Université de Reims Champagne-Ardenne
École Doctorale Mathématiques, Physique, Sciences du Numérique et de l'ingénieur, MPSNI, ED 620
Laboratoire de Mathématiques de Reims, LMR, CNRS, UMR 9008

THÈSE

Pour obtenir le grade de

Docteur de l'Université de Reims Champagne-Ardenne

Discipline : Mathématiques Appliquées

Présentée et soutenue publiquement par

Laurie SAPIA

le 12 décembre 2025

MODÉLISATION PRÉDICTIVE DES FLUX DE PATIENTS DANS LES SERVICES ACCUEIL D'URGENCES DU GRAND EST

Thèse dirigée par Amor KEZIOU
Co-encadrée par Philippe REGNAULT

JURY

Directeur de thèse	Amor KEZIOU	Université de Reims Champagne-Ardenne
Co-Encadrant	Philippe REGNAULT	Université de Reims Champagne-Ardenne
Rapporteur	Salim BOUZEBDA	Université de Technologie de Compiègne
Rapportrice	Emilie LEBARBIER	Université Paris-Nanterre
Examinatrice	Mitra FOULADIRAD	Aix-Marseille université, Centrale Méditerranée
Examinatrice	Emmanuelle GAUTHERAT	Université de Reims Champagne-Ardenne
Invitée	Céline GIGET	Association Est-RESCUE

Remerciements

Je souhaite exprimer ma profonde gratitude à toutes les personnes qui ont contribué à la réalisation de cette thèse.

Tout d'abord, je tiens à remercier chaleureusement mes directeurs de thèse, Amor Keziou et Philippe Regnault. Je suis reconnaissante d'avoir bénéficié de votre encadrement et d'avoir bénéficé de vos précieux avis depuis le master. Merci à Amor Keziou de m'avoir initié à la recherche avec le mémoire de master et merci à Philippe Regnault de m'avoir fait découvrir les projets statistiques. Merci à vous deux de vos conseils avisés, de votre patience, d'avoir toujours été disponible pour répondre à mes questions et enfin de votre soutien tout au long de ces années.

Je tiens tout particulièrement à remercier Salim Bouzebda et Emilie Lebarbier pour avoir accepté de rapporter ma thèse et pour le temps précieux qu'ils ont consacré à la lecture attentive de ce manuscrit. Je remercie également les membres du jury, Mitra Fouladirat et Emmanuelle Gautherat, pour avoir accepté d'évaluer cette thèse.

Je souhaite également remercier l'équipe Est-RESCUE, dont les échanges constructifs ont enrichi ma réflexion tout au long de ce travail. Un grand merci aussi aux membres du CSI, pour leurs remarques pertinentes et leurs conseils avisés, qui ont contribué à améliorer la qualité de cette recherche.

Je souhaite exprimer ma gratitude aux professeurs et enseignants du Laboratoire de Mathématiques de Reims (LMR) pour leurs enseignements de qualité et leur implication. Je tiens également à remercier tout particulièrement mes professeurs de master, qui ont su éveiller en moi l'intérêt pour la statistique. Un grand merci à Christelle pour son aide précieuse au quotidien, et à Guillaume pour avoir toujours répondu présent pour me sortir de mes impasses. Je remercie tous les doctorants que j'ai rencontrés

durant ces belles années de thèse au LMR pour leur bonne humeur, leurs conseils précieux et les moments conviviaux partagés. Je remercie donc les anciens doctorants : Quentin, Ana, Pierre, Benjamin et Lulu, et je remercie également les doctorants actuels : Lisa, Jade, Charlotte, Killian, Esther, Perrine, Hugo et Yoann.

Je remercie de tout cœur mes amies Lilou, Cyrielle et Alicia pour leur soutien indéfectible et leur présence précieuse tout au long de cette aventure. Merci également à Margaux, ma précieuse collègue en statistique.

Enfin, je tiens à exprimer ma reconnaissance à ma famille et à mes amis pour leur soutien moral et leur patience. Un merci tout particulier à Jérémy pour sa compréhension et son encouragement constant tout au long de cette aventure.

Résumé

Les tensions hospitalières dans les Services Accueil d'Urgences (SAUs) ont augmenté ces dernières années et sont largement médiatisées. Des facteurs organisationnels, structurels ou conjoncturels peuvent expliquer cette situation, qui doivent être pris en compte pour anticiper les flux de patients. Nous proposons dans cette thèse un modèle additif généralisé des flux journaliers intégrant des informations exogènes (telles que les périodes de vacances scolaires) dans lequel la tendance est une fonction polynomiale et la saisonnalité est une combinaison d'harmoniques. Ce modèle est d'abord ajusté individuellement pour chaque SAU de la région Grand Est à partir des historiques quotidiens de fréquentation fournis par l'association Est-RESCUE. Une extension multivariée du modèle est ensuite proposée, combinant des techniques de clustering et de régression PLS. Les performances de prédiction sont évaluées et comparées aux modèles classiques SARIMA, qui sont la référence couramment utilisée dans la littérature.

Mots-clés : Flux de patients des services d'urgence, modèle SARIMA, modèle additif généralisé, séries temporelles, prédiction.

Abstract

Hospital tensions in the French emergency departments (ED) have increased last years and are widely publicized. Organizational, structural or cyclical factors can explain this situation ; they must be taken into account to anticipate the flows of patients in the EDs. In this PhD thesis, we propose a generalized additive model of daily flows incorporating exogenous information (such as school holiday periods) in which trend is a polynomial function and seasonality is a combination of harmonics. This model is first adjusted individually for each ED of the French Grand Est area from the daily records of frequentation provided by the association Est-RESCUE. A multivariate version of the model is then proposed, combining clustering and PLS regression technique. The prediction performances are evaluated and compared to the classical SARIMA models, which are the reference commonly used in the literature.

Keywords : Emergency departments' patients flow, SARIMA model, generalized additive model, time-series, forecasting.

Table des matières

Table des matières vii

Table des figures ix

Liste des tableaux xii

Liste des abréviations xiii

Liste des services accueil d'urgences de la région Grand Est xvii

Introduction 1

1 Contexte 5

1.1 Réponse organisationnelle à la sur-fréquentation des services d'accueil d'urgences 5

1.2 Partenariat Est-RESCUE - URCA 7

1.3 Des résumés de passages aux urgences aux historiques de fréquentation 8

1.4 Modélisation des historiques de fréquentation 12

2 Modélisation univariée des flux de passage 17

2.1 Notations 17

2.2 Présentation des modèles 18

2.2.1 SARIMA et SNAIVE 18

2.2.2 PROPHET 20

2.2.3 Modèles Additifs Généralisés avec tendance polynomiale et saisonnalité basée sur des harmoniques empiriques 22

2.2.4 Comparaisons formelles des modèles 24

2.3 Sélection de modèles et ajustement aux données 25

2.3.1 Sélection de modèles SARIMA 26

2.3.2 Sélection de modèles TSHR et estimation des paramètres 27

2.4 Comparaison des performances de prévision des modèles 30
2.4.1 Les mesures de performance de prévision d'un modèle 30
2.4.2 Estimation de l'erreur de prévision d'un modèle par validation croisée par fractionnement séquentiel 31
2.5 Annexe - Présentation des variantes du modèle TSHR et des procédures d'estimation 33
2.5.1 Variantes du modèle TSHR 33
2.5.2 Variantes de la méthode de sélection de modèle 34
2.5.3 Variantes de la méthode de validation croisée 35

3 Ajustement de modèles aux historiques de fréquentation des SAUs et comparaison de leurs performances de prévision 37
3.1 Déroulement de la procédure d'ajustement du modèle TSHR ; exemple avec le SAU du « Pôle Public St Louis » 38
3.2 Analyse des résidus : exemple avec le SAU du « CHU de Metz » 40
3.3 Comparaison des performances de prévision des modèles SNAIVE, SARIMA, PROPHET et TSHR 43
3.4 Analyse a posteriori des modèles 45

4 Modélisation multivariée des flux via clustering et régression PLS 51
4.1 Clustering des historiques de fréquentation 51
4.1.1 Méthodes de clustering et modalités de comparaison de leurs performances 52
4.1.2 Résultats du clustering 58
4.2 Modélisation multivariée via régression PLS par cluster 60
4.2.1 Le modèle de régression PLS 60
4.2.2 Application du modèle de régression PLS pour une modélisation multivariée des séries temporelles 63
4.2.3 Évaluation de la performance de prévision des trois modèles multivariés 65
4.3 Annexe : description de la composition des classes 66

Bibliographie 75

Table des figures

1.1 Représentation schématique du fonctionnement de la collecte et de l'analyse des données d'un RPU. La transmission des RPUs de l'émetteur au récepteur est figurée par des traits en pointillés oranges avec des flèches bleus pointant vers le récepteur. Source : [Est-Rescue, a] 9

1.2 Série temporelle représentant les flux de patients aux urgences du CHU de Metz . . 10

1.3 Périmètre territorial du Groupement Hospitalier de Territoire de rattachement. Source : [Est-Rescue, b] 12

2.1 Illustration de la méthode SSCV pour l'estimation de l'erreur théorique de prévision MAE(H), pour $H = 7$, $T = 1008$ et $T_{\max} = 1096$. 33

3.1 Évolution de l'EQM d'ajustement (à gauche) et de la valeur du BIC (à droite) en fonction du degré de la tendance polynomiale au cours de la première étape de la procédure de sélection du modèle (en haut), et du nombre d'harmoniques dans la composante saisonnière au cours de la deuxième étape (en bas). 39

3.2 Périodogramme des résidus (2.13) après suppression de la tendance polynomiale, pour le SAU Pôle Public St Louis. La ligne horizontale en pointillés bleus correspond au seuil déterminant les fréquences les plus significatives. 40

3.3 Ajustement de l'historique de fréquentation du SAU Pôle Public St Louis par les composantes T, TS, TSH et TSHR (à gauche, de haut en bas). La fonction d'autocorrélation des résidus correspondants pour chaque modèle est présentée à droite. 41

3.4 Fonctions d'autocorrélation (ACF) des résidus des modèles, de gauche à droite, PROPHET, TSHR et SARIMA. 42

3.5 Analyse de la normalité des résidus pour les trois modèles, de gauche à droite, PROPHET, TSHR et SARIMA. En haut : QQ-plots des résidus. En bas : histogrammes des résidus avec densités superposées : la courbe continue rouge correspond à l'estimateur à noyau de la densité, la courbe noire en pointillés représente la densité de la loi normale. 43

3.6 Comportement de la séquence de moyennes partielles 44
3.7 Distributions pour l'ensemble des SAUs des valeurs retenues pour le nombre d'harmoniques, le paramètre p et le paramètre q du modèle TSHR 46
3.8 Diagrammes en barres, représentant le degré optimal du polynôme noté K (à gauche) et le nombre d'harmoniques optimal noté m (à droite) ; identifiés pour chaque SAU, en distinguant les SAUs pédiatriques des non pédiatriques. Les nombres d'harmoniques ont été regroupés en quatre classes pour améliorer la lisibilité. 47
3.9 Diagramme en barres représentant le nombre d'harmoniques optimal (noté m) pour chaque SAU, distinguant les établissements privés et les établissements publics. Les nombres d'harmoniques ont été regroupés en quatre classes pour améliorer la lisibilité. 48
3.10 Diagramme en barres représentant le nombre d'harmoniques optimal (noté m) chaque SAU, distinguant les établissements avec et sans SMUR. Les nombres d'harmoniques ont été regroupés en quatre classes pour améliorer la lisibilité. 49
3.11 Diagrammes en barres représentant, à gauche, le degré polynômial optimal K, et à droite, le nombre d'harmoniques optimal m, pour chaque SAU, en distinguant les SAUs suivant la gravité des cas à traiter. Les nombres d'harmoniques ont été regroupés en quatre classes afin de faciliter la lecture. 49
3.12 Diagrammes en barres représentant, à gauche, le degré polynômial optimal K, et à droite, le nombre d'harmoniques optimal m, pour chaque SAU, en fonction de leur appartenance au GHT. Les nombres d'harmoniques ont été regroupés en quatre classes afin de faciliter la lecture. 50
3.13 Carte des SAUs de la région Grand Est, indiquant pour chaque service le degré polynômial optimal K et le nombre d'harmoniques optimal m. Les services pédiatriques sont mis en évidence en fond vert. 50

4.1 Courbes de silhouette des onze méthodes de clustering appliquées aux historiques de fréquentation centrés et réduits. Les lignes horizontales hachurées rouges identifient les valeurs maximales de chacune des courbes tandis que les lignes verticales en pointillés bleus identifient les nombres de classes pour lesquels ces maximum sont atteints. 59
4.2 Positions géographiques des SAUs. La couleur d'un point permet d'identifier la classe à laquelle est alloué le SAU correspondant. La taille du point est proportionnelle à la fréquentation hebdomadaire moyenne sur la période 2016-2018. La forme du point identifie le type d'établissement. 69

4.3 Distributions des fréquentations quotidiennes moyennes des SAUs selon leur classe. 70
4.4 Qualités de représentations des SAUs dans le plan principal associé à l'ACP des historiques de fréquentation centrés réduits. 71
4.5 Corrélations de la première composante principale avec les fréquentations quotidiennes pour les années 2016 à 2018. 72
4.6 Corrélations de la deuxième composante principale avec les fréquentations quotidiennes pour les années 2016 à 2018. 73
4.7 Projections des cinquante et un SAUs sur le plan principal associé à l'ACP des historiques de fréquentation centrés et réduits. Les axes 1 et 2 représentent respectivement 6,98 % et 5,97 % de l'inertie. Les couleurs identifient les classes d'appartenances des SAUs. 74

Liste des tableaux

1.1 Synthèse des articles portant sur la modélisation des flux de patients aux urgences médicales. 14

2.1 Ensemble $\mathbb{H}$ des six types de périodes de vacances et jours fériés et les dates correspondantes pour l'année 2016 (les dates pour les années 2017 et 2018 sont omises). . 23

2.2 Caractéristiques formelles distinguant les modèles TSHR et SARIMA. 25

2.3 Plages de valeurs pour les hyperparamètres des modèles SARIMA et TSHR définis dans les parties 2.2.1 et 2.2.3. 26

2.4 Mesures de performance de prévision. 31

3.1 MAE et WMAPE agrégés (moyenne des 51 SAUs) aux horizons de prévision de 7 à 28 jours pour les modèles SNAIVE, SARIMA, PROPHET et TSHR. 45

3.2 MAE de prévision à 7 jours des quatre modèles pour les sept SAUs de la Marne. . . 45

3.3 Les hyperparamètres retenus lors de la procédure de sélection du modèle TSHR par validation croisée séquentielle pour les sept SAUs de la Marne ($\widehat{K}$: degré du polynôme modélisant la tendance, $\widehat{m}$: nombre d'harmoniques de la composantes saisonnière, $(\widehat{p}, \widehat{q})$: les hyperparamètres de l'ARMA modélisant le terme résiduel). . 46

4.1 Distances usuelles utilisées pour le clustering de séries temporelles. 54

4.2 Mesures de dissimilarité utilisées dans les méthodes de clustering des SAUs basés sur la CAH. 54

4.3 Principales caractéristiques des trois méthodes de partitionnement de type k-means utilisées pour le clustering des SAUs. 55

4.4 Ensemble des méthodes de clustering et leurs caractéristiques comparées dans le cadre de la définition des classes de SAUs. On renvoie à la table 4.2 pour la définition des acronymes des mesures de dissimilarité. 58

4.5 Les hyperparamètres optimaux des modèles multivariés associés aux trois classes de SAUs. 64

4.6 MAE et WMAPE agrégés (moyenne des SAUs appartenant à la classe) aux horizons de prévision de 7 à 28 jours pour les trois modèles multivariés. 65
4.7 Composition et taille des trois classes obtenues à l'issue de la procédure de clustering par k-means. 67

Liste des abréviations

ACF	Fonction d'autocorrélation
ACP	Analyse en Composantes Principales
AR	Modèle autorégressif
ARIMA	Modèle autorégressif intégré à moyenne mobile
ARMA	Modèle autorégressif à moyenne mobile
ARS	Agence régionale de santé
CAH	Classification Ascendante Hiérarchique
CH	Centre Hospitalier
CHR	Centre Hospitalier Régional
CHU	Centre Hospitalier Universitaire
Cli	Clinique
CV	Validation croisée
DTW	Déformation temporelle dynamique
Est-RESCUE	Réseau et observatoire des urgences du Grand Est
GAM	Modèle additif généralisé
GARCH	Modèle autorégressif généralisé à hétéroscédasticité conditionnelle
GHSA	Groupement Hospitalier Sud-Ardennes
GHT	Groupement hospitalier de territoire
HCli	Hôpital Clinique
HET	Hôpital en tension
IRM	Imagerie par résonance magnétique
MA	Modèle à moyenne mobile
MAE	Erreur absolue moyenne
MAPE	Erreur absolue moyenne en pourcentage
MCO	Moindres carrés ordinaires
ORU	Organisation des réseaux d'urgence
PAM	Partitionnement autour des médoïdes
PLS	Moindres carrés partiels
PLSR	Régression par moindres carrés partiels

RLM	Régression linéaire multiple
RMSE	Racine de l'erreur quadratique moyenne
RNN	Réseau de neurones récurrents
RPU	Résumés de passage aux urgences
SARIMA	Modèle autorégressif intégré à moyenne mobile saisonnier
SAU	Service accueil d'urgences
SMUR	Service mobile d'urgence et de réanimation
SSCV	Validation croisée par fractionnement séquentiel
TSH	Modèles additifs généralisés avec tendance, saisonnalité et vacances
TSHR	Modèles additifs généralisés avec tendance, saisonnalité, vacances et ARMA
UPGMA	Méthode des paires non pondérées avec moyenne arithmétique
WMAPE	Erreur absolue moyenne pondérée en pourcentage
WPGMC	Méthode des paires pondérées avec moyenne centroidale

Liste des services accueil d'urgences de la région Grand Est étudiés

1 CH Altkirch
2 CH Bar-le-Duc
3 CH Briey
4 CH Chalons
5 CH Charleville adultes
6 CH Charleville pédiatrie
7 CH Chaumont
8 CH Epernay
9 CH Epinal
10 CH Forbach
11 CH Guebwiller
12 CH Haguenau
13 CH Langres
14 CH Lunéville
15 CH Mont St Martin
16 CH Mulhouse Hôpital Emile Muller adultes
17 CH Mulhouse Hôpital Emile Muller pédiatrie
18 CH Neufchateau
19 CH Pont à Mousson
20 CH Remiremont
21 CH Sarrebourg
22 CH Sarreguemines
23 CH Saverne
24 CH Sedan
25 CH Sélestat
26 CH St Dié

27 CH St Dizier
28 CH Toul
29 CH Troyes
30 CH Verdun
31 CH Vitry
32 CH Vittel
33 CH Wissembourg
34 CHR Metz
35 CHR Thionville
36 CHU Nancy Brabois pédiatrie
37 CHU Nancy Central adultes
38 CHU Reims adultes
39 CHU Reims pédiatrie
40 CHU Strasbourg Hautepierre adultes
41 CHU Strasbourg Hautepierre pédiatrie
42 CHU Strasbourg NHC
43 Cli Courlancy
44 Cli Courlancy Reims-Bezannes
45 Cli Gentilly
46 Cli Sainte Anne
47 Diaconat Fonderie
48 GHSA Rethel Vouziers
49 HCli Claude Bernard
50 Hôpital de St Avold
51 Pôle public St Louis

Introduction

Les Services Accueil d'Urgences (SAUs) jouent un rôle central dans le système de santé en assurant la prise en charge immédiate des patients nécessitant une intervention rapide. Cependant, ces dernières années, la surcharge et les tensions dans ces services se sont intensifiées, phénomène largement médiatisé et préoccupant pour les décideurs et les professionnels de santé. Plusieurs facteurs peuvent expliquer ces tensions : des contraintes organisationnelles (effectifs, horaires, répartition des patients), des facteurs structurels (taille et localisation des établissements, capacité d'accueil, vieillissement de la population, désertification médicale), ainsi que des événements conjoncturels ou saisonniers (épidémies, périodes de vacances scolaires, variations climatiques) influençant directement l'affluence. Comprendre et anticiper les flux de patients dans les SAUs est ainsi devenu un enjeu majeur, non seulement pour améliorer la qualité des soins et la sécurité des patients, mais également pour optimiser la gestion des ressources hospitalières.

Dans ce contexte, la modélisation et la prévision des flux journaliers apparaissent comme des outils essentiels permettant de prendre en compte à la fois la tendance générale, la saisonnalité et les effets des facteurs exogènes. Cette thèse s'inscrit dans cette perspective en proposant des méthodes basées sur des modèles additifs généralisés et des extensions multivariées, visant à prédire avec précision l'affluence dans les SAUs et à comparer ces approches à des modèles classiques comme SARIMA, largement utilisés dans la littérature pour l'analyse des séries temporelles hospitalières.

Les SAUs sont confrontés à une augmentation des tensions hospitalières, nécessitant une gestion efficace des ressources et une coordination entre établissements. Les Agences Régionales de Santé (ARS) s'appuient sur des indicateurs quotidiens issus des Résumés de Passage aux Urgences (RPU) pour suivre l'activité, réorienter les patients si nécessaire et déclencher des plans d'urgence tels que le plan HET ou le Plan blanc. La collecte et l'analyse de ces données sont centralisées par les Observatoires Régionaux des Urgences (ORU), dont Est-RESCUE est

le représentant pour la région Grand Est. Ce dernier synthétise les données des SAUs, identifie les situations à risque et alerte les ARS pour faciliter la prise de décision.

Pour améliorer l'anticipation des flux de patients, un partenariat a été établi entre Est-RESCUE et l'Université de Reims Champagne-Ardenne (URCA). Sur les 59 SAUs initialement étudiés, 51 ont été retenus pour leur cohérence et la disponibilité des données. Les RPUs ont été agrégés pour produire des séries temporelles quotidiennes, complétées par des variables descriptives propres à chaque service (type d'établissement, spécialisation, présence d'équipements ou de SMUR, etc.). Ces éléments de contexte sont présentés de façon détaillés dans le chapitre 1.

Cette thèse est structurée autour des contributions principales suivantes :

- nous introduisons un modèle additif généralisé (GAM) adapté aux historiques de fréquentation des SAUs, nommé TSHR en référence aux quatre composantes dont il est la somme : une tendance polynomiale, une composante saisonnière prenant la forme d'un polynôme trigonométrique, une composante événementielle combinaison linéaire d'indicatrices et un processus résiduel autorégressif à moyenne mobile (ARMA). L'ajustement du modèle TSHR à une série temporelle se fait en quatre étape imbriquées, chaque composante du modèle étant ajusté sur les résidus de l'étape précédente. La sélection des hyperparamètres suit une approche similaire, en utilisant le critère d'information bayésien (BIC). Les principaux atouts du modèle TSHR résident dans 1) l'efficacité computationnelle des procédures d'ajustement et de sélection de modèle, 2) une interprétabilité claire et 3) une extension aux séries multivariées ;

- une méthode de validation croisée par fractionnement séquentiel (SSCV) adaptée aux séries temporelles est introduite, permettant, notamment, d'estimer les performances de prévision de modèles stochastiques. Cette méthode consiste à générer plusieurs couples d'entraînement et de test en avançant d'un jour dans la série, permettant une estimation robuste des erreurs de prévision sur différents horizons (H = 7, 14, 21, 28 jours) ;

- une extension multivariée du modèle TSHR reposant sur une régression linéaire des moindres carrés partiels (PLS) des historiques de fréquentation en fonction des composantes T, S et H du modèle TSHR ajusté sur le représentant moyen d'un groupe de SAU ;

- l'ajustement de ces modèles aux historiques de fréquentation de 51 SAUs de la région Grand Est et la comparaison des leurs performances de prévision via la procédure SSCV

avec celles de modèles usuels de la littérature, les modèles SNAIVE, SARIMA et PROPHET. Pour l'évaluation des modèles, plusieurs mesures de performance sont utilisées : MAE, MAPE, WMAPE, MSE et RMSE. L'ajustement du modèle multivarié a nécessité une étape préalable de clustering des historiques de fréquentation, permettant d'identifier trois classes de SAUs selon leurs régimes d'évolution de fréquentation ;

- une analyse descriptive contextualisée des caractéristiques des modèles ajustés aux historiques de fréquentation.

Le modèle TSHR et les modèles SNAIVE, SARIMA et PROPHET auxquels il est comparé sont présentés dans le chapitre 2. Dans ce même chapitre, sont également présentées les procédures d'ajustement et de sélection de modèles TSHR à une série temporelle et la procédure de validation croisée par fractionnement séquentiel. Enfin, une annexe identifie et caractérise plusieurs variantes de GAM qui ont été testés pour modéliser les historiques de fréquentation et dont les évolutions incrémentales ont abouti au modèle TSHR.

L'ajustement et la sélection de modèles TSHR aux historiques de fréquentation des 51 SAUs du Grand Est sont présentés dans le chapitre 3. L'apport de chaque composante y est explicité, les résidus du modèle ajusté sont étudiés, validant l'usage de ces modèles dans ce contexte. Les performances de prévision des modèles retenus pour les 51 SAUs sont comparées sur une période de test aux performances des modèles SNAIVE, SARIMA et PROPHET. Enfin, une analyse descriptive des hyperparamètres des modèles TSHR sélectionnés pour les 51 SAUs est présentée, croisant ces hyperparamètres avec les caractéristiques administratives et géographiques des établissements, permettant de faire émerger des ressemblances entre établissements et constituant un prémisse à la procédure de clustering présenté dans la première partie du chapitre 4.

Le dernier chapitre est consacré à la modélisation multivariée des historiques de fréquentation. La procédure de clustering permet de regrouper les 51 SAUs entre trois classes de 11, 21 et 19 établissements. Pour chacune de ces classes, un modèle multivarié est ajusté aux historiques de fréquentation des SAUs lui appartenant. Ce modèle est obtenu par une régression PLS des historiques sur les variables définissant les composantes T, S et H du modèle TSHR ajusté à la série moyenne des historiques de fréquentations des SAUs. Cette procédure aboutit à des performances de prévision comparables aux prévisions fournies par les modèles univariés, mais pour des temps d'exécution drastiquement réduits.

Chapitre 1

Contexte

1.1 Réponse organisationnelle à la sur-fréquentation des services d'accueil d'urgences

La tension hospitalière est une situation exceptionnelle résultant d'un déséquilibre entre les ressources disponibles de l'hôpital et le flux de patients dans ses services. Les services d'accueil d'urgences (SAUs) sont particulièrement touchés par la tension hospitalière, dans plusieurs pays ; voir par exemple [Trzeciak and Rivers, 2003], [Cooke et al., 2004a], [Hing, 2012], [Wartelle et al., 2022]. En France, la tension hospitalière dans les SAUs a été particulièrement marquée durant la pandémie de Covid-19, et elle demeure un enjeu majeur depuis, comme l'illustrent les restrictions d'accès à certains SAUs durant les périodes de forte tension, telles que les vacances d'été, phénomène largement relayé par les médias ; voir par exemple [Dagnet, 2023], [Quentin and Vildy, 2024], [Battaglia, 2022], [Williamson, 2023]. Elle peut également avoir un impact financier sur les hôpitaux, comme le rapporte [Foley et al., 2011]. La tension hospitalière a un impact négatif sur la qualité des soins. Les patients attendent plus longtemps pour avoir accès à un traitement, ce qui peut empirer leur état de santé ; voir, par exemple, [Boyle et al., 2012], [Sun et al., 2013] et [Cooke et al., 2004b]. Sans oublier que le personnel hospitalier peut avoir une charge de travaille excessive, ce qui peut engendrer de la fatigue et une diminution de la qualité des expertises prodiguées ([Sutherland et al., 2023]).

Pour limiter les effets néfastes de la surpopulation, des mesures ont été prises. En effet, en 2019, l'organisation des urgences françaises et leur financement public ont été réformés. Le plan de réforme, présenté par le ministère de la Santé ([sante.gouv, 2019]), introduit des soins alternatifs, améliore la qualité des soins et optimise le flux des patients. Il propose des me-

sures telles que la médecine à distance, l'amélioration des consultations sans rendez-vous et le renforcement des compétences des professionnels de santé, dans le but de garantir un accès rapide et approprié aux services de santé pour tous les citoyens. La note d'information du syndicat [de l'Hospitalisatoin Privée MCO, 2023] explique la réforme du financement des services d'urgence, qui passe d'une approche basée sur l'activité à une approche qui prend en compte les besoins de la population pour plus d'équité. De nouveaux mécanismes de financement, tels que la dotation supplémentaire de qualité et la dotation basée sur la population, sont introduits pour améliorer la qualité et l'équité du financement des soins d'urgence, tandis que des initiatives comme le « forfait patient d'urgence » simplifient la facturation et augmentent les revenus des soins de santé privés.

De telles tensions dans les SAUs peuvent résulter de multiples facteurs. Certains sont liés à des événements remarquables — dits conjoncturels — comme les épidémies saisonnières, les événements climatiques extrêmes ou d'autres événements plus imprévisibles comme des accidents majeurs, provoquant un afflux massif de patients sur une courte période ; voir par exemple [Pealat et al., 2019]. D'autres sont des facteurs structurels, comme le vieillissement de la population, qui entraîne une augmentation de l'afflux de personnes âgées, dont la prise en charge est notoirement plus complexe ; voir par exemple [Mielczarek, 2013]. Ce phénomène peut être aggravé dans les zones rurales par une désertification médicale : la dégradation du tissu médical local entraîne une prise en charge déficiente des soins courants, puis un afflux accru de patients vers les structures hospitalières ; voir par exemple [Wartelle et al., 2022]. D'autres facteurs encore relèvent de l'organisation : les SAUs peuvent manquer de ressources humaines ou matérielles. De nombreuses études ont analysé la gestion des ressources hospitalières [Coskun and Erol, 2010] afin d'optimiser les moyens disponibles tout en maintenant un haut niveau de qualité des soins ; voir par exemple [Kwak and Lee, 1997], [Lin et al., 2008], [Headrick and Morgan, 1988], [Ajmi et al., 2013] et [Luo et al., 2016].

En France, les pouvoirs publics ont pris en compte la tension hospitalière depuis plusieurs décennies à travers la mise en place de plans organisationnels, tels que le « Plan blanc » et le dispositif « Hôpital en Tension » (HET) ([Ammiratti et al., 2019], [Pulsy, 2024], [Est-RESCUE, 2025]). Ces plans visent à améliorer la réactivité des structures médicales face à des situations critiques. Ils s'appuient sur des réseaux de pilotage géographiquement structurés et hiérarchisés, coordonnés notamment par les Agences Régionales de Santé (ARS). Ces réseaux ont pour missions le recensement précis des ressources humaines (comme la disponibilité des équipes médicales ou paramédicales) et matérielles (nombre de lits disponibles, respirateurs, etc.), ainsi que le suivi en temps réel de l'activité hospitalière. Par exemple, en cas d'afflux massif de pa-

tients ou de saturation des capacités hospitalières, les ARS peuvent réorienter les patients vers d'autres établissements ou déclencher un plan HET. Ces actions s'appuient sur une collecte quotidienne de données et le calcul d'indicateurs de performance permettant une coordination efficace entre les différents niveaux de gestion du système de santé.

En particulier, l'activité des SAUs est suivie par un réseau d'observatoires régionaux nommés « Observatoires Régionaux des Urgences » (ORU). Les ORUs s'intègrent dans le dispositif national du Système d'Information des Urgences (SI Urgences), piloté par la Direction Générale de l'Offre de Soins (DGOS) et Santé publique France, afin d'uniformiser la collecte et l'analyse des données à l'échelle nationale. Ils ont pour mission de « collecter, analyser et partager les données issues des SAUs et de la médecine de premier recours dans une zone donnée, ainsi que de fournir une expertise en médecine d'urgence », passant par la production de rapports, de tableaux de bord d'activité ainsi que des analyses thématiques (sur les motifs de recours, les prises en charge pédiatriques, ou encore les délais d'attente). Ils constituent l'interface entre les SAUs et les structures administratives , dont les ARSs, responsables du déclenchement des plans HET ou « Plan blanc ». En situation de crise sanitaire, comme lors de la pandémie de COVID-19, cette organisation a joué un rôle central dans le pilotage en temps réel des capacités hospitalières et l'adaptation des politiques publiques.

Le déclenchement du plan HET peut être motivé par le dépassement de certains indicateurs d'activité ([Ammiratti et al., 2019]). Chaque SAU transmet des Résumés de Passage aux Urgences (RPU), qui sont des synthèses administratives standardisées retraçant la prise en charge de chaque patient ; voir la partie 1.3 pour plus de détails. Ces RPUs sont centralisés par des plate-formes régionales de coordination numérique, telles que Pulsy dans la région Grand Est, qui assurent la transmission sécurisée et automatisée de ces données vers l'ORU de rattachement – l'association Est-RESCUE pour la région Grand Est. Ce dernier synthétise ces données sous la forme d'indicateurs associés à la tension hospitalière, identifie les situations à risque (comme une augmentation soudaine de l'activité ou une saturation des capacités), puis alerte, si besoin, l'Agence Régionale de Santé (ARS), qui dispose du pouvoir décisionnel pour activer les dispositifs comme le plan HET ou le Plan blanc.

1.2 Partenariat Est-RESCUE - URCA

L'ORU de la région Grand Est s'appelle Est-RESCUE. Il s'agit d'une association créée en 2017 suite à la fusion des ORUs des anciennes régions de Champagne-Ardenne, Lorraine et Alsace. Dans le cadre de ses activités, Est-RESCUE suit notamment six indicateurs associés à la tension

hospitalière, parmi lesquels figure le nombre de patients attendus dans les SAUs à J+1 ([Pulsy, 2024]). L'intégration de prévisions à plus long terme permettrait de favoriser l'anticipation et la réactivité des services de santé.

Dans cette optique, un partenariat a été établi entre Est-RESCUE et l'Université de Reims Champagne-Ardenne (URCA), Est-RESCUE ne disposant pas du temps nécessaire ni des ressources pour la modélisation des flux de passage aux urgences. Dès 2018, une première collaboration a été formalisée entre Est-RESCUE et le master SEP (Statistique pour l'Évaluation et la Prévision) de l'URCA, sous la forme d'un contrat de partenariat. En 2021, ce contrat a été reconduit puis un nouveau a été mis en place pour encadrer une thèse de doctorat portant sur la modélisation des flux de patients dans les SAUs du Grand Est. L'objectif de ce travail de doctorat est de modéliser les historiques de fréquentation des SAUs de la région Grand Est afin de prévoir les flux de patients à un horizon temporel de sept jours, voire davantage. Ce partenariat vise à apporter une plus-value scientifique et opérationnelle : il permet de coupler la connaissance fine du terrain d'Est-RESCUE avec l'expertise méthodologique de l'URCA en modélisation prédictive. Les modèles développés peuvent non seulement enrichir les outils de surveillance existants, mais aussi soutenir la prise de décision des structures de pilotage en fournissant des indicateurs prospectifs fiables. Ils contribuent ainsi à l'optimisation des ressources hospitalières, à l'anticipation des pics d'activité et, in fine, à l'amélioration de la qualité des soins.

1.3 Des résumés de passages aux urgences aux historiques de fréquentation

Un Résumé de Passage aux Urgences (RPU) est un recueil de données administratives et médicales créé à l'arrivée d'un patient dans un SAU et complété jusqu'à sa sortie du service. La production et la transmission des RPUs sont encadrées par la loi. Depuis l'arrêté du 24 juillet 2013, la transmission de ces données vers les ORUs est obligatoire pour tous les SAUs. Ce dispositif vise à alimenter un système national de surveillance et d'analyse des urgences, piloté par Santé publique France et la la Direction Générale de l'Offre de Soins.

Les RPUs comprennent des données démographiques (âge, sexe, emplacement géographique approximatif du lieu de résidence), administratives (date, heure et durée de la prise en charge, moyen de transport utilisé pour rejoindre le SAU, etc) et médicales (indicateur de gravité de l'état à l'entrée, orientation du patient, diagnostic, etc). Ces données sont saisies à l'aide du

système d'information hospitalier, par le personnel médical et administratif, comme illustré dans la Figure 1.1.

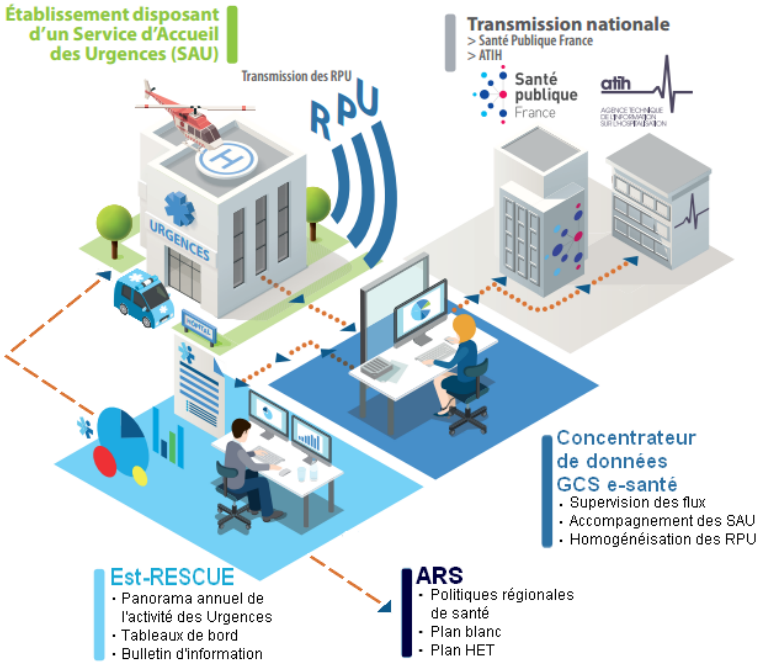

FIGURE 1.1 : Représentation schématique du fonctionnement de la collecte et de l'analyse des données d'un RPU. La transmission des RPUs de l'émetteur au récepteur est figurée par des traits en pointillés oranges avec des flèches bleus pointant vers le récepteur. Source : [Est-Rescue, a]

Le Grand Est compte 59 SAUs récoltant les RPUs des patients qu'ils prennent en charge. Une fois collectés, les RPUs sont transmis à un concentrateur des données, un organisme régional qui a pour mission de les harmoniser, en lien avec un ORU, comme schématisé dans la Figure 1.1. Ce processus permet de normaliser les données entre les services, même si elles sont déjà standardisées au niveau des établissements. Ces RPUs deviennent ainsi exploitables à l'échelle régionale et nationale. Cela permet de s'assurer que les données provenant de différents établissements de santé sont cohérentes, facilitant ainsi l'analyse globale. Ensuite, les données sont prêtes à être transmises aux ORUs, comme Est-RESCUE.

Dans le cadre du partenariat Est-RESCUE - URCA, Est-RESCUE met à notre disposition des versions simplifiées, notamment expurgées des informations identifiantes et des données médicales, des RPUs des 59 SAUs du Grand Est sur une période de 3 ans allant du 1er janvier 2016 au 31 décembre 2018. Ces RPUs sont ensuite agrégés pour obtenir 59 séries temporelles (une par SAU) correspondant aux flux quotidiens de patients dans les SAUs ; dans la suite de ce document, nous désignons ces séries par « historiques de fréquentation ». Un exemple de série temporelle est présenté à la figure 1.2.

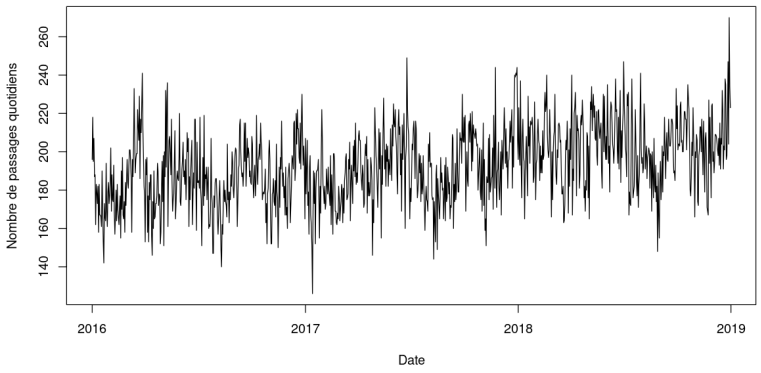

Figure 1.2 : Série temporelle représentant les flux de patients aux urgences du CHU de Metz

Certains des SAUs ont été écartés de l'étude pour diverses raisons. L'une de ces raisons concerne la nature même de ces services. En effet, les services d'urgence CCOM Illkirch et Diaconat Roosevelt ne sont pas des urgences de médecine générale, mais exclusivement des urgences spécialisés en médecine de la main. Le service du GHAM Romilly a également été supprimé, car il ne s'agit pas d'un service d'urgence, mais d'une unité des Structures Mobiles d'Urgence et de Réanimation (SMUR). Enfin, le HIA Legouest (hôpital d'instruction des armées) a été écarté car présentant des caractéristiques de fréquentation très distinctes des autres services.

Quatre SAUs supplémentaires ont été exclus de l'étude en raison de données incomplètes sur la période de trois ans, dues à des ouvertures récentes, des fusions ou des divisions entre anciens services. Ainsi, le service du GHAM Sézanne n'a commencé qu'en octobre 2017, offrant seulement un an et trois mois de données. De même, la Clinique Rhéna a ouvert en mars 2017, après la fusion de deux établissements. Concernant le CH de Colmar, le service pédiatrique n'a été ouvert qu'en septembre 2016. Avant cette date, un seul flux de données regroupait à la fois les patients adultes et pédiatriques. À partir de septembre 2016, deux flux distincts sont disponibles (adultes et pédiatrique), mais il est impossible de reconstituer rétrospectivement la séparation entre ces deux groupes sur la période antérieure. Ainsi, les données exploitables pour Colmar ne débutent qu'à partir de septembre 2016. Afin d'assurer la cohérence des données, les services adultes et pédiatrique de Colmar ont donc été exclus de l'analyse.

Enfin, deux dates manquaient pour le service adultes du CH Mulhouse (15 janvier et 14 avril 2016) ; elles ont été imputées par la moyenne des mêmes dates des autres années disponibles. Au total, 51 services d'urgence ont été retenus, correspondant à 51 séries temporelles de 1096 observations chacune.

Cette période de trois ans a été découpée en une période d'apprentissage de 1015 jours, allant du 1er janvier 2016 au 11 octobre 2018, et une période de test du 12 octobre 2018 au 31 décembre 2018. Le seuil de séparation entre ces deux périodes a été choisi de manière à ce que les deux périodes, d'apprentissage et de test, contiennent des semaines aux dynamiques de fréquentation variées, telles que des semaines de vacances et hors vacances, ou encore des épidémies de gastro-entérite ; voir la partie 2.2.3 pour plus de détails.

En plus des historiques de fréquentation, le jeu de données, fourni par Est-RESCUE, regroupe les caractéristiques suivantes, propres à chaque SAU :

- la région ;
- la localisation géographique (longitude et latitude) ;
- l'âge moyen des patients accueillis ;
- le Groupement Hospitalier de Territoire (GHT) correspond à un périmètre territorial qui réunit plusieurs établissements de santé publics autour d'un projet médical partagé. Il est paraît pertinent de l'utiliser comme référence car il reflète le territoire de coopération hospitalière dans lequel s'inscrit le service étudié. La carte correspondante est présentée dans l'image n°1.3 ;
- Un classement des services a été établi selon le nombre moyen de passages journaliers. Pour cela, nous avons d'abord calculé la moyenne quotidienne du nombre de passages pour SAU. Ensuite, les SAUs ont été répartis en quatre classes de taille égale, allant des services les moins fréquentés aux plus fréquentés : la première classe regroupe les SAUs à faible activité, tandis que la quatrième correspond aux services à très forte activité ;
- le niveau de prise en charge des traumatismes sévères (utilisé pour évaluer la capacité à traiter des patients gravement traumatisés) ;
- l'accès ou non à une IRM ;
- la spécialisation pédiatrique ou non ;
- le type d'établissement (hôpital, hôpital régional, CHU, clinique, hôpital d'instruction des armées) ;
- le caractère privé ou public de l'établissement ;
- la présence ou non d'un SMUR ;

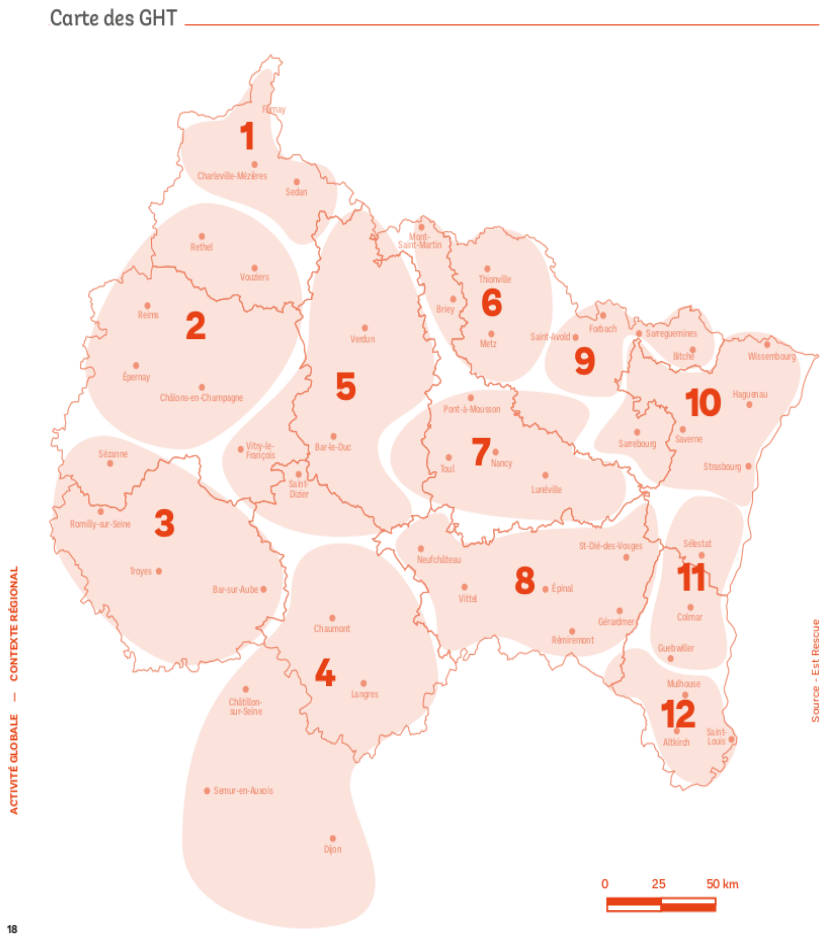

Figure 1.3 : Périmètre territorial du Groupement Hospitalier de Territoire de rattachement. Source : [Est-Rescue, b]

- la possibilité ou non d'interprétation d'images médicales à distance.

En résumé, la base utilisée contient, pour chacun des 59 services d'urgence de la région Grand Est, les données de fréquentation sur les années 2016 à 2018, ainsi qu'un ensemble de variables descriptives facilitant l'analyse des flux de patients et la caractérisation des structures concernées.

1.4 Modélisation des historiques de fréquentation

La construction de modèles prédictifs des flux de patients dans les SAUs permet d'anticiper et d'optimiser les ressources à mobiliser. La littérature sur le sujet est abondante. [De Gooijer and Hyndman, 2006] propose une revue d'articles publiés entre 1980 et 2006 sur la modélisation des flux dans les SAUs ; voir aussi [Gul and Celik, 2020] pour une revue plus récente portant sur plus de 120 articles consacrés à la prévision des flux de patients en SAUs.

Les modèles ARMA et SARIMA sont couramment utilisés dans la littérature pour modéliser les flux de patients et restent les références auxquelles sont comparés les autres modèles. À notre connaissance, la première utilisation de ces modèles dans ce contexte et due à [Milner, 1988]. Le Tableau 1.1 présente un aperçu non exhaustif de travaux de modélisation de flux de patients utilisant des modèles ARMA/SARIMA. [Shi et al., 2011] a utilisé un modèle ARMA intégrant des facteurs socio-économiques, météorologiques et organisationnels pour prédire les flux de patients. [Earnest et al., 2005] modélisent par un ARMA le nombre de lits occupés sur la base de deux mois et demi de données. [Schull et al., 2005] concluent que la grippe a une influence sur les personnes âgées, ce qui met en évidence un effet saisonnier. Deux études récentes se concentrent sur les flux de patients aux urgences en France : [Kadri et al., 2014] utilisent un modèle ARMA pour prédire à horizon de 7 jours, le nombre d'admissions quotidiennes aux urgences pédiatriques de Lille ; [Afilal et al., 2016] comparent le modèle ARMA à un modèle hybride basé sur les séries de Fourier, en se basant sur cinq ans de données quotidiennes pour les urgences de Troyes.

Le modèle SARIMA a été largement utilisé, notamment à Singapour ; [Sun et al., 2009] appliquent un modèle SARIMA à des données de passages quotidiens aux urgences, en distinguant trois catégories de gravité de l'état des patients, et en intégrant des variables exogènes comme la qualité de l'air, l'humidité, etc. Le modèle optimal est identifié pour chaque catégorie de gravité. Aux États-Unis, [Jones et al., 2009] mènent une analyse multivariée avec un modèle VAR ; [Bergs et al., 2014] utilisent un lissage exponentiel automatisé pour prédire les passages mensuels ; [Champion et al., 2007] comparent les modèles ARIMA et de lissage exponentiel sur des données mensuelles d'un service d'urgence à Victoria, en Australie.

De nombreux autres modèles ont été explorés pour modéliser les flux de patients dans les services d'urgences à travers le monde. [Aboagye-Sarfo et al., 2015] compare le modèle ARMA et le modèle Holt-Winters ; [Rocha and Rodrigues, 2021], au Portugal, utilisent des réseaux neuronaux récurrents (RNN) ; [Jones et al., 2002], au Royaume-Uni, utilisent le modèle GARCH avec des données sur la température extérieure et la grippe pour prédire le nombre de lits occupés ; [Boyle et al., 2008] utilisent des modèles de régression ; [Jones et al., 2009] aux États-Unis réalisent une analyse multivariée en utilisant un modèle VAR ; [Bergs et al., 2014] utilisent le lissage exponentiel automatisé pour prédire les passages mensuels.

Dans le cadre du Master SEP, plusieurs étudiants ont déjà exploré la problématique de la modélisation ou de l'analyse des flux hospitaliers, en particulier dans les SAUs. [Sagno, 2020] ajuste un modèle SARIMA$(1,0,1)(0,1,1)_7$ à l'historique de fréquentation du SAU adultes du CHU de

Référence	Année	Pays - État/Ville	Modèle	Temporalité	Cible	Période de test et/ou Horizon de prévision	Erreur de prévision
[Milner, 1988]	1988	Royaume-Uni - Trent	ARIMA	Annuelle, 1974–1985	3 types : nouvelles visites, "re-visites" et visites totales	1986 à 1994	visites totales en 1986 : 4,3% (MAPE)
[Jones et al., 2002]	2002	Royaume-Uni - Londres	GARCH(2,0), couplé avec la température et les données sur la grippe	Quotidienne, 1993–1999	lits occupés	32 jours	3% (RMS)
[Earnest et al., 2005]	2005	Singapour	ARIMA(1,0,3)	Quotidienne, du 14 mars au 31 mai 2003	lits occupés pendant une épidémie de SRAS	21 avril au 31 mai	8,6% (MAPE)
[Champion et al., 2007]	2007	Australie - Victoria	Lissage exponentiel (ES) ARMA(0,1,1)	Mensuelle, 2000–2005	admissions aux urgences	5 mois	3,3% (MSE) pour ES 3,9% (RMSE) pour ARMA
[Boyle et al., 2008]	2008	Australie - Queensland	Régression avec variables explicatives (dont mois de l'année en variable muette)	Quotidienne, mais agrégée mensuellement pour la régression, du 01/07/2002 au 30/06/2007	2 hôpitaux, données d'admissions aux urgences	6 mois	1,79% (MAPE)
[Sun et al., 2009]	2009	Singapour	SARIMA par catégorie, avec variables exogènes (qualité de l'air, humidité, température moyenne)	Quotidienne, juillet 2005 à mars 2008	3 catégories de gravité des patients	6 mois	selon la catégorie : de 7% à 17% (MAPE)
[Jones et al., 2009]	2009	États-Unis	VAR avec variables exogènes (données opérationnelles des urgences). Prétraitement : saisonnalité et autocorrélation supprimées. Comparaison avec modèle Holt-Winters.	Horaire, 3 hôpitaux en 2006, 7920 heures (= 330 jours) en apprentissage	nombre d'admissions, nombre de lits occupés	840 heures (= 35 jours)	entre 2 et 4 (MAE) selon l'horizon (1h à 24h)
[Boyle et al., 2012]	2012	Australie - Queensland	ARIMA, régression, lissage exponentiel, prise en compte des jours fériés	Quotidienne, juillet 2002 à juillet 2007	2 hôpitaux pour l'apprentissage et 27 pour le test	2005–2008	admissions (MAPE) : mensuel : 2%, journalier : 11%, toutes les 4h : 38%, horaire : 50%
[Shi et al., 2011]	2011	Taïwan	ARIMA avec facteurs socio-économiques, météorologiques et organisationnels	Mensuelle, janvier 2005 à septembre 2009	4 types : pédiatrique, traumatologique, non-trauma et recettes	9 mois	(MAPE) : 29% pédiatrique, 12% trauma, 20% non-trauma, 23% recettes
[Kadri et al., 2014]	2014	France / Lille	ARMA(2,0,1)	Quotidienne, 2012	pédiatrique	7 jours	3,79 (RMSE)
[Bergs et al., 2014]	2014	Belgique	Lissage exponentiel automatisé	Mensuelle, 2005–2011	4 hôpitaux (5 ans d'apprentissage chacun)	12 mois	2,6% à 4,8% (MAPE)
[Aboagye-Sarfo et al., 2015]	2015	Australie	VARMA et modèle Holt-Winters	Mensuelle, 2006–2013	différents services d'urgences (dont pédiatrie, privé), avec nombre d'entrées par âge (4 groupes) et gravité, lieu de traitement	24 mois	1,8% à 4,5% (MAPE) selon le sous-groupe
[Afilal et al., 2016]	2016	France / Troyes	ARMA, modèle hybride avec séries de Fourier Modèle hybride avec séries de Fourier	Quotidienne, 2010–2014 5 ans : 2010–2014	flux de patients par pathologie	apprentissage : 2010–2013, test : 2014	Total (WMAPE) : 91,1% pour les modèles long terme et 91,8% court terme
[Rocha and Rodrigues, 2021]	2021	Portugal	Lissage exponentiel avec décomposition SARIMA Réseaux de neurones récurrents : AR-NN et RNN-1L XGBoost	Intervalles variés : 4h, 8h, 24h 2009–2018	admissions aux urgences, 1 hôpital	apprentissage : 2009–2016, validation : 2017, test : 2018	Erreur horaire (MAE) : 4,9 3,9 AR-NN : 4,4 et RNN-1L : 3,6 3,5

Table 1.1 : Synthèse des articles portant sur la modélisation des flux de patients aux urgences médicales.

Reims. Il a obtenu un taux d'erreur d'environ 7% pour un horizon de prédiction de 7 jours. Ce travail illustre l'efficacité de l'intégration de la saisonnalité hebdomadaire dans la modélisation des flux d'urgence. [Devie, 2020] réalise une analyse descriptive des flux de patients au sein du CHU Châlons-en-Champagne. Son étude s'est attachée à caractériser finement les profils de fréquentation contribuant ainsi à une meilleure compréhension des dynamiques locales propres à un SAU donné.

Pour finir, [Alloux, 2021] compare deux modèles SARIMA, de saisonnalité 7 et 365 pour modéliser et prédire les flux dans le SAU adultes du CHU de Reims. Le modèle SARIMA$(5,1,0)(0,1,0)_365$ a été retenu, affichant un taux d'erreur de 8% pour un horizon de prédiction de 7 jours. Son travail met en lumière les limites des approches classiques face à la forte variabilité des données. L'analyse souligne l'importance d'une bonne décomposition des séries temporelles ainsi que le rôle des effets calendaires et épidémiques dans la dynamique des flux.

Ces travaux constituent des étapes clés dans le développement d'une expertise académique autour des urgences hospitalières, en lien étroit avec les enjeux de terrain portés par Est-RESCUE et l'ARS Grand Est.

L'objectif de cette thèse est de modéliser et de prédire les flux de passages dans une cinquantaine de SAUs de la région Grand Est, à des horizons de 7, 14, 21 et 28 jours. L'étude débute par une approche univariée, à l'aide de modèles classiques de séries temporelles, permettant d'établir une base comparative. Dans ce cadre, nous proposons notamment un modèle additif généralisé (GAM) pour les flux quotidiens de patients dans les SAUs. Ce modèle intègre plusieurs composantes : une tendance polynomiale, une saisonnalité construite comme la superposition d'harmoniques apprises à partir des données, ainsi qu'une composante événementielle permettant de capturer l'influence de périodes particulières comme les vacances scolaires ou certains événements ponctuels. Enfin, une modélisation multivariée est proposée, fondée sur les hyperparamètres identifiés précédemment du modèle TSHR, dans le but de tirer parti des liens entre services et d'améliorer la précision des prévisions.

Le corps de la thèse est structuré en trois chapitres principaux, correspondant aux différentes étapes de la modélisation des flux de passages dans les services d'urgence.

Le chapitre 2 est consacré à la modélisation univariée des flux. Il présente les modèles mobilisés, notamment SARIMA, Prophet et le modèle TSHR développé dans cette thèse, ainsi que leurs fondements méthodologiques. Chaque modèle est décrit en détail afin de poser les bases de leur mise en oeuvre. Le chapitre 3 applique ces modèles à l'ensemble des données dispo-

nibles. Il décrit les procédures d'ajustement et de sélection des hyperparamètres mises en place, ainsi qu'une procédure d'évaluation de la performance de prévision par validation croisée séquentielle. Les performances prédictives sont ensuite analysées, ce qui permet de comparer les modèles entre eux et d'identifier les configurations les plus pertinentes. Une analyse multivariée a posteriori des modèles les plus performants est également réalisée afin d'explorer les corrélations potentielles entre services. Le chapitre 4 propose une modélisation multivariée des flux de passages. Une étape préalable de clustering permet d'identifier des groupes homogènes de SAUs, caractérisés par des dynamiques temporelles similaires. Pour chaque groupe, une version multi-variée du modèle TSHR est développée pour la modélisation des flux correspondants via une régression PLS pour des ensembles de régresseurs bien choisis.

Chapitre 2

Modélisation univariée des flux de passage

Dans ce chapitre, nous présentons le modèle additif généralisé univarié nommé TSHR que nous proposons pour modéliser les flux de patients des services d'urgence (partie 2.2.3), ainsi que les trois autres modèles auxquels il sera comparé, à savoir les modèles SNAIVE, SARIMA (partie 2.2.1) et PROPHET (partie 2.2.2).

2.1 Notations

Introduisons les notations qui seront utilisées dans ce chapitre et les suivants. L'ensemble des Services Accueil d'Urgences (SAUs) est noté $\mathcal{D}$. Nous désignons de manière générique par d tout élément de $\mathcal{D}$, c'est-à-dire, tout SAU. Pour tout d, la suite des nombres quotidiens successifs de patients pris en charge par le SAU d pendant une période fixe de T jours, T étant un entier positif, est notée

$$(x_{1:T}^{(d)}) = (x_t^{(d)})_{t\in\{1,\dots,T\}}.$$

Nous supposons implicitement que le jour $t = 1$ correspond au 1er janvier 2016, la première date pour laquelle nous disposons d'enregistrements dans notre ensemble de données. Les enregistrements disponibles s'arrêtent au 31 décembre 2018, correspondant à $T = T_{\max} := 1096$ jours.

Comme première étape dans la modélisation des flux de patients, nous supposons que, pour tout $d \in \mathcal{D}$, la suite $(x_{1:T}^{(d)})$ correspond aux T premières coordonnées de la réalisation d'un

processus stochastique

$$\mathbf{X}^{(d)} = (X_t^{(d)})_{t\in\mathbb{N}^*},$$

défini sur un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$, c'est-à-dire,

$$x_t^{(d)} = X_t^{(d)}(\omega), \quad t = 1, \dots, T,$$

où $\omega \in \Omega$. Ainsi, par modélisation des flux de patients, nous entendons décrire mathématiquement le comportement stochastique des processus $\mathbf{X}^{(d)}$, $d \in \mathcal{D}$.

Lorsque nous ne prêtons pas une attention particulière à un service donné, nous omettons l'exposant $^{(d)}$ dans les notations précédentes. Ainsi, $(x_t)_{t\in\{1,\dots,T\}}$ désignera la suite des nombres quotidiens de patients dans un service générique et non spécifié, et $\mathbf{X} := (X_t)_{t\in\mathbb{N}^*}$, sa version stochastique.

Nous considérerons souvent des sous-suites de $(x_t)_{t\in\{1,\dots,T\}}$ ou $(X_t)_{t\in\mathbb{N}^*}$. En particulier, pour tout $t_1, t_2 \in \{1, \dots, T_{\max}\}$, avec $t_1 < t_2$, nous notons $x_{t_1:t_2} := (x_{t_1}, x_{t_1+1}, \dots, x_{t_2-1}, x_{t_2})$ la suite des valeurs successives de x_t entre t_1 et t_2. Des notations similaires sont utilisées pour les suites extraites de $\mathbf{X}$.

2.2 Présentation des modèles

2.2.1 SARIMA et SNAIVE

Les modèles ARMA, ARIMA et SARIMA sont largement utilisés dans de nombreux domaines pour modéliser des séries temporelles. Une littérature prolifique existe sur les aspects théoriques et pratiques de ces modèles stochastiques. Une présentation précise et mathématiquement rigoureuse des modèles ARMA et SARIMA ainsi que des questions liées aux problèmes d'estimation et de prévision peuvent être trouvées dans [Brockwell and Davis, 1991] et [Brockwell and Davis, 2016]. Comme mentionné précédemment dans la partie 1.4, les modèles ARMA et SARIMA ont été largement utilisés pour modéliser les flux de patients dans les services d'urgences hospitaliers ; voir le Tableau 1.1 pour un aperçu.

L'acronyme ARMA est la concaténation de AR pour « autoregressive » (auto-régressif) et MA pour « moving average » (moyenne mobile). Heuristiquement, un processus stochastique suit un modèle ARMA s'il est défini par une relation récurrente linéaire sur ses coordonnées (la partie AR) et sur les coordonnées du processus résiduel (la partie MA). Les modèles ARMA sont indexés par deux hyperparamètres, p et q, représentant les ordres des relations récurrentes

linéaires des parties AR et MA, respectivement. Formellement, un processus stochastique $\mathbf{X} = (X_t)_{t\in\mathbb{Z}}$ suit un modèle ARMA(p,q), $p, q \in \mathbb{N}^*$, s'il existe des paramètres $\phi = (\phi_1, \ldots, \phi_p) \in \mathbb{R}^p$ et $\theta = (\theta_1, \ldots, \theta_q) \in \mathbb{R}^q$ avec $\phi_p \neq 0$ et $\theta_q \neq 0$, $c \in \mathbb{R}$ et un processus de bruit blanc $\boldsymbol{\epsilon} = (\epsilon_t)_{t\in\mathbb{Z}}$ (i.e., une suite de variables aléatoires, centrées et de variance finie σ^2), tels que

$$X_t = c + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}, \quad t \in \mathbb{Z}. \tag{2.1}$$

Pour plus de commodité, la relation précédente (2.1) est présentée généralement de manière plus compacte en utilisant les notations de la théorie des opérateurs, à savoir

$$\phi(B)\mathbf{X} = c + \theta(B)\boldsymbol{\epsilon}, \tag{2.2}$$

avec $\phi(B) = 1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p$ et $\theta(B) = 1+\theta_1 B+\theta_2 B^2+\cdots+\theta_q B^q$ sont des polynômes de l'opérateur de décalage "arrière" B agissant sur une suite en décalant ses coordonnées, c'est-à-dire $B(X_t)_{t\in\mathbb{Z}} = (X_{t-1})_{t\in\mathbb{Z}}$.

Par extension les modèles ARMA de paramètres $(p, 0)$ et $(0, q)$ désignent respectivement des modèles AR(p) and MA(q), correspondant, à $\theta(B) = 1$ et $\Phi(B) = 1$ dans (2.2). Un modèle ARMA(0,0) désigne un bruit blanc (une suite de variables aléatoires non corrélées, centrées de variance finie).

Sous certaines hypothèses supplémentaires sur ses paramètres ϕ et θ, qui ne sont pas détaillées ici mais simplement désignées par le terme “causalité”, un modèle ARMA est (faiblement) stationnaire, signifiant que toutes les coordonnées X_t ont la même espérance (indépendante de t) et la même fonction d'auto-corrélation (indépendante du temps t, dépendant uniquement du décalage temporel τ).

Le « I » des modèles ARIMA signifie « intégré ». Une suite $\mathbf{X} = (X_t)_{t\in\mathbb{Z}}$ est un processus stochastique suivant un modèle ARIMA avec les hyperparamètres (p, d, q), noté ARIMA(p, d, q), si le processus stochastique dérivé $\mathbf{Y} = (1 - B)^d\mathbf{X}$ suit un modèle ARMA(p, q) causal. L'opérateur $(1 - B)^d$ est un opérateur de dérivation discrète itérée. La suite $\mathbf{X}$ peut présenter une tendance linéaire ou plus complexe la rendant non stationnaire. L'application de $(1 - B)^d$ à $\mathbf{X}$ rend le processus dérivé $\mathbf{Y} = (1 - B)^d\mathbf{X}$ un processus stationnaire, pour lequel les méthodes d'estimation et de prévision peuvent être appliquées, puis étendues à $\mathbf{X}$ par « intégration » de $\mathbf{Y}$. Formellement, un processus ARIMA(p, d, q) $\mathbf{X}$ est décrit par la relation suivante

$$\phi(B)(1 - B)^d\mathbf{X} = \theta(B)\boldsymbol{\epsilon}. \tag{2.3}$$

Enfin, les modèles SARIMA incluent une composante saisonnière dans la relation récursive de $\mathbf{X}$. Les modèles SARIMA sont indexés par un ensemble de sept hyperparamètres entiers

p, d, q, P, D, Q, s, souvent notés $(p, d, q)(P, D, Q)_s$, où s représente la saisonnalité — ou la périodicité — de la suite, par exemple $s = 7$ pour une saisonnalité hebdomadaire. Une suite $\mathbf{X}$ est un processus SARIMA$(p, d, q)(P, D, Q)_s$ si la suite $\mathbf{Y} := (1 - B)^d(1 - B^s)^D\mathbf{X}$ est une suite ARMA causale. Un tel processus est décrit par la relation récursive

$$\phi(B)\Phi(B^s)(1 - B)^d(1 - B^s)^D\mathbf{X} = \theta(B)\Theta(B^s)\boldsymbol{\epsilon}, \tag{2.4}$$

où ϕ, Φ, θ et Θ sont des polynômes de degrés respectifs p, P, q et Q, dont le terme constant est égal à 1, les autres coefficients $(\phi_1, \ldots, \phi_p)$, $(\Phi_1, \ldots, \Phi_P)$, $(\theta_1, \ldots, \theta_q)$ et $(\Theta_1, \ldots, \Theta_Q)$ sont les paramètres du modèle. Les triplets d'hyperparamètres (p, d, q) et (P, D, Q) décrivent la complexité des parties non périodiques et périodiques, respectivement, dans la relation récursive.

Le modèle particulier, SARIMA$(0, 0, 0)(0, 1, 0)_s$, s'appelle le modèle SNAIVE de paramètre (de saisonnalité) s.

Les modèles SARIMA sont très populaires en raison de leur forme simple et de leur capacité à prendre en compte des motifs généraux ou adaptés aux comportements humains, tels que les tendances et la saisonnalité hebdomadaire (ou toute autre périodicité). L'ajustement de ces modèles aux données est pris en charge par de nombreux logiciels, dont R ([R Core Team, 2024]). Des procédures de sélection de modèles sont également spécifiquement conçues et implémentées ; voir par exemple la fonction `auto.arima()` du package `{forecast}` ([Hyndman and Khandakar, 2008]). Ces modèles bénéficient également d'une base théorique solide, permettant la construction d'intervalles de confiance.

En résumé, les familles de modèles ARMA et SARIMA dépendent d'hyperparamètres, respectivement (p, q) et $(p, d, q)(P, D, Q)_s$, où $p, d, q, P, D, Q \in \mathbb{N}$ et $s \in \mathbb{N}^*$. Pour des hyperparamètres donnés, les modèles ARMA et SARIMA sont caractérisés par des paramètres respectivement (ϕ, θ) et $(\phi, \theta, \Phi, \Theta)$, où $\phi \in \mathbb{R}^p$, $\theta \in \mathbb{R}^q$, $\Phi \in \mathbb{R}^P$ et $\Theta \in \mathbb{R}^Q$.

Nous décrivons dans la section 2.3 la procédure que nous avons suivie pour ajuster les modèles ARMA ou SARIMA à nos données.

2.2.2 PROPHET

PROPHET est un modèle additif généralisé (GAM) introduit par [Taylor and Letham, 2018]. Il décompose additivement une série temporelle en quatre termes : une composante de tendance, une composante saisonnière, une composante événementielle (jours fériés/événements), et un résidu. Précisément, un processus stochastique $\mathbf{X}$ suit un modèle PROPHET s'il existe des

fonctions $\mathbf{G} = (G_t)_t$, $\mathbf{S} = (S_t)_t$, $\mathbf{H} = (H_t)_t$ et un bruit blanc gaussien $\boldsymbol{\epsilon} = (\epsilon_t)_t$ tels que

$$X_t = G_t + S_t + H_t + \epsilon_t, \quad t \in \mathbb{N}^*. \tag{2.5}$$

La tendance G_t peut être une fonction linéaire par morceaux du temps t, ou une fonction logistique modélisant une croissance non linéaire saturante

$$G_t = \frac{C(t)}{1 + \exp(-k(t)(t-m))}, \quad t \in \mathbb{N}^*,$$

où m est un paramètre de décalage, C est la borne supérieure correspondant à la limite de saturation, et k est le taux de croissance pouvant dépendre du temps. Ce dernier est supposé être constant par morceaux, chaque longueur de morceau étant soit dérivée de connaissances expertes, soit tirée d'une distribution de Laplace (paramètre par défaut).

La composante saisonnière S_t est définie par

$$S_t = \sum_{k=1}^{N} \left(a_k \cos\left(\frac{2\pi k t}{P}\right) + b_k \sin\left(\frac{2\pi k t}{P}\right) \right), \quad t \in \mathbb{N}^*,$$

où $N \in \mathbb{N}^*$ un hyperparamètre, a_k, b_k avec $k = 1, \ldots, N$, $2N$ paramètres réels, et P la période saisonnière (par exemple $P = 7$ pour des effets hebdomadaires). Plusieurs périodes peuvent être superposées si nécessaire.

Enfin, la composante jours fériés/événements est une combinaison linéaire de fonctions indicatrices

$$H_t = \sum_{h \in \mathcal{H}} b_h \mathbb{1}_{D_h}(t), \quad t \in \mathbb{N}^*,$$

où $\mathcal{H}$ est un ensemble d'événements/jours fériés, chaque D_h étant associé à un temps fériés/événements t_h ou une fenêtre temporelle autour de celui-ci, i.e., $D_h = t_h \pm w$, avec $w \in \mathbb{N}$ la taille de la fenêtre, nulle par défaut.

Une procédure bayésienne est proposée, dans [Taylor and Letham, 2018] et [Rafferty, 2023], pour ajuster le modèle PROPHET (2.5) aux séries temporelles et effectuer des prévisions ; celle-ci n'est pas présentée ici pour des raisons de concision. Cette procédure est prise en charge par une implémentation open-source pour les langages Python et R ; voir par exemple la fonction `prophet()` du package `{prophet}` de R ; voir [Taylor and Letham, 2021].

Dans notre application, nous avons utilisé les valeurs par défaut des hyperparamètres, pour l'ajustement du modèle PROPHET aux flux de patients des services accueil d'urgences ; voir chapitre 3.

2.2.3 Modèles Additifs Généralisés avec tendance polynomiale et saisonnalité basée sur des harmoniques empiriques

Nous proposons ici un autre modèle GAM, nommé TSHR dans la suite, en référence à ses quatre composants. Un processus stochastique $\mathbf{X}$ suit un modèle TSHR s'il se décompose en la somme de quatre termes,

$$\mathbf{X} = \mathbf{T} + \mathbf{S} + \mathbf{H} + \mathbf{R}, \tag{2.6}$$

tels que

- $\mathbf{T} = (T_t)_{t\in\mathbb{N}^*}$ modélise la tendance à moyen et long terme du processus par une fonction polynomiale de t. Précisément, étant donnés

  $$P_0(t) = 1, P_1(t), \ldots, P_K(t), \tag{2.7}$$

  $K+1$ polynômes orthogonaux de degrés croissants $K = 0, \ldots, K_{\max}$, nous supposons que

  $$T_t := T_t^{(K)}(\mathbf{a}) := \sum_{k=0}^{K} a_k P_k(t), \quad t \in \mathbb{N}^*, \tag{2.8}$$

  où $\mathbf{a} := (a_0, a_1, \ldots, a_K) \in \mathbb{R}^{K+1}$.

- $\mathbf{S} = (S_t)_{t\in\mathbb{N}^*}$ modélise le comportement saisonnier du processus. Nous supposons que cette composante saisonnière est une somme d'harmoniques. Précisément,

  $$S_t := S_t(F_{(m)}, \mathbf{b}_{(m)}, \mathbf{c}_{(m)}) := \sum_{f\in F_{(m)}} (b_f \cos(2\pi f t) + c_f \sin(2\pi f t)), \quad t \in \mathbb{N}^*, \tag{2.9}$$

  où $F_{(m)}$ est un sous-ensemble constitué de m fréquences et $\mathbf{b} := (b_f)_{f\in F_{(m)}} \in \mathbb{R}^m$ et $\mathbf{c} := (c_f)_{f\in F_{(m)}} \in \mathbb{R}^m$ sont des paramètres réels. Dans le contexte de la modélisation des séries temporelles, $F_{(m)}$ est un sous-ensemble, composé de m éléments, de l'ensemble

  $$\left\{\frac{1}{T}, \frac{2}{T}, \ldots, \frac{\lfloor (T-1)/2 \rfloor}{T}\right\}, \tag{2.10}$$

  T étant la taille de la série temporelle d'entraînement. La notation $\lfloor x \rfloor$ désigne la partie entière de x, pour tout nombre réel x . Le terme constant dans (2.9) est omis car il est pris en compte par la première composante (2.8);

- $\mathbf{H} = (H_t)_{t\in\mathbb{N}^*}$ modélise les effets locaux (dans le temps) des événements à des dates spécifiques. Précisément,

  $$H_t := H_t(\mathcal{H}, \mathbf{d}_{\mathcal{H}}) := \begin{cases} \sum_{h\in\mathcal{H}} d_h \mathbb{1}_h(t) & \text{si } \mathcal{H} \neq \emptyset, \\ 0 & \text{si } \mathcal{H} = \emptyset, \end{cases}, \quad t \in \mathbb{N}^*, \tag{2.11}$$

où $\mathcal{H}$ est un sous-ensemble d'un ensemble $\mathbb{H}$ d'événements/jours fériés spécifiques, tels que les vacances scolaires, le jour de Noël, les fêtes nationales, etc. Pour tout $h \in \mathbb{H}$, $\mathbb{1}_h(\cdot)$ est la fonction indicatrice de la période h, c'est-à-dire que $\mathbb{1}_h(t) = 1$ si $t \in h$ et 0 ailleurs. $\mathbf{d}_\mathcal{H} := (d_h)_{h\in\mathcal{H}}$ est un paramètre à valeurs dans $\mathbb{R}^{n_\mathcal{H}}$ quantifiant les flux de patients spécifiques aux périodes de jours fériés $h \in \mathcal{H}$, où $n_\mathcal{H}$ est la taille de l'ensemble $\mathcal{H}$. À titre d'exemple, le tableau 2.1 présente l'ensemble $\mathbb{H}$ des périodes de vacances et jours fériés considérés (restreintes à l'année 2016).

Jours fériés	Périodes
Vacances de Noël	17/12 à 31/12 et 01/01 à 03/01
Vacances d'hiver	06/02 à 21/02
Vacances de printemps	02/04 à 17/04
Vacances d'été	06/07 à 01/09
Vacances de la Toussaint	20/10 à 02/11
Jours fériés	01/01, 28/03, 01/05, 05/05, 08/05, 16/05, 14/07, 15/08, 01/11, 11/11 et 25/12

TABLE 2.1 : Ensemble $\mathbb{H}$ des six types de périodes de vacances et jours fériés et les dates correspondantes pour l'année 2016 (les dates pour les années 2017 et 2018 sont omises).

- Le terme résiduel $\mathbf{R} = (R_t)_{t\in\mathbb{N}^*}$, avec $R_t := X_t - T_t - S_t - H_t$, est supposé être un processus stochastique de type ARMA(p,q), à savoir

$$R_t = \sum_{i=1}^{p} \phi_i \, R_{t-i} + \epsilon_t + \sum_{j=1}^{q} \theta_j \, \epsilon_{t-j},$$

où $\phi := (\phi_1, \ldots, \phi_p) \in \mathbb{R}^p$, $\theta := (\theta_1, \ldots, \theta_q) \in \mathbb{R}^q$ et $(\epsilon_t)_t$ est un bruit blanc.

En résumé, la famille des modèles TSHR dépend des hyperparamètres $K \in \mathbb{N}$, le degré maximal de la fonction polynomiale qui modélise la tendance à moyen et long terme, $F_{(m)}$, le sous-ensemble des harmoniques à conserver dans la partie saisonnière, le sous-ensemble $\mathcal{H}$ de l'ensemble $\mathbb{H}$ des périodes de jours fériés/événements et $(p, q) \in \mathbb{N}^2$ les hyperparamètres des résidus ARMA $\mathbf{R}$. Les paramètres du modèle TSHR(K, $F_{(m)}$, $\mathcal{H}$, p, q) sont $\mathbf{a}_{(K)} := a_{0:K} \in \mathbb{R}^{K+1}$ les coefficients de la tendance polynomiale, $(\mathbf{b}_{(m)}, \mathbf{c}_{(m)}) \in \mathbb{R}^{2m}$ les amplitudes des effets périodiques, $\mathbf{d}_\mathcal{H} = d_{1:n_\mathcal{H}} \in \mathbb{R}^{n_\mathcal{H}}$ les effets des périodes de vacances/événements, et $\phi \in \mathbb{R}^p, \theta \in \mathbb{R}^q$ les paramètres de la composante ARMA R_t.

Rappelons que les modèles doivent être ajustés de manière univariée à la suite des flux de chaque service d'urgence. Cela signifie que les hyperparamètres et les paramètres des quatre composantes sont spécifiques à chaque service accueil d'urgence.

Les prévisions seront effectuées à un certain horizon, c'est-à-dire le nombre (désigné ci-après par H) de jours futurs à prévoir à l'aide d'un modèle ajusté sur les données historiques, disons prévision à $D + H$ jours. Dans la partie 3.3, les performances de prévision du modèle proposé TSHR seront comparées à celles des modèles classiques SARIMA, PROPHET et SNAIVE, pour différentes valeurs de l'horizon H ; dans nos applications, $H = 7, 14, 21$ et 28.

2.2.4 Comparaisons formelles des modèles

Avant de comparer les performances de prévision des trois modèles TSHR, PROPHET et SARIMA, dans la section 3.3, nous soulignons brièvement leurs caractéristiques théoriques discriminantes. D'une part, les avantages des GAM (PROPHET et TSHR), incluent

- la capacité de prendre en compte l'incidence d'événements ponctuels éventuels sur les flux de patients (composante **H**) ;
- les tendances à moyen et long terme peuvent être décrites avec précision et peuvent intégrer des connaissances d'experts (choix de la forme de la tendance) ;
- possibilité de modéliser une saisonnalité multiple ;
- les paramètres sont faciles à interpréter.

D'autre part, les avantages des SARIMA sont les suivants

- l'effet saisonnier peut évoluer dans le temps ;
- la capacité de prendre explicitement en compte la dépendance temporelle historique, en utilisant les termes AR et MA dans son modèle ;
- la disponibilité d'implémentations pour l'ajustement, la sélection de modèles et les procédures de prévision dans de nombreux langages de programmation, ce qui rend leur utilisation facile à reproduire.

PROPHET et TSHR diffèrent dans la manière dont la tendance, la saisonnalité, les jours fériés et les résidus sont gérés. Précisément,

- les points de rupture dans la tendance de PROPHET offrent la possibilité d'ajuster l'évolution de la tendance avec les connaissances d'experts et d'introduire des irrégularités ;

- les périodes de saisonnalité dans PROPHET doivent être définies par l'utilisateur, tandis que les périodes les plus pertinentes sont extraites des données pour TSHR ;
- PROPHET permet d'étendre les effets des jours fériés sur un nombre prescrit de jours avant et après la période des jours fériés ;
- les résidus ARMA dans TSHR peuvent être considérés comme une modélisation de deuxième passage pour capturer les effets de l'auto-corrélation ;
- Pour notre modèle TSHR proposé, les étapes d'ajustement (estimation des paramètres), de sélection de modèle et de prévision sont très rapides comparées à celles des modèles PROPHET et SARIMA. De plus, TSH+ARMA(0,0) peut être étendu à la modélisation de séries temporelles multivariées ; voir chapitre 4.

TSHR	SARIMA
• Prise en compte d'événements exogènes. • Flexibilité sur la forme des tendances à moyen et long terme permettant d'intégrer les connaissances d'experts. • Modélisation simple de la saisonnalité multiple. • Paramètres faciles à interpréter	• Effet saisonnier évolue dans le temps. • C'est un modèle classique, qui peut être facilement reproduit dans d'autres langages de programmation. • SARIMA prend explicitement en compte la dépendance temporelle historique, en utilisant les termes AR et MA dans son modèle.

TABLE 2.2 : Caractéristiques formelles distinguant les modèles TSHR et SARIMA.

2.3 Sélection de modèles et ajustement aux données

Les trois familles de modèles décrites dans la Section 2.2 dépendent d'hyperparamètres et de paramètres. Ajuster ces modèles aux historiques de fréquentation des SAUs requiert de choisir de manière appropriée les hyperparamètres de ces familles et estimer les paramètres correspondants de manière cohérente. Ces deux opérations (choix des hyperparamètres, estimation des paramètres) doivent faire face à deux objectifs adverses : d'une part, un modèle avec un grand nombre de paramètres (c'est-à-dire de grandes valeurs des hyperparamètres) devrait théoriquement réussir à s'adapter précisément aux données. D'autre part, un modèle aussi grand conduit généralement à un sur-ajustement (le modèle choisi est tellement adapté aux données à partir desquelles il a été entraîné qu'il ne représente plus correctement le phénomène de manière générique), ce qui, dans la pratique, conduit à de mauvaises performances en matière de prévision. L'arbitrage est classiquement géré en deux étapes

1. Étape d'estimation : Pour chaque valeur possible des hyperparamètres d'une famille de

modèles donnée, les paramètres du modèle sont estimés à partir d'un ensemble de données d'entraînement.

2. Étape de sélection du modèle : Ensuite, le « meilleur » choix d'hyperparamètres est retenu. Ce que l'on entend par « meilleur » dépend de la famille de modèles considérée et du critère de sélection utilisé. Pour les modèles SARIMA et TSHR, la sélection du modèle est effectuée en minimisant une fonction de coût incluant un terme de pénalisation croissant avec le nombre de paramètres du modèle (il s'agit typiquement de l'opposé du double de la log-vraisemblance avec une pénalisation de type BIC). Pour PROPHET, les valeurs par défaut des hyperparamètres fournies dans l'implémentation du package R ont été conservées, l'étape d'estimation étant coûteuse en temps de calcul (se basant sur une approche bayésienne) ; voir [Rafferty, 2023].

Cette procédure en deux étapes est coûteuse en temps de calcul, en particulier pour le modèle SARIMA (voir la section 2.3.1), et nécessite donc de restreindre les valeurs possibles des hyperparamètres des modèles à un ensemble réduit de valeurs ; Le Tableau 2.3 identifie de manière synthétique les hyperparamètres des modèle SARIMA et TSHR, ainsi que leurs valeurs considérées. La méthode de sélection de modèles utilisée pour les modèles SARIMA et TSHR sont explicitement décrites dans les sections 2.3.1 et 2.3.2, respectivement.

	SARIMA							TSHR				
Hyperparamètres	p	d	q	P	D	Q	s	K	m	$n_{\mathcal{H}}$	p	q
Valeur minimale	0	0	0	0	0	0	7	0	0	0	0	0
Valeur maximale	6	1	6	3	1	3	7	4	30	6	3	3

Table 2.3 : Plages de valeurs pour les hyperparamètres des modèles SARIMA et TSHR définis dans les parties 2.2.1 et 2.2.3.

2.3.1 Sélection de modèles SARIMA

La sélection des modèles SARIMA a été effectuée à l'aide d'une procédure classique de sélection basée sur le Critère d'Information Bayésien (BIC), suivie d'une étape de post-traitement

- Pour chaque SAU $d \in \mathcal{D}$ et pour chaque valeur possible des hyperparamètres $(p, d, q)(P, D, Q)_s$, voir le Tableau 2.3, les estimateurs du maximum de vraisemblance des paramètres du modèle SARIMA sont calculés pour la suite d'entraînement $\mathbf{x}_{1:T}^{(d)}$, avec $T = 1008$ grâce à la fonction `arima()` du package `{stats}` du logiciel statistique R ([R Core Team, 2024]).

- Les modèles sont ensuite filtrés pour ne retenir que ceux pour lesquels l'algorithme de maximisation de la vraisemblance converge. Parmi les modèles ainsi sélectionnés, celui ayant le BIC le plus faible est retenu.

Il convient de noter que la première étape de cette procédure de sélection de modèles, appliquée aux $\#\mathcal{D} = 51$ SAUs de la région Grand Est, est chronophage. Les calculs ont été réalisés à l'aide d'une procédure parallélisée en mémoire partagée sur un processeur unique doté de 28 cœurs du centre HPC régional ROMEO[1], nécessitant environ 15 heures de calcul.

2.3.2 Sélection de modèles TSHR et estimation des paramètres

La procédure de sélection de modèle TSHR se compose de quatre étapes imbriquées, consistant à choisir successivement les hyperparamètres de chaque composante additive (tendance, saisonnalité, jours fériés et ARMA), en tenant compte des résultats de l'étape précédente et en ajustant progressivement le choix des hyperparamètres sur les résidus de l'étape antérieure :

1) Sélection de K (le degré du polynôme de la tendance). Pour chaque degré du polynôme $K = 0, 1, \ldots, K_{max}$, avec $K_{max} = 4$ dans notre application, on utilise la méthode des moindres carrés ordinaires (MCO) pour estimer les paramètres de la composante de tendance $T_t^{(K)}$, à partir de la séquence d'apprentissage $x_{1:T}$, avec $T = 1008$ dans notre application, comme suit :

$$\widehat{\mathbf{a}}_{(K)} := \underset{\mathbf{a}_{(K)} \in \mathbb{R}^{K+1}}{\arg\min} \sum_{t=1}^{T} \left(x_t - \sum_{k=0}^{K} a_k P_k(t) \right)^2 .$$

Puis, on calcule pour chaque $K = 0, 1, \ldots, K_{max}$, le critère d'information bayésien (BIC) :

$$\mathrm{BIC}(K) := T\,(1 + \log(2\pi)) + T\,\log\left(\frac{\sum_{t=1}^{T} \widehat{\epsilon}_t^2}{T}\right) + \log(T)\,(1 + K), \tag{2.12}$$

où

$$\widehat{\epsilon}_t = x_t - \sum_{k=0}^{K} \widehat{a}_k P_k(t), \quad t = 1, \ldots, T.$$

La valeur « optimale » $\widehat{K}$ du degré K est alors définie par :

$$\widehat{K} := \underset{K \in \{0,1,\ldots,K_{\max}\}}{\arg\min} \mathrm{BIC}(K).$$

[1] Le centre HPC ROMEO est géré par l'Université de Reims Champagne-Ardenne depuis 2002. Il fournit aux chercheurs de l'Université et aux partenaires industriels une plateforme de calcul intensif de renommée mondiale. ROMEO2018 se compose de 117 serveurs équipés de processeurs Intel® Skylake, 6132 processeurs à 2,6 GHz pour un total de 3276 coeurs, de 280 accélérateurs Nvidia P100 SXM2 interconnectés par la technologie NVlink, et de mémoires DDR4 de 2667 MT/s pour une capacité totale de mémoire distribuée de 15,3 To.

2) Sélection des harmoniques de la composante saisonnière S_t. Elle s'effectue à partir de la série des résidus, notée $y_{1:T}$ obtenue, en retirant à l'historique des fréquentations $x_{1:T}$, la tendance polynomiale sélectionnée à l'étape précédente :

$$y_t := x_t - T_t^{(\widehat{K})}(\widehat{\mathbf{a}}_{(\widehat{K})}) = x_t - \sum_{k=0}^{\widehat{K}} \widehat{a}_k P_k(t), \quad t = 1, \dots, T. \tag{2.13}$$

On commence par calculer la transformée de Fourier discrète (TFD) de la séquence $y_{1:T}$,

$$\mathrm{F}(k/T) := \frac{1}{T} \sum_{t=1}^{T} y_t \, \exp\left(-2\pi j \frac{k}{T}(t-1)\right), \quad k = 0, 1, \dots, T-1,$$

où j est le nombre imaginaire tel que $j^2 = -1$. Remarquons que, $F(0/T) = \frac{1}{T}\sum_{t=1}^{T} y_t = 0$. Les fréquences k/T, $k = 1, \dots, \lfloor (T-1)/2 \rfloor$, sont ensuite classées (par ordre décroissant) selon leur « niveau d'énergie » (module du nombre complexe $F(k/T)$, $k = 1, \dots, \lfloor (T-1)/2 \rfloor$)

$$|\mathrm{F}(k/T)|, \quad k = 1, \dots, \lfloor (T-1)/2 \rfloor. \tag{2.14}$$

On note alors

$$f_1, f_2, \dots, f_{\lfloor (T-1)/2 \rfloor}, \tag{2.15}$$

la suite finie des fréquences ordonnées, et

$$F_{(m)} := \{f_1, \dots, f_m\}, \quad m = 1, \dots, m_{\max}, \tag{2.16}$$

la suite croissante des ensembles de fréquences correspondants, où $m_{\max}$ est le nombre maximal de fréquences considérées (on choisit $m_{\max} = 30$ dans notre application).

Cela permet de construire des modèles croissants pour la composante saisonnière $\mathbf{S}$:

$$S_t = S_t(F_{(m)}, \mathbf{b}_{(m)}, \mathbf{c}_{(m)}) = \sum_{i=1}^{m} \left(b_i \cos(2\pi f_i \, t) + c_i \sin(2\pi f_i \, t)\right), \quad t = 1, \dots, T, \quad m = 1, \dots, m_{\max}. \tag{2.17}$$

Pour $m = 0$, on pose $S_t(F_{(0)}, \mathbf{c}_{(0)}, \mathbf{d}_{(0)}) = 0$, pour tout $t = 1, \dots, T$.

Le problème revient alors à sélectionner le nombre optimal de fréquences $m \in \{0, \dots, m_{\max}\}$. Pour chaque $m = 0, \dots, m_{\max}$, on estime les paramètres $\mathbf{a}_{(\widehat{K})} := (a_0, \dots, a_{\widehat{K}})$, $\mathbf{b}_{(m)} := (b_1, \dots, b_m)$ et $\mathbf{c}_{(m)} := (c_1, \dots, c_m)$ par la méthode MCO, soit

$$(\widehat{\mathbf{a}}_{(\widehat{K})}, \widehat{\mathbf{b}}_m, \widehat{\mathbf{c}}_m) := \underset{\mathbf{a}_{(\widehat{K})} \in \mathbb{R}^{\widehat{K}+1}, (\mathbf{b}_{(m)}, \mathbf{c}_{(m)}) \in \mathbb{R}^{2m}}{\arg\min} \sum_{t=1}^{T} \left(x_t - \sum_{k=0}^{\widehat{K}} a_k P_k(t) - S_t(F_{(m)}, \mathbf{b}_{(m)}, \mathbf{c}_{(m)}) \right)^2.$$

On calcule ensuite, pour chaque $m = 0, \dots, m_{max}$, la valeur $\mathrm{BIC}(m)$ du critère BIC (2.12) avec

$$\widehat{\epsilon}_t = x_t - \sum_{k=0}^{\widehat{K}} \widehat{a}_k P_k(t) - \sum_{i=1}^{m} \left(\widehat{b}_i \cos(2\pi f_i t) + \widehat{c}_i \sin(2\pi f_i t)\right).$$

Enfin, le choix optimal du nombre de fréquences est défini par

$$\widehat{m} := \underset{m\in\{0,\dots,m_{\max}\}}{\arg\min}\ \mathrm{BIC}(m). \tag{2.18}$$

3) Sélection du sous-ensemble de variables exogènes liées aux jours fériés. On considère ici un ensemble $\mathbb{H}$ composé de six variables liées aux jours fériés (voir le Tableau 2.1). Pour chaque sous-ensemble $\mathcal{H}$ de $\mathbb{H}$, y compris l'ensemble vide $\mathcal{H} = \emptyset$, on estime les paramètres $\mathbf{a}_{(\widehat{K})} := (a_0, \dots, a_{\widehat{K}})$, $\mathbf{b}_{(\widehat{m})} := (b_1, \dots, b_{\widehat{m}})$, $\mathbf{c}_{(\widehat{m})} := (c_1, \dots, c_{\widehat{m}})$ et $\mathbf{d}_{\mathcal{H}} := (d_h)_{h\in\mathcal{H}}$ par la méthode des MCO, soit

$$(\widehat{\mathbf{a}}_{(\widehat{K})}, \widehat{\mathbf{b}}_{(\widehat{m})}, \widehat{\mathbf{c}}_{(\widehat{m})}, \widehat{\mathbf{d}}_{\mathcal{H}}) := \underset{\mathbf{a}_{(\widehat{K})}\in\mathbb{R}^{\widehat{K}+1}, (\mathbf{b}_{(\widehat{m})}, \mathbf{c}_{(\widehat{m})})\in\mathbb{R}^{2\widehat{m}}, \mathbf{d}_{\mathcal{H}}\in\mathbb{R}^{n_{\mathcal{H}}}}{\arg\min} \sum_{t=1}^{T} \left(x_t - T_t^{(\widehat{K})}(\mathbf{a}_{\widehat{K}}) - S_t(F_{(\widehat{m})}, \mathbf{b}_{\widehat{m}}, \mathbf{c}_{\widehat{m}}) - H_t(\mathcal{H}, \mathbf{d}_{\mathcal{H}}) \right)^2,$$

où $H_t(\mathcal{H}, \mathbf{d}_{\mathcal{H}})$ est donné par (2.11). Le sous-ensemble optimal de variables de jours fériés est défini par

$$\widehat{\mathcal{H}} := \underset{\mathcal{H}\subset\mathbb{H}}{\arg\min}\, \mathrm{BIC}(\mathcal{H}),$$

où $\mathrm{BIC}(\mathcal{H})$ est donné par (2.12) avec

$$\widehat{\epsilon}_t = x_t - \sum_{k=0}^{\widehat{K}} \widehat{a}_k P_k(t) - S_t(F_{(\widehat{m})}, \widehat{\mathbf{b}}_{(\widehat{m})}, \widehat{\mathbf{c}}_{(\widehat{m})}) - H_t(\mathcal{H}, \widehat{\mathbf{d}}_{\mathcal{H}}), \quad t = 1, \dots, T.$$

4) Sélection du modèle ARMA pour le terme résiduel. On définit

$$z_t := x_t - \left(T_t^{(\widehat{K})}(\widehat{\mathbf{a}}_{\widehat{K}}) + S_t(F_{(\widehat{m})}, \widehat{\mathbf{b}}_{\widehat{m}}, \widehat{\mathbf{c}}_{\widehat{m}}) + H_t(\widehat{\mathcal{H}}, \widehat{\mathbf{d}}_{\widehat{H}}) \right), \quad t = 1, \dots, T.$$

On suppose que la suite des résidus $z_{1:T}$ est une réalisation d'un processus ARMA(p, q), noté $R_t(\phi, \theta, p, q)$, associé à un bruit blanc gaussien centré de variance finie σ^2, avec $p \in \{0, \dots, p_{\max}\}$ et $q \in \{0, \dots, q_{\max}\}$. Le critère BIC du modèle ARMA(p, q) est donné par

$$\mathrm{BIC}(p,q) = (T-p-q)\log\left(\frac{T\,\widehat{\sigma^2}}{T-p-q}\right) + T(1+\log(\sqrt{2\pi})) + (p+q)\log\left(\frac{(\sum_{t=1}^{T} z_t^2) - T\,\widehat{\sigma^2}}{p+q}\right),$$

où $\widehat{\sigma^2}$ est l'estimateur du maximum de vraisemblance de σ^2 associé à la série $z_{1:T}$. Le choix optimal des hyperparamètres p and q est défini par :

$$(\widehat{p}, \widehat{q}) := \underset{p\in\{0,\dots,p_{\max}\}, q\in\{0,\dots,q_{\max}\}}{\arg\min}\ \mathrm{BIC}(p,q).$$

On obtient alors le modèle TSHR suivant :

$$X_t = T_t^{(\widehat{K})}(\widehat{\mathbf{a}}_{\widehat{K}}) + S_t(F_{(\widehat{m})}, \widehat{\mathbf{b}}_{\widehat{K}}, \widehat{\mathbf{c}}_{\widehat{K}}) + H_t(\widehat{\mathcal{H}}, \widehat{\mathbf{d}}_{\widehat{H}}) + R_t(\widehat{\phi}, \widehat{\theta}, \widehat{p}, \widehat{q}), \quad t = 1, \dots, T, \tag{2.19}$$

où $\widehat{\phi} := (\widehat{\phi}_1, \ldots, \widehat{\phi}_{\widehat{p}})$ et $\widehat{\theta} := (\widehat{\theta}_1, \ldots, \widehat{\theta}_{\widehat{q}})$ sont les estimations du maximum de vraisemblance des paramètres $(\phi_1, \ldots, \phi_{\widehat{p}})$ et $(\theta_1, \ldots, \theta_{\widehat{q}})$ du modèle optimal ARMA($\widehat{p}$, $\widehat{q}$), obtenus à partir de la suite $z_{1:T}$.

2.4 Comparaison des performances de prévision des modèles

L'objectif de cette partie est de comparer les quatre modèles SNAIVE, SARIMA, PROPHET et TSHR, en terme de qualité de prévision, à différents horizon : $H = 7$, 14, 21 et 28, dans notre application. Nous rappelons d'abord, dans la sous-section 2.4.1 ci-après, les mesures de qualité de prévision les plus connues de la littérature, en distinguant bien les deux versions, stochastique et théorique, de ces mesures. Nous proposons dans la sous-section 2.4.2, pour l'estimation des versions théoriques de ces mesures, une nouvelle procédure ad hoc inspirée du paradigme de la validation croisée, adaptée aux séries temporelles. Les modèles seront alors comparés en se basant sur les mesures de performances estimées via la méthode proposée.

2.4.1 Les mesures de performance de prévision d'un modèle

Pour évaluer la performance de prévision d'un modèle, à un certain horizon H, cinq mesures sont couramment utilisées : Erreur Absolue Moyenne (Mean Absolute Error, MAE), Erreur Absolue Moyenne en Pourcentage (Mean Absolute Percentage Error, MAPE), Erreur Absolue Moyenne Pondérée en Pourcentage (Weighted Mean Absolute Percentage Error, WMAPE), Erreur quadratique moyenne (Mean Squared Error, MSE) et Racine de l'erreur quadratique moyenne (Root Mean Squared Error, RMSE), voir par exemple [Gul and Celik, 2020]. Nous rappelons ci-après leurs expressions stochastiques comme leurs versions théoriques formelles ; la section 2.4.2 qui suit détaille une procédure d'estimation de la version théorique de la mesure de performance particulière MAE (erreur absolue moyenne).

Soit $X_{1:T}$ la suite des T premières coordonnées d'un processus stochastique (d'apprentissage), $T > 0$, à partir desquelles les valeurs prédites $\widehat{X}_t(T)$, $t = T+1, \ldots, T+H$, de $X_{T+1}, \ldots, X_{T+H}$, sont construites à l'aide d'un modèle prédictif. Les versions stochastiques des cinq mesures de performance de prévision sont définies dans le tableau 2.4. En particulier, la MAE (stochas-

tique) est définie par

$$\mathrm{MAE}(T,H) = \frac{1}{H}\sum_{h=1}^{H} |X_{T+h} - \widehat{X}_{T+h}(T)|. \tag{2.20}$$

Si le processus "résiduel" $(X_{T+h} - \widehat{X}_{T+h}(T))_{T \geq T_{train}}$ est stationnaire, T_{train} étant la taille de la série temporelle d'apprentissage, l'espérance mathématique de cette mesure, appelée version attendue (ou théorique) dans le tableau 2.4, dépend de h et ne dépend plus de T et, évidemment, n'est plus stochastique, ce qui en fait une mesure adéquate de performance de prévision. La MAE théorique est explicitement donnée par

$$\mathrm{MAE}(H) := \mathbb{E}\left\{\mathrm{MAE}(T,H)\right\} = \frac{1}{H}\sum_{h=1}^{H} \mathbb{E}\left\{|X_{T+h} - \widehat{X}_{T+h}(T)|\right\}, \tag{2.21}$$

où $\mathbb{E}\{\cdot\}$ représente l'espérance mathématique.

Mesure	Version stochastique	Version théorique
MAE	$\frac{1}{H}\sum_{h=1}^{H} \lvert X_{T+h} - \widehat{X}_{T+h}(T)\rvert$	$\frac{1}{H}\sum_{h=1}^{H} \mathbb{E}\left\{\lvert X_{T+h} - \widehat{X}_{T+h}(T)\rvert\right\}$
MAPE	$\frac{1}{H}\sum_{h=1}^{H} \frac{\lvert X_{T+h} - \widehat{X}_{T+h}(T)\rvert}{\lvert X_{T+h}\rvert}$	$\frac{1}{H}\sum_{h=1}^{H} \mathbb{E}\left\{\frac{\lvert X_{T+h} - \widehat{X}_{T+h}(T)\rvert}{\lvert X_{T+h}\rvert}\right\}$
WMAPE	$\frac{\sum_{h=1}^{H} \lvert X_{T+h} - \widehat{X}_{T+h}(T)\rvert}{\sum_{h=1}^{H} \lvert X_{T+h}\rvert}$	$\mathbb{E}\left\{\frac{\sum_{h=1}^{H} \lvert X_{T+h} - \widehat{X}_{T+h}(T)\rvert}{\sum_{h=1}^{H} \lvert X_{T+h}\rvert}\right\}$
MSE	$\frac{1}{H}\sum_{h=1}^{H} (X_{T+h} - \widehat{X}_{T+h}(T))^2$	$\frac{1}{H}\sum_{h=1}^{H} \mathbb{E}\left\{\left(X_{T+h} - \widehat{X}_{T+h}(T)\right)\right\}$
RMSE	$\sqrt{\frac{1}{H}\sum_{h=1}^{H} (X_{T+h} - \widehat{X}_{T+h}(T))^2}$	$\mathbb{E}\left\{\sqrt{\frac{1}{H}\sum_{h=1}^{H} (X_{T+h} - \widehat{X}_{T+h}(T))^2}\right\}$

TABLE 2.4 : Mesures de performance de prévision.

Classiquement, ces mesures sont estimées sur la base d'une trajectoire unique $x_{1:(T+H)}$ de $X_{1:(T+H)}$, ce qui donne des estimations médiocres des mesures théoriques de performance de prévision. Dans la sous-section 2.4.2 ci-après, nous présentons une procédure d'estimation de la mesure théorique de performance de prévision MAE basée sur une procédure de validation croisée adaptée à l'aspect séquentiel et la dépendance des séries temporelles.

2.4.2 Estimation de l'erreur de prévision d'un modèle par validation croisée par fractionnement séquentiel

Afin d'estimer l'erreur théorique de prévision d'un modèle (à un horizon donné $D + H$), voir (2.21), nous proposons une procédure de validation croisée adaptée à la nature séquentielle des données. La structure de dépendance temporelle des données rend la procédure classique de validation croisée K-fold inadéquate. Nous introduisons ici une procédure de validation croisée appelée « validation croisée par fractionnement séquentiel/sequential splitting cross-validation » (SSCV). Contrairement aux méthodes de validation croisée classiques pour les

données i.i.d., la présente méthode tient compte de la dépendance temporelle et de l'aspect séquentiel des données. Nous procédons de la manière suivante. La série temporelle $x_{1:T_{\max}}$, $T_{\max} = 1096$ dans notre application, est divisée plusieurs fois pour générer n_c couples

$$(\mathbf{z}_j^{\text{train}}, \mathbf{z}_j^{\text{test}}), \quad j = 1, \ldots, n_c,$$

de sous-suites d'entraînement et de test. Précisément, nous fixons $T = 1008$, pour la première, et extrayons séquentiellement des sous-suites de longueur croissante à partir de $x_{1:T_{\max}}$, en avançant d'un jour, pour définir

$$\mathbf{z}_j^{\text{train}} = x_{1:T+j-1}, \quad \mathbf{z}_j^{\text{test}} = x_{T+j:T+j+H-1}, \qquad j = 1, \ldots, n_c = T_{\max} - H - T + 1.$$

Dans notre application, nous avons choisi $T = 1008$, ce qui permet de construire un nombre suffisant n_c (par exemple, pour $H = 7$, nous avons $n_c = T_{\max} - H - T + 1 = 82$) de couples de suites d'entraînement et de test pour évaluer les mesures théoriques de performance des modèles en terme de prévision à $D+H$. Nous proposons alors l'estimateur suivant de la mesure théorique $\text{MAE}(H)$ de prévision (2.21) :

$$\widehat{\text{MAE}}(H) := \frac{1}{n_c} \sum_{j=1}^{n_c} \frac{1}{H} \sum_{h=1}^{H} \left| x_{T+h+j-1} - \widehat{x}_{T+h+j-1}(T+j-1) \right|, \tag{2.22}$$

$\widehat{x}_{T+h+j-1}(T+j-1)$ étant la valeur prédite de $x_{T+h+j-1}$, donnée par le modèle en cours d'évaluation, utilisant les données historiques $x_{t:(T+j-1)}$, jusqu'au temps $T+j-1$. La figure 2.1 illustre la procédure d'estimation SSCV de la mesure théorique de prévision $\text{MAE}(H)$, avec $H = 7$, $T = 1008$ et $T_{\max} = 1096$.

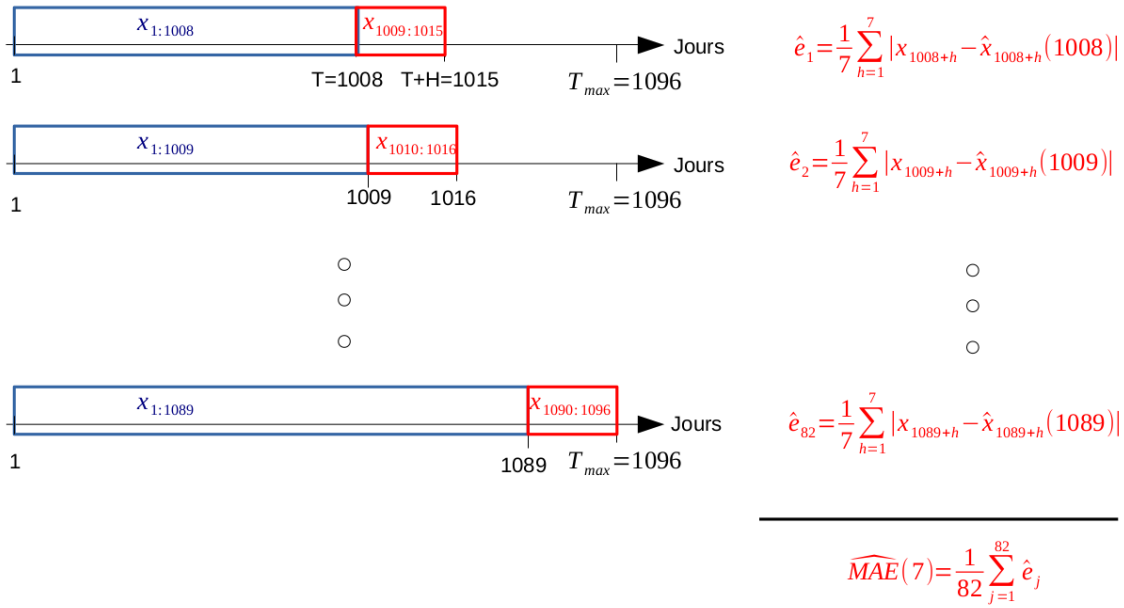

FIGURE 2.1 : Illustration de la méthode SSCV pour l'estimation de l'erreur théorique de prévision MAE(H), pour $H = 7$, $T = 1008$ et $T_{\max} = 1096$.

2.5 Annexe - Présentation des variantes du modèle TSHR et des procédures d'estimation

Dans cette partie, on répertorie des versions alternatives du modèle TSHR ou des procédures d'estimation et de sélection de modèle présentées précédemment. Ces alternatives ont été introduites au fil des essais, constituant des tentatives d'arbitrage entre deux objectifs adverses : 1) la performance de prévision et 2) le temps d'exécution des procédures d'estimation et de sélection de modèle. Le modèle et la procédure présentés dans les parties 2.2.3 et 2.3.2 sont les variantes aboutissant à des performances de prévision parmi les meilleures obtenues, pour une exécution n'excédant pas quelques secondes (par SAU) sur un ordinateur de bureau.

2.5.1 Variantes du modèle TSHR

La première variante de modèle additif généralisé testé a été un modèle de type (TH)SR, c'est-à-dire, une première composante regroupant tendance polynomiale et effets ponctuels, une seconde composante saisonnière, la dernière étant un ARMA. Ce premier modèle présentait un défaut d'"identifiabilité"; les deux premières composantes pouvant chacune couvrir des propriétés saisonnières de la série. Du point de vue de l'ajustement et de la sélection de modèle, cela s'est traduit par une forte instabilité des hyperparamètres associés à ces deux composantes, lorsque ces modèles étaient ajustés pour des extraits de taille variable des historiques de fréquentation d'un SAU.

Le modèle TSHR présenté dans la partie 2.2.3 sépare les composantes **T** et **H** et priorise la composante saisonnière **S** de sorte à ce que la procédure de sélection de modèle par étapes imbriquées favorise la composante **S** avec un grand nombre d'harmoniques et capte ainsi l'essentiel des propriétés saisonnières de la série. La composante **H** « résiduelle » capte alors les variations ponctuelles de fréquentation dues aux événements couverts, indépendamment de l'aspect saisonnier des dates de ces événements.

Cette troisième composante **H** prend la forme d'une courbe en escalier, en raison de l'utilisation d'indicatrices pour représenter ces variables. Cette irrégularité ne permet pas de capter une réalité plus subtile, où l'effet d'un événement (jour férié ou autre) peut avoir un impact, éventuellement plus modéré, un ou plusieurs jours avant ou après la date. Afin de corriger cet aspect, une méthode de lissage a été testée, en utilisant un noyau gaussien centré sur la date de l'événement. Toutefois, cette approche n'a pas été retenue dans le modèle final, en raison de la complexité du paramétrage et du temps de calcul élevé qu'elle impliquait. En effet, chaque événement ponctuel influence le modèle de manière différente, ce qui nécessitait de définir un écart-type spécifique pour chacun d'eux, avec une plage de valeurs possibles très large.

2.5.2 Variantes de la méthode de sélection de modèle

La méthode de sélection des composantes tendancielle et saisonnière présentée dans la partie 2.3.2 se fait selon une progression hiérarchisée : les polynômes de degré K considérés sont des combinaisons linéaires des K premiers polynômes orthogonaux, les polynômes trigonométriques de la partie saisonnière sont ordonnés selon le niveau d'énergie des harmoniques qui les définissent. Une variante testée a consisté en une recherche exhaustive : pour la composante tendancielle, plutôt que de sélectionner le degré optimal de manière séquentielle et croissante (degré 1, degré 2, etc.), une recherche exhaustive a été mise en place, dans laquelle on explore l'ensemble des polynômes selon le nombre de coefficients non nuls : un polynôme de taille 3 peut inclure des degrés non consécutifs tels que 2, 5 et 8. De même pour la composante saisonnière, on cherche la meilleur combinaison de m harmoniques, sans les hiérarchiser par niveau d'énergie. Cette procédure, plus flexible dans le choix du modèle, a été abandonnée en raison de son temps d'exécution dissuasif.

La méthode de sélection de modèle TSHR présentée dans la partie 2.3.2 repose sur l'utilisation du critère BIC pour chaque composante du modèle. Nous avons également testé une alternative, reposant toujours sur la sélection séquentielle des hyperparamètres des composantes : au lieu de choisir les hyperparamètres minimisant le critère BIC, nous choisissons les hyperpa-

ramètres minimisant l'erreur de prévision estimée grâce à la procédure de validation croisée par fractionnement séquentiel présentée dans la partie 2.4.2 (ou la variante présentée dans la partie ci-dessous).

Cette procédure de sélection de modèle s'accompagne d'une difficulté qui nous a conduit à privilégier la sélection de modèle via le critère BIC : les temps d'exécution deviennent prohibitifs, notamment lorsque l'on procède à une recherche exhaustive des hyperparamètres.

2.5.3 Variantes de la méthode de validation croisée

La première variante de la procédure de validation croisée par fractionnement séquentiel que l'on a testée consistait à prendre pour « folds » des fenêtres temporelles de taille constante de 840 jours puis à faire glisser cette fenêtre de sept jours. La taille de 840 jours a été choisie pour deux raisons :

- elle permet, pour cette première variante de la procédure de SSCV, de disposer d'un nombre raisonnable de « folds » ; explicitement, $(1008 - 840)/7 = 24$ folds ;
- dans les premiers temps, l'estimation de l'ensemble $\mathcal{F}_{m_{\max}}$ des harmoniques utilisées dans la partie saisonnière du modèle était faite en classant les harmoniques de plus grande énergie moyenne sur les folds de la procédure de SSCV. Dans cette optique, le choix d'une taille de 840 jours présente deux intérêts : 1) 840 est un multiple de 7, ce qui garantit que l'harmonique correspondant à la saisonnalité hebdomadaire est bien présente et 2) 840 est multiple d'une puissance de 2 (2^3), ce qui constitue un atout pour l'exécution de l'algorithme de transformée de Fourier rapide.

Nous avons finalement privilégié une fenêtre glissante de taille croissante, avançant d'un jour à chaque fold. Ces choix permettent d'améliorer la précision de l'estimation des erreurs de prévision espérées et ainsi d'améliorer la fiabilité de la comparaison des modèles.

Chapitre 3

Ajustement de modèles aux historiques de fréquentation des SAUs et comparaison de leurs performances de prévision

Dans ce chapitre, les quatre familles de modèles introduits au chapitre précédent sont ajustés à l'historique de fréquentation (restreint à la période d'entraînement) de chaque SAU. La procédure d'ajustement et de sélection de modèles TSHR est illustrée et détaillée pour l'un des SAUs dans la partie 3.1. En particulier, y sont identifiés les contributions des quatre composantes de ce modèle additif ; les résidus de chacune des quatre étapes successives de l'ajustement sont présentés et brièvement analysés. Une analyse rigoureuse des résidus à l'issue de la quatrième étape est présentée pour l'un des SAUs dans la partie 3.2. Les performances de prévisions des quatre familles de modèles pour les 51 SAUs sont présentées et comparées dans la partie 3.3. Enfin, une analyse descriptive a posteriori des hyper-paramètres retenus pour les 51 SAUs est présentée dans la partie 3.4.

3.1 Déroulement de la procédure d'ajustement du modèle TSHR ; exemple avec le SAU du « Pôle Public St Louis »

À titre d'illustration et d'exemple, nous présentons certaines caractéristiques de la procédure de sélection de modèles décrite dans la section 2.3.2 du chapitre 2, lorsqu'elle est appliquée à la série temporelle des passages quotidiens aux urgences du « Pôle Public St Louis ». Plus précisément, l'évolution, en fonction du degré K de la tendance polynomiale $\mathbf{T}$, de l'erreur quadratique moyenne d'ajustement et des valeurs du critère BIC, correspondant à la première étape de la procédure, est présentée dans la partie supérieure de la figure 3.1. L'évolution, en fonction du nombre d'harmoniques de la composante saisonnière $\mathbf{S}$, de l'erreur quadratique moyenne d'ajustement et des valeurs du critère BIC, correspondant à la deuxième étape, est présentée dans la partie inférieure de la figure 3.1.

L'évolution de l'erreur quadratique moyenne d'ajustement (MSE) (à gauche) et des valeurs du BIC (à droite) pour la tendance polynomiale (en haut) et la composante saisonnière harmonique (en bas), en fonction respectivement de K et m, suit une tendance qualitativement classique : le MSE d'ajustement diminue naturellement avec la complexité du modèle, tandis que les valeurs du BIC suivent globalement un schéma de type « décroissance puis croissance », les valeurs minimales correspondant à un compromis entre biais et variance du modèle.

La figure 3.2 présente le périodogramme (à gauche) de la séquence des résidus (2.13) après la première étape. La ligne horizontale en pointillés bleus correspond au niveau d'énergie des $\widehat{m}$ harmoniques les plus élevées de cette séquence résiduelle, $\widehat{m}$ étant le nombre optimal d'harmoniques défini par (2.18).

L'ensemble $F_{(\widehat{m})}$ des fréquences utilisées dans la composante saisonnière du modèle TSHR est représenté par les lignes verticales rouges en pointillés. Il convient de noter que ces fréquences comprennent $1/7$ (correspondant à la saisonnalité hebdomadaire) et certains de ses multiples ou diviseurs. On les retrouves pour tous les SAUs, et pas seulement pour le SAU du Pôle Public St Louis.

La figure 3.3 (partie gauche) présente les composantes imbriquées ajustées (de haut en bas) T, TS, TSH et TSHR du modèle TSHR sur la série d'apprentissage $x_{1:T}$, avec $T = 1008$, pour le service accueil d'urgences du « Pole Public St Louis » . Les fonctions d'auto-corrélation (ACF) des résidus associés aux composantes T, TS, TSH et TSHR sont présentées dans la partie de droite.

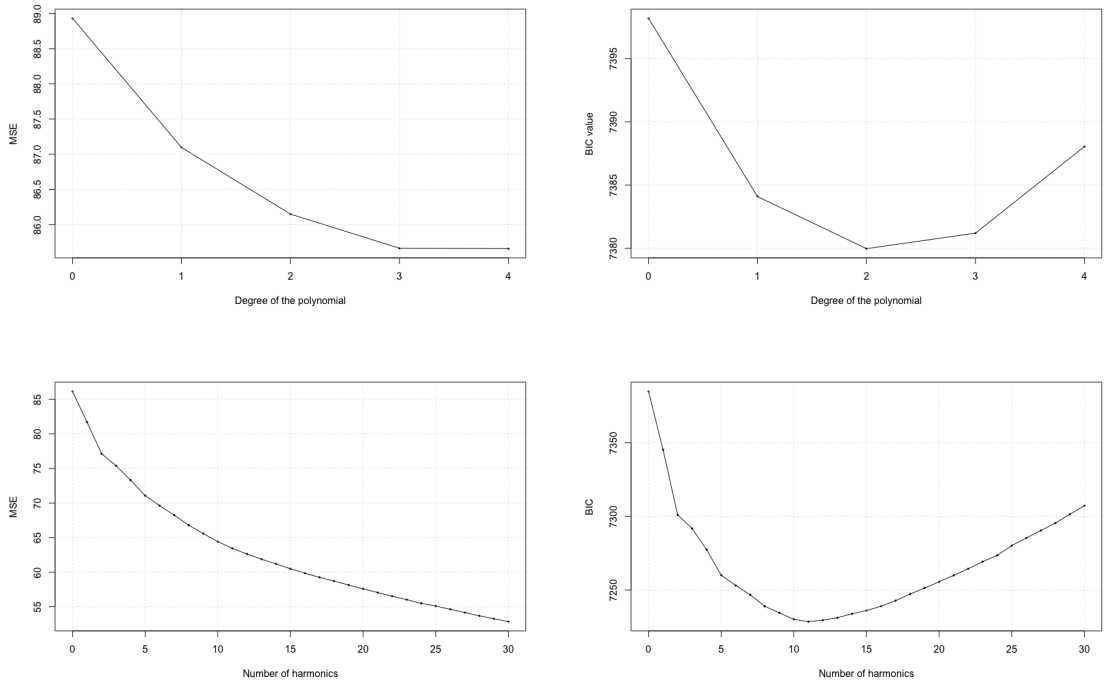

FIGURE 3.1 : Évolution de l'EQM d'ajustement (à gauche) et de la valeur du BIC (à droite) en fonction du degré de la tendance polynomiale au cours de la première étape de la procédure de sélection du modèle (en haut), et du nombre d'harmoniques dans la composante saisonnière au cours de la deuxième étape (en bas).

Comme prévu, la tendance polynomiale, de degré 2 ici, correspond à la tendance générale mais ne parvient pas à ajuster les phénomènes saisonniers et locaux. L'ACF de ses résidus montre clairement que de nombreuses auto-corrélations significatives n'ont pas été prises en compte par la composante tendancielle. L'ajustement par la composante TS est bien meilleur, avec un petit nombre d'auto-corrélations à court terme restant significatives. L'intégration de la composante exogène des vacances ne modifie pas significativement l'ajustement dans ce cas, car la partie saisonnière a déjà pris en compte les effets des vacances, en particulier des vacances d'été. Enfin, l'ajustement des résidus de la composante TSH par un processus ARMA permet d'obtenir une composante TSHR globale sans auto-corrélations significatives.

D'après le graphique ACF (figure 3.2, en bas à droite), aucun effet périodique significatif ne subsiste dans les résidus résultant de l'ajustement par le modèle TSHR, ce qui souligne la capacité du modèle TSHR à capturer correctement les périodicités du flux de patients dans les SAUs. En outre, les quatre étapes sont nécessaires pour décorréler correctement les résidus, comme le montrent explicitement les tracés ACF sur la droite de la figure 3.2.

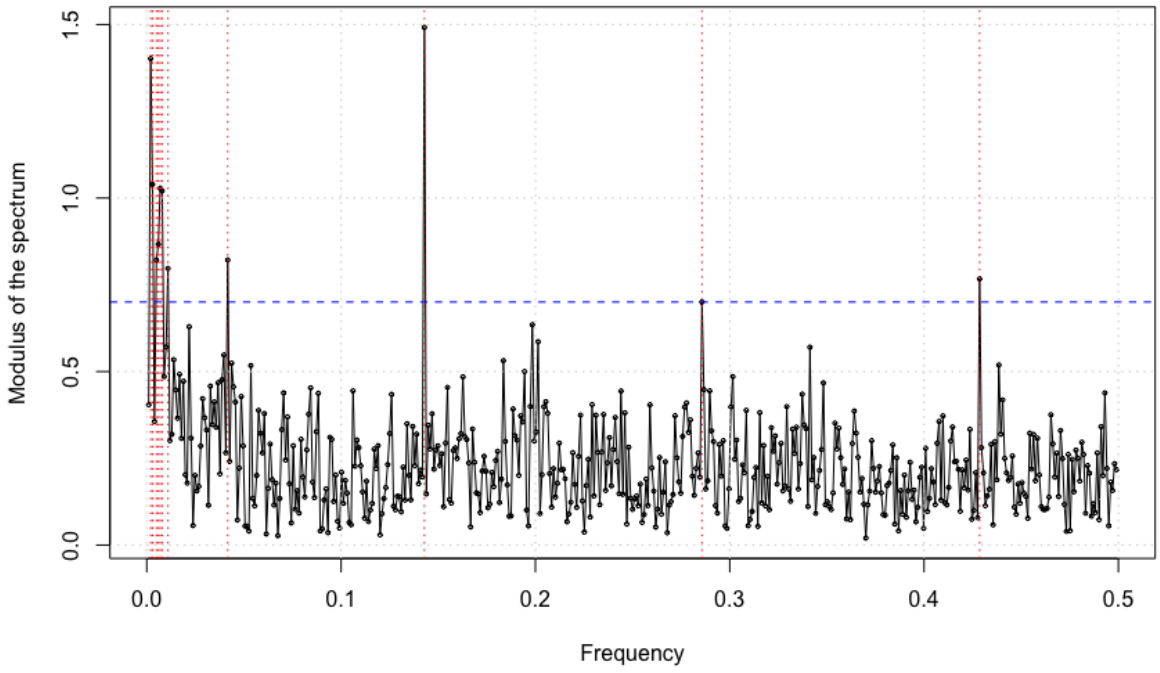

Figure 3.2 : Périodogramme des résidus (2.13) après suppression de la tendance polynomiale, pour le SAU Pôle Public St Louis. La ligne horizontale en pointillés bleus correspond au seuil déterminant les fréquences les plus significatives.

3.2 Analyse des résidus : exemple avec le SAU du « CHU de Metz »

La série résiduelle après l'ajustement par l'un des modèles PROPHET, SARIMA et TSHR est supposée être dépourvue de tendance et de saisonnalité. L'évaluation de la qualité de la modélisation, en plus de sa performance prédictive, passe par l'analyse des résidus, qui doivent idéalement se comporter comme un bruit blanc gaussien, c'est-à-dire, être non corrélés, homoscédastiques et distribués selon une loi normale. Ces trois conditions vont être vérifiées pour les trois modèles TSHR, SARIMA et PROPHET, à l'aide de graphiques, et de statistiques de tests appropriées.

La figure 3.4 présente les tracés des fonctions d'autocorrélation des suites résiduelles associées aux trois modèles. On voit que l'ACF de la série résiduelle du modèle PROPHET, contrairement aux deux autres modèles, TSHR et SARIMA, montre des autocorrélations à court terme significatives.

Le test de Durbin-Watson [Durbin and Watson, 1971] a ensuite été appliqué, à l'aide de la

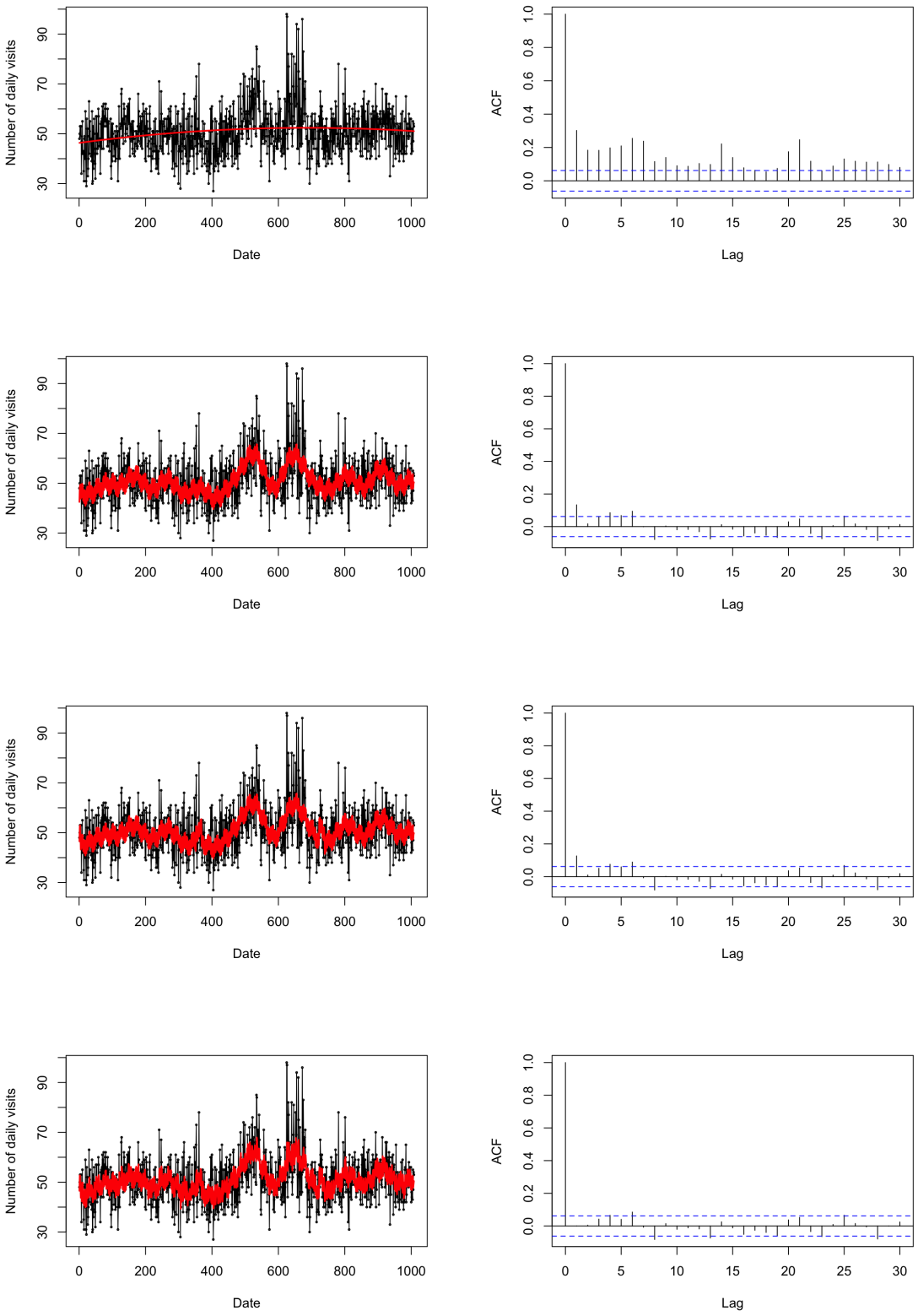

FIGURE 3.3 : Ajustement de l'historique de fréquentation du SAU Pôle Public St Louis par les composantes T, TS, TSH et TSHR (à gauche, de haut en bas). La fonction d'autocorrélation des résidus correspondants pour chaque modèle est présentée à droite.

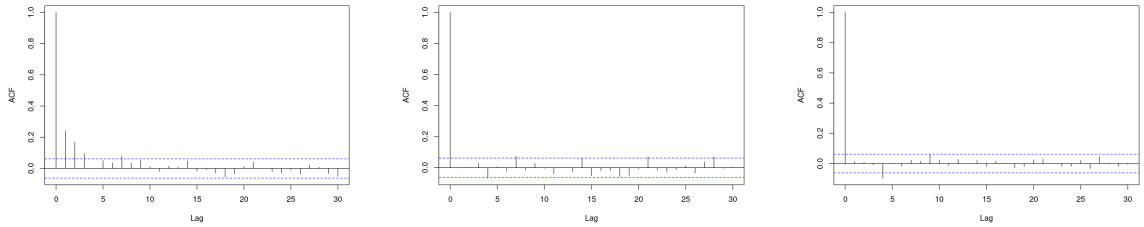

Figure 3.4 : Fonctions d'autocorrélation (ACF) des résidus des modèles, de gauche à droite, PROPHET, TSHR et SARIMA.

fonction `dwtest()` du package `{lmtest}` de R, pour tester l'hypothèse nulle selon laquelle les résidus sont non corrélés. Nous avons obtenu les valeurs suivantes de P-value : $0,994$ pour le modèle TSHR, $0,321$ pour le modèle SARIMA et $2,919 \times 10^{-14}$ pour le modèle PROPHET. L'hypothèse nulle de non corrélation des résidus est donc rejetée pour le modèle PROPHET ; pas d'autocorrélation significative pour le modèle TSHR ni, dans une moindre mesure, pour le modèle SARIMA. Ces résultats sont cohérentes avec les graphiques des fonctions d'autocorrélation des trois modèles ; voir 3.4.

Ces conclusions ont été consolidées également par l'application du test de type "portmanteau" de Ljung-Box [Ljung and Box, 1978] ; pour laquelle on a utilisé la fonction `box.test()` du package `{stats}` de R. Ce test a été appliqué aux résidus de chaque modèle avec un retard de 1. Il indique une p-value de 3×10^{-14} pour le modèle PROPHET, traduisant une autocorrélation résiduelle d'ordre 1 très significative. À l'inverse, les modèles TSHR (p-value = 0,95) et SARIMA (p-value = 0,67) présentent des résidus non autocorrélés (à l'ordre 1). L'application du test aux ordres supérieurs, avec un retard de 10, donne une p-value de 2.2×10^{-16} pour le modèle PROPHET, 0.23 pour le modèle TSHR et 0.13 pour SARIMA, confirmant la non corrélation des erreurs pour ces deux derniers modèles, aux ordres supérieurs.

La constance de la variance des résidus a été testée à l'aide du test de Breusch-Pagan [Breusch and Pagan, 1979]. Pour le modèle PROPHET, la p-value est inférieur au seuil de 5%, ce qui conduit au rejet de l'hypothèse d'homoscédasticité, traduisant une variance non constante. Concernant les deux autres modèles, les p-values sont très proche du seuil ce qui ne permet pas de conclure.

La normalité des résidus a été vérifiée à l'aide d'outils graphiques et statistiques. Les QQ-plots montrent un alignement globalement satisfaisant des points le long de la droite de référence, ne permettant pas d'identifier des déviations à la normalité des résidus. Les histogrammes présentent également une forme symétrique et centrée autour de zéro, sans asymétrie marquée.

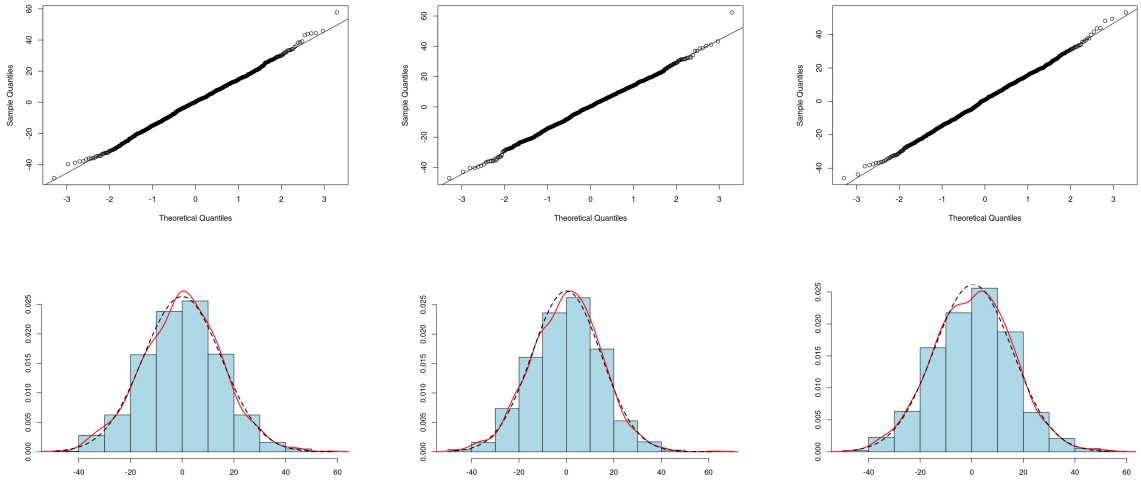

FIGURE 3.5 : Analyse de la normalité des résidus pour les trois modèles, de gauche à droite, PROPHET, TSHR et SARIMA. En haut : QQ-plots des résidus. En bas : histogrammes des résidus avec densités superposées : la courbe continue rouge correspond à l'estimateur à noyau de la densité, la courbe noire en pointillés représente la densité de la loi normale.

Ces graphiques se trouvent dans la figure 3.5. Et enfin, le test de Shapiro-Wilk [Shapiro and Wilk, 1965] a été utilisé pour confirmer ces observations. Dans tous les cas, les p-values étant supérieures à 0,05, l'hypothèse de normalité ne peut être rejetée. Les résidus des trois modèles sont donc considérés comme normalement distribués.

3.3 Comparaison des performances de prévision des modèles SNAIVE, SARIMA, PROPHET et TSHR

Cette section présente des comparaisons de performance en matière de prévision pour les quatre modèles étudiés. Les comparaisons sont effectuées en estimant pour chaque modèle l'erreur absolue moyenne de prévision (MAE) – voir la section 2.4.1 pour une définition précise de cette mesure de performance et d'autres mesures classiques – pour différents horizons temporels ($H = 7, 14, 21, 28$). Ces estimations sont réalisées selon une procédure ad hoc inspirée du paradigme de la validation croisée, appelée validation croisée séquentielle (SSCV), présentée à la section 2.4.2.

La figure 3.6 illustre la cohérence de la méthode SSCV pour l'estimation de la MAE(H) attendue, avec $H = 7$, du modèle TSHR, pour le SAU CHR Metz. En fait, nous pouvons voir sur la

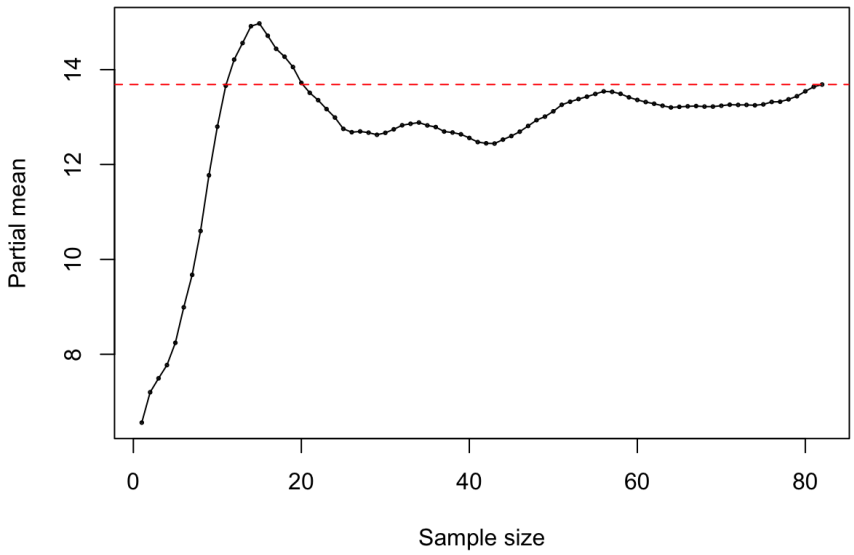

Figure 3.6 : Comportement de la séquence de moyennes partielles

figure que la séquence des moyennes partielles

$$\left(\frac{1}{n}\sum_{j=1}^{n}\frac{1}{H}\sum_{h=1}^{H}|x_{T+h+j-1}-\widehat{x}_{T+h+j-1}(T)|\right)_{n\geq 1}$$

converge vers la moyenne totale (ligne pointillée en rouge) lorsque le nombre de termes n augmente. Un comportement similaire a été observé pour les autres SAUs.

Les modèles SARIMA (avec la période $s = 7$) et TSHR ont été sélectionnés individuellement pour chacun des 51 SAUs selon les méthodologies respectives décrites dans les parties 2.3.1 et 2.3.2, sur la base de la période d'entraînement de $T = 1008$ jours, du 1er janvier 2016 au 4 octobre 2018. Les deux modèles SNAIVE et PROPHET ont été appliqués avec la période $s = 7$. Ensuite, le MAE de chaque modèle (SNAIVE, SARIMA, PROPHET et TSHR), pour quatre horizons de prévision ($H = 7, 14, 21, 28$), a été estimé à l'aide du SSVC ci-dessus sur la base des séries temporelles couvrant la période de test, du 5 octobre 2018 au 31 décembre 2018. Afin d'exprimer l'erreur de prévision en pourcentage, nous calculons également le WMAPE « estimé » :

$$\widehat{\mathrm{WMAPE}}(H) := \frac{1}{n_c}\sum_{j=1}^{n_c}\frac{\sum_{h=1}^{H}|x_{T+h+j-1}-\widehat{x}_{T+h+j-1}(T)|}{\sum_{h=1}^{H}|x_{T+h+j-1}|}. \tag{3.1}$$

Il est évident qu'il n'est pas possible de présenter ici de manière exhaustive les performances de prévision des quatre familles de modèles pour les 51 SAUs et les quatre horizons de prévision. Au lieu de cela, nous présentons d'abord une vue d'ensemble des performances de prévision

des quatre familles de modèles pour quatre horizons de prévision, obtenues en agrégeant les résultats pour les 51 SAUs. Ensuite, nous nous concentrons sur les performances de prévision à l'horizon de 7 jours des quatre modèles ajustés pour les sept SAUs du département de la Marne.

Le tableau 3.1 présente les performances de prévision agrégées (moyenne sur les 51 SAUs) de 7 à 28 jours (MAE et WMAPE estimés) des modèles SNAIVE, SARIMA, PROPHET et TSHR pour les quatre horizons de prévision considérés.

Mesures de performance des prévisions	MAE				WMAPE (%)			
Horizon de prévision H	7	14	21	28	7	14	21	28
SNAIVE	10. 13	10.37	10.39	10.41	15.30	15.74	15.83	15. 86
SARIMA	7.63	7.76	7.84	7.91	11.54	11.76	11.92	12.05
PROPHET	7.55	7.56	7.57	7. 58	11.24	11.29	11.33	11.35
TSHR	7.99	8.03	8.01	7.99	11.95	12.03	12.04	12.04

Table 3.1 : MAE et WMAPE agrégés (moyenne des 51 SAUs) aux horizons de prévision de 7 à 28 jours pour les modèles SNAIVE, SARIMA, PROPHET et TSHR.

Le tableau 3.2 présente les performances de prévision des modèles SNAIVE, SARIMA, PROPHET et TSHR à l'horizon de prévision (H=7), pour les sept SAUs de la Marne.

SAUs	Chalons	Epernay	Vitry	Reims-ad.	Reims-ped.	Courlancy	Reims-B.
SNAIVE	12.18	8.47	6.31	13.98	12.20	6.71	7.54
SARIMA	7.65	6.33	5.00	9.91	9.19	5.14	5.80
PROPHET	7.63	6.15	4.81	9.85	8.39	4.85	5.62
TSHR	7.94	6.83	5.09	12.58	10.52	4.56	5.99

Table 3.2 : MAE de prévision à 7 jours des quatre modèles pour les sept SAUs de la Marne.

Les tableaux 3.1 et 3.2 montrent que le modèle SNAIVE est nettement moins efficace en termes de prévision que les modèles SARIMA, PROPHET et TSHR. Ces trois derniers modèles présentent des performances de prévision équivalentes. Il est à noter que le modèle TSHR, en plus de sa simplicité, nécessite un temps de calcul beaucoup plus faible que les deux modèles SARIMA et PROPHET.

3.4 Analyse a posteriori des modèles

Cette partie consiste en une étude a posteriori des modèles optimaux TSHR et SARIMA. Plus précisément, elle vise à examiner les hyperparamètres optimaux de ces modèles, et les croiser

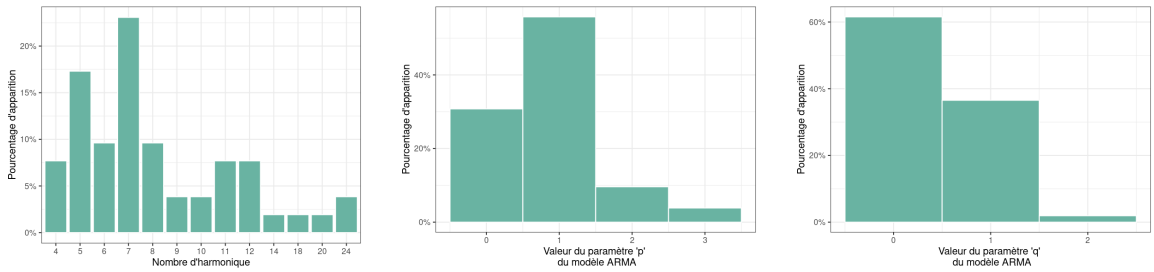

Figure 3.7 : Distributions pour l'ensemble des SAUs des valeurs retenues pour le nombre d'harmoniques, le paramètre p et le paramètre q du modèle TSHR

avec les informations dont nous disposons sur les SAUs.

Tout d'abord concernant le modèle SARIMA, le modèle d'hyperparamètres $(1,0,1)(0,\ 1,\ 1)_7$ revient dans 96% des cas.

Le tableau 3.3 présente les hyperparamètres sélectionnés pour les modèles TSHR des sept SAUs de la Marne (la présentation des résultats a été limitée à ces derniers pour des raisons de concision).

SAU	Chalons	Epernay	Vitry	Reims-ad.	Reims-p.	Courlancy	Reims-B.
$\widehat{K}$	1	0	1	0	0	1	3
$\widehat{m}$	6	10	6	8	20	12	7
$(\widehat{p},\widehat{q})$	(2,0)	(3,0)	(1,1)	(0,0)	(3,0)	(1,0)	(1,1)

Table 3.3 : Les hyperparamètres retenus lors de la procédure de sélection du modèle TSHR par validation croisée séquentielle pour les sept SAUs de la Marne ($\widehat{K}$: degré du polynôme modélisant la tendance, $\widehat{m}$: nombre d'harmoniques de la composantes saisonnière, $(\widehat{p},\widehat{q})$: les hyperparamètres de l'ARMA modélisant le terme résiduel).

Un coup d'œil rapide à ce tableau, et plus généralement à l'ensemble des SAUs disponibles, montre que la tendance est modélisée par un polynôme de degré 0 ou 1 dans environ 90% des SAUs.

En ce qui concerne les harmoniques sélectionnées, il y a des similitudes entre les SAUs. La saisonnalité sept et ses multiples sont systématiquement présentes. On retrouve souvent les périodes $7 \times \frac{1}{2}$, $7 \times \frac{1}{3}$, 7×1, 7×30, etc. Par ailleurs, les graphiques (Figure 3.7) montrent que, dans environ la moitié des SAUs étudiés, sept harmoniques sont suffisantes pour représenter adéquatement la composante saisonnière des séries.

Enfin, les valeurs des hyperparamètres du modèle ARMA sont comprises entre 0 et 3. L'étude

montre que, pour environ 80% des séries considérées, les valeurs de p et q sélectionnées ne dépassent pas 1. Cela suggère qu'une modélisation reposant sur des valeurs faibles de p et q est, dans la majorité des cas, suffisante pour caractériser le processus résiduel R_t. De plus, dans certains cas, ce dernier est assimilé à du bruit blanc, c'est le cas par exemple, pour le SAU Reims adulte).

Une fois les hyperparamètres des modèles TSHR sélectionnés pour l'ensemble des SAUs, il est possible de les interpréter à l'aide des informations supplémentaires disponibles pour chaque service, permettant ainsi de mieux les caractériser. Les informations supplémentaires dont nous disposons sur les SAUs sont présentées à la partie 1.2.

On met ainsi en évidence certaines régularités dans le choix des hyperparamètres des modèles, en fonction des caractéristiques propres aux SAUs.

Tout d'abord, les modèles retenus pour les SAUs pédiatriques présentent de fortes similarités. Plus précisément, les SAUs pédiatriques se distinguent par des degrés de polynômes qui ne sont pas parmi les plus élevés, mais présentent en revanche un nombre d'harmoniques relativement important, voir la figure 3.8. Les polynômes modérés observés dans les services pédiatriques indiquent que la tendance globale est moins marquée. Comme les polynômes n'apportent qu'une contribution limitée, le nombre élevé d'harmoniques reflète l'essentiel de la variabilité quotidienne, hebdomadaire, ou saisonnière, liée notamment aux épidémies infantiles, et c'est donc principalement elles qui expliquent les fluctuations des flux.

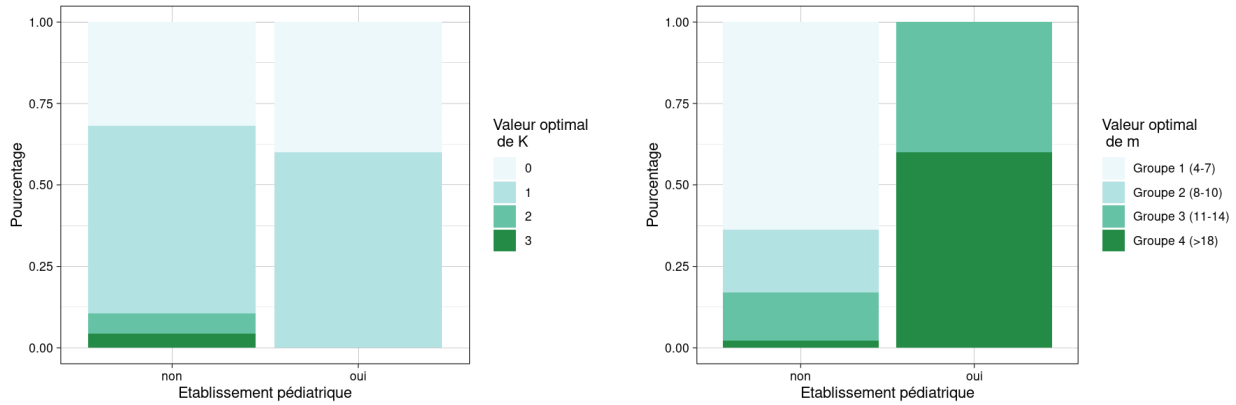

FIGURE 3.8 : Diagrammes en barres, représentant le degré optimal du polynôme noté K (à gauche) et le nombre d'harmoniques optimal noté m (à droite) ; identifiés pour chaque SAU, en distinguant les SAUs pédiatriques des non pédiatriques. Les nombres d'harmoniques ont été regroupés en quatre classes pour améliorer la lisibilité.

À l'inverse, ce sont uniquement les SAUs publics qui apparaissent parmi ceux ayant un très grand nombre d'harmoniques. Le faible nombre d'harmoniques des SAUs privés peut suggérer

des flux plus réguliers, moins soumis à des variations saisonnières ou journalières fortes. Cela pourrait aussi refléter une patientèle plus stable ou des services avec moins de pics ponctuels ; voir la figure 3.9.

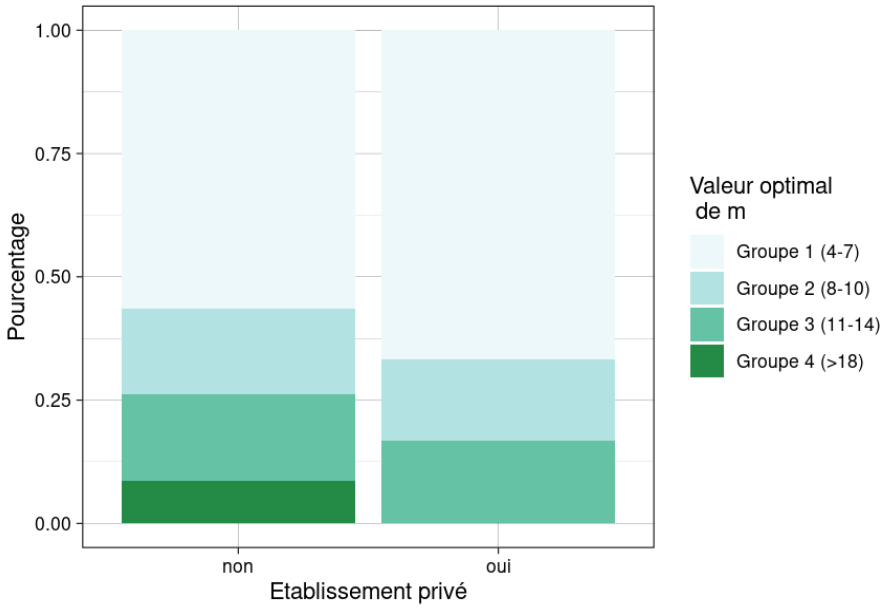

Figure 3.9 : Diagramme en barres représentant le nombre d'harmoniques optimal (noté m) pour chaque SAU, distinguant les établissements privés et les établissements publics. Les nombres d'harmoniques ont été regroupés en quatre classes pour améliorer la lisibilité.

Les SAUs sans SMUR tendent à avoir davantage d'harmoniques. Ils accueillent tous les patients sans tri préalable par un service d'urgence mobile, ce qui peut entraîner des flux plus irréguliers et moins prévisibles, avec des pics ponctuels plus fréquents, voir la figure 3.10.

Concernant le grade traumatologique, les SAUs recevant les cas les moins graves présentent généralement des degrés polynomiaux plus faibles, tandis que ceux prenant en charge des cas plus sévères tendent à être associés à un nombre d'harmoniques plus limité, comme le montre la figure 3.11.

Par ailleurs, certaines tendances se dégagent au niveau des GHT : les GHT n°2, 3, 8 et 12 affichent des degrés polynomiaux plus importants que les autres, tandis que les GHT n°4, 5, 8 et 11 sont associés à un nombre d'harmoniques plus limité. L'examen de la carte des SAUs du Grand Est – voir les figures 1.3 et 3.13 – permet de mettre en perspective les résultats obtenus. Les GHT présentant des degrés polynomiaux plus élevés que les autres sont situés dans des zones urbaines de la région Grand Est, correspondant à des bassins de population denses et concentrés autour de grandes agglomérations comme Reims ou Metz. De plus, les GHT n°2 et

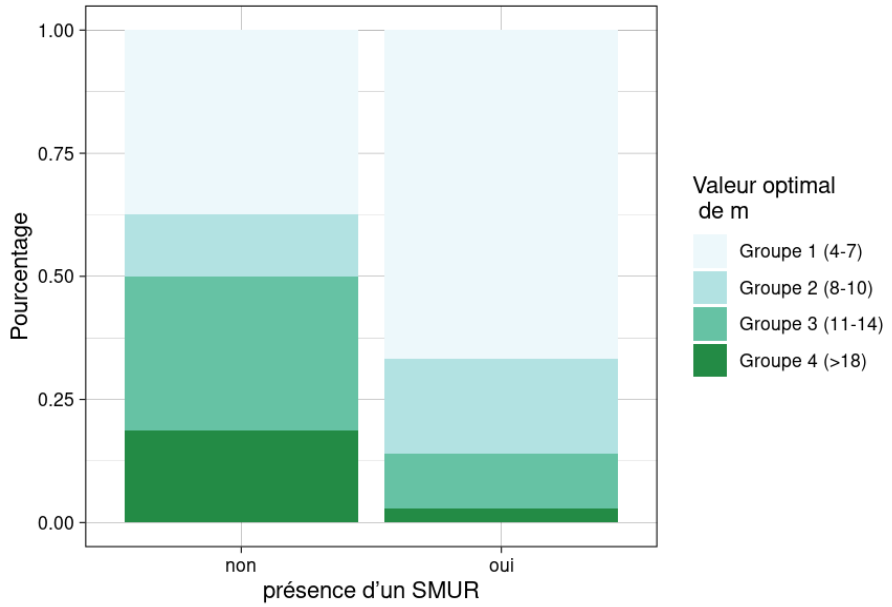

FIGURE 3.10 : Diagramme en barres représentant le nombre d'harmoniques optimal (noté m) chaque SAU, distinguant les établissements avec et sans SMUR. Les nombres d'harmoniques ont été regroupés en quatre classes pour améliorer la lisibilité.

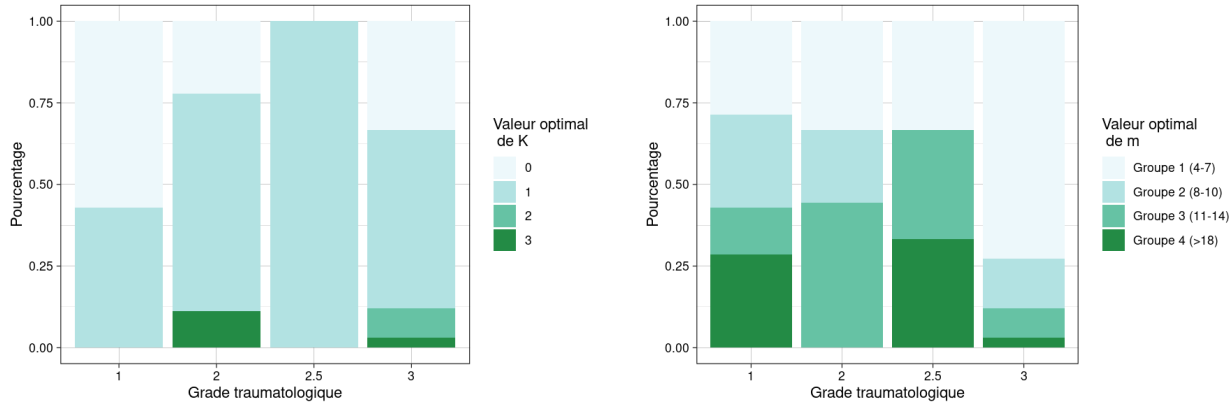

FIGURE 3.11 : Diagrammes en barres représentant, à gauche, le degré polynômial optimal K, et à droite, le nombre d'harmoniques optimal m, pour chaque SAU, en distinguant les SAUs suivant la gravité des cas à traiter. Les nombres d'harmoniques ont été regroupés en quatre classes afin de faciliter la lecture.

n°3 sont situés à l'ouest de la région, tandis que les GHT n°8 et n°12 se trouvent dans la partie sud est du Grand Est. En ce qui concerne les GHT n°4,5,8 et 11, ils sont situés dans le sud de la région et au centre.

Enfin, les GHT n°2, 6, 9 et 11 se démarquent par des valeurs plus élevées du paramètre p de l'ARMA, suggérant une dynamique temporelle différente.

De cette carte, on observe que, globalement, des SAUs géographiquement proches ont des va-

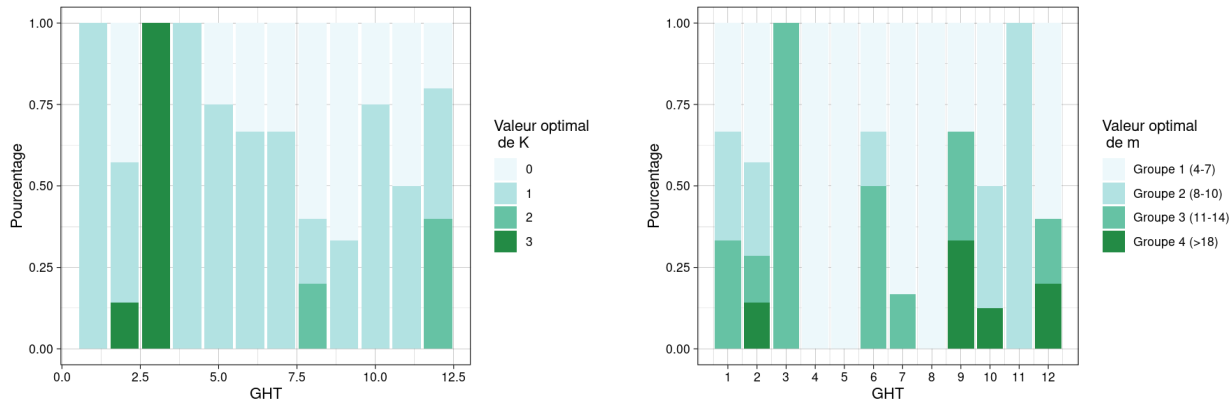

Figure 3.12 : Diagrammes en barres représentant, à gauche, le degré polynômial optimal K, et à droite, le nombre d'harmoniques optimal m, pour chaque SAU, en fonction de leur appartenance au GHT. Les nombres d'harmoniques ont été regroupés en quatre classes afin de faciliter la lecture.

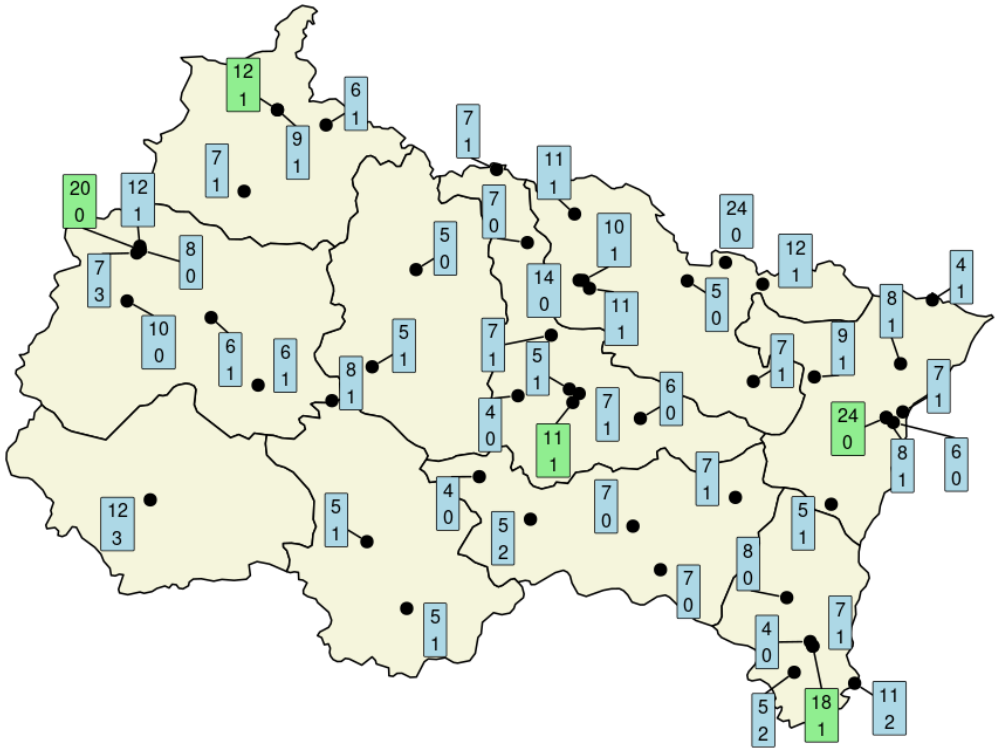

Figure 3.13 : Carte des SAUs de la région Grand Est, indiquant pour chaque service le degré polynômial optimal K et le nombre d'harmoniques optimal m. Les services pédiatriques sont mis en évidence en fond vert.

leurs de (K,m) proches.

Chapitre 4

Modélisation multivariée des flux via clustering et régression PLS

Ce chapitre présente une méthode de prévision des flux de patients dans les 51 SAUs de la région Grand Est basée sur une modélisation multivariée des flux. Afin de réduire la dimension du problème (le nombre de séries à modéliser simultanément), une étape préliminaire de clustering des historiques de flux des SAUs a été réalisée. Elle a permis d'identifier trois classes, chacune regroupant des SAUs partageant des caractéristiques temporelles de fréquentation similaires, différentes de celles des autres classes. Pour chaque classe, un modèle additif généralisé multivarié inspiré du modèle TSHR univarié, introduit dans le chapitre 2, est ensuite ajusté aux flux des SAUs appartenant à la classe, selon une procédure de régression PLS multivariée.

Les étapes de clustering et de modélisation multivariée par cluster sont détaillées dans les parties 4.1 et 4.2, respectivement.

4.1 Clustering des historiques de fréquentation

Commençons par rappeler quelques notations : $\mathbf{x}^{(d)} = x^{(d)}_{1:T_{\max}}$ est la série temporelle des nombres de passages dans le SAU $d \in \mathcal{D}$ entre le premier janvier 2016 et le 31 décembre 2018, qu'on nomme également historique de fréquentation du SAU d, où $\mathcal{D}$ est l'ensemble des 51 SAUs du Grand Est pour lesquels on dispose d'un historique complet ; voir la partie 1.3 du chapitre 1.

Pour tout SAU $d \in \mathcal{D}$, on note $\mathbf{z}^{(d)} = (\mathbf{x}^{(d)} - \overline{\mathbf{x}}^{(d)})/\sigma(\mathbf{x}^{(d)})$ la version centrée et réduite de $\mathbf{x}^{(d)}$, où

$\overline{\mathbf{x}}^{(d)} = \frac{1}{T_{\max}} \sum_{t=1}^{T_{\max}} x_t^{(d)}$ et $\sigma(\mathbf{x}^{(d)}) = \sqrt{\frac{1}{T_{\max}} \sum_{t=1}^{T_{\max}} (x_t^{(d)} - \overline{\mathbf{x}}^{(d)})^2}$ sont respectivement les moyenne et écart-type du nombre de passages quotidiens dans le SAU d sur la période d'étude.

Par « clustering des historiques de fréquentation », on entend associer à chaque SAU d un indice $k \in \{1, \ldots, K\}$, avec $K \in \mathbb{N}^*$, de sorte que les établissements qui partagent le même indice présentent des motifs temporels de fréquentation similaires, tout en étant distincts de ceux des établissements auxquels sont associés un autre indice.

Formellement, le résultat d'une méthode de clustering est un couple constitué d'une valeur $K \in \mathbb{N}^*$, avec $K < |\mathcal{D}| = 51$, et d'une fonction Cl de $\mathcal{D}$ dans $\{1, \cdots, K\}$. De nombreuses méthodes de clustering sont compatibles avec la nature numérique temporelle des historiques de fréquentation des SAUs ; plusieurs d'entre elles ont été utilisées et comparées. On s'attache dans la partie 4.1.1 à décrire les principales caractéristiques de ces méthodes et les modalités de leur comparaison. Les résultats obtenus sont présentés dans la partie 4.1.2.

4.1.1 Méthodes de clustering et modalités de comparaison de leurs performances

Les méthodes utilisées partagent un paradigme commun : un SAU d est identifié au point $\mathbf{z}^{(d)}$ de $\mathbb{R}^{T_{\max}}$ muni d'une distance notée génériquement dist. Les K groupes de SAUs sont formés de sorte que les distances entre les SAUs du groupe, ou entre les SAUs et un point de $\mathbb{R}^{T_{\max}}$ jouant le rôle de représentant du groupe, soient faibles et que les distances entre les groupes soient importantes.

Le choix de la version centrée et réduite $\mathbf{z}^{(d)}$ de l'historique de fréquentation pour représenter un SAU plutôt que l'historique lui-même favorise la création de groupes de SAUs partageant une proximité en termes de dynamique temporelle de fréquentation plutôt qu'en terme de volume de fréquentation. Plus concrètement, on souhaite regrouper ensemble des SAUs avec des patterns de fréquentation communs (pics de fréquentation pendant les week-ends, creux pendant les vacances, etc), sans que les volumétries soient nécessairement similaires. En effet, l'étape de modélisation multivariée qui va suivre le clustering vise à appréhender finement cette dynamique temporelle ; créer des clusters en donnant beaucoup d'importance à la volumétrie de fréquentation semble donc moins adapté.

On utilise dans la suite deux familles de méthodes de clustering : les méthodes hiérarchiques et les méthodes dérivées de l'algorithme des k-means. Les deux familles dépendent de la distance dont on munit $\mathbb{R}^{T_{\max}}$. La table 4.1 synthétise les principales caractéristiques des distances

utilisées dans la suite. La distance euclidienne a une importance particulière ; il s'agit de la distance « usuelle » préfigurant nos intuitions géométriques. Elle est associée à un produit scalaire. Ces caractéristiques offrent aux méthodes de clustering utilisant cette distance une grande interprétabilité, au prix d'un éventuel manque de pertinence pour des données s'écartant du paradigme gaussien, en particulier lorsque des valeurs extrêmes sont présentes. La distance de Manhattan, norme 1 ou norme L^1, constitue une alternative qui peut être adaptée à de telles données. Enfin, la distance DTW (*Dynamic Time Warping* - déformation temporelle dynamique) introduite par [Sakoe and Chiba, 1978], abusivement appelée distance – elle ne vérifie pas l'inégalité triangulaire – est spécifiquement adaptée au caractère séquentiel des séries temporelles. Elle juge de la proximité structurelle – régime de variations similaire – entre deux séries temporelles. Formellement, elle est définie comme la somme des plus courtes distances entre points des deux séries, la somme portant sur un sous-ensemble des couples. Dans notre contexte, la distance DTW entre deux historiques de fréquentation $\mathbf{z}^{(\mathbf{d_1})}$ et $\mathbf{z}^{(\mathbf{d_2})}$ est donnée par

$$\mathrm{DTW}(\mathbf{z}^{(\mathbf{d_1})}, \mathbf{z}^{(\mathbf{d_2})}) = \min_{c \in \mathcal{C}} \sum_{(t,s)\in c} |z_t^{(d_1)} - z_s^{(d_2)}|, \tag{4.1}$$

où $\mathcal{C}$ est l'ensemble des chemins $c = ((t_1, s_1), \cdots, (t_k, s_k))$ formés de k couples (t_i, s_i) d'instants (avec $T_{\max} \leq k$) vérifiant :

- les valeurs de départ et de fin des séries sont comparées, soit $(t_1, s_1) = (1, 1)$ et $(t_k, s_k) = (T_{\max}, T_{\max})$;
- tout chemin forme une suite croissante d'instants en chaque coordonnée et chaque instant apparaît au moins une fois dans chaque , soit

$$\begin{cases} t_{k-1} \leq t_k \leq t_{k-1} + 1, \\ s_{k-1} \leq s_k \leq s_{k-1} + 1. \end{cases}$$

L'ensemble $\mathcal{C}$ des chemins possibles est très grand, rendant inopérante la recherche du minimum définissant la distance DTW par exploration exhaustive des cas. Le calcul de ce minimum est effectué dans les implémentations usuelles par programmation dynamique ; voir par exemple [Tavenard, 2021].

Les méthodes hiérarchiques consistent à créer une suite croissante ou décroissante de partitions de l'ensemble des SAUs. Dans l'approche ascendante – on parle de classification ascendante hiérarchique (CAH) – on construit une suite décroissante de partitions partant de la partition la plus fine, constituée des singletons, $\mathcal{P}_0 = \{\{d\}, d \in \mathcal{D}\}$. A chaque itération i

Distance	Expression	Cadre applicatif	Comportement
Euclidienne (L^2)	$\left(\sum_t (x_t - y_t)^2\right)^{1/2}$	Données numériques	Adaptée aux données gaussiennes
Manhattan (L^1)	$\sum_t \lvert x_t - y_t \rvert$	Données numériques	Plus robuste que L^2 aux valeurs extrêmes
DTW	Voir Eq. (4.1)	Données temporelles	Proximité du régime de variations temporelles

Table 4.1 : Distances usuelles utilisées pour le clustering de séries temporelles.

de l'algorithme CAH, une nouvelle partition $\mathcal{P}_i$ est créée, constituée de $|\mathcal{D}| - i$ parties, obtenue en fusionnant deux parties de la partition précédente $\mathcal{P}_{i-1}$. Les deux parties à fusionner sont sélectionnées de sorte à minimiser une mesure de dissimilarité entre parties, quantifiant l'éloignement géométrique entre ces parties et dépendant de la distance sur l'espace $\mathbb{R}^{T_{\max}}$. La dissimilarité de Ward ou MISSQ (*Minimum Increase of Sum of SQuares*) – introduite dans [Ward, 1963] – est associée au carré de la norme euclidienne $\|\cdot\|$; elle quantifie l'écart quadratique entre les barycentres de deux parties A, B d'une partition $\mathcal{P}_i$, $i = 1, \cdots, D$, explicitement donné par

$$\frac{|A|.|B|}{|A \cup B|}\|\mu_A - \mu_B\|^2 = \sum_{x \in A \cup B} \|x - \mu_{A\cup B}\|^2 - \sum_{x \in A} \|x - \mu_A\|^2 - \sum_{x \in B} \|x - \mu_B\|^2,$$

où μ_A, μ_B et $\mu_{A\cup B}$ sont les barycentres des parties A, B et $A \cup B$, respectivement. La dissimilarité moyenne (*average*) ou UPGMA (*Unweighted Pair Group Method with Arithmetic mean*) entre deux parties A, B est égale à la moyenne des distances entre les éléments des deux parties. La dissimilartité médiane (*median*) ou WPGMC (Weighted Pair Group Method with Centroid average) entre A, B est égale au carré de la distance euclidienne entre les vecteurs des médianes des parties A et B.

La table 4.2 synthétise les caractéristique de ces trois mesures de dissimilarité que l'on utilisera dans le cadre de la classification des SAUs.

Dissimilarité	**Expression**
Ward.D2 ou MISSQ	$\frac{\lvert A\rvert.\lvert B\rvert}{\lvert A\cup B\rvert}\|\mu_A - \mu_B\|^2$
Average ou UPGMA	$\frac{1}{\lvert A\rvert.\lvert B\rvert}\sum_{x\in A}\sum_{y\in B}\mathrm{dist}(x,y)$
Median ou WPGMC	$\|\mathrm{Med}_A - \mathrm{Med}_B\|^2$

Table 4.2 : Mesures de dissimilarité utilisées dans les méthodes de clustering des SAUs basés sur la CAH.

Après $|D| - 1$ itérations, nous obtenons une suite décroissante de partitions $\mathcal{P}_0, \ldots, \mathcal{P}_{|D|-1}$ de tailles respectives $|\mathcal{D}|, \ldots, 1$ que l'on visualise sous forme d'un dendogramme sur lequel on présente les dissimilarités entre les parties fusionnées à chaque itération sur l'axe vertical. À tout

choix du nombre de classes K, correspond une partition $\mathcal{P}_{|D|-K}$ dont les éléments constituent les K classes. La partition produite par une CAH dépend donc de deux caractéristiques : la distance sur $\mathbb{R}^{T_{\max}}$ et la mesure de dissimilarité. L'un des enjeux de la partie 4.1.2 est d'optimiser le choix de ces caractéristiques. L'algorithme CAH est implémenté dans le package `{stats}` de R (installé par défaut avec le logiciel) sous la forme de la fonction `hclust()`. Cette dernière construit la suite de partitions $\mathcal{P}_0, \ldots, \mathcal{P}_{|D|-1}$. Une méthode `cut()`, spécifique aux objets de classe `hclust` produits par la fonction `hclust()`, permet de sélectionner la partition souhaitée selon le nombre de parties – donc, selon le nombre de classes de SAUs. La fonction `hclust()` prend comme argument principal la matrice des distances entre les SAUs, qui peut être obtenue, selon la distance choisie par les fonctions `dist()` (du package `{stats}`) ou `dtwDist()` (du package `{dtw}`).

Les méthodes dérivées de l'algorithme des k-means sont des algorithmes itératifs d'allocation des SAUs au sein de K classes, le nombre K de ces classes étant fixé a priori. Une étape d'initialisation consiste à choisir un représentant $r_k^{(0)}$ pour chaque classe $k \in \{1, \ldots, K\}$, c'est-à-dire un point de $\mathbb{R}^{T_{\max}}$ jouant le rôle de référence pour la classe; selon les méthodes, ces représentants peuvent être choisis aléatoirement ou en minimisant un critère, éventuellement en se limitant aux données observées $\mathbf{z}^{(d)}$. Chaque SAU – identifié au point $\mathbf{z}^{(\mathbf{d})} \in \mathbb{R}^{T_{\max}}$ – est ensuite alloué à la classe k dont le représentant est le plus proche de $\mathbf{z}^{(\mathbf{d})}$ selon une distance sur $\mathbb{R}^{T_{\max}}$. À l'issue de cette étape d'initialisation, on itère les deux étapes suivantes, jusqu'à convergence : 1) on met à jour le représentant de chaque classe, tenant compte des SAUs alloués à la classe puis 2) on met à jour l'allocation de chaque SAU en l'associant à la classe dont le nouveau représentant est le plus proche. Le procédé calculatoire du représentant d'une classe au cours de ces itérations est propre à chaque algorithme et peut dépendre du choix de la distance sur $\mathbb{R}^{T_{\max}}$, comme spécifié dans la Table 4.3. Ces méthodes dépendent de deux caractéristiques qui conditionnent le résultat : la distance sur $\mathbb{R}^{T_{\max}}$ et le procédé calculatoire du représentant.

Méthode	Distance	Représentant	Caractéristiques
k-means	L2 au carré	Moyenne	Adapté à des clusters isotropes (sphériques) et de même taille.
k-medians	L1	Vecteur des médianes	Robuste face à la présence de valeurs extrêmes.
k-medoids	À choisir	Médoïd = un point des données	Adapté à tout type de données.

TABLE 4.3 : Principales caractéristiques des trois méthodes de partitionnement de type k-means utilisées pour le clustering des SAUs.

Trois de ces méthodes sont utilisées par la suite : k-means, k-medians et k-medoids. La méthode des k-means est historiquement la plus ancienne, introduite indépendamment par [MacQueen,

1967] et [Lloyd, 1982]. L'éloignement entre deux points de $\mathbb{R}^{T_{\max}}$ est mesuré par le carré de la distance euclidienne ; le représentant d'une classe est le barycentre des points appartenant à la classe. La distance et le mode de calcul du représentant font des k-means une méthode de clustering adaptée aux données dont les classes sont isotropes (même variance dans toutes les dimensions), sensible aux valeurs extrêmes. Également, la composition des classes en sortie des k-means dépend fortement de l'initialisation des représentants des classes. Une variante appelée k-means++, introduite par [Arthur and Vassilvitskii, 2007], insère une étape entre l'initialisation aléatoire des représentants et la mise à jour itérative des classes et de leurs représentants. Cette étape consiste à remplacer les représentants aléatoires initiaux par K points parmi les d points $\mathbf{z}^{(d)}$, choisis aléatoirement selon une loi associant à chacun une probabilité proportionnelle au carré de la distance euclidienne entre ce point et le représentant initial le plus proche.

Les k-medians utilisent la distance L1 et prennent pour représentant d'une classe le vecteur des médianes univariées des SAUs de la classe – c'est-à-dire, les vecteurs dont les composantes sont les médianes, calculées pour chaque jour, des nombres de passages des SAUs de la classe. Ces choix de distance et représentants rendent la méthode robuste à la présence de valeurs extrêmes.

Les méthodes des k-means et k-medians disposent d'implémentations dans le logiciel R, sous la forme des fonctions `kmeans()` et `kcca()` des packages `{stats}` et `{flexclust}`. Ce dernier permet l'initialisation des représentants selon la méthode k-means++ (aussi bien pour les k-means que pour les k-medians).

Enfin, les k-medoids offrent une plus grande flexibilité sur le choix de la distance utilisée, en le dissociant du choix des représentants. Ces derniers sont choisis parmi les points $\mathbf{z}^{(d)}$, $d \in \mathcal{D}$, de sorte à minimiser la somme des distances aux autres points de la classe. Il s'agit d'un problème algorithmiquement complexe – voir [Hsu and Nemhauser, 1979] – pour lequel plusieurs heuristiques ont été développées. Parmi celles-ci, les algorithmes PAM (*Partitioning Around Medoid* et CLARA (*Clustering LARge Applications*) ; voir [kau, 1990]. Le premier consiste à initialiser le choix des représentants en les sélectionnant itérativement selon une approche gloutonne puis à optimiser ce choix en procédant à des échanges entre ces représentants et des points du cluster qu'ils définissent. Sa complexité est de l'ordre de $O(|D|^2)$. Le second consiste à utiliser PAM sur des sous-ensembles du jeu de données brutes de sorte à atteindre une complexité algorithmique de l'ordre de $O(|D|)$, au prix d'un choix des représentants moins précis. Ces deux algorithmes sont implémentés en R, sous la forme des fonctions `pam()` et `clara()` du package

`{cluster}`. Dans le contexte des SAUs, dont le nombre est limité, l'algorithme PAM est tout à fait adapté.

Le choix a priori d'une méthode de clustering et des caractéristiques associées (distance, mesure de dissimilarité ou nature des représentants) adaptée aux historiques de fréquentation des SAUs n'a rien d'évident. On a adopté une approche exploratoire visant à comparer les résultats des méthodes précédemment décrites pour plusieurs choix de leurs caractéristiques selon une mesure de l'homogénéité des classes, appelée silhouette, introduite par [Rousseeuw, 1987]. Étant donnés un nombre de classes K et une fonction Cl d'allocation aux classes, on associe à chaque SAU d son coefficient de silhouette, défini par

$$\mathrm{Sil}(d|K,\mathrm{Cl}) = \frac{\mathrm{out}(d) - \mathrm{in}(d)}{\max\{\mathrm{out}(d), \mathrm{in}(d)\}},$$

où $\mathrm{in}(d)$ et $\mathrm{out}(d)$ sont respectivement la moyenne des distances de d aux SAUs de la même classe et la moyenne des distances de d aux SAUs de la classe voisine :

$$\begin{aligned} \mathrm{in}(d) &= \frac{1}{|\mathrm{C}_{k_d}| - 1} \sum_{d' \in \mathcal{C}_{k_d}, d' \neq d} \mathrm{dist}(d, d'), \\ \mathrm{out}(d) &= \min_{k \neq k_d} \frac{1}{|\mathcal{C}_k|} \sum_{d' \in \mathcal{C}_k} \mathrm{dist}(d, d'), \end{aligned}$$

avec $\mathcal{C}_{k_d}$ la classe contenant d et dist la distance sur $\mathbb{R}^{T_{\max}}$. Pour tout SAU $d \in \mathcal{D}$, son coefficient de silhouette $\mathrm{Sil}(d|K,\mathrm{Cl})$ est compris entre -1 et 1. Un SAU est d'autant mieux classé que son coefficient de silhouette est proche de 1. Une valeur négative indique qu'il existe une autre classe que celle à laquelle est alloué le SAU dont les éléments lui ressemblent plus, suggérant un manque de pertinence de l'allocation $\mathrm{Cl}(d)$. La moyenne des coefficients de silhouette des SAUs, appelée coefficient de silhouette de la méthode de clustering (K, Cl), donnée par

$$\mathrm{Sil}(K,\mathrm{Cl}) = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \mathrm{Sil}(d|K,\mathrm{Cl}), \tag{4.2}$$

quantifie la pertinence de la méthode.

Le coefficient de silhouette permet de comparer les performances de plusieurs couples (K, Cl) puis de les hiérarchiser. Concrètement, les couples (K, Cl) que nous considérons dans la suite sont obtenues en appliquant les méthodes de clustering CAH, k-means, k-means++, K-medians et k-medoids pour différents choix de distance et de mesure de dissimilarlité ou de représentant des classes décrits dans la table 4.4. On note Meth l'ensemble de ces méthodes. À chaque méthode $\mathcal{M}$, on associe une courbe de silhouette, associant à chaque entier K, la valeur du coefficient de silhouette $\mathrm{Sil}(K, \mathrm{Cl}_{\mathcal{M}})$ de la fonction d'allocation $\mathrm{Cl}_{\mathcal{M}}$ aux K classes fournie par la méthode $\mathcal{M}$. On restreint le nombre de classes K à l'ensemble $\{3, \ldots, 10\}$, afin réaliser un

compromis entre nécessité de réduire le nombre de SAUs à prendre en compte simultanément dans la modélisation multivariée (donc disposer de classes de faibles effectifs) et la capacité à tirer profit des corrélations entre SAUs d'une même classe. La méthode de clustering (K^*, Cl^*) que l'on retient pour définir les classes de SAUs utilisées dans la partie 4.2 est celle maximisant le coefficient de silhouette :

$$(K^*, \mathrm{Cl}^*) = \underset{(K,\mathcal{M})\in\{3,\cdots,10\}\times\mathrm{Meth}}{\arg\max} \mathrm{Sil}(K, \mathrm{Cl}_{\mathcal{M}}).$$

Méthode	Distance	Dissimilarité/Représentant
CAH	L2	Ward.D2
CAH	L2	UPGMA
CAH	L1	WPGMC
CAH	DTW	WPGMC
k-means	L2 au carré	Moyenne
k-means++	L2 au carré	Moyenne
k-medianes	L1	Vecteur des médianes
k-medoids (PAM)	L2	Médoïde
k-medoids (PAM)	L2 au carré	Médoïde
k-medoids (PAM)	L1	Médoïde
k-medoids (PAM)	DTW	Médoïde

Table 4.4 : Ensemble des méthodes de clustering et leurs caractéristiques comparées dans le cadre de la définition des classes de SAUs. On renvoie à la table 4.2 pour la définition des acronymes des mesures de dissimilarité.

4.1.2 Résultats du clustering

Les courbes de silhouette associées aux méthodes de clustering listées dans la table 4.4 sont présentées par la figure 4.1. Parmi les onze méthodes explorées, sept présentent une courbe de silhouette qualitativement décroissante, avec un maximum pour $K = 3$ classes, deux (CAH avec distance L2 et dissimilarité moyenne puis k-means++) un maximum pour $K = 4$ classes. Les k-medians présentent une courbe de silhouette instable, avec un maximum en $K = 6$ tandis que et k-medoids avec distance DTW ont une courbe de silhouette quasi-constante, avec un maximum en $K = 7$.

La procédure de clustering optimale (K^*, Cl^*) est celle associée aux k-means, pour 3 classes. Elle est suivie par les k-medoids avec distance L2 au carré. Ci-dessous, une brève description de la composition des trois classes résultant de la procédure optimale, s'appuyant sur une analyse descriptive présentée en annexe de ce chapitre.

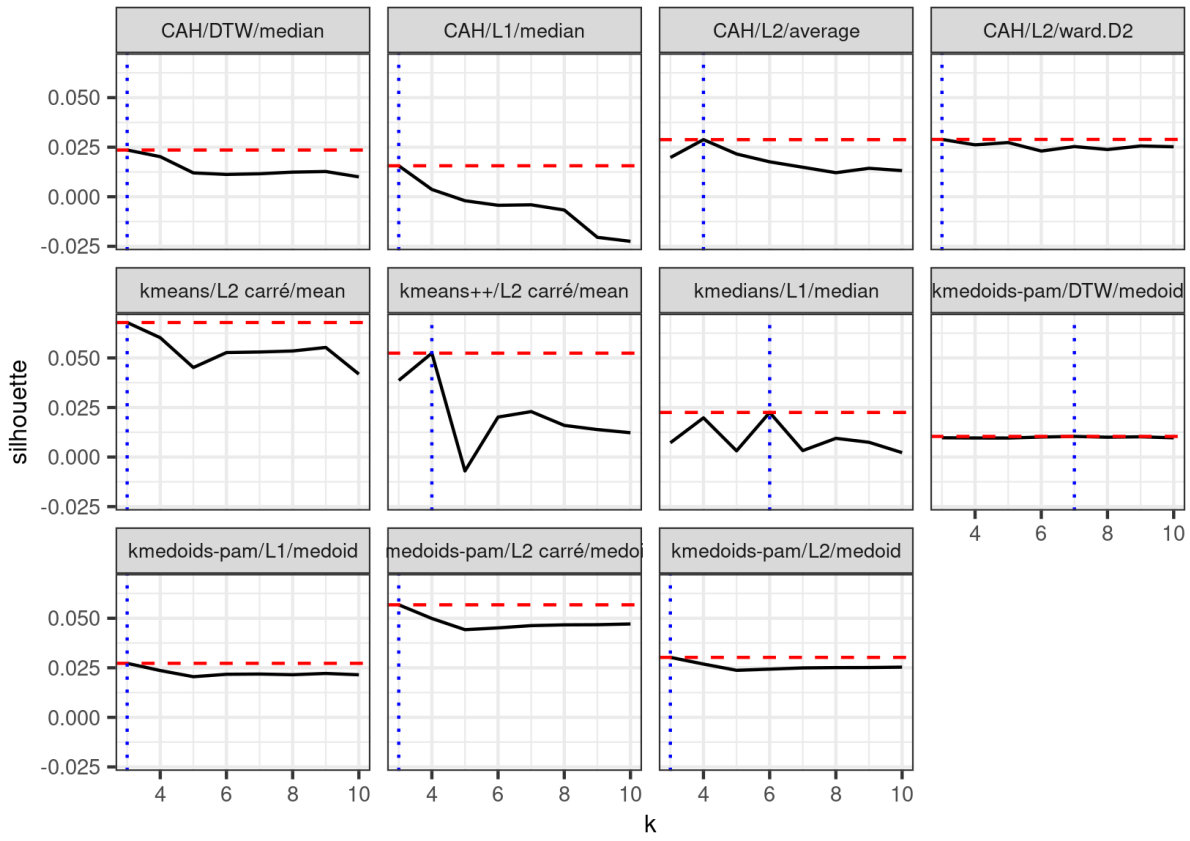

FIGURE 4.1 : Courbes de silhouette des onze méthodes de clustering appliquées aux historiques de fréquentation centrés et réduits. Les lignes horizontales hachurées rouges identifient les valeurs maximales de chacune des courbes tandis que les lignes verticales en pointillés bleus identifient les nombres de classes pour lesquels ces maximum sont atteints.

- Classe 1 : composée de dix-neuf SAUs concentrés dans les régions de Lorraine et d'Alsace (dix-huit sur dix-neuf). Les SAUs de cette classe ont tendance à présenter de plus faibles fréquentations (que les autres SAUs) les jours de semaine en hiver et de plus fortes fréquentations les week-ends de printemps et mois de juillet et août.

- Classe 2 : composée de onze SAUs. Elle regroupe les cinq SAUs pédiatriques de la région Grand Est, ainsi que quatre SAUs parmi les cinq que compte la Moselle marqués par des flux de passage parmi les plus importants de la région, ainsi que les CHs de Châlons (Marne) et Épinal (Vosges). Ces établissements partagent des tendances de fréquentation similaires, avec des afflux de patients plus importants les week-ends d'automne et d'hiver, un moindre afflux pendant les vacances scolaires ; ces tendances variant subtilement en intensité et en durée selon les années.

- Classe 3 : composée de vingt et un SAUs. Elle regroupe onze des quatorze établissements de Champagne-Ardenne. Quatre des six cliniques de la région appartiennent à

cette classe. Les SAUs de cette classe ont des tendances faiblement décroissantes de fréquentation au cours du temps. Egalement une tendance de moindre fréquentation les week-ends d'automne et d'hiver et des fréquentations plus élevées pendant les vacances scolaires.

4.2 Modélisation multivariée via régression PLS par cluster

Nous considérons dans cette partie, le problème de modélisation multivariée de l'ensemble des séries temporelles appartenant à un même cluster, l'objectif étant d'améliorer la robustesse et la performance de prévision des modèles univariés, en exploitant l'information apportée par chacune des séries, et de réduire le temps de calcul lors des étapes de sélection des hyperparamètres et d'ajustement des modèles univariés TSHR.

Dans la suite, nous considérons une classe quelconque (parmi les trois classes obtenues dans la partie 4.1), que l'on note C, le nombre de séries temporelles appartenant à la classe C sera noté d. Pour rappel, la méthode de clustering employée, voir partie 4.1, a permis de distinguer trois classes homogènes C_1, C_2 et C_3, respectivement, d'effectif $d_1 = 19$, $d_2 = 11$ et $d_3 = 21$. Dans cette approche, nous utilisons le modèle classique de régression des moindres carrés partielles (Partial Least Squares Regression, PLSR), avec un ensemble de régresseurs bien choisi, issu de l'application de la procédure de sélection de modèles TSHR univariés, du chapitre 2, pour la modélisation de la série moyenne (représentant de la classe).

Nous présentons brièvement le modèle de régression PLS dans la partie 4.2.1, nous précisons ensuite la manière de construire l'ensemble des régresseurs.

4.2.1 Le modèle de régression PLS

Le modèle de régression PLS généralise et combine les techniques de l'analyse en composantes principales et la régression linéaire multiple, largement utilisée pour la prévision d'un ensemble de variables réponses (pour la version multivariée) en fonction d'un grand nombre de variables régresseurs. Elle a été introduite par [Wold, 1966], et généralisée aux cas multivarié et de grande dimension par [Martens and Naes, 1984], [Martens and Næs, 1992]. Un autre point de vue est donnée à la méthode, comme modèle statistique, dans les travaux de [Helland, 1990] et [Frank and Friedman, 1993]. Il existe dans la littérature plusieurs versions de

la méthode PLS, selon la dimensionnalité (une ou plusieurs variables réponses à modéliser) ou la finalité poursuivie [Tenenhaus, 1998] : la régression PLS univariée, les méthodes PLS multivariées pour l'analyse canonique, l'analyse discriminante et la régression. C'est cette dernière que l'on utilise pour la modélisation des séries temporelles multivariées. Nous rappelons ci-dessous l'algorithme permettant la construction de ce modèle. Introduisons d'abord les notations suivantes. Notons

$$\mathbb{Y} = [y_{i,j}]_{i=1,\ldots,n;\, j=1,\ldots,d} \tag{4.3}$$

le tableau des données centrées réduites, $n \times d$, constitué de n, $n \geq 1$, observations décrites par d, $d > 1$, variables réponses $Y_1, \ldots, Y_d$. Notons

$$\mathbb{X} = [x_{i,j}]_{i=1,\ldots,n;\, j=1,\ldots,p} \tag{4.4}$$

le tableau des données centrées réduites, $n \times p$, constitué de n observations décrites par p, $p \geq 1$, variables régresseurs $X_1, \ldots, X_p$. Dans le cas d'une seule variable réponse (i.e., $d = 1$ et par conséquent $\mathbb{Y}$ est un vecteur colonne de dimension n) et d'une matrice $\mathbb{X}$ des variables régresseurs de plein rang (i.e. p, et cela suppose que $n \geq p$), le problème se ramène à une modélisation par régression linéaire multiple (RLM) classique. Si la matrice $\mathbb{X}$ n'est pas de plein rang, due au problème de multicolinéarité ou du fait que le nombre de régresseurs est large par rapport au nombre d'observations, le modèle RLM n'est plus adapté. Plusieurs approche existent pour surmonter cette difficulté, parmi lesquelles, les méthodes de régression régularisées de type Lasso ou Ridge, voir e.g. [Tibshirani, 1996], ou les modèles de régression sur composantes principales du tableau de données $\mathbb{X}$ des régresseurs. L'orthogonnalité des composantes principales répond au problème de multicolinéarité des régresseurs. Une réduction de dimension peut être réalisée en conservant un nombre réduit de composantes principales, plus petit que le nombre de régresseurs. Néanmoins, les composantes principales sont pertinentes pour expliquer et résumer l'information apporter par l'ensemble de tous les régresseurs, et ne sont pas forcément pertinentes pour les variables réponses. L'un des rôles de la régression PLS est de répondre à cette limite. On cherche à construire r, $1 \leq r \leq p$, composantes orthogonales (combinaisons linéaires des variables régresseurs)

$$\underset{(n\times 1)}{\mathbf{t}_h} = \underset{(n\times p)}{\mathbb{X}}\, \underset{(p\times 1)}{\mathbf{a}_h}\,, \quad h = 1, \ldots, r, \tag{4.5}$$

et r composantes orthogonales (combinaisons linéaires des variables réponses)

$$\underset{(n\times 1)}{\mathbf{u}_h} = \underset{(n\times d)}{\mathbb{Y}}\, \underset{(d\times 1)}{\mathbf{b}_h}\,, \quad h = 1, \ldots, r, \tag{4.6}$$

explicatives de leur propre groupe, et bien corrélées entre elles. Elles sont obtenues par maximisation du carré de la covariance

$$\{\mathbf{a}_h, \mathbf{b}_h\} := \underset{\mathbf{a}\in\mathbb{R}^p, \mathbf{b}\in\mathbb{R}^d : \|\mathbf{a}\|=\|\mathbf{b}\|=1}{\arg\max} \operatorname{cov}^2(\mathbb{X}\mathbf{a}, \mathbb{Y}\mathbf{b}) \tag{4.7}$$

$$= \underset{\mathbf{a}\in\mathbb{R}^p, \mathbf{b}\in\mathbb{R}^d : \|\mathbf{a}\|=\|\mathbf{b}\|=1}{\arg\max} \operatorname{cor}^2(\mathbb{X}\mathbf{a}, \mathbb{Y}\mathbf{b}) \operatorname{var}(\mathbb{X}\mathbf{a}) \operatorname{var}(\mathbb{Y}\mathbf{b}), \tag{4.8}$$

pour $h = 1$, puis par itération sous contraintes d'orthogonalité, pour les composantes suivantes. On déduit de la relation (4.8) que la régression PLS réalise un compromis entre l'analyse en composantes principales et la régression. Ce calcul itératif est réalisé souvent par l'algorithme NIPALS (Non Linear Iterative PArtial Least Square), que nous rappelons ici :

Algorithm 1 NIPALS

$\mathbb{Y}$: matrice $n \times d$ des données centrées réduites (associée aux variables réponses)
$\mathbb{X}$: matrice $n \times p$ des données centrées réduites (associée aux variables régresseurs)
Initialiser $\mathbf{u}_1$ par la première colonne de $\mathbb{Y}$ ($\mathbf{u}_1 = \mathbb{Y}[, 1]$)
for $h = 1, \ldots, r$ **do**
  while convergence non atteinte **do**
    $\mathbf{a}_h = \mathbb{X}'\mathbf{u}_h/(\mathbf{u}_h'\mathbf{u}_h)$
    $\mathbf{a}_h = \mathbf{a}_h/(\mathbf{a}_h'\mathbf{a}_h)$ (normalisation)
    $\mathbf{t}_h = \mathbb{X}\mathbf{a}_h$ (variable latente associée aux régresseurs)
    $\mathbf{b}_h = \mathbb{Y}'\mathbf{t}_h/(\mathbf{t}_h'\mathbf{t}_h)$
    $\mathbf{b}_h = \mathbf{b}_h/(\mathbf{b}_h'\mathbf{b}_h)$ (normalisation)
    $\mathbf{u}_h = \mathbb{Y}'\mathbf{b}_h$ (variable latente associée aux variables réponses)
  end while
  $\mathbf{c}_h = \mathbb{X}'\mathbf{t}_h/(\mathbf{t}_h'\mathbf{t}_h)$ (régression partielle de $\mathbb{X}$ sur $\mathbf{t}_h$)
  $\mathbb{X} \leftarrow \mathbb{X} - \mathbf{t}_h\mathbf{c}_h'$ (Résidus de $\mathbb{X}$)
  $\mathbb{Y} \leftarrow \mathbb{Y} - \mathbf{t}_h\mathbf{b}_h'$ (Résidus de $\mathbb{Y}$)
end for

Les composantes orthogonales $\{\mathbf{t}_1, \ldots, \mathbf{t}_r\}$ et $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ s'appellent les variables latentes, associées au tableau des régresseurs $\mathbb{X}$ et tableau des variables réponses $\mathbb{Y}$, respectivement. Les vecteurs de $\mathbb{R}^p$, $\{\mathbf{a}_1, \ldots, \mathbf{a}_r\}$, et les vecteurs de $\mathbb{R}^d$, $\{\mathbf{b}_1, \ldots, \mathbf{b}_r\}$, s'appellent les vecteurs "loadings", de normes égalent à 1, mais non nécessairement orthogonaux. Enfin, un modèle de régression multidimensionnelle de $\mathbb{Y}$ sur les variables latentes

$$\underset{(n\times r)}{\mathbb{T}} := [\mathbf{t}_1 \cdots \mathbf{t}_r], \tag{4.9}$$

associées aux régresseurs, est réalisé

$$\begin{aligned} \underset{(n\times d)}{\mathbb{Y}} &= \underset{(n\times r)}{\mathbb{T}}\ \underset{(r\times d)}{\widehat{\mathbb{B}}} + \underset{(n\times d)}{\mathbb{E}}, \\ &=: \underset{(n\times d)}{\widehat{\mathbb{Y}}} + \underset{(n\times d)}{\mathbb{E}} \end{aligned} \tag{4.10}$$

où $\widehat{\mathbb{B}}$ est l'estimateur MCO des paramètres du modèle de régression, $\underset{(n\times d)}{\widehat{\mathbb{Y}}}$ est le tableau des valeurs ajustées, $\mathbb{E}$ étant la matrice des résidus.

Le choix du nombre de composantes r est effectué par minimisation de l'erreur de prévision, estimée par validation croisée.

La méthode PLS est implémentée dans le package `{pls}`, voir [Liland et al., 2024], du logiciel R, sous la forme de la fonction `mvr()`, offrant la possibilité de sa mise en oeuvre en mode canonique ou régression, et la possibilité de choisir le nombre de composantes principales selon différentes méthodes de validation croisée.

4.2.2 Application du modèle de régression PLS pour une modélisation multivariée des séries temporelles

Considérons l'ensemble C constitué des d historiques de fréquentation des SAUs appartenant à un même cluster notés

$$x^{(j)}_{1:T_{\max}}, \quad j = 1, \ldots, d, \quad \text{avec } T_{\max} = 1096,$$

i.e.,

$$C := \left\{ x^{(j)}_{1:T_{\max}}; \quad T_{\max} = 1096, \quad j = 1, \ldots, d \right\}. \tag{4.11}$$

Notons

$$C_{\text{app}} := \left\{ x^{(j)}_{1:T_{\text{app}}}; \quad T_{\text{app}} = 1008, \quad j = 1, \ldots, d \right\}, \tag{4.12}$$

l'ensemble des séries temporelles d'apprentissage associées et

$$\overline{x}_t := \frac{1}{d} \sum_{j=1}^{d} x^{(j)}_t; \quad t = 1, \ldots, T_{\text{app}}, \tag{4.13}$$

la série temporelle moyenne, représentant de la classe C_{app}. L'objectif étant de construire un modèle multivariée de prévision de l'ensemble des séries temporelles appartenant à la classe, le tableau des variables réponses est choisi naturellement de la forme

$$\mathbb{Y} = \left[x^{(j)}_{1:T_{\text{app}}} \right]_{j=1,\ldots,d}, \tag{4.14}$$

constitué ainsi de $n = T_{\text{app}} = 1008$ lignes et de d colonnes, d étant le nombre de séries temporelles à modéliser simultanément. Dans cette représentation, chacune des séries temporelles, $x^{(j)}_{1:T_{\text{app}}}$, $j = 1, \ldots, d$, est disposée en colonne (un vecteur colonne de dimension $n = T_{\text{app}} = 1008$). Pour construire le tableau de données associées aux variables régresseurs, on procède comme suit. On modélise la série moyenne (univarié) (4.13) ci-dessus à l'aide d'un modèle

TSH + ARMA$(0,0)$, en suivant la méthode de sélection des hyperparamètres proposée dans la section (2.3.2) du chapitre 2. Notons alors $\widehat{K}$, $\widehat{m}$ et $\widehat{\mathcal{H}}$, les hyperparamètres optimaux du modèle TSH + ARMA$(0,0)$, donnant, respectivement, le degré de la tendance polynomiale, le nombre d'harmoniques de la composante saisonnière et le sous-ensemble de variables vacances/évènements sélectionné. Le tableau des variables régresseurs est alors choisi comme suit

$$\mathbb{X} = \left[P_1(t), \ldots, P_{\widehat{K}}(t), \cos(2\pi f_1 t), \sin(2\pi f_1 t), \ldots, \cos(2\pi f_{\widehat{m}} t), \sin(2\pi f_{\widehat{m}} t), \widehat{\mathcal{H}}\right]_{t=1,\ldots,T_{\text{app}}}, \quad (4.15)$$

où $P_1(t), \ldots, P_{\widehat{K}}(t)$ sont les $\widehat{K}$ polynômes orthogonaux introduits dans (2.7), $f_1, \ldots, f_{\widehat{m}}$ sont les fréquences ordonnées définies dans (2.15) et $\widehat{\mathcal{H}}$ est le tableau de données associées au sous-ensemble de variables vacances/évènements sélectionné. Ainsi, le tableau $\mathbb{X}$ est constitué de $n = T_{\text{app}} = 1008$ lignes et de $p = \widehat{K} + 2\widehat{m} + n_{\widehat{\mathcal{H}}}$ colonnes, $n_{\widehat{\mathcal{H}}}$ étant le nombre de variables vacances/évènements sélectionnées.

Enfin, par application de la méthode PLS multivariée de régression, aux versions centrées et réduites (notons-les encore $\mathbb{Y}$ et $\mathbb{X}$) des deux tableaux de données (4.14) et (4.15) ci-dessus, nous obtenons le modèle multivarié suivant (à r composantes)

$$\begin{aligned} \mathbb{Y} &= \mathbb{T}\,\widehat{\mathbb{B}} + \mathbb{E}, \\ &=: \widehat{\mathbb{Y}} + \mathbb{E} \end{aligned} \quad (4.16)$$

où $\widehat{\mathbb{B}}$ est l'estimateur MCO des paramètres du modèle de régression, $\mathbb{T}$ est le tableau des variables latentes associées aux variables régresseurs (4.15), $\widehat{\mathbb{Y}} := \mathbb{T}\,\widehat{\mathbb{B}}$ est le tableau des séries temporelles ajustées, $\mathbb{E}$ étant la matrice des résidus. Le nombre r de composantes est sélectionné en minimisant l'erreur de prévision, estimée par la méthode de validation croisée leave-one-out.

Le tableau 4.5 résume les hyperparamètres optimaux obtenus $\widehat{K}$, $\widehat{m}$, $n_{\widehat{\mathcal{H}}}$ et r, lors de la construction du modèle (4.16) pour les trois classes C_1, C_2 et C_3.

Classe, Effectif	$\widehat{K}$	$\widehat{m}$	$n_{\widehat{\mathcal{H}}}$	r
C_1, $d_1 = 19$	0	22	0	5
C_2, $d_2 = 11$	1	28	1	9
C_3, $d_3 = 21$	1	29	1	9

Table 4.5 : Les hyperparamètres optimaux des modèles multivariés associés aux trois classes de SAUs.

4.2.3 Évaluation de la performance de prévision des trois modèles multivariés

Nous évaluons ici la performance de prévision des trois modèles multivariés (un modèle par classe), présentés dans la section (4.2.2) précédente, à différent horizons ($H = 7$, 14, 21 et 28), en utilisant la méthode de validation croisée séquentielle (2.4.2). Nous reportons les résultats dans le tableau 4.6.

Mesures de performance des prévisions	MAE				WMAPE			
Horizon de prévision H	7	14	21	28	7	14	21	28
Cluster C_1, $d_1 = 19$	7.47	7.48	7.45	7.45	12.62	12.68	12.66	12.65
Cluster C_2, $d_2 = 11$	10.34	10.36	10.29	10.27	12.14	12.27	12.29	12.27
Cluster C_3, $d_3 = 21$	8.04	8.07	8.02	7.97	12.27	12.35	12.35	12.31

TABLE 4.6 : MAE et WMAPE agrégés (moyenne des SAUs appartenant à la classe) aux horizons de prévision de 7 à 28 jours pour les trois modèles multivariés.

Ces résultats montrent une efficacité de prévision équivalente en terme de l'erreur de prévision MAE, et une légère baisse de performance en terme de l'erreur de prévision WMAPE, en comparaison avec l'approche univarié des chapitres 2 et 3, voir table 3.1. Cependant, cette approche multivariée a permis de réduire de manière très importante le temps de calcul. En effet, cette approche multivariée utilise la procédure de sélection 2.3.2 des hyperparamètres pour modéliser un nombre réduit de séries, seulement trois séries (les trois séries types des trois classes), contrairement à l'approche univariée qui nécessite l'application de la procédure de sélection des hyperparamètres pour chacune des 51 séries des 51 SAUs.

4.3 Annexe : description de la composition des classes

La table 4.7 identifie la répartition des SAUs au sein des trois classes retenues à l'issue de la procédure de clustering par k-means. Les trois classes contiennent respectivement dix-neuf, onze et vingt et un SAUs.

La répartition géographique des SAUs des classes est illustrée par la figure 4.2. Les classes 3 et 1 se différencient géographiquement par une opposition Est-Ouest. En effet, onze établissement sur les quatorze de Champagne-Ardenne (Ardennes, Aube, Haute-Marne et Marne) sont dans la classe 3, un seul (Langres) dans le cluster 1, possédant un score de silhouette négatif. La classe 1 regroupe neuf établissements sur les quatorze d'Alsace (Bas et Haut-Rhin). La classe 2 regroupe quatre des cinq établissements de Moselle.

Sont également figurés sur la carte 4.2, les types des SAUS (CH, CHR, CHU, Clinique) ainsi que leur caractère exclusivement pédiatrique ou non. La classe 2 regroupe les cinq établissements pédiatriques. La classe 3 regroupe quatre cliniques sur les six de la région Grand Est (disposant d'un SAU).

La distribution des fréquentations quotidiennes moyennes par classe est illustrée sous forme de boîtes à moustaches par la figure 4.3. La classe 2 se distingue des deux autres classes par la présence d'établissements présentant des fréquentations hebdomadaires moyennes globalement plus élevées, avec une forte hétérogénéité due à la présence d'établissements à très forte fréquentation (CHRs de Metz et Thionville et CH de Forbach, en Moselle, service pédiatrique du CHU de Strasbourg).

Nous avons réalisé une analyse en composantes principales (ACP) des historiques de fréquentations centrés réduits $z_{1:T}^{(d)}$, $d \in \mathcal{D}$, consistant à diagonaliser la matrice des corrélations entre ces historiques.

Le plan principal associé à cette ACP représente une proportion modérée de l'inertie (environ 13 %), mais tout de même appréciable au regard de la dimension de la matrice de corrélation (de taille 51×51). Les qualités de représentation de SAUs dans ce plan sont présentées par la figure 4.4. Les établissements pédiatriques (classe 2) sont les mieux représentés dans ce plan avec des qualités comprises entre $0,3$ et $0,5$. Les autres SAUs de la classes 2 présentent des qualités inférieures à $0,25$. Dans les classes 1 et 3, seuls cinq établissements présentent des qualités de représentation supérieures à $0,2$ (le CH St-Dié et service adultes du CHU de Nancy pour la classe 1, les deux cliniques Courlancy et le CH de Sedan pour la classe 3).

Classe	Taille	SAU
1	19	CH Altkirch CH Bar-le-Duc CH Briey CH Guebwiller CH Haguenau CH Langres CH Mulhouse Hôpital Emile Muller adultes CH Pont à Mousson CH Remiremont CH Sarrebourg CH St Dié CH Sélestat CH Toul CH Vittel CH Wissembourg CHU Nancy Central adultes CHU Strasbourg NHC Cli Sainte Anne Pôle public St Louis
2	11	CH Chalons CH Charleville pediatrie CH Epinal CH Forbach CH Mulhouse Hôpital Emile Muller pédiatrie CHR Metz CHR Thionville CHU Nancy Brabois pediatrie CHU Reims pediatrie CHU Strasbourg Hautepierre pédiatrie HCli Claude Bernard
3	21	CH Charleville adultes CH Chaumont CH Epernay CH Lunéville CH Mont St Martin CH Neufchateau CH Sarreguemines CH Saverne CH Sedan CH St Dizier CH Troyes CH Verdun CH Vitry CHU Reims adultes CHU Strasbourg Hautepierre adultes Cli Courlancy Cli Courlancy Reims-Bezannes Cli Gentilly Diaconat Fonderie GHSA Rethel Vouziers Hôpital de St Avold

TABLE 4.7 : Composition et taille des trois classes obtenues à l'issue de la procédure de clustering par k-means.

Les corrélations des composantes principales associées aux deux premiers axes avec les 1096 fréquentations hebdomadaires (centrées réduites par établissement) sont présentées sous forme

de cartes de chaleur structurées en calendrier annuel pour les années 2016 à 2018 - figures 4.5 pour la première dimension de l'ACP et 4.6 pour la deuxième dimension.

Le premier axe de l'ACP fait apparaître une dérive croissante de la fréquentation avec le temps, ce premier ayant tendance à être négativement corrélé avec les dates les plus anciennes et positivement corrélé avec les plus récentes. À cette tendance, s'ajoute une opposition entre faibles (à gauche dans le plan principal) et fortes (à droite) fréquentations les week-ends d'automne et d'hiver, fortes (à gauche) et faibles (à droite) fréquentation lors des périodes scolaires.

Le second axe est négativement corrélé aux jours de semaine d'hiver, positivement corrélés avec les week-ends de printemps et les mois de juillet et août.

La projection des SAUs sur le plan principal associé à cette ACP est présentée par la figure **??**. Le premier axe oppose les classes 3 (à gauche) et 2 (à droite). Le second oppose la classe 1 (en haut) aux deux autres classes. Les établissements pédiatriques se détachent en bas à droite. Croisée avec l'interprétation des corrélations des composantes avec les fréquentations quotidiennes, cette répartition suggère les tendances suivantes : Les SAUs de la classe 1 ont tendance à présenter de plus faibles fréquentations les jours de semaine en hiver et de plus fortes fréquentations les week-ends de printemps et mois de juillet et août. La classe 2 présente une tendance de fréquentation croissante au cours du temps, avec des fréquentations plus élevées les week-ends d'automne et d'hiver et une moindre fréquentation pendant les vacances scolaires. La classe 3 présente des établissements avec des tendances faiblement décroissantes de fréquentation au cours du temps. Également une tendance de moindre fréquentation les week-ends d'automne et d'hiver et des fréquentations plus élevées pendant les vacances scolaires.

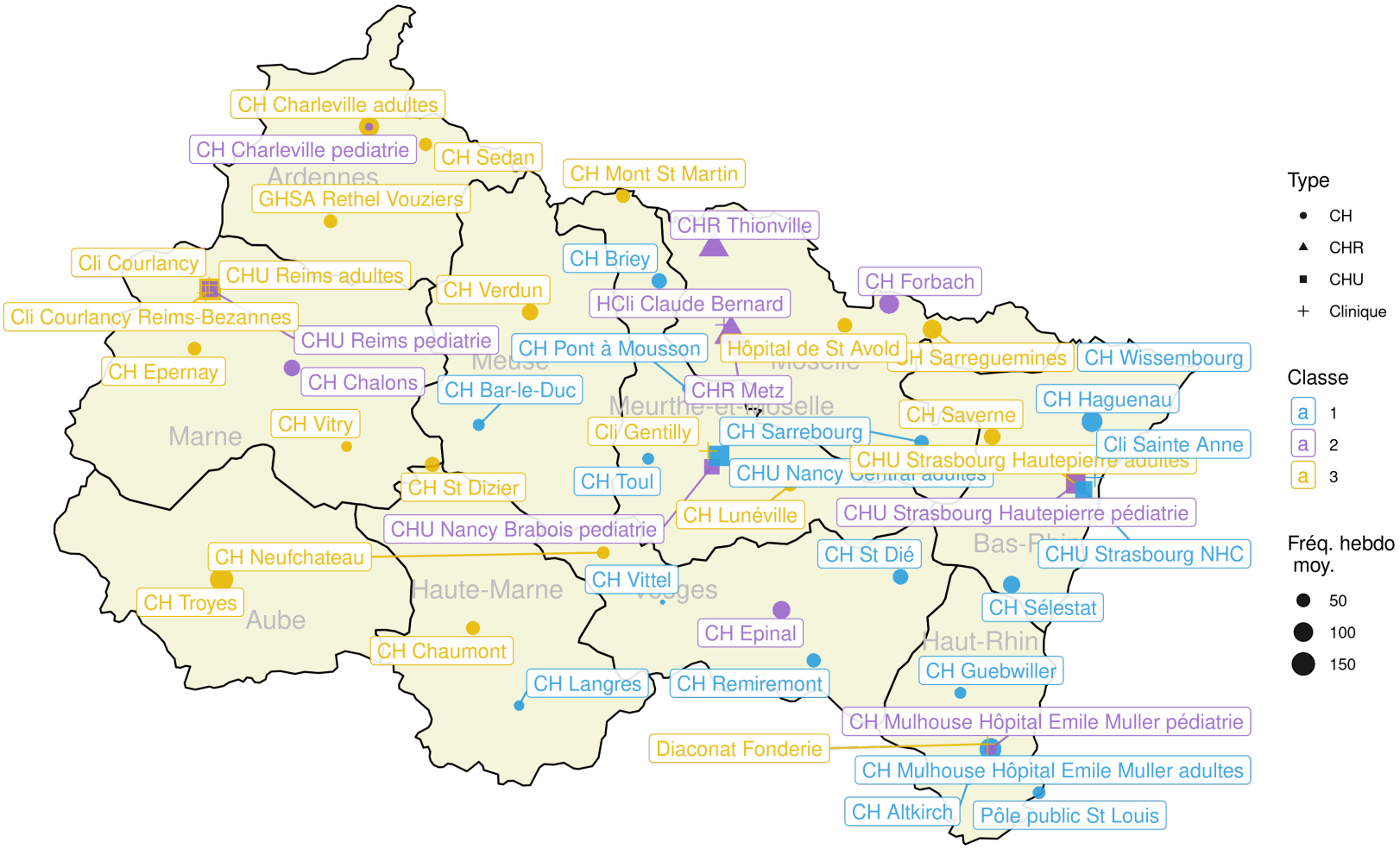

FIGURE 4.2 : Positions géographiques des SAUs. La couleur d'un point permet d'identifier la classe à laquelle est alloué le SAU correspondant. La taille du point est proportionnelle à la fréquentation hebdomadaire moyenne sur la période 2016-2018. La forme du point identifie le type d'établissement.

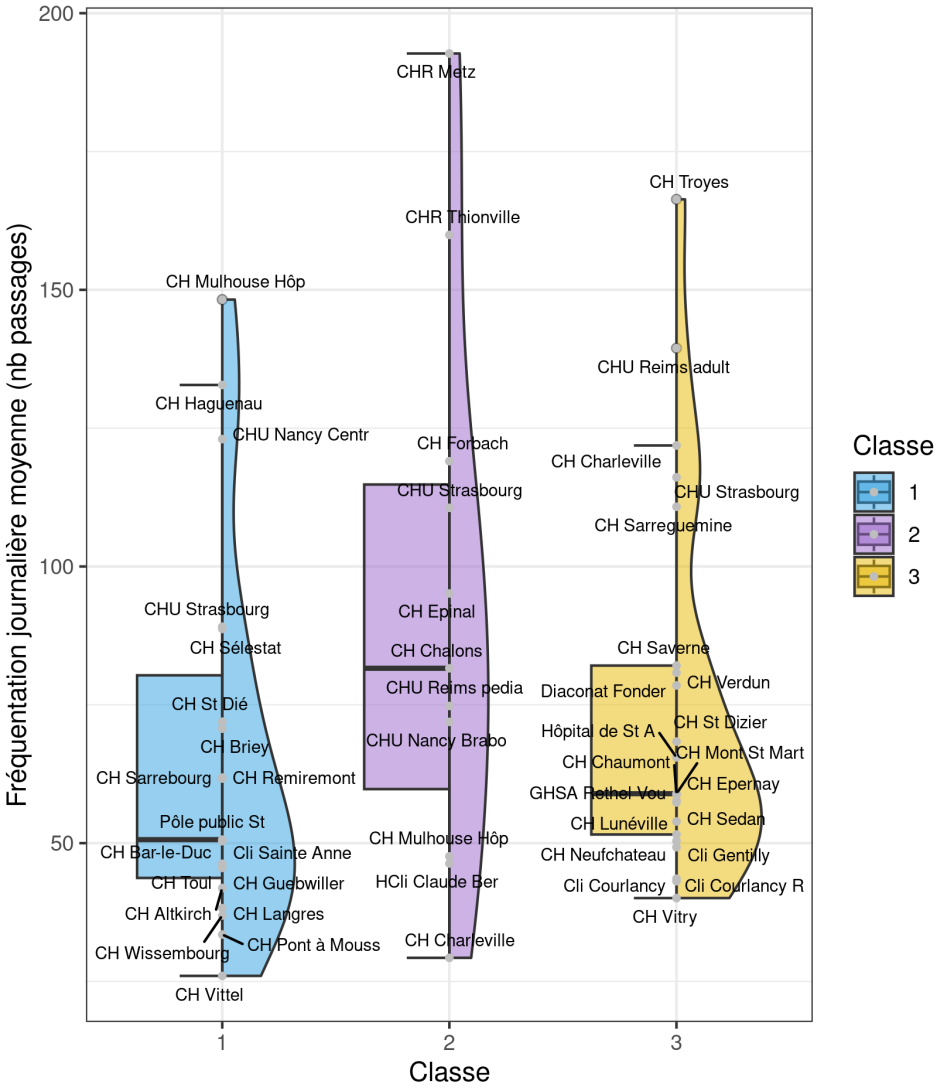

Figure 4.3 : Distributions des fréquentations quotidiennes moyennes des SAUs selon leur classe.

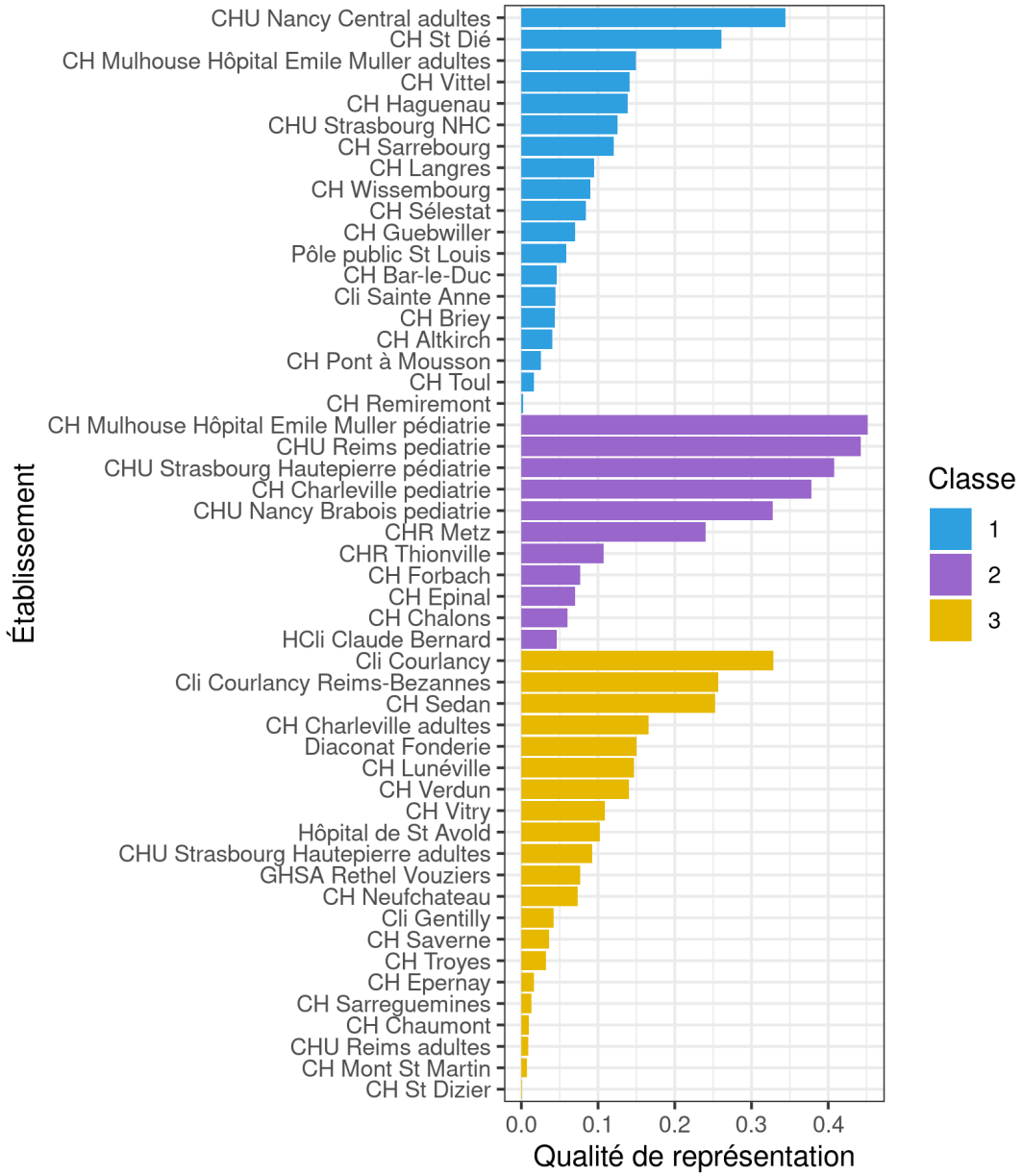

FIGURE 4.4 : Qualités de représentations des SAUs dans le plan principal associé à l'ACP des historiques de fréquentation centrés réduits.

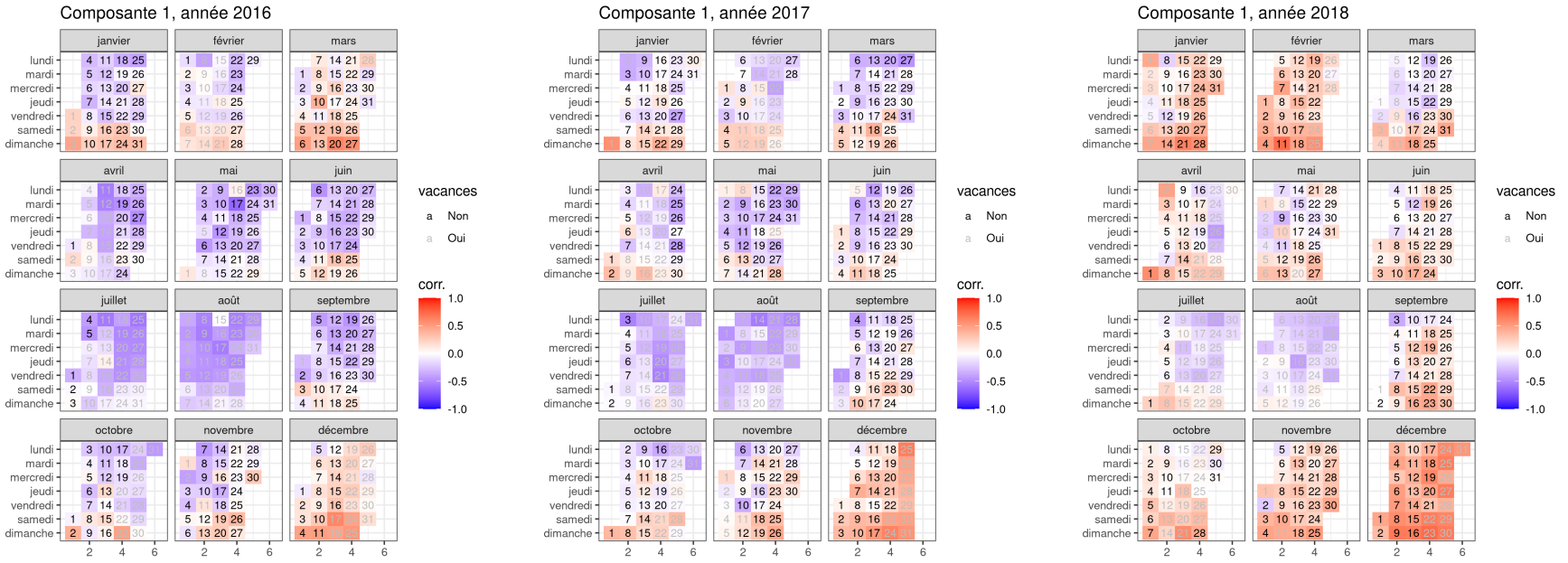

Figure 4.5 : Corrélations de la première composante principale avec les fréquentations quotidiennes pour les années 2016 à 2018.

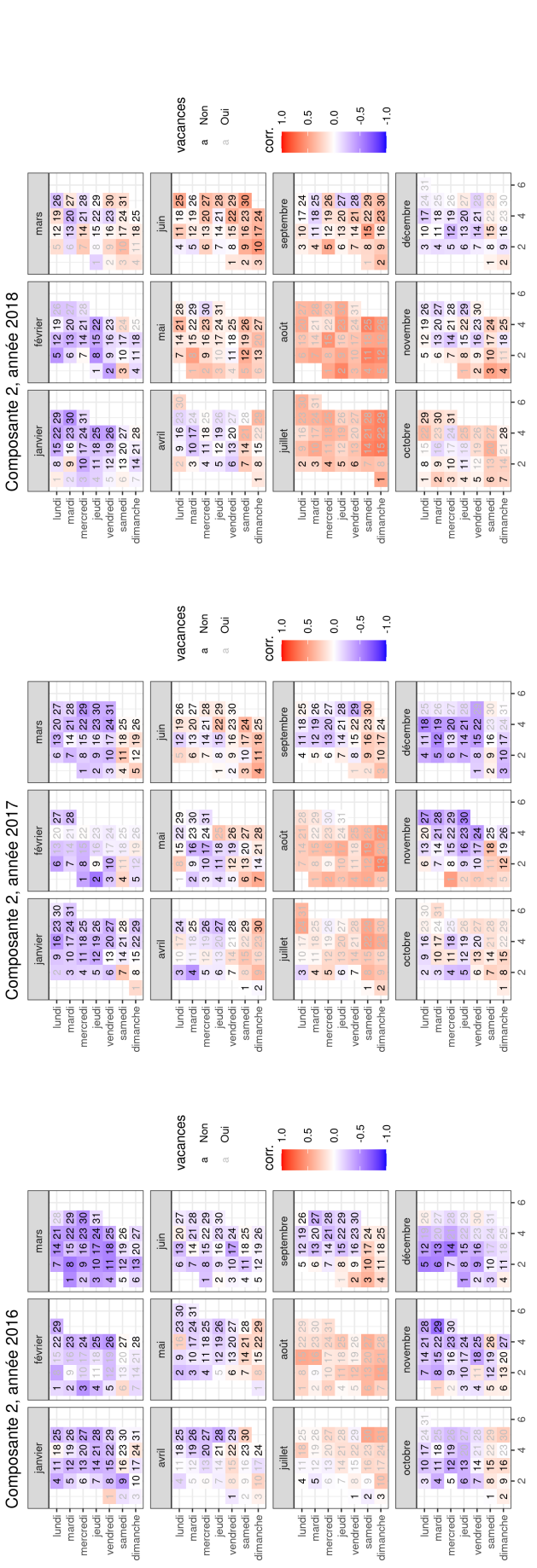

FIGURE 4.6 : Corrélations de la deuxième composante principale avec les fréquentations quotidiennes pour les années 2016 à 2018.

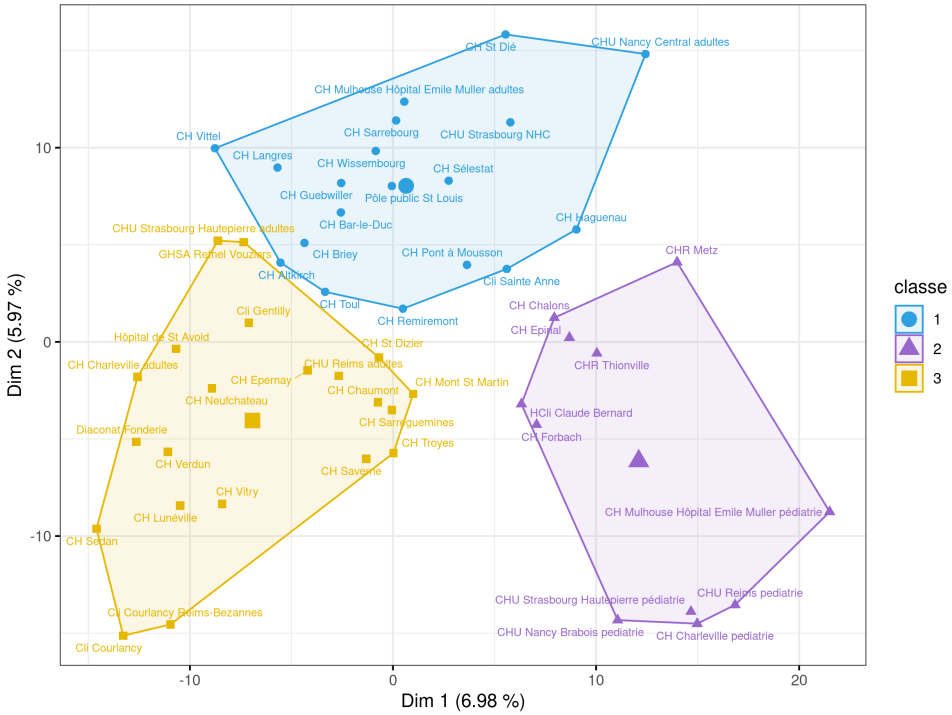

Figure 4.7 : Projections des cinquante et un SAUs sur le plan principal associé à l'ACP des historiques de fréquentation centrés et réduits. Les axes 1 et 2 représentent respectivement 6,98 % et 5,97 % de l'inertie. Les couleurs identifient les classes d'appartenances des SAUs.

Bibliographie

[kau, 1990] (1990). Partitioning Around Medoids (Program PAM). In *Finding Groups in Data*, pages 68–125. John Wiley & Sons, Ltd.

[Aboagye-Sarfo et al., 2015] Aboagye-Sarfo, P., Mai, Q., Sanfilippo, F. M., Preen, D. B., Stewart, L. M., and Fatovich, D. M. (2015). A comparison of multivariate and univariate time series approaches to modelling and forecasting emergency department demand in western australia. *Journal of biomedical informatics*, 57 :62–73.

[Afilal et al., 2016] Afilal, M., Yalaoui, F., Dugardin, F., Amodeo, L., Laplanche, D., and Blua, P. (2016). Forecasting the emergency department patients flow. *Journal of medical systems*, 40 :1–18.

[Ajmi et al., 2013] Ajmi, I., Zgaya, H., and Hammadi, S. (2013). Optimized workflow for the healthcare logistic : Case of the pediatric emergency department. pages 77–84.

[Alloux, 2021] Alloux, M. (2021). Modélisation et prédiction de la fréquentation des urgences pédiatriques dans le grand est. Master's thesis, Université de Reims Champagne-Ardenne, Master SEP 1ère année.

[Ammiratti et al., 2019] Ammiratti, C., Bernard, M., Bertrand, C., Bouisse, M., Chevallier, J.-R., De La Coussaye, J.-E., Di Palma, M., Fessler, J.-M., Meyer, G., Nivet, L., Pillet, C., and Regouby, C. (2019). L'établissemenent de santé en tension - Les éléments opérationnels de veille de l'activité quotidienne dans les établissements de santé et les seuils de déclenchement du plan blanc. Guide, Ministère de la santé et des solidarités.

[Arthur and Vassilvitskii, 2007] Arthur, D. and Vassilvitskii, S. (2007). k-means++ : The Advantages of Careful Seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics Philadelphia, PA, USA.

[Battaglia, 2022] Battaglia, M. (2022). How French emergency rooms are dealing with the stressful holiday season.

[Bergs et al., 2014] Bergs, J., Heerinckx, P., and Verelst, S. (2014). Knowing what to expect, forecasting monthly emergency department visits : A time-series analysis. *International emergency nursing*, 22(2) :112–115.

[Boyle et al., 2012] Boyle, A., Beniuk, K., Higginson, I., and Atkinson, P. (2012). *Emergency medicine international*, 2012.

[Boyle et al., 2008] Boyle, J., Wallis, M., Jessup, M., Crilly, J., Lind, J., Miller, P., and Fitzgerald, G. (2008). Regression forecasting of patient admission data. pages 3819–3822.

[Breusch and Pagan, 1979] Breusch, T. S. and Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica : Journal of the econometric society*, pages 1287–1294.

[Brockwell and Davis, 1991] Brockwell, P. J. and Davis, R. A. (1991). *Time series : theory and methods*. Springer Series in Statistics. Springer-Verlag, New York, second edition.

[Brockwell and Davis, 2016] Brockwell, P. J. and Davis, R. A. (2016). *Introduction to time series and forecasting*. Springer Texts in Statistics. Springer, [Cham], third edition.

[Champion et al., 2007] Champion, R., Kinsman, L. D., Lee, G. A., Masman, K. A., May, E. A., Mills, T. M., Taylor, M. D., Thomas, P. R., and Williams, R. J. (2007). Forecasting emergency department presentations. *Australian Health Review*, 31(1) :83–90.

[Cooke et al., 2004a] Cooke, M., Wilson, S., Halsall, J., and Roalfe, A. (2004a). Total time in english accident and emergency departments is related to bed occupancy. *Emergency Medicine Journal*, 21(5) :575–576.

[Cooke et al., 2004b] Cooke, M., Wilson, S., Halsall, J., and Roalfe, A. (2004b). Total time in english accident and emergency departments is related to bed occupancy. *Emergency Medicine Journal*, 21(5) :575–576.

[Coskun and Erol, 2010] Coskun, N. and Erol, R. (2010). An optimization model for locating and sizing emergency medical service stations. *Journal of medical systems*, 34 :43–49.

[Dagnet, 2023] Dagnet, A.-L. (2023). Hôpital : près d'un service d'urgences sur deux a fermé au moins une fois cet été, selon une enquête de Samu-Urgences de France.

[De Gooijer and Hyndman, 2006] De Gooijer, J. G. and Hyndman, R. J. (2006). 25 years of time series forecasting. *International journal of forecasting*, 22(3) :443–473.

[de l'Hospitalisatoin Privée MCO, 2023] de l'Hospitalisatoin Privée MCO, F. (2023). Financement de la médecine d'urgence.

[Devie, 2020] Devie, B. (2020). Caractérisation de la patientèle du service des urgences de châlons-en-champagne. Master's thesis, Université de Reims Champagne-Ardenne, Master SEP 1ère année.

[Durbin and Watson, 1971] Durbin, J. and Watson, G. S. (1971). Testing for serial correlation in least squares regression. iii. *Biometrika*, 58(1) :1–19.

[Earnest et al., 2005] Earnest, A., Chen, M. I., Ng, D., and Sin, L. Y. (2005). Using autoregressive integrated moving average (arima) models to predict and monitor the number of beds occupied during a sars outbreak in a tertiary hospital in singapore. *BMC Health Services Research*, 5 :1–8.

[Est-Rescue, a] Est-Rescue. Panorama 2016.

[Est-Rescue, b] Est-Rescue. Panorama 2020.

[Est-RESCUE, 2025] Est-RESCUE (2025). Hôpital en Tension.

[Foley et al., 2011] Foley, M., Kifaieh, N., and Mallon, W. K. (2011). Financial impact of emergency department crowding. *Western Journal of Emergency Medicine*, 12(2) :192.

[Frank and Friedman, 1993] Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2) :109–135.

[Gul and Celik, 2020] Gul, M. and Celik, E. (2020). An exhaustive review and analysis on applications of statistical forecasting in hospital emergency departments. *Health Systems*, 9(4) :263–284.

[Headrick and Morgan, 1988] Headrick, R. W. and Morgan, G. W. (1988). Resource allocation in multifacility emergency medical service systems. *Journal of Medical Systems*, 12 :121–128.

[Helland, 1990] Helland, I. S. (1990). Partial least squares regression and statistical models. *Scandinavian journal of statistics*, pages 97–114.

[Hing, 2012] Hing, E. (2012). *Wait time for treatment in hospital emergency departments, 2009*. Number 102. US Department of Health and Human Services, Centers for Disease Control and ….

[Hsu and Nemhauser, 1979] Hsu, W.-L. and Nemhauser, G. L. (1979). Easy and hard bottleneck location problems. *Discrete Applied Mathematics*, 1(3) :209–215.

[Hyndman and Khandakar, 2008] Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting : the forecast package for R. *Journal of Statistical Software*, 26(3) :1–22.

[Jones et al., 2002] Jones, S. A., Joy, M. P., and Pearson, J. (2002). Forecasting demand of emergency care. *Health care management science*, 5 :297–305.

[Jones et al., 2009] Jones, S. S., Evans, R. S., Allen, T. L., Thomas, A., Haug, P. J., Welch, S. J., and Snow, G. L. (2009). A multivariate time series approach to modeling and forecasting demand in the emergency department. *Journal of biomedical informatics*, 42(1) :123–139.

[Kadri et al., 2014] Kadri, F., Chaabane, S., Harrou, F., and Tahon, C. (2014). Modélisation et prévision des flux quotidiens des patients aux urgences hospitalières en utilisant l'analyse de séries chronologiques. In *7ème conférence de Gestion et Ingénierie des Systèmes Hospitaliers (GISEH)*, pages 1–8.

[Kwak and Lee, 1997] Kwak, N. and Lee, C. (1997). A linear goal programming model for human resource allocation in a health-care organization. *Journal of Medical Systems*, 21 :129–140.

[Liland et al., 2024] Liland, K. H., Mevik, B.-H., and Wehrens, R. (2024). *pls : Partial Least Squares and Principal Component Regression*. R package version 2.8-5.

[Lin et al., 2008] Lin, B. Y.-J., Hsu, C.-P. C., Chao, M.-C., Luh, S.-P., Hung, S.-W., and Breen, G.-M. (2008). Physician and nurse job climates in hospital-based emergency departments in taiwan : management and implications. *Journal of medical systems*, 32 :269–281.

[Ljung and Box, 1978] Ljung, G. M. and Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2) :297–303.

[Lloyd, 1982] Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2) :129–137.

[Luo et al., 2016] Luo, L., Luo, Y., You, Y., Cheng, Y., Shi, Y., and Gong, R. (2016). A mip model for rolling horizon surgery scheduling. *Journal of medical systems*, 40 :1–7.

[MacQueen, 1967] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1 : Statistics*, volume 5.1, pages 281–298. University of California Press.

[Martens and Naes, 1984] Martens, H. and Naes, T. (1984). Multivariate calibration. In *Chemometrics : mathematics and statistics in chemistry*, pages 147–156. Springer.

[Martens and Næs, 1992] Martens, H. and Næs, T. (1992). *Multivariate calibration*. John Wiley & Sons.

[Mielczarek, 2013] Mielczarek, B. (2013). Estimating future demand for hospital emergency services at the regional level. pages 2386–2397.

[Milner, 1988] Milner, P. (1988). Forecasting the demand on accident and emergency departments in health districts in the trent region. *Statistics in medicine*, 7(10) :1061–1072.

[Pealat et al., 2019] Pealat, C., Bouleux, G., and Cheutet, V. (2019). Extracting Most Impacting Emergency Department Patient Flow By Embedding Laboratory-confirmed and Clinical Diagnosis on The Stiefel Manifold. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 1–4. ISSN : 2641-3604.

[Pulsy, 2024] Pulsy (2024). Dispositif HET.

[Quentin and Vildy, 2024] Quentin, O. and Vildy, N. (2024). "C'est ingérable !", les urgences de l'hôpital du Mans en grève illimitée.

[R Core Team, 2024] R Core Team (2024). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Rafferty, 2023] Rafferty, G. (2023). *Forecasting Time Series Data with Prophet : Build, improve, and optimize time series forecasting models using Meta's advanced forecasting tool*. Packt Publishing Ltd.

[Rocha and Rodrigues, 2021] Rocha, C. N. and Rodrigues, F. (2021). Forecasting emergency department admissions. *Intelligent Data Analysis*, 25(6) :1579–1601.

[Rousseeuw, 1987] Rousseeuw, P. J. (1987). Silhouettes : A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20 :53–65.

[Sagno, 2020] Sagno, I. (2020). Modélisation des flux de patients au service d'urgences du chu de reims. Master's thesis, Université de Reims Champagne-Ardenne, Master SEP 1ère année.

[Sakoe and Chiba, 1978] Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1) :43–49.

[sante.gouv, 2019] sante.gouv (2019). Pacte de refondation des ugences.

[Schull et al., 2005] Schull, M. J., Mamdani, M. M., and Fang, J. (2005). Influenza and emergency department utilization by elders. *Academic emergency medicine*, 12(4) :338–344.

[Shapiro and Wilk, 1965] Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4) :591–611.

[Shi et al., 2011] Shi, H.-Y., Tsai, J.-T., Ho, W.-H., and Lee, K.-T. (2011). Autoregressive integrated moving average model for long-term prediction of emergency department revenue and visitor volume. 3 :979–982.

[Sun et al., 2013] Sun, B. C., Hsia, R. Y., Weiss, R. E., Zingmond, D., Liang, L.-J., Han, W., McCreath, H., and Asch, S. M. (2013). Effect of emergency department crowding on outcomes of admitted patients. *Annals of emergency medicine*, 61(6) :605–611.

[Sun et al., 2009] Sun, Y., Heng, B. H., Seow, Y. T., and Seow, E. (2009). Forecasting daily attendances at an emergency department to aid resource planning. *BMC emergency medicine*, 9 :1–9.

[Sutherland et al., 2023] Sutherland, C., Smallwood, A., Wootten, T., and Redfern, N. (2023). Fatigue and its impact on performance and health. *British Journal of Hospital Medicine*, 84(2) :1–8.

[Tavenard, 2021] Tavenard, R. (2021). An introduction to dynamic time warping. https://rtavenar.github.io/blog/dtw.html.

[Taylor and Letham, 2021] Taylor, S. and Letham, B. (2021). *prophet : Automatic Forecasting Procedure*. R package version 1.0.

[Taylor and Letham, 2018] Taylor, S. J. and Letham, B. (2018). Forecasting at scale. *Amer. Statist.*, 72(1) :37–45.

[Tenenhaus, 1998] Tenenhaus, M. (1998). *La régression PLS : théorie et pratique*. Editions technip.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 58(1) :267–288.

[Trzeciak and Rivers, 2003] Trzeciak, S. and Rivers, E. P. (2003). Emergency department overcrowding in the united states : an emerging threat to patient safety and public health. *Emergency medicine journal*, 20(5) :402–405.

[Ward, 1963] Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301) :236–244.

[Wartelle et al., 2022] Wartelle, A., Mourad-Chehade, F., Yalaoui, F., Questiaux, H., Monneret, T., Soliveau, G., Chrusciel, J., Duclos, A., Laplanche, D., and Sanchez, S. (2022). Multimorbidity clustering of the emergency department patient flow : Impact analysis of new unscheduled care clinics. *PLOS ONE*, 17(1) :e0262914.

[Williamson, 2023] Williamson, L. (2023). France's health system under pressure of increasing demands.

[Wold, 1966] Wold, H. (1966). Estimation of principal components and related models by iterative least squares. *Multivariate analysis*, pages 391–420.

Modélisation Prédictive des Flux de Patients dans les Services Accueil d'Urgences du Grand Est

Les tensions hospitalières dans les Services Accueil d'Urgences (SAUs) ont augmenté ces dernières années et sont largement médiatisées. Des facteurs organisationnels, structurels ou conjoncturels peuvent expliquer cette situation, qui doivent être pris en compte pour anticiper les flux de patients. Nous proposons dans cette thèse un modèle additif généralisé des flux journaliers intégrant des informations exogènes (telles que les périodes de vacances scolaires) dans lequel la tendance est une fonction polynomiale et la saisonnalité est une combinaison d'harmoniques. Ce modèle est d'abord ajusté individuellement pour chaque SAU de la région Grand Est à partir des historiques quotidiens de fréquentation fournis par l'association Est-RESCUE. Une extension multivariée du modèle est ensuite proposée, combinant des techniques de clustering et de régression PLS. Les performances de prédiction sont évaluées et comparées aux modèles classiques SARIMA, qui sont la référence couramment utilisée dans la littérature.

Mots-clés : Flux de patients aux services accueil d'urgences, modèle SARIMA, modèle additif généralisé, séries temporelles, modélisation prédictive multivarié

Modeling and forecasting patient flows in emergency departments

Hospital tensions in the French emergency departments (ED) have increased last years and are widelypublicized. Organizational, structural or cyclical factors can explain this situation ; they must be taken into account to anticipate of the flows of patients in the EDs. In this PhD thesis, we propose a generalized additive model of daily flows incorporating exogenous information (such as school holiday periods) in which trend is a polynomial function and seasonality is a combination of harmonics. This model is first adjusted individually for each ED of the French Grand Est area from the daily records of frequentation provided by the association Est-RESCUE. A multivariate version of the model is then proposed, combining clustering and PLS regression technique. The prediction performances are evaluated and compared to the classical SARIMA models, which are the reference commonly used in the literature.

Keywords : Emergency departments' patients flow, SARIMA model, generalized additive model, time-series, multivariate forecasting modeling

Discipline : Mathématiques appliquées.
Spécialité : Statistique

Université de Reims Champagne-Ardenne
LMR - UMR 9008