# EA MAP471

Prédiction des flux d'arrivées dans les services d'urgences d'Île-de-France

Eloïse Berthier, Guillaume Dalle
Encadrés par Emmanuel Bacry, Youcef Sebiat

14 décembre 2017

# Sommaire

# Introduction

La fréquentation des services médicaux d'urgences est une donnée-clef pour comprendre l'efficacité de prise en charge des patients et la qualité du travail médical.

Dans le modèle conceptuel proposé par [Asplin et al., 2003] pour analyser la surpopulation aux urgences, la demande en soins urgents fait partie du domaine de l'*input*, mais influe également sur le déroulement du *throughput* (séjour du patient dans le service) ainsi que sur l'*output* (sortie du patient et suivi postérieur).

Un nombre excessif de patients peut entraîner des conséquences néfastes ([Hoot and Aronsky, 2008], [Bernstein et al., 2009]), parmi lesquelles :

— une qualité de traitement amoindrie
— un taux de mortalité augmenté
— des insuffisances matérielles (disponibilité des lits)
— des délais de transport allongés (reroutage d'ambulances)
— des délais de prise en charge allongés (disponibilité des médecins et infirmiers)
— des fugues de patients
— des pertes financières pour l'hôpital

Pour soulager la surcharge et résoudre en partie ces problèmes, une des voies explorées est celle de la prédiction des arrivées.

En effet, une meilleure compréhension de la demande en soins et des facteurs qui l'influencent peut permettre de détecter plus facilement les périodes de tension, et d'adapter la rotation du personnel. Un outil de prévision des arrivées à court ou moyen terme peut, quant à lui, permettre une optimisation de l'usage des ressources physiques, et une meilleure organisation du transport.

Le Centre de Mathématiques Appliquées de l'École polytechnique a été retenu en collaboration avec l'Unité de Recherche Clinique Paris IDF Ouest pour un appel d'offre PREPS, afin de répondre à cette problématique d'analyse et de prédiction des flux de patients. Le présent rapport s'inscrit dans ce projet, son but est de présenter les résultats obtenus dans le cadre d'un Enseignement d'Approfondissement de 3ème année.

# 1 Données

## 1.1 Arrivées de patients

Les données dont nous disposons proviennent de l'Agence Régionale de Santé (ARS) d'Île-de-France. Elles présentent l'ensemble des saisies correspondant à l'entrée d'un patient dans un service d'urgences d'Île-de-France, pour les années 2014 et 2015.

Afin de respecter la confidentialité nécessaire au traitement de données à caractère personnel, certaines données sont bruitées. Par exemple, le nom du centre hospitalier n'est pas renseigné, ce qui empêche toute analyse géographique.

### 1.1.1 Contenu

Pour chaque patient enregistré lors de son arrivée dans un service, les données fournissent :
- des informations temporelles : les dates et heures d'entrée et de sortie ;
- des informations personnelles : sexe, code postal de résidence, âge du patient ;
- des informations circonstancielles : mode de transport, code de circonstance ;
- des informations liées à la prise en charge : motif du recours, classification de gravité, orientation.

Notre objectif est de modéliser et de prédire la charge à laquelle seront confrontés les services d'urgences. Compte-tenu des données dont nous disposons, nous ne pouvons réaliser cette analyse que sur une échelle agrégée (c'est-à-dire tous les centres d'Île-de-France). Nous prendront comme unité de mesure le nombre de patients qui arrivent aux urgences dans un intervalle de temps donné.

### 1.1.2 Qualité des données

Le rapport fourni par l'ARS et une première explorations des données nous permettent de dresser plusieurs constats quant à la qualité des données dont nous disposons.

*Constat 1* : Une partie des données de l'année 2015 sont manquantes : la lecture du fichier csv s'arrête avant d'atteindre le nombre de lignes donné dans le rapport récapitulatif.

*Constat 2* : Les années 2014 et 2015 ont des profils très différents. L'année 2015 présente de grandes variations temporelles par rapport à 2014.

*Constat 3* : Parmi les données sus-citées, seules certaines sont saisies avec une qualité suffisante. Celles pour lesquelles le champ est renseigné dans plus de 80% des cas sont : la date et heure d'entrée, la date et heure de sortie, le sexe, le code postal de résidence, le mode de sortie, le mode de transport et le code de gravité. De plus, même parmi les données fortement renseignées, les modalités de saisie semblent avoir été modifiées au cours de l'année. C'est le cas par exemple pour le code gravité et le code de circonstance (voir figure 1).

*Constat 4* : Certains centres ont commencé à saisir des données au milieu de l'intervalle d'étude, certains arrêtent la saisie au cours de l'année, et d'autres ne saisissent pas pendant plusieurs mois. Ces irrégularités dans les différents centres font que la quantité agrégée de patients est très fortement biaisée par le début ou l'arrêt de saisie de plusieurs centres chaque mois. Ces effets ne se compensent pas entre eux car le nombre global de centres qui saisissent augmente au cours de l'année 2014.

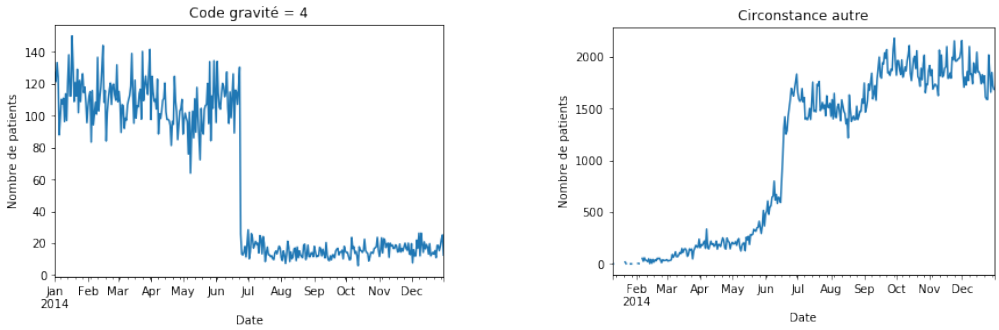


FIGURE 1 – Nombre de saisies du code de gravité "4" et du code de circonstance "autre" en 2014 (après filtrage des centres).

### 1.1.3 Nettoyage des données

En tenant compte de ces observations, nous avons donc décidé d'une part de ne travailler que sur l'année 2014, et d'autre part de ne pas considérer les données de gravité ou de circonstances pour nos analyses.

De plus, pour résoudre le problème des différences de remplissage entre centres, nous avons procédé à un filtrage sur les centres. L'idée est de retirer les centres que l'on considère comme non fiables sur l'année.

Ce filtrage ne peut pas être effectué directement, car nous ne pouvons pas attribuer une entrée de patient à un centre. Le rapport de l'ARS fournit néanmoins un tableau présentant le nombre d'entrées saisies par centre et par mois. Chaque mois, nous pouvons donc supprimer l'influence des centres non fiables sur le nombre agrégé mensuel de patients en soustrayant le nombre mensuel d'entrées saisies par ces centres.

Concrètement, nous calculons 12 ratios (un par mois), égaux à la proportion mensuelle d'entrées saisies par des centres valides. Puis nous multiplions le nombre total d'entrées sur un intervalle de temps (3h, jour, semaine...) par le ratio du mois correspondant.

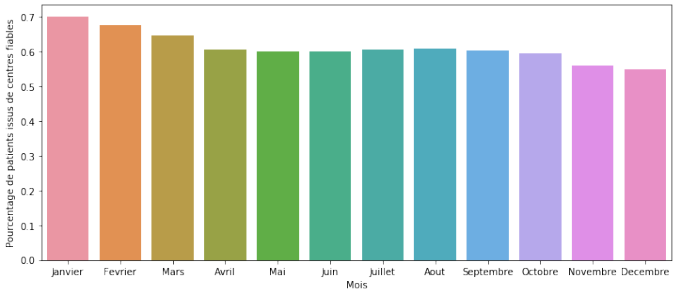


FIGURE 2 – Filtrage des données pour l'année 2014

Un centre est jugé non fiable si un des mois comporte plus de 5 jours non renseignés, ou si la différence relative entre l'année 2014 et l'année 2015 dépasse 30%. Nous avions tenté d'introduire un critère de cohérence entre les mois de l'année, pour exclure les centres dont les données

patients mois par mois seraient trop disparates mais toujours non nulles. Cependant, un filtrage de ce type peut agir de deux façons :

— Retirer les centres qui rentrent leurs données de façon incohérente au cours de l'année
— Retirer les centres qui ont bel et bien des entrées de patients "incohérentes" au cours de l'année

Par exemple, en 2014, le service de pédiatrie de Rambouillet a enregistré 1099 patients en décembre, contre seulement 413 en août. Cela semble une différence conséquente, mais comment savoir alors si cela relève de l'erreur humaine (personnels différents certains mois, peut-être moins habitués au logiciel), ou bien d'un réel phénomène sur la demande en soins (typiquement, l'absence de nombreux enfants pendant les vacances d'été), ou encore d'un effet dû au fonctionnement du service (moins de personnels aussi pendant l'été). Dans le doute, nous avons préféré renoncer à ce critère de "lissage".

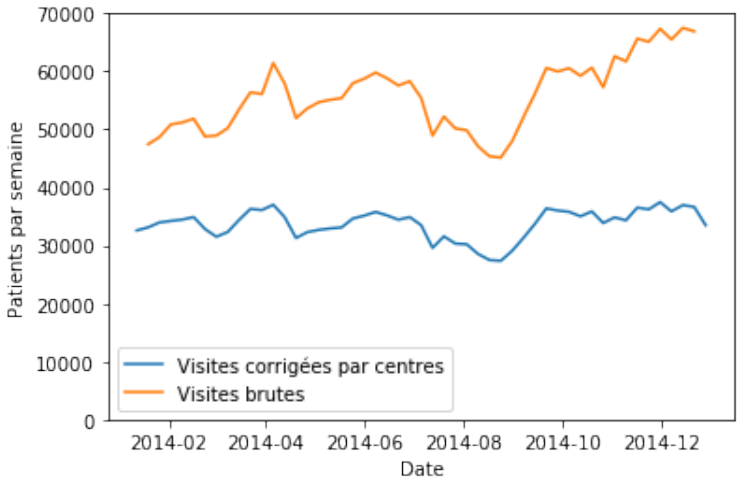


FIGURE 3 – Nombre de patients par semaine, avant et après filtrage des centres

Par ce filtrage, nous obtenons des ratios décroissants dans l'année, de 70% à 55%, ce qui signifie que nous ne gardons en moyenne qu'environ 60% des entrées renseignées (voir figure 2). Ce filtrage nous a donc permis de supprimer l'influence de l'arrivée progressive de nouveaux centres. En effet, il a annulé la tendance croissante qui existait dans les données originales (voir figure 3).

### 1.1.4 Régularités temporelles

Nous commençons notre étude par une exploration visuelle des données, afin de repérer d'éventuelles périodicités.

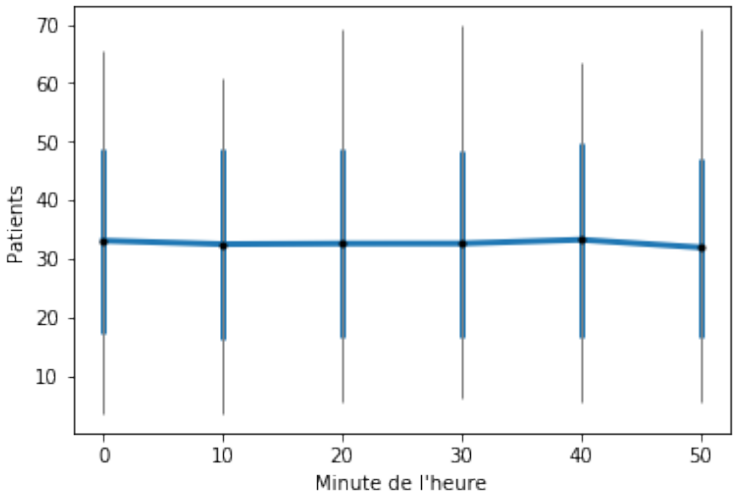


FIGURE 4 – Nombre moyen de patients par tranche de 10 min

Nous vérifions dans un premier temps que le remplissage au sein d'une heure n'est pas asymétrique. En effet, cela pourrait arriver si des changements d'équipes aux heures pleines reportaient ou avançaient la saisie des données.

Pour contrôler cet effet, nous effectuons un test ANOVA testant l'hypothèse $H_0$ d'égalité des moyennes des échantillons. Ce test requiert de faire les hypothèses raisonnables d'indépendance, de normalité et d'homoscédasticité des échantillons. Ce test renvoie une p-valeur supérieure à 0,99. Ainsi, nous ne pouvons pas rejeter $H_0$ et nous ne détectons aucun effet significatif.

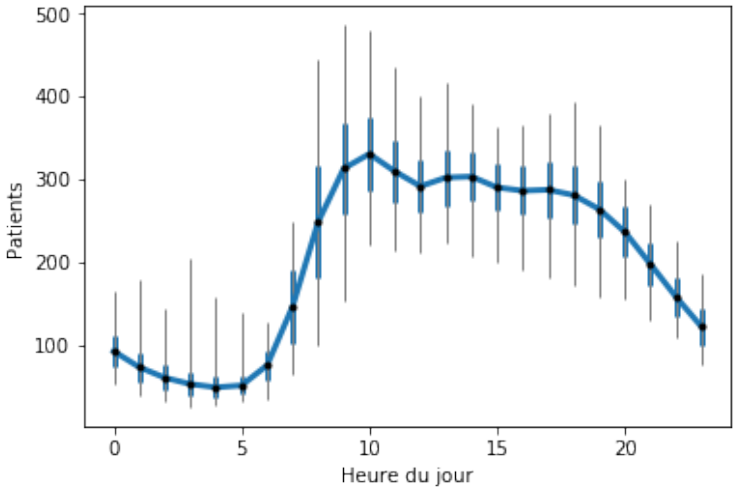


FIGURE 5 – Nombre moyen de patients par tranche d'une heure

A l'échelle d'une journée, l'évolution des entrées est assez régulière sur l'année, avec des différences significatives selon l'heure de la journée. Cette fois, nous ne pouvons pas utiliser le test

ANOVA car les variances ne sont pas égales entre les différents groupes (voir les écarts-types en bleu sur la figure 5). Nous utilisons donc le test non paramétrique de Kruskal-Wallis, qui fait comme seule hypothèse l'indépendance des échantillons. Le test pour les 24 échantillons renvoie une p-valeur nulle. Ainsi, il y a bien des variations du nombre de patients entrés au cours de la journée (ce qui est très intuitif).

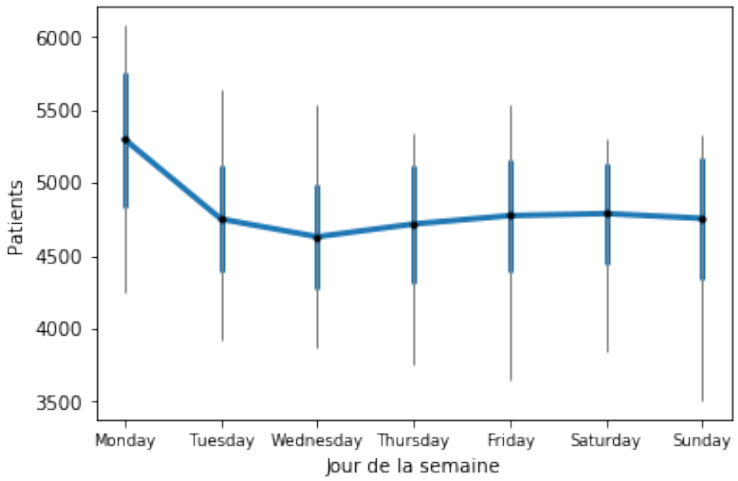


FIGURE 6 – Nombre moyen de patients par jour

Enfin, à l'échelle d'une semaine, le même test renvoie une p-valeur nulle pour l'égalité des 7 moyennes journalières. Plus précisément, testons l'hypothèse $H_0$ d'égalité des moyennes entre deux jours de la semaine. Le tableau suivant présente les p-valeurs de ces tests. Nous voyons donc que seuls le lundi et le mercredi sont significativement différents des autres jours.

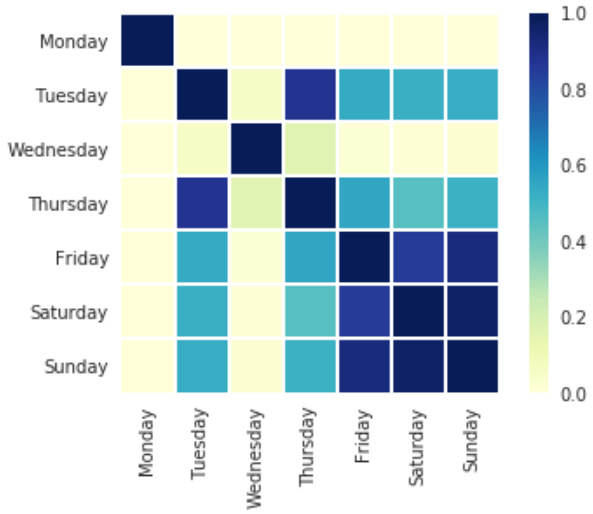


FIGURE 7 – p-valeurs du test de Kruskal-Wallis pour l'hypothèse $H_0$ d'égalité des moyennes sur l'année entre deux jours de la semaine.

## 1.2 Données exogènes

Si les données endogènes d'arrivée des patients peuvent être suffisantes pour construire un modèle prédictif intéressant, considérer des données exogènes peut apporter une plus-value à la fois sur la qualité de la prédiction et sur son interprétation [Wargon et al., 2009]. Un travail conséquent a donc été mené pour rassembler des données de sources diverses pouvant influer sur la demande en soins.

Pour effectuer la prédiction, toutes les variables exogènes non-binaires ont été centrées et réduites. De plus, pour les variables non calendaires, nous avons décidé d'introduire des décalages dans le temps : pour prédire le nombre de patients au temps $t$ (mesuré en tranches), nous pouvions utiliser la valeur de la variable exogène au temps $t$, mais aussi $t-1$, $t-2$, etc. Nous avons aussi utilisé les moyennes sur les 24h précédentes comme variable d'intérêt.

### 1.2.1 Calendrier

Le premier type de données exogènes est celui que l'on trouve dans le calendrier. En effet, comme évoqué plus haut, on observe de fortes variations selon les jours, qu'il a fallu tenter d'expliquer.

On a donc récolté les données suivantes :

— Mois, jours de la semaine : Ces données ont été générées grâce aux fonctionnalités de la librairie Python `pandas` [1]

— Vacances scolaires : Les vacances scolaires pour la région parisienne ont été récupérées sur le site d'Index Education [2], qui les utilise en interne pour ses logiciels pédagogiques

1. https ://pandas.pydata.org
2. http ://www.index-education.com/fr/

— Jours fériés : Les jours fériés en France ont été récoltés grâce aux fonctionnalités de la librairie Python `workalendar` [3]

Ces données ont été traduites comme des variables binaires : une pour chaque jour de la semaine et chaque mois de l'année, une variable indiquant les week-ends, une indiquant les vacances scolaires, et finalement des variables indiquant les jours pré- ou post-vacances et pré- ou post-jours fériés.

### 1.2.2 Météo

Un autre type de données intéressant est celui des variables météorologiques. En effet, on pourrait penser que le cycle des saisons influe sur la fréquentation des urgences, ou encore que les patients sont plus enclins à y aller les jours de pluie.

Les relevés météo ont été téléchargés sur le site des données publiques de Météo France [4], pour la station de mesure d'Orly. Les champs conservés sont : température, humidité, nébulosité, pression atmosphérique, vitesse du vent.

D'autre part, vu la fréquence croissante des pics de pollution subis par l'agglomération parisienne, une prise en compte de la qualité de l'air semblait intéressante. Les valeurs de pollution pour différents polluants (CO, NO2, O3, particules fines de diamètres $< 10\ \mu m$ et $< 2.5\ \mu m$) ont donc été récoltées sur le site d'AirParif [5]

### 1.2.3 Activités humaines

Les passages aux urgences sont intimement reliés à la notion globale d'activité de la population. Afin d'estimer cette variable mal définie, nous avons fait des recherches sur le site Open Data Paris [6].

Une base de données d'évènements de la vie parisienne était proposée (qui aurait pu permettre d'étudier l'effet "lendemain de soirée"), mais de par son remplissage lacunaire elle nous a semblé insuffisante.

En revanche, les données de tous les horodateurs de l'agglomération sur l'année 2014 étaient disponibles, et nous avons choisi de les exploiter en tant que mesure, certes imparfaite, de la circulation automobile et plus généralement du rythme de vie journalier des parisiens. Le seul problème avec ces données se rencontre au mois d'août, durant lequel les parcmètres étaient gratuits cette année-là. Nous avons donc reproduit, pour combler ce vide, les valeurs de la dernière semaine de juillet.

### 1.2.4 Santé publique

Un facteur évident d'influence des visites aux urgences se trouve dans les phénomènes sanitaires globaux tels que les épidémies. Le site d'open data du gouvernement [7] propose des relevés pour mesurer l'acuité des épidémies de grippe et de gastro-entérite sur le territoire.

Pour ces variables, il nous a semblé que le manque possible de centralisation des données et la difficulté d'accès à ces informations rendait le suivi en temps réel et la prédiction difficiles.

3. https ://pypi.python.org/pypi/workalendar/0.1
4. https ://donneespubliques.meteofrance.fr
5. http ://airparif.asso.fr
6. https ://opendata.paris.fr/page/home/
7. https ://www.data.gouv.fr/fr/

C'est pourquoi nous les avons remplacées à chaque instant par leur moyenne glissante sur la semaine précédente, afin de simuler la difficulté de mesure immédiate.

# 2 Prédiction

## 2.1 Scénario

Au delà de la simple exploration des données, notre objectif est de modéliser et de prévoir l'afflux de patients aux urgences. Pour cela, nous considérons un scénario plausible pour un système d'information qui pourrait être mis en production dans un hôpital.

Le but de ce système serait de prédire pendant la nuit la courbe d'affluence du lendemain. Pour cela, l'utilisateur disposerait de l'historique des entrées passées (via la base de données actuelle), ainsi que des données exogènes listées plus haut. Remarque importante : nous avons supposé que ces données exogènes étaient également disponibles tranche par tranche tout au long de la journée à venir, autrement dit qu'elles pouvaient être elles-mêmes prédites afin de baser là-dessus la prédiction du nombre de patients. Cette hypothèse semble évidente pour les données calendaires (on sait à coup sûr si demain sera un mardi), raisonnable dans le cadre des prévisions de type météo, un peu moins raisonnable sur les données comme les transactions de parcmètres, et franchement invraisemblable sur les épidémies (c'est pourquoi dans ce dernier cas nous avons lissé le signal, voir ci-dessus).

Après le nettoyage décrit plus haut, nous agrégeons les entrées sur tous les centres, au sein d'un même intervalle de temps. Pour ce faire, nous avons prmis un échantillonnage de nos données selon des intervalles de largeur comprise entre 2h et 6h. En effet, il s'agit d'un compromis entre un intervalle petit qui contraint les capacités de prédiction, et un intervalle grand qui empêche de détecter des variations dans une journée. Pour nos tests, nous avons choisi 3h.

Cette échelle temporelle du scénario est guidée à la fois par l'utilité pratique d'un tel système, et par les limitations intrinsèques des modèles que nous utiliserons. Par exemple, un modèle de type ARMA avec un ordre raisonnable pourra prédire sur une journée mais difficilement sur une semaine.

Enfin, nous jugerons nos modèles selon deux grands critères :
— la précision, c'est-à-dire sa capacité à prédire fidèlement le nombre de patients ;
— la transparence, c'est-à-dire la simplicité et l'intelligibilité des méthodes utilisées.

## 2.2 Définition des modèles

Dans tous les modèles, on appellera "données endogènes" la suite des valeurs du nombre de patients, échantillonnée sur des intervalles de taille fixe, et "données exogènes" toutes les données externes rassemblées par diverses méthodes détaillées plus haut.

### 2.2.1 Séries temporelles

Les modèles de séries temporelles sont des outils de base dans la prédiction d'une suite de valeurs modélisées comme des variables aléatoires. Pour gagner une compréhension suffisante du domaine sans trop entrer dans les détails techniques, nous nous sommes basés sur des cours universitaires tels que [Viano and Philippe, 1999].

La modélisation, l'estimation des paramètres et la prédiction ont été effectuées grâce à la librairie `statsmodels`[8].

Le modèle de base considéré est le modèle ARMA (*Auto-Regressive Moving Average*). Si $(Y_t)_{t\in\mathbb{N}}$ est la série temporelle étudiée, une modélisation ARMA suppose que $Y_t$ vérifie l'équation de récurrence suivante :

$$Y_t + \sum_{i=1}^{p} \alpha_i Y_{t-i} = \varepsilon_t - \sum_{j=1}^{q} \beta_{t-j}\varepsilon_{t-j}$$

où $(\varepsilon_t)_{t\in\mathbb{N}}$ est un bruit blanc gaussien. L'idée est que la valeur du nombre de patients à l'instant $t$ peut dépendre à la fois du nombre de patients aux instants précédents et d'un facteur externe non observé, lui aussi pris en compte à différents instants du passé (facteur qu'on pourrait interpréter comme le nombre de gens qui se blessent ou tombent malades).

Paramètres à régler : $p$ (profondeur du modèle autorégressif) et $q$ (profondeur des moyennes mobiles) + tous les coefficients $\alpha$ et $\beta$

L'extension naturelle du modèle ARMA est l'ARIMA, où le I signifie *Integrated* : on ajoute au processus une tendance approchée par une fonction polynomiale du temps. Si cette approche est intéressante pour les séries qui montrent une croissance régulière sous-jacente, notre échelle de temps (de l'ordre de l'année) est trop courte pour observer ce genre d'effets, dus par exemple à l'augmentation de la population. Il existe bel et bien une tendance, une *baseline* sur l'année, mais la modéliser comme polynôme du temps serait peu instructif. On va donc utiliser d'autres extensions du modèle.

Nous passons ainsi au modèle SARMA, où le S signifie *Seasonal*. L'idée est d'introduire une dépendance plus distante dans le temps, pour modéliser une saisonnalité en plus des variations à très court terme. En pratique, cela se fait par l'ajout d'un deuxième modèle ARMA, plaqué sur la série saisonnière.
Dans notre approche, les arrivées de patients étant regroupées sur des intervalles de temps de quelques heures, la saisonnalité ajoutée sera celle de la journée. En effet, c'est l'échelle où se produisent les variations les plus conséquentes, comme on l'a observé plus haut. Bien que les modèles SARMA s'étendent à plusieurs périodes, nous avons choisi de traiter la deuxième saisonnalité notable du problème, de période hebdomadaire, comme une variable exogène.

Paramètres à régler : $s$ (période de la saisonnalité), $P$ (profondeur du modèle autorégressif saisonnier) et $Q$ (profondeur des moyennes mobiles saisonnières) + tous les coefficients $\alpha$ et $\beta$ saisonniers.

Enfin, la version la plus sophistiquée du modèle sera le SARMAX, où le X signifie *External*. Cela revient à ajouter dans la modélisation précédente l'influence linéaire de variables exogènes telles que celles que nous avons récoltées : données météorologiques, épidémiologiques, climatiques ou calendaires. Les jours de la semaine seront alors inclus dans ce type de modélisation, sous forme de variables binaires.

En pratique, si le modèle SARMAX est bel et bien présent dans la librairie `statsmodels`, l'ajout des coefficients correspondant à de nombreuses variables exogènes conduit à un temps

8. http ://www.statsmodels.org/stable/index.html

d'estimation par maximum de vraisemblance déraisonnable. Nous avons donc implémenté notre propre variante, selon le principe suivant :

1. Décomposer le signal $Y_t = T_t + S_t + N_t$ où $T$ est la tendance, $S$ est la saisonnalité journalière, et $N$ est le bruit. Cette décomposition est faite par moyennes mobiles.
2. Estimer $T_t$ par une régression ridge à l'aide des variables exogènes, notons $\hat{T}_t \sim T_t$ le modèle obtenu
3. Estimer $Y_t - \hat{T}_t = (T_t - \hat{T}_t) + S_t + N_t \sim S_t + N_t$ à l'aide d'un SARMA, notons $\hat{Z}_t \sim (Y_t - \hat{T}_t)$ le modèle obtenu
4. La prédiction est alors la somme des deux : $\hat{T}_t + \hat{Z}_t \sim Y_t$

Paramètres à régler : choix des variables externes à inclure, en particulier de leur nombre.

### 2.2.2 Machine learning

Nous comparons le modèle présenté à deux autres modèles classiques d'apprentissage automatique : les réseaux de neurones et les forêts aléatoires. Ces modèles sont plus simples, dans le sens où nous n'avons pas à décrire explicitement une relation de dépendance linéaire entre données endogènes et exogènes. Nous fournissons l'ensemble, ou seulement un sous-ensemble des données exogènes en entrée, pour prédire l'affluence.

Néanmoins, un tel modèle ne tient pas compte de la dépendance temporelle des données, or l'affluence au temps $t$ dépend de l'affluence aux temps précédents $t - k$. Pour obtenir une comparaison équitable avec les modèles précédents, nous devons donc leur fournir en entrée les données d'affluence aux temps précédents.

Ce passage de données pose une difficulté pratique qui n'est pas rencontrée avec les séries temporelles, dont la prédiction se fait au fur et à mesure : le modèle de machine learning, avec sa prédiction point par point, ne doit pas avoir accès à des données futures. Dans notre scénario, le système prédit à minuit la courbe d'affluence de la journée suivante, par pas de 3h. Il ne dispose à ce moment que des données d'affluence de la journée passée, jusqu'à minuit. Pour prédire l'affluence à 3h du matin le lendemain, on disposera donc des données réelles d'affluence à 0h et avant. Mais pour prédire à 15h le lendemain, on ne dispose pas encore de l'affluence réelle à 12h. On ne peut donc pas prévoir d'un seul coup toutes les valeurs par tranche sur la journée à venir. En revanche, on peut prévoir tranche par tranche, et à chaque pas, ajouter à la matrice de données exogènes les valeurs endogènes prédites juste avant par le modèle.

Finalement, les données endogènes fournies à chaque instant sont :
— la valeur d'affluence réelle si elle est connue au moment de la prédiction,
— sinon, la valeur d'affluence prédite par le modèle pour les temps précédents.
Avec une telle procédure, nous approchons le fonctionnement des modèles classiques de séries temporelles, en nous affranchissant des hypothèses de linéarité.

Paramètres à régler : choix des données externes à inclure, ordre $p$ du schéma "autorégressif" simulé (nombre d'antériorités rajoutées aux données exogènes), paramètres propres au modèle d'apprentissage (fonction d'activation, paramètre de régularisation, couches de neurones pour les réseaux de neurones ; nombre d'arbres, critère d'évaluation pour les forêts aléatoires).

## 2.3 Méthode de validation

Pour évaluer les modèles et optimiser leurs paramètres, une méthode très répandue en apprentissage automatique est celle de la validation croisée. Malheureusement, la structure de dépendance des données temporelles ne permet pas d'isoler un morceau au hasard, d'entraîner sur le reste et d'évaluer le modèle sur ce morceau, comme le souligne [Bergmeir and Benítez, 2012]. Il a donc fallu utiliser un autre système.

L'idée est d'exploiter les données sur un intervalle de temps, puis de prédire sur les instants suivants. Dans le scénario que nous avons choisi, la prédiction s'effectue sur la période d'une journée dans le futur.

Le mode de validation théorique est donc le suivant :
— Exploiter les valeurs sur les $k$ premiers jours pour prédire les valeurs du $k+1$-ème (et stocker le résultat)
— Exploiter les valeurs sur les $k+1$ premiers jours pour prédire sur le $k+2$-ème (et stocker le résultat)
— Itérer sur $k$ jusqu'à atteindre la fin de l'année

Attention cependant, "exploiter" peut vouloir dire deux choses ici :

1. A chaque itération, observer les valeurs précédentes et réestimer tous les paramètres du modèle à l'aide de ces données, pour ensuite effectuer la prédiction
2. A chaque itération, observer les valeurs précédentes mais effectuer la prédiction avec un modèle déjà entraîné sur toute une année

L'idéal aurait été d'entraîner le modèle sur l'intégralité de l'année 2014, puis d'effectuer la validation sur les jours de 2015 en gardant les paramètres calculés. Cela aurait permis d'avoir un ensemble d'apprentissage complet (d'une année), tout en gardant un ensemble de validation indépendant. Mais puisque les données de 2015 étaient incomplètes, nous avons du nous restreindre à travailler sur la seule année 2014.

### 2.3.1 Une méthode honnête qui ne marche pas

Notre première approche a donc été de ré-entraîner le modèle à chaque fois (méthode 1). Cette approche permet d'être "honnête" dans le sens où on ne fait pas usage de données futures au moment de la prédiction. Elle implique cependant deux restrictions à notre méthode de validation :

1. Il est déraisonnable de prédire dès les premiers jours de l'année, car l'ensemble d'apprentissage sera trop petit pour avoir un modèle efficace. Nous avons donc placé le point de départ de la prédiction au début de février dans la plupart des cas.
2. Il est déraisonnable de prédire sur tous les jours de l'année, car cela implique des temps de calcul trop importants pour effectuer l'apprentissage sur tous les intervalles $[\![1, k]\!]$. Nous avons donc opté pour une validation sur un ensemble aléatoire de jours.

La figure 8 illustre ce principe à petite échelle. On exclut les `days_skipped` $= 10$ premiers jours de la prédiction. Ensuite, on sélectionne `random_subset` $= 4$ jours parmi les `days_predicted` $= 10$ jours suivants, pour chacun on fait courir l'ensemble d'apprentissage jusqu'au point rouge et on prédit sur la journée suivante.

En prédiction réelle, `days_skipped` = 31 (le mois de janvier est uniquement consacré à l'apprentissage), `days_predicted` = 365 - `days_skipped` et `random_subset` = 30 (on évalue la prédiction sur 30 journées tirées aléatoirement dans les mois de février à décembre).

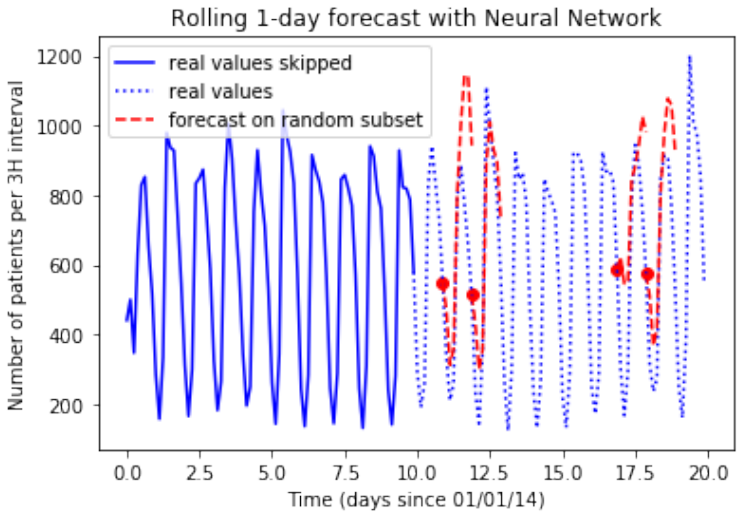


FIGURE 8 – Validation randomisée

Pourquoi cette méthode ne marche-t-elle pas ? L'explication est que puisque nous ne validons que sur un sous-ensemble aléatoire, les erreurs calculées ne sont que des estimations de l'erreur empirique réelle. Et cette estimation s'accompagne d'intervalles de confiance sur l'erreur trop importants pour pouvoir conclure et comparer précisément les modèles. La méthodologie présentée plus haut, choisie à la fois pour être honnête et permettre un temps de calcul raisonnable, n'est donc pas adaptée pour la sélection du meilleur modèle.

A titre d'exemple, on peut voir figure 9 les valeurs de l'erreur absolue moyenne pour les modèles de type ARMA(p,q) avec $(p, q) \in [\![0, 3]\!]^2$.

Chaque point correspond à un modèle ARMA testé selon notre méthodologie. L'ordonnée des points donne les valeurs de l'erreur absolue moyenne sur les 30 jours de prédiction. Notons que l'erreur journalière est elle même la moyenne des erreurs sur les 8 tranches de prédiction de 3h, mais nous la considérons individuellement car notre unité fondamentale est la prédiction sur une journée entière, et les différentes tranches horaires ne sont pas indépendantes entre elles.

Les lignes en pointillés sont les intervalles de confiance à 95% pour l'erreur moyenne absolue, obtenus comme suit. Supposons que l'erreur sur une journée suive une loi normale $\mathcal{N}(\mu, \sigma)$ (ce qui est arbitraire mais facilite la vie). Supposons de plus que les erreurs sur les différentes journées soient indépendantes (ce qui est un peu abusif). Alors l'erreur absolue $\Delta$ suit une loi normale repliée, et l'erreur absolue moyenne dans notre validation est une moyenne de 30 lois normales repliées.
Les intervalles de confiance dans ce cas n'ont pas de forme explicite très agréable, mais étant données les mesures des paramètres empiriques $\mu$ et $\sigma$ de l'erreur, on peut déterminer numériquement un intervalle de confiance. Il suffit de simuler un grand nombre de fois un ensemble de

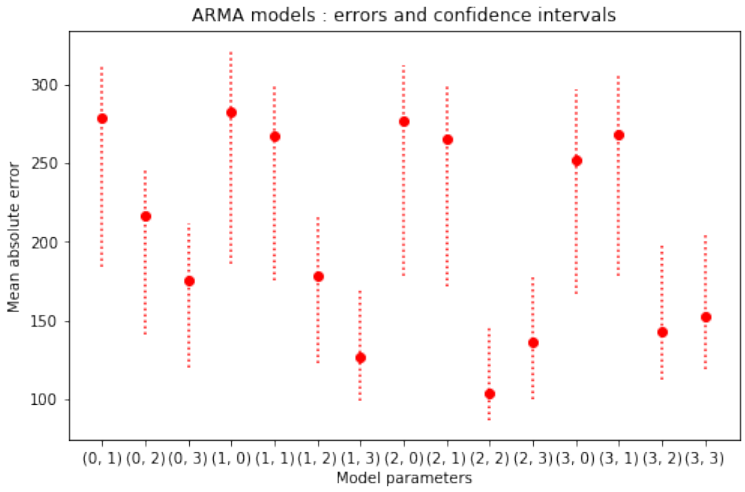


FIGURE 9 – Erreurs absolues moyennes avec intervalles de confiance à 95%

30 variables suivant la loi $|\mathcal{N}(\mu, \sigma)|$, de faire la moyenne de chacun de ces ensembles et d'examiner les quantiles de ces moyennes.

On peut donc constater que si l'ARMA(2,2) semble être le meilleur de sa catégorie, il ne se distingue pas clairement des ARMA(1,3) et ARMA(2,3). D'où la nécessité d'un changement de méthode pour tirer des conclusions sans appel.

#### 2.3.2 Une méthode malhonnête qui marche

Notre seconde approche a donc été d'entraîner les modèles sur toute l'année pour obtenir leurs paramètres, tout en gardant le même système de prédiction des courbes d'affluence jour par jour pour la validation (méthode 1).

Ainsi, plus besoin de réapprendre à chaque nouveau jour de prédiction (ce qui permet des temps de calcul fortement réduits), et plus besoin de laisser passer un ou plusieurs mois au début de l'année avant de commencer l'évaluation : quelques jours suffisent pour laisser aux modèles la possibilité de voir les valeurs passées dont ils ont besoin.

C'est donc sur cette base que seront présentés les résultats, qui ne contiendront donc plus d'aléa intrinsèque à la mesure. Rappelons toutefois que les différentes validations de modèles peuvent conserver un caractère imprévisible, par exemple dans le succès ou non des méthodes d'optimisation pour l'estimation des paramètres par maximum de vraisemblance.

# 3 Résultats

## 3.1 Performances de prédiction

### 3.1.1 Indicateurs mesurés

Nous avons évalué nos modèles avec les paramètres globaux suivants :

— Intervalle de temps d'échantillonnage (pour données exogènes et endogènes : `sampling_freq` = 3h. Ce qui donne une période journalière `period` = 24 / 3 = 8
— Nombre de jours passés avant le début de la prédiction : `days_skipped` = 7 (première semaine de janvier)

Une fois obtenues, pour toutes les tranches de 3h, les prédictions $Y_{\text{pred}}$, les comparer aux valeurs réelles $Y$ permet de quantifier la qualité du modèle en mesurant les indices suivants :
— Sur l'erreur simple $Y_{\text{pred}} - Y$ : moyenne, écart-type
— Sur l'erreur absolue $|Y_{\text{pred}} - Y|$ : moyenne, écart-type, quantiles à 90% et 95%, maximum
— Sur l'erreur absolue relative $\frac{|Y_{\text{pred}} - Y|}{|Y|}$ : moyenne, écart-type, quantiles à 90% et 95%, maximum
— Coefficient $R^2$

Dans la perspective d'une utilisation future en milieu hospitalier, plusieurs indicateurs peuvent être jugés utiles.

Ainsi l'étude de l'erreur, et en particulier de sa moyenne, peut nous montrer si le modèle a tendance à sous-estimer ou à sur-estimer le flot de patients, mais les effets de compensation la disqualifient comme mesure fiable de la qualité de notre modèle. Quant à l'erreur absolue relative, elle peut être utile dans certains contextes, mais le fait qu'elle pondère moins les patients arrivés en période de grande affluence nous paraît dangereux : en effet, si on suppose un personnel constant, un patient supplémentaire sera peut-être une contrainte plus importante si le service est déjà plein que s'il est à moitié vide. C'est pourquoi nous avons préféré baser nos analyses et notre sélection de prédicteur sur l'erreur absolue. Cependant, pour des besoins de visualisation, l'erreur relative pourra être sélectionnée par endroits.

### 3.1.2 Meilleurs modèles de séries temporelles

Du côté des modèles de séries temporelles, nous avons pu comparer les modèles sur la grille de paramètres du tableau 3.1.2.

| Modèle | $p$ | $q$ | $s$ | $P$ | $Q$ |
|---|---|---|---|---|---|
| ARMA | 1...15 | 1...15 | - | - | - |
| SARMA | 1...7 | 1...7 | 8 | 1...4 | 1...4 |

TABLE 1 – Paramètres testés pour les modèles ARMA et SARMA

Des résultats partiels en terme d'erreur absolue relative moyenne sont synthétisés dans la figure 10.

On note de manière générale que pour les modèles ARMA simples, les performances augmentent globalement avec le nombre de paramètres, ce qui suggère qu'on se trouve du bon côté du dilemme biais-variance (on peut encore diminuer le biais en complexifiant le modèle sans voir trop augmenter la variance). On voit aussi un décrochage notable lorsque $p$ ou $q$ dépassent 7 : cela traduit le fait qu'on commence à regarder dans le passé au delà d'une journée (8 intervalles de 3h), et qu'on obtient donc brusquement une information intéressante traduisant la périodicité journalière. Enfin, les taches sur le bas du tableau sont sans doute liées à des problèmes de convergence dans l'estimation des paramètres, résultant en des prédictions extrêmement fausses.

Si on ajoute les composantes saisonnières, le meilleur modèle est le SARMA(1,3)(3,3). On remarque que rajouter des coefficients d'ARMA simple apporte peu face aux bénéfices saisonniers. En revanche si on avait poussé l'exploration plus loin pour les valeurs de $P$ et $Q$ (décalages

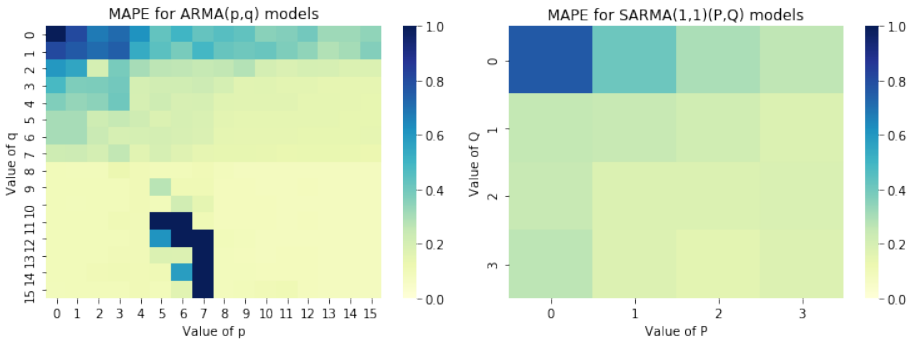


FIGURE 10 – Erreur absolue relative moyenne pour différents modèles ARMA et SARMA

saisonniers), on aurait sans doute pu améliorer les performances.

Partant de ce modèle de base, nous l'avons ensuite transformé en SARMAX, pour faire varier le nombre de features externes. Celles-ci ont été ordonnées par la méthode de la régression lasso (détaillée plus bas), puis ajoutées une à une, afin d'étudier l'évolution de la précision. On obtient le graphe de la figure 11.

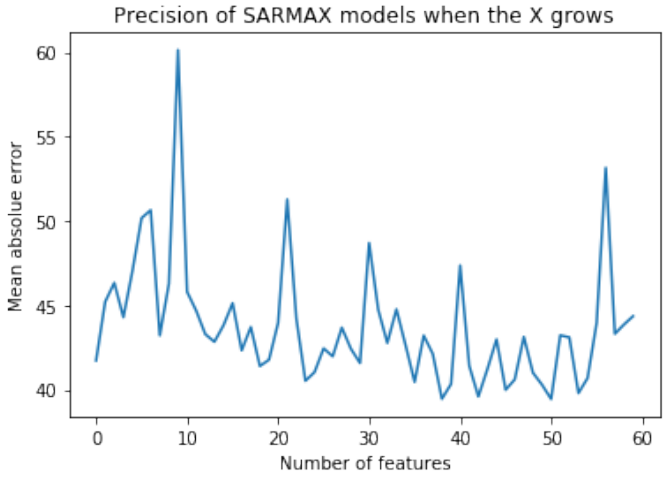


FIGURE 11 – Erreur absolue moyenne d'un SARMAX en fonction du nombre de features

La relation entre le nombre de features et la précision n'est pas monotone, on en déduit donc que l'évaluation de leur importance par lasso n'est pas forcément pertinente pour notre modèle. Cependant les autres méthodes (information mutuelle et arbres de décision) présentent des résultats similaires.

Une méthode plus exhaustive consisterait en un test complet des sous-ensembles de features jusqu'à une certaine taille, ou bien en un algorithme d'optimisation (type algorithme génétique) pour sélectionner l'ensemble optimal de features. Pour des raisons de temps de calcul, nous n'avons pas effectué cette analyse, et nous nous contentons donc pour l'instant du meilleur sous-ensemble sur cette courbe, qui comprend 50 features externes.

En définitive, on a testé trois catégories de modèles, de complexité croissante mais pas forcément inclues l'une dans l'autre : pour des raisons de calcul, tous les paramètres (p,q) testés en ARMA n'ont pas pu l'être en SARMA, et de par notre méthode particulière d'implémentation du SARMAX, celui-ci ne s'apparente pas simplement à un SARMA avec des coefficients en plus. Les champions de chaque catégorie sont récapitulés dans la table 2.

| Modèle | Erreur absolue moyenne |
|---|---|
| ARMA(15,14) | 43.2 |
| SARMA(1,3)(3,3) | 41.7 |
| SARMAX(1,3)(3,3) - 50 features | 39.4 |

TABLE 2 – Meilleurs modèles de séries temporelles

On constate que l'ajout de variables externes ne permet malheureusement qu'une légère amélioration de la performance. Cela peut être du à la façon dont celles-ci sont employées au sein de nos modèles SARMAX modifiés, ou bien simplement à leur faible pouvoir prédictif supplémentaire par rapport au modèle saisonnier.

### 3.1.3 Meilleurs modèles de machine learning

**Réseaux de neurones**

Nous avons testé différents paramètres, liés soit au réseau lui-même (fonction d'activation, paramètre de régularisation $\alpha$, forme du réseau), soit aux données en entrée (choix des caractéristiques externes, et des antériorités des données endogènes). Nous avons dans un premier temps repéré les ordres de grandeur de paramètres qui fournissaient des résultats satisfaisants. Cela a notamment permis de choisir la fonction d'activation ReLU plutôt que la fonction sigmoïde ou identité.

Cette première étude a permis d'effectuer ensuite une recherche selon la grille suivante :

| Paramètre | Valeurs |
|---|---|
| $\alpha$ | $10^{-i} \quad i = 2, 3, 4$ |
| réseau | (5,), (10,), (20,), (30,), (50,), (75,), (100,), (10,5), (20,20), (60,50) |
| features exogènes | toutes (une centaine), les 20 plus importantes |
| antériorités | [], [1, 2], [1, 2, 8], [1, 2, 3, 4, 5, 6, 7, 8], [8, 16, 24], [1, 2, 3, 8, 9, 10, 16, 17, 18] |

TABLE 3 – Grilles de paramètres pour les réseaux de neurones

Le tableau ci-après présente les trois meilleurs modèles pour l'erreur absolue moyenne :

| Modèle | Erreur absolue moyenne |
|---|---|
| ReLU, (100,), $\alpha = 0.001$, toutes features exogènes, antériorités : [1, 2, 3, 4, 5, 6, 7, 8] | 38.44 |
| ReLU, (50,), $\alpha = 0.01$, toutes features exogènes, antériorités : [1, 2, 3, 8, 9, 10, 16, 17, 18] | 38.49 |
| ReLU, (75,), $\alpha = 0.01$, toutes features exogènes, antériorités : [1, 2, 3, 8, 9, 10, 16, 17, 18] | 38.64 |

TABLE 4 – Meilleurs modèles de réseaux de neurones

#### Random forests

Nous avons procédé de manière analogue pour les random forests. Nous réglons d'abord le paramètre `criterion` des arbres à "mae" pour mean absolute error. Nous aurions pu choisir "mse" si nous utilisions ce critère d'évaluation. Nous pouvons également régler le nombre d'estimateurs, c'est-à-dire le nombre d'arbres qui composent notre modèle. Les réglages des données en entrée sont les mêmes : choix des caractéristiques externes, et des antériorités des données endogènes.

Nous effectuons ensuite une recherche selon la grille suivante :

| Paramètre | Valeurs |
|---|---|
| nombre d'estimateurs | 5, 10, 20 |
| features exogènes | toutes (une centaine), les 20 plus importantes |
| antériorités | [], [1, 2], [1, 2, 8], [1, 2, 3, 4, 5, 6, 7, 8], [8, 16, 24], [1, 2, 3, 8, 9, 10, 16, 17, 18] |

TABLE 5 – Grilles de paramètres pour les forêts aléatoires

Le tableau ci-après présente les trois meilleurs modèles pour l'erreur absolue moyenne :

| Modèle | Erreur absolue moyenne |
|---|---|
| 20 estimateurs, toutes features exogènes, antériorités : [8, 16, 24] | 11.7 |
| 10 estimateurs, toutes features exogènes, antériorités : [8, 16, 24] | 12.3 |
| 20 estimateurs, toutes features exogènes, antériorités : [1, 2, 3, 8, 9, 10, 16, 17, 18] | 13.1 |

TABLE 6 – Meilleurs modèles de forêts aléatoires

On note que les résultats sont bien meilleurs que tous les autres modèles, mais vu notre méthode "malhonnête" de validation, cela peut être dû à la proverbiale tendance des forêts aléatoires à l'overfitting.

## 3.2 Interprétation qualitative

Les données exogènes nous permettent de générer de nombreuses caractéristiques que nous pouvons potentiellement inclure dans notre modèle de prédiction pour l'enrichir. Or comme on l'a vu, afin de garder une taille de modèle raisonnable et de ne pas utiliser des données non signifiantes, il faut parfois sélectionner les caractéristiques les plus pertinentes. Cette étude a de plus une utilité pratique indéniable, puisqu'elle met en évidence les meilleurs variables d'influence exogènes de la fréquentation des urgences. C'est ce point que nous abordons à présent.

Pour d'autres analyses des importances relatives de différentes variables exogènes, on pourra notamment consulter [Xu et al., 2013].

### 3.2.1 Mesures générales d'importance

Dans un premier temps, nous pouvons étudier de façon générale l'importance relative des composantes exogènes pour déterminer les nombres de patients par tranches de 3h. Nous avons

utilisé pour ce faire trois méthodes différentes :

*Méthode de la régression lasso.* Nous effectuons des régressions lasso de $Y$, le nombre de patients sur $X$, vecteur contenant toutes les caractéristiques normalisées, pour différents paramètres de régularisation $\alpha$. Quand $\alpha$ augmente, les coefficients $\hat{\beta}_i$ de la régression pour la caractéristique $i$ deviennent progressivement nuls. Nous prenons comme estimateur de l'importance du prédicteur $X_i$ :

$$\alpha_i^\star = \min\{\alpha > 0,\ \hat{\beta}_i = 0\}$$

Cette méthode fait encore l'hypothèse qu'il existe une dépendance linéaire entre les $X_i$ et $Y$.

*Méthode de l'information mutuelle.* Pour chaque prédicteur $X_i$, nous estimons sur les données l'information mutuelle entre le prédicteur $X_i$ et le nombre de patients $Y$ :

$$I(X_i, Y) = \sum_{y \in Y} \sum_{x \in X_i} p(x, y) \log \left( \frac{p(x, y)}{p(x)p(y)} \right)$$

Cette méthode a l'avantage d'être non paramétrique et de ne faire aucune hypothèse sur les lois de $X_i$ ou $Y$.

*Méthode de l'arbre de décision.* Après ajustement aux données pour une tâche de régression, l'arbre de décision peut produire la liste des importances de chaque caractéristiques, calculées comme la réduction totale normalisée de l'erreur apportée par cette caractéristique (on parle également d'importance de Gini).

La figure 12 montre les 30 caractéristiques les plus significatives parmi les 90 caractéristiques exogènes, pour chacune des trois méthodes. Les variables nommées "V-i" désignent la mesure de la variable V à l'instant $t - i$. Si l'on retrouve globalement les mêmes groupes de données, chaque méthode produit un classement significativement différent. Il paraît alors utile d'adapter la méthode de sélection de features au modèle utilisé pour prédire le nombre de patients. Par exemple, il est plus logique d'utiliser les données sélectionnées par lasso pour une régression linéaire, ou la méthode d'arbre de décision pour un modèle de forêts aléatoires.

Cette analyse nous permet de repérer des dépendances subtiles, qui n'apparaissent pas dans l'analyse simple des corrélations effectuée par la suite. En effet, les méthodes d'information mutuelle et d'arbre de décision sélectionnent fortement les données d'épidémies (grippe et gastro-entérite), alors que leur corrélation avec le nombre de patients est faible. Nous pouvons donc supposer que cette relation existe bel et bien, mais qu'elle n'est pas linéaire.

Ces critères d'importance doivent néanmoins être considérés avec précaution. Notamment, certaines variables sont éliminées arbitrairement par la méthode de l'arbre de décision. Cette méthode a la particularité de sélectionner, pour deux producteurs très corrélés, uniquement l'un d'entre eux.

### 3.2.2 Sens des corrélations entre variables exogènes

Nous allons maintenant passer à une analyse des corrélations, plus réductrice car linéaire, mais porteuse d'un sens additionnel de par son signe. Pour simplifier, dans la suite nous analyserons les données exogènes sans décalages, en considérant seulement les valeurs prises au temps

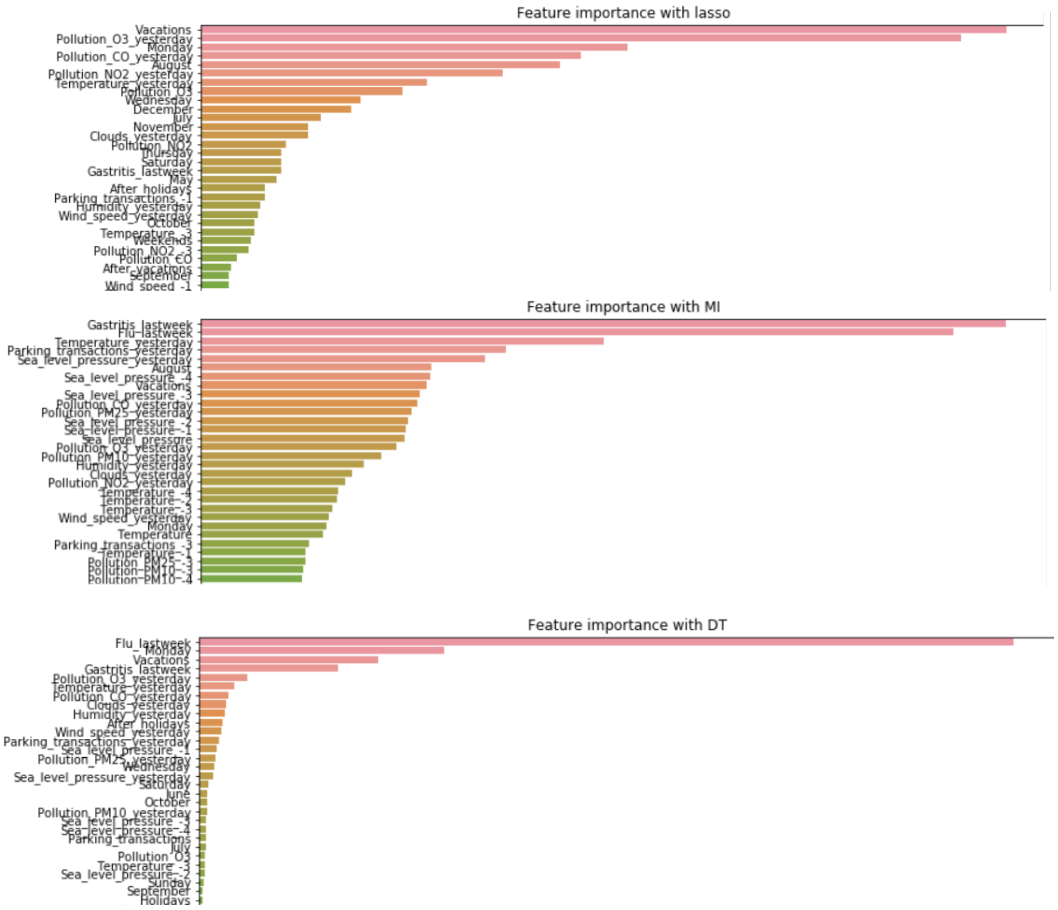


FIGURE 12 – Les 30 caractéristiques les plus importantes sélectionnées par chacune des trois méthodes.

$t$.

Il est intéressant tout d'abord d'examiner les relations qu'entretiennent les variables exogènes entre elles. En effet, nous avons remarqué au cours de notre étude que certaines des variables que nous utilisons représentent peu ou prou la même information. Pour le vérifier, nous pouvons tracer la matrice des corrélations des variables exogènes (en excluant les variables de mois ou de jour qui encombrent la matrice).

Le graphique 13 de ces corrélations appelle quelques remarques :

— Les transactions de parcmètres reflètent bien l'activité humaine à l'échelle d'une journée, avec notamment des diminutions durant les week-ends et les vacances.

— Le mauvais temps (pluie, température) est corrélé avec les épidémies (et avec les mois hivernaux qui n'apparaissent pas ici).

— La pollution O3 est négativement corrélée aux autres types de pollution. Ce comportement différent d'O3 par rapport aux autres polluants peut s'expliquer par le fait qu'il s'agit d'un polluant secondaire (par opposition aux autres, qui sont primaires), qui pro-

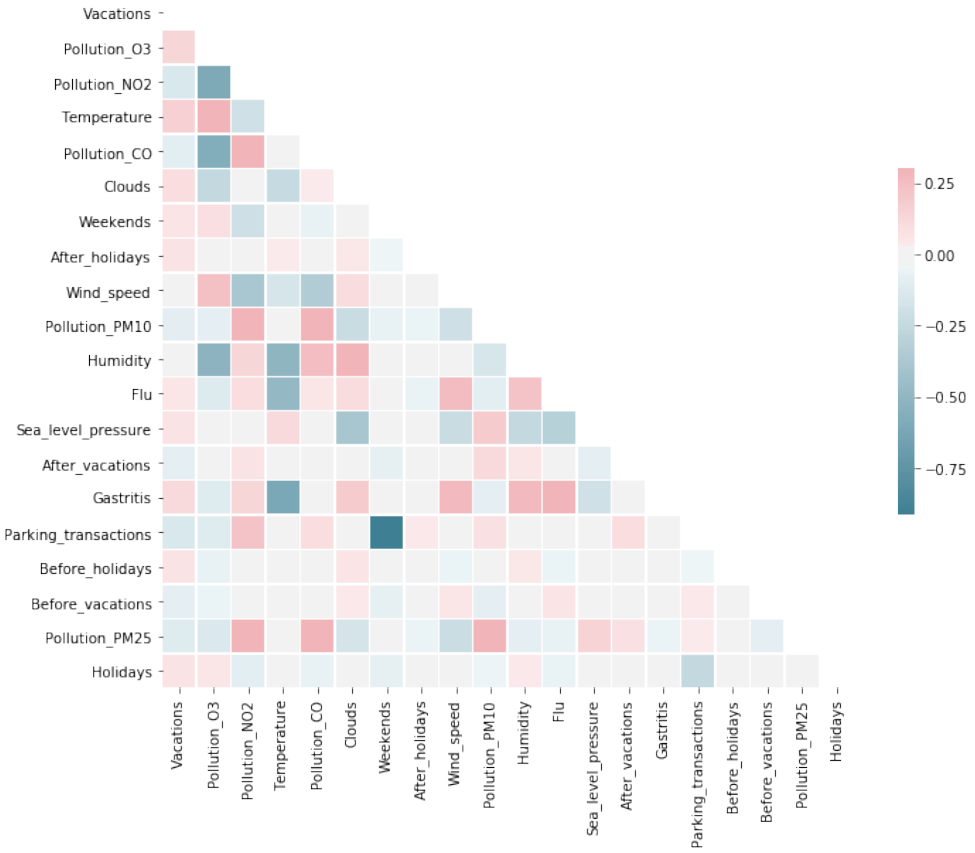


FIGURE 13 – Corrélations entre variables exogènes

vient de la réaction de l'oxygène au contact d'oxydes d'azote et d'hydrocarbures, en présence de rayonnement ultra-violet solaire et d'une température élevée [9].

### 3.2.3 Sens des corrélations avec la variable endogène

Enfin, nous pouvons analyser les corrélations entre les features externes et le nombre de patients. Mais pour cela, il faut différencier deux échelles de temps : celle de la journée (ou plus généralement de la tendance), et celle des variations intra-journalières. Ces deux échelles fourniront des informations différentes, d'abord du point de vue de la modélisation, ensuite du point de vue de l'interprétation.

Ainsi, pour notre modèle de SARMAX modifié, nous avons confié la tâche de prédire la tendance à une régression linéaire sur le signal moyenné, et les variations journalières ont été capturées par le modèle SARMA. A l'inverse, les modèles de machine learning seront intéressés à la fois par les variables qui prédisent bien la baseline jour après jour, et par les variables capables d'expliquer la variance à l'intérieur d'une journée.

Pour décider de la significativité des corrélations, nous utilisons l'échelle heuristique de Co-

9. https ://www.airparif.asso.fr/pollution/differents-polluants

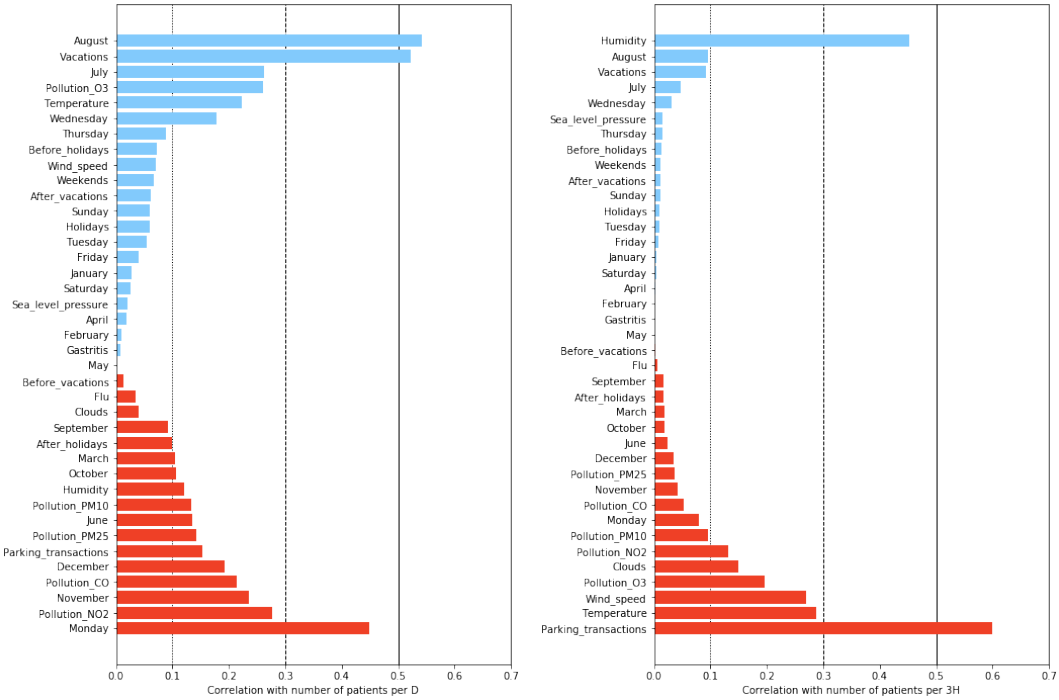


FIGURE 14 – Corrélations variables exogènes / patients, échelle journée et échelle 3h

hen (la corrélation est considérée faible si elle dépasse 0.1, moyenne si elle dépasse 0.3 et forte au delà de 0.5). Nous pouvons donc repérer les données significativement corrélées au nombre de patients (voir figure 14) et examiner leur sens : corrélations positives en rouge, négatives en bleu.

A l'échelle des tranches de 3h, une variable en particulier se détache : les transactions de parcmètres. Notre explication est qu'il ne faut pas y chercher une causalité : cette variable est simplement le reflet d'un cycle journalier d'activité humaine. En revanche cette interprétation est moins valable pour les variables météo, également assez corrélées avec les arrivées aux urgences : si le taux d'humidité présente lui aussi un profil journalier caractéristique (figure 15), d'autres telles que la vitesse du vent ou la nébulosité sont moins périodiques sur 24h (figure 16). On est donc poussés à conclure que les urgences se remplissent plus par mauvais temps.

En étudiant les corrélations sur les moyennes jour par jour, on voit une influence beaucoup plus grande des variables calendaires. Les urgences sont plus fréquentées le lundi (peut-être car les gens y vont moins le week-end et reportent les urgences moins essentielles au début de la semaine) et moins fréquentées le mercredi (pas d'explication claire là-dessus, peut-être faut-il y voir l'effet du jour libéré des enfants en primaire et au collège). Sur le plan des mois, on voit nettement le creux des vacances et des mois d'été (particulièrement en août), et à l'inverse un remplissage accentué durant les mois d'hiver, peut-être plus propices aux maladies. Enfin, le fait de tomber après un jour férié (variable "After_holidays") est également assez déterminant.

Le cas de la température est intéressant, puisqu'à l'échelle de 3h elle est positivement corrélée au nombre de patients (ce qui peut simplement s'expliquer par le fait qu'ils viennent plus durant la journée que durant la nuit), tandis qu'à l'échelle de la journée elle est négativement corrélée

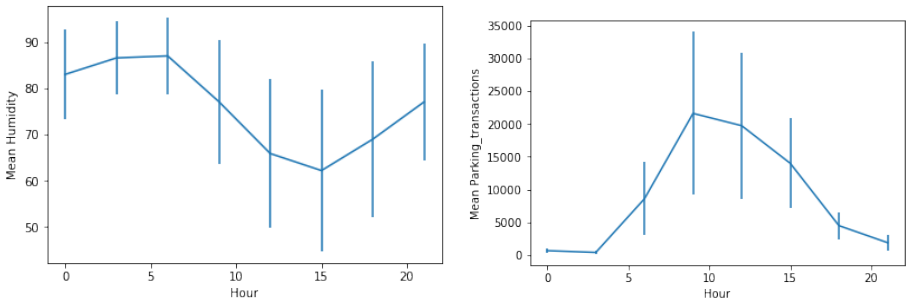


FIGURE 15 – Des variables externes au cycle journalier marqué

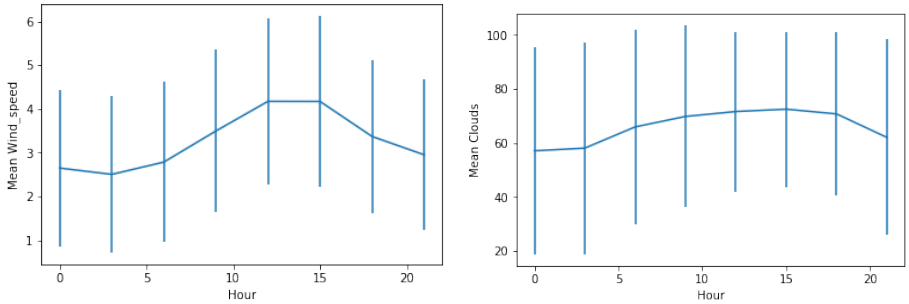


FIGURE 16 – Des variables externes au cycle journalier moins marqué

(potentiellement parce que les patients viennent plus en hiver).

Dans les deux classements, on distingue une relation positive entre la pollution et les arrivées aux urgences (le cas particulier de l'ozone était relevé plus haut). Étant donnée la situation déplorable de la qualité de l'air en Île-de-France, ceci peut être imputé, au moins en partie, à des hausses dans les troubles respiratoires dus aux particules fines. Mais d'autres interprétations sont possibles : par exemple, beaucoup d'épisodes de pollution extrême se produisent en hiver, qui est aussi une saison propice à la demande en soins d'urgences comme on l'a vu plus haut.

Faute de données plus exploitables, nous n'avons malheureusement pas pu mener d'analyses poussées en séparant les patients selon leur degré de gravité ou la raison de leur arrivée. C'est pourquoi cette analyse des corrélations ne peut être que superficielle. En effet, comme le souligne par exemple [Hoot and Aronsky, 2008], les visites non urgentes et les "*frequent flyers*" sont une cause majeure de remplissage des services. Or pour ces visites, on peut avancer que les facteurs environnementaux ou calendaires auront plus d'influence. Le fait qu'elles ne soient pas aisément séparables conduit sans doute à surestimer le rôle et la capacité prédictive des variables exogènes dans les cas vraiment urgents.

## 3.3 Pistes à poursuivre

### 3.3.1 Amélioration des modèles

Les résultats présentés ci-dessus sont plutôt bons mais pas encore tout à fait satisfaisants. En effet, en termes d'erreur absolue relative, on peine à descendre en dessous de 8%, ce qui

peut représenter une quantité conséquente de patients, et donc une surcharge imprévue pour les capacités d'accueil du service. Afin de les améliorer, la première voie envisageable est purement technique : elle réside dans de meilleurs modèles de prédiction.

Dans le domaine des séries temporelles, la variation que nous avons proposée sur le thème du modèle SARMAX est aisément modulable. Elle permet entre autres de remplacer la régression ridge dans l'estimation de la tendance par un modèle d'apprentissage plus sophistiqué, rassemblant ainsi le "meilleur des deux mondes". Dans la même veine, on peut aussi étudier d'autres méthodes d'extraction de la tendance que les moyennes mobiles, ou encore tenter l'ajout d'une deuxième saisonnalité (hebdomadaire) directement dans le modèle au lieu de la traiter à l'aide de variables catégorielles externes.

Une autre piste serait le changement du type de modèle pour passer sur une modélisation de type Poisson, plus adaptée aux séries temporelles qui comptent des évènements (comme l'arrivée de patients). L'idée est de considérer que ce n'est pas $Y_t$ lui-même qui suit une relation linéaire en fonction des valeurs précédentes : $Y_t$ est modélisé comme suivant une loi $\mathcal{P}(\lambda_t)$, et c'est $\lambda_t$ qui est déterminé par une relation de récurrence.

Concernant les méthodes de machine learning, une piste à creuser serait celle des réseaux de neurones récurrents. Ces structures neuronales munies de cycles permettent en effet d'incorporer les dépendances temporelles à l'intérieur du modèle (au lieu de dupliquer articiellement les valeurs endogènes) afin de permettre l'apprentissage.

### 3.3.2 Traitement des données et scénario

Dans un autre registre, nous avons effectué beaucoup de choix arbitraires au sujet du traitement des données, qui étaient pour l'essentiel dus à des sources peu fiables. Ces choix pourraient être remis en cause, et les sources en question pourraient être améliorées.

Par exemple, les données endogènes étaient faussées par le comportement imprévisible des centres, que nous n'avons pas pu corriger précisément faute de pouvoir assigner un patient à un centre. Beaucoup de relevés pour les variables exogènes présentaient des périodes de trous, que nous avons la plupart du temps comblées par interpolation.

De plus, la prédiction pour la journée à venir nécessite de connaître les valeurs à venir des variables d'influence, tranche par tranche. Nous avons supposé cela possible pour la météo et la pollution, où des modèles prédictifs efficaces sont utilisés quotidiennement et faciles d'accès, mais plus difficile pour les épidémies. Ces hypothèses sont sans doute discutables.

Une autre source d'amélioration du processus réside dans le scénario choisi. L'intervalle de 3h résulte d'une décision de notre part, mais il faudrait étudier le bon compromis entre taille de la tranche et précision de la prédiction afin de trouver l'échelle optimale. La contrainte de prédiction jour par jour, si elle semble réaliste, émane quant à elle de la nécessité de ne pas s'avancer trop loin dans le futur. En effet, les modèles de séries temporelles comme les modèles de machine learning se basaient sur une prédiction de tranche pour prédire la tranche suivante, empilant ainsi les incertitudes. Si finalement l'usage dicte que les prévisions dans chaque centre se feront à la semaine, il faudra revoir les méthodes employées et s'assurer qu'elles soient robustes pour une prévision à moyen terme.

### 3.3.3 Ajustement pour les applications pratiques

Ce qui nous amène à la question des applications. En effet, pour étendre ce travail dans une perspective à plus long terme, il est également nécessaire de dialoguer avec les personnels médicaux, afin de comprendre de façon plus juste leurs besoins. Il est envisageable, par exemple, que le nombre de patients ne soit pas une mesure pertinente de la charge d'un service. Une autre possibilité serait de pondérer ces nombres de patients selon la gravité des cas (un cas grave sollicitera plus de soins et donc plus de personnels qu'un cas bénin). Ou bien on pourrait considérer la charge comme le nombre de patients entrants, pondéré par le temps passé dans le service (cela permettant de prendre en compte la tension sur les ressources matérielles comme les lits, et non seulement sur les ressources humaines).

En outre, il serait intéressant, si les contraintes de protection des données personnelles permettent d'avoir accès à cette colonne, de mener une analyse géographique sur les différents centres et les affluences respectives dans chacun. Pour l'instant, le travail sur des données agrégées ne permet aucune considération spatiale, et pourtant cette dimension est essentielle dans les applications futures. En effet le nombre de patients arrivant en un lieu donné dépend beaucoup de sa localisation, de son accessibilité, du type de services et de soins proposés, de la population des quartiers alentours, etc. Les pistes d'extension de l'étude sont donc multiples.

Et pour conclure, si cet outil doit un jour être mis en service dans des conditions réelles, le mécanisme de prédiction doit être assorti d'une estimation d'incertitudes, question qu'il faudra creuser plus en détail si l'on sort des modèles traditionnels de séries temporelles où l'estimation de l'incertitude est justifiée de façon théorique.

De manière générale, le processus dans son ensemble doit être absolument transparent, à la fois dans son fonctionnement et dans l'interprétation de ses résultats. Il faut donc veiller, malgré la complexification des modèles, à ce qu'ils restent aisément explicables et communicables à ceux qui seront les utilisateurs principaux.

# Conclusion

La plupart des grandes organisations, qu'elles soient des entreprises ou des établissements publics, produisent aujourd'hui de grands volumes de données. Depuis quelques années, nous pouvons mesurer la création de valeur qui peut être retirée d'une exploitation adéquate de ces données : meilleure connaissance des clients, optimisation des moyens et des processus de travail, automatisation de certaines tâches. Il en résulte globalement une amélioration de la productivité et du service rendu. La valorisation de ces données métiers pose néanmoins deux grands défis.

Le premier est celui de la juste collecte de données à partir de systèmes d'informations conçus pour répondre à un besoin métier. C'est le cas pour les données d'admission aux urgences dont il est question dans ce rapport. En effet, celles-ci sont extraites de bases de données qui servent avant tout à prendre en compte les patients au moment de leur arrivée plutôt qu'à conserver des informations utiles pour une étude ultérieure. L'Unité de Recherche Clinique Paris IDF Ouest a débuté la collecte des données en 2014, ce qui explique la qualité inégale des données que nous avons exploitées. Cette qualité s'améliorera certainement significativement dans les prochaines années.

Le second défi est celui de la modélisation. S'il est possible d'obtenir certains résultats à partir des données seules, une interaction entre le data scientist et l'acteur métier produira des analyses beaucoup plus riches. Le mode de fonctionnement d'un service d'urgences, la manière de saisir les données, le "savoir métier" acquis par un professionnel, ainsi que la compréhension de son besoin sont autant de paramètres qui pourront aider le chercheur dans son travail de modélisation. C'est à ce prix que l'on produira un service utile aux professionnels de la santé et à leurs patients.

# Références


Brent R. Asplin, David J. Magid, Karin V. Rhodes, Leif I. Solberg, Nicole Lurie, and Carlos A. Camargo. A conceptual model of emergency department crowding. *Annals of Emergency Medicine*, 42(2) :173–180, August 2003. ISSN 0196-0644. doi : 10.1067/mem.2003.302. URL http://www.sciencedirect.com/science/article/pii/S019606440300444X.

Christoph Bergmeir and José M. Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191(Supplement C) :192 – 213, 2012. ISSN 0020-0255. doi : https://doi.org/10.1016/j.ins.2011.12.028. URL http://www.sciencedirect.com/science/article/pii/S0020025511006773. Data Mining for Software Trustworthiness.

Steven L. Bernstein, Dominik Aronsky, Reena Duseja, Stephen Epstein, Dan Handel, Ula Hwang, Melissa McCarthy, K. John McConnell, Jesse M. Pines, Niels Rathlev, Robert Schafermeyer, Frank Zwemer, Michael Schull, Brent R. Asplin, and Emergency Department Crowding Task Force Society for Academic Emergency Medicine. The Effect of Emergency Department Crowding on Clinically Oriented Outcomes. *Academic Emergency Medicine*, 16 (1) :1–10, January 2009. ISSN 1553-2712. doi : 10.1111/j.1553-2712.2008.00295.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1553-2712.2008.00295.x/abstract.

Nathan R. Hoot and Dominik Aronsky. Systematic Review of Emergency Department Crowding : Causes, Effects, and Solutions. *Annals of Emergency Medicine*, 52(2) :126–136.e1, August 2008. ISSN 0196-0644. doi : 10.1016/j.annemergmed.2008.03.014. URL http://www.sciencedirect.com/science/article/pii/S0196064408006069.

M.-C. Viano and A. Philippe. *Maîtrise d'Économétrie : Cours de Séries Temporelles*. Université de Sciences et Technologies de Lille, 1999.

M. Wargon, B. Guidet, T. D. Hoang, and G. Hejblum. A systematic review of models for forecasting the number of emergency department visits. *Emergency Medicine Journal*, 26 (6) :395–399, June 2009. ISSN 1472-0205, 1472-0213. doi : 10.1136/emj.2008.062380. URL http://emj.bmj.com/content/26/6/395.

M. Xu, T. C. Wong, and K. S. Chin. Modeling daily patient arrivals at Emergency Department and quantifying the relative importance of contributing variables using artificial neural network. *Decision Support Systems*, 54(3) :1488–1498, February 2013. ISSN 0167-9236. doi : 10.1016/j.dss.2012.12.019. URL http://www.sciencedirect.com/science/article/pii/S0167923612003855.