésurveillance médicale, la forte corrélation entre le niveau d'activité et la fréquence cardiaque incite à considérer le même nombre d'intervalles de discrétisation pour ces deux paramètres. Les connaissances empiriques sur ces paramètres guident par ailleurs le choix du nombre de ces intervalles. Le tableau II.4.2 de classification des travaux physiques d'après la fréquence cardiaque suggère en effet de considérer au plus quatre intervalles de valeurs dans le cadre d'activités à domicile, du repos à une activité modérée. Les fréquences cardiaques associées varient alors jusqu'à environ 125 battements par minute en moyenne – ce qui est bien en accord avec la borne maximum – égale à 120 – définie pour la normalisation des valeurs de fréquence cardiaque.

Avec ces contraintes, les intervalles de discrétisation obtenus pour les différents paramètres par application de l'algorithme des $k-$plus proches voisins sur les données expérimentales d'une personne sont présentés dans le tableau 6.1. Les résultats obtenus sont globalement en accord avec les estimations empiriques. Un niveau d'activité considéré comme de repos au niveau expérimental correspond cependant plutôt à une activité estimée "très légère" empiriquement, avec par conséquent une limite expérimentale supplémentaire entre 75 et 100 bpm – à environ 90 bpm – pour séparer expérimentalement une activité très légère d'une activité légère. Un exemple de résultat de la discrétisation est obtenu sur le graphe (c) de la Fig. 6.8.

	Connaissances empiriques		Expérimentation	
	Niveau d'activité	**Fréquence cardiaque**	**Niveau d'activité**	**Fréquence cardiaque**
1	*Repos*	≈ 65	< 1.8	< 75
2	*Très léger*	< 75	1.8 à 3.8	75 à 92
3	*Léger*	75 à 100	3.8 à 7	92 à 104
4	*Modéré*	100 à 125	> 7	104 à 120

TAB. 6.1 – Intervalles de discrétisation du niveau d'activité et de la fréquence cardiaque.
Le tableau présente les intervalles de discrétisation des paramètres quantitatifs observés dans le cadre de la télésurveillance médicale à domicile – le niveau d'activité (échelle arbitraire) et la fréquence cardiaque (battements par minute) – en comparaison aux délimitations empiriques des valeurs.

Agrégation

Nécessité de l'agrégation

Dans le contexte d'une étude "haut niveau" par rapport aux données exploitées, une simple discrétisation ne suffit pas forcément à une abstraction des séquences temporelles appropriée aux objectifs de décision. Les séquences discrètes contiennent en effet encore des variations non significatives à l'échelle de la décision. Sur la séquence discrète proposée sur la figure 6.8 (graphe (c)), on constate par exemple que des passages des niveaux d'activité discrets 1 à 2 apparaissent subrepticement pendant la nuit. Excepté celui ayant lieu à environ 1 heure du matin, associé au lever de la personne et à un important pic de fréquence cardiaque, les autres n'ont pas de sens au regard des données initiales et de l'échelle de décision.

Seuil maximum de distance

L'agrégation est réalisée par une contrainte sur la distance minimum d'une succession de vecteurs. Dans notre contexte expérimental, on décide de fixer un seuil maximum de distance minimum nul, soit $\sigma = 0$. Cela signifie qu'on agrège les vecteurs successifs tant qu'ils ne présentent pas de

variations significatives, en supposant alors qu'ils sont représentatifs d'une même "action" réalisée : les valeurs des paramètres qualitatifs sont identiques, et celles des paramètres quantitatifs varient au plus dans les intervalles de discrétisation adjacents.
Une autre valeur, supérieure, de ce seuil ne semble pas vraiment appropriée. Particulièrement si on considère le cas des paramètres qualitatifs, un changement dans la pièce occupée ou la posture correspond intuitivement le plus souvent à un changement dans la tâche élémentaire réalisée. On définit par ailleurs peu d'intervalles de discrétisation pour les paramètres quantitatifs, mais leur délimitation doit être significative en terme des situations possibles observables au niveau de décision. Par conséquent, un seuil plus élevé pour l'agrégation ne semble pas non plus pertinent au regard de ces paramètres car il diminuerait l'importance de variations discrètes durables et/ou au-delà des intervalles adjacents de valeurs.

Validité des résultats
Si on observe dans ces conditions les résultats de l'agrégation sur l'exemple de la figure 6.8 (graphe (d)), on remarque que les variations non significatives du niveau d'activité pendant la nuit n'apparaissent plus, alors que la variation correspondant au lever de la personne vers 1h00 est bien mis en évidence. De même au cours de la journée, les périodes caractéristiques observées sur les données initiales (graphe (a)) sont globalement celles qui apparaissent sur la séquence de données agrégées (graphe (d)). On remarque également que les passages rapides aux toilettes sont supprimés à l'issue de l'abstraction. Bien que cette activité fasse partie des activités de base de la vie quotidienne, elle ne peut cependant pas être considérée et étudiée à la même échelle de décision que les autres activités de base telle que l'alimentation ou le sommeil. Les passages aux toilettes sont en effet des tâches élémentaires éventuellement interruptrices des autres activités, que l'on va donc chercher à "gommer" à ce niveau d'étude pour bien reconnaître la réalisation de la tâche principale. Une analyse spécifique ou à une autre échelle permettra de les prendre en compte. Par contre, si la personne commence à passer beaucoup de temps aux toilettes et de façon répétée, alors ces événements ne seront plus gommés par l'abstraction – voir l'exemple d'un plus long passage aux toilettes sur les dernières minutes du graphe (d) – et les activités principales risquent de ne plus être reconnues, signe d'une situation inhabituelle de la personne.

En plus de l'extraction des tendances significatives de variation, l'agrégation temporelle résulte en une réduction de la dimension des séquence étudiées dans le temps. Dans le contexte expérimental présenté, une séquence analysée correspondant à environ 29 jours, soit la succession de *41624 vecteurs discrets* (période d'une minute), est par exemple abstraite en la succession de *1525 symboles*, soit en moyenne environ 52 symboles représentant les activités de la personne pendant une journée.

Finalement, on a ainsi déterminé dans notre contexte expérimental des valeurs qui semblent appropriées pour l'ensemble des paramètres relatifs à l'abstraction des séquences initiales.

6.2.3 Fouille de données

La fouille de données comporte plusieurs étapes successives :

1. **Fouille de caractères** pour l'extraction des sous-séquences récurrentes – les **tentatives de motifs** (voir paragraphe 5.2), qui comprend :
   - Projections aléatoires,
   - Examen de la matrice de collisions,
   - Synthèse des tentatives de motifs ;
2. **Classification** de ces sous-séquences en **motifs** (voir paragraphe 5.3).

Les paramètres impliqués à ces étapes sont différemment contraints *a priori* et spécifiquement influents sur les performances du système. On identifie ainsi deux types de paramètres :

- **D'une part des paramètres dits *méthodologiques*** dont on peut définir *a priori* une valeur appropriée dans un contexte donné : (1) nombre de symboles, (2) dimension de projection, (3) nombre de projections ;
- **D'autre part des paramètres dits *de réglage*** correspondant aux différents seuils à adapter en fonction des caractéristiques des données : (1) seuil minimum de collisions et (2) seuil maximum de distance.

Les paragraphes suivants définissent l'ensemble de ces paramètres et le choix de valeurs ou d'intervalles de valeurs appropriés.

Nombre de symboles définissant les sous-séquences de base

Définition

Il s'agit d'estimer le nombre de symboles approprié à la définition de la longueur des *sous-séquences de base* de la séquence initiale, utilisées pour les projections aléatoires. Cette longueur doit correspondre au nombre minimum de symboles définissant toute instance d'un motif, en terme de l'abstraction de la sous-séquence correspondante. Ce paramètre dépend ainsi du niveau de représentation des séquences initiales, et particulièrement du niveau de la contrainte d'agrégation de vecteurs successifs définissant les symboles.

Choix d'une valeur

Dans notre contexte, le nombre de symboles considéré s'interprète ainsi comme le minimum d'"actions" élémentaires réalisées successivement et qui définissent une activité de la quotidienne. Il a été défini intuitivement au nombre de *4 symboles* lors de la présentation des données expérimentales (voir paragraphe 6.1.2). Sélectionner une valeur inférieure, telle que trois symboles, définissant une sous-séquence de base peut être approprié, mais à l'inverse une valeur supérieure risque de masquer certaines instances de motifs.

Dimension de projection

Définition

Le dimension de projection définit le nombre de symboles et le nombre de composantes de ces symboles considérés pour la comparaison des sous-séquences de base à chaque projection aléatoire. Le réglage de ce paramètre est ainsi réalisé en fonction du taux de "bruit" que l'on souhaite autoriser entre deux sous-séquences considérées similaires.

Choix d'une valeur

Dans notre contexte, une sous-séquence de base comprend 4 symboles, composés eux-mêmes de 4 dimensions. Les instances d'un motif peuvent par ailleurs être très bruitées. On expérimente alors plusieurs dimensions de projections dans le cadre de la comparaison avec une séquence de référence des séquences 1 à 18 utilisées pour l'évaluation de la mesure de similarité (voir paragraphe 6.2.1) : les séquences 1 à 8 sont similaires à la séquence de référence (classe 0), contrairement aux séquences 9 à 18 (classe 1). Un exemple des pourcentages de collisions obtenus entre ces séquences suite à la réalisation de 100 projections aléatoires, avec différentes dimensions de projection, est présenté sur le tableau 6.2.

Globalement, quelle que soit la classe de la séquence comparée à la séquence de référence, le pourcentage maximum de collisions entre les sous-séquences de base augmente lorsque la dimension de projection diminue – la contrainte de collision est relaxée puisqu'on compare moins de symboles et/ ou de paramètres à chaque projection. Par contre, l'écart entre les valeurs

moyenne et médiane du pourcentage de collisions des séquences de chaque classe est sensiblement identique quelle que soit la dimension de projection, voire légèrement inférieur dans le cas de la dimension la plus faible.

Étant donnés ces résultats et l'intuition d'un taux de collisions qui doit rester faible entre des séquences non similaires (classe 1), la dimension de projection qui semble la plus significative consiste à ne comparer à chaque projection que 3 des 4 symboles, sur 3 de ses 4 dimensions. Ceci implique qu'une collision entre sous-séquences est définie par l'égalité de 9 des 16 valeurs discrètes qui la composent.

Dimension de projection	**% Moyen de collisions**			**% Médian de collisions**		
	Classe 0	Classe 1	Écart	Classe 0	Classe 1	Écart
3 × 3	**46.1**	**8.5**	**37.6**	**45.5**	**3.5**	**42.0**
2 × 3	56.2	18.3	37.9	61.5	14.5	47.0
3 × 2	58.2	20.3	37.9	64.0	21.0	43.0
2 × 2	67.5	34.5	33.0	78.5	39.5	39.0

Tab. 6.2 – Pourcentages de collisions observés entre les séquences des classes 0 et 1 et la séquence de référence selon la dimension de projection.
Le tableau présente les pourcentages moyens et médians observés lors de la réalisation de 100 projections aléatoires, ainsi que l'écart entre les pourcentages de collisions correspondant à chacune des classes.
La dimension du masque de projection correspond au nombre de symboles projetés × le nombre de paramètres projetés par symbole. Dans un contexte particulièrement bruité, on considère qu'il faut supporter la présence du bruit à la fois sur les symboles et sur les paramètres qui les composent. La dimension de projection est ainsi d'au plus 3 × 3.

Nombre de projections réalisées pour la construction de la matrice de collisions

Définition

Le réglage de ce paramètre doit assurer d'une part qu'on n'observe pas par hasard un grand nombre de collisions (spécificité), et d'autre part que deux sous-séquences qu'on suppose similaires correspondent bien à un grand nombre de collisions (sensibilité). Plus les sous-séquences de base sont longues, les symboles de grande dimension et le taux de bruit élevé, alors plus le nombre de projections qu'il est nécessaire de réaliser pour obtenir des résultats significatifs est important. Si ces paramètres sont bien définis, on peut alors choisir le nombre de projections approprié.

Choix d'une valeur

Dans notre contexte de projection de 3 des 4 symboles des sous-séquences de base, sur 3 de leurs 4 dimensions, on choisit alors assez largement de réaliser 100 projections aléatoires pour la construction d'une matrice de collisions. Le tableau 6.3 illustre ce choix par la comparaison des pourcentages de collisions moyens observés pour les séquences de chaque classe et pour différents nombres de projections allant de 50 à 200. Les résultats présentés sont relativement similaires quel que soit le nombre de projections, justifiant *a priori* la pertinence de chacun d'entre eux.
D'autres paramètres clés de la fouille de données influencent très fortement les performances du système, mais ne peuvent pas être déterminés si simplement. Leur détermination dépend en effet des contraintes de décision, des caractéristiques des séquences analysées et en particulier du taux de bruit dans les données.

Nombre de projections	% Moyen de collisions	
	Classe 0	Classe 1
50	44.2	7.8
100	**46.1**	**8.5**
150	44.1	8.5
200	44.1	7.7

TAB. 6.3 – Pourcentages moyens de collisions observés entre les séquences des classes 0 et 1 et la séquence de référence, selon le nombre de projections réalisées.
Le tableau présente les résultats obtenus avec une dimension de projection de 4×4 vers 3×3.

Seuil minimum de collisions

Définition

Ce seuil fixe la borne inférieure des nombres de collisions considérés significatifs, c'est-à-dire représentatifs d'une "vraie" similarité entre les sous-séquences de base correspondantes. Ce seuil ne doit pas être trop élevé pour ne pas manquer des instances de motifs dès cette étape d'analyse. On préfère sélectionner trop de sous-séquences – éliminées aux étapes suivantes du processus de décision – que de manquer dès l'examen des collisions certaines instances de motifs. Il faut cependant veiller à ne pas perdre tout l'intérêt des projections – baisser ce seuil à l'extrême revient à étudier la similarité deux à deux de toutes les sous-séquences possibles d'un nombre donné de symboles.

Choix d'une valeur

Dans notre contexte, la figure 6.9 illustre le choix du seuil minimum de collisions. On compare les résultats obtenus avec différentes dimensions de projection pour montrer qu'une diminution de dimension ne résout pas le problème d'adaptation à d'importants taux de bruits dans les séquences comparées : les observations réalisées sont en effet complètement similaires dans leurs tendances relatives. Par exemple, les sous-séquences qui posent problème sont les mêmes, dans les mêmes positions relatives.

Tout d'abord, l'exemple proposé met en évidence la nécessité d'une seconde étape de décision pour affiner l'identification des sous-séquences récurrentes après l'examen de la matrice de collisions. Dans ce contexte particulièrement bruité, aucun seuil de collisions ne permet en effet de classer correctement les séquences expérimentales. Suivant le "meilleur seuil", les séquences de la classe 0 qui sont mal classées ou à la limite de ce seuil sont aussi particulièrement bruitées, incluant soit une combinaison de différents types de modifications "normales" (séquences 5 et 6), soit un taux de bruit dans les valeurs particulièrement élevé (séquence 8). La séquence 17 est par contre la seule mal classée de la classe 1. Elle correspond à la séquence de référence bruitée uniquement par de trop longues interruptions, d'où le fort nombre de collisions obligatoirement observé pour les sous-séquences identiques par ailleurs. Dans tous ces cas, la définition d'un seuil maximum sur la mesure de distance réelle doit alors affiner la classification.

Le seuil minimum de collisions doit ainsi *idéalement* correspondre au plus faible nombre de collisions observé entre des séquences similaires, en fonction du taux de bruit possible. Dans l'exemple proposé, si on considère que le cas de la séquence 8 est assez extrême (taux de bruit égal à 1), un seuil minimum de collisions de l'ordre de 0 à 5% semble alors approprié à un contexte bruité. La figure 6.10 confirme qu'en moyenne, sur un grand nombre d'expérimentations, un seuil même très faible permet de réduire considérablement le nombre de sous-séquences de

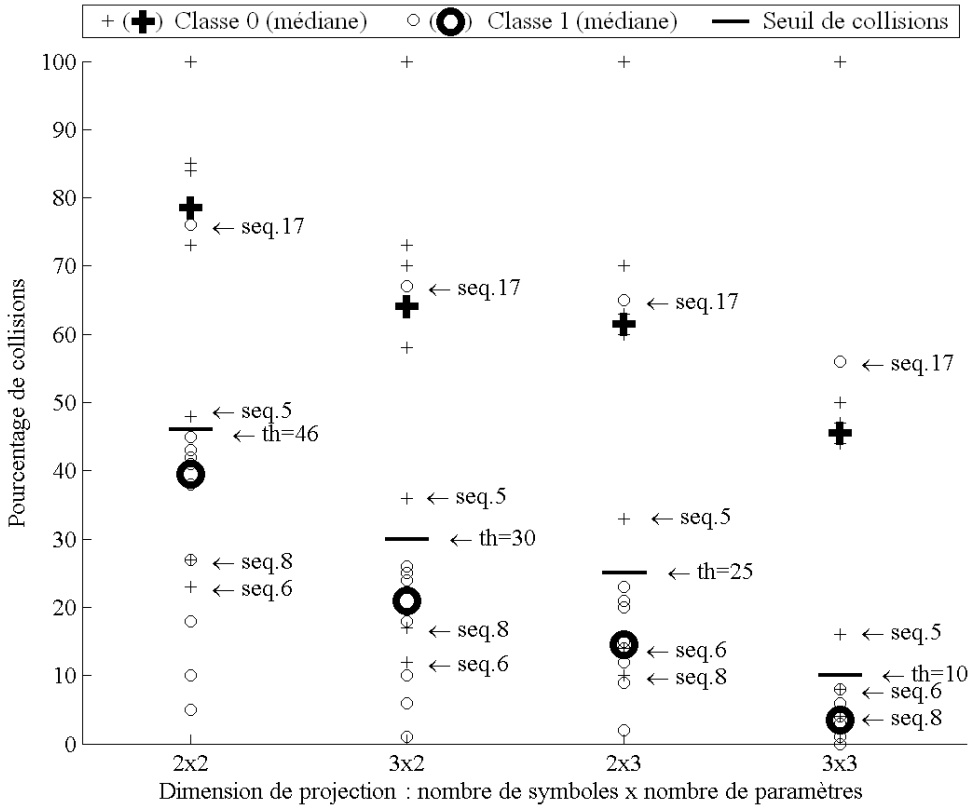


FIG. 6.9 – Sélection du "meilleur" seuil minimum de collisions en fonction des collisions observées pour les séquences des classes 0 et 1 avec la séquence de référence, et de la dimension de projection. Le graphe présente le pourcentage maximum de collisions obtenu entre les sous-séquences de base de la séquence de référence d'une part (séquence 0), et celles des différentes séquences des classes 0 (séquences 1 à 8) et 1 (séquences 9 à 18) d'autre part. On positionne par ailleurs le "meilleur" seuil de collisions pour la classification de ces séquences (th), en indiquant les séquences dont la classification est alors erronée. Enfin, ces résultats ont été obtenus par la réalisation de 100 projections, et sont présentés pour différentes dimensions de projection.

base à comparer "réellement". Si à la limite on ne considère que les couples de sous-séquences correspondant à un nombre de collisions strictement positif, on restreint déjà l'espace de recherche des sous-séquences récurrentes à 6% de l'espace initial ; avec un seuil à 5%, ces sous-séquences sont recherchées dans à peine plus de 2% de l'espace complet, et dans 1% de cet espace à 10%.

L'hypothèse qu'une sous-séquence récurrente apparaît plusieurs fois dans une séquence analysée permet cependant probablement d'augmenter ce seuil afin de gagner en spécificité tout en conservant une bonne sensibilité de l'identification et de la classification des sous-séquences similaires. En effet, dans le processus d'extraction de motifs, si la similarité de deux instances est manquée au niveau de l'observation des collisions mais que celle de chacune d'entre elles avec une troisième instance ne l'est par contre pas, toutes les tentatives de motifs seront néanmoins identifiées. Une mesure de distance appropriée permet alors de regrouper les trois instances au moment de la classification. D'après les résultats présentés sur la figure 6.9, on peut envisager l'expérimentation de valeurs du seuil minimum de collisions comprises environ entre 0 et 40 % du nombre de projections réalisées.

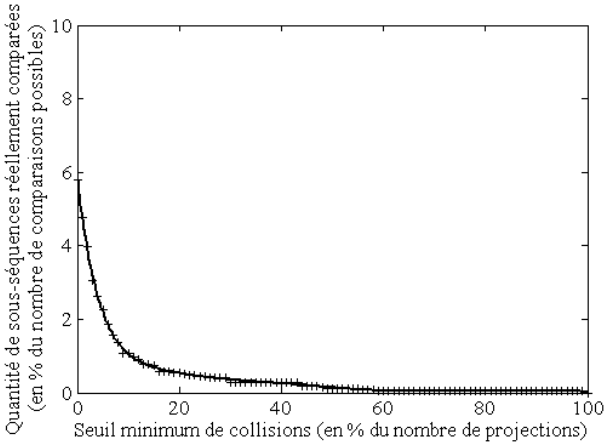


FIG. 6.10 – Réduction de la dimension de l'espace de recherche de sous-séquences récurrentes par la fouille de caractères.
Le graphe présente le pourcentage du nombre de couples de sous-séquences de bases possibles qui sont comparés effectivement à l'issue de l'examen de la matrice de collisions. Les sous-séquences sélectionnées doivent correspondre à un nombre de collisions *strictement supérieur* au seuil proposé.

Seuil maximum de distance

Définition

Ce seuil fixe la borne supérieure sur la distance réelle entre deux sous-séquences pour les considérer similaires, c'est-à-dire éventuellement représentatives d'un même motif. Il y a un compromis à trouver dans la détermination de ce seuil afin de considérer comme effectivement similaires toutes les sous-séquences représentatives d'un même motif, même en présence de bruit (sensibilité), sans inclure de sous-séquences relativement proches mais non représentatives de ce même motif (spécificité).

Choix d'une valeur

D'après l'analyse des résultats présentés sur la figure 6.6, un seuil maximum de distance compris environ entre 0.2 et 0.6 semble approprié si la valeur de ϵ_{LCSS} est environ entre 0.2 et 0.6 elle aussi, et dans le cas où $\delta_{LCSS} = 40$. Une faible valeur de ϵ_{LCSS} contraint fortement la similarité de deux points et induit d'importantes valeurs de distance, si bien que le seuil de distance doit être plus élevé, et réciproquement.

Dans le contexte où $\epsilon_{LCSS} = 0.3$, le seuil maximum de distance peut être compris environ entre 0.25 et 0.45 si la valeur de δ_{LCSS} varie entre 35 et 65. Les intervalles de valeurs possibles sont ainsi assez larges, et les influences relatives de ces paramètres rendent par ailleurs plus complexe le choix des "meilleures" valeurs. Il est ainsi nécessaire d'étudier les performances du système pour différentes valeurs de ces paramètres afin de mieux appréhender les critères de leur sélection.

Finalement, il semble approprié d'expérimenter des valeurs du seuil maximum de distance entre 0.2 et 0.6.

Les seuils de collisions et de distance sont ainsi des paramètres clés du système, même si la détermination d'un seuil de collisions approprié apparaît moins critique et difficile que celle du seuil de distance. C'est en effet ce dernier qui permet finalement d'identifier les sous-séquences récurrentes.

6.2.4 Synthèse sur la qualité de la méthode

Cette étude préliminaire de la qualité de la méthode permet d'évaluer la pertinence des différentes étapes proposées pour l'extraction de motifs, et de mettre en évidence les valeurs possibles des paramètres du système.

D'après les paragraphes précédents, les quatre paramètres clés particulièrement influents sur les performances du système sont les **différents seuils** impliqués d'une part pour la définition de la mesure de distance réelle entre séquences temporelles – ϵ_{LCSS} et δ_{LCSS} – et d'autre part pour la sélection des tentatives de motifs et leur classification en motifs – seuils minimum de collisions et maximum de distance. Les autres paramètres dits *méthodologiques* peuvent être déterminés *a priori*, si bien qu'on fixe dans ces expérimentations d'emblée leur valeur.

On définit ensuite les intervalles de variation possible des paramètres clés du système. Les valeurs proposées à l'issue de l'étude sur la qualité de la méthode ont été définies dans un contexte particulièrement bruité. Les intervalles de variation proposés ne sont néanmoins qu'*indicatifs*, issus de l'observation de quelques exemples, et il convient de les considérer un peu plus largement pour l'étude de la qualité des résultats. En particulier, une différence notable dans le contexte expérimental d'extraction de motifs par rapport à ces études préliminaires est liée à la localisation des sous-séquences récurrentes dans la séquence initiale. Lors du processus d'extraction, la comparaison des séquences n'est pas forcément réalisée suivant la localisation exacte des sous-séquences récurrentes dans le temps.

Cette dernière caractéristique nécessite sûrement une définition plus souple en particulier de l'écart maximum autorisé dans le temps entre deux points similaires – δ_{LCSS} – pour atteindre une bonne sensibilité du système, tout en essayant de préserver sa spécificité. Par contre, le seuil minimum de collisions peut être envisagé au contraire plus restrictif, c'est-à-dire plus élevé. En effet, si deux instances d'un même motif présentent une grande variabilité l'une par rapport à l'autre en terme du nombre de collisions, on peut supposer qu'elles ne diffèrent pas autant de toutes les instances de ce motif et qu'elles pourront néanmoins être identifiées comme tentatives de motifs. La bonne adéquation du seuil de distance permet ensuite de bien les regrouper. Ainsi on gagne à la fois en sensibilité et en spécificité pour la classification.

On a également mis en évidence dans cette section l'influence relative de plusieurs de ces paramètres, justifiant la complexité de leur définition, et en particulier :

- d_{max} et δ_{LCSS},
- d_{max} et ϵ_{LCSS}.

Il ne semble par ailleurs pas approprié d'exploiter l'influence relatives des paramètres δ_{LCSS} et ϵ_{LCSS}. Ces deux paramètres doivent en effet répondre spécifiquement à certains types de bruit : respectivement, les déformations temporelles et interruptions d'une part, et la variabilité dans les valeurs d'autre part. Il convient alors plutôt de les définir indépendamment, en fonction des différents types et taux de bruit possibles dans les données observées.

Le tableau 6.4 résume finalement les valeurs ou intervalles de variation définis *a priori* pour les différents paramètres impliqués dans la méthode proposée pour l'extraction non supervisée de motifs temporels. Les astérisques signalent les paramètres considérés comme particulièrement influents et critiques sur les performances du système, et dont une valeur appropriée ne peut pas être déterminée facilement *a priori*.

Paramètre	**Valeurs**	**Définition**
Mesure de similarité		
δ_{LCSS}*	[35 65]	Écart temporel maximum entre deux points similaires
ϵ_{LCSS}*	[0.2 0.6]	Écart maximum sur les valeurs de deux points similaires
Abstraction		
Filtre moyen		Moyenne mobile sur une fenêtre pondérée, symétrique, sans réduction temporelle *a priori*
n_{mod}	4	Nombre d'intervalles de discrétisation des paramètres quantitatifs
dd_{max}	0	Seuil maximum de distance minimum sur les séquences discrètes pour l'agrégation
Projections aléatoires		
w_{symb}	4	Nombre de symboles
w_{proj}	3	Nombre de symboles projetés
p_{proj}	3	Nombre de paramètres projetés par symbole
n_{proj}	100	Nombre de projections
Examen des collisions		
c_{min}*	[0.0 0.4]	Seuil minimum de collisions
d_{max}*	[0.2 0.6]	Seuil maximum de distance

Tab. 6.4 – Paramètres de la méthode proposée pour l'extraction de motifs.

6.3 Qualité des résultats

L'évaluation de la qualité des résultats obtenus est réalisée en expérimentant la méthode proposée pour la recherche de motifs sur des séquences pour lesquelles on connaît *a priori* les caractéristiques des motifs et la localisation temporelle de leurs instances. Les performances du système sont alors évaluées en terme des sensibilité et spécificité de l'identification des tentatives de motifs d'une part, et de leur classification en motifs d'autre part (voir paragraphe 6.1.4). On s'intéresse en particulier au contexte de la présence d'une grande quantité de bruit entre les instances d'un motif. L'objectif est ainsi d'évaluer l'efficacité et la résistance au bruit du système proposé, et de poser ses limites.
Les expérimentations sont réalisées en plusieurs étapes :

- **6.3.2. Évaluation des performances en fonction des paramètres clés du système**, dans un contexte faiblement bruité, et pour l'analyse de séquences pour lesquelles on connaît *a priori* les motifs qu'elles contiennent : on observe en particulier l'évolution des indices de sensibilité et spécificité en fonction des paramètres clés du système, et leurs influences relatives sur les performances.
- **6.3.3. Test de Sensibilité.** Évaluation de la résistance du système au bruit présent entre les instances d'un motif, c'est-à-dire aux modifications "normales" de comportement. On observe en particulier l'évolution des indices de sensibilité et spécificité en fonction des taux de bruit introduits entre les instances de motifs, dans une configuration donnée du système.
- **6.3.4. Test de Spécificité.** Éstimation des performances du système en présence de modifications inquiétantes de comportement. On observe en particulier sa capacité à ne pas détecter les instances "dégradées" d'un motif.

On rappelle d'abord en **6.3.1** les **exigences et paramètres clés du système**.

6.3.1 Exigences et paramètres clés du système

Exigence de performance des étapes de l'extraction de motifs

En considération de la méthode d'identification des sous-séquences récurrentes dans une séquence initiale – en terme d'une succession de symboles qui les représente – la localisation de ces sous-séquences ne peut pas être très précise, ce qui est par ailleurs en adéquation avec les objectifs de décision : identifier la présence de comportements récurrents plus que la délimitation précise de leurs instants d'occurrence. Ceci implique une assez large tolérance sur les valeurs de sensibilité et de spécificité relatives à l'identification des tentatives de motifs, ces indices étant quant à eux calculés assez précisément. On considère en effet la comparaison point à point des tentatives de motifs identifiées avec les véritables délimitations des instances de motifs.

Les exigences sont par contre importantes pour ce qui concerne la classification de ces sous-séquences en motifs. La construction d'un profil comportemental fiable nécessite en effet de reconnaître tous les comportements récurrents et de savoir les regrouper de façon significative, même si on ne sait pas identifier précisément leurs instants d'occurrence. On veille ainsi particulièrement au maintien de fortes valeurs de sensibilité et de spécificité de la classification. Idéalement, on souhaite ces indices à leur valeur maximum, égale à 1.

Ces exigences de la classification nécessitent indirectement des indices corrects de performance pour la localisation des sous-séquences récurrentes. En effet, une identification trop "grossière" de ces sous-séquences entraîne probablement de plus grandes distances calculées entre sous-séquences correspondant pourtant aux instances d'un même motif : pas assez de points

effectivement représentatifs de l'instance d'un motif et trop de points correspondant à des "non motifs" sont considérés dans les différentes sous-séquences comparées, ce qui augmente les mesures de distances. Leur classification risque alors d'être mauvaise à cause d'un dépassement plus probable du seuil maximum de distance.

Finalement, on privilégie de bonnes performances de la classification, particulièrement en terme de spécificité : on ne souhaite caractériser que des comportements effectivement récurrents. Cet objectif semble néanmoins nécessiter une bonne sensibilité de l'identification des sous-séquences récurrentes.

Dans les expérimentations suivantes, les résultats sont présentés sous forme de *courbes COR* (*Caractéristique Opérationnelle du Receveur*), appelées encore courbes *ROC* en anglais (*Receiver Operating Characteristic*). La méthode COR, proposée par Provost et Fawcett [89], exploite la relation entre les taux de bonnes détections et de fausses alarmes pour évaluer les performances d'un système de décision. Selon les expérimentations, on représente ainsi les couples des indices de performance (*sensibilité*, 1−*spécificité*) – c'est-à-dire, les taux de bonnes détection et de fausses alarmes – obtenus pour différentes valeurs des paramètres ou taux de bruits entre les instances des motifs.

Paramètres clés des expérimentations

Les expérimentations sont d'abord réalisées simplement sur des séquences de données correspondant à 2 journées d'enregistrement, à la fréquence d'une donnée toutes les minutes, et dans lesquelles on insère 6 instances d'un seul motif. Pour chaque test de l'expérimentation, les motifs insérés sont sélectionnés aléatoirement dans les séquences générées par le processus de simulation et représentatives d'une personne à domicile. On réalise ainsi plusieurs tests, avec pour chacun l'expérimentation de plusieurs valeurs des paramètres du système et/ ou de plusieurs taux de bruit entre les instances des motifs, et on présente finalement les performances moyennes obtenues.

D'après les paragraphes précédents, les quatre paramètres clés, peu contraints, et particulièrement influents sur les performances du système sont les **différents seuils** impliqués d'une part pour la définition de la mesure de distance réelle entre séquences temporelles, et d'autre part pour la sélection des tentatives de motifs et leur classification en motifs.

- δ_{LCSS}, Écart temporel maximum entre deux points similaires ;
- ϵ_{LCSS}, Écart maximum sur les valeurs de deux points similaires ;
- c_{min}, Seuil minimum de collisions ;
- d_{max}, Seuil maximum de distance.

6.3.2 Performances et paramètres clés du système

Dans l'objectif de déterminer le "meilleur" ensemble de paramètres à considérer, les premières études concernent l'influence de leurs valeurs sur les performances du système. On vérifie d'abord la pertinence des résultats pour l'extraction de motifs dans un contexte non bruité, afin de s'assurer que la méthode proposée peut réellement être efficace. On ajoute ensuite une quantité "raisonnable" de bruit entre les instances de motifs pour étudier plus précisément l'influence des paramètres clés du système sur ses performances. On étudie enfin l'influence relative de certains paramètres pour préciser une configuration appropriée du système.

On utilise pour ces expérimentations une configuration par défaut du processus d'extraction de motifs, définie *a priori*, et on fait varier successivement chaque paramètre clé : les seuils minimum de collisions (c_{min}) et maximum de distance (d_{max}), et les écarts maximums de similarité,

sur les valeurs (ϵ_{LCSS}) et dans le temps (δ_{LCSS}). En fonction des résultats de l'analyse de la qualité de la méthode, on propose des valeurs par défaut qui semblent *a priori* appropriées à un contexte plus ou moins bruité. Le tableau 6.5 présente ces valeurs par défaut et les intervalles de variation sélectionnés pour chacun de ces paramètres afin d'évaluer leur influence sur les performances. Les autres paramètres du système sont positionnés aux valeurs estimées *a priori* appropriées au contexte expérimental d'après l'analyse de la qualité de la méthode.

- **Le seuil minimum de collisions** (c_{min}) est défini de façon sensiblement restrictive d'après les remarques de cette section, pour atteindre une bonne spécificité de l'identification des sous-séquences récurrentes et selon l'hypothèse d'une bonne sensibilité malgré tout du fait de la similarité de chaque sous-séquence avec plusieurs autres. On choisit ainsi arbitrairement **0.2**.
- **Les seuils de distance et de similarité sur les valeurs** (d_{max} et ϵ_{LCSS}) sont définis aux valeurs maximum admises dans la perspective de résistance à un contexte bruité, soit **0.6**. D'après les études précédentes, le seuil de distance doit cependant diminuer si le seuil de similarité sur les valeurs augmente, et réciproquement. Ce couple de valeurs n'est pas conséquent pas forcément le plus adapté, et ces premières expérimentations devraient permettre d'évaluer la pertinence de ce choix et l'adaptation éventuelle des valeurs.
- **L'écart temporel de similarité** (δ_{LCSS}) est défini légèrement supérieur à la borne supérieure (65) pour donner de la souplesse dans la précision de l'identification des sous-séquences récurrentes dans la séquence initiale : on choisit arbitrairement une valeur égale à **90** minutes.

Ces valeurs *a priori* ne sont ainsi en rien limitatives pour la définition de paramètres effectivement appropriés à notre contexte. Elles permettent de disposer d'un ensemble de paramètres par défaut comme base de l'expérimentation de leur influence sur les performances du système. L'objectif n'est ainsi pas à ce niveau de définir le meilleur ensemble de paramètres mais d'identifier pour chacun une valeur "réaliste". Ces premières expérimentations de la qualité des résultats doivent justement permettre de préciser un réglage adapté au contexte d'étude.

Paramètre	**Valeur par défaut**	**Intervalle de variation**
c_{min}	0.2	[0.0 0.6]
d_{max}	0.6	[0.2 0.8]
δ_{LCSS}	90	[30 120]
ϵ_{LCSS}	0.6	[0.2 0.8]

TAB. 6.5 – Valeurs par défaut et variation des paramètres de réglage pour l'observation de leur influence sur les performances du système.

Les expérimentations réalisées dans ce contexte donnent globalement de bons résultats en terme de sensibilité et de spécificité de l'identification et de la classification des sous-séquences récurrentes. On présente successivement l'évolution des performances du système en contexte non bruité, puis faiblement perturbé, en fonction de différentes valeurs des paramètres. On étudie ensuite les variations relatives des différents paramètres pour définir enfin un ensemble de valeurs *a priori* approprié pour les paramètres clés du système.

Efficacité du système en contexte non bruité

La figure 6.11 présente les quatre graphes correspondant aux courbes COR relatives à l'évolution des performances du système – identification des tentatives de motifs et classification en motifs –

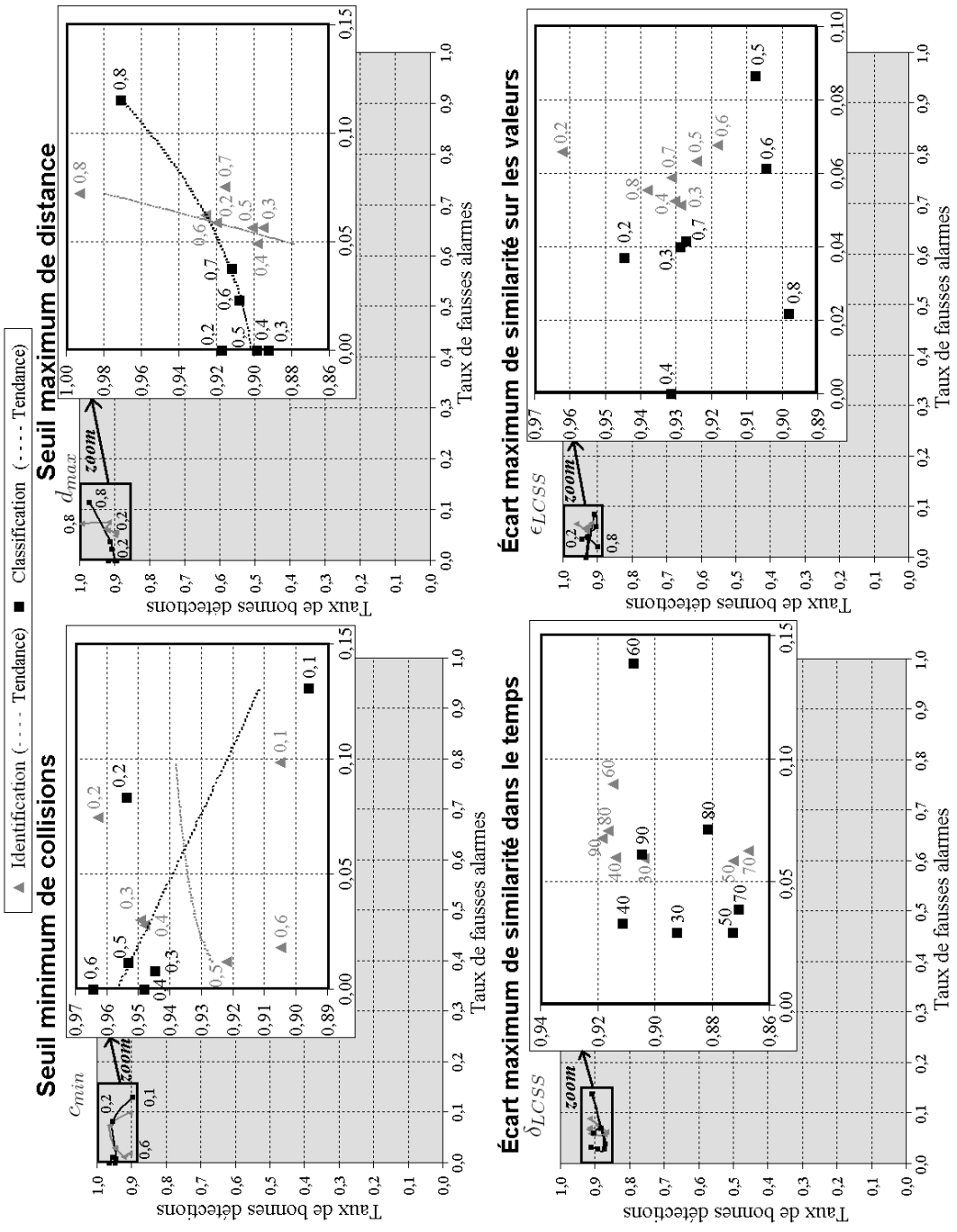


FIG. 6.11 – Performances de l'identification des tentatives de motifs et de leur classification en motifs dans un contexte non bruité.
Chaque graphe correspond à l'étude d'un des paramètres clés du système. Il présente les performances moyennes de l'identification des tentatives de motifs (triangles) et de leur classifications en motifs (carrés). Les étiquettes de données indiquent pour chaque point du graphe la valeur associée du paramètre étudié, les autres paramètres étant positionnés aux valeurs par défaut. Les sous-graphes correspondent à un zoom sur la zone de variation des indices de performance.

selon les valeurs de chaque paramètre clé dans un contexte non bruité. De manière générale, on constate de très bonnes performances des deux étapes d'extraction et de classification des motifs. Les couples des taux de bonnes détections (sensibilité) et de fausses alarmes (1−spécificité) correspondent en effet à des points situés dans le coin en haut à gauche de chaque graphe, c'est-à-dire qu'ils sont respectivement proches de 1 et 0. Par conséquent, les variations observées pour les indices de performance en fonction des valeurs de chaque paramètre ne sont pas forcément significatives, puisque limitées à un espace assez restreint de bonnes valeurs de performances. Par ailleurs, les indices de fractionnement observés sont quasiment toujours maximum, égaux à 1, signifiant que les instances de motifs sont identifiées pour chacune par une seule tentative de motifs, sans fractionnement.

Globalement, on constate ainsi que les valeurs des paramètres n'ont finalement pas beaucoup d'influence sur les performances du système qui restent en toutes circonstances très bonnes. Dans un contexte non bruité, l'examen des collisions permet en effet presque à lui seul de réaliser de correctes identification et classification des sous-séquences récurrentes tant que le seuil minimum de collisions est suffisamment élevé : le seuil le plus faible (0.1) donne les moins bonnes performances. Les sous-séquences identifiées et étendues d'après un critère de collisions suffisamment restrictif ne correspondent alors quasiment qu'à des sous-séquences effectivement similaires, d'où la faible incidence des seuils de distance et de similarité sur les performances.

Une analyse plus fine des courbes de la figure 6.11 (voir les zooms proposés sur cette figure) permet cependant d'identifier peut-être des tendances de variation des performances, particulièrement en fonction des seuils minimum de collision et maximum de distance. Ces hypothèses sont vérifiées dans un contexte d'expérimentation faiblement bruité qui les met bien mieux en évidence, et ainsi détaillées dans le paragraphe suivant.

Influence des paramètres clés du système sur ses performances

Puisque le système d'extraction de motifs donne de bonnes performances pour l'extraction de motifs en contexte non bruité, on étudie maintenant la qualité des résultats lorsqu'on introduit entre les instances de chaque motif une quantité "raisonnable" de bruit, selon les indices suivants (voir annexe F.2.2 pour le détail de l'introduction de ces bruits) :

- Taux de bruit dans les valeurs : 0.2 ;
- Taux de déformation temporelle : 0.1 ;
- Taux d'interruption : 0.05.

Les paramètres clés du système sont positionnés aux valeurs par défaut définies *a priori*, puis on fait varier successivement les valeurs de chacun d'eux. La figure 6.12 présente les performances du système obtenues dans ce contexte, et la figure 6.13 l'évolution des indices de fractionnement des instances de motifs reconnues. Même si on parvient dans certaines configurations à des indices de performance corrects, les résultats obtenus sont globalement beaucoup moins bons que dans un contexte non bruité. Concernant l'évolution des indices de fractionnement, les tendances éventuellement identifiées s'expliquent difficilement et il est probable qu'elles ne soient pas significatives. Plus généralement, on constate que les indices restent assez proches de 1, c'est-à-dire que les instances de motifs sont le plus souvent reconnues dans leur intégralité, chacune en une seule tentative de motif.

L'idée est alors d'étudier chaque paramètre à partir des résultats de la figure 6.12 afin d'évaluer s'il est possible d'identifier une configuration du système qui contribue aux "meilleures" performances en particulier de la classification, dernière étape particulièrement critique. Les influences générales des paramètres sur les performances du système supposées d'après l'analyse en contexte non bruité sont confirmées et affinées dans cette étude. Une analyse de l'influence relative des

paramètres permet ensuite d'identifier une configuration par défaut qui semble appropriée pour le système.

Seuil minimum de collisions (c_{min})

Un seuil minimum de collisions croissant – c'est-à-dire progressivement plus restrictif – permet d'augmenter la spécificité mais diminue la sensibilité de l'identification des tentatives de motifs. La diminution de sensibilité s'explique par la comparaison, dans le contexte d'un seuil minimum de collisions élevé, d'un nombre restreint de sous-séquences susceptibles d'être récurrentes, associée à une moins large extension potentielle de ces sous-séquences. Par ailleurs, l'augmentation de la spécificité est due à la considération de moins de sous-séquences comme susceptibles d'être récurrentes, incluant par conséquent moins de sous-séquences qui pourraient l'être "par hasard". Les seuils importants de distance et de similarité sur les valeurs sélectionnés pour cette expérimentation contribuent probablement à la sensibilité de ce paramètre vis-à-vis des performances du système. Des valeurs plus restrictives du seuil de distance notamment doivent pouvoir diminuer son influence.

Au niveau de la classification, l'augmentation du seuil minimum de collisions permet globalement une meilleure classification, tant en terme des indices de sensibilité que de spécificité. L'interprétation *a priori* étonnante d'une meilleure sensibilité est due à la récurrence des sous-séquences représentatives des motifs. Une instance pas ou mal reconnue comme similaire à une autre par la contrainte d'un trop grand nombre de collisions ne l'est pas forcément avec toutes les autres instances de ce motif. Par conséquent, même si la localisation des tentatives de motif n'est pas très précise et probablement "sous-étendue" par rapport aux instances effectives du motif, on parvient néanmoins à identifier la plupart des occurrences. Par ailleurs, une augmentation de la contrainte sur le nombre de collisions résulte forcément en une meilleure spécificité.

D'après les résultats observés, un seuil minimum de collisions défini autour de 0.3 semble pouvoir donner des performances appropriées à notre contexte : les tentatives de motifs sont globalement "sous-étendues" mais malgré tout relativement bien identifiées et correctement classées. L'analyse des performances avec des **seuils de similarité plus restrictifs** pour la définition des mesures de distance (dans cette expérimentation, $\epsilon_{LCSS} = 0.6$ et $\delta_{LCSS} = 90$) doit par ailleurs permettre d'évaluer la possibilité de *diminuer légèrement* ce seuil minimum de collisions. L'objectif est d'augmenter un peu la précision de l'identification des tentatives de motifs, et peut-être par conséquent les performances de la classification. Un **seuil maximum de distance plus restrictif** (dans cette expérimentation, $d_{max} = 0.6$) agit probablement également dans ce sens.

Ainsi, un **seuil minimum de collisions** efficace semble devoir se situer approximativement **entre 0.1 et 0.3** avec des valeurs légèrement plus restrictives des autres paramètres par rapport à la configuration par défaut.

Seuil maximum de distance (d_{max})

Pour ce qui concerne l'identification des tentatives de motifs, on observe qu'une augmentation du seuil maximum de distance induit une meilleure sensibilité en même temps qu'une moins bonne spécificité. La contrainte de similarité entre deux séquences est en effet plus souple, et autorise en conséquence également le rapprochement de sous-séquences qui ne doivent pas être considérées similaires au regard des objectifs de décision.

Au niveau de la classification, si on considère l'expérimentation de seuils compris entre 0.4 et 0.8, on constate globalement une amélioration de la sensibilité et une diminution de la spécificité avec un seuil maximum de distance croissant. On considère en effet progressivement plus de sous-séquences comme effectivement similaires, et par conséquent on associe bien les sous-séquences effectivement représentatives d'un même motif en même temps que d'autres qui leur ressemblent

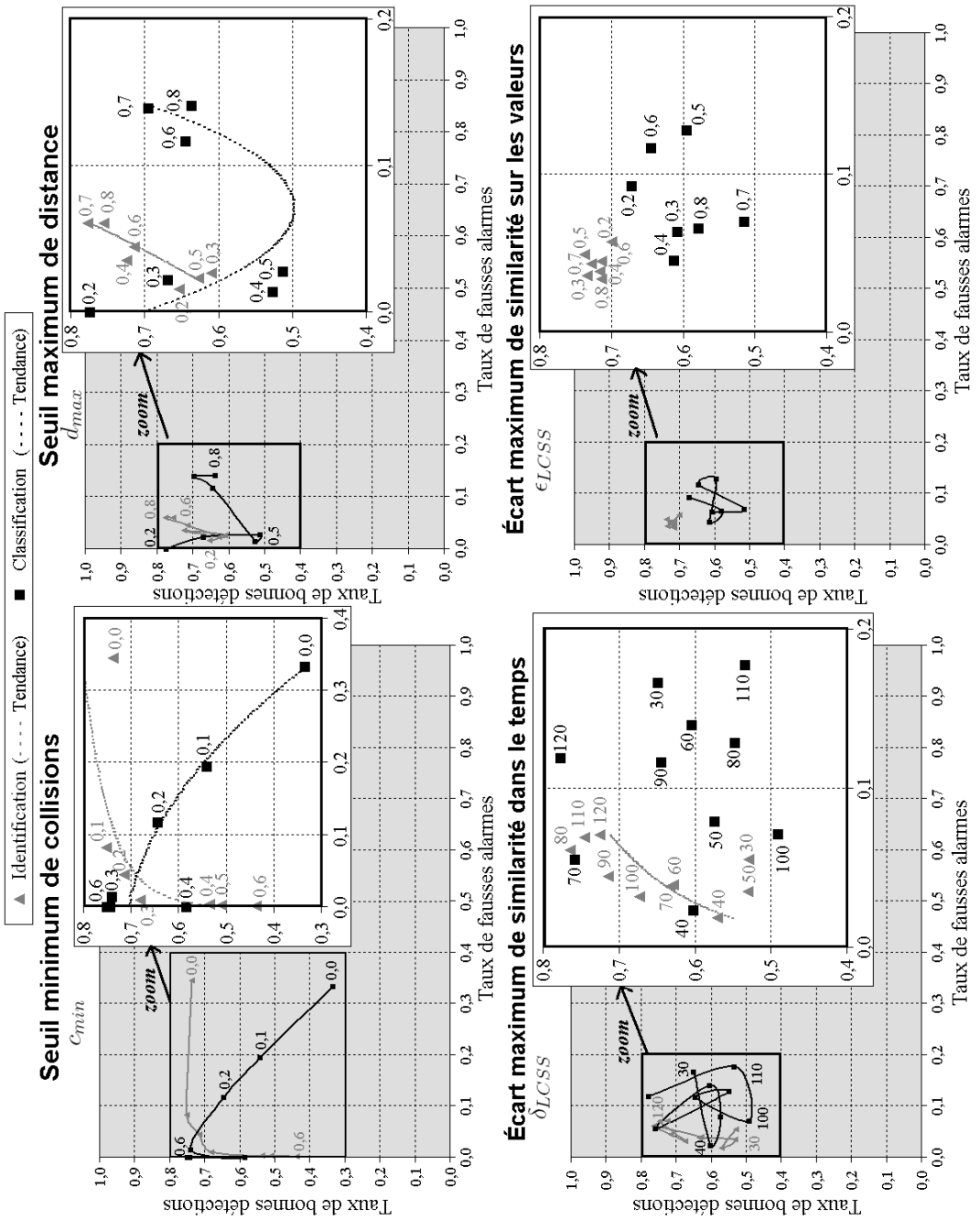


FIG. 6.12 – Performances de l'identification des tentatives de motifs et de leur classification en motifs dans un contexte faiblement bruité.
On introduit entre les instances de chaque motif une quantité "raisonnable" de bruit selon des taux de 0.2 pour les valeurs, 0.1 pour les déformations dans le temps, et 0.05 pour les interruptions.
Chaque graphe correspond à l'étude d'un des paramètres clés du système. Il présente les performances moyennes de l'identification des tentatives de motifs (triangles) et de leur classifications en motifs (carrés). Les étiquettes de données indiquent pour chaque point du graphe la valeur associée du paramètre étudié, les autres paramètres étant positionnés aux valeurs par défaut. Les sous-graphes correspondent à un zoom sur la zone de variation des indices de performance.

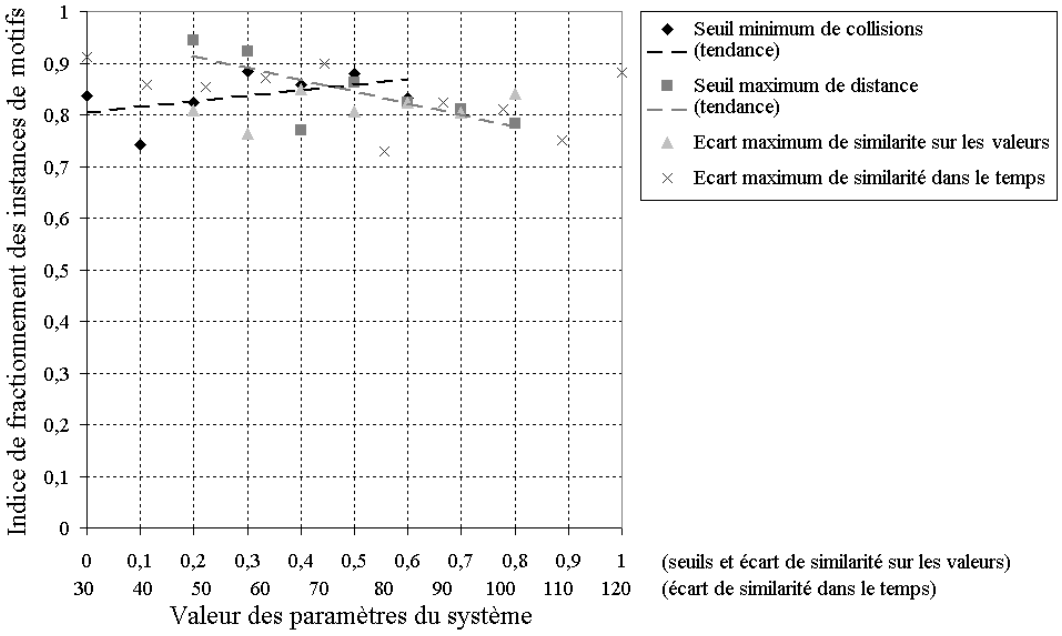


FIG. 6.13 – Indices de fractionnement des instances de motifs dans un contexte faiblement bruité, en fonction des paramètres clés du système.
Un indice maximum correspond à la reconnaissance intégrale des instances de motifs, sans fractionnement en plusieurs tentatives de motifs associés alors à une ou plusieurs classes.

"par hasard".
Un seuil maximum de distance plus restrictif (entre 0.2 et 0.3) permet sur ce principe d'augmenter encore la spécificité, tout en augmentant également la sensibilité de la classification. Une interprétation est liée à la précision de l'identification des tentatives de motifs qui ne prennent alors en compte quasiment que des points qui appartiennent effectivement aux instances de motifs. Les distances entre ces sous-séquences deux à deux sont ainsi relativement faibles et leur classification selon un seuil maximum de distance faible est efficace en terme des indices de sensibilité et de spécificité.
Avec la considération d'un seuil supérieur de distance, l'identification des tentatives de motifs inclut probablement plus de points qui ne correspondent pas systématiquement aux instances de motifs. Les distances calculées sont alors plus élevées, et les sous-séquences représentatives d'un même motif probablement regroupées en plusieurs classes, dont certaines ne sont peut-être pas significatives au regard de la décision en terme de leur effectif. Étant donnée une contrainte sur la dimension d'une classe, certaines sous-séquences ne sont alors plus considérées comme effectivement représentatives de l'instance d'un motif, ce qui diminue l'indice de sensibilité. Pour augmenter la sensibilité, on peut alors augmenter encore le seuil de distance, ce qui a cependant pour conséquence de diminuer en même temps la spécificité.

On identifie ainsi comme un "puits des indices de performance" de la classification pour des valeurs intermédiaires du seuil maximum de distance. On s'oriente ainsi plutôt vers une valeur assez restrictive de ce seuil, approximativement entre 0.2 et 0.3. Cependant, une **valeur plus restrictive de la contrainte de similarité sur les valeurs** (dans cette expérimentation, $\epsilon_{LCSS} = 0.6$) doit peut-être permettre d'*augmenter un peu ce seuil* pour une meilleure résistance au bruit sans grande influence sur la spécificité.

Ainsi, un **seuil maximum de distance** efficace semble devoir se situer approximativement **entre 0.2 et 0.5**, avec des valeurs légèrement plus restrictives des autres paramètres par rapport à la configuration par défaut.

Écarts maximums de similarité sur les valeurs (ϵ_{LCSS}) et dans le temps (δ_{LCSS})
La définition des écarts maximum de similarité sur les valeurs et dans le temps utilisés pour les calculs de distance n'a pas de conséquence évidente sur les performances dans un contexte où les paramètres par défaut du système ont été choisis bien peu restrictifs. Les performances du système ne sont malgré tout pas catastrophiques à cause de la contrainte de similarité sur les paramètres qualitatifs dans la mesure de distance. Celle-ci impose une borne supérieure au nombre de points similaires entre deux sous-séquences quels que soient les contraintes de similarité sur les paramètres quantitatifs (ϵ_{LCSS} et δ_{LCSS}). Les distances restent ainsi malgré tout relativement élevées entre des sous-séquences non similaires. Sur la figure 6.6 de l'analyse de la qualité de la méthode, on constate en effet qu'*a priori* la distance entre deux sous-séquences différentes est en moyenne supérieure à 0.6 quelles que soient les valeurs des paramètres de cette mesure.

Une telle situation n'est pas du tout satisfaisante puisqu'elle tend à diminuer voire, à l'extrême, à supprimer la prise en compte des variations des paramètres quantitatifs. Dans ce contexte, une dégradation du comportement de ces paramètres ne sera jamais détectée. Cette configuration du système n'est ainsi pas justifiée même en prévision de l'adaptation à un contexte particulièrement bruité. Afin d'assurer la prise en compte des paramètres quantitatifs et la génération de faibles distances uniquement entre sous-séquences effectivement similaires au regard de l'ensemble des paramètres, il est finalement indispensable de définir des écarts maximum de similarité plus restrictifs (ϵ_{LCSS} et δ_{LCSS}). Ces valeurs de paramètres doivent êtres associées à un seuil maximum de distance (d_{max}) relativement faible mais adapté aux mesures de distances globalement plus élevées dans le contexte de seuils de similarité plus contraignants. Le seuil minimum de collisions (c_{min}) peut également être alors défini assez faible pour améliorer l'identification des tentatives de motifs.

Influence relative des différents paramètres

Ces premières expérimentations ont mis en évidence la complexité de l'influence des différents paramètres sur les performances de l'extraction de motifs. Il est ainsi difficile de mettre en évidence le "meilleur" ensemble de valeurs pour la configuration du système. Les valeurs relatives des différents paramètres semblent également particulièrement importantes. Si un paramètre n'est pas défini de façon suffisamment restrictive, les autres doivent l'être pour tenter de compenser et ne pas trop dégrader les performances du système. Ces relations entre les paramètres sont cependant relativement complexes, d'autant plus que chaque paramètre n'intervient pas forcément qu'à une seule étape de l'extraction de motifs. Les mesures de distance, et par conséquent également le seuil maximum de distance, sont utilisés pour l'identification et l'extension des motifs lors de l'examen de la matrice de collisions, mais aussi pour la classification des tentatives de motifs.

Afin de définir finalement des valeurs appropriées des différents paramètres, on étudie alors leur influence relative sur les performances de l'identification et de la classification. On essaie cependant déjà de limiter les intervalles de variation aux valeurs les plus probables suite à l'analyse de la qualité de la méthode et des premières expérimentations de la qualité des résultats. Les intervalles de variation alors considérés sont décrits dans le tableau 6.6. Ils correspondent aux contraintes suivantes :

- **Un écart maximum de similarité restrictif dans les valeurs** : approximativement, ϵ_{LCSS} peut être compris entre 0.3 et 0.4 ;

- **un écart maximum de similarité restrictif mais qui reste cependant assez souple dans le temps** : on choisit par exemple $\delta_{LCSS} = 60$, ce qui autorise un écart d'au plus une heure dans le temps entre deux points similaires ;
- **un seuil maximum de distance relativement faible mais adapté aux mesures de distances ainsi réalisées**, qui sont globalement plus élevées étant donné les valeurs plus restrictives de ϵ_{LCSS} et δ_{LCSS} : approximativement d_{max} peut être compris entre 0.4 et 0.5 ;
- **un seuil minimum de collisions relativement faible**, particulièrement dans le contexte de valeurs plus restrictives des autres paramètres : approximativement, c_{min} peut être compris entre 0.1 et 0.2.

Paramètre	Intervalle de variation
c_{min}	[0.1 0.2]
d_{max}	[0.4 0.5]
δ_{LCSS}	60
ϵ_{LCSS}	[0.3 0.4]

TAB. 6.6 – Variation des paramètres de réglage pour l'étude de leur influence relative sur les performances du système.

On étudie en particulier l'influence relative des seuils minimum de collisions (c_{min}) et maximum de distance (d_{max}) et de l'écart maximum de similarité sur les valeurs (ϵ_{LCSS}). Les performances moyennes obtenues dans ces configurations du système sont présentées sur la figure 6.14. Les résultats montrent de nouveau qu'il est difficile d'extraire des tendances d'évolution des performances en fonction des valeurs de chaque paramètre. Les indices de classification sont particulièrement critiques dans notre contexte. Parmi les configurations expérimentées, on remarque que les meilleurs indices sont obtenus quand $c_{min} = 0.1$, $d_{max} = 0.4$ et $\epsilon_{LCSS} = 0.4$, sachant que $\delta_{LCSS} = 60$. Pour ce triplet de valeurs, les performances de l'identification des motifs ne sont pas les meilleures mais restent relativement bonnes.

Une analyse plus complète pourrait être réalisée par l'étude de la distribution des indices de sensibilité et de spécificité dans l'espace des paramètres clés du système. On aurait peut-être mieux mis en évidence leurs influences relatives et le "meilleur" réglage du système. Cependant, compte tenu de l'absence de données réelles pour cette expérimentation et du manque de connaissances expertes disponibles, ce niveau de finesse de l'analyse n'est pas forcément justifié à cette étape d'expérimentation et de validation.

Définition de valeurs appropriées des différents paramètres

D'après les résultats précédents, une configuration appropriée du système est définie par défaut selon les valeurs de paramètres du tableau 6.7.
Les performances obtenues en moyenne pour un grand nombre de tests sont présentées dans le tableau 6.8. Les indices de sensibilité et de spécificité observés sont globalement meilleurs pour l'identification des tentatives de motifs que pour leur classification. Dans chacun des cas, la spécificité est aussi en moyenne supérieure à la sensibilité. Ces tendances peuvent être interprétées comme suit :

- **Pour l'identification des tentatives de motifs**, les sous-séquences récurrentes identifiées sont plus souvent "sous-étendues" par rapport aux véritables instances de motifs que l'inverse.

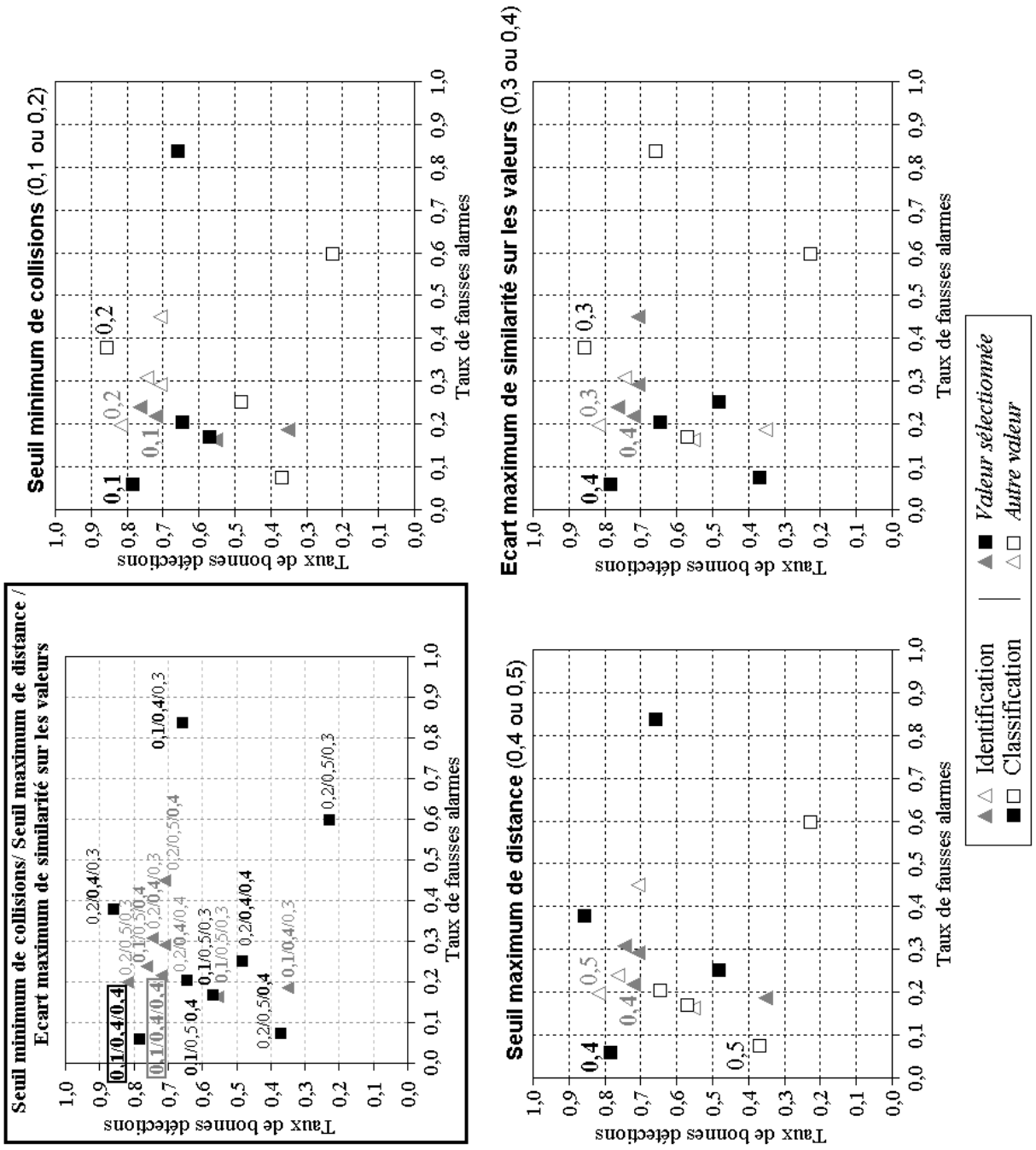


FIG. 6.14 – Performances de l'identification des tentatives de motifs et de leur classification en motifs relativement pour différentes valeurs des seuils minimum de collisions (c_{min}) et maximum de distance (d_{max}) et de l'écart maximum de similarité sur les valeurs (ϵ_{LCSS}).
On introduit entre les instances de chaque motif une quantité "raisonnable" de bruit selon des taux de 0.2 pour les valeurs, 0.1 pour les déformations dans le temps, et 0.05 pour les interruptions. Les graphes présentent les performances moyennes de l'identification des tentatives de motifs (triangles) et de leur classifications en motifs (carrés).
Le graphe encadré (en haut à gauche) présente les performances moyennes étiquetées selon le triplet correspondant de valeurs c_{min}/ d_{max}/ ϵ_{LCSS}, avec $\delta_{LCSS} = 60$.
Pour une meilleure lisibilité, les trois autres graphes présentent ces mêmes performances moyennes mais étiquetées pour chaque paramètre étudié en fonction des valeurs sélectionnées par défaut (formes pleines) et les autres valeurs expérimentées (formes vides).

– **Pour la classification en motifs**, des résultats non parfaits correspondent plutôt à une ou plusieurs instances "manquées" et écartées d'une classe principale contenant les autres instances du même motif qu'à la prise en compte de sous-séquences non représentatives de ces classes.

Paramètre	Valeur par défaut
c_{min}	0.1
d_{max}	0.4
δ_{LCSS}	60
ϵ_{LCSS}	0.4

TAB. 6.7 – Valeurs par défaut des paramètres de réglage du système.

On constate également que les instances de motifs sont en moyenne peu fractionnées dans leur reconnaissance. Chaque instance est identifiée comme une seule tentative de motif dans 70% des cas.

On remarque enfin qu'il existe une certaine variabilité dans les indices de performance, particulièrement pour ce qui concerne la classification, et même plus précisément pour la sensibilité de la classification. Une classification parfaite des tentatives de motifs est réalisée dans environ 20% des cas. Mais par ailleurs, la sensibilité de la classification peut aussi être nulle. Il s'agit des cas où les instances d'un motif, même si elles sont reconnues, ne sont pas toutes regroupées dans une même classe, si bien qu'elles peuvent constituer plusieurs classes de faible effectif qui ne sont par conséquent pas considérées comme significatives. Cette variabilité dans les performances peut être en partie liée à la sélection aléatoire des motifs considérés. Certains types de motifs sont probablement plus "faciles" à identifier que d'autres, selon leurs caractéristiques en terme d'alternances de valeurs dans les paramètres, de stationnarité des valeurs au long de l'activité correspondante, etc.

Finalement, ces résultats montrent qu'il est effectivement possible d'identifier correctement et de classer parfaitement les instances d'un motif par la méthode proposée. Il existe cependant une certaine variabilité dans les résultats qui laisse supposer une diversité des motifs sélectionnés aléatoirement. Pour la validation des performances de la méthode, on doit ainsi s'assurer d'une expérimentation prenant en compte uniquement des motifs complètement significatifs et interprétables au regard des contraintes et objectifs de l'extraction de sous-séquences récurrentes. Cela nécessite l'analyse de données réelles, enregistrées dans notre contexte dans un véritable contexte de télésurveillance médicale à domicile pour des personnes ayant annotées leurs activités quotidiennes.

Un exemple d'une classification parfaite à l'issue de l'extraction de motifs est présenté sur la figure 6.15. Les instances du motif insérées sont presque parfaitement délimitées et correctement regroupées en une seule classe. La caractérisation des motifs extraits correspond par ailleurs bien aux tendances générales de variation observées pour le motif initial. On retrouve en particulier l'idée des "pics" de valeurs mais lissés par la méthode de reconstruction du motif à partir des instances bruitées identifiées.

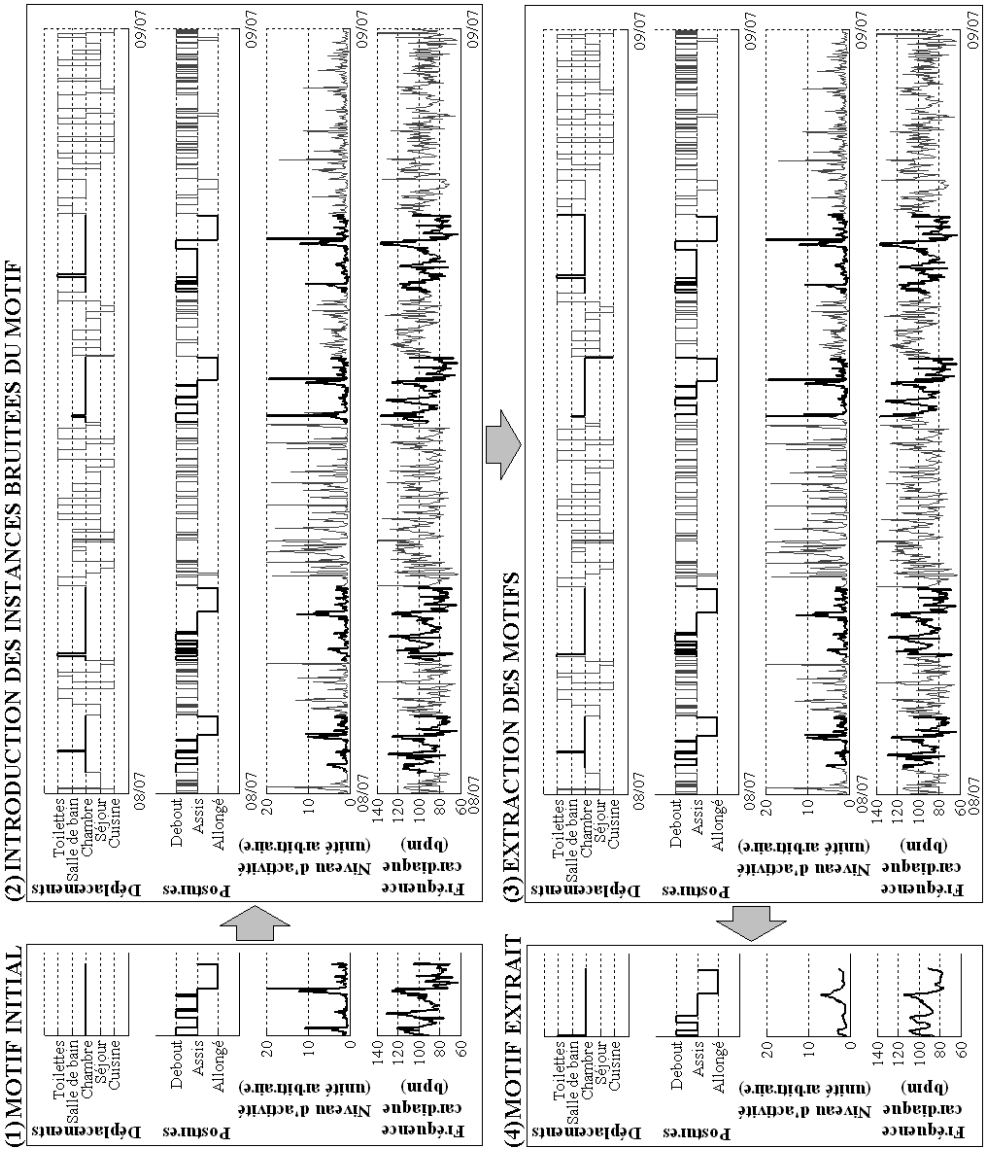


FIG. 6.15 – Illustration de l'extraction des instances d'un motif insérés initialement dans une séquence de données simulée à partir de déplacements aléatoires.
Les étapes présentées sont les suivantes :

(1) On définit *a priori* un motif, c'est-à-dire une sous-séquence à quatre dimensions (déplacements, postures, niveau d'activité, fréquence cardiaque) issue des données simulées pour une personne télésurveillée à domicile ;

(2) Des instances bruitées de ce motif sont introduites dans une séquence de données simulée à partir de déplacements aléatoires (en gras sur le graphe) ;

(3) Le processus d'extraction de motifs proposé identifie et regroupe dans une même classe les sous-séquences récurrentes représentées en gras sur le graphe correspondant ;

(4) La caractérisation du représentant de cette classe permet d'identifier le motif présent dans la séquence initiale.

	Identification		**Classification**		**Fractionnement**
Indices	S_e	S_p	S_e	S_p	λ
Moyenne	**0.71**	**0.92**	**0.66**	**0.79**	**0.89**
Écart-type	0.18	0.07	0.34	0.26	0.19
Indices parfaits	–	–	35%	60%	70%
Classification parfaite			**20%**		

TAB. 6.8 – Indices moyens de performance de l'approche proposée pour l'extraction de motifs dans la configuration par défaut du système et dans un contexte "raisonnablement" bruité.
On présente les performances du système en termes de la sensibilité (S_e) et de la spécificité (S_p) de l'identification des tentatives de motifs et de leur classification en motifs, ainsi que l'indice de fractionnement (λ) de la reconnaissance des motifs. Tous ces indices sont des réels compris entre 0 et 1, les valeurs maximum correspondant aux performances idéales.

6.3.3 Test de Sensibilité : modifications "normales" de comportement

Dans la configuration par défaut du système identifiée au paragraphe précédent, on étudie la résistance du système à la présence de bruit entre les instances des motifs. Par défaut, les taux de bruit sont nuls, et on fait ensuite varier successivement les taux suivants :

- **Variabilité dans les valeurs** : entre 0.0 et 1.0 ;
- **Déformation temporelle** : entre 0.0 et 0.5 – c'est le taux de déformation dans le temps par rapport à la durée initiale du motif ;
- **Interruptions** : entre 0.0 et 0.5 – c'est le rapport de la durée d'une interruption insérée en fonction de celle du motif.

Les résultats obtenus pour les performances du système sont présentés sur la figure 6.16. Globalement, on constate que la présence de bruit entre les instances de motifs affecte bien uniquement les performances en terme de la sensibilité (et non de la spécificité) de l'identification et de la classification des sous-séquences récurrentes. Les résultats restent généralement assez bons, particulièrement pour l'identification des tentatives de motifs. Il semble néanmoins qu'une faible dégradation de la sensibilité de cette identification résulte en de plus lourdes conséquence sur les indices de performances de la classification. Une mauvaise localisation des sous-séquences récurrentes induit en effet de plus larges mesures de distance qui peuvent perturber la classification. Le processus d'extraction de motifs semble malgré cela particulièrement bien résistant à la présence de bruit dans les valeurs et de déformations dans le temps, un peu moins à la présence de longues interruptions.

Dans les cas particuliers des déformations temporelles et interruptions, la contrainte sur l'écart maximum de similarité dans le temps (δ_{LCSS}) entraîne volontairement la non-détection des instances de motifs trop déformées ou comprenant des interruptions prolongées. Ces importantes déformations remettent en cause la structure même des activités et peuvent être considérées comme inquiétantes (cas par exemple d'interruptions répétées par de longs passages aux toilettes). Une déformation correspondant à un taux de 0.5 signifie que la durée de l'activité est prolongée au maximum de la moitié de sa durée, ce qui peut être déjà considéré comme étonnant. De même une interruption dont la durée correspond au maximum à la moitié de la durée de l'activité (taux de 0.5 également) ne doit pas forcément non plus être considérée comme ne perturbant pas son déroulement. On pressent qu'il est néanmoins difficile dans certains cas de bien délimiter les modifications "normales" de celles qui doivent être considérées comme inquiétantes. D'autres facteurs doivent peut-être être pris en compte (lieu de l'interruption par exemple, etc.)

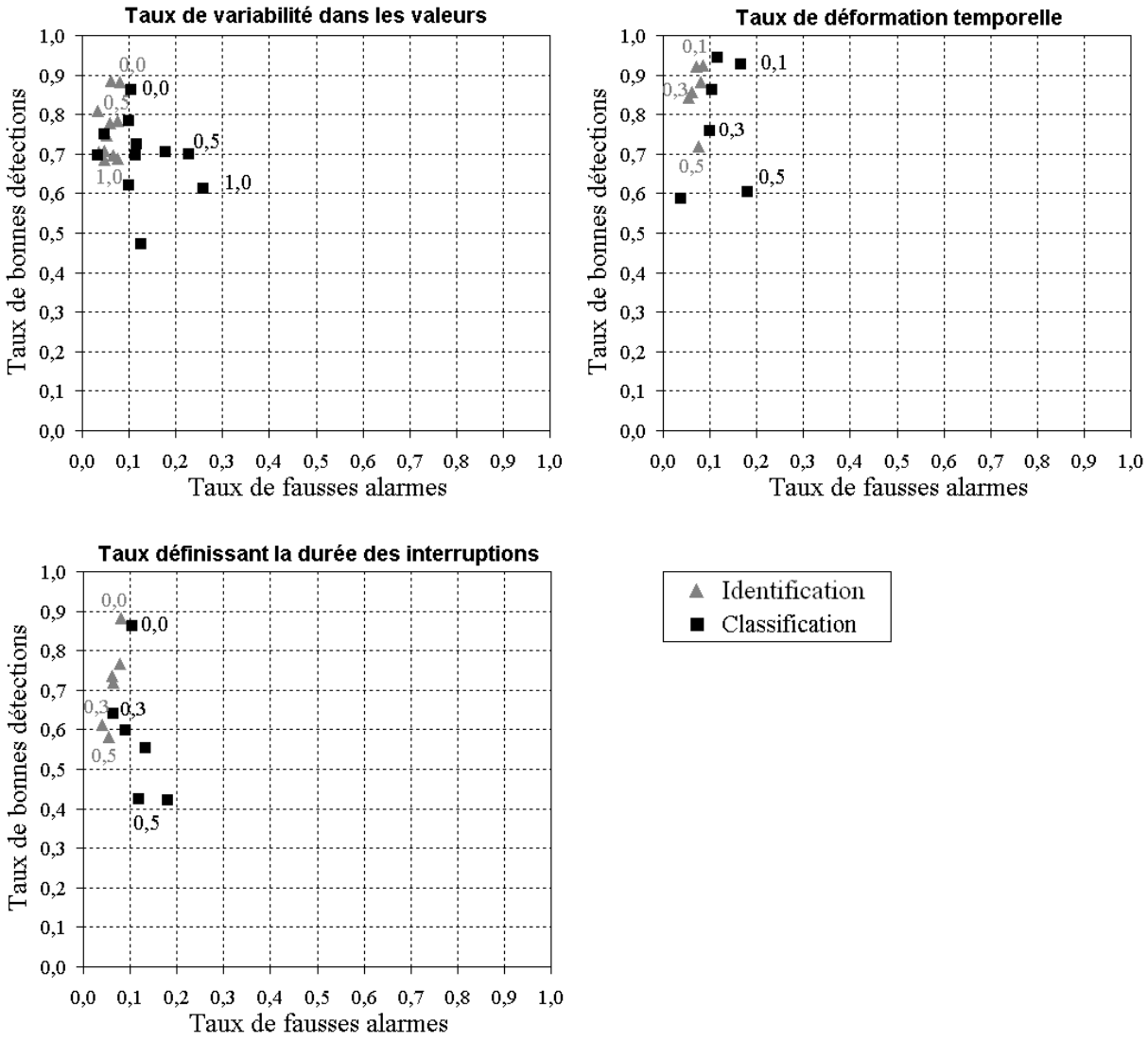


FIG. 6.16 – Performances de l'identification des tentatives de motifs et de leur classification en motifs en fonction des taux de bruit insérés entre les instances de motifs.
Chaque graphe correspond à l'étude d'un des paramètres clés du système. Il présente les performances moyennes de l'identification des tentatives de motifs (triangles) et de leur classifications en motifs (carrés). Les étiquettes de données indiquent, pour les points du graphe les plus significatifs en terme de la mise en évidence de tendances de variation, le taux du bruit considéré, les autres types de bruit n'étant alors pas pris en compte.

6.3.4 Test de Spécificité : modifications inquiétantes de comportement

Dans la configuration par défaut identifiée pour le système, on évalue enfin la spécificité de l'approche proposée pour l'extraction de motifs en terme de la non détection des instances dégradées d'un motif. Compte tenu du risque mentionné au paragraphe 6.3.2 d'une extraction de motifs gouvernée par les paramètres qualitatifs, on se place pour cette évaluation dans le cas particulier d'une déviation des valeurs d'un paramètre quantitatif. L'exemple proposé concerne l'augmentation progressive de la fréquence cardiaque moyenne quelle que soit l'activité réalisée. Ce type de déviation apparaît par exemple dans le cas de l'installation d'une insuffisance cardiaque : la fréquence cardiaque au repos peut alors monter jusqu'à 90 ou 100 battements par minute au lieu des 70 habituels (voir annexe B).

Au niveau de la génération d'instances d'un motif représentatives de ce type d'évolution inquiétante, on modifie progressivement certains paramètres de la simulation selon des taux de 10, 20, 30 puis 40 % d'augmentation par rapport aux valeurs habituelles, de la façon suivante :

- Augmentation de la valeur moyenne de la fréquence cardiaque au repos (*Mésor*), évoluant alors d'environ 70 bpm pour un individu "moyen" à 77, 84, 91 puis 98 bpm ;
- Dégradation de la relation approximativement linéaire entre le niveau d'activité et la fréquence cardiaque vers une fonction quasiment constante pour la génération de fréquences cardiaques élevées quelle que soit l'activité réalisée : selon les mêmes rapports de déviation, on introduit une diminution progressive de la pente de la fonction linéaire caractérisant cette dépendance en même qu'une augmentation de l'ordonnée à l'origine.

Un exemple de l'insertion d'instances "normalement" modifiées (c'est-à-dire, bruitées) puis progressivement dégradées d'un motif dans ce contexte est présenté sur le figure 6.17. Chaque motif est sélectionné aléatoirement dans une séquence de données générée par le processus de simulation pour un individu "moyen". On génère ensuite une séquence de données à partir de déplacements aléatoires correspondant à trois journées d'enregistrement. On introduit dans chacune des deux premières journées quatre instances bruitées du motif considéré. La dernière journée contient elle les quatre instances progressivement dégradées de ce même motif. On réalise ainsi plusieurs expérimentations successives.

Les résultats obtenus lors de l'application de l'algorithme d'extraction de motifs sont exprimés sur le tableau 6.9 en terme du pourcentage d'instances "normalement modifiées" puis dégradées de motifs reconnues et regroupées dans une même classe. On remarque que les instances correspondant aux modifications inquiétantes de comportement sont rarement associées à la classe des instances dites "normales" du motif. Ce taux de reconnaissance diminue par ailleurs progressivement avec une dégradation de plus en plus importantes des motifs. Ces constatations permettent de conclure à une bonne spécificité de la méthode proposée pour l'extraction de motifs.

Modifications	**Normales**	**Inquiétantes**			
Taux de modification	*0%*	*10%*	*20%*	*30%*	*40%*
Taux de reconnaissance	**77.5%**	**25%**	**10%**	**10%**	**5%**

Tab. 6.9 – Résultats de l'extraction de motifs à partir de séquences contenant des instances normalement modifiées puis dégradées d'un même motif.
Le tableau présente en moyenne le pourcentage des instances d'un motif reconnues et regroupées dans une même classe dite "normale". On considère d'une part les taux de reconnaissance des instances normalement modifiées, et d'autre part celui des instances correspondant à une modification inquiétante du comportement, avec différents taux de dégradation.

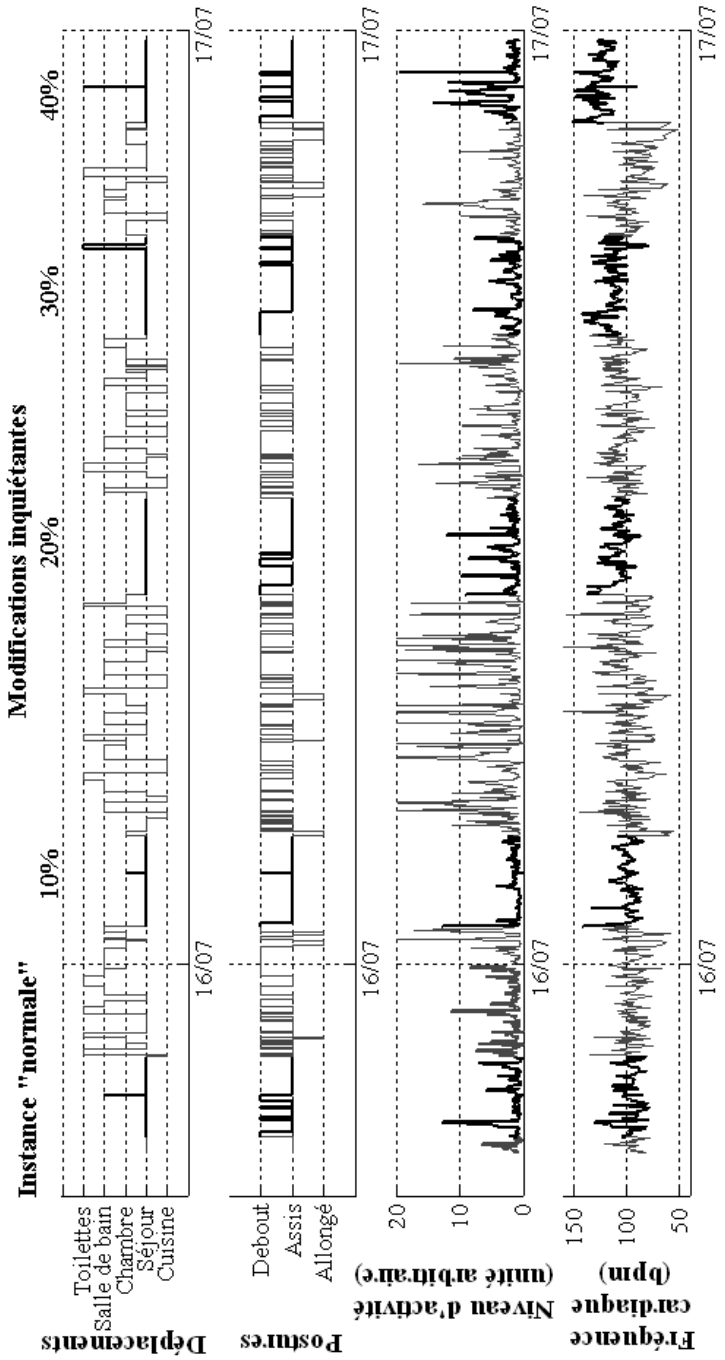


FIG. 6.17 – Illustration de l'introduction d'une modification inquiétante de comportement entre les instances d'un motif.

Le graphe présente une séquence à quatre dimensions issue de la simulation incluant (1) les déplacements, (2) les postures, (3) le niveau d'activité et (4) la fréquence cardiaque. Les instances d'un motif (en gras sur le graphe) sont insérées dans une séquence simulée à partir de déplacements aléatoires selon des modifications de plus en plus inquiétantes d'une sous-séquence de base représentant le motif. Elles correspondent à une augmentation globale de la fréquence cardiaque indépendamment de l'activité, avec des taux croissants de 10 à 40% d'augmentation.

6.4 Validation de la simulation par la décision

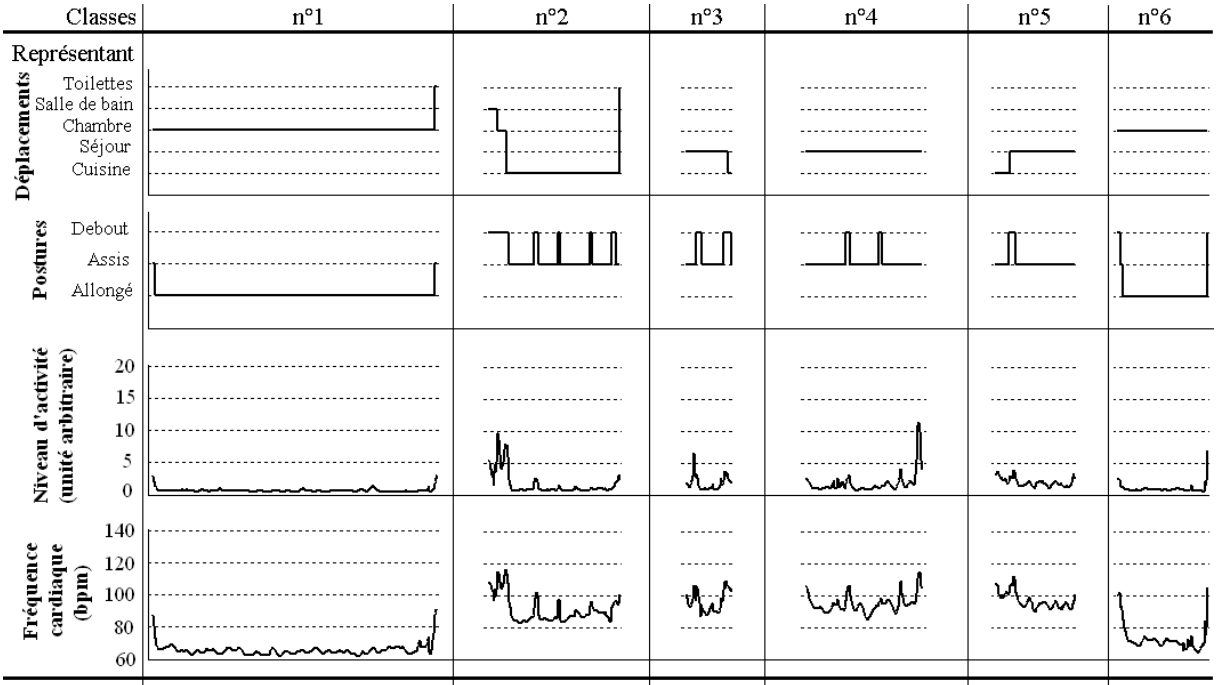
Dans notre contexte expérimental, on ne dispose pas d'enregistrements réels issus de la télésurveillance médicale d'une personne à domicile pour vérifier la pertinence de la méthode proposée pour l'identification des comportements récurrents d'une personne dans sa vie quotidienne. On propose cependant d'observer les résultats issus de l'extraction de motifs sur une séquence de données générée par le processus de simulation dans les conditions habituelles de vie d'un individu "moyen".

La figure 6.18 présente les caractéristiques des motifs identifiés à partir de l'analyse d'une séquence correspondant à la télésurveillance d'une personne pendant sept journées consécutives, du matin de la première au soir de la dernière. On constate que les motifs identifiés sont tout à fait interprétables *a posteriori* en terme de la réalisation d'activités récurrentes de la vie quotidienne. Le nombre d'instances identifiées pour les motifs correspondant à ces activités est cependant souvent inférieur au nombre d'occurrences attendues, c'est-à-dire le plus souvent à une fois par jour. C'est par exemple le cas de la nuit de sommeil qui devrait alors être identifiée six fois pendant la période d'apprentissage. Ces remarques remettent en partie en cause soit le réalisme des séquences générées par le processus de simulation, soit le réglage ou l'ajustement du système d'apprentissage. Un cycle de raffinement de l'un et/ ou l'autre de ces processus nécessite cependant les données expérimentales issues d'un système réel pour identifier plus précisément les contraintes supplémentaires et les évolutions à apporter.

6.5 Synthèse

Dans ce chapitre, on a proposé un **processus expérimental** pour l'évaluation de la pertinence de la méthode proposée pour l'expérimentation de l'extraction de motifs à partir de séquences de données multidimensionnelles et hétérogènes. On a pour cela notamment redéfini des indices de sensibilité et de spécificité appropriés à notre contexte d'apprentissage non supervisé.
L'évaluation de la **qualité de la méthode** permet d'abord de valider séparément chaque étape du processus. On détermine alors les valeurs appropriées d'un ensemble de paramètres de la méthode.
Les autres paramètres dits *de réglage* sont définis par l'analyse de la **qualité des résultats**. Les valeurs sélectionnées correspondent aux meilleures performances obtenues dans le contexte de l'application. On évalue alors en particulier la *sensibilité* (contexte de modifications "normales" de comportement) et la *spécificité* (contexte de modifications inquiétantes)

Le système ainsi configuré et expérimenté est enfin utilisé pour l'extraction de motifs à partir de séquences de données issues de la simulation. On vérifie ainsi qu'il est possible d'extraire des sous-séquences représentatives de comportements qu'on sait interpréter *a posteriori* en terme de la réalisation de certaines activités de la vie quotidienne.



Classes	n°1	n°2	n°3	n°4	n°5	n°6
Effectif	**3**	**5**	**5**	**3**	**10**	**5**
Durée	**7h30**	**3h30**	**1h15**	**3h**	**2h**	**2h30**
Instant d'occurrence	**vers minuit**	**vers 9h**	**vers 13h ou 20h**	**vers 15h**	**entre 16h et 19h**	**vers 14h ou 21h**
Fréquence moyenne reconnue	tous les deux jours	presque tous les matins	un jour sur deux	un jour sur deux	une à deux fois par jour	un jour sur deux
Interprétation de l'activité	***Nuit de sommeil***	***Toilette du matin Petite activité dans la cuisine et repas***	***Petite période calme au salon***	***Longue période calme au salon suivie d'une activité plus intense***	***Repas rapide ou collation suivie d'une activité calme***	***Période courte de repos***

FIG. 6.18 – Caractéristiques des motifs identifiés à partir de l'analyse d'une séquence simulée pour un individu "moyen" pendant sept journées consécutives.
Les séquences de données correspondant au représentant moyen de chaque classe significative extraite (effectif supérieur ou égal à 3 sous-séquences) sont présentées dans le haut de la figure. Dans le bas de la figure, on précise pour chaque classe les caractéristiques moyennes : effectif de la classe, durée et instant d'occurrence des instances du motif, fréquence moyenne de la reconnaissance du motif dans chacune des sept journées considérées pour l'apprentissage – par exemple, "un jour sur deux" est ainsi une considération moyenne et ne correspond pas à la reconnaissance exacte du motif tous les deux jours. On donne enfin une interprétation possible *a posteriori* de l'activité habituelle correspondante.

7

Discussion et Conclusion

Dans cette partie, une approche générique pour l'apprentissage des motifs contenus dans une séquence de données a été proposée. Son expérimentation a été réalisée dans le contexte de la télésurveillance médicale à domicile, à partir des séquences de données générées par le processus de simulation défini dans la partie II. Les résultats obtenus permettent de mettre en avant les points forts et les qualités innovantes de la méthode par rapport aux autres travaux effectués dans des domaines similaires ou connexes, en même temps que sa complexité et les perspectives d'amélioration de son efficacité. Ce chapitre s'articule ainsi autour de trois points fondamentaux concernant d'abord les caractéristiques et l'originalité de l'approche proposée, puis les perspectives issues des résultats obtenus lors de son expérimentation. On discute enfin l'application possible de cette méthode à d'autres niveaux de détails et dans le contexte d'autres applications.

Caractéristiques fondamentales de la méthode proposée

La méthode proposée pour l'extraction de motifs de séquences temporelles permet tout d'abord de prendre en compte simultanément **plusieurs paramètres** pouvant être **hétérogènes** – quantitatifs ou qualitatifs, à modalités ordonnées ou non – pour décrire une situation observée. On a donc dû en particulier définir une **mesure de similarité** adaptée à ce type de séquences. Dans notre contexte d'application particulièrement bruité, on a vérifié la pertinence du choix d'une **méthode non métrique** de mesure de similarité – fondée sur la plus longue sous-séquence commune – par rapport à une méthode métrique (distance DTW). Les travaux déjà réalisés autour de ce type de distances ont été ensuite étendus pour la prise en compte de composantes hétérogènes.

Un autre point important est l'aspect complètement non supervisé, tant pour l'identification des instances de motifs que pour leur classification. On se place dans un cadre où on ne dispose d'aucun *a priori* sur les motifs à extraire. Par ailleurs les séquences temporelles observées peuvent contenir à la fois des *motifs* et des *non motifs*.

Un intérêt particulier de la méthode proposée est qu'elle ne nécessite pas un grand ensemble de données d'apprentissage : dès qu'un motif a deux représentants dans la séquence étudiée, il peut être identifié. Dans notre application, si on suppose par exemple qu'une activité habituelle a lieu presque tous les jours, une séquence d'apprentissage correspondant aux données enregistrées pendant une semaine à dix jours permet alors d'avoir déjà une bonne idée des comportements réguliers de la personne. Une conséquence de l'identification de sous-séquences récurrentes sur peu d'instances des différents motifs est la considération d'un certain nombre de "classes parasites", correspondant "au hasard" de deux comportements similaires. Ces classes ont un faible effectif et leur représentant correspond le plus souvent à une séquence de courte durée. Elles sont ainsi

facilement supprimées par des seuils sur la durée d'une sous-séquence et l'effectif d'une classe définissant des tentatives de motifs et des motifs significatifs.

En terme des caractéristiques des instances de motifs, on a proposé une méthode qui prend en compte une grande variabilité possible entre les instances d'un même motif, tant au niveau temporel – déformations temporelles, interruptions – que dans les valeurs des différents paramètres au cours de la réalisation d'une même activité. L'utilisation comme première étape de recherche de sous-séquences récurrentes de l'algorithme de projections aléatoires permet notamment de tolérer d'importants taux de bruit entre les instances d'un même motif. Par contre, l'implémentation de cet algorithme dans le cadre des travaux de Chiu *et al.* [28] n'autorise pas les déformations temporelles entre sous-séquences similaires. Dans notre cas, l'étape préliminaire d'abstraction des données brutes permet d'exprimer une séquence initiale comme une succession de symboles pouvant correspondre à des durées différentes. Par conséquent, les sous-séquences de base considérées pour la réalisation des projections, même si elles comportent le même nombre de symboles, ne correspondent pas pour autant à la même durée, si bien que les instances de motifs peuvent être déformées dans le temps. Une étape suivante d'extension des sous-séquences récurrentes de base identifiées suite aux projections aléatoires permet par ailleurs de reconnaître des instances "complètes" de motifs. Une méthode de synthèse des tentatives de motifs a ensuite été proposée dans le contexte de l'analyse des résultats des projections aléatoires, pour répondre aux critères de fréquence, signification et non redondance des tentatives de motifs.

Pour ce qui concerne l'architecture du processus, on a construit une méthode qui permet d'analyser des séquences de données bas niveau pour l'identification de motifs significatifs à haut niveau de décision. Cela implique de mettre en place un système à plusieurs niveaux d'extraction des informations pertinentes par rapport aux objectifs de décision, jusqu'à la découverte de la localisation des instances de motifs et leur classification en motifs. On a ainsi défini ou redéfini selon les cas plusieurs étapes de traitement :

(1) La représentation des signaux, considérée en fait comme une véritable étape d'abstraction des données brutes ;

(2) La fouille de caractères pour l'identification des sous-séquences qui sont les "meilleures candidates" pour correspondre effectivement à des instances de motifs – les *tentatives de motifs* ;

(3) La catégorisation de ces sous-séquences pour l'identifications des *motifs*.

Expérimentation et validation des résultats

L'analyse de la qualité de la méthode et des résultats a mis en évidence à la fois les potentialités et la complexité de mise en pratique de l'approche proposée. Le manque de données expérimentales issues d'un système réel de télésurveillance médicale à domicile n'a cependant pas permis une validation complète des résultats obtenus.

Tout d'abord, les valeurs appropriées d'un certain nombre de paramètres impliqués dans le processus doivent être déterminées *a priori*. Cette démarche nécessite plusieurs études préliminaires dans le contexte de l'application considérée. L'identification de valeurs pertinentes pour les autres paramètres dits "de réglage" n'est pas non plus très simple compte tenu de leurs inter-dépendances et de leur implication à différents niveaux de l'analyse.

Dans une configuration par défaut du système, on a néanmoins montré les potentialités de la méthode pour l'extraction et la classification des instances de motifs présents dans les données initiales. En particulier, l'extraction de motifs sur les séquences de données issues du processus de simulation a finalement mis en évidence l'identification des comportements récurrents d'une personne à domicile. Cependant, même si l'expérimentation prouve qu'il est possible de reconnaître

les activités de la vie quotidienne sur des séquences de données simulées, rien ne valide véritablement l'efficacité de la méthode pour la construction effective d'un profil comportemental. Cette démarche de validation nécessite des enregistrements dont on ne dispose pas, en provenance d'un système réel de télésurveillance médicale à domicile, et pendant des périodes durant lesquelles les personnes ont annoté les activités récurrentes de leur vie quotidienne.

L'expérimentation du système sur les données issues de la simulation fait cependant apparaître des perspectives d'amélioration du système à différents niveaux de l'analyse des séquences temporelles.

Au niveau de l'abstraction, on peut envisager par exemple de donner des poids différents aux paramètres considérés pour l'agrégation des données successives similaires sur l'axe temporel. Une meilleure validation de cette étape nécessite par ailleurs des séquences de données réelles pour la comparaison des tendances de variation extraites et des activités effectivement réalisées pendant les périodes correspondant aux différents symboles.

Au niveau de l'identification des tentatives de motifs, la démarche proposée ne permet pas toujours d'identifier les instances complètes des motifs. La tendance observée sur les performances correspond à une "sous-extension" des sous-séquences de base identifiées comme récurrentes. Il serait par conséquent nécessaire, mais difficilement envisageable, de relaxer certaines contraintes liées à l'extension. Les seuls paramètres considérés à ce niveau sont en effet les seuils minimum de collisions et maximum de distance déjà optimisés par une analyse comparative de la qualité des résultats pour différentes valeurs possibles. Une autre méthode intéressante à expérimenter consisterait à identifier initialement, à partir de l'examen de la matrice de collisions, uniquement des sous-séquences récurrentes correspondant à des sous-séquences de base, sans extension possible. Une conséquence est la redéfinition nécessaire de la démarche de synthèse de ces sous-séquences en tentatives de motifs pour satisfaire les contraintes de fréquence, signification et non redondance. Les caractéristiques des sous-séquences récurrentes identifiées sont en effet très différentes dans ce nouveau contexte. Par exemple, on ne peut alors plus faire l'hypothèse qu'une classe significative de ces sous-séquences, dont le représentant définit une tentative de motif, ne doit contenir que des sous-séquences qui se superposent toutes mutuellement.

Au niveau de la classification en motifs, on constate que les erreurs correspondent le plus souvent à une ou plusieurs instances effectivement identifiées mais mal classées car leur distance avec l'une des autres sous-séquences associée pourtant au même motif est supérieure au seuil maximum de distance. Si on augmente ce seuil, on risque cependant de prendre en compte trop de sous-séquences non significatives ou correspondant aux instances dégradées d'un motif. L'étape précédente d'analyse permettant d'extraire les sous-séquences les plus significatives en termes de l'identification des motifs, la contrainte sur les mesures de distance observées pour la classification peut probablement être moins stricte à ce niveau de l'analyse. Une démarche intéressante serait ainsi d'expérimenter la dissociation des seuils maximum de distance pris en compte lors de ces deux étapes, en augmentant légèrement celui correspondant à la classification.

Perspectives d'application

La généricité de la méthode proposée doit permettre de l'appliquer à différents niveaux de détail de l'analyse de la situation d'une personne à domicile d'une part, et à différentes applications de surveillance d'autre part.

Choix du niveau de détail

Les niveaux de détail considérés dans le cadre de ces travaux ne sont pas limitatifs pour l'utilisation de la méthode proposée. En effet, le réglage des différents paramètres impliqués dans le

processus permet de modifier le niveau de décision. La conséquence est l'identification de motifs correspondant à des niveaux de détail différents sur les axes des valeurs et du temps.
Les *paramètres de l'abstraction* définissent en particulier l'échelle de temps considérée pour l'analyse, en permettant l'extraction de tendances de variation significatives à cette l'échelle. Le nombre d'intervalles de discrétisation ou les caractéristiques du filtrage moyen des données brutes permettent par exemple d'ajuster la précision de l'analyse des tendances de variation sur l'axe temporel.
La modification des **paramètres de la fouille de données** définit plutôt le niveau de précision recherché entre les instances d'un même motif. On peut par exemple changer la variabilité autorisée dans les valeurs et dans le temps entre les instances d'un même motif en influant sur les paramètres définissant la mesure de distance et les différents seuils. Le nombre de symboles définissant une sous-séquence de base pour la réalisation des projections aléatoires permet quant à lui d'agir sur la durée minimum d'une instance significative.

Une autre façon de changer le niveau de décision est de considérer des données initiales à différentes échelles de temps. Par exemple on peut considérer des données très hautes fréquences (enregistrées toutes les secondes ou millisecondes par exemple, pour l'étude d'un comportement particulièrement précis dans sa réalisation), ou à l'inverse des données enregistrées seulement une fois par jour, par exemple pour étudier l'évolution de l'état de santé d'une personne en termes de l'évolution quotidienne du poids, des pressions artérielles et d'un paramètre global qualifiant le niveau d'activité par exemple. Un paramètre global d'activité peut typiquement être issu d'une analyse à un plus bas niveau de décision.

Choix de l'application
Toujours en terme de généricité, la méthode proposée a aussi été définie de façon a être utilisée pour n'importe quelle application qui concerne l'extraction de caractères habituels – sous-séquences récurrentes – à partir de l'observation de l'évolution dans le temps d'un ensemble possiblement hétérogène de paramètres complémentaires représentatifs des situations habituelles. Les paramètres considérés peuvent être soit quantitatifs, soit qualitatifs, à modalités ordonnées ou non.

Il est cependant nécessaire de définir pour la méthode d'extraction de motifs les paramètres les plus adaptés à chaque contexte considéré, et de bien valider leur choix aux différentes étapes du processus. Cette démarche doit être abordée selon la méthodologie proposée au début de ce document (voir Fig. I.3.2).

Conclusion et Perspectives

Ce travail de thèse a permis une première analyse de l'apprentissage des habitudes de vie d'une personne à domicile à partir de la fusion de données enregistrées par un ensemble de capteurs. Tout écart par rapport à ce profil comportemental est susceptible de correspondre à une situation inquiétante ou critique. La démarche proposée comprend (I) la définition du contexte et des objectifs de la problématique, (II) la mise en place d'un processus de simulation de données adaptées à sa résolution et (III) la définition d'un système d'extraction de motifs dans des séquences de données multidimensionnelles et hétérogènes permettant d'identifier les comportements récurrents d'une personne à domicile dans ses activités de la vie quotidienne.

La formulation de la problématique d'étude a mis en évidence la nécessité de bien définir les contraintes et objectifs de résolution à la fois pour la mise un place d'un processus de simulation et pour la construction d'un système de décision. Ces trois éléments forment ainsi un cycle de résolution d'un problème qui peut être itéré tant que les résultats de la décision ne sont pas en adéquation avec les objectifs fixés.

La mise en place d'un processus de simulation dans ce contexte a en particulier montré la complexité des variations individuelles et conjointes des paramètres étudiés, le manque de connaissances *a priori* disponibles, et la diversité des profils de comportement et des situations rencontrées. Une démarche incrémentale et hybride a été proposée dans ce contexte pour la construction d'un modèle et son expérimentation. La validation des résultats obtenus n'a cependant pu être réalisée qu'en partie étant donné le manque de données expérimentales issues d'un système réel de télésurveillance. On estime cependant "suffisant" leur niveau de validité au "haut niveau" d'analyse considéré dans le cadre de l'observation de l'évolution de la situation d'une personne à long terme. Même si les séquences générées ne sont pas précisément celles enregistrées potentiellement par un système réel, elles permettent néanmoins de disposer d'une grande quantité de données relativement réalistes, représentatives de différents profils de personnes et types de situations, et appropriées à l'étude du profil comportemental d'une personne.

L'apprentissage du profil comportemental d'une personne est ensuite réalisé à partir des séquences de données issues de la simulation par l'extraction des motifs ou classes de sous-séquences récurrentes représentatives des comportements habituels dans la vie quotidienne. On propose une approche complètement non supervisée, permettant l'identification de motifs multidimensionnels dans des séquences temporelles de données hétérogènes, et autorisant différents types de bruit entre les instances d'un même motif : déformation temporelle, variabilité dans les valeurs et interruptions. Les résultats de l'expérimentation montrent les potentialités de la méthode pour la reconnaissance de motifs en contexte bruité (sensibilité) et la non détection des instances anormalement modifiées d'un motif (spécificité). L'analyse de séquences de données issues de processus de simulation et représentatives des conditions habituelles de vie d'un individu donné permet alors d'identifier un ensemble de comportements récurrents que l'on sait interpréter *a posteriori* en termes des activités de la vie quotidienne. Il reste cependant à valider la pertinence de son application à des données enregistrées d'un système réel.

Les approches proposées dans ce travail pour la simulation et la décision ne pourront être complètement validées que par leur expérimentation sur des séquences de données enregistrées en environnement réel de télésurveillance. L'évaluation de la pertinence du processus de simulation de données et de l'identification de comportements récurrents dans ce contexte permettra en particulier de savoir si un autre cycle supplémentaire de résolution du problème est nécessaire. Il peut s'agir d'un raffinement du processus de simulation et/ ou de la méthode d'extraction de motifs.

Les perspectives à plus long terme de ce travail concernent d'abord la caractérisation du profil comportemental d'une personne en terme d'une succession de *motifs* et de *non-motifs* : des activités régulières, identifiées grâce au processus d'apprentissage proposé, suivies de périodes où les activités réalisées sont peu prévisibles. Le profil de comportement peut alors être caractérisé à l'aide de graphes temporels, par points (chroniques) ou par intervalles (algèbre de Allen). Il s'agit de définir des événements (réalisation d'une activité) et des contraintes entre ces événements : ordre de réalisation des activités de la vie quotidienne, instants de réalisation, intervalle de temps entre chaque activité, etc. Les habitudes de vie quotidienne d'une personne peuvent alors être représentées par un ou plusieurs graphes, ayant peut-être chacun une probabilité d'occurrence spécifique.

La démarche suivante concerne la détection de situations inquiétantes ou critiques par comparaison des séquences de données enregistrées à tout moment à ce profil comportemental caractérisé dans des conditions habituelles de vie de la personne. La mise en place de ce processus de décision comporte deux étapes principales : (1) identification des motifs définis par apprentissage dans les nouvelles séquences enregistrées et (2) comparaison de leurs caractéristiques d'occurrence (durée, ordre, fréquence, etc.) au profil comportemental de la personne.

La première de ces deux étapes est en fait une extraction supervisée de motifs dans des séquences de données temporelles. Elle peut être réalisée sur le même principe que l'approche proposée dans un contexte non supervisé d'extraction de motifs. On introduit simplement une différence au niveau de la fouille de caractères pour l'identification des sous-séquences les plus susceptibles de correspondre à un comportement récurrent d'une personne. Dans le cas d'une extraction supervisée, pour laquelle on connaît *a priori* un représentant de chaque motif, on construit la matrice de collisions à partir de la comparaison des projections aléatoires de toutes les sous-séquences de base issues de la séquence initiale d'une part, et des sous-séquences de base constituant chaque motif à identifier d'autre part. Une forte valeur de collisions entre une sous-séquence et le représentant d'un motif laisse alors supposer que la sous-séquence correspondante est une nouvelle instance de ce motif. On peut ainsi représenter les séquences observées en temps réel par des graphes temporel correspondant aux successions de motifs et de non-motifs et leurs contraintes temporelles d'occurrence.

La seconde étape consiste alors à comparer les graphes temporels construits au cours de l'observation de la personne dans sa vie quotidienne aux graphes représentatifs de son comportement habituel et issus de la période d'apprentissage. Les situations inquiétantes correspondent à une mauvaise adéquation de ces deux ensembles des graphes. La mise au point de cette étape de décision nécessite de prendre en compte des séquences de données représentatives de dégradations du comportement de la personne. Il est alors intéressant d'évaluer la sensibilité et la spécificité de la détection des situations inquiétantes, ainsi que le temps de détection de ses situations à partir du moment où des perturbations significatives sont intégrées dans les séquences temporelles.

On peut ensuite envisager de prendre progressivement en compte de nouvelles données dans le système d'apprentissage, tout en supprimant la considération de données considérées trop anciennes, pour la mise à jour du profil comportemental en fonction des évolutions possibles des habitudes de vie quotidienne d'une personne.

Finalement, la validation de ces étapes d'identification du profil comportement d'une personne à domicile et de détection des situations critiques ne pourra être réalisée complètement que par l'expérimentation des algorithmes correspondant avec des séquences de données enregistrées dans un environnement réel de télésurveillance médicale. Il est également indispensable de la compléter par la contribution d'experts du domaine, c'est-à-dire notamment de praticiens hospitaliers de différentes spécialités, telle que la gériatrie, particulièrement concernés par la prise en charge médicale possible de leurs patients à domicile.

Une autre perspective intéressante de ces travaux est d'expérimenter l'utilisation de la méthode générique proposée pour l'extraction non supervisée de motifs multidimensionnels et hétérogènes à d'autres niveaux de détail des données et de la décision, et pour d'autres applications de surveillance.

Bibliographie

[1] H. Afsarmanesh, V. Guevara-Masis, L.O. Hertzberger, "Federated management of information for TeleCARE," in *Proc. of the 1st International Workshop on Tele-Care and Collaborative Virtual Communities in Elderly Care*, Porto, Portugal, April 2004.

[2] R. Agrawal, K. Sawhney, K. Shim, "Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases," in *Proc. of the 21st International Conference on Very Large Data Bases (VLDB)*, Zurich, Switzerland, September 1995, pp. 490–501.

[3] C.M. Antunes and A.L. Oliveira, "Temporal Data Mining : an overview," in *Proc. of the Workshop on Temporal Data Mining, at the 7th International Conference on Knowledge Discovery and Data Mining (KDD'01)*, San Francisco, CA, August 2001, pp. 1–15.

[4] A. Apostolico, "String editing and longest common subsequence," *G. Rozenberg and A. Salomaa, editors, Handbook of Formal Languages*, Springer Verlag, vol. 2, pp. 361–398, Berlin, 1997.

[5] L. Bajolle, "E-médecine : Amélioration, Optimisation et Humanisation de la médecine de ville par l'usage de l'internet et des nouvelles technologies," *Thèse de doctorat en médecine de l'Université Joseph Fourier*, Grenoble, janvier 2002.

[6] T. Bass, "Intrusion detection systems & multisensor data fusion," *Communications ACM*, vol. 43(4), pp. 99-105, 2000.

[7] L. Bergroth, H. Hakonen, and T. Raita, "A survey of Longest Common Subsequence Algorithms," in *Proc. of the 7th International Symposium on String Processing Information Retrieval (SPIRE'00)*, Coruna, Spain, 2000, pp. 39–48.

[8] D. Berndt & J. Clifford, "Using Dynamic Time Warping to find Patterns in Time Series," in *Proc. of AAAI, Workshop on Knowledge Discovery in Databases,* Seattle, Washington, 1994.

[9] G. Blessed, B. Tomlinson, M. Roth, "The association between quantitative measures of dementia and of senile change in the cerebral gray matter of elderly subjects," *Br J Psychiatry*, vol. 114, pp. 797–811, 1968.

[10] I. Bloch, "Fusion d'informations en traitement du signal et des images," *Edit. Hermès Science*, Paris, France, 318 p., janvier 2003.

[11] A.F. Bobick, "Movement, activity, and action : the role of knowledge in the perception of motion," *Philosophical Transactions*, vol. 352, pp. 1257–1265, 1997.

[12] S.G. Bonner, "Assisted Interactive Dwelling House," in *Proc. of the 3rd TIDE Congress : Technology for Inclusive Design and Equality Improving the Quality of Life fir the European Citizen*, Helsinki, Finland, 1998.

[13] D. Buonomano, "Ethique et télémédecine, aspects spécifiques du questionnement éthique appliqué à la télémédecine," *Télémédecine en Gérontologie*, A. Franco, M. Frossard, et C. Montani (eds), Serdi, Paris, 2000, pp. 209–217.

[14] L. Bourdon, A. Buguet, M. Cucherat, M.W. Radomski, "Use of a Spreadsheet Program for Circadian Analysis of Bilogical/Physiologial Data," *Aviation, Space, and Environmental Medicine*, vol. 66(8), pp. 787–791, August 1995.

[15] B.R. Bracio, W. Horn, P.F. Dietmar, "Sensor fusion in biomedical systems," in *Proc. of the 19th IEEE-EMBS*, Chicago, USA, 1997, Clausthal (ed), pp. 1387–1390.

[16] L. Brouha, "La physiologie de l'homme au travail," *Edit. corpo. prof.*, Paris, France, 1 vol., 94 p.

[17] J. Buhler and M. Tompa, "Finding motifs using random projections," *Journal of Computational Biology*, vol. 9(2), pp. 225–242, 2002.

[18] K. Cameron, K. Hughes, K. Doughty, "Reducing fall incidence in community elders by telecare using predictive systems," in *Proc. of the 19th Annual International Conference of the IEEE Engineering in Medicine & Biology Society*, Chicago, USA, 1997.

[19] J. Cannady, "Artificial Neural Networks for misuse detection," in *Proc. of the 21st National Information Systems Security Conference*, Crystal City, Virginia, 1998, pp. 441–454.

[20] M. Cauville, "Diagnostic, soins et prévention par la télémédecine : explications de J. Demongeot," *Siences et Technologies*, pp. 32–34, 1999.

[21] B.G. Celler, T. Hesketh, W. Earnshaw, E. Ilsar, "An instrumentation system for the remote monitoring of changes in functional health status of the elderly at home," in *Proc. of the 16th Annual IEEE Engineering in Medicine and Biology Society*, Baltimore, USA, 1994, pp. 908–909.

[22] B.G. Celler, W. Earnshaw, E. Ilsar, L. Betbeder-Matibet, M.F. Harris, R. Clark, T. Hesketh, N.H. Lovell, "Remote monitoring of health status of the elderly at home. A multidisciplinary project on aging at the university of South Wales," *Int J Biomed Comput*, vol. 40, pp. 147–155, 1995.

[23] B.G. Celler, W. Earnshaw, E. Ilsar, "Remote monitoring of the elderly at home : preliminary results of a pilot project at the University of N.S.W.," *J Biomed Eng – Applications, Basis and Communications*, vol.9, pp. 134–140, 1997.

[24] M. Chan, H. Bocquet, E. Campo, T. Val, J. Pous, "Alarm communication network to help carers of the elderly for safety purposes : a survey of a project," *Int J Rehabil Res*, vol. 22, pp. 131-136, 1999.

[25] K. Chan, A.W. Fu, "Efficient time series matching by wavelets," in *Proc. of the 15th IEEE International Conference on Data Engineering*, Sydney, Australia, Mar. 23–26, 1999, pp. 126–133.

[26] S. Charbonnier, J.P. Siché, J.M. Mallion, "Toward a portable blood pressure recorder device equipped with accelerometers," *Medical Engineering and Physics*, vol. 21, pp. 343-352, 1999.

[27] M. Chavent, "A monothetic clustering method," *Pattern Recognition Letters*, vol. 19, pp. 989–996, 1998.

[28] B. Chiu, E. Keogh, and S. Lonardi, "Probabilistic Discovery of Time Series Motifs," in *Proc. of the 9th ACM International Conference on Knowledge Discovery and Data Mining (KDD'03)*, Washington DC., August 2003, pp. 493–498.

[29] A. Choudhri, L. Kagal, A. Joshi, T. Finin, Y. Yesha, "PatientService : A system for Electronic Patient Record Redaction and Delivery in Pervasive Environments," in *Proc. of the 5th International Workshop on Enterprise Networking and Computing in Healthcare Industry*, Santa Monica, CA, 2003, pp. 41–47.

[30] E.-H. Christensen and P. Högberg, "Steady-state, O_2-deficit and O_2-debt at severe work," *Arbeitsphysiol.*, Berlin, vol. 14, pp. 251–154, 1950.

[31] Conservatoire National des Arts et Métiers, "Support de cours d'Ergonomie," Paris, *http ://www.cnam.fr/ergonomie/.*

[32] L. Chwif, R.J. Paul, "On simulation model complexity," in *Proc. of the 32nd conference on Winter simulation*, Orlando, Florida, 2000, pp. 449–455.

[33] G. Das, D. Gunopulos, and H. Mannila, "Finding similar time series," *Principles of Data Mining and Knowledge Discovery*, vol. 19, pp. 88–100, 1997.

[34] G. Das, K. Lin, H. Mannila, G. Renganathan, and P. Smyth, "Rule discovery from time series," in *Proc. of the 4th International Conference on Knowledge Discovery and Data Mining (KDD'98)*, New-York, NY, August 27-31, 1998, pp. 16–22.

[35] R. DeJong, O. Osterlund, G. Roy, "Measurement of quality-of-life changes in patients with Alzheimer's disease," *Clin ther*, vol. 11, pp. 545–554, 1989.

[36] K. Doughty, R. Isak, P.J. King, P. Smith, G. Williams, "MIDAS – Miniature Intelligent Domiciliary Alarm System - a practical application of telecare," in *Proc. of the 1st Joint BMES/EMBS Conf Serving Humanity, Advancing Technology*, Atlanta, USA, 1999, pp. 691.

[37] G. Elger, B. Furugren, ""SmartBo" – An ICT an computer-based demonstration home for disabled people," in *Proc. of the 3rd TIDE Congress : Technology for Inclusive Design and Equality Improving the Quality of Life for the European Citizen*, Helsinki, Finland, 1998.

[38] J. Finkelstein, G. O'Connor, R.H. Friedman, "Development and implementation of the Home Asthma Telemonitoring (HAT) system to facilitate asthma self-care," in *Proc. of the 10th World Congress on Medical Informatics (MEDINFO)*, London, UK, 2001, pp. 810–814.

[39] A. Franco, "La télémédecine au service de l'autonomie," *La revue de médecine interne*, vol. 24(suppl. 4), pp. 390s–393s.

[40] W. Frawley, G. Piatetsky-Shapiro, C. Matheus, "Knowledge Discovery in Databases : An Overview," *AI Magazine*, vol. 13(3), pp. 57–70, 1992.

[41] A. Galata, N. Johnson, D.C. Hogg, "Learning behaviour models of human activities," in *Proc of the British Machine Vision Conference*, Nottingham, 1999.

[42] S. Gatepaille et S. Brunessaux, "Dossier I.A. et Fusion de données," *Bulletin de l'Association Française pour l'Intelligence Artificielle (AFIA)*, vol. 24, pp. 21–30, Janvier 1996.

[43] M. Gavrilov, D. Anguelov, P. Indyk, R. Motwani, "Mining The Stock Market : Which Measure Is Best ?," in *Proc. of the 6th ACM International Conference on Knowledge Discovery and Data Mining*, Boston, MA, Aug. 20–23, 2000, pp. 487–496.

[44] A.K. Gosh, A. Schwartzbard, M. Schatz, "Learning program behavior profiles for intrusion detection," in *Proc. of the 1st USENIX Workshop on Intrusion Detection and Network Monitoring*, Santa Clara, CA, 1999, pp. 51–62.

[45] K.Z. Haigh et H.A. Yanco, "Automation as Caregiver : A Survey Of Issues and Technologies," in *Automation as Caregiver : The Role of Intelligent Technology in Elder Care*, AAAI Press, Edmonton, Alberta, pp. 39–53, 2002.

[46] F. Halberg, "Chronobiology," *Ann Rev Physiol*, vol. 31, pp. 675–725, 1969.

[47] D. S. Hirschberg, "Algorithms for the longest common subsequence problem," *JACM*, vol. 24 (4), pp. 664–675, 1977.

[48] P. Hong, T.S. Huang, "Learning to extract multi-temporal signal patterns from a temporal signal sequence," in *Proc. of the 15th International Conference on Pattern Recognition*, Barcelona, Spain, September 3-7, 2000.

[49] P. Hong, S.R. Ray, T.S. Huang, "A new scheme for extracting multi temporal sequence patterns," in *Proc. of the International Joint Conference on Neural Networks*, Washington DC., USA, July 10-16, 1999.

[50] F. Höppner, "Discovery of Temporal Patterns – Learning Rules about the Qualitative Behaviour of Time Series," in *Proc. of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01)*, Freiburg, Germany, 2001, pp. 192–203.

[51] F. Höppner, "Time Series Abstraction Methods – A survey," in *Proc. of the GI Jahrestagung Informatik, Workshop on Knowledge Discovery in Databases*, Dortmund, Germany, 2002, pp. 777–786.

[52] J. W. Hunt & T. G. Szymanski, "A fast algorithm for computing longest common subsequences," *CACM*, vol. 20 (5), pp. 350–353, 1977.

[53] S. Katz, A.B. Ford, R.W. Moscokowitz, B.A. Jackson, M.W. Jaffe, "The index of ADL : a standardized measure of biological and psychosocial function," *J Am Med Assoc*, vol. 185, pp. 914–919, 1963.

[54] W.D. Kelton, "Perspectives on Simulation Research and Practice," *ORSA Journal on Computing*, vol. 6, pp. 318–328, 1994.

[55] E.J. Keogh and M.J. Pazzani, "An enhaced representation of time series which allows fast and accurate classification, clustering and relevance feedback," in *Proc. of the 4th International Conference on Knowledge Discovery and Data Mining (KDD'98)*, New-York, NY, August 27-31, 1998, pp. 239-241.

[56] E. Keogh, M. Pazzani, "Scaling up Dynamic Time Warping for Datamining Applications," in *Proc. of the 21st Int. Conf. on Very Large Databases*, Boston, MA, 2000, pp. 285–289.

[57] E. Keogh, M. Pazzani, "A Simple Dimensionality Reduction Technique for Fast Similarity Search in Large Time Series Databases," in *Proc. of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications*, April 18–20, 2000, pp. 122–133.

[58] E. Keogh, K. Chakrabarti, S. Mehrotra, M. Pazzani, "Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Database," in *Proc. of the ACM SIGMOD Conference on Management of Data*, Santa Barbara, CA, May 21–24, 2001, pp. 151–162.

[59] E. Keogh, K. Chakrabarti, M. Pazzani, S. Mehrotm, "Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases," *Journal of Knowledge and Information Systems*, pp. 263–286, 2000.

[60] F. Korn, H.V. Jagadish, C. Faloutsos, "Efficiently Suporting Ad Hoc Queries in Large Datasets of Time," in *Proc. of ACM SIGMOD*, Tucson, AZ, May 1997, pp. 289–300.

[61] J. B. Kruskall & M. Liberman, "The symmetric time warping algorithm : From continuous to discrete," *Time Warps, String Edits and Macromolecules*, Addison-Wesley, 1983.

[62] D. Kudenko and H. Hirsh, "Feature generation for sequence categorization," in *Proc. of the 15th Nat'l Conf. Artificial Intelligence (AAAI'98)*, AAAI Press, Menlo Park, California, pp. 733-739.

[63] T. Lane, C.E. Brodley, "Temporal sequence learning and data reduction for anomaly detection," *ACM Transactions on information and System Security*, vol. 2, pp. 295–331, 1999.

[64] M.P. Lawton, E. Brody, "Assessment of older people : self-maintaining and instrumental activities of daily living," *Gerontologist*, pp. 179–186, 1969.

[65] S.L. Lee, S.J. Chun, D.H. Kim, J.H. Lee, and C.W. Chung, "Similarity Search for Multidimensional Data Sequences," in *Proc. of the 16th IEEE International Conference on Data Engineering (ICDE'00)*, California, 2000, pp. 599–608.

[66] W. Lee, S.J. Stolfo, "Data mining approaches for intrusion detection," in *Proc. of the 7th USENIX Security Symposium*, San Antonio, Texas, 1998.

[67] N. Lesh, M.J. Zaki, M. Ogihara, "Scalable Feature Mining for Sequential Data," *IEEE Intelligent Systems*, vol. 15(2), pp. 48–56, March/April 2000.

[68] N. Lesh, M.J. Zaki, M. Ogihara, "Mining features for sequence classification," in *Proc. of the 5th International Conference on Knowledge Discovery and Data Mining (KDD'99)*, August 1999, pp. 342–346.

[69] J. Lin, E. Keogh, P. Patel, and S. Lonardi, "Finding motifs in time series," in *Proc. of the 2nd Workshop on Temporal Data Mining, at the 8th International Conference on Knowledge Discovery and Data Mining (KDD'02)*, Edmonton, Alberta, Canada, July 2002, pp. 53–68.

[70] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic Representation of Time Series, with Implications for Streaming Algorithms," in *Proc. of the ACM Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD'03)*, San Diego, CA, 2003, pp. 2–11.

[71] L. Lind, E. Sundvall, H. Ahlfeldt, "Experiences from development of home health care applications based on emerging Java technology," in *Proc. of the 10th World Congress on Medical Informatics (MEDINFO)*, London, UK, 2001, pp. 830–834.

[72] N. Meratnia, R.A. de By, "Aggregation and Comparison of Trajectories," in *Proc. of the 10th ACM International Symposium on Advances in geographic information systems*, McLean, Virginia, USA, November 8-9, 2002, pp. 49–54.

[73] E. Micheli-Tzanakou, "Supervised and Unsupervised Pattern Recognition : Feature Extraction and Computational Intelligence," Eds : Boca Raton, FL : CRC, 371 pp., 2000, ISBN : 0-8493-2278-2, *reviewed by* K. Chen, *IEEE Transaction on Neural Networks*, vol. 12(3), pp. 644–647, 2001.

[74] H. Monod et M. Pottier, "Adaptations respiratoires et circulatoires du travail musculaire," *Précis de physiologie du travail, Notions d'Ergonomie*, 2nd ed., J. Scherrer et al., Ed. Paris : Masson, pp. 159–204, 1981.

[75] B. Moon, I.F. Vega Lopez, and V. Immanuel, "Efficient Algorithms for Large-Scale Temporal Aggregation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15(3), pp. 744–759, May/June 2003.

[76] B. Najafi, K. Aminian, A. Paraschiv-Ionescu, F. Loew, C.J. Büla et P. Robert, "Ambulatory System for Human Motion Analysis Using a Kinematic Sensor : Monitoring of Daily Physical Activity in the Elderly," *IEEE Transactions on Biomedical Engineering*, vol. 50(6), pp. 711–723, June 2003.

[77] A. Nanopoulos, R. Alcock, Y. Manolopoulos, "Feature-based classification of time-series data," *International Journal of Computer Research, Special Issue : Information processing and technology*, vol. 10(3), pp. 49–61, 2001.

[78] A. Nemo, "La télémédecine : Faire voyager les informations plutôt que le malade," *Journal du Téléphone*, pp. 4, 1994.

[79] N. Noury, T. Hervé, V. Rialle, G. Virone, E. Mercier, "Monitoring behavior in home using smart fall sensor and position sensors," in *Proc. of the 1st IEEE-EMBS on Microtechnologies in Medicine and Biology*, Lyon, France, 2000, pp. 607–610.

[80] N. Noury, P. Barralon, G. Virone, P. Boissy, M. Hamel, P. Rumeau, "A Smart Sensor Based on Rules and its Evaluation in Daily Routines," in *Proc of the IEEE-EMBC*, Cancun-Mexico, septembre 2003, pp. 3286–3289.

[81] M. Ogawa, T. Togawa, "Attempt at monitoring health status in the home," in *Proc. of the 1st IEEE-EMBS on Microtechnologies in Medicine and Biology*, Lyon, France, 2000, Dittmar and Beebe (eds), pp. 552–556.

[82] T.I. Ören, "Simulation : Taxonomy," *Systems and Control Encyclopedia*, M.G. Singh, Ed. Oxford (England) : Pergammon Press, pp. 4411–4414, 1987.

[83] S. Park, S. Kim, W.W. Chu, " Segment-Based Approach for Subsequence Searches in Sequence Databases," in *Proc. of the 16th ACM Symposium on Applied Computing*, Las Vegas, NV, march 11–14, 2001, pp. 248–252.

[84] L.W. Pedretti, "Occupational Therapy. Practice skills for physical dysfunction," Chapter 25, D. Foderaro, "Cardiac dysfunction," 2nd edition, Eds : The C.V. Mosby Company, St Louis, Missouri, 1985.

[85] P.H.P. Peeters, "Design criteria for an automatic safety-alarm system for elderly," *Technol Health Care*, vol. 8, pp. 81–91, 2000.

[86] N. Pelletier-Fleury, "Analyse économique des difficultés de diffusion de la télémédecine," *Actes des 22ème Journées des Économistes Français de la Santé*, 1998.

[87] R. Pfeffer, T. Kurosaki, C. Harrah *et al.*, "Measurement of Functional Activities in older adults in the community," *J Gerontol*, vol. 37, pp. 323–329, 1982.

[88] H. Pigot, B. Lefebvre, J.-G. Meunier, B. Kerhervé, A. Mayers, S. Giroux, "The role of intelligent habitats in upholding elders in residence," in *Proc. of the 5th International Conference on Simulations in Biomedicine*, Slovenia, April 2003, pp. 497–506.

[89] F. Provost & T. Fawcett, "Analysis and visualization of classifier performance : Comparison under imprecise class and cost distribution," in *Proc. of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD-97)*, Huntington Beach, CA, 1997, AAAI Press, pp. 43–48.

[90] D. Rafiei, A. Mendelzon, "Efficient Retrieval of Similar Time Sequences Using DFT," in *Proc. of the 5th International Conference on Foundations of Data Organization and Algorithms*, Kobe, Japan, Nov. 12–13, 1998.

[91] L.R. Ramos, E.J. Simoes, M.S. Albert, "Dependence in activities of daily living and cognitive impairment strongly predicted mortality in older urban residents in Brazil : a 2-year follow-up," *J Am Geriatr Soc*, vol. 49(9), pp. 1168–1175, 2001.

[92] V. Rialle, N. Noury, T. Hervé, "An experimental health smart home and its distributed Internet-based Information and Communication System : first steps of a research project," in *Proc. of the 10th World Congress on Medical Informatics (MEDINFO)*, Londres, 2001, the Patel *et al.* (eds), pp. 1479–1483.

[93] V. Rialle, N. Noury, J. Fayn, M. Chan, E. Campo, L. Bajolle, J.P. Thomesse, "Health smart home information systems : concepts and illustrations," in *Proc. of the 3rd International Workshop on Enterprise Networking and Computing in Health Care Industry (HEALTHCOM)*, L'Aquila, Italy, 2001, pp. 99–103.

[94] V. Rialle, F. Duchêne, N. Noury, L. Bajolle, J. Demongeot, "Health "Smart" Home : Information Technology for Patients at Home," *Telemedicine Journal and E-Health*, vol. 8(4), pp. 395–410, Winter 2002.

[95] V. Rialle, P. Rumeau et C. Hervé, "Éléments pour une méthodologie d'analyse éthique des technologies d'aide au maintien à domicile de personnes en perte d'autonomie," *Actes du colloque Éthique Numérique*, Saint-Cyr-sur-mer, France, 12–14 mai 2003, L'Harmattan (ed), Paris, 2004.

[96] J.F. Roddick, M. Spiliopoulou, "A survey of Temporal Knowledge Discovery Paradigms and Methods," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14(4), pp. 750–767, July/August 2002.

[97] A. Roth, Z. Carthy, M. Benedk, "Telemedicine in emergency home care – the Shahal experience," *J Telemed Telecare*, vol. 3(1), pp. 58–60, 1997.

[98] M.J. Rodriguez, M.T. Arredondo, F. del Pozo, E.J. Gomez, A. Martinez, A. Dopico, "A home telecare management system," *J Telemed Telecare*, vol. 1, pp. 86–94, 1995.

[99] G. Saporta, "Probabilités, Analyse des Données et Statistique," *Edit. Technip.*, Paris, France, 1 vol., 493 p., 1990.

[100] R.G. Sargent, "Validation and verification of simulation models," in *Proc. of the 31st conference on Winter simulation*, Phoenix, Arizona, United States, 1999, pp. 39–48.

[101] R.E. Shannon,"Introduction to the Art and Science of simulation," *Proc. of the 30th Winter Simulation Conference*, Washington, D.C., United States, 1998, pp. 7–14.

[102] A-S. Silvent, C. Garbay, P.-Y. Carry et M. Dojat, "Rôle des données, informations et connaissances dans la construction de scénarios médicaux," *Extraction et gestion des connaissances (EGC'2003)*, M.-S. Hacid, Y. Kodratoff, D. Boulanger (eds), Hermès-Lavoisier, Lyon, 2003, vol. 17(1–3), pp. 207–212.

[103] C. Suarez, "La télémédecine : quelle légitimité d'une innovation radicale pour les professionnels de santé ? ," *Revue de l'Institut de Recherches Économiques et Sociales (IRES)*, vol. 39, 2002.

[104] O. Sueda, M. Ide, A. Honma, M. Yamaguchi, "Smart House in Tokushima," in *Proc. of the 5th European Conference for the Advancement of Assistive Technology*, Düsseldorf, Germany, 1999.

[105] S. Teunisse, M.M.A. Derix, H. van Crevel, "Assessing the severity of dementia : patient and caregiver," *Arch Neurol*, vol. 48, pp. 274–277, 1991.

[106] J.P. Thomesse, "TISSAD : Technologies de l'Information Intégrées aux Services des Soins À Domicile," *Télémédecine et e-santé*, R. Beuscart, P. Zweigenbaum, A. Venot, et P. Degoulet (eds), Springer–Verlag, collection "Informatique et Santé", Paris, France, pp. 27–34, 2002.

[107] Y. Tsai, "The constrained Longest Common Subsequence Problem," *Information Processing Lettres*, vol. 88(4), pp. 173–176, 2003.

[108] A. Van Berlo, "A "smart" model house as research and demonstration tool for telematics development," in *Proc. of the 3rd TIDE Congress : Technology for Inclusive Design and Equality Improving the Quality of Life fir the European Citizen*, Helsinki, Finland, 1998.

[109] G. Virone, B. Lefebvre, N. Noury, J. Demongeot, "Modeling and Computer Simulation of Physiological Rhythms and Behaviors at Home for Data Fusion Programs in a Telecare System," in *Proc 5th International Workshop on Enterprise Networking and Computing in Healthcare Industry*, Santa Monica, CA, USA, 2003, pp. 111–117.

[110] M. Vlachos, G. Kollios, ans G. Gunopulos, "Discovering Similar Multidimensional Trajectories," in *Proc. of the 18th International Conference on Data Engineering (ICDE'02)*, San Jose, CA, February 2002, pp. 673–684.

[111] R.A. Wagner et M.J. Fischer, "The string-to-string correction problem," *J. ACM*, vol. 21(1), pp. 168–173, 1974.

[112] R. Washburn, K. Smith, A. Jette, C. Janney, "The Physical Activity Scale for the elderly (PASE) : development and evaluation," *J Clin Epidemiol*, vol. 52(4), pp. 153–162, 1993.

[113] G. Williams, K. Doughty, D.A. Bradley, "A system approach to achieving CarerNet – an integrated and intelligent telecare system," *IEEE Trans Biomed Eng*, vol. 2, pp. 1–9, 1998.

[114] G. Williams, K. Doughty, D.A. Bradley, "Distributed intelligent nodes as information filters in advanced telecare systems," in *Proc. of the 21st Annual IEEE Engineering in Medicine & Biology Society*, Atlanta, USA, 1999, pp. 703.

[115] G. Williams, K. Doughty, K. Cameron, D.A. Bradley, "A smart fall and activity monitor for telecare applications," in *Proc. of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Hong-Kong, 1998, Chang and Zhang (eds), pp. 1151–1154.

[116] B. Yi, C. Faloutsos, "Fast time sequence indexing for arbitrary L_p norms," in *Proc. of the 26th International Conference on Very Large Databases*, Cairo, Egypt, Sept. 10–14, 2000, pp. 385–394.

A

Journée type d'une personne âgée

Ci-dessous la journée type d'une dame âgée fragile de 87 ans, veuve, telle que proposée par Pierre Rumeau, praticien hospitalier en gériatrie au CHU La Grave-Casselardit à Toulouse.

▷	6h30	Réveil
▷	6h45	Lever
▷	6h47	**WC**
▷	6h49	Se lave les mains
▷	6h52	Enfile sa robe de chambre
▷	6h54	*Prépare son petit déjeuner* : Café au lait avec Nescafé et lait bouilli dans une casserole sur la gazinière
▷	7h00	**Prise du petit déjeuner**, avec deux tartines beurrées, pain de la veille ou biscottes et *médicaments*
▷	7h30	*Petite vaisselle*
▷	7h35	**WC**, selles
▷	7h50	**Toilette** : Visage, haut du dos, creux axillaires, dents et dentier
▷	8h15	*S'habille*
▷	8h30 à 9h40	*Courses* importantes ou *ménage*
▷	10h00	Messe
▷	10h40	Retour, achète le pain, prend et ouvre le courrier en rentrant
▷	11h00	*Prépare le repas* et met la table au salon – Télévision
▷	11h45 à 12h30	**Repas**
▷	12h32	*Vaisselle*
▷	13h00	Actualités puis feuilleton allemand, 5 minutes de pause aux **WC**
▷	16h00	Visite d'amis, thé, discussions
▷	18h00	*Papiers* si besoin, lecture du courrier (salon), puis *préparation du repas* du soir, léger, bouillon, yaourt
▷	19h00	**Repas** du soir devant la télévision
▷	20h00	*Coup de téléphone*, durée 30 minutes : enfants/ petits-enfants/ etc.
▷	20h50	**Toilette** du soir (rapide)
▷	21h00 à 23h00	Télévision, puis lecture au lit
▷	22h00	*Somnifère* Entre la télévision et le lit, passage aux **WC**
▷	Vers 4h00	Un lever nocturne

Cette journée type d'une personne âgée montre bien la présence des **activités de base** de la vie quotidienne (en **gras**) – (1) dormir, (2) faire sa toilette, (3) s'alimenter, (4) aller au cabinet. On observe également dans cet exemple la présence habituelle d'activités plus complexes (en *italique*) telles que préparer un repas, faire les courses, faire le ménage, utiliser le téléphone, prendre ses médicaments.

B

Situations inquiétantes à domicile

Cette annexe présente quelques situations inquiétantes d'une personne âgée à domicile qui nous ont été décrites par Pierre Rumeau, praticien hospitalier en gériatrie au CHU La Grave-Casselardit à Toulouse.

Les scénarios proposés correspondent à des perturbations très spécifiques, mais sensibles, qui s'installent progressivement sur un moyen terme : (1) l'infection urinaire, (2) l'insuffisance cardiaque, et (3) la dépression. Ils sont décrits en terme de l'évolution de la pathologie et des conséquences sur la vie quotidienne de la personne concernée.

B.1 Infection urinaire

Une infection urinaire ne s'installe pas progressivement mais apparaît au contraire d'un coup. Une personne concernée par cette infection va beaucoup plus souvent aux toilettes : 10 et 12 fois par jour, avec jusqu'à 2 levers nocturnes. L'apparition possible d'une incontinence implique parfois des passages supplémentaires dans la chambre et la salle de bain après les passages aux toilettes. La personne est gênée et bouge beaucoup sur place, se lève parfois brutalement et marche plus vite. Des interruptions sont possibles même en pleine activité pour se rendre aux toilettes.

Spontanément, une personne souffrant d'une infection urinaire a tendance à boire moins. On peut observer une *tachycardie* – accélération du rythme cardiaque – environ 2 minutes avant le passage aux toilettes. Par exemple, une fréquence cardiaque habituelle de 70 battements par minute (bpm) dans un contexte donné est alors enregistrée plutôt à 80–90 bpm, voire plus.

Dans le cas d'une *pyélonéphrite* – infection urinaire profonde – la personne âgée est "abattue", avec la fièvre. Elle passe alors ses journées allongée ou semi-allongée, le plus souvent sur son lit.

B.2 Insuffisance cardiaque

L'insuffisance cardiaque s'installe progressivement, sur 1 mois environ, et les premiers signes notables apparaissent 15 jours plus avant. Elle peut exceptionnelle s'installer brutalement s'il y a en plus une infection.

Globalement, une personne concernée bouge moins, mange moins, se déplace moins vite, et chute éventuellement plus souvent. La position allongée n'est plus supportée, le sommeil est par conséquent mauvais, et la personne passe moins de temps au lit pour plus de temps dans un fauteuil au séjour. Quand elle est dans son lit, elle choisit alors une position semi-assise mais bouge beaucoup.

Au niveau physiologique, le coeur accélère progressivement. La fréquence cardiaque au repos peut monter jusqu'à 90 ou 100 battements par minute au lieu des 70 habituels. On observe également de la *dyspnée* – "difficulté à respirer", s'accompagnant d'une sensation de gêne ou d'oppression. La fréquence respiratoire normale inférieure à 12 inspirations par minute peut alors monter jusqu'à 20 à 25. La personne affectée prend également du poids à cause de l'apparition d'œdèmes – "Accumulation anormale de liquide séreux dans les espaces intercellulaires du tissu conjonctif".

B.3 Dépression

La dépression chez les personnes âgées survient suite à une facteur déclenchant, tel qu'une petite fille qui s'éloigne, et s'installe en moins d'une semaine (5–6 jours). L'angoisse peut cependant être déjà présente avant, se traduisant par des problèmes de sommeil.

Globalement, l'activité diminue, la personne concernée sort moins, se désinvestit. Elle devient *apathique* – ne réagit pas, paraît sans volonté, sans énergie. Elle fait moins la cuisine, les plats sont moins élaborés. Elle mange même parfois sans se mettre à table.

La personne est également souvent couchée. Par exemple une sieste habituellement d'1/2 heure à 1 heure dure parfois jusqu'à 4 heures. Par contre la personne se couche tard (vers minuit) et se lève tôt (vers 5h00) : l'angoisse les empêche de dormir. Le sommeil est plus agité, avec des mouvements plus réguliers que ceux d'un rythme de sommeil habituel. Malgré le manque de sommeil, la personne peut par contre rester plus longtemps dans sa chambre, voire même au lit (par exemple, jusque vers midi).

La dépression peut également être associée à une constipation. La personne ne va pas plus souvent aux toilettes mais par contre y passe plus de temps, jusqu'à 30 minutes.

Au niveau physiologique, la tension augmente, de même que la fréquence cardiaque, mais très légèrement : de 70 bpm au repos à 75 ou 80.

Les scénarios présentés pour décrirent ces pathologies montrent bien la relation étroite entre l'installation d'une pathologie à moyen terme et la réalisation des activités de la vie quotidienne. Certains paramètres physiologiques qui peuvent être relativement simples à mesurer – tension, fréquence cardiaque ou respiratoire – connaissent alors également des variations dégradées par rapport aux conditions habituelles de vie de la personne.

C

Tests statistiques sur la nature d'un échantillon

Ce chapitre détaille quelques tests d'ajustement utilisés pour vérifier qu'un échantillon provient ou non d'une variable aléatoire de distribution connue. Les informations présentées dans les paragraphes suivants ont essentiellement été extraites de [99].

C.1 Test d'ajustement graphique

Un test d'ajustement graphique permet de vérifier rapidement la compatibilité d'un échantillon avec l'hypothèse d'une distribution connue. Pour la plupart des lois de probabilité, une transformation fonctionnelle simple permet en effet de représenter la courbe de répartition par une droite. Si la taille de l'échantillon considérée est suffisamment grande, la fonction de répartition empirique – courbe des fréquences cumulées – diffère peu de la fonction de répartition théorique, et on peut alors vérifier simplement l'adéquation des données au modèle en comparant ces fonctions à une droite sur un papier à échelles fonctionnelles.

Dans le cas d'une loi de probabilité de *Laplace-Gauss* – $LG(m, \sigma)$ – pour une variable aléatoire X, la fonction de répartition n'a pas d'expression mathématique simple. On utilise alors les propriétés de la variable aléatoire centrée-réduite associée, $U = \frac{X-m}{\sigma}$. Si la distribution des valeurs empiriques de X, $\{x_i\}_{1 \leq i \leq n}$, est réellement gaussienne, il existe une relation linéaire entre x_i et u_i^*, où les u_i^* sont les valeurs de la variable aléatoire centrée réduite correspondant aux fréquences cumulées empiriques $\frac{effectif < x_i}{n}$. En effet, u_i^* doit alors peut différer de $\frac{x_i - m}{\sigma}$.

Ces méthodes empiriques ne permettent cependant pas de préciser les risques d'erreur.

C.2 Test de Kolmogorov-Smirnov

Le test de Kolmogorov-Smirnov est un test non paramétrique qui peut être utilisé comme test d'ajustement d'un échantillon de fonction cumulative $F_n^*(z)$ à une distribution entièrement spécifiée de fonction de répartition $F_o(z)$, ou pour déterminer si les fonctions de répartition de deux échantillons indépendants, $F_{n,X}^{1*}(z)$ et $F_{m,Y}^*(z)$, sont identiques.

La statistique de test est l'écart vertical maximal observé entre la répartition empirique F_n^* et la répartition théorique F_o, $D_n = sup\,|F_n^*(z) - F_o(z)|$, ou entre les deux répartitions empiriques $F_{n,X}^*$ et $F_{m,Y}^*$, $D_n = sup\left|F_{n,X}^*(z) - F_{m,Y}^*(z)\right|$. La fonction D_n est asymptotiquement distribuée

comme suit :

$$P(\sqrt{n}D_n < y) \rightarrow \sum_{-\infty}^{+\infty}(-1)^k exp(-2k^2y^2) = K(y).$$

La fonction $K(y)$ a été tabulée et fournit donc un test de l'hypothèse H_o – $F_n^*(z) = F_o(z)$ ou $F_{n,X}^*(z) = F_{m,Y}^*(z)$ pour tout z – qui dépend de la taille n de l'échantillon et du niveau de signification du test, α. Par exemple au seuil $\alpha = 0.05$ et si $n > 80$, la région critique est $D_n > \frac{1.3581}{\sqrt{n}}$.

C.3 Test de Lilliefors

Ce test compare la fonction de répartition d'une variable aléatoire X quelconque à celle d'une loi normale de moyenne m et d'écart-type σ non spécifiés.

La statistique de test est analogue à celle du test de Kolmogorov-Smirnov : il s'agit de l'écart vertical maximal $D_n = sup\left|\Phi(x) - \hat{F}_n(x)\right|$ observé entre une fonction de répartition empirique $\hat{F}_n(x)$ et la fonction de répartition de la loi normale $\Phi(x)$, toutes deux centrées et réduites. La fonction de répartition empirique "centrée-réduite", $\hat{F}_n(x)$, est estimée à partir de la moyenne empirique du n-échantillon $X_1, X_2, \ldots X_n$, $\bar{X} = \sum_{i=1}^n X_i/n$, et de l'écart-type empirique S défini par $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$. $\hat{F}_n(x)$ est alors la fonction de répartition des variables Z_i définies par : $Z_i = \frac{X_i - \bar{X}}{S}$.

Le test compare alors la distribution empirique de X à une distribution normale de mêmes moyenne et écart-type. Les estimateurs de la moyenne et de la variance introduisent cependant un aléa supplémentaire par rapport à la situation du test de Kolmogorov dont il faut tenir compte.

C.4 Application

Un des avantages des tests de Kolmogorov-Smirnov et de Lilliefors est leur possible utilisation quel que soit l'effectif de l'échantillon. Ces tests sont cependant de plus en plus considérés comme des méthodes assez pauvres, et pas forcément recommandées selon le contexte d'utilisation. Dans le cas d'une estimation grossière de distributions empiriques, ils donnent malgré tout une bonne idée de la nature d'un échantillon.

Dans notre contexte de modélisation, un test d'ajustement graphique est particulièrement utilisé pour vérifier une forte de présomption de normalité sur une distribution donnée, confirmée par un test de Lilliefors. Le test de Kolmogorov-Smirnov est quant à lui particulièrement utile à la vérification de la paramétrisation correcte d'une distribution expérimentale. On compare alors les distributions générées par le modèle à la distribution expérimentale.

D

Données expérimentales

Les données expérimentales dont on dispose correspondent à 8 sujets dont les activités quotidiennes ont été observées pendant deux périodes non consécutives de 24 heures. Ces données ont été enregistrées dans le cadre des travaux de recherche de Sylvie Charbonnier du Laboratoire d'Automatique de Grenoble (LAG). Les sujets portent un accéléromètre triaxial au niveau du torse, ainsi qu'un appareil portatif de mesure de l'ECG. On peut ainsi extraire en moyenne toutes les minutes les mesures suivantes :

- **Niveau d'activité**, défini sur une échelle arbitraire, et correspondant à la moyenne toutes les minutes de l'accélération enregistrée suivant l'axe antérieur-postérieur.
- **Fréquence cardiaque moyenne**, calculée toutes les minutes à partir des données enregistrées par l'appareil de mesure de l'ECG.

Par ailleurs, les sujets ont annoté leurs activités toutes les 15 minutes, parmi un ensemble de 14 activités exigeant des niveaux d'énergie différents : (1) dormir, (2) être allongé, (3) être assis calmement, (4) être assis et discuter, (5) être assis et travailler, (6) manger, (7) être debout, (8) être debout et travailler, (9) marcher doucement, (10) marcher vite, (11) faire de la bicyclette, (12) courir, (13) monter et (14) descendre les escaliers.

Les figures présentées ci-après présentent les données expérimentales enregistrées pour les 8 sujets de l'expérimentation. Ces enregistrements montrent une grande variabilité des données selon les sujets, mais aussi selon la situation. Par exemple, des niveaux d'activité régulièrement autour de 20 à 30 sont en particulier enregistrés pour le sujet 2 ou 8, alors que les valeurs correspondant aux sujets 3 et 4 ne dépassent qu'à peine 10. Concernant les situations rencontrées, on constate par exemple pour le sujet 1 une importante activité nocturne, mais uniquement dans le contexte du second enregistrement. On en pressent clairement les conséquences par exemple sur l'observation des variations de la fréquence cardiaque : on observe sur le deuxième enregistrement une tendance globale à l'augmentation des valeurs pendant la nuit, ce qui n'est pas le cas sur le premier enregistrement pour ce sujet.

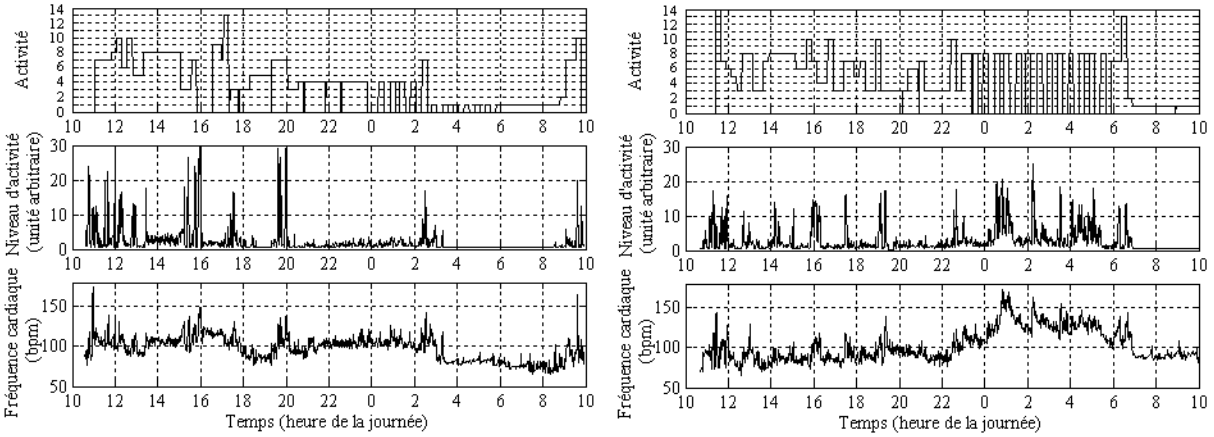


FIG. D.1 – Enregistrements du sujet 1.

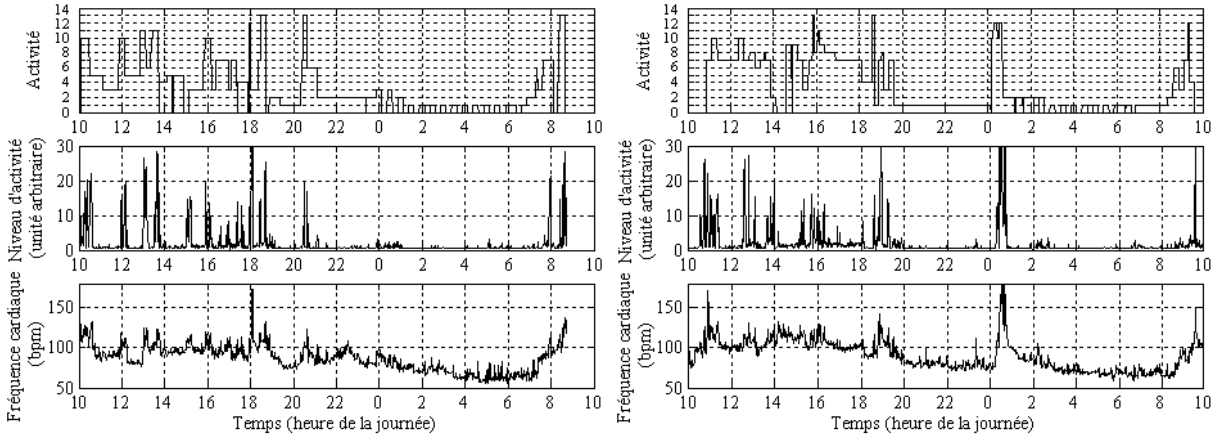


FIG. D.2 – Enregistrements du sujet 2.

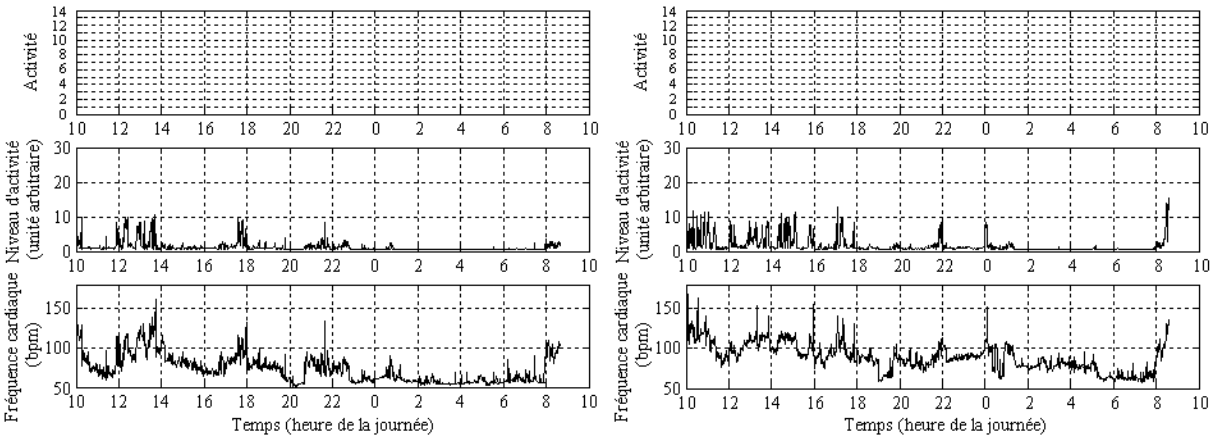


FIG. D.3 – Enregistrements du sujet 3.

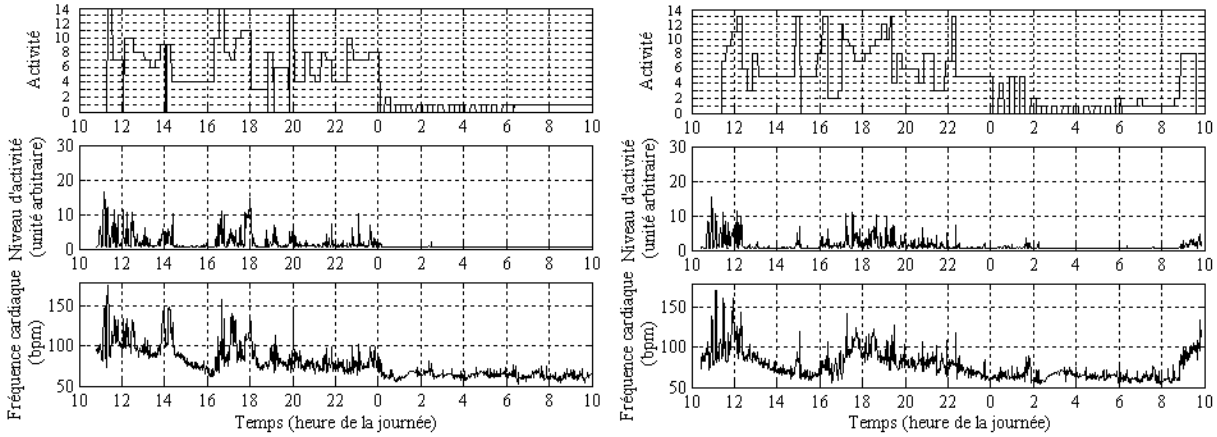


FIG. D.4 – Enregistrements du sujet 4.

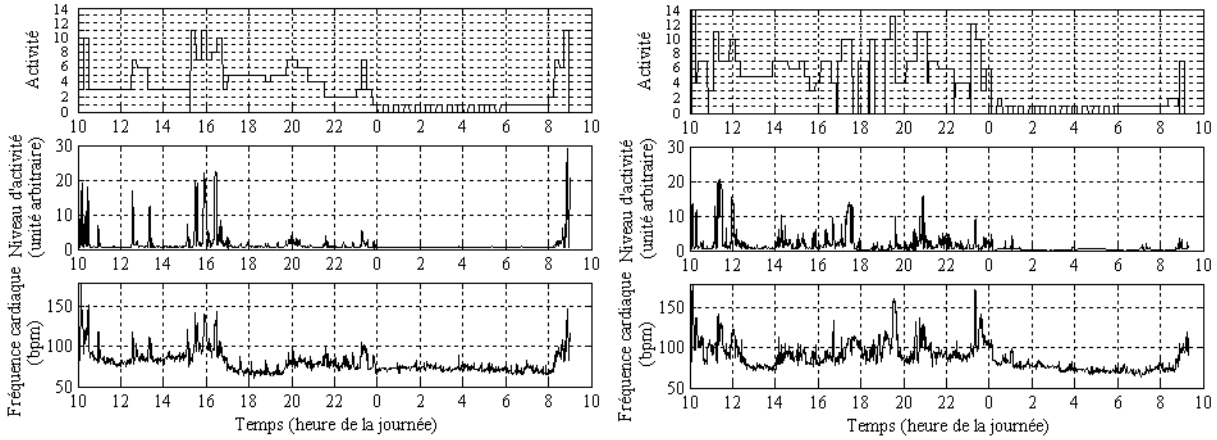


FIG. D.5 – Enregistrements du sujet 5.

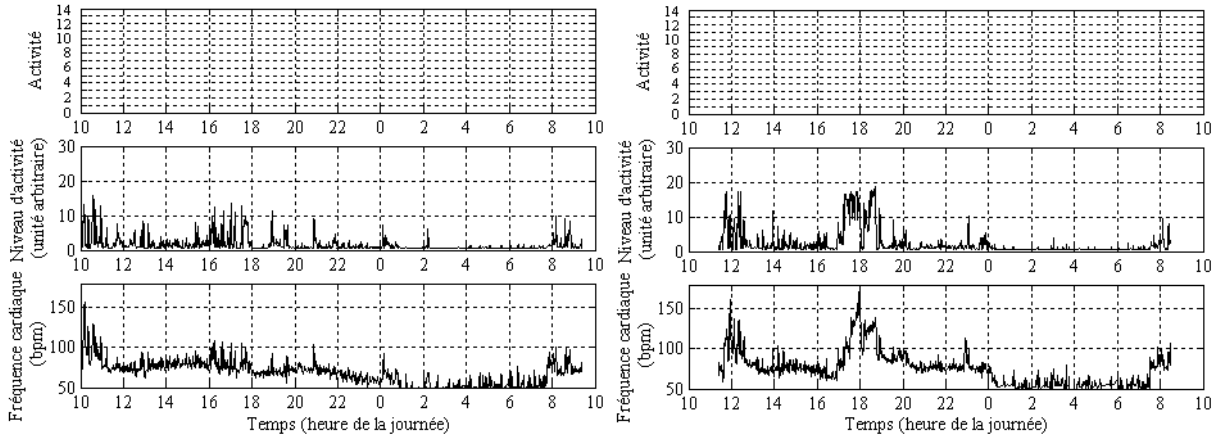


FIG. D.6 – Enregistrements du sujet 6.

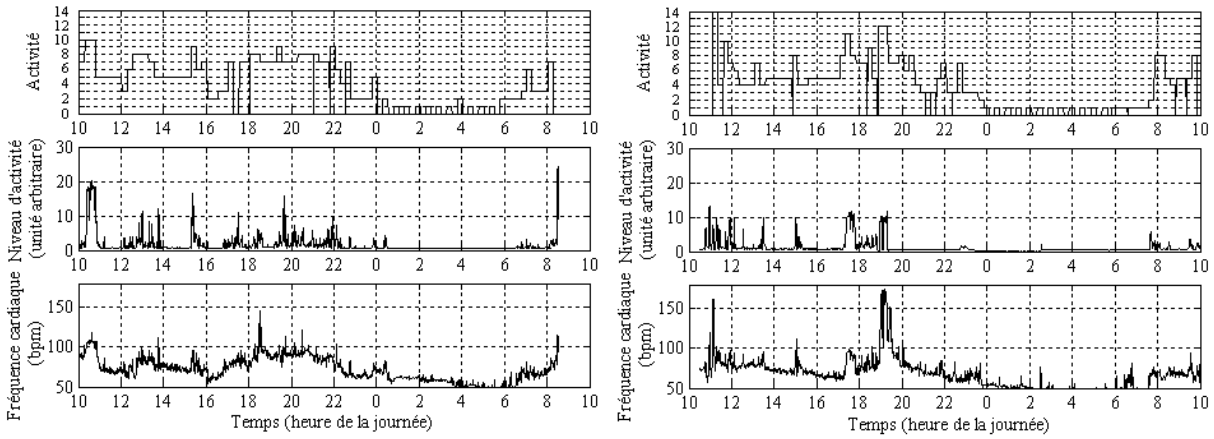


FIG. D.7 – Enregistrements du sujet 7.

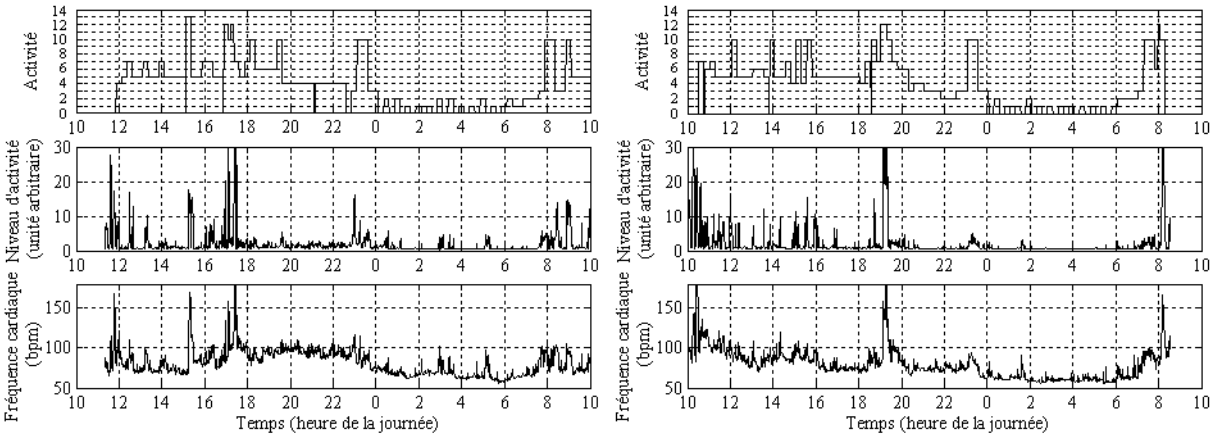


FIG. D.8 – Enregistrements du sujet 8.

E

Analyse des données de modélisation

Les données expérimentales dont on dispose correspondent à 8 sujets dont les activités quotidiennes ont été observées pendant deux journées non consécutives. Seule la moitié de ces données sont dédiées à la conception et à la validation du modèle de simulation – les enregistrements des 4 premiers sujets, appelées données de modélisation. Les enregistrements correspondant aux autres sujets sont utilisés pour la validation des séquences générées par le processus de simulation – données de validation. Afin de donner plus de force à la validation, il est important de ne pas intégrer les données issues d'un même sujet à la fois pour la modélisation et la validation des résultats produits par le modèle.

Par ailleurs, on ne considère pas les données des sujets 3 et 6, dont les activités ne sont pas annotées, dans les analyses qui nécessitent cette information.

Enfin, pour se rapprocher du contexte de la télésurveillance médicale à domicile, on ne considère que les données correspondant à des activités d'intensité faible à modérée, et en supposant que la personne habite dans une maison de plein pied, c'est-à-dire : dormir, être allongé, être assis calmement, être assis et discuter, être assis et travailler, manger, être debout, être debout et travailler et marcher doucement.

E.1 Niveau d'activité

E.1.1 Caractéristiques relatives des distributions

Les distributions expérimentales observées pour le niveau d'activité en fonction du type de mouvement – en posture allongée, assise, debout, ou en marchant – sont présentées sur les figures E.1 à E.4. Ces distributions correspondent à un mélange d'une distribution normale pour les bas niveaux d'activité, et d'une distribution exponentielle pour les hauts niveaux. Elles sont présentées toutes à la même échelle pour mettre en évidence la variabilité des niveaux d'activité observés due à la posture.

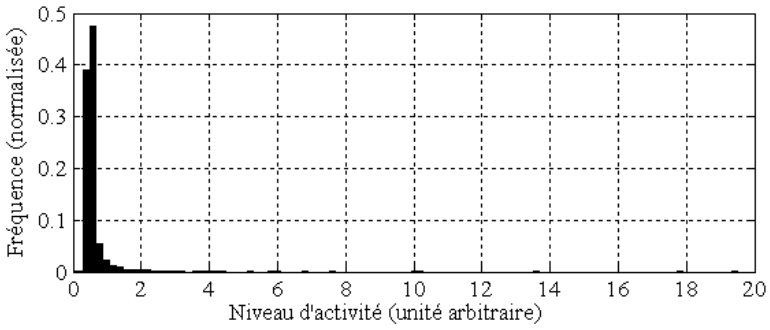


FIG. E.1 – Distribution des valeurs de niveau d'activité observées alors que la personne effectue des mouvements en posture allongée.

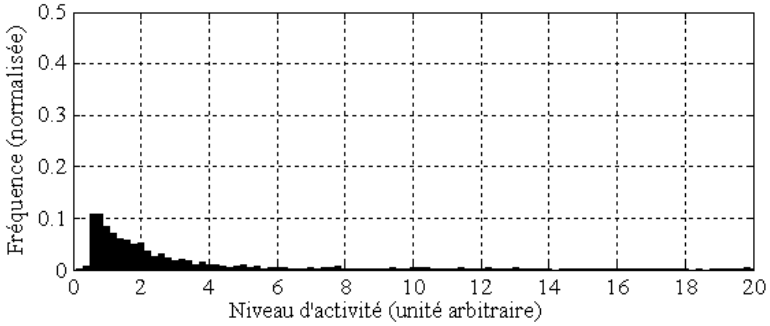


FIG. E.2 – Distribution des valeurs de niveau d'activité observées alors que la personne effectue des mouvements en posture assise.

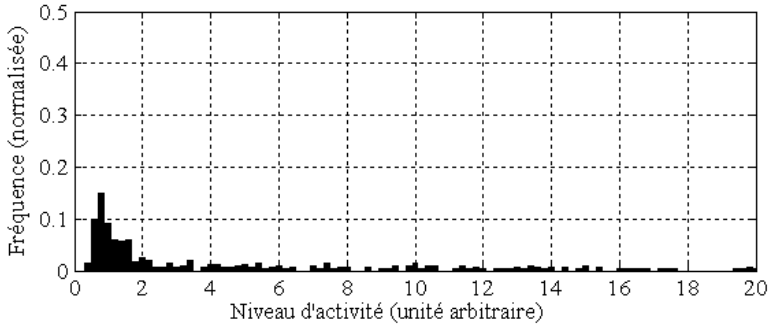


FIG. E.3 – Distribution des valeurs de niveau d'activité observées alors que la personne effectue des mouvements en posture debout.

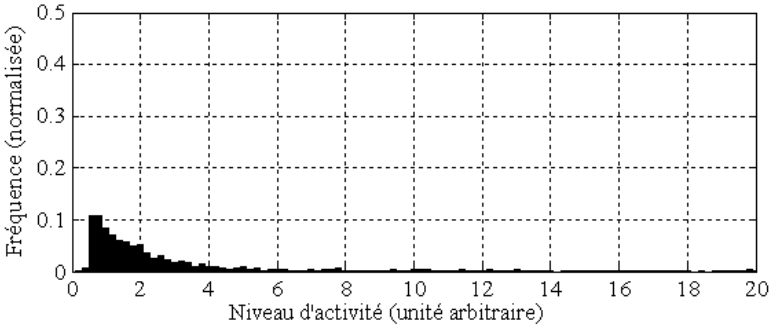


FIG. E.4 – Distribution des valeurs de niveau d'activité observées alors que la personne marche.

E.1.2 Caractéristiques intrinsèques des distributions

Les figures E.5 à E.8 représentent sur le graphe du haut la distribution expérimentale, et sur celui du bas un exemple de distribution simulée après estimation des paramètres des lois normale et exponentielle. Ces distributions sont toutes deux à deux similaires d'après le test de Kolmogorov-Smirnov. Elles sont présentées à des échelles spécifiques suivant le type de posture, selon la représentation qui met le plus en évidence les caractéristiques de chacune de ses distributions.

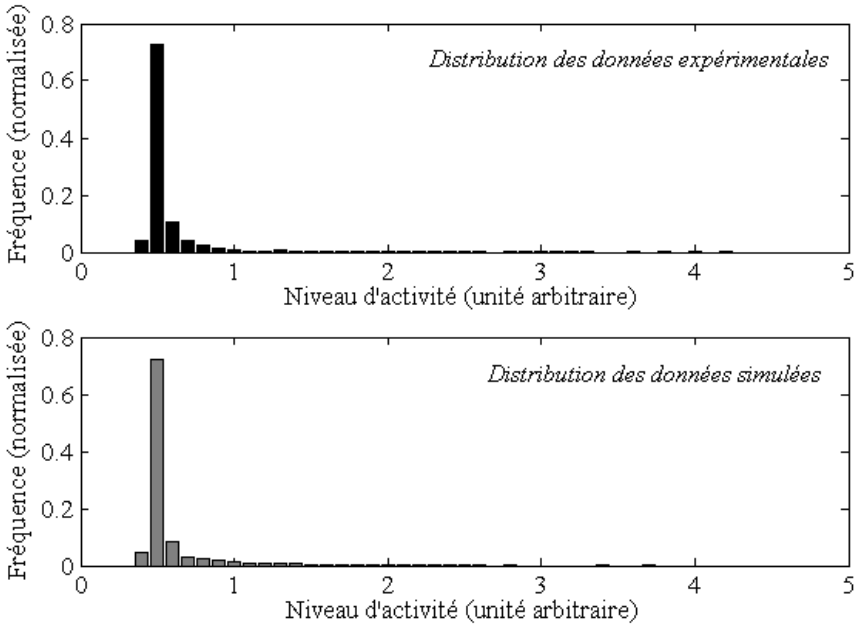


FIG. E.5 – Distribution des valeurs de niveau d'activité observées alors que la personne effectue des mouvements en posture allongée.

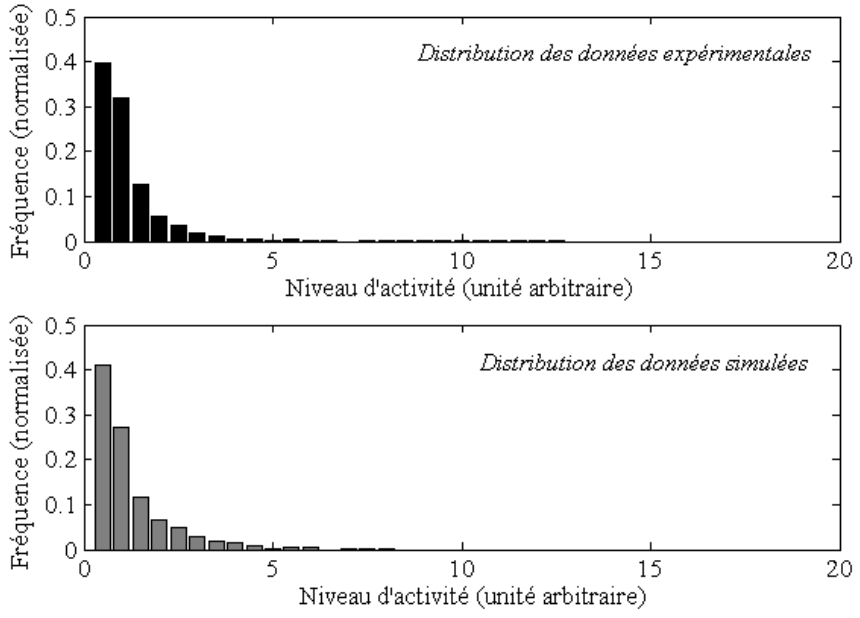


FIG. E.6 – Distribution des valeurs de niveau d'activité observées alors que la personne effectue des mouvements en posture assise.

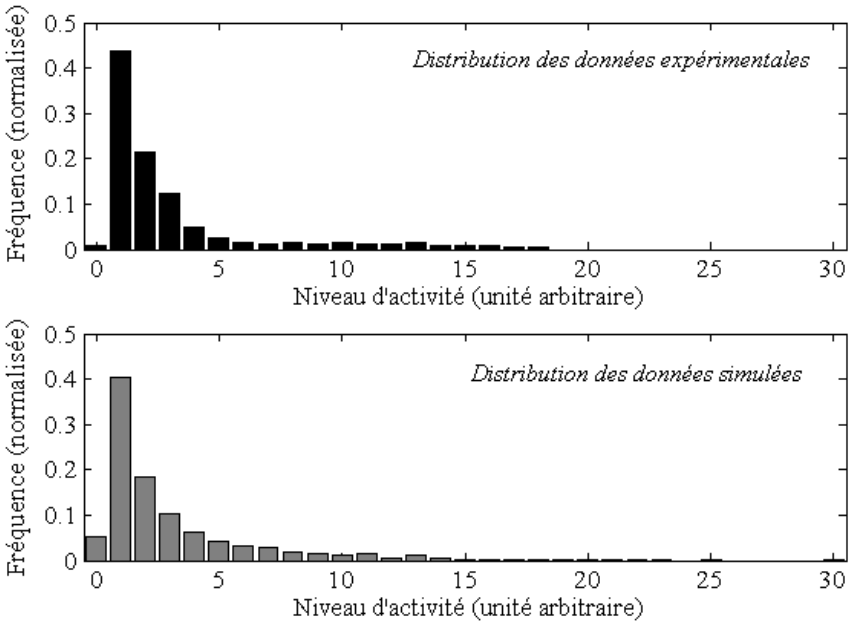


FIG. E.7 – Distribution des valeurs de niveau d'activité observées alors que la personne effectue des mouvements en posture debout.

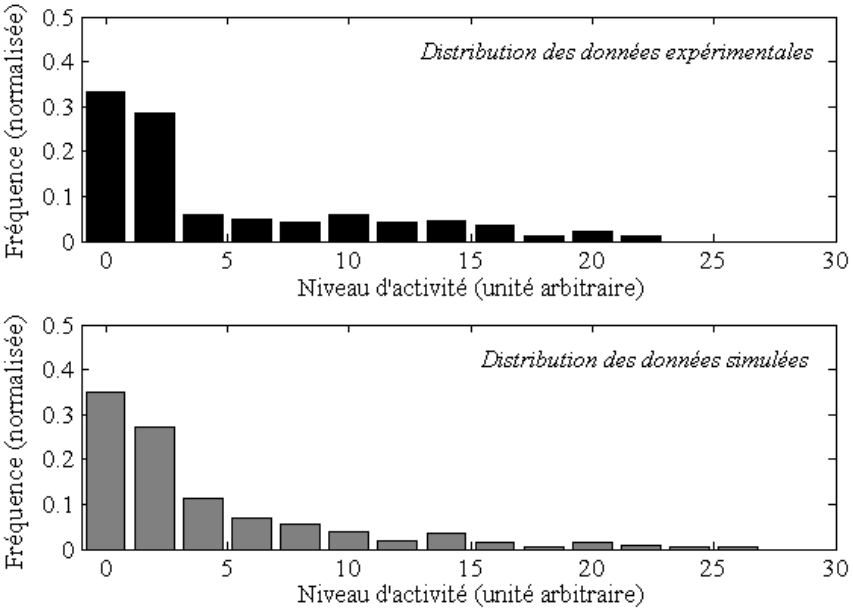


FIG. E.8 – Distribution des valeurs de niveau d'activité observées alors que la personne marche.

E.2 Fréquence cardiaque de repos

Le rythme circadien de la fréquence cardiaque est estimé pour chaque sujet par la méthode du cosinor, qui fait l'hypothèse de variations de repos sinusoïdales. La fréquence cardiaque de repos est alors définie pour chaque sujet pour une équation de type (E.1), où t est le temps exprimé en heures.

$$Fc_{repos}(t) = M + A \cdot sin\left(\frac{2\pi}{24}t + \phi\right). \tag{E.1}$$

On estime ainsi les paramètres suivants :

(1) ***MESOR,*** M *– Midline Estimating Statistic Of Rythm*, c'est-à-dire le niveau moyen autour duquel les valeurs oscillent ;

(2) **Amplitude,** A – Mesure de l'étendue des variations possibles dans les valeurs ;

(3) **Acrophase,** $A\phi$ – Moment de la journée où l'approximation sinusoïdale du rythme atteint la valeur maximale, où ϕ est la phase des variations, en unités trigonométriques.

L'estimation de la "meilleure sinusoïde" – qui induit l'erreur minimale au sens des moindres carrés – est réalisée à partir des données enregistrées pour chaque sujet au cours de 24 heures et qui correspondent à de faibles niveaux d'activité – inférieurs à 0.6. On ne choisit pas un seuil plus faible afin de disposer de suffisamment de données pour l'estimation. Par ailleurs, on prend pour référence la fréquence cardiaque de repos en position allongée. Comme on ne dispose pas de valeurs de fréquence cardiaque au repos allongé tout au long de la journée, on divise les valeurs observées pour de faibles niveaux d'activité en position assise par 1.1, et en position debout par 1.25, pour tenir compte du coût cardiaque de la posture (voir paragraphe 4.3.1). Des fréquences cardiaques supérieurs à de réelles valeurs de repos sont alors peut-être prises en compte. Ainsi, la succession des valeurs prises en compte est lissée en calculant à chaque instant – toutes les minutes – la moyenne des fréquences cardiaques enregistrées sur une fenêtre glissante de 30 minutes autour de ce point. La meilleure sinusoïde correspondant à ces données est estimée en faisant varier respectivement le MESOR et l'amplitude autour de la moyenne et l'amplitude des variations observées.

Le tableau E.1 présente les valeurs de paramètres obtenus pour l'ensemble des sujets ayant annoté leurs activités, la période des variations étant de 24 heures. On considère à la fois les données de modélisation et de validation pour l'estimation des rythmes circadiens car toutes ces données ont besoin d'être normalisées en fonction des variations de repos – estimation du coût cardiaque. Par ailleurs on réalise pour chaque sujet deux estimations des caractéristiques des variations de repos : les enregistrements n'étant pas consécutifs, il est en effet possible que les rythmes biologiques des sujets se soient décalés entre les deux expérimentations – un rythme biologique peut se décaler d'au plus 2 heures par jour.

Sujet n°	**Jour n°**	**MESOR (bpm)**	**Amplitude (bpm)**	**Acrophase (heure)**
1	1	77.8	9.5	00h00
	2	86	11.4	03h22
2	1	75.7	9.8	18h16
	2	76.5	6.4	19h02
4	1	64.3	1.5	17h53
	2	61.8	1.5	12h55
5	1	69.3	3.0	04h54
	2	75.2	4.9	22h05
7	1	59.8	5.6	17h53
	2	57.0	11.2	17h07
8	1	66.3	2.5	19h25
	2	66.0	5.5	16h00
Moyenne		**69.6**	**6.1**	**16h00**
Référence[46, 74]		**65**	[**0.7**,**13.9**]	[**10h40**,**19h30**]

Tab. E.1 – Paramètres des variations sinusoïdales de repos des sujets de l'expérimentation.

E.3 Coût cardiaque d'une activité

L'analyse du coût cardiaque de l'activité, en fonction du niveau d'activité et pour chaque type de posture, est réalisée en moyenne pour l'ensemble des sujets. On constate principalement qu'il existe bien en moyenne une relation linéaire entre le coût cardiaque d'une activité et le logarithme de la moyenne du niveau d'activité observé pendant les deux minutes précédentes, pour chaque type de posture. On vérifie cependant que cette relation linéaire est bien observée pour tous les sujets, avec pour un même niveau d'activité un coût cardiaque d'autant plus élevé que l'effort requis par la posture est important. Les résultats obtenus pour les trois sujets expérimentaux dont les données sont exploitées pour la modélisation – sujets 1, 2 et 4 – sont présentés sur la figure E.9.

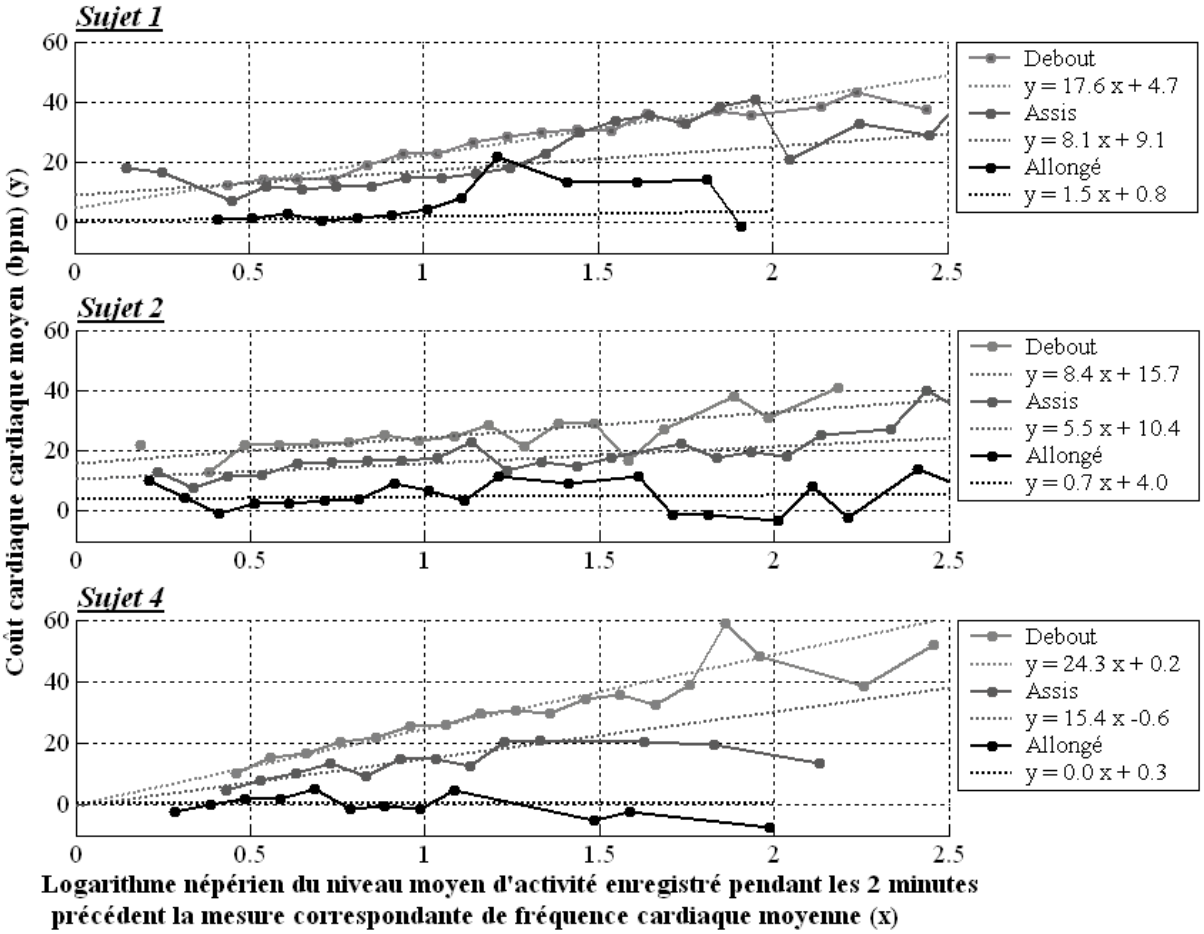


FIG. E.9 – Moyenne du coût cardiaque en fonction des niveaux d'activité observés, pour chaque type de posture.
Les points représentent les moyennes calculées sur les données expérimentales, et les droites en pointillés la régression linéaire effectuée sur tout ou partie de ces points, suivant leur pertinence. De haut en bas, les graphes correspondent respectivement aux sujets 1, 2 et 4.

E.4 Influence de l'activité végétative sur la fréquence cardiaque

Certains sujets ayant annoté les périodes d'alimentation, on est en mesure d'étudier leur influence sur la fréquence cardiaque pendant les trois heures d'activité végétative qui suivent la prise d'un repas. D'après [74], la fréquence cardiaque augmente pendant ces périodes. La figure E.10 présente les données expérimentales enregistrées pour le premier sujet pendant la deuxième journée de surveillance. Elle met en évidence d'une part les périodes d'activité végétative – inférées à partir des annotations du sujet sur ses activités – et d'autre part le coût cardiaque résiduel une fois soustraits aux variations de la fréquence cardiaque (1) le rythme de repos et (2) le coût cardiaque de la posture et du niveau d'activité.

Le graphe présenté montre qu'il n'est pas possible de mettre en évidence une quelconque augmentation de la fréquence cardiaque pendant l'activité végétative. Il s'agit probablement d'une influence trop légère pour qu'elle puisse être observée dans notre contexte, compte tenu de l'ensemble des autres facteurs d'influence non explicités et de l'imprécision de l'estimation des variations de repos et du coût cardiaque des activités et postures.

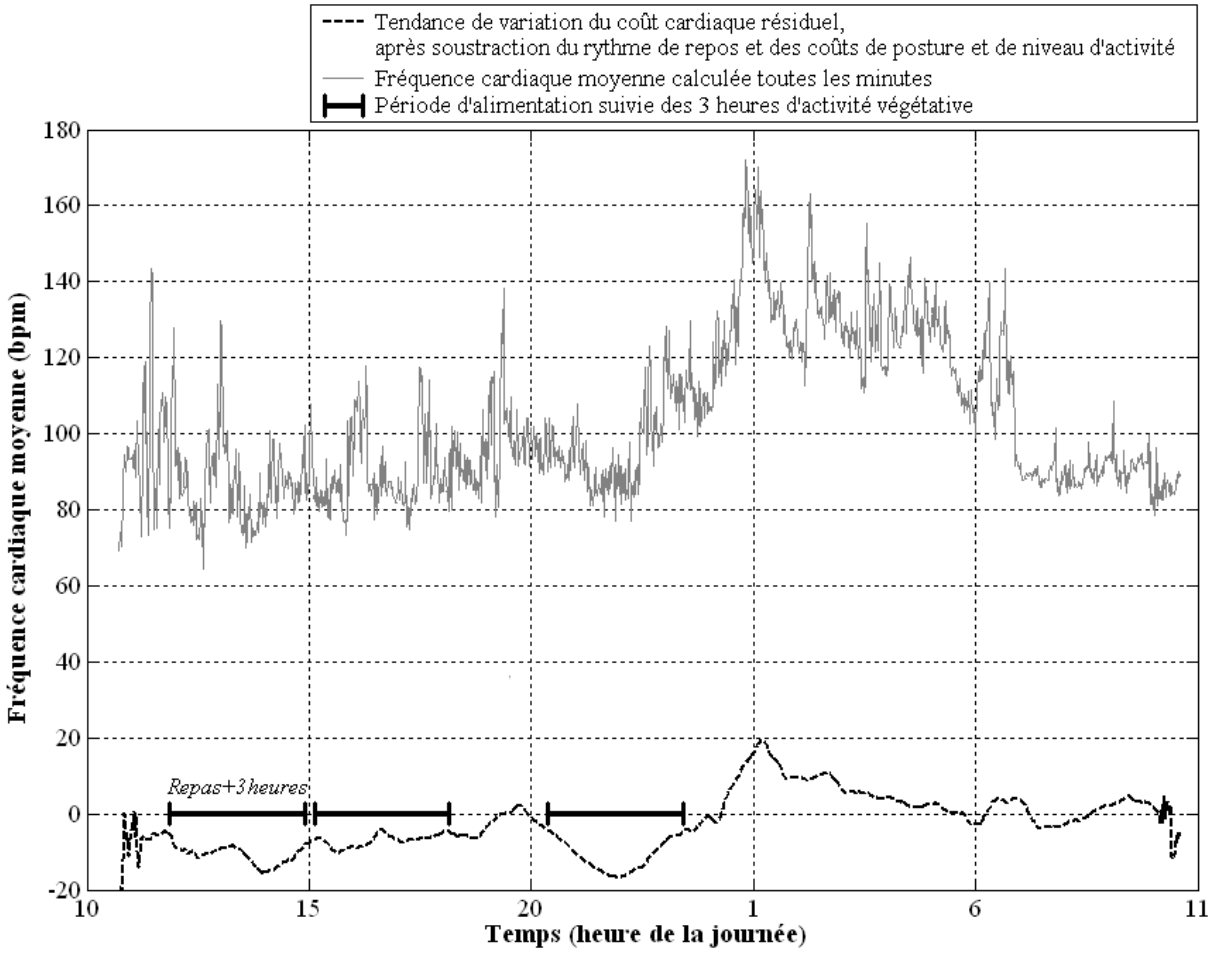


FIG. E.10 – Observation des tendances de variation du coût cardiaque en fonction des périodes d'alimentation sur la deuxième journée de surveillance du premier sujet de l'expérimentation.

F

Implémentation du modèle de simulation

L'implémentation du modèle proposé pour la simulation de paramètres enregistrés à domicile – déplacements, postures, niveau d'activité et fréquence cardiaque – est réalisée avec le logiciel `MATLAB`. Les paragraphes suivants présentent (1) le principe de l'implémentation, (2) l'interface et l'implémentation de la simulation, (3) la structure d'un fichier de paramétrisation et enfin (4) la structure globale du programme de simulation.

F.1 Principe d'implémentation

Étant donnée la complexité du modèle au niveau du nombre de paramètres qui doivent être définis *a priori*, on utilise un ensemble de fichiers au format "texte" pour la définition des valeurs courantes utilisées pour la simulation et des valeurs par défaut. Une interface graphique permet la modification de ces paramètres, et ainsi lit et écrit dans ces fichiers. Le processus de simulation, une fois lancé, peut alors récupérer la valeur de chaque paramètre dans ces fichiers. À la fin de la simulation, les données générées sont sauvegardées dans un fichier de données de `MATLAB` (fichier '.dat'), et éventuellement au format "texte", un fichier par paramètre simulé.

La figure F.1 présente un schéma synthétisant le principe général de l'implémentation du processus de simulation.

F.2 Interface et implémentation de la simulation

Le processus de simulation est composé de deux interfaces qui concernent d'une part la simulation d'un individu donné dans une certaine situation, et d'autre part la simulation de modifications "normales" de comportement.

F.2.1 Simulation d'un individu donné dans une certaine situation

Une interface graphique permet d'agir "facilement" sur le grand nombre de paramètres de la simulation. Cette interface présente *a priori* des valeurs par défaut définies soit intuitivement – éventuellement après tâtonnement jusqu'à la génération de séquences qui semblent réalistes, soit à partir des résultats d'analyses sur les données expérimentales. Ces valeurs peuvent être modifiées pour la génération de plusieurs profils de personnes et types de situation, puis sauvegardées, et à n'importe quel moment réinitialisées aux valeurs par défaut.

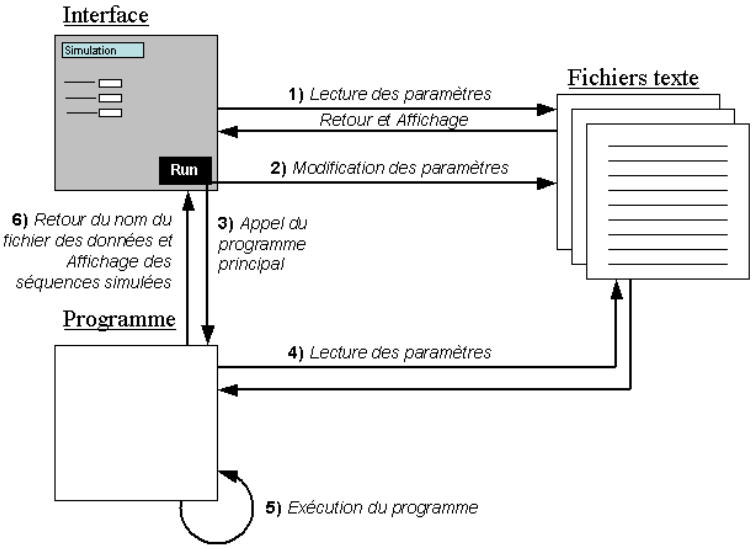


FIG. F.1 – Principe de l'implémentation du processus de simulation.

Cette interface est présentée sur la figure F.2. Les paramètres sont classés en fonction du sous-modèle de simulation sur lequel ils agissent. Certains paramètres complexes, tels que les probabilités de transition entre les différentes postures selon la pièce considérée et le moment de la journée, doivent être entrés directement dans un fichier "texte", qu'il est possible d'éditer à partir de l'interface graphique (voir paragraphe F.3).

F.2.2 Simulation de modifications "normales" de comportement

Description de l'interface

Une interface graphique permet d'introduire du bruit dans la réalisation de certaines activités de la vie quotidienne, correspondant à des modifications "normales" de comportement d'une personne. Les sous-séquences représentatives initialement de ces activités sont appelées les *motifs*. L'idée de cet utilitaire est d'introduire facilement des instances bruitées de ces motifs dans des séquences de données générées par le processus de simulation à partir d'une séquence de déplacements aléatoires. L'objectif est alors d'expérimenter l'extraction des instances récurrentes de ces motifs, en étant capable d'évaluer les performances du système puisqu'on connaît *a priori* les motifs présents et la localisation de leurs instances.

La figure F.3 présente cette interface pour l'introduction de motifs bruités dans une séquence de données. Les paramètres de l'introduction des bruits sont classés selon le type de bruit qu'ils concernent : déformations temporelles, variabilité dans les valeurs, ou interruptions. À l'issu du processus, il est possible de visualiser les séquences générées (dont les données sont sauvegardées dans un fichier de données `MATLAB`), les motifs considérés et les indices des instances introduites.

Principe d'introduction de bruit sur l'instance d'un motif

On présente dans ce paragraphe le détail de l'implémentation des différents types de bruit possibles sur l'instance d'un motif : déformation temporelle, variabilité dans les valeurs, ou interruption. Leur définition est détaillée dans le paragraphe II.6.5.2.

Dans les descriptions qui suivent, les variables *rand* et *randn* correspondent à des fonctions génératrices d'un nombre aléatoire selon une distribution respectivement *uniforme* sur $[0, 1]$, et *normale*, de moyenne nulle et d'écart-type égal à 1. La variable *round* est par ailleurs une fonction d'arrondi à l'entier le plus proche, $floor$ étant l'arrondi à l'entier inférieur. Toutes les durées sont exprimées en nombre de points d'une sous-séquence.

Interruption

On choisit les notations suivantes pour les paramètres définissant une interruption :

- **Durée maximum d'une interruption,** comme un rapport de la durée totale de la sous-séquence correspondant à la réalisation de la tâche principale. C'est ainsi un réel compris entre 0 et 1, noté r_{int}.
- **Nombre d'interruptions,** si on considère qu'une activité principale peut-être interrompue plusieurs fois. C'est ainsi un entier positif ou nul.

Le principe d'insertion d'interruptions comprend alors les étapes suivantes :

- **Détermination des durées d'interruption,** notée d_{int}, aléatoirement selon une distribution uniforme en fonction de la durée de la séquence principale considérée (d_{act}) et du taux définissant la durée maximum d'une interruption :

$$d_{int} = round(r_{int} \times d_{act} \times rand).$$

- **Détermination des instants d'interruption,** notés t_{int}, aléatoirement selon une distribution uniforme en fonction de la durée de l'activité principale (d_{act}) :

$$t_{int} = round(d_{act} \times rand).$$

  On s'assure par ailleurs que les interruptions ne sont ni au début ni à la fin de la séquence représentant l'activité principale, et qu'elles sont toutes disjointes dans le temps.
- **Simulation de l'interruption.** On suppose que l'interruption d'une activité mène la personne dans une pièce différente de celle(s) de l'activité principale avant et après l'interruption. Cette pièce est sélectionnée aléatoirement selon une loi uniforme fonction du nombre des autres pièces du domicile. La séquence de données représentative de l'interruption est alors générée par le processus de simulation à partir des données de déplacement correspondant à la présence de la personne dans la pièce de l'interruption à l'instant et pendant la durée de celle-ci.

Déformation temporelle

On choisit les notations suivantes pour les paramètres définissant une déformation temporelle :

- **Taux de variation de la durée de réalisation de l'activité principale,** défini par un réel noté r_{temp}, positif ou négatif selon qu'il s'agit respectivement d'un étirement ou d'une compression dans le temps de la durée de réalisation de l'activité principale.
- **Durée minimale d'une activité dite principale,** définie par un entier positif ou nul, afin de ne pas compresser à un extrême non significatif la réalisation d'une activité.

Le principe de déformation temporelle comprend alors les étapes suivantes :

- **Détermination de la nouvelle durée de l'activité**, notée d'_{act}, en fonction de la durée initiale (d_{act}) à laquelle on ajoute une variation de durée (positive ou négative) déterminée aléatoirement selon une distribution gaussienne. Toutes le durées sont exprimées ici en nombre de points des séquences, et correspondent ainsi à des valeurs entières. La moyenne

d'une variation de durée est considérée nulle et l'écart-type est égal à la durée calculée selon le taux de variation r_{temp} à partir de la durée de référence de l'activité :

$$d_{act}^{'} = round(d_{act} + r_{temp} \times d_{act} \times randn).$$

La nouvelle durée d'une activité est par ailleurs bornée en limite inférieure par la durée minimale d'une activité principale.

- **Détermination du nombre de points à ajouter ou supprimer,** noté nb_{chgt}, défini comme un relatif positif ou négatif selon la variation de durée par rapport à la durée de référence de l'activité :

$$nb_{chgt} = d_{act}^{'} - d_{act}.$$

- **Détermination des instants d'ajout ou de suppression,** notés t_{chgt}, définis aléatoirement selon une loi uniforme parmi les instants de définition la séquence de référence :

$$t_{chgt} = round(rand \times d_{act}).$$

Afin de préserver la continuité de l'activité réalisée, on s'assure par ailleurs de ne pas ajouter ou supprimer plusieurs points successifs.

- **Introduction des déformations,** en ajoutant ou supprimant des points aux instants sélectionnés. Chaque point ajouté correspond à la valeur moyenne des deux points entre lesquels il s'insère.

Variabilité dans les valeurs

On note r_{bruit} le paramètre définissant le taux de variabilité possible des valeurs observées. Ce taux est interprété différemment pour chaque type de paramètre, selon qu'il est qualitatif ou quantitatif, et toujours de manière à préserver la tendance globale des variations.
Le principe d'introduction de variabilité dans les valeurs (ou *bruit moyenne fréquence*) est présenté en deux étapes, concernant l'introduction de bruit dans les valeurs des paramètres quantitatifs d'une part, et qualitatifs d'autre part. Pour ce qui concerne les paramètres qualitatifs dans notre contexte, seules les successions de postures peuvent être modifiées à ce niveau, un changement dans les déplacements observés au cours d'une activité est en effet considéré comme une interruption. On modifie par ailleurs en conséquence les valeurs de fréquence cardiaque, puisqu'une partie du coût cardiaque d'une activité est imputé directement à la posture.

- **Paramètres quantitatifs : niveau d'activité et fréquence cardiaque**
  Le principe d'introduction de variabilité dans les valeurs est le suivant :
  - **Détermination de l'écart-type "bruité",** noté σ_{bruit}, définissant la variation possible des valeurs suivant le taux de variabilité prédéfini, r_{bruit}, par rapport à l'écart-type observé sur la distribution des valeurs représentatives de l'activité, noté σ :

$$\sigma_{bruit} = r_{bruit} \times \sigma.$$

  - **Introduction de bruit dans les valeurs,** selon une distribution normale de moyenne nulle et d'écart-type égal à l'écart-type bruité. Pour chaque valeur v, on définit alors une nouvelle valeur $v^{'}$ telle que :

$$v^{'} = v + randn \times \sigma_{bruit}.$$

  - **Réorganisation temporelle des valeurs,** selon le principe général de continuité physique, afin de retrouver et préserver les tendance globales de variation.

- **Paramètres qualitatifs : posture**.
  Le principe d'introduction de variabilité dans les valeurs est le suivant :
  - **Détermination des intervalles à considérer pour l'introduction de bruit**. Ce sont les intervalles de temps sur lesquels la valeur du paramètre correspond à la même modalité, et dont la durée (notée d_{obs}) est supérieure à la durée moyenne pendant laquelle cette modalité est habituellement maintenue (notée d_{mod}) pour l'ajout d'alternances sur de longues périodes de maintien d'une même posture, ou inférieure à cette durée pour la suppression de courtes alternances :

    $$d_{obs} >= d_{mod} \text{ ou } d_{obs} < d_{mod}.$$

  - **Détermination du nombre de changements de modalité possibles,** en cas d'introduction d'une alternance supplémentaire – cas où $d_{obs} >= d_{mod}$. Ce nombre est noté nb_{chgt} et est égal au rapport de la durée effective de maintien de la modalité sur la durée moyenne observée :

    $$nb_{chgt} = floor(\frac{d_{obs}}{d_{mod}}).$$

  - **Définition des changements de modalité effectifs**, en fonction du taux de bruit r_{bruit}. Pour chaque changement possible, on sélectionne aléatoirement un nombre selon une distribution uniforme sur [0,1] et on détermine si le changement de modalité (ajout ou suppression) est effectif selon la condition suivante :

    $$rand < r_{bruit}.$$

    Puis, dans le cas où un alternance supplémentaire est effectivement introduite ($d_{obs} >= d_{mod}$), on réalise les deux étapes suivantes :
  - **Détermination de l'instant d'occurrence d'un changement**, en cas d'introduction d'une alternance supplémentaire. Cet instant est noté t_{chgt} et est déterminé aléatoirement selon une loi uniforme en fonction de la durée de l'intervalle considéré (d_{mod}), débutant à l'instant t_{ini} :

    $$t_{chgt} = t_{ini} + round(d_{mod} \times rand).$$

  - **Détermination de la durée du changement**, notée d_{chgt}, aléatoirement selon une loi uniforme en fonction de la durée de l'intervalle considéré (d_{mod}) et du taux de bruit autorisé (r_{bruit}) :

    $$d_{chgt} = round(r_{bruit} \times d_{mod} \times rand).$$

    Enfin, dans tous les cas :
  - **Définition des nouvelles modalités.** Les changements de posture sont définis vers la posture "debout" lorsque la personne est assise, et réciproquement. Un changement en posture allongée est par contre considéré comme une interruption puisqu'elle correspond le plus souvent à de longues périodes de sommeil.
  - **Prise en compte du coût cardiaque de la posture.** Lorsque la posture est modifiée, on modifie en conséquence les valeurs de fréquence cardiaque pour tenir compte d'un coût cardiaque différent requis par la nouvelle posture (voir paragraphe 5.2.4).

F.3 Structure des fichiers de paramétrisation

On associe un ou plusieurs fichiers à l'implémentation de chaque sous-modèle de simulation, contenant la définition des paramètres qu'il met en jeu. Chaque fichier contenant une fonction

"clé", correspondant à une partie majeure du processus de simulation, lit dans un fichier "texte" les paramètres qu'elle met en jeu. Le nom de ce fichier est connu automatiquement par une syntaxe prédéfinie contenant le nom de la fonction considérée. Ces fichiers "texte" sont la plupart du temps directement gérés par les interfaces des différents processus. Les valeurs des paramètres sont spécifiées dans un format "type XML", tel que présenté sur l'exemple ci-dessous de définition des paramètres de variation de la fréquence cardiaque de repos.

```
% ===== Variations de la fréquence cardiaque de repos =====
% Période d'échantillonnage des données simulées [1]
<peMinFc>1</peMinFc>
% Paramètres de la sinusoïde définissant les variations de repos
% - période en heures [24]
<p>24</p>
% - acrophase : heure du maximum de la sinusoïde (HH:MM) [16:00]
<sphi>16:00</sphi>
% - mesor (m) :  fréquence cardiaque moyenne, en bpm [70]
<m>70</m>
% - amplitude (a), en bpm [6]
<a>6</a>
```

Certains autres fichiers "texte" sont utilisés pour la définition de paramètres complexes. Ci-dessous l'exemple d'un fichier de définition des probabilités de transition entre les postures dans la cuisine, en fonction des périodes de la journée.

```
% Pièce : Cuisine
% Positions : Debout - Assis - Allongé
% Période 1: Nuit
0.5 0.5 0.0
0.5 0.5 0.0
0.0 0.1 0.9
% Période 2: Lever
0.5 0.5 0.0
0.5 0.5 0.0
0.0 0.9 0.1
% Période 3: Matinée
0.4 0.6 0.0
0.3 0.7 0.0
0.0 0.9 0.1
...
```

Les séquences de déplacements générées par le simulateur implémenté par G. Virone sont importé de la lecture d'un fichier "texte" au format XML, dont un exemple est présenté ci-dessous.

```
<ID>Florence</ID><sexe>f</sexe><age>28</age>
<commentaire>Conditions normales de la vie quotidienne</commentaire>
<date>01/02/2000</date><heure>07:38:00</heure><entrée>WC</entrée>
<date>01/02/2000</date><heure>07:46:00</heure><entrée>cuisine</entrée>
<date>01/02/2000</date><heure>08:22:00</heure><entrée>douche</entrée>
<date>01/02/2000</date><heure>08:32:00</heure><entrée>chambre</entrée> ...
```

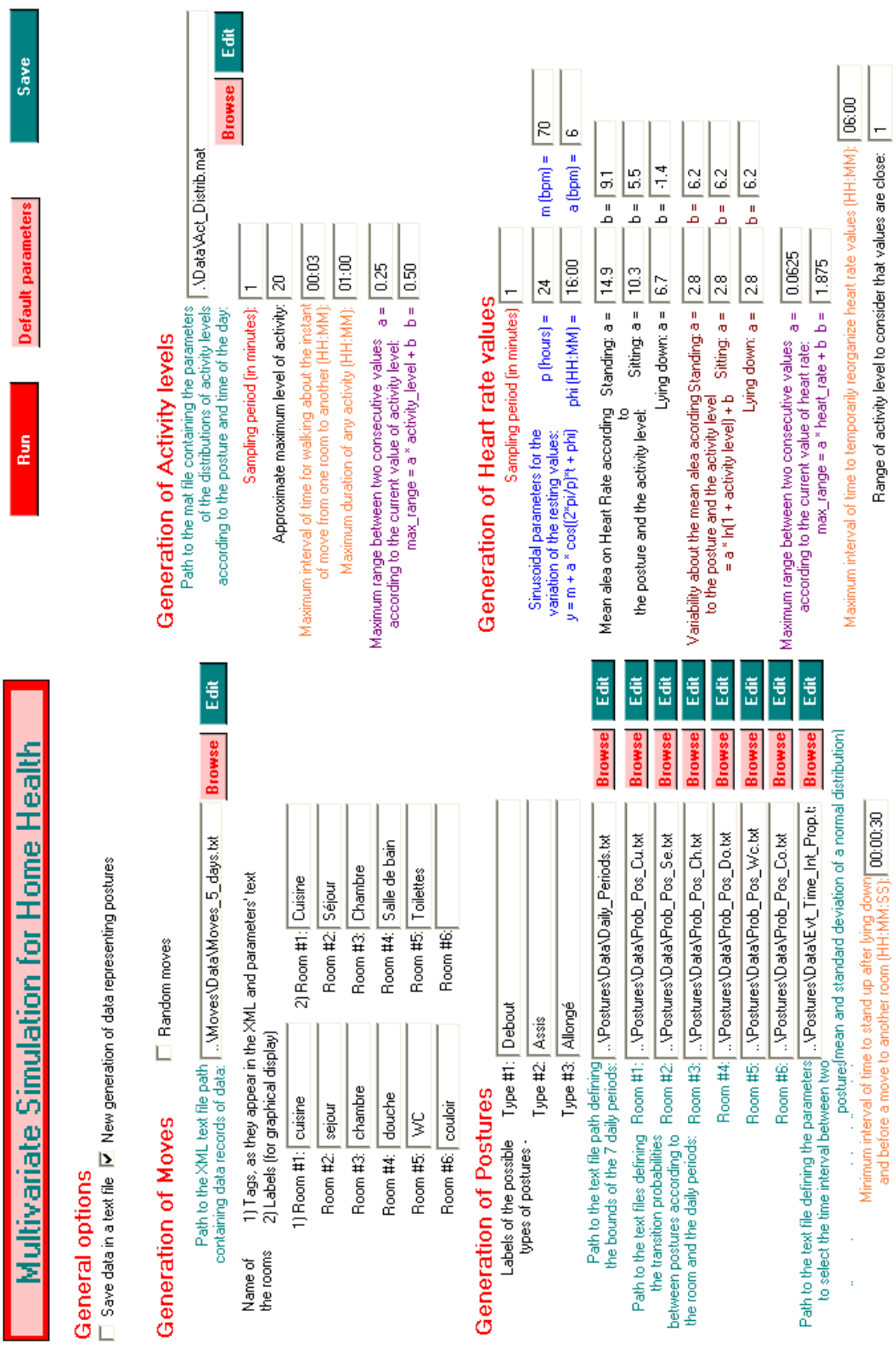


FIG. F.2 – Interface du processus de simulation.

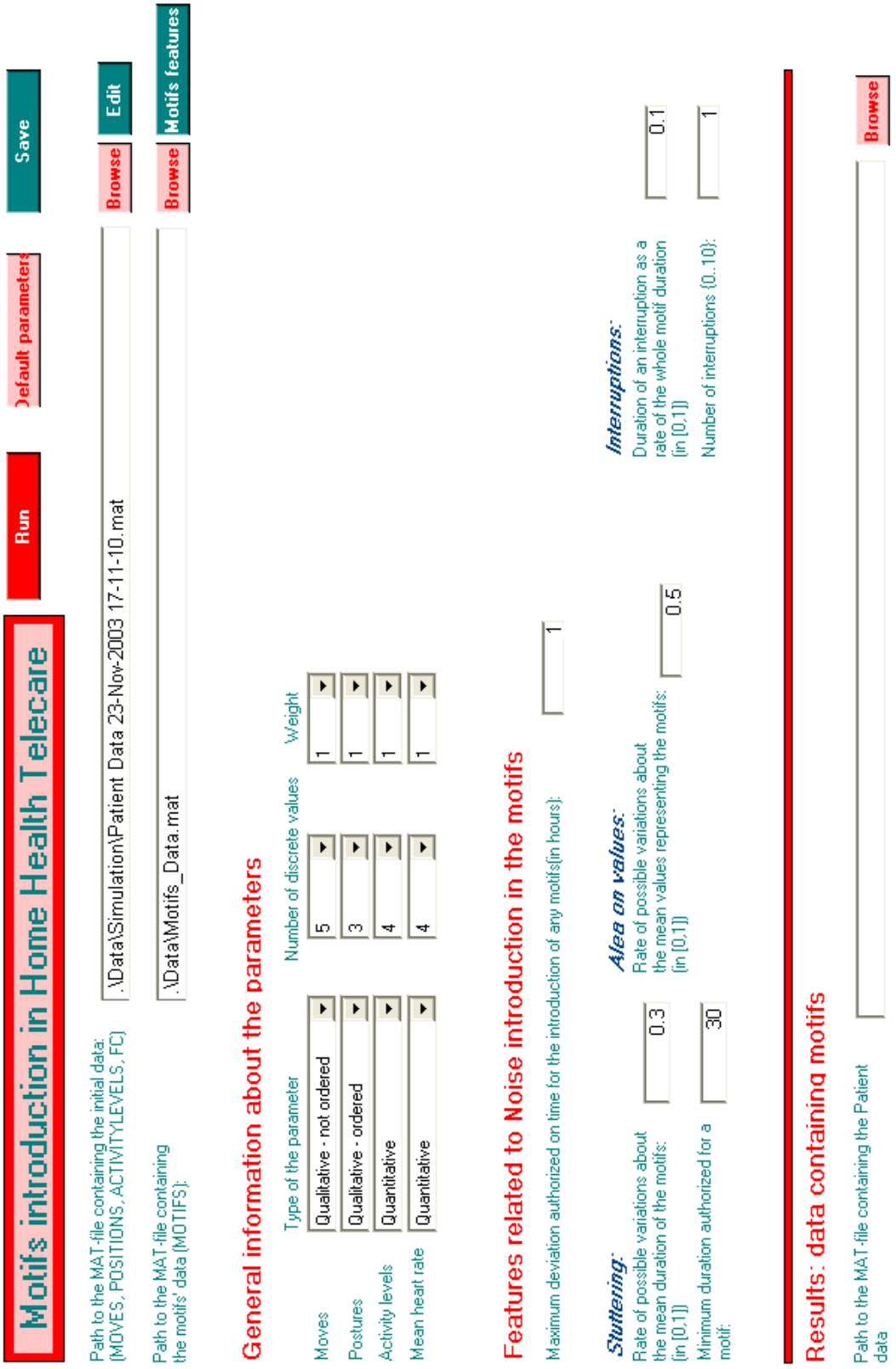


FIG. F.3 – Interface d'introduction de motifs bruités dans une séquence de données.

F.4 Structure globale du programme de simulation

La structure globale du programme de simulation est complètement *séquentielle*. Il s'agit d'appeler l'une après l'autre les fonctions réalisant les étapes "clés" de la simulation, et correspondant successivement à la génération des différents paramètres, dans l'ordre : déplacements, postures, niveau d'activité et fréquence cardiaque. Chaque fonction appelée prend en paramètre les résultats de l'appel à la fonction précédente, et fournit de la même manières les résultats de son exécution en paramètre de la fonction suivante.

Les grandes étapes de l'exécution de la simulation sont ainsi les suivantes :

```
% Fonction de la simulation (retourne le nom du fichier de sauvegarde)
function fileName = getPatientData()

% ==== Initialisation des variables globales (les paramètres simulés)
global MOVES = [];
global POSTURES = [];
global ACTIVITYLEVELS = [];
global FC = []; ...

% ==== Lecture des paramètres dans le fichier texte approprié
fidInput = fopen(strcat('Input_',mfilename,'.txt')); ...
TAGVALS = readFileInfosTags(fidInput, TAGNAMES); ...
fclose(fidInput);

% ==== Importation des séquences de déplacement simulées par G. Virone
importMoves;

% ==== Génération de postures appropriées
setPosture;

% ==== Génération de niveaux d'activité appropriées
setActLevels;

% ==== Génération des valeurs de repos de la fréquence cardiaque
setRestingFc;

% ==== Prise en compte du coût cardiaque des postures et niveaux d'activité
setAleasFc;

% ==== Sauvegarde des résultats de la simulation
fileName = ...
save(fileName,'PATIENT','MOVES','POSTURES','ACTIVITYLEVELS','FC',...);

% ==== Ecriture éventuelle des résultats sous format texte
writeTableInFile(MOVES,'MOVES',...); ...
```

G

Implémentation de l'extraction de motifs

L'implémentation du modèle proposé pour l'extraction de motifs est réalisée sous le logiciel `MATLAB`. Le principe général d'implémentation et de gestion des données et paramètres est le même que pour l'implémentation du processus de simulation et n'est donc pas détaillé de nouveau dans cette annexe (voir annexe F). Les paragraphes suivants présentent ainsi uniquement (1) l'interface de l'extraction de motifs et (2) la structure globale du programme d'extraction.

G.1 Interface de l'extraction de motifs

Le processus d'extraction de motifs comprend deux interfaces qui concernent d'une part l'extraction de motifs pour la définition des paramètres et l'exécution du système, et d'autre part l'analyse des résultats et des performances.

G.1.1 Définition des paramètres de l'extraction

Une interface graphique permet d'agir facilement sur les paramètres de l'extraction de motifs (voir Fig. G.1). Les paramètres sont présentés en quatre classes selon l'étape d'extraction sur laquelle ils agissent : (1) Informations générales sur les paramètres étudiés, (2) Mesure de distance $LCSS$, (3) Prétraitements et abstraction (filtrage, réduction temporelle, discrétisation, agrégation) et (4) Extraction des tentatives de motifs (projections aléatoires et examen de la matrice de collisions). Les dernières étapes de classification des tentatives de motifs en motifs et de détermination d'un représentant de chaque classe ne nécessitent par contre pas de paramètres spécifiques.

G.1.2 Analyse des résultats et performances du système

Une interface graphique permet d'analyser les résultats et performances du système une fois réalisée l'exécution du processus d'extraction (voir Fig. G.2). Ces résultats sont alors sauvegardés dans un fichier texte. L'interface proposée comporte plusieurs sections et sous-sections correspondant aux différentes étapes de l'extraction.
La première section concerne les étapes du **processus d'extraction**, incluant :

- **Des informations générales** sur le déroulement du processus (caractéristiques et taille des données analysées, nombre de paramètres considérés, temps d'exécution, etc.) ;
- **Le prétraitement** et l'abstraction des données brutes, avec notamment la possibilité de visualiser les séquences à l'issue de chaque étape ;

- **L'extraction des tentatives de motifs,** incluant la possibilité de visualiser l'ensemble des sous-séquences de base, la matrice de collisions, les tentatives de motifs identifiés et la matrice des distances entre ces sous-séquences récurrentes ;
- **La classification en motifs,** incluant des informations sur les classes identifiées, les sous-séquences qu'elles contiennent, leurs représentants, et la possibilité de visualiser les sous-séquences de chaque classe dans la séquence initiale. Des facilités de recherche de la classe d'une tentative de motifs donnée, ou de définition de contraintes sur l'effectif d'une classe et la durée des sous-séquences qu'elle contient sont également proposées.

La seconde section concerne l'analyse des performances du système par comparaison s'il est disponible à un fichier de données contenant les caractéristiques et la localisation des motifs initialement présents dans la séquence étudiée. L'analyse de performance peut être réalisée avec des contraintes de seuil minimum sur la durée d'une séquence et l'effectif d'une classe pour que celles-ci soient représentatives. Les indices de sensibilité et spécificité de l'indentification des tentatives de motifs d'une part et de leur classification en motifs d'autre part sont alors présentés.

G.1.3 Évaluation à grande échelle des performances du système

Un processus d'analyse des performances de l'extraction à grande échelle est également implémenté, sans interface graphique. Il permet de réaliser successivement plusieurs tests d'extraction de motifs. Un test correspond à l'analyse d'un jeu de motifs (sélectionnés aléatoirement dans une journée "typique" de vie quotidienne d'une personne) insérés dans une séquence correspondant à des déplacements aléatoires. L'expérimentation de différents paramètres d'instanciation des motifs et d'extraction est possible pour chaque test. Les performances obtenues sont sauvegardées dans une table de données `MATLAB` en même temps que les caractéristiques de l'introduction des motifs et de leur extraction. Ces tables sont ensuite exploitées pour évaluer l'efficacité du système et sa résistance au bruit.

G.2 Structure globale du programme d'extraction de motifs

La structure globale du programme d'extraction de motifs est complètement *séquentielle*. Il s'agit d'appeler l'une après l'autre les fonctions réalisant les étapes "clés" du système, et correspondant successivement au prétraitement et à l'abstraction des données brutes, à l'extraction des tentatives de motifs, puis à leur classification en motifs. Chaque fonction appelée prend en paramètre les résultats de l'appel à la fonction précédente, et fournit de la même manières les résultats de son exécution en paramètre de la fonction suivante.

Les grandes étapes de l'extraction de motifs sont ainsi les suivantes :

```
% Fonction de l'extraction de motifs
% Retourne le nom du fichier de sauvegarde et le temps d'exécution
function analysisRes = analyzingPatientData()

% ==== Initialisation des variables globales (les paramètres étudiés)
global MOVES = [];
global POSTURES = [];
global ACTIVITYLEVELS = [];
global FC = []; ...

% ==== Lecture des paramètres dans le fichier texte approprié
```

```
fidInput = fopen(strcat('Input_',mfilename,'.txt')); ...
TAGVALS = readFileInfosTags(fidInput, TAGNAMES); ...
fclose(fidInput);

% ====
% Initialisation d'une structure pour sauvegarder les valeurs
% des paramètres clés de l'extraction
PARAMS = struct('samplePeriod',1,'minDiscreteDist',0,'n_symb',0,'n_proj',0,
'p_proj',0,'nb_proj',0,'c_min',0,'d_max',0,'epsilonLCSS',0,'deltaLCSS',0,...);

% ==== Importation des séquences des paramètres analysés
load(patientFile,'-mat');

% ==== Alignement temporel des données des différents paramètres
[RAW_DATA, samplePeriod] = buildDataTable(MOVES,POSTURES,ACTIVITYLEVELS,FC,...);

% ==== Prétraitement de la séquence
[PRETREATED_DATA, PARAMS] = pretreatData(RAW_DATA, PARAMS,... );

% ==== Discrétisation
[DISCRETE_DATA, BREAKPOINTS, DISCRETE_DIST] = discretizeData(PRETREATED_DATA);

% ==== Agrégation DISCRETE_DATA
[AGGR_REG_DATA, PARAMS] = aggrRegularities(
       DISCRETE_DATA, PRETREATED_DATA, BREAKPOINTS, DISCRETE_DIST, PARAMS);

% ==== Projections aléatoires
[ASH_TABLE, ASH_TABLE_IDX, COLLISION_TABLE, PARAMS] = getProjections(
       AGGR_REG_DATA, PRETREATED_DATA, PARAMS);

% ==== Extraction des tentatives de motifs
[TENTATIVE_MOTIFS, DIST, PARAMS] = extractTentativeMotifs(
       ASH_TABLE, ASH_TABLE_IDX, COLLISION_TABLE, PRETREATED_DATA, PARAMS);

% ==== Classification en motifs
[CLASS_MOTIFS, nb_Kc] = extractMotifs(DIST);

% ==== Calcul du représentant de chaque classe
MOTIFS_SPEC = characterizeMotifs(
       CLASS_MOTIFS, TENTATIVE_MOTIFS, ASH_TABLE_IDX, PRETREATED_DATA, PARAMS);

% ==== sauvegarde des résultats
save(outFileName, 'MOTIFS_SPEC','CLASS_MOTIFS','DIST',...);

% ==== Variables de sorties de la fonction
analysisRes.outFileName = outFileName;
analysisRes.timeExec = timeExec;
```

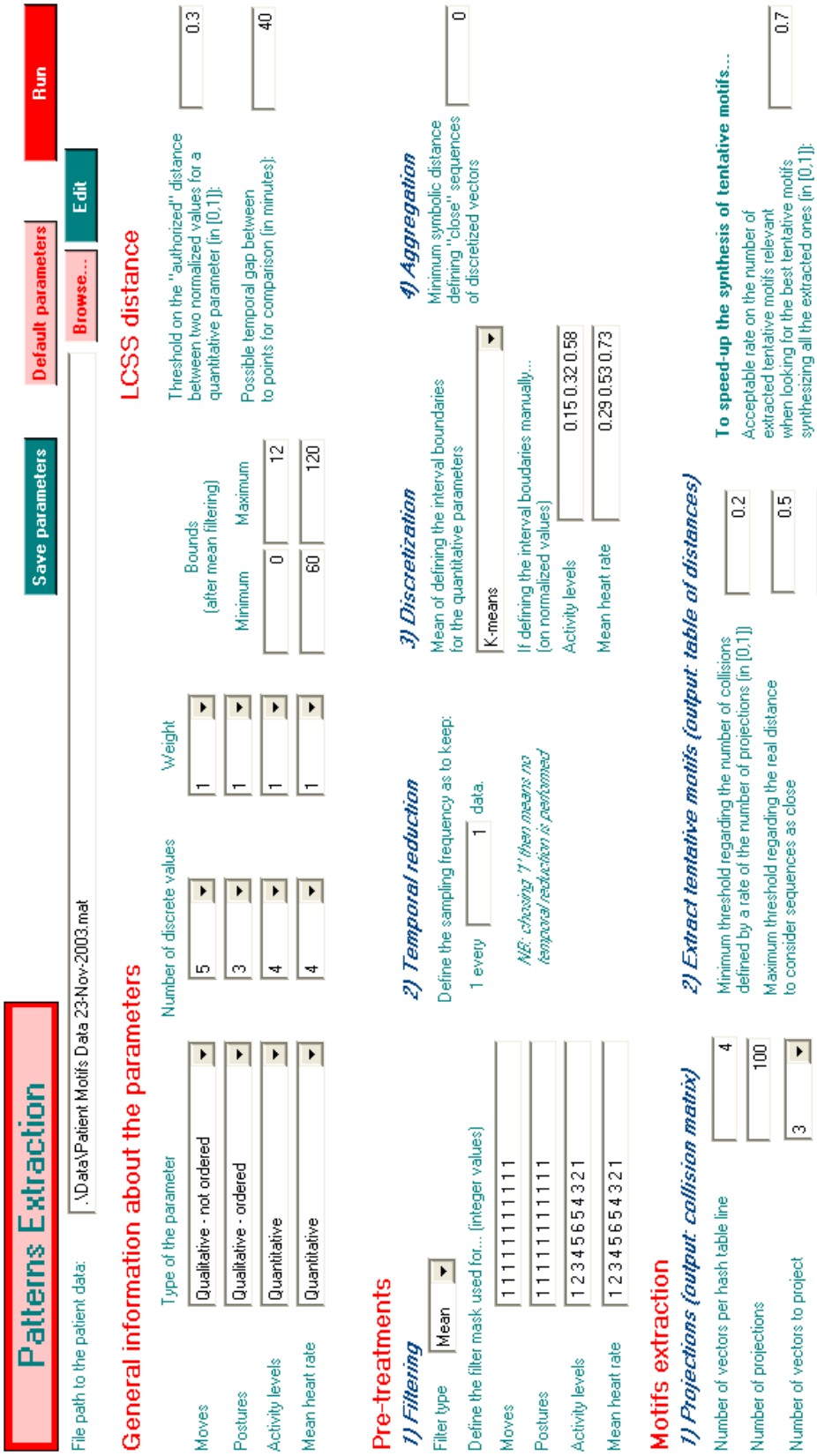


FIG. G.1 – Interface du processus d'extraction de motifs.

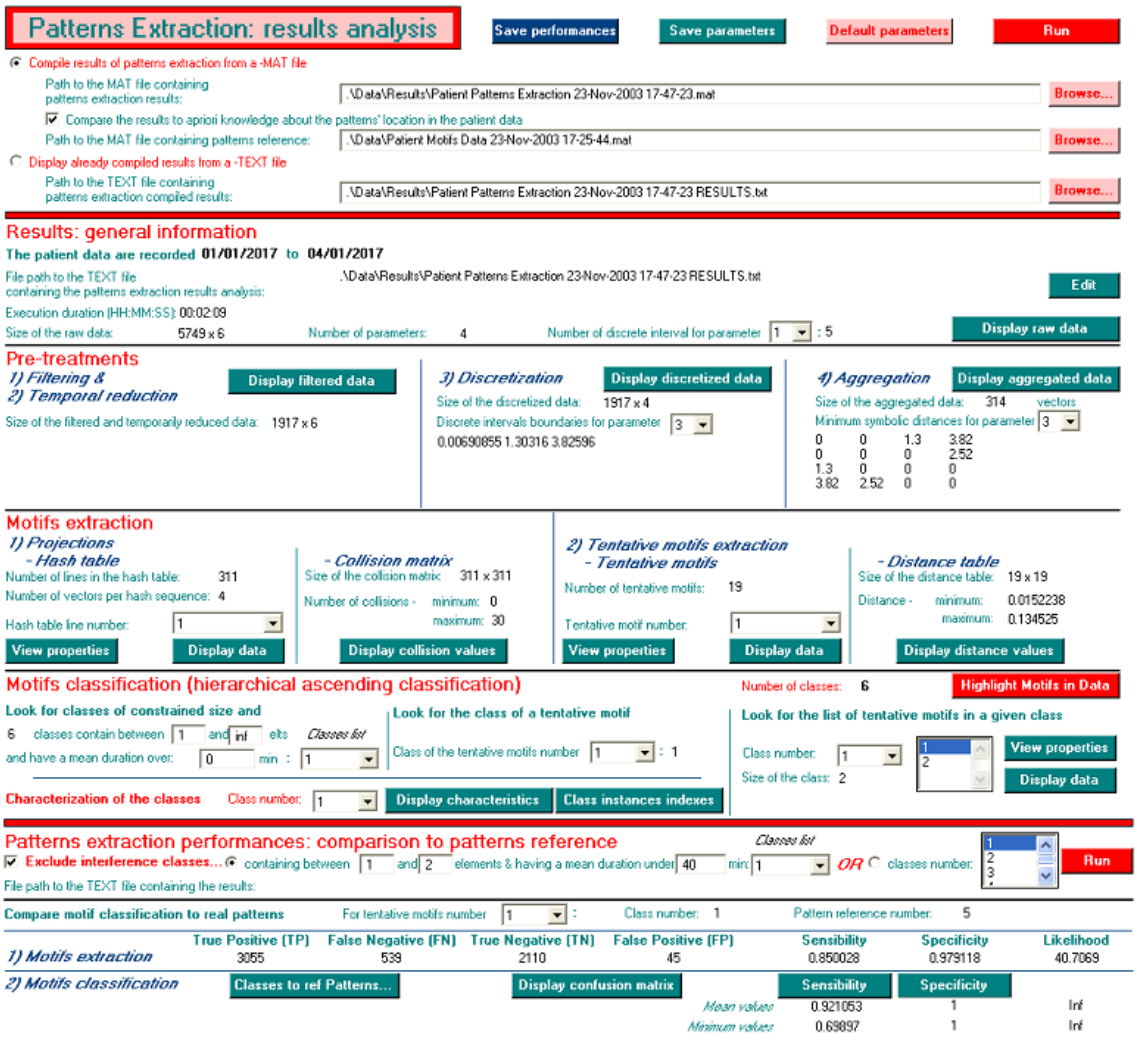


FIG. G.2 – Interface de l'analyse des résultats et des performances de l'extraction.

H

Plus Longue Sous-séquence Commune

H.1 Notion de plus longue sous-séquence commune

La plus longue sous-séquence commune à deux séquences – notée *LCSS* ou encore *LCS*, du terme anglophone *Longest Common Subsequence* – est impliquée dans le calcul de la distance entre ces deux séquences, tel que décrit au paragraphe III.4.2. Si on note $LCSS_{\delta,\epsilon}(A, B)$ la longueur de la plus longue sous-séquence commune à deux séquences A et B, la distance entre ces deux séquences, notée $D_{\delta,\epsilon}(A, B)$, est en effet calculée selon l'équation suivante [110] :

$$D_{\delta,\epsilon}(A, B) = 1 - \frac{LCSS_{\delta,\epsilon}(A, B)}{min(n, m)},$$

où δ et ϵ sont respectivement les écarts maximum autorisés pour les valeurs et le temps entre deux points similaires.

Plus généralement, la comparaison de séquences est nécessaire dans de nombreux environnements de recherche tels que : les *correcteurs orthographiques*, afin d'identifier le mot d'un dictionnaire qui ressemble le plus à un mot donné ; les *gestionnaires de versions*, afin de stocker de façon compacte les versions successives sous la forme de l'enregistrement des différences par rapport à la version initiale ; la *biologie moléculaire*, pour la comparaison de séquences ADN par exemple. Le problème de recherche de la plus longue sous-séquence commune à deux séquences est ainsi largement abordé depuis de nombreuses années, comme en témoignent l'article publié par Wagner et Fischer en 1975 [111], jusqu'à celui publié par exemple par Tsai en 2003 [107]. Dans [7], Bergroth *et al.* proposent une revue des algorithmes de résolution de ce problème.

Les paragraphes suivants présentent les concepts de base et quelques algorithmes de recherche de la plus longue sous-séquence commune que l'on présente dans le cadre de la comparaison de deux mots dont les caractères appartiennent à un certain alphabet. L'application de ces algorithmes au contexte de comparaison de deux trajectoires multidimensionnelles et hétérogènes ne modifie que la contrainte de similarité entre deux points ou caractères. Dans le cas de la comparaison de deux mots, la similarité de deux caractères est équivalente à leur égalité, alors qu'elle correspond à une notion plus complexe lorsqu'il s'agit de deux points – similarité des valeurs de chaque composante selon certains seuils de proximité (sur les valeurs et dans le temps) définis *a priori* (voir paragraphe III.4.2).

H.2 Concepts de base

Considérons deux mots X et Y, de longueur respective $|X| = m$ et $|Y| = n \geq m$, tels que $X = x_1x_2 \dots x_m$ et $Y = y_1y_2 \dots y_n$. Une sous-séquence commune à X et Y peut être obtenue en supprimant certains points de X ou de Y. Les algorithmes de recherche d'une plus longue sous-séquence commune aux séquences X et Y, notées aussi $X = X_m$ et $Y = Y_n$, se basent tous sur un ensemble de propriétés décrites ci-dessous. Si on note $Z_k = z_1z_2 \dots z_k$, une des plus longues sous-séquences communes, de longueur k telle que $k \leq min(m, n)$, les propriétés suivantes sont vérifiées :

1. Si $x_m = y_n$ alors $z_k = x_m = y_n$ et $Z_{k-1} = z_1z_2 \dots z_{k-1}$ est la plus longue sous-séquence commune à X_{m-1} et Y_{n-1} ;
2. Si $x_m \neq y_n$ alors $z_k \neq x_m$ implique que Z_k est une plus longue sous-séquence commune à X_{m-1} et Y_n ;
3. Si $x_m \neq y_n$ alors $z_k \neq y_n$ implique que Z_k est une plus longue sous-séquence commune à X_m et Y_{n-1}.

L'algorithme de recherche d'une plus longue sous-séquence commune obtenu le plus directement à partir de ces propriétés est alors l'algorithme récursif défini par la formule (H.1) :

$$L(i,j) = \begin{cases} 0 & \text{si } i = 0 \text{ ou } j = 0, \\ L(i-1, j-1) + 1 & \text{si } x_i = y_j, \\ max\{L(i-1,j), L(i,j-1)\} & \text{si } x_i \neq y_j, \end{cases} \tag{H.1}$$

où $L(i,j)$, $1 \leq i \leq m$ et $1 \leq j \leq n$, est la longueur de la plus longue sous-séquence commune à $X_i = x_1x_2 \dots x_i$ et $Y_j = y_1y_2 \dots y_j$. Ainsi, la longueur de celle commune à $X = X_m$ et $Y = Y_n$ est $L(m,n)$. La fonction récursive `LCS(X,i,Y,j)` qui calcule la longueur de la plus longue sous-séquence commune à X_i et Y_j est alors définie par l'algorithme ci-dessous :

```
LCS(X,i,Y,j)
si i = 0 ou j = 0
alors retourner 0
sinon si X[i] = Y[j]
      alors retourner 1 + LCS(X,i-1,Y,j-1)
      sinon retourner max(LCS(X,i-1,Y,j),LCS(X,i,Y,j-1))
      fin
fin
```

Cet algorithme est cependant de complexité au moins exponentielle dans le cas où les deux sous-séquences n'ont pas d'éléments en commun. De nombreuses recherches concernent ainsi la définition d'algorithmes rapides pour la résolution du problème [4, 47, 52]. Les paragraphes suivants décrivent en particulier un algorithme de résolution par programmation dynamique [111], puis un algorithme plus efficace proposé par Hirschberg à partir de la définition du contour de L [47]. Enfin nous présentons une variante de [47] proposée par Apostolico [4] et implémentée dans le cadre de notre application.

H.3 Algorithme de résolution par programmation dynamique

Une des premières approches de résolution du problème de recherche de la plus longue sous-séquence commune a été proposé par plusieurs chercheurs dont Wagner et Fischer [111] en 1975.

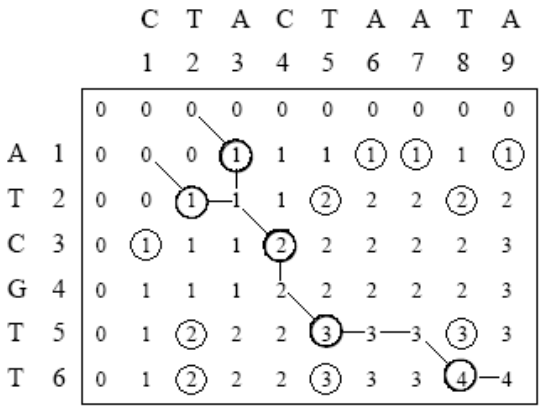


FIG. H.1 – Construction du tableau L pour la comparaison des deux mots $x = ATCGTT$ et $y = CTACTAATA$.
La valeur maximum du tableau donne la longueur de la plus longue sous-séquence commune à x et y, c'est-à-dire 4 dans cet exemple. Les valeurs $L(i, j)$ entourées dans le tableau correspondent aux caractères similaires, $x_i = y_j$. Les chemins tracés permettent de remonter la construction de L pour identifier alors les plus longs sous-mots communs, "ACTT" et "TCTT".

C'est un algorithme de résolution par programmation dynamique qui repose sur la même fonction de base que l'algorithme récursif – formule (H.1) – mais conserve les valeurs successives de L dans un tableau.

À partir d'un tableau $L[0 \dots m, 0 \dots n]$ et de deux mots $x[1 \dots m]$ et $y[1 \dots n]$, l'algorithme permettant alors la construction de ce tableau, L, est décrit ci-dessous.

```
pour i=1 à m faire
    pour j = 1 à n faire  si x[i]=y[j]
                          alors L[i,j] := L[i-1,j-1] + 1
                          sinon L[i,j] := max{L[i-1,j],L[i,j-1]};
                          fin
    fin
fin
```

La figure H.1 illustre la construction du tableau L pour la comparaison des mots $X = ATCGTT$ et $Y = CTACTAATA$. $L(n, m)$, c'est-à-dire $L(6, 9)$ dans cet exemple, est la valeur maximum du tableau et longueur de la plus longue sous-séquence commune à X et Y – $L(6, 9) = 4$.

La détermination du ou des plus longs mots communs à X et Y nécessite de remonter dans la construction du tableau L pour retrouver les prédécesseurs de chaque valeur à partir de $L(n, m)$. À chaque fois qu'on rencontre l'égalité de caractères de X et Y, $x_i = y_j$ avec $1 \leq i \leq m$ et y_j, $1 \leq j \leq n$, c'est que la deuxième règle du système d'équations (H.1) a été appliquée et le caractère correspondant appartient à une sous-séquence commune de X et Y. On remonte ainsi dans la construction de L jusqu'à trouver une longueur nulle. En général plusieurs chemins sont possibles car il n'y a pas systématiquement une unique plus longues sous-séquence commune. Par exemple sur l'illustration de la figure H.1 on identifie deux plus longs sous-mots communs aux mots X et Y considérés.

La complexité de cet algorithme est en $O(mn)$, due aux deux boucles imbriquées, et d'autres méthodes bien plus efficaces en terme de temps et d'espace, mais aussi plus complexes, ont été

développées. En particulier, Hirschberg [47] et Hunt et Szymanski [52] ont exploité les propriétés du tableau L pour définir le "contour" de L, à la base du développement d'algorithmes rapides.

H.4 Algorithme de résolution par contours

Puisque la plus longue sous-séquence commune à deux séquences X et Y consiste en une succession de points communs à ces deux sous-séquences, la restriction du tableau L construit par programmation dynamique à l'étude de ces points communs – les valeurs entourées sur l'exemple de la figure H.1 – ne doit pas faire perdre d'informations nécessaire à la résolution du problème. Les régularités observées dans le tableau L suggèrent par ailleurs des propriétés structurelles à exploiter pour la résolution du problème, et en particulier on note les caractéristiques suivantes :

$$\begin{cases} L(i-1,j-1) \leq L(i,j-1), \\ L(i-1,j-1) \leq L(i-1,j), \\ |L(i,j) - L(i-1,j-1)| \leq 1. \end{cases}$$

Les algorithmes proposés par Hirschberg [47] ou encore Hunt et Szymanski [52] sont ainsi fondés sur ces observations et exploitent en particulier une représentation efficace du tableau L par ses *contours*, comme présenté sur l'exemple de la figure H.2. Cette représentation nécessite de définir les concepts suivants, où r est la longueur de la plus longue sous-séquence commune à deux séquences X et Y.

- **Similarités** : Un couple (i,j) définit une similarité entre les séquences X et Y si $x_i = y_j$.
- **Ensemble des similarités** : L'ensemble de tous les points similaires, noté M est tel que $M = \{(i,j)/x_i = y_j, 1 \leq i \leq m \cap y_j, 1 \leq j \leq n\}$. Les valeurs entourée de la figure H.2 correspondent à l'ensemble des similarités observées.
- **Classe de similarités** : Chaque similarité observée appartient à une classe notée C_k et telle que $C_k = \{(i,j)/(i,j) \in M \cap L(i,j) = k, 1 \leq k \leq r\}$. Il est par ailleurs souvent pratique de définir une "pseudo-classe" $C_0 = (0,0)$. Comme chaque similarité est associée à exactement une classe, ces classe définissent une partition de l'ensemble M des similarités.
- $k-$**Similarité** : Couple (i,j) définissant une similarité de la classe C_k. On dit alors que k est le rang de (i,j). Les valeurs entourées sur la figure H.2 égales à un entier k donné définissent la classe C_k.
- **Similarité** $k-$**dominante** : Couple (i,j) définissant une similarité de la classe C_k telle que pour n'importe quel autre couple $(i^{'},j^{'})$ de rang k soit $i^{'} > i$ et $j^{'} \leq j$, soit $i^{'} \leq i$ et $j^{'} > j$. Les valeurs doublement entourées sur la figure H.2 égales à un entier k définissent les similarités $k-$dominantes.
- **Contours** : Limites des régions où les valeurs $L(i,j)$ sont égales. Les similarités k-dominantes se situent juste sous le $k^{ème}$ contour. Les coins de chaque $k^{ème}$ contour correspondent à la localisation des similarités $k-$dominantes.

La connaissance des similarités $k-$dominantes suffit finalement à la définition des contours, et par conséquence à la résolution du problème de recherche de la plus longue sous-séquence commune. Plusieurs stratégies ont été ensuite développées autour de cette réflexion. En particulier,

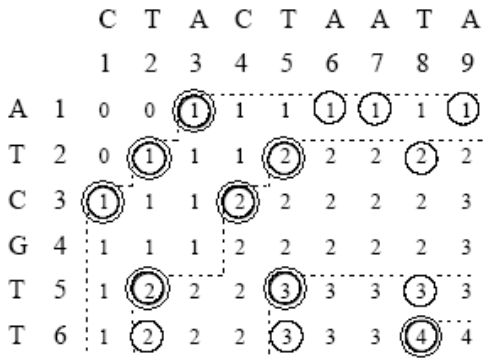


FIG. H.2 – Tableau $L[1 \ldots m, 1 \ldots n]$ résultant de la programmation dynamique.
Les valeurs $L(i,j)$ correspondant à des caractères identiques entre X et Y – $x_i = y_j$ – sont entourées, celles qui correspondent plus spécifiquement à une similarité $k-$dominante le sont doublement, et les régions dans lesquelles les valeurs de $L(i,j)$ sont identiques sont séparées par des contours en pointillés.

l'algorithme de Hirschberg [47] consiste à rechercher ligne par ligne en référence au tableau L les similarités $k-$dominantes qu'il contient. Le principe est décrit ci-dessous.

- On suppose que les similarités $(k-1)-$dominantes ont été découvertes pour un k donné, $0 \leq k \leq r-1$, en examinant la partie de L située au-dessus ou à gauche du $(k-1)^{ème}$ contour, inclus.
- La recherche du $k^{ème}$ contour est alors réalisée en explorant les valeurs non encore explorées de L, de la droite vers la gauche et de haut en bas, jusqu'à ce qu'une similarité soit rencontrée sur une ligne i donnée. La similarité rencontrée la plus à gauche est alors une similarité $k-$dominante (i,j) avec la plus petite valeur pour i.
- La recherche continue alors à la ligne suivante à gauche de cette similarité, et ce processus se répète pour toutes les lignes successives jusqu'à ce que le $k^{ème}$ contour soit identifié complètement.

L'application de ce principe proposée par Hischberg [47] permet d'atteindre une complexité $O(nr + nlogn)$ en temps.

H.5 Algorithme implémenté dans notre application

Dans le cadre de notre application on utilise un algorithme similaire à celui proposé par Hirschberg, issu de [4]. D'autres méthodes encore plus rapides et efficaces ont été développées, mais les temps de calculs obtenus en utilisant l'algorithme que l'on va décrire sont considérés comme suffisamment rapides pour les premières expérimentations.

H.5.1 Présentation de l'algorithme

L'algorithme implémenté nécessite quelques prétraitements pour identifier, pour chaque symbole distinct σ_p de la séquence X, le nombre d'occurrences de σ_p dans la séquence Y, et la liste $\sigma_p - OCC$ des positions, dans l'ordre croissant, des occurrences de ces symboles dans Y. D'après Hirschberg, on peut ensuite définir les contours à partir des $\sigma_p - OCC$ plutôt qu'en parcourant le tableau L.

Apostolico [4] propose un algorithme réalisé en r étapes, r étant la longueur d'une plus longue sous-séquence commune à X et Y. Chaque $k^{ème}$ étape a pour objectif l'identification des similarités $k-$dominantes. Cet algorithme utilise les tableaux ou listes de valeurs définies ci-dessous.

- `X[1..m]`et `Y[1..n]` : Séquences comparées, contenant respectivement `m` et `n` symboles.
- `x_OCC[1..m]`: Tableau définissant pour chaque symbole `X[i]` de `X` la liste des positions dans `Y` de symboles identiques : par exemple `x_OCC[i]=[3;7]` signifie que `X[i]=Y[3]=Y[7]`.
- `PEBBLE[..]` : Tableau d'entiers, initialisés à 1. Pour tout entier `i`de `[1..m]`, alors `PEBBLE[i]` pointe : (a) soit sur une entrée de la liste `x_OCC[i]` des positions de symboles identiques à `X[i]` dans la séquence `Y`; on a alors `x_OCC[i][PEBBLE[i]]` dans `[1..n]` et il est dit *actif* ; (b) soit `x_OCC[i][PEBBLE[i]]` vaut `n+1`, et il est dit *inactif*.
- `RANK[0..r]` : Tableau contenant pour chaque valeur k la liste des couples (i, j) correspondant aux similarités $k-$dominantes.
- `SYMB[1..m]` : Tableau auxiliaire défini pour chaque ligne `i` qui permet d'accéder directement à l'indice dans la liste `x_OCC[i]` de la position courante considérée pour `Y`, c'est-à-dire tel que `SYMB[i][t]=p` si `Y[t]=Y[x_OCC[i][p]]=X[i]`. La valeur de `SYMB[i][n+1]` correspond alors au nombre de symboles de `Y` similaires plus `1`, c'est-à-dire au dernier indice de la liste `x_OCC[i]`.

On définit également les deux variables suivantes, `T` et `t`.

- `T` : Indice dans `Y` de la dernière similarité dominante enregistrée lors d'une étape `k` donnée. Elle correspond ainsi à un *seuil maximum* sur les indices de `Y` pour le test de l'enregistrement possible comme similarité `k`-dominante d'une similarité observée à un moment donné entre un certain symbole `X[i]` de `X` et un symbole similaire `Y[t]` de `Y`. La valeur de `t` correspond ainsi une valeur du tableau `x_OCC[i]`contenant les indices dans `Y` des symboles similaires à `X[i]`. Comme à chaque instant de l'étape `k`, `PEBBLE[i]` contient l'indice courant, dans la liste `x_OCC[i]`, de la similarité étudiée entre `X[i]` et `Y[t]`, alors on doit vérifier `x_OCC[i][PEBBLE[i]]<T` pour enregistrer cette similarité comme `k`-dominante.

- `t` : Pour chaque ligne `i`, indice dans `Y` de la valeur initiale du seuil `T` au début de l'examen des points de `Y` similaires à `X[i]`. Lors d'une étape `k` donnée, et pour la considération de la ligne `i`, `t` vaut ainsi soit `n+1` si aucune similarité `k`-dominante n'a encore été enregistrée, ou bien une valeur strictement inférieure et égale à l'indice dans `Y` de la dernière similarité `k`-dominante enregistrée lors de l'examen des lignes `1` à `i-1`. La variable `t` permet ainsi de positionner la valeur suivante de `PEBBLE[i]` pour la réalisation de l'étape `k+1`, une fois l'étape `k` réalisée sur la ligne `i`. L'objectif est de pointer l'indice d'un symbole de `Y` identique à `X[i]` et qui soit au moins égal à l'indice de la dernière similarité `k`-dominante enregistrée. La contrainte sur `PEBBLE[i]` est alors `x_OCC[i][PEBBLE[i]]<t`.

Le processus de recherche se termine lorsqu'il n'y a plus aucun `PEBBLE[i]` qui est *actif*. L'algorithme proposé par Apostolico et extrait de [4] est présenté ci-dessous.

```
pour i = 1 à m faire PEBBLE[i] = 1;
k = 0;
tant que "il y a un PEBBLE actif" faire
   % Début de l'étape k+1 de recherche des similarités (k+1)-dominantes
   T = n+1; k = k+1; RANK[k] = [];
   pour i = k à m faire
      % Avance des PEBBLE
      t = T;
      si x_OCC[i][PEBBLE[i]] < T
      alors
         % Enregistrement d'une similarité dominante
         RANK[k] = RANK[k] U [i, x_OCC[i][PEBBLE[i]]];
         % Mise à jour du seuil T
         T = x_OCC[i][PEBBLE[i]];
      fin
      % Avance du PEBBLE si nécessaire
      si X[i] = Y[t]
      alors
         PEBBLE[i] = SYMB[i][t] + 1;
      sinon tant que x_OCC[i][PEBBLE[i]] < t faire
         PEBBLE[i] = PEBBLE[i] + 1;
      fin
   fin
fin
```

H.5.2 Application sur un exemple

La figure H.3 illustre le fonctionnement de cet algorithme en indiquant les positions initiales des `PEBBLE[i]` pour chaque étape `k` de recherche de similarités $k-$dominantes.

Initialisations

Au début de l'exécution, pour chaque `i`, `PEBBLE[i]=1` c'est-à-dire qu'il pointe le premier indice de `Y` correspondant à un symbole identique à `X[i]` : `Y[x_OCC[i][PEBBLE[i]]]=X[i]`. Puis, `PEBBLE[i]` se déplace dans la liste `x_OCC[i]` jusqu'au dernier indice correspondant à `x_OCC[i][PEBBLE[i]]=n+1` lorsque le `PEBBLE` devient *inactif*, c'est-à-dire lorsque plus aucune des similarités enregistrées dans la liste `x_OCC[i]` ne peut être considérée comme dominante, à n'importe quel rang `k`.

La variable `T` indique quant à elle l'indice dans `Y` de la dernière similarité dominante enregistrée lors d'une étape `k` donnée. Au début de chaque étape `k` de recherche, la variable `T` est ainsi initialisée à `n+1`, indiquant qu'aucune similarité `k`-dominante n'a encore été enregistrée.

Recherche des similarités 1-dominantes : $k = 1$

Considérons la recherche des similarités `1`-dominantes sur l'exemple proposé, soit `k=1`. Pour chaque ligne `i`, de `1` à `m`, il faut alors déplacer les `PEBBLE[i]` en fonction des similarités observées entre `X[i]` et les symboles de `Y`, enregistrées dans le tableau `x_OCC[i]` afin d'identifier celles qui sont `1`-dominantes. Pour chaque ligne `i`, une similarité entre `X[i]` et `Y[x_OCC[i][PEBBLE[i]]]` est `1`-dominante et peut être ajoutée à la liste `RANK[1]` si `x_OCC[i][PEBBLE[i]]<T`, où `T` enregistre l'indice dans `Y` de la dernière similarité `1`-dominante enregistrée.

Examen de la première ligne : $i = 1$
Pour `i=1`, on a `X[1]=A`et les symboles similaires dans `Y` sont aux indices `x_OCC[1]=[3;6;7;9;10]`, incluant en dernier `x_OCC[1][5]=10` (`n+1`) pour rendre le pointeur sur la liste `x_OCC[1]` inactif. On a également `PEBBLE[1]=1` et `T=n+1`. Ainsi, le premier symbole de `Y` similaire à `X[1]`, soit `Y[x_OCC[1][PEBBLE[1]]]=Y[x_OCC[1][1]]=Y[3]` est enregistré comme `1`-dominant et ajouté à la liste `RANK[1]`. `T` est alors mis à jour à la valeur `x_OCC[1][PEBBLE[1]]=3` pour ne pas enregistrer de similarités `1`-dominantes correspond à des symboles de `Y` dont l'indice serait supérieur à `3`. On incrémente ensuite la valeur de `PEBBLE[1]` pour continuer le parcours de la liste `x_OCC[1]`. Comme on n'avait pas encore enregistré de similarité dominante, `PEBBLE[1]` est positionné de façon à pointer l'indice `10` de `Y` pour la recherche des similarités dominantes suivantes : on n'identifiera donc plus de similarités dominantes sur cette ligne.

Examen de la deuxième ligne : $i = 2$
On étudie ensuite la ligne suivante, soit `i=2`, où `X[2]=T`. On a alors `x_OCC[2]=[2;5;8;10]`, et initialement `PEBBLE[2]=1`, `T=3` et `t=T=3`. Comme il existe un symbole de `Y` identique à `X[2]` à un indice strictement inférieur à `T=3`, on l'enregistre comme `1`-dominant. On repositionne ensuite `PEBBLE[2]` de façon à pointer à indice dans `Y` , correspondant à un autre symbole identique à `X[2]`, mais qui soit au moins égal à `t=3`, soit `PEBBLE[2]=5`.

Itérations suivantes : $i > 2$***, puis*** $k > 1$
On itère de la même façon ce processus pour toutes les lignes `i` jusqu'à ce que tous les `PEBBLE[i]` soient inactifs. La longueur de la plus longue sous-séquence commune à `X` et `Y` correspond alors au rang `r` de la dernière similarité dominante enregistrée. Les plus longues sous-séquences communes correspondent alors aux successions possibles, pour chaque `k` allant de `1` à `r`, d'un couple similaire `k`-dominant contenu dans la liste `RANK[k]`. Si le couple `k`-dominant `(i,j)`est sélectionné à l'indice `k`, le couple `(k-1)`-dominant `(i',j')`sélectionné à l'indice `k` doit vérifier `i'<i` et `j'<j`.

H.5.3 Contexte de notre application

Dans le contexte de notre application, l'implémentation de cet algorithme est modifiée uniquement du point de vue de la définition de la similarité de deux symboles. Les symboles considérés sont en effet des vecteurs, et la similarité de deux points doit ainsi être vérifiée sur chaque composante, en fonction d'un seuil définissant l'écart maximum autorisé entre deux valeurs similaires dans le cas d'une composante de type quantitatif.

Par ailleurs, on utilise également un seuil maximum de proximité dans le temps entre deux points similaires. On limite ainsi pour chaque symbole x_i la dimension de la liste des indices j dans Y des valeurs similaires à x_i. L'étude des similarités possibles entre les points de X et Y est alors largement restreinte dès que l'écart maximum autorisé dans le temps est suffisamment faible. On limite alors à un certain intervalle les indices j possibles pour Y en fonction de l'indice i considéré pour X.

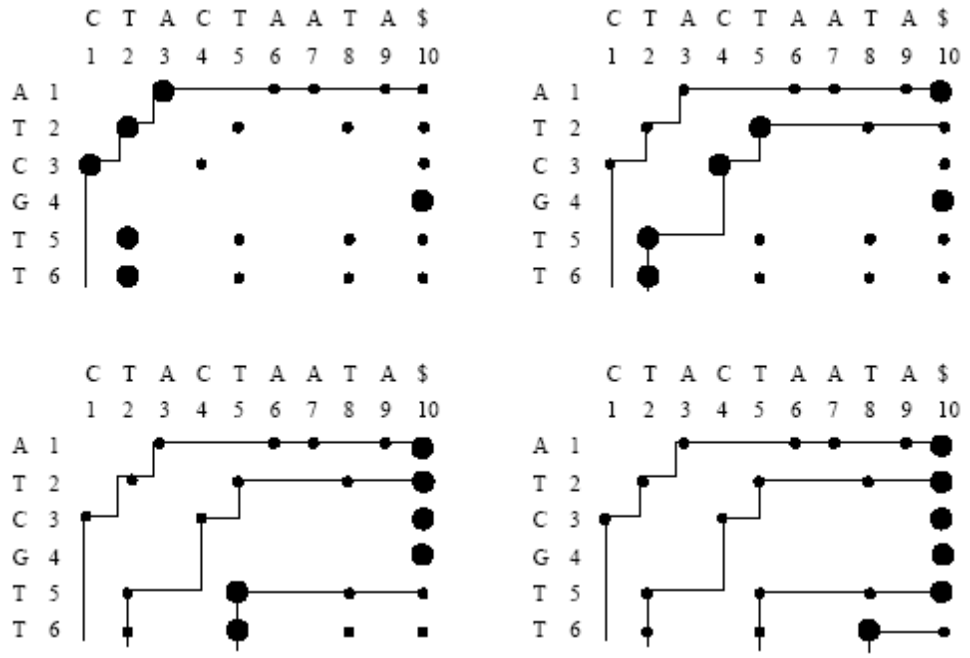


FIG. H.3 – Fonctionnement de l'algorithme de recherche de la plus longue sous-séquence commune. [4]
Les points noirs correspondent aux similarités observées entre X et Y, et les ronds noirs indiquent les positions occupées pour chaque ligne i par les $PEBBLE[i]$ au début de chacune des 4 étapes d'identification pour chacune des similarités $k-$dominantes, $1 \leq k \leq 4$.

I

Mesure de distance selon le principe *DTW*

I.1 Principe

La présentation ci-dessous du principe de calcul d'une distance selon la méthode *DTW* – "*Dynamic Time Warping*" – est issue de [56]. Elle a été introduite initialement à la communauté de fouille de données par Berndt et Clifford [8]. L'objectif de cette distance est d'apporter une solution à la sensibilité d'une distance Euclidienne à une distorsion de l'axe du temps. Le principe est alors d'aligner les séquences comparées, c'est-à-dire de trouver successivement les couples de points de chacune des deux séquences qui correspondent à une distance globale minimum calculée alors selon le principe d'une distance *Euclidienne*.

I.2 Calcul de la distance

Considérons deux séries temporelles à une dimension Q et C, de longueur respective n et m, définies par :

$$Q = q_1, q_2, \ldots, q_i, \ldots, q_n$$

$$C = c_1, c_2, \ldots, c_j, \ldots, c_m$$

Pour aligner les deux séquences par le principe *DTW*, on construit une matrice de dimension $n \times m$ dont l'élément (i, j) contient la distance Euclidienne entre les points q_i et c_j, notée $d(q_i, c_j)$:

$$d(q_i, c_j) = (q_i - c_j)^2.$$

Un alignement possible pour la comparaison des séquences Q et C correspond alors à un chemin W suivant des éléments contigus de cette matrice. Le $k^{ème}$ élément de W est défini par $w_k = (i, j)_k$, tel que :

$$W = w_1, w_2, \ldots, w_k, \ldots, w_K \text{ où } max(m, n) \leq L \leq m + n - 1.$$

La définition d'un chemin W doit également vérifier les contraintes suivante :

- **Conditions aux limites.** $w_1 = (1, 1)$ et $w_K = (m, n)$, ce qui signifie que le chemin doit commencer et finir aux points diagonalement opposés de la matrice.
- **Continuité.** Si $w_k = (a, b)$, alors $w_{k-1} = (a', b')$ où $a - a' \leq 1$ et $b - b' \leq 1$. Cette contrainte restreint l'évolution d'un chemin vers les éléments adjacents de la matrice.

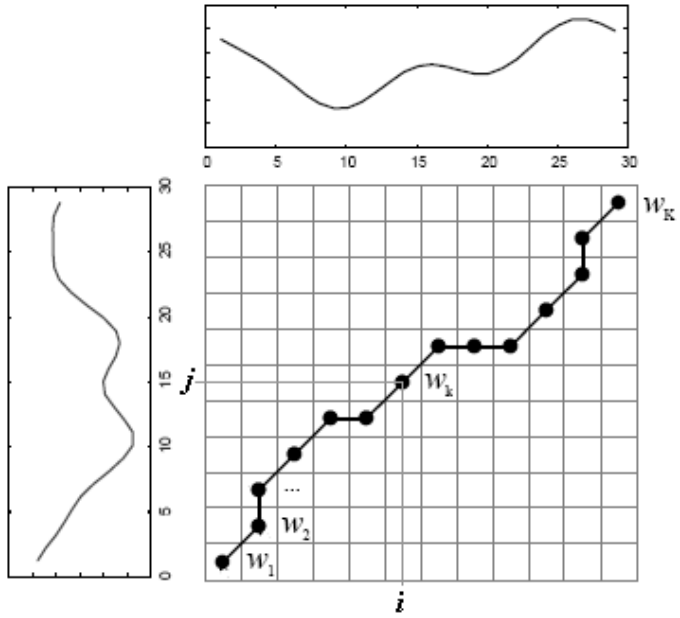


FIG. I.1 – Illustration sur un exemple de la recherche d'un chemin DTW.

- **Monotonie.** Si $w_k = (a, b)$, alors $w_{k-1} = (a', b')$ où $a - a' \geq 0$ et $b - b' \geq 0$. Cela contraint les points successifs du chemin à être répartis de façon monotone dans le temps.

Dans ces conditions, il y a un nombre exponentiel de chemins W pour parcourir les éléments de la matrice. On s'intéresse finalement uniquement à celui qui minimise la distance, notée alors $DTW(Q, C)$, entre la succession des couples de points qui le constitue :

$$DTW(Q, C) = min\left\{\sqrt{\frac{\sum_{k=1}^{K} w_k}{K}}\right\} = min\left\{\sqrt{\frac{\sum_{k=1}^{K}(q_{i_k} - c_{j_k})^2}{K}}\right\}.$$

Le facteur K compense le fait que les chemins appropriés à la comparaison de deux séquences peuvent avoir des longueurs différentes. Le principe de cette distance est illustré sur la figure I.1.

La distance Euclidienne peut alors être considérée comme un cas particulier de la distance DTW, selon les contraintes suivantes :

- Les séquences comparées doivent être de même longueur : $n = m$;
- Le choix du chemin W est alors contraint de façon à ce que chacun de ses éléments $w_k = (i, j)_k$ soit tel que : $i = j = k$.

I.3 Calcul effectif

Le calcul effectif de cette distance dite DTW est réalisé très efficacement à l'aide de la programmation dynamique. L'objectif est alors d'évaluer la distance cumulée au niveau de la comparaison du couple de points (i, j), notée $\gamma(i, j)$, comme la distance entre les points i et j à laquelle s'ajoute le minimum des distances cumulées correspondant aux points adjacents, tel que décrit par l'équation suivante :

$$\gamma(i, j) = d(q_i, c_j) + min\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1).$$

J

Liste des publications

J.1 Revues

V. Rialle, F. Duchêne, N. Noury, L. Bajolle, J. Demongeot, "Health "Smart" Home : Information Technology for Patients at Home," *Telemedicine Journal and E-Health*, vol. 8(4), pp. 395–410, Winter 2002.

J.2 Congrès internationaux avec actes et comité de lecture

F. Duchêne, C. Garbay, V. Rialle, "Similarity Measure for Heterogeneous Multivariate Time-series," *Proc. of the 12th European Signal Processing Conference (EUSIPCO)*, Vienna, Austria, 7-10 septembre, 2004.

F. Duchêne, C. Garbay, V. Rialle, "An Hybrid Knowledge-Based Methodology for Multivariate Simulation in Home Health Telecare," *Proc. of the Joint Workshop Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP) of the 9th Artificial Intelligence in Medicine Europe conference (AIME)*, Cyprus, 19-22 octobre 2003, pp. 87–94.

F. Duchêne, C. Garbay, V. Rialle, "An Hybrid Refinement Methodology for Multivariate Simulation in Home Health Telecare," *Proc of the 5th International Workshop on Enterprise Networking and Computing in Healthcare Industry*, Santa Monica, CA, 6-7 juin 2003, pp. 101–110.

F. Duchêne, V. Rialle, N. Noury, "Home Health Telecare : Proposal of an Architecture for Patient Monitoring and Critical Situation Detection," *Proc. of the 4th International Workshop on Enterprise Networking and Computing in Healthcare Industry*, Nancy, France, juin 2002, pp. 105–108.

J.3 Congrès nationaux avec actes et comité de lecture

F. Duchêne, V. Rialle, N. Noury, "Télésurveillance médicale à domicile : Proposition d'une architecture pour un système de détection de situations critiques et de décision sur l'état d'un patient," *Actes des 9ème Journées Francophones d'Informatique Médicale*, A.M. Grant, J.P. Fortin, L. Mathieu (eds), Soqibs, Sherbrooke, Canada, mai 2002, pp. 451–461.

Résumé

Le développement des systèmes de télésurveillance médicale à domicile est fondamental face au vieillissement de la population et aux capacités limitées d'admission dans les hôpitaux et centres spécialisés. Ce travail de thèse concerne particulièrement la conception d'un assistant intelligent pour l'analyse des données hétérogènes collectées par des capteurs au domicile afin de détecter, voire prévenir, l'occurrence de situations inquiétantes. Il s'agit de concevoir un système d'apprentissage des habitudes de vie d'une personne, tout écart par rapport à ce profil comportemental étant considéré comme critique. L'étude proposée concerne d'une part la conception d'un processus de simulation pour la génération de grandes quantités de données appropriées au contexte expérimental. D'autre part, une méthode générique pour l'extraction non supervisée de motifs dans des séquences temporelles multidimensionnelles et hétérogènes est proposée puis expérimentée dans le contexte de l'identification des comportements récurrents d'une personne dans ses activités quotidiennes. On évalue en particulier les indices de sensibilité (tolérance aux modifications normales de comportement) et de spécificité (rejet des modifications inquiétantes) du système. L'application du système d'apprentissage aux séquences générées par la simulation permet également de vérifier l'extraction possible de comportements récurrents interprétés *a posteriori* en terme de la réalisation d'activités de la vie quotidienne.

Mots-clés: Télésurveillance médicale à domicile, Fusion de données hétérogènes, Simulation multivariée, Analyse de données multidimensionnelles, Fouille de séries temporelles, Apprentissage non supervisé, Motifs temporels.

Abstract

The development of medical remote care applications is crucial due to the general aging of the population and the restricted number of possible admissions to hospital, residential or nursing homes. This thesis deals with the concept of "smart assistants" used to analyze large heterogeneous data sets collected at home from sensors, in order to detect and prevent unusual situations that may give serious cause for concern. The aim is to learn meaningful patterns representative of the person's daily living habits. Any deviation from this learned profile is then considered as an unexpected situation. The study concerns on one hand the conception of a multivariate simulation process for generating large amount of data relevant to our experimental context. On the other hand, we build a generic unsupervised learning process for mining heterogeneous multivariate time-series and identifying temporal patterns. The proposed approach is applied in the context of identifying recurrent behaviors during activities of daily living. We especially evaluate the sensibility (tolerance of normal changes in behavior) and specificity (rejection of abnormal changes) of the system. This learning method is applied to the output of the simulation process and shows its relevance in extracting recurrents behaviors that can be *a posteriori* interpreted as some activities of daily living.

Keywords: Home health telecare, Heterogeneous data fusion, Multivariate simulation, Multivariate data analysis, Time-series mining, Unsupervised learning, Temporal patterns.