Mémoire présenté le : 08/03/2022

pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA et l'admission à l'Institut des Actuaires

Par : Fabien Travaillot

Titre Comprendre et prédire l'absentéisme grâce au Machine Learning et Forecasting

Confidentialité : ☐ NON ☒ OUI (Durée : ☐ 1 an ☒ 2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membre présents du jury de l'Institut des Actuaires signature

David DUBOIS
Anaëlle LE BERRE
Sylvain CARACO

Membres présents du jury de l'ISFA

Anne EYRAUD-LOISEL
Esterina MASIELLO

Entreprise :

Nom : ADDACTIS

Signature :

Directeur de mémoire en entreprise :

Nom : Cécile Paradis

Signature :

Invité :

Nom :

Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels *(après expiration de l'éventuel délai de confidentialité)*

Signature du responsable entreprise

Signature du candidat

RESUME

Depuis ces dernières années, l'absentéisme au sein des entreprises du secteur privé a connu une croissance et ce, dans quasiment l'ensemble des domaines d'activité. Les coûts liés à ce phénomène ont des impacts importants, que ce soit en termes d'organisation, de rendement pour les entreprises, mais également en termes budgétaire ou en gestion du risque pour les assureurs et la Sécurité Sociale. Comprendre et prédire l'absentéisme semble donc inévitable afin de maîtriser au mieux ce risque pouvant potentiellement s'intensifier dans un contexte d'épidémie mondiale touchant ainsi l'ensemble des acteurs.

Afin de proposer une étude sur la compréhension et la prédiction de l'absentéisme, ce risque est étudié à l'aide du taux d'absentéisme. Les travaux de ce mémoire ont été réalisés sur un portefeuille du secteur de l'industrie, avec un effectif de près de 30 000 individus et un historique de 3 ans. Cette étude vise à mettre en place de nouvelles méthodes statistiques permettant une surveillance du risque plus fine, rendue possible avec l'arrivée de la Déclaration Sociale Nominative.

L'utilisation des outils de *Machine Learning* dans une première partie, comme les arbres *CART* ou encore les modèles de forêts aléatoires, permet dans un premier temps de segmenter le portefeuille selon le risque d'absentéisme, d'étudier les variables les plus influentes dans la modélisation du taux d'absentéisme et de proposer de premières prédictions de ce taux. Cependant l'absentéisme est un phénomène qui évolue au cours du temps. La rigidité des modèles présentés implique un manque de robustesse pourtant nécessaire pour leur exploitation au cours du temps.

L'implémentation de modèles basés sur les séries temporelles permet ainsi d'étudier la prédiction de l'absentéisme sous un nouvel angle en s'intéressant à la structure d'évolution du taux d'absentéisme au cours du temps. L'ajout de variables explicatives appelées régresseurs apportant de l'explicabilité aux prédictions du taux d'absentéisme permet ainsi d'obtenir de nouvelles modélisations plus flexibles définies à partir de mesures d'erreur adaptées à la problématique.

Cette nouvelle façon d'étudier le risque absentéisme permet d'intégrer des facteurs exogènes à la vie de l'entreprise, pouvant influencer le taux d'absentéisme au cours du temps. La situation actuelle d'épidémie avec l'émergence de nouveaux variants, des périodes de confinements ou encore de couvre-feux peuvent ainsi être intégrés dans la modélisation en prenant en compte la temporalité de ces phénomènes pour décrire et anticiper au mieux l'absentéisme.

MOTS CLES

Déclaration Sociale Nominative ; Taux d'absentéisme ; Forêts aléatoires ; Séries temporelles ; Modèle ARIMAX.

ABSTRACT

In recent years, absenteeism in private sector companies has increased in almost all areas of activity. The costs related to this phenomenon have important impacts, whether in terms of organization, performance for companies, but also in terms of budget or risk management for insurers and Social Security. Understanding and predicting absenteeism therefore seems inevitable in order to better control this risk, which could potentially intensify in the context of a global epidemic affecting all players.

In order to propose a study in the understanding and prediction of absenteeism, this risk is studied via the absenteeism rate. The study was carried out on a portfolio of the industry sector, with a workforce of nearly 30,000 individuals and a 3-year history. It aims at implementing new statistical methods allowing a more accurate risk monitoring, made possible with the arrival of the Nominative Social Declaration.

The use of Machine Learning tools in a first part, such as CART trees or random forest models, allows to segment the portfolio according to the risk of absenteeism, to study the most influential variables in the modeling of the absenteeism rate and to propose first predictions of this rate. However, since absenteeism is a phenomenon that evolves over time, the rigidity of the models presented implies a lack of robustness, which is necessary for their exploitation over time.

The implementation of models based on time series allows us to study the problem from a new angle by focusing on the evolutionary structure of the absenteeism rate over time. The addition of explanatory variables called regressors, which make the predictions of the absenteeism rate more explicable, allows us to obtain new, more flexible models defined from error measures adapted to the problem.

This new way of studying the risk of absenteeism makes it possible to integrate factors that are exogenous to the life of the company and that can influence the absenteeism rate over time. The current epidemic situation with the emergence of new variants, periods of confinement or curfews can thus be integrated into the modeling by taking into account the temporality of these phenomena in order to describe and anticipate absenteeism as well as possible.

KEYWORDS

Nominative Social Declaration; Absenteeism rate; Random Forest; Time series; ARIMAX model.

REMERCIEMENTS

Je tiens tout d'abord à remercier mes collègues d'Addactis, en particulier Alexandra Barral et Jean-Pascal Hermet qui ont contribué pour leurs idées et leurs remarques à définir l'orientation de mon mémoire, pour leurs fortes implications, leurs patiences, et pour leurs encouragements à terminer celui-ci. Je leur en serai pour toujours infiniment reconnaissant. Je remercie également Nabil Rachdi pour nos échanges techniques et ses conseils sur des points parfois « touchy ». Je remercie également Cécile Paradis de m'avoir permis de réaliser ce mémoire dans l'équipe Data Life&Health chez Addactis.

Je tiens également à remercier Stéphane Loisel en tant que tuteur pédagogique à l'ISFA de ce mémoire, pour les échanges ayant été réalisés au cours de l'année et également pour sa validation pour présenter ce mémoire.

Je remercie très chaleureusement l'ensemble des personnes qui m'auront apporté leur soutien durant cette période difficile, je pense qu'ils se reconnaîtront. En particulier, je tiens à remercier l'ensemble de ma famille, mais plus encore mes parents qui ont toujours été présents. J'aimerais pouvoir leur rendre au centuple tout ce qu'ils ont pu m'apporter, je ne pourrai jamais assez les remercier. Ce mémoire leur est entièrement dédié.

Je n'oublierai pas Clara et Iris mes premières connaissances à l'ISFA et collègues chez Addactis. Je les remercie pour tous ces moments passés ensemble, nos échanges, nos fous rires, et bien évidemment leur soutien. Pour finir je remercierai spécialement Solène pour son soutien précieux, ses relectures et ses conseils sans qui je n'aurais pu finir ce mémoire.

SOMMAIRE

Introduction

Depuis ces dernières années, **l'absentéisme au sein des entreprises n'a cessé d'augmenter** et ce, dans l'ensemble des secteurs d'activité. Si le taux d'absentéisme moyen était de 4,59% en 2016, celui-ci a progressé pour atteindre les **5,11% courant 2019** en France. Pourtant, une forme d'acceptation des entreprises à l'égard de l'absentéisme, en particulier pour les arrêts longs, se fait ressentir, à tel point que certains acteurs ne les prennent plus en compte dans leur calcul du taux d'absentéisme.

Cependant, **le risque absentéisme a bien un coût.** En France, ce coût est **estimé autour de 108 milliards d'euros** par l'Institut Sapiens. Il est souvent décomposé en deux types de coûts distincts : les coûts directs et les coûts indirects. Les arrêts maladies traduisent les coûts directs pour l'entreprise, évalués aux alentours de 60 milliards d'euros, tandis que les coûts indirects se traduisent par la perte de productivité ou l'intensification des besoins de recrutement ; des coûts plus difficiles à détecter mais pourtant qui ne sont pas sans gravité.

Puisque ce **phénomène d'absentéisme représente un manque à gagner** pour la productivité des entreprises, mais également pour la Sécurité Sociale et les organismes assureurs, il semble légitime de proposer des **méthodes d'analyse, de prédiction, d'explicabilité et de mise en place de mesures de prévention afin de maitriser ce risque.**

La première partie de ce mémoire présentera **le contexte d'étude et les différentes données qui ont été utilisées**. En outre, une définition de l'absentéisme ainsi qu'une présentation du contexte actuel et du système de protection sociale en France seront proposées. Une base de données issue d'un organisme assureur sera exploitée afin de mettre en lumière la place de l'absentéisme dans ce portefeuille. Une section sur la présentation et l'explication du traitement des données utilisées sera ensuite abordée tout en mettant en avant l'importance de la confidentialité des données et des réglementations en vigueur.

Les prédictions du taux d'absentéisme par les méthodes de Machine Learning seront au cœur de la deuxième partie des travaux. La variable à prédire, à savoir le taux d'absentéisme, sera alors présentée afin d'aboutir à la construction de modèles de Machine Learning, à l'aide **d'outils usuels de modélisation des comportements clients**. Les résultats obtenus seront ensuite étudiés en mettant en avant les limites de ces modélisations qui ne prennent pas assez en compte la temporalité des données lors des différentes prédictions effectuées.

Les méthodes de Forecasting viendront ensuite compléter ce mémoire dans une troisième partie. Cette partie sera l'occasion de présenter et de comparer ces méthodes avec celles utilisées dans la deuxième partie de ce mémoire. La théorie des séries temporelles sera introduite, pour finalement utiliser ces outils, **donner des prédictions mais également de l'explicabilité à ces résultats. Cette méthode innovante de l'étude de l'absentéisme prenant en compte la temporalité** montrera le potentiel d'une telle approche dans **le suivi de ce risque au cours du temps,** comparé aux méthodes classiques précédentes**.**

Avertissement préalable à la lecture du mémoire :

Les travaux suivants, présentés dans ce mémoire, mettent en œuvre une modélisation du taux d'absentéisme à partir de modèles de *Machine Learning* et de modèles plus dynamiques de séries temporelles. Les données, utilisées à ces fins, sont des données fournies par un organisme assureur sur les 3 exercices : 2017, 2018 et 2019. Ces données sont présentées au cours de la première partie de ce mémoire.

Les données relatives aux arrêts de travail et permettant de réaliser les travaux, proviennent des **Données Sociales Nominatives (DSN), entrées en vigueur au 1er janvier 2017**. **La montée en charge de ces éléments étant progressive sur cette première année**, il est important d'identifier dès à présent qu'elles ne contiennent pas l'ensemble des informations réelles du portefeuille observé (nous observons d'ailleurs que le nombre d'arrêts de travail recensés sur 2017 est croissant)**.** L'utilisation de ces données dans un contexte de modélisation du taux d'absentéisme peut **engendrer un biais sur les résultats**, en particulier lors de la modélisation via *Machine Learning* et la prédiction du taux d'absentéisme d'une année sur l'autre.

Une première approche consistant à ne pas prendre en compte ces données 2017 aurait pu être mise en œuvre afin de ne pas biaiser les résultats lors de la prédiction du taux d'absentéisme. Cependant, l'objectif de ce mémoire étant de montrer **le caractère statique de la modélisation via les modèles de *Machine Learning***, pour basculer sur une modélisation plus dynamique du taux d'absentéisme : **l'utilisation d'un historique d'au moins 3 années était nécessaire.** De plus, les autres données 2017 étant complètes (consommation santé, caractéristiques des individus, données Open Data, …), l'utilisation de ces autres **données permet de vérifier la stationnarité des données sur 2018 et 2019 pour les travaux de projection**, et compenser les prédictions réalisées sur 2019. Enfin, les données de 2017 étant **moins impactantes lors des projections du modèle de série temporelle** (car un apprentissage moins long dans le temps par l'étude des données passées précédentes à court terme), celles-ci sont conservées.

Le lecteur restera donc vigilant lors de la lecture des résultats du modèle de *Machine Learning* : ce dernier pourrait présenter des **faiblesses lors de la prédiction du taux d'absentéisme, d'une part du fait du caractère statique de la modélisation, et d'autre part du fait du biais présent sur les données arrêt de travail 2017.**

1. CONTEXTE ENJEUX ET CADRE DE L'ETUDE REALISEE

L'absentéisme est un **phénomène de société** bien trop souvent **négligé par les entreprises**. La présentation d'une définition de l'absentéisme, des indicateurs utilisés ainsi que de statistiques provenant de différentes études montreront que l'absentéisme est **un réel problème** avec de **nombreux enjeux**, aux conséquences parfois minimisées et pourtant bien présentes.

Ce phénomène touchant une plus grande part de la population française d'année en année, la présentation des **différents acteurs touchés** par ce fléau permettra par la suite de montrer **les actions menées** par chacun d'entre eux pour **minimiser les coûts pesant sur les salariés**.

Enfin, **la présentation des données** qui seront utilisées dans le cadre de cette étude sur l'absentéisme conclura cette partie **en introduisant les modélisations** qui seront présentées par la suite.

1.1. L'absentéisme en France

1.1.1 L'absentéisme, un sujet souvent négligé

Comme aux prémices d'un rhume, l'absentéisme semble ne pas avoir de grandes conséquences sur les entreprises.

- Un coût souvent jugé non significatif
- Un absentéisme élevé uniquement dans les secteurs faisant intervenir un travail physique
- Des absences récurrentes sur un unique groupe particulier d'individus
- L'idée d'une certaine fatalité liée à l'absentéisme qui ne peut être réduit …

Toutes ces affirmations sont des raisons souvent avancées par les employeurs pour ne pas s'intéresser à ce problème.

Pourtant, plus le temps passe, plus **l'absentéisme engendre des conséquences directes et indirectes non négligeables**, devenant finalement un problème majeur aux lourdes conséquences sur le bon fonctionnement et la bonne santé de l'entreprise. De plus, ces fausses affirmations peuvent vite être écartées aux vues de nombreux rapports et données statistiques sur l'absentéisme.

Un coût sous-estimé ? Selon l'étude menée par l'Institut Sapiens [5] en 2018, **le coût total** de l'absentéisme avoisinerait les **108 milliards d'euros par an aux entreprises et à l'Etat**. Equivalent à presque deux fois le budget du ministère de l'éducation nationale, il impacte ainsi la croissance française. Ce montant prend en compte **les coûts cachés de l'absentéisme** : ceux qui ne sont pas comptabilisés dans les comptes de résultat ou dans les budgets. Ces coûts sont de ce fait trop souvent minimisés à l'échelle de l'entreprise. Ces coûts cachés pourraient être calculés

comme étant la somme des sursalaires (dus aux versements de l'employeur d'un maintien de salaire en fonction des conventions sociales en vigueur dans l'entreprise), du surtemps (dus aux actes effectués par les agents présents pour réguler l'activité en raison des absences), des surconsommations (dus aux achats supplémentaires de services pour pallier les absences) et la non-production (liée au travail non effectué).

Un niveau d'absentéisme lié uniquement à la pénibilité du travail ? Selon le 11ème baromètre de l'absentéisme de 2019[2] proposé par Ayming-AG2R LA MONDIALE, le taux d'absentéisme du secteur industrie-BTP (4,26%) était inférieur à celui du secteur des services (5,26%). De ce fait, **les raisons d'un absentéisme élevé ne sont pas entièrement liées à la pénibilité du travail,** une assertion trop souvent répandue. Les raisons de l'absentéisme sont multiples et doivent faire l'objet d'une étude approfondie de l'entreprise pour en comprendre les tenants et les aboutissants.

Des absences uniquement liées à un groupe particulier d'individus ? Certaines absences récurrentes peuvent effectivement provenir d'un même groupe d'individus au sein de l'entreprise, ou ayant des caractéristiques communes. Cependant, **l'absentéisme peut toucher l'ensemble des salariés et avoir des causes différentes.** Ainsi, l'étude portée par Ayming en 2019 [2] a montré que l'absentéisme était plus présent chez les salariés les plus âgées. Cependant, depuis quelques années, une dégradation de l'absentéisme chez les jeunes salariés a été mesurée, avec **une hausse de 23% des absences de plus de 90 jours pour les moins de 40 ans.** L'absentéisme peut ainsi toucher l'ensemble d'un portefeuille de salariés. Une segmentation des profils d'absentéisme doit donc être réalisée afin de pouvoir étudier profil par profil les raisons et les caractéristiques de chaque absence, pour permettre d'agir sur les différents profils d'absentéisme avec une solution de prévention adaptée. (cf. partie 2.2.2)

L'absentéisme, une fatalité ? L'absentéisme au travail est souvent perçu comme étant un phénomène qui ne peut être endigué. Jugées comme étant la conséquence d'évènements non maîtrisables, les absences semblent inévitables dans le monde de l'entreprise. Bien que **certaines absences soient effectivement en lien avec des causes incontrôlables (telles que les épidémies de grippe, de gastro…), d'autres en revanche sont liées à des facteurs endogènes à l'entreprise**. Certaines études montrent que **plus de la moitié des absences** appartiennent à cette dernière catégorie. Des mesures de prévention pourraient être alors mises en place afin de contrôler et diminuer cet absentéisme. Par exemple dans le secteur du BTP, l'Organisme Professionnel de Prévention du Bâtiment et des Travaux Publics (OPPBTP) permet d'améliorer le cadre de travail et la sécurité des salariés contre d'éventuels accidents. Cette prise en compte de la prévention dans le monde du travail peut justifier un taux d'absentéisme inférieur dans le secteur industrie-BTP qui est habitué à améliorer les conditions de travail pour accroître la sécurité des salariés. A l'inverse, le secteur des services, dont l'absentéisme est plutôt lié à la méthode de management des équipes et le recours à l'intérim, doit également jongler avec des contraintes psychologiques qui sont plus difficiles à prendre en compte.

L'absentéisme au travail est devenu un réel enjeu économique et de société dont les conséquences ne doivent pas être prises à la légère. Si les entreprises sont évidemment les

premiers acteurs touchés par ce phénomène, il ne faut pas oublier que **l'absentéisme a également un coût pour les assureurs**, mais également pour les régimes d'assurance maladie dont la situation financière s'est dégradée depuis ces dernières années. Les études et les actions mises en place pour diminuer l'absentéisme sont aujourd'hui nécessaires pour maîtriser les conséquences économiques touchant différents acteurs de la société. A terme, les objectifs sont de pouvoir offrir de meilleures conditions de travail aux salariés par le biais de la prévention, d'améliorer la santé financière de l'assurance maladie en diminuant les prestations versées en lien avec les absences au travail mais également d'aider les assureurs en diminuant leurs ratios de S/P (sinistres/primes). Ainsi diminués, de meilleurs S/P permettraient un abaissement des primes d'assurance envers les entreprises, ce qui permettrait de créer un cercle vertueux entre ces différents acteurs.

D'année en année, la problématique de l'absentéisme devient de plus en plus considérée et prioritaire, aux enjeux multiples. **L'accès à de nouvelles données notamment pour les assureurs, permettant d'étudier plus finement ce phénomène, offre de nouvelles opportunités de maîtrise de ce risque**.

Depuis peu, le législateur a mis en place une nouvelle démarche de déclaration et de paiement des cotisations sociales des salariés. **La Déclaration Sociale Nominative (DSN)** vient remplacer la déclaration annuelle des données sociales unifiée (DADS-U) ainsi que la déclaration obligatoire d'emploi de travailleurs handicapés (DOETH). Cette nouvelle déclaration permet de simplifier les processus administratifs et de réduire le nombre de démarches et de données transmises en fiabilisant le processus à l'aide d'un document **unique et dématérialisé**. Mise en vigueur progressivement à partir de 2017, la DSN concerne l'ensemble des entreprises du secteur privé[1] qui emploient des salariés[2]. Les informations remplies dans la DSN sont alors de deux sortes :

- **Les données concernant la paie du salarié (poste, ancienneté, cotisations…)**
- **Les évènements concernant les périodes d'activité du salarié**, à savoir les arrêts de travail, maladies, périodes de maternité, périodes de paternité ou encore la fin de contrat

La déclaration de ce dernier point est un atout majeur pour l'étude de l'absentéisme au sein des entreprises. En effet, comme dit précédemment, la fiabilisation des données par le biais de cette unique déclaration permet ainsi d'obtenir **des informations fiables et chronologiques** des évènements passés des salariés au sein de l'entreprise, en particulier des données d'arrêts de travail **transmises aux assureurs dès le premier jour de l'arrêt et non plus après expiration**

[1] La DSN sera généralisée au domaine de la fonction publique au 1er janvier 2022
[2] hormis quelques exceptions comme les marins-pêcheurs, dockers ou fonctionnaires en détachement auprès d'un établissement privé qui doivent continuer à remplir la DADS-U

de la franchise comme c'était le cas auparavant. L'étude et le traitement des données DSN permettraient ainsi de **mettre en place une étude de l'absentéisme au sein de chaque entreprise à partir de ces données fiabilisées et mises à jour à minima mensuellement.**

Comprendre l'absentéisme au travers d'études statistiques, de modélisations ainsi que de prédictions est donc la motivation première de ce mémoire qui est rendue possible grâce à l'usage de données retraitées à partir de données DSN. La première étape consiste alors à définir un cadre précis d'étude afin de pouvoir par la suite comprendre et modéliser l'absentéisme au sein d'une entreprise.

1.1.2 Définition et indicateurs de l'absentéisme

Si l'absentéisme se rapporte au mot « absence », cette notion nécessite une définition et un cadre d'étude précis. En effet, la manière de définir et de comptabiliser l'absentéisme n'est pas unique. Au sein d'une entreprise, l'absentéisme peut être défini suivant la définition du dictionnaire [10] ci-dessous :

> « Absentéisme, fait d'être absent du lieu de travail, (…) où, pour des raisons de travail, de participation à une action ou autre, la présence est obligatoire. »

Il reste alors à définir précisément les causes valables d'une absence au sein d'une entreprise. En effet les salariés d'une entreprise ne sont pas présents tous les jours de l'année. Les week-ends, les congés payés, les RTT, les formations ou encore les grèves sont autant de cas où le salarié peut être absent de son lieu de travail. Pourtant ces jours d'absences ne sont généralement pas pris en compte dans le cadre de l'absentéisme.

Afin de valider un jour d'absence au sein de l'entreprise, le mieux est de revenir à la législation en vigueur. La DSN permettant de prendre en compte les évènements liés aux arrêts de travail, il est alors possible de comptabiliser comme étant de l'absentéisme, toute absence appartenant à la liste des **motifs d'arrêt possibles écrits dans la DSN**. Les différents motifs d'arrêt sont les suivants :

- Maladie
- Maladie professionnelle
- Maternité / Adoption
- Paternité / Accueil de l'enfant
- Congé à la suite d'un accident de trajet
- Congé à la suite d'un accident de travail
- Femme enceinte dispensée de travail

À la suite de cette définition de l'absentéisme et des différents motifs d'arrêt désignés comme étant valide pour prétendre à une absence au sein de l'entreprise, il est nécessaire **d'introduire des indicateurs**. L'objectif est de pouvoir **étudier l'absentéisme sous différents angles** afin d'avoir une vision claire des enjeux sous-jacents et par la suite de proposer des solutions.

Outre le fait de comptabiliser le nombre d'arrêts pour chaque motif sur une certaine période et sur un groupe d'individus, il est possible d'ajouter d'autres informations en lien avec ces absences. **L'absentéisme est une notion propre au temps.** Quelle est la durée de l'arrêt ? Quel jour a débuté l'arrêt de travail ? Combien de fois l'individu a été absent sur une période ? Toutes ces questions nécessitent une réponse à l'aide d'indicateurs prenant en compte la notion de temporalité. De manière non exhaustive, les indicateurs les plus utilisés pour étudier l'absentéisme sont les suivants :

- **La fréquence**

Cet indicateur permet de mesurer **l'occurrence des arrêts** sur une certaine période et pour un groupe de salariés

$$Fréquence = \frac{Nombre\ d'arrêts}{Somme\ des\ expositions^{3}\ salariés}$$

- **La gravité**

Cet indicateur permet de mesurer **la durée moyenne des arrêts**

$$Gravité = \frac{Nombre\ de\ jours\ d'arrêt}{Nombre\ d'arrêts}$$

Il est également possible d'étudier la distribution de ces durées d'arrêt au lieu d'analyser uniquement la durée moyenne

- **La proportion d'absence**

Cet indicateur permet de calculer **la proportion d'individus ayant eu au moins un arrêt de travail** sur la période et le groupe d'individus d'étude

$$Proportion\ d'absence = \frac{Nombre\ d'individus\ avec\ au\ moins\ un\ arrêt}{Nombre\ total\ d'individus}$$

- **Le taux d'absentéisme**

On définit pour l'ensemble de ce mémoire **le taux d'absentéisme** comme suit :

[3] Exposition d'un salarié = durée de présence du salarié dans le portefeuille, une exposition d'un an dans le portefeuille est égale à 1

$$Taux\ d'absentéisme = \frac{nombre\ de\ jours\ en\ arrêt}{nombre\ de\ jours\ d'exposition}$$

Le nombre de **jours d'exposition correspond au nombre de jours où le salarié est enregistré comme faisant partie de l'entreprise**. En outre, les week-ends et les congés sont également pris en compte. Ce taux d'absentéisme peut alors être calculé **individuellement,** mais également **collectivement** en sommant les jours de présences et d'arrêts sur les ensembles des individus impliqués, et ce pour **n'importe quelle période** définie. Cet indicateur est un indicateur global et permettant d'apprécier l'absentéisme au sein d'un effectif.

L'analyse combinée de ces différents indicateurs peut ainsi aider à la compréhension de l'absentéisme au sein d'une entreprise. Par exemple, une gravité faible avec une fréquence et une exposition élevée peut révéler une surcharge de travail ou des restructurations ponctuelles ou mal vécues. A l'inverse une gravité plus élevée avec une fréquence plus faible pourrait indiquer un vieillissement de l'effectif et la possible apparition de pathologies[4] en lien ou non avec l'environnement professionnel. Ces assertions doivent évidemment faire l'objet à chaque fois d'une vérification à l'aide d'une étude statistique de la population étudiée ainsi que des conditions de travail pour comprendre davantage les raisons de l'absentéisme.

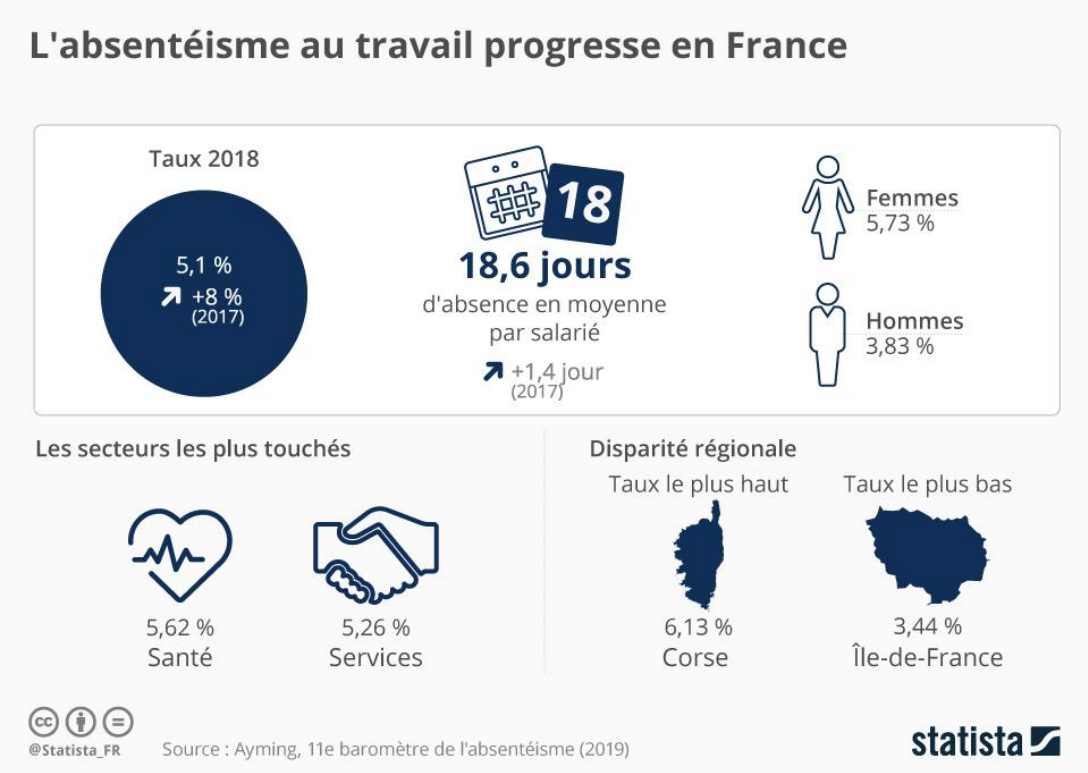
L'ensemble de ces indicateurs sont généralement utilisés lors d'études sur le niveau d'absentéisme global en France. Ces études permettent de connaître le niveau d'absentéisme dans le pays mais également de remarquer les points de vigilance à avoir lors de l'étude d'une entreprise en particulier. Ces études doivent être poursuivies au cours du temps afin de prendre en compte leurs évolutions, l'absentéisme étant un risque fluctuant.

[4] En raison de la réglementation RGPD (cf. partie 1.3.5), ces données ne sont évidemment pas étudiées et ne sont pas connues

1.1.3 Statistiques sur l'absentéisme en France

Connaître l'état de santé de la France par rapport à l'absentéisme, c'est étudier ce phénomène selon deux angles bien distincts. L'absentéisme touche en premier les entreprises par les absences des salariés. Ces absences ont donc des conséquences sur la productivité, la croissance économique et la bonne santé de l'ensemble des entreprises. Cependant, l'absentéisme touche également l'Assurance Maladie. Etudier l'état de santé de la France en termes d'absentéisme, revient donc également à étudier l'impact sur le système de l'Assurance Maladie.

Figure 1 : Statistiques principales du 11ème baromètre de l'absentéisme (Ayming – AG2R LM) [4]



Sur le plan des entreprises et ce depuis ces dernières années, **le taux d'absentéisme en France poursuit son augmentation dans l'ensemble des secteurs d'activité.** Une donnée assez préoccupante quand l'étude de l'Institut Sapiens [5] montre que l'absentéisme **coûte environ 108 milliards d'euros par an en 2018** aux entreprises, à l'Etat et impactant ainsi la croissance française.

Si le taux global d'absentéisme en France était de 4,59% en 2016, celui-ci a augmenté jusqu'à atteindre 5,11% en 2019. La gravité des absences a également augmenté passant à 18,7 jours d'absence par an et par salarié tout secteur confondu en 2019 contre 17,2 jours en 2017. Ces deux premiers indicateurs sont déjà révélateurs d'une dégradation de l'absentéisme sur l'ensemble du territoire.

Le secteur le plus touché par ces problèmes d'absentéisme est le secteur de la santé avec un taux de 5,62% en 2018, expliqué par les fortes contraintes à la fois physiques et psychiques endurées par les salariés de ce secteur. Même le secteur d'activité des services qui avait vu une baisse de son taux d'absentéisme aux alentours de 2017, a vu son taux augmenter l'année d'après. En revanche des secteurs d'activité répondant à des attentes physiques tels que l'industrie et le BTP détiennent un taux d'absentéisme inférieur malgré la pénibilité du travail demandé, ce taux étant de 4,26%. L'absentéisme semble donc avoir des causes plus complexes, outre le fait de la pénibilité du travail. Un constat qui va certainement s'accentuer avec l'arrivée de la COVID-19 en raison du contexte sanitaire, économique et financier.

Des différences se font sentir entre les régions mais également entre les genres. Le taux d'absentéisme est en augmentation sur l'ensemble du territoire français à l'exception de la Bretagne passant de 5,14% à 4,28% de 2017 à 2018 et de la Corse de 6,99% à 6,13%. Toutefois la Corse reste la région la plus touchée par ce phénomène, à l'inverse de l'Ile-de-France qui détient le taux le plus bas avec 3,44% en 2018. L'absentéisme touche différemment les hommes et les femmes, avec 3,83% pour les hommes contre 5,73% pour les femmes ce qui peut éventuellement être expliqué à la fois par des statuts plus précaires à savoir des temps partiels, des postes générateurs de problèmes de santé et des arrêts maladie liés à la grossesse. De plus un phénomène sociétal toujours d'actualité fait que les femmes continuent de gérer plus largement les charges domestiques.

L'âge reste tout de même une variable discriminante du taux d'absentéisme, avec une corrélation positive entre ces deux données. Le taux chez les moins de 25 ans est de 2,48% contre 7,40% pour les salariés de plus de 40 ans en moyenne. Cela s'explique par le fait que les personnes les plus âgées ont des arrêts plus longs.

Ces dernières années, une **nette dégradation de l'absentéisme a été observée pour les arrêts de plus de 90 jours avec une augmentation de 23% chez les salariés de moins de 40 ans**, contre 9% pour les salariés de plus de 40 ans. L'absentéisme est un risque préoccupant touchant également les populations les plus jeunes, à savoir en-dessous de 30 ans.

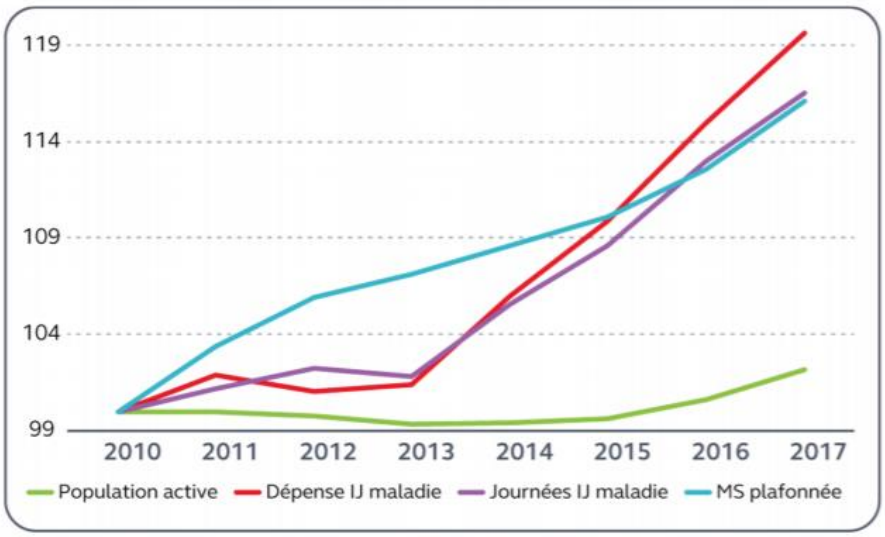
Or ces absences de longues durées ne sont pas assez prises en compte par les entreprises, que ce soit sur le pilotage du risque ou la réinsertion des individus après ces longues absences. Elles doivent alors agir, au risque de voir ce phénomène s'accentuer dans les prochaines années.

Au niveau de l'Assurance Maladie, chargée de verser les indemnités journalières aux salariés dont l'état de santé ne permet pas de continuer le travail, son rapport avec l'absentéisme découle irrémédiablement du bilan fait précédemment sur l'absentéisme au sein des entreprises. En effet, depuis 2013, **l'équilibre financier de l'Assurance Maladie s'est dégradé** du fait de l'augmentation des dépenses d'indemnisation des arrêts maladie. En 2017, le régime général a indemnisé près de 6,97 millions d'arrêts maladie, 1,05 millions d'arrêts pour accident du travail ou maladie professionnelle (AT-MP) et 635 000 arrêts pour maternité représentant **12,9 milliards d'euros d'indemnités au total pour ces arrêts.**

Entre 2013 et 2017, les indemnités versées par l'Assurance Maladie pour AT-MP ont augmenté de 3,8% par an en moyenne, à l'instar des indemnités pour maladie qui ont augmenté de 4,2% par an en moyenne. Cette augmentation des indemnités est bien plus rapide que celle prévue par l'objectif national de dépenses d'assurance maladie (ONDAM), de 2,1% par an en moyenne pour les indemnités maladie.

Dans la même dynamique, depuis 2013, **l'augmentation des dépenses en indemnités journalières maladie dépasse celle de la masse salariale plafonnée à 1,8 SMIC**. De plus, à la vue de la faible augmentation de la population active entre 2010 et 2017, il est alors important de constater que **l'augmentation des dépenses est due à la dégradation de deux facteurs de l'absentéisme : la durée des absences et leur fréquence**.

Figure 2 : Evolution des dépenses d'indemnités journalières maladie, du nombre de journées indemnisées et de la masse salariale plafonnée à 1,8 SMIC (2010-2017, base 100 en 2010)

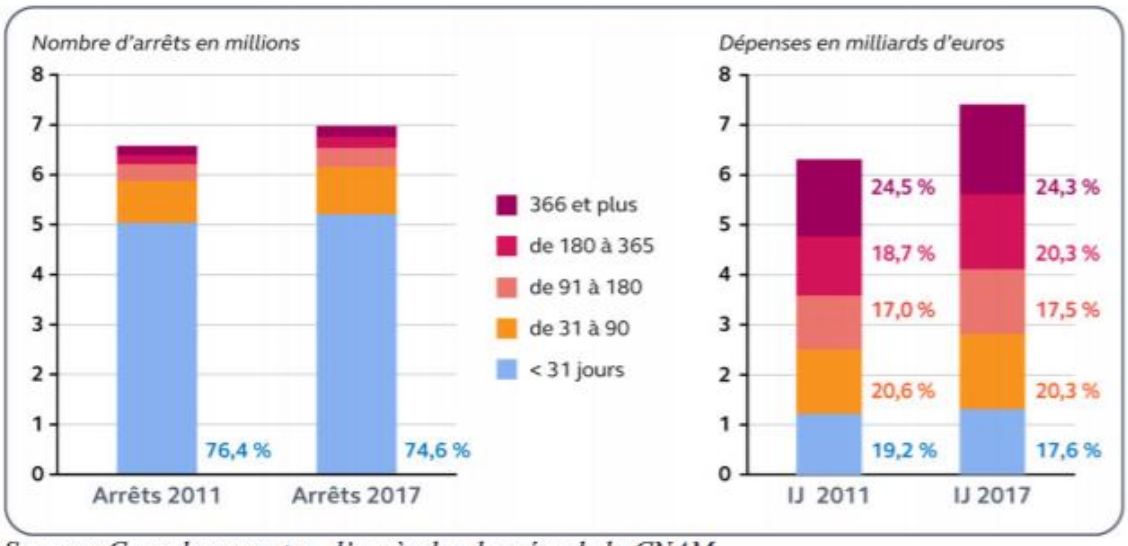


Source : Cour des comptes, d'après des données de l'Insee, de la DSS et de la CNAM.

La durée moyenne d'indemnisation pour l'ensemble des arrêts déclarés par un salarié a effectivement augmenté de 9% entre 2013 et 2017 pour atteindre une moyenne de 47,4 jours (à savoir le nombre de journées d'arrêt divisé par le nombre d'arrêts calculé sur une année). **La durée moyenne de chaque arrêt, elle, s'est accentuée de 7,4% pour atteindre 33,5 jours.** Sur cette même période la fréquence moyenne d'un individu est passée de 1,39 à 1,42. La proportion d'absence a également augmenté puisque la proportion d'assurés a évolué de 26,7% en 2013 pour 27,1% en 2017. L'ensemble de ces indicateurs montrent ainsi une dégradation de l'absentéisme : **les arrêts de travail durent plus longtemps, sont indemnisés plus longtemps et sont plus fréquents**. Ces dégradations expliquent donc le déclin de l'équilibre financier de l'assurance maladie qui doit continuer à verser des indemnités journalières à une population toujours plus touchée et absente de la société.

Cet allongement des durées d'absences concentre ainsi de plus en plus les dépenses de l'assurance maladie **vers des absences longues de plus de 90 jours**. Or, si les absences de courte durée restent les plus fréquentes, ce sont bien **les arrêts de longue durée qui coûtent le plus**, avec par exemple 24% des dépenses (soit 1,8 milliard d'euros) uniquement pour les arrêts de plus de 1 an. Cette concentration des dépenses qui tend de plus en plus vers les absences de longue durée année après année peut être visionnée à l'aide des données de la Caisse Nationale de l'Assurance Maladie entre 2011 et 2017 :

Figure 3 : Evolution entre 2011 et 2017 du nombre d'arrêts et des IJ versées en fonction de la durée des arrêts



Source : Cour des comptes, d'après des données de la CNAM.

L'évolution négative du bilan financier de l'Assurance Maladie, mais également de l'absentéisme de façon générale en France, que ce soit en fréquence, gravité ou exposition doit alerter. Ces statistiques montrent que l'absentéisme est **un phénomène qui s'accélère** et ayant des conséquences non négligeables sur l'impact financier à la fois des entreprises mais également des grandes institutions étatiques.

De plus, l'absentéisme et l'allongement des durées d'absence peut également être l'indicateur d'un **mal-être croissant des salariés** au sein des entreprises. En 2016**, les traitements psychotropes ainsi que les maladies psychiatriques étaient classés en 2ème pathologie** [6] derrière les hospitalisations ponctuelles en termes de dépenses d'indemnités journalières maladie avec près de 22,9% des dépenses totales pour les arrêts de plus de 180 jours. Comme vu précédemment, **les arrêts de longues durées sont les plus coûteux** à l'assurance maladie, et pourtant **certaines de ces absences pourraient être prises en charge** pour en diminuer le nombre ou la durée. En effet, même si les raisons de l'absentéisme sont multiples, les absences de longue durée sont souvent liées aux conditions de travail. Des solutions sont alors possibles afin de contenir cet absentéisme au travers de la prévention.

1.1.4 Traiter le problème de l'absentéisme

Bien que la gestion du risque absentéisme semble complexe à première vue par la multiplicité des causes, certaines actions sont possibles afin de prévenir ou diminuer ces absences au sein de l'entreprise. Dans la stratégie visant à diminuer le nombre d'arrêts ainsi que les coûts engendrés, l'absentéisme peut être divisé en deux catégories :

- L'absentéisme **incompressible**
- L'absentéisme **compressible**

L'absentéisme incompressible fait référence à l'ensemble des absences qui ne peuvent être réduites, ou du moins dont les mesures de prévention seraient beaucoup trop complexes à mettre en place, relevant davantage d'enjeux de santé publique. Elles caractérisent généralement des absences récurrentes sur lesquelles l'employeur ne peut agir directement. Les épidémies saisonnières ou encore les congés maternité sont généralement pris en exemple dans la définition de cet absentéisme incompressible qui est souvent la conséquence de causes exogènes. Néanmoins, pour les absences prévisibles tels que les congés maternité, des actions en prévision de futurs remplacements ou de la gestion des équipes peuvent être mises en place pour réduire les coûts futurs. De même, pour les absences a priori imprévisibles comme les arrêts dus aux épidémies saisonnières, des modèles prédictifs basés sur ces périodes cycliques peuvent tout de même être mis en place, malgré l'incertitude existant autour de ces phénomènes.

A l'inverse, les absences liées à la charge de travail, à l'épuisement dû au travail intensif ou aux potentiels risques que le salarié peut rencontrer sur son lieu de travail, sont des absences faisant partie de ce qu'on appelle **l'absentéisme compressible**. En mettant en place certains leviers ou mesures de prévention, cet absentéisme pourrait être évité en particulier pour les motifs suivants, représentant environ 52% des arrêts :

- Maladie professionnelle
- Accident de travail
- Accident de trajet
- Absences injustifiées

En négligeant ce problème, les salariés victimes de ces absences s'éloignent de plus en plus de leur entreprise et de ce fait la réinsertion devient de plus en plus compliquée, ce qui engendre davantage d'absences. Un vrai travail d'étude de l'absentéisme au sein même de l'entreprise doit alors être effectué afin de comprendre, agir et prévenir.

En premier lieu, il est primordial de mener un diagnostic face au risque absentéisme rencontré :

1. **Analyser les tendances générales d'évolution de l'absentéisme** pour voir si l'entreprise est confrontée à une augmentation de son taux d'absentéisme et avec quelle rapidité.
2. **Etudier les différents indicateurs de l'absentéisme (exposition, fréquence, gravité)** afin de comprendre le profil d'absentéisme rencontré dans l'entreprise.
3. Etudier **les secteurs d'activités et le type de population présents dans l'entreprise** afin de comprendre les corrélations entre conditions de travail, individus et causes de l'absentéisme.

Le diagnostic ayant été fait sur l'entreprise concernée, il est primordial de revenir sur les causes de l'absentéisme. Celles-ci proviennent généralement de trois cadres distincts mais pas nécessairement indépendants :

- **Les caractéristiques du travail**
- **Le contexte socio-organisationnel**
- **La vie hors du travail et les caractéristiques socio-démographiques**

Si les absences sont causées par **les conditions de travail**, l'absentéisme est généralement compressible. En effet, les absences sont alors causées par des contraintes physiques causant soit des accidents du travail, des arrêts maladie ou une usure professionnelle liés aux contraintes du travail. Des plans d'action de prévention sont possibles afin d'améliorer la qualité de travail des salariés en diminuant la pénibilité de celui-ci.

Le contexte socio-organisationnel fait référence non pas au travail en lui-même mais à la situation économique de l'entreprise, aux relations internes et au mode de management appliqué au sein de l'équipe. En raison d'une mauvaise gestion de l'un de ces paramètres, les salariés peuvent de ce fait ressentir un sentiment de retrait ou de démotivation ce qui engendre généralement des arrêts courts et répétés dans le temps. Selon l'analyse effectuée par l'Institut de socio-économie des entreprises et des organisations (ISEOR), **99% des cas d'absentéisme compressible ont pour origine la méthode de management.** Les causes les plus répandues au travers du management sont les suivants :

- **Des conditions psychologiques dégradées** par un manque de considération en lien avec le travail demandé
- **Une mauvaise définition du travail** à effectuer impactant les motivations du salarié
- **Une communication absente** ou trop peu constructive
- **L'absence de formations** professionnelles adaptées aux réalités du métier

- **Une mauvaise gestion du temps**
- **Une méconnaissance de la stratégie** adoptée par l'équipe ou l'entreprise

Il est important en matière de prévention d'améliorer les relations au sein de l'équipe, proposer des dialogues, des formations d'encadrement, une politique de rémunération stimulante mais également de présenter de façon régulière les perspectives de l'entreprise afin que le salarié puisse se sentir concerné par un projet commun.

Dernière catégorie, **la plus exogène au monde de l'entreprise, la vie hors-travail** et les caractéristiques socio-démographiques. Ce point fait allusion aux difficultés personnelles que le salarié peut rencontrer tous les jours, dans sa vie de famille ou face à sa vie sociale ou économique. Ces absences sont ainsi plus compliquées à gérer par la nature confidentielle des raisons qui poussent à être absent. Une possibilité est d'aménager les horaires de travail afin que le salarié puisse se sentir plus à l'aise entre sa vie privée et professionnelle.

Ces quelques actions peuvent permettre de faire diminuer le taux d'absentéisme en mettant en place des mesures de prévention qui viennent limiter les absences avant que celles-ci ne se produisent. **Si les mesures de prévention permettent de réduire l'absentéisme, les mesures de réinsertion au travail après de longues absences favorisent également cette réduction.** Elles motivent le salarié à ne pas se sentir à l'écart ou en retrait des perspectives de l'entreprise. L'absentéisme appelle l'absentéisme. De ce fait, un entretien avec le manager, une adaptation du poste de travail, des aménagements horaires à son retour peuvent permettre au salarié de se sentir soutenu après son absence, permettant ainsi de diminuer encore une fois l'absentéisme.

Enfin, si la plupart des entreprises ne se sentent pas concernées par l'absentéisme au travail considérant que les absences font parties intégrantes de la vie professionnelle, certaines en revanche tentent de plus en plus de se prémunir face à ce risque. Elles mettent en place **un suivi dynamique de leur taux d'absentéisme** afin de le comprendre et de mettre en place des mesures de prévention adaptées.

L'absentéisme et ses conséquences sont donc au cœur des nouvelles préoccupations des employeurs, des assureurs mais également du système de protection sociale français : trois acteurs dont il est nécessaire de présenter, faisant face à un risque grandissant et de plus en plus coûteux.

1.2. Le système de protection sociale français face à l'absentéisme

1.2.1 Le régime de base

a. Les débuts de la Sécurité Sociale

Si les premières vraies actions de sécurité sociale apparaissent dès le 17ème siècle avec la création du premier « régime de retraite » pour les marins sous Colbert, ce n'est qu'au milieu de

la moitié du 20ème siècle que le terme de Sécurité Sociale apparait. Auparavant, plusieurs initiatives en matière de sécurité sociale avaient fait leur apparition en France.

Ce sont finalement les ordonnances du 4 et 19 octobre 1945 promulguées par le gouvernement du Général de Gaulle qui finalisent la création d'un système de sécurité sociale en France. Ce système se base sur le modèle « bismarckien », où la gestion est tenue par des partenaires sociaux et dont le financement se fait par le biais de cotisations à la charge des employeurs et des salariés. Les principaux objectifs sont d'obtenir un unique régime couvrant la majorité de la population française tout en couvrant davantage de risques. Avec l'ordonnance du 4 octobre 1945, le système de sécurité sociale en France devient un réseau coordonné de caisses, substituant les multiples organismes préexistants. Cependant certains régimes spéciaux refusent de s'intégrer à ce nouveau régime. C'est le cas par exemple pour les cheminots et les marins, restant dans un cadre « transitoire », toujours d'actualité. Il faudra attendre 1946 pour reconnaître le droit à la protection de la santé, la sécurité matérielle, le repos et les loisirs pour tous dans la Constitution de la IVème République.

Tout au long de la seconde moitié du 20ème siècle et jusqu'au début du 21ème siècle, d'autres améliorations au nom de la Sécurité Sociale verront le jour, toujours dans l'optique de protéger la population. Ce système est alors divisé en différents régimes et branches afin d'encadrer et structurer les aides apportées aux Français face aux aléas de la vie.

b. Les régimes

Trois grands régimes sont créés à la suite des ordonnances de 1945 :

- **Le régime général**, propre aux salariés et travailleurs assimilés à des salariés, concernant environ 80% de la population.
- **Le régime des travailleurs non-salariés non agricoles**, propre aux artisans, commerçants et professions libérales.
- **Le régime agricole** : propre aux exploitants et salariés agricoles, ainsi que certains secteurs rattachés à l'agriculture.

Cependant certains régimes appelés **régimes spéciaux** se distinguent des trois grands régimes principaux mentionnés ci-dessus. De manière non exhaustive, ces principaux régimes spéciaux sont :

- Caisse de prévoyance et de retraite SNCF (CPR SNCF)
- Caisse nationale militaire de sécurité sociale (CNMSS)
- Caisse des français à l'étranger (CFE)
- Régime Alsace-Moselle

Le régime Alsace-Moselle est un régime local. Sa particularité se fonde sur un critère géographique, et non selon un critère de secteur d'activité comparativement aux autres régimes

spéciaux. Environ 1,4 millions de cotisants et 2,2 millions de bénéficiaires relèvent de ce régime complémentaire et obligatoire. Les taux de remboursement appliqués sont différents en fonction du régime auquel l'individu est rattaché.

Le financement de la protection sociale était principalement réglé par des cotisations. A partir des années 1990, ce financement se diversifie avec la loi de finance de 1991 instaurant la Contribution Sociale Généralisée (CSG), imposée sur tous les revenus des contribuables (salaires, pensions, allocations de chômage…) et d'autres impôts.

Aujourd'hui ces ressources peuvent être réparties en trois catégories :

- **Les cotisations sociales** représentant environ **60% des recettes**. Ces cotisations sont alignées sur le travail du salarié au travers de la part salariale et de l'employeur, et sur les revenus de toute nature.
- **La CSG** représentant environ **20% des recettes.**
- **Les autres impôts et taxes** représentent environ **13% des recettes**. Elles correspondent aux divers prélèvements de nature fiscale, contributions et taxes affectées au financement de la Sécurité Sociale

D'autres ressources participent au financement provenant d'autres organismes comme les fonds de solidarité vieillesse ou d'autres régimes. Ces divers financements permettent de mettre en place des structures spécialisées dans des domaines d'intervention spécifiques.

c. Structure et domaines d'intervention

Bien que la Sécurité Sociale soit un système unique, celle-ci est divisée en un ensemble d'Institutions qui ont pour fonction principale de protéger les individus des conséquences d'évènements en lien avec des risques sociaux. Les prestations de ces Institutions sont donc de nature financière à l'intention des bénéficiaires qui potentiellement seraient confrontés à l'un de ces risques généralement coûteux.

Cette division a été instaurée par l'ordonnance de 1967, conduisant à la séparation de la Sécurité Sociale en **5 branches principales autonomes**. Chaque branche est alors responsable de ses ressources et de ses dépenses selon leur domaine d'intervention.

- La branche Maladie gérée par l'Assurance Maladie
- La branche Accident du travail et Maladies professionnelles gérée par l'Assurance Maladie
- La branche Retraite gérée par l'Assurance Retraite
- La branche Famille gérée par les Allocations familiales
- La branche Recouvrement gérée par les URSSAF

Depuis l'adoption du projet de loi de financement de la Sécurité Sociale pour 2021 du 30 novembre 2020, une dernière branche est venue s'ajouter aux différentes branches préexistantes : la branche Dépendance.

La branche Maladie

La branche Maladie est en charge de tous les risques liés à la santé des assurés, elle garantit également l'accès aux soins. Celle-ci a également pour rôle de mettre en place des mesures de prévention et d'améliorer l'accès aux soins, a fortiori pour les plus démunis. Ces actions de prévention s'inscrivent dans une logique de minimisation des dépenses tout en améliorant l'état de santé de la population française. Les différents risques couverts par la branche Maladie sont les suivants :

- Maladie
- Maternité
- Invalidité
- Décès

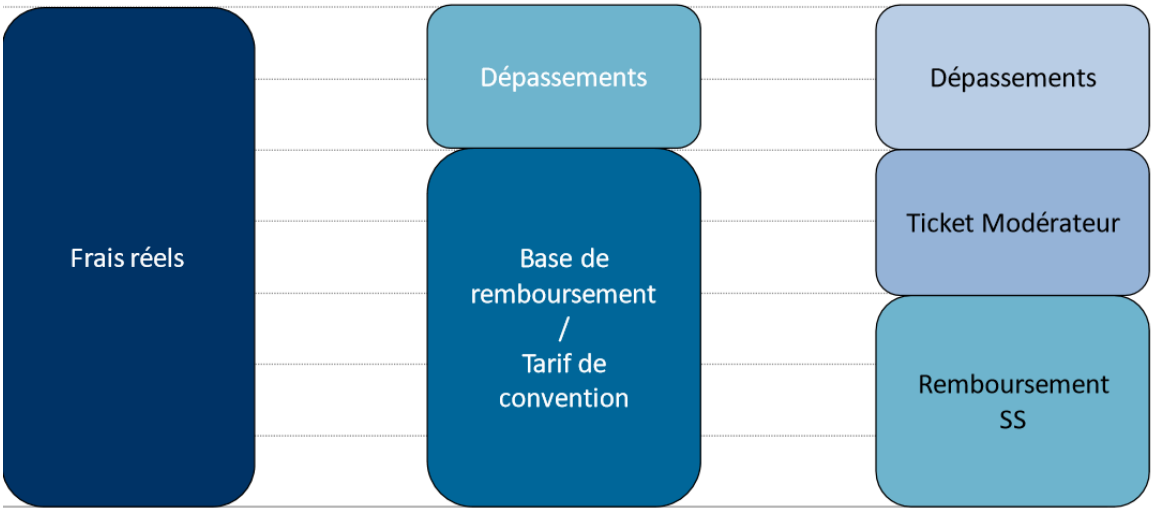
Au niveau structurel, la branche Maladie est gérée par la Caisse Nationale de l'Assurance Maladie (CNAM). Celle-ci se décompose ensuite en un réseau divisé en différents organismes. Les caisses primaires d'assurance maladie (CPAM), les caisses générales de sécurité sociale en Outre-Mer (CGSS), les directions régionales du service médical (DRSM), les caisses d'assurance retraite et de la santé au travail (CARSAT) et les unions de gestion des établissements de caisse d'assurance maladie (UGECAM). Ces diverses caisses sont réparties sur l'ensemble du territoire afin de pouvoir donner à la population française une couverture face à l'ensemble des risques relatifs à la santé dont ils peuvent être confrontés. Des prestations sont alors accordées en cas d'évènements coûteux envers les bénéficiaires. En assurance maladie, deux types de prestations peuvent être allouées :

- Les **prestations en nature** (art. L.322-1 et suivants du Code de la Sécurité Sociale), qui correspondent aux remboursements de frais de santé si ces actes de santé figurent dans la liste des actes santé ou médicaments remboursables et si ces soins sont dispensés par un établissement ou un praticien habilité. En règle générale, le bénéficiaire doit d'abord avancer ces frais avant de pouvoir toucher ces remboursements, le terme de « tiers-payant » est employé dans le cas contraire. Les frais de santé peuvent alors être décomposés comme suit :

  Les frais réels correspondent au montant total de la prestation. Ces frais réels sont ensuite divisés en deux parties, la base de remboursement (BR) prise en compte dans le calcul du remboursement de la Sécurité Sociale et les dépassements non pris en charge par la Sécurité Sociale. A partir de la base de remboursement, un pourcentage est appliqué afin de calculer le remboursement réel de la Sécurité Sociale, le reste étant

à la charge du bénéficiaire sous le nom de Ticket Modérateur sauf si le bénéficiaire est en Affection de Longue Durée (ALD). Ce Ticket Modérateur peut alors être pris en charge par une complémentaire.

Figure 4 : Décomposition des prestations en nature avec le remboursement de la Sécurité Sociale



Exemple sur une consultation d'un généraliste conventionné :

- **La consultation coûte 25€,** correspondant aux frais réels
- La base de remboursement est de 25€, **pris en charge à 70%** par la Sécurité Sociale moins le forfait de **1 euro payé par l'assuré au titre de la participation forfaitaire.** Cette participation forfaitaire ne peut être prise en charge par une complémentaire. Ainsi le montant remboursé vaut : (25*0,7) - 1=16,50€
- Le **reste à charge** est donc de 25-16.50=8,50€ dont 7,50€ de **ticket modérateur**

- Les **prestations en espèce** (art. L.323-1 et suivants du Code de la Sécurité Sociale), compensant la perte d'un revenu pour les personnes devant cesser leur activité professionnelle pour raison de santé. Ces revenus de remplacement correspondent aux indemnités journalières en cas d'arrêt maladie ou durant les congés maternité et paternité, ou à une pension d'invalidité, lorsque l'assuré présente une invalidité réduisant sa capacité de travail ou de gains.

Toute personne résidant en France de manière stable et régulière est soit affiliée à un régime en qualité d'assuré ou d'ayant droit, soit peut bénéficier de la Complémentaire Santé Solidaire (CSS) anciennement appelée Couverture Maladie Universelle (CMU).

La branche Accidents du travail et Maladies professionnelles

La branche Accidents du travail et Maladies professionnelles (AT-MP) est en charge des risques rencontrés dans le monde du travail, que ce soient **les accidents de travail, les accidents de trajet ou les maladies professionnelles**. Son rôle est donc de pouvoir à la fois indemniser les victimes de ces risques, mettre en place une politique de prévention et fixer les contributions au bon financement de ce système.

Au niveau structurel, la branche AT-MP est gérée par la Caisse Nationale de l'Assurance Maladie (CNAM) et les Caisses d'Assurance Retraite et de la Santé du Travail (CARSAT). Comme dit précédemment, ces organismes indemnisent les victimes pour les risques suivants :

- **Accident du travail**
- **Accident de trajet**
- **Maladie professionnelle**

Sont donc bénéficiaires de ces indemnités toute personne relevant de la législation des accidents du travail selon l'article L.411-1 du Code de la Sécurité Sociale, « toute personne salariée ou travaillant à quelque titre ou en quelque lieu que ce soit, pour un ou plusieurs employeurs ou chefs d'entreprise ». À la suite d'un accident ou d'une maladie professionnelle, la victime a alors la garantie d'avoir une prise en charge totale de la part de la Sécurité Sociale pour les soins et actions de rééducation fonctionnelle mais également de la réinsertion professionnelle. Si ces risques engendrent des arrêts de travail, alors des indemnités journalières seront versés afin de compenser les pertes de revenus. Enfin si l'incapacité de travailler dépasse une durée de 3 ans ou s'il y a réduction définitive de la capacité de travail approuvée par l'avis du contrôle médical de la Sécurité Sociale, alors la victime pourra recevoir une rente d'invalidité (taux d'incapacité permanente > 10%) ou un capital (taux < 10%) proportionnellement au degré d'invalidité. Enfin en cas de décès, les ayant-droits toucheront une rente dont le montant total des rentes versées est plafonné à 85 % du salaire annuel de l'assuré décédé.

Les descriptions de la branche Maladie ainsi que celle de la branche Accidents du travail et Maladies professionnelles montrent que la prise en charge de l'absentéisme par la Sécurité Sociale se fait par le biais **des indemnités journalières**. Cependant, des conditions relatives à l'ouverture des droits, au calcul et au montant maximal des indemnités journalières existent en fonction du type d'arrêt.

Les conditions d'ouverture des droits aux indemnités journalières pour les causes maladie, maternité et AT-MP sont résumées dans le tableau suivant :

Tableau 1 : Conditions d'ouverture aux droits des indemnités journalières

Cause de l'arrêt de travail	Conditions d'ouverture des droits
Maladie	**<u>Pour les arrêts inférieurs à une durée de 6 mois</u>** Avoir travaillé au moins 150 heures de travail dans les 90 jours avant l'arrêt de travail ou Avoir cotisé sur un salaire supérieur ou égal à 1 015 SMIC horaires dans les 6 derniers mois civils avant l'arrêt de travail **<u>Pour les arrêts supérieurs ou égaux à 6 mois</u>** Reconnaître au minimum 12 mois d'immatriculation à la Sécurité Sociale à la date d'arrêt et • au moins 600 heures de travail dans les 365 jours précédents ou • avoir cotisé sur un salaire supérieur ou égal à 2 030 SMIC horaires pendant les 6 derniers mois civils avant l'arrêt de travail
Maternité	Reconnaître au minimum 10 mois d'immatriculation à la Sécurité Sociale à la date de l'arrêt et • au moins 150 heures de travail pendant les 90 derniers jours ou • avoir cotisé sur un salaire supérieur ou égal à 1 015 SMIC horaires pendant les 6 derniers mois civils
	La cause de l'arrêt étant liée au travail, aucune condition n'est demandée pour justifier une

AT-MP	ouverture de droits aux indemnités journalières

Si le salarié remplit les conditions d'ouverture des droits, celui-ci pourra obtenir des indemnités journalières (IJ) à la suite de son absence. Ces indemnités sont calculées sur une base de calcul qui diffère entre les arrêts maladies ou maternité et les arrêts AT-MP. Les IJ pour maladie et maternité sont calculées sur la base du salaire brut des 3 mois précédant l'arrêt ou des 12 mois précédents si l'activité est discontinue. Dans le cas des AT-MP, les IJ sont calculées sur la salaire brut du mois précédant l'arrêt.

Enfin, **les modes de calcul ainsi que les montants limites des IJ** pour chaque type d'arrêt sont répertoriés dans le tableau suivant :

Tableau 2 : Calculs et montants limites des IJ

Cause de l'arrêt de travail	Calcul de l'indemnité	Montant maximal journalier en (2020 et 2021)
Maladie	50% du salaire brut plafonné à 1,8 SMIC (la majoration de 66,6% pour au moins 3 enfants à charge n'existant plus)	Maximum : 45,55€ en 2020 46,00€ en 2021
Maternité	Salaire journalier de base dans la limite du plafond de la Sécurité Sociale diminué d'un taux forfaitaire de 21%	Minimum : 9,53€ Maximum : 89,03€
AT-MP	**Durant les 28 premiers jours :** 60% du salaire journalier de base plafonné à 0,834% du plafond annuel de la Sécurité Sociale diminué du taux forfaitaire de 21% **Les jours suivants :** Le taux de 60% passe à 80%	**Durant les 28 premiers jours :** 205,84€ **Les jours suivants :** 274,46€

Quelques éléments à ajouter à la suite de ces informations. Les IJ pour maternité sont versées également les samedis, dimanches et jours fériés. Les IJ pour AT-MP sont versées jusqu'à la guérison ou la consolidation de l'état de santé du salarié. La Sécurité Sociale indemnise les arrêts maladie après une période de carence de 3 jours. Ceux-ci sont également indemnisés pendant 360 jours au plus par période glissante de trois. En revanche, les arrêts en affection de longue durée (ALD) sont indemnisés pendant une durée maximale de 3 ans. Si une autre ALD est détectée ou que le salarié a repris le travail pendant au moins un an, celui-ci pourra bénéficier à nouveau d'une durée maximale d'indemnisation de 3 ans.

Malgré ce système d'indemnisations journalières qui permet aux salariés en incapacité d'obtenir des revenus pour compenser l'absence au travail, ces IJ peuvent paraître insuffisantes. De ce fait les régimes complémentaires ainsi que les Employeurs permettent d'obtenir un complément d'indemnisation durant ces périodes d'absence.

1.2.2 Les régimes complémentaires

a. Le marché de l'assurance complémentaire

Différents acteurs interviennent sur le marché de l'assurance complémentaire, à savoir les organismes d'assurance, les courtiers et les gestionnaires. Ces différents acteurs travaillent la plupart du temps en collaboration afin de pouvoir proposer de nouveaux produits d'assurance répondant aux demandes des assurés face à des évènements aléatoires de la vie généralement coûteux.

Les organismes assureurs

Les organismes assureurs sont décomposés en trois grandes familles :

- **Les Mutuelles**
- **Les Institutions de Prévoyance**
- **Les Sociétés d'assurance**

Si ces trois familles ont une réglementation commune en matière technique et financière, ces différents organismes possèdent toutefois **leur propre fonctionnement et un code qui leur est propre**. A noter que la grande différence entre ces trois types d'organisme réside dans les domaines où elles interviennent de manière privilégiée.

Régies par le code de la Mutualité, **les Mutuelles** sont des groupements à but non lucratif. Leur domaine d'intervention privilégié réside dans la couverture des Frais de Santé. Néanmoins depuis la généralisation de la complémentaire santé de 2016, les Mutuelles ont développé des offres pour diversifier leurs portefeuilles en Prévoyance lourde, à savoir le décès et arrêt de travail. Comme son nom l'indique, les Mutuelles utilisent la mutualisation des risques afin de pouvoir honorer leurs engagements auprès de leurs assurés. Elles sont contrôlées par les adhérents eux-mêmes, possédant chacun une voix dans le cadre des Assemblées Générales. Depuis ces dernières

années, le nombre de Mutuelles a considérablement baissé du fait des regroupements effectués entre-elles passant de 5 000 à 450 Mutuelles en l'espace de 10 ans.

Régies par le Livre IX du Code de la Sécurité Sociale, **les Institutions de Prévoyance** interviennent dans les offres de Prévoyance collective. Ce sont des personnes morales de droit privé, et tout comme les Mutuelles, à but non lucratif. Par l'article L931-1 du code de la Sécurité Sociale, ces institutions sont administrées paritairement par des membres adhérents et des membres participants.

Enfin, **les Sociétés d'Assurances** sont régies comme son nom l'indique par le code des assurances. Les Sociétés d'Assurances peuvent prendre deux types de formes juridiques, soit en étant une Société anonyme, soit en se désignant comme une Société d'assurance mutuelle gérée par leurs sociétaires par le biais d'élections revendiquant ainsi des valeurs mutualistes. A la différence des Mutuelles et des Institutions de Prévoyance, ces organismes peuvent intervenir dans tous les domaines de l'assurance, de la Prévoyance à la Santé en passant par la Retraite ou encore l'assurance de biens.

Les courtiers et comparateurs d'assurance

Les courtiers sont des travailleurs indépendants contrairement aux agents généraux qui eux sont tenus par un contrat pour le compte d'une compagnie d'assurance. Le courtier se place donc entre l'entreprise et les organismes assureurs afin de proposer une offre d'assurance au client-entreprise tout en étudiant l'ensemble des possibilités d'offre sur le marché. Le courtier est alors rémunéré par le biais d'une commission d'apport en pourcentage du montant des primes.

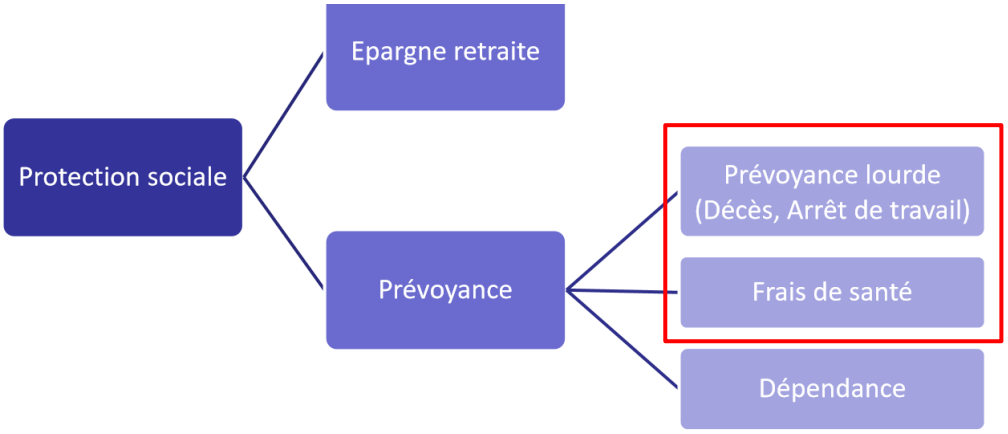
Les gestionnaires

Comme son nom l'indique, les gestionnaires sont présents afin de gérer administrativement les contrats d'assurance que ce soit en Prévoyance, Santé, Contrats Emprunteurs et autres. Leur objectif est donc de vérifier la bonne tenue du contrat, en termes de réception des primes versées par l'assuré et du paiement des diverses prestations promises par l'assureur. Il existe deux types de gestionnaires, les gestionnaires indépendants et les gestionnaires rattachés de près ou de loin à un cabinet de courtage. Contrairement au courtier, le gestionnaire n'est donc pas un intermédiaire du contrat mais bien une tierce-partie.

b. Les garanties de prévoyance complémentaire

Le périmètre de la Protection Sociale complémentaire peut être représenté par le schéma suivant :

Figure 5 : Périmètre de la protection sociale complémentaire



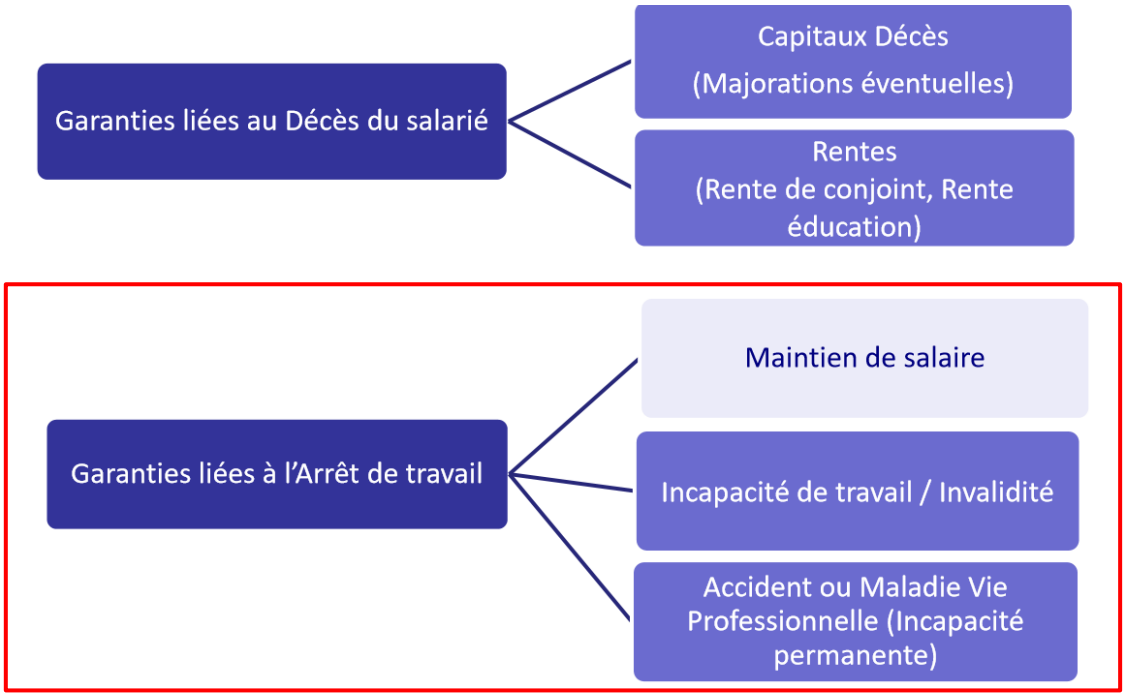
Ce mémoire traitant du risque absentéisme, seules les parties sur la santé et la prévoyance lourde seront développées ici.

Les frais de santé

Au niveau des frais de santé, la garantie complémentaire Santé intervient en complément de la Sécurité Sociale pour permettre à l'assuré d'avoir un reste à charge moindre ou encore de proposer un remboursement si la Sécurité Sociale ne prend pas en charge certains actes Santé. Plusieurs niveaux de garanties sont proposées dans les offres du marché allant de la garantie entrée de gamme avec prise en charge du Ticket Mordérateur jusqu'à la garantie la plus haute avec une prise en charge à 100% des Frais réels en complément du remboursement de la Sécurité Sociale.

Au niveau de la Prévoyance lourde, les diverses garanties peuvent être résumées comme suit :

Figure 6 : Les différentes garanties de la Prévoyance lourde



Incapacité de travail

La garantie incapacité de travail permet au salarié en arrêt de percevoir des indemnités journalières afin de compenser sa perte de salaire. Ces indemnités journalières viennent compléter celles de la Sécurité Sociale et le complément de revenu versé par l'employeur (obligation mise en place par la loi de mensualisation de 1978). Les trois principaux paramètres définissant une garantie incapacité sont la franchise, le montant et l'assiette de calcul. Il existe différents types de franchises :

- **La franchise ferme ou continue**, où l'arrêt de travail commence à être indemnisé après x jours d'absence continu
- **La franchise discontinue**, où l'arrêt de travail commence à être indemnisé dès que x jours d'absence sont comptabilisés dans l'année
- **La franchise rétroactive**, où l'arrêt de travail commence à être indemnisé après x jours d'absence continus. Les indemnités correspondant à la période de franchise seront alors versées rétroactivement à l'issu de cette période

Invalidité

La rente d'invalidité compense, en totalité ou en partie, la perte de revenu d'un assuré déclaré invalide. Elle vient compléter la pension d'invalidité versée par la Sécurité sociale. Le montant de cette rente d'invalidité peut être fixe, ou bien correspondre à un pourcentage du dernier salaire de l'assuré. Le plus souvent son montant dépend également des catégories d'invalidité fixées par la Sécurité sociale.

c. Les différents types de contrat de Prévoyance complémentaire sur le marché

Contrats uniques ou avec options

Il existe en matière de Prévoyance lourde, deux familles de régimes, les régimes uniques où les garanties sont identiques pour tous les individus, et les régimes à choix d'option. La grande différence entre ces deux types de contrat est que le régime unique couvre les assurés de façon identique s'ils appartiennent à une même catégorie, tandis que le régime avec choix d'option permet de choisir des niveaux de garanties différents, tout en garantissant une couverture actuariellement équivalente pour tous.

Contrats obligatoires et facultatifs

La plupart des régimes d'Entreprise sont des régimes collectifs à adhésion obligatoire. L'Entreprise peut toutefois prévoir des garanties optionnelles à adhésion facultative pour les salariés qui souhaiteraient bénéficier de garanties améliorées. Cependant, l'entreprise qui participerait au financement d'un tel régime, ne pourrait pas bénéficier des exonérations de charges sociales sur la cotisation employeur ni des avantages fiscaux prévus. De ce fait, les

entreprises ne participent généralement pas au financement de ces régimes à adhésion facultative. L'intérêt pour l'entreprise de prévoir ce type de régime, est de faire bénéficier ses salariés d'un contrat groupe, souscrit à priori dans des conditions financièrement avantageuses.

1.2.3 Les obligations de l'Employeur

Si le régime de base et les complémentaires permettent aux individus de faire face à des évènements risqués coûteux, l'employeur doit également respecter certaines obligations en termes de protection sociale. Ces obligations ont été ajoutées au fur et à mesure par le législateur afin de protéger les salariés.

La loi de mensualisation

Une de ces obligations, est le maintien de salaire par l'employeur d'après **la loi de mensualisation de 1978**, améliorée par l'Accord National Interprofessionnel du 11 janvier 2008. Pour honorer cet engagement, l'employeur peut utiliser sa propre trésorerie, confier cet engagement à un organisme assureur par le biais d'un contrat de mensualisation ou inclure cette prestation dans les prestations garanties par le contrat de Prévoyance.

Pour pouvoir bénéficier de cette protection, le salarié doit attester de minimum 1 an d'ancienneté dans l'entreprise. Si l'arrêt de travail n'est pas lié à un accident du travail ou à une maladie professionnelle, une franchise de 7 jours est alors appliquée. Généralement une partie de cette franchise est prise en charge par la Sécurité Sociale au bout de 3 jours de carences. A noter que certaines conventions collectives prennent en charge les jours de franchise dès le premier jour d'arrêt. Passé ce délai de 7 jours, l'Employeur complète les indemnités de la Sécurité Sociale à hauteur de 90% puis à hauteur de 66,66% de la rémunération brute que le salarié aurait dû obtenir s'il avait continué à travailler. Les durées de maintien du salaire par l'Employeur dépendent de l'ancienneté du salarié et sont répertoriées dans le tableau suivant :

Tableau 3 : Pourcentage et durée du maintien de salaire par l'Employeur en fonction de l'ancienneté du salarié

Ancienneté	90% du salaire brut	66,66% du salaire brut
1 à 6 ans	30 jours	30 jours
6 à 11 ans	40 jours	40 jours
11 à 16 ans	50 jours	50 jours
16 à 21 ans	60 jours	60 jours
21 à 26 ans	70 jours	70 jours
26 à 31 ans	80 jours	80 jours
Supérieure à 31 ans	90 jours (maximum)	90 jours (maximum)

La CCN des Cadres de 1947

D'autres mesures prises par le législateur permettent de protéger les salariés. Par exemple la Convention Collective Nationale des Cadres du 14 mars 1947 rend obligatoire le taux de cotisation versé par les employeurs comme le montre l'article suivant : « Les employeurs s'engagent à verser, pour tout bénéficiaire visé aux articles 4 et 4 bis de la Convention ou à l'annexe IV à cette Convention, une cotisation à leur charge exclusive, égale à 1,50 % de la tranche de rémunération inférieure au plafond fixé pour les cotisations de Sécurité sociale. » En cas de non-respect de cette obligation, l'Employeur pourra alors avoir une sanction de l'ordre de 3 PASS (Plafond Annuel de la Sécurité Sociale) à sa charge.

La loi Evin de 1989

La loi Evin du 31/12/1989 est la première loi spécifique en Prévoyance qui permet de garantir aux salariés une protection sociale au sein de l'entreprise. Par cette loi le législateur crée pour la première fois un ensemble de règles applicables à toutes les familles d'assureurs que ce soient les Mutuelles, les Institutions de Prévoyance ou les Compagnies. Les principales dispositions sont les suivantes :

- Obligation de prendre en charge les suites de pathologies antérieures
- Obligation par les assureurs de proposer la poursuite de la couverture santé aux anciens salariés
- Maintien du paiement des rentes au niveau atteint en cas de résiliation
- Interdiction à l'employeur d'imposer de façon unilatérale une quote-part de cotisation à un régime de prévoyance
- Obligation de remettre une notice résumant les garanties
- Consultation du Comité d'Entreprise obligatoire en ce qui concerne le régime de prévoyance
- Obligation de remettre un rapport sur les résultats du contrat mis en place

La loi du 8 août 1994

La loi du 8 août 1994 vient ensuite poursuivre les dispositions prises par la loi Evin. Originaire d'une directive européenne en matière d'assurance, cette loi crée à la fois le premier corps de règles relatives à la protection sociale des salariés mais également des règles de fonctionnement aux Institutions de Prévoyance dans un premier temps, qui seront ensuite appliquées aux Mutuelles en 2001. Les principales dispositions sont les suivantes :

- Création des fondements juridiques de toute couverture collective
- Obligation d'une période de réexamen maximale de 5 ans pour les conventions collectives de branche ou d'entreprise qui désignent un organisme assureur

- Poursuite de la revalorisation des prestations en cours de service en cas résiliation du contrat ainsi que le maintien de la garantie décès

Les conventions collectives peuvent alors être mises en place par le biais de 3 modalités, par convention ou accord collectif, par référendum ou par décision unilatérale du chef d'entreprise.

Continuité de la loi Evin

La loi du 17 juillet 2001 vient compléter la loi Evin de 1989 en obligeant les assureurs à maintenir aux salariés en incapacité de travail ou en invalidité et percevant des prestations de la Sécurité Sociale le bénéfice de la garantie décès au-delà de la résiliation du contrat par l'entreprise.

L'Accord National Interprofessionnel

Enfin, par l'Accord National Interprofessionnel (ANI) du 11 janvier 2008, le législateur accorde aux salariés la portabilité des droits permettant ainsi aux salariés de conserver le bénéfice des garanties prévues au contrat collectif en vigueur dans l'ancienne entreprise que ce soit en Frais de Santé ou en Prévoyance. Ces droits sont également applicables aux ayant-droits du salarié bénéficiant de ces garanties. Cependant quelques conditions doivent être remplies :

- Être victime d'une rupture de contrat de travail qui ne soit pas la conséquence d'une faute lourde
- Avoir droit à la prise en charge par l'assurance chômage du fait de cette rupture
- Avoir ouvert des droits à une couverture complémentaire chez le dernier Employeur

La durée du maintien des garanties complémentaires pendant la période de chômage est alors égale à la durée du dernier contrat de travail, ou le cas échéant, des derniers contrats de travail lorsqu'ils sont consécutifs chez le même employeur. Cette durée exprimée en mois ne peut alors dépasser la limite de 12 mois. Au-delà de cette période, le maintien des garanties cesse. Le maintien des garanties peut également cesser lors de la reprise d'un nouvel emploi par le salarié ou si celui-ci fait valoir son droit de renonciation. A noter que ce droit au maintien des garanties est à titre gratuit pour le salarié et pris en charge par le contrat collectif de l'entreprise.

La présentation des enjeux de l'absentéisme ainsi que des moyens mis en place pour permettre aux salariés d'obtenir un revenu emmènent à présent à effectuer une étude de l'absentéisme sur des données réelles. Cette étude va permettre de mettre en place une réflexion et une méthodologie de travail afin de comprendre, modéliser et prédire le phénomène d'absentéisme sur une période et un périmètre précis. Une première étape de présentation des données utilisées dans la suite de ce mémoire est alors nécessaire.

Rappel de l'avertissement présenté en introduction :

Les données relatives aux arrêts de travail et permettant de réaliser les travaux, proviennent des **Données Sociales Nominatives (DSN), entrées en vigueur au 1er janvier 2017**. **La montée en charge de ces éléments étant progressive sur cette première année**, il est important d'identifier dès à présent qu'elles ne contiennent pas l'ensemble des informations réelles du portefeuille observé (nous observons d'ailleurs que le nombre d'arrêts de travail recensés sur 2017 est croissant)**.** L'utilisation de ces données dans un contexte de modélisation du taux d'absentéisme peut **engendrer un biais sur les résultats**, en particulier lors de la modélisation via *Machine Learning* et la prédiction du taux d'absentéisme d'une année sur l'autre.

Une première approche consistant à ne pas prendre en compte ces données 2017 aurait pu être mise en œuvre afin de ne pas biaiser les résultats lors de la prédiction du taux d'absentéisme. Cependant, l'objectif de ce mémoire étant de montrer **le caractère statique de la modélisation via les modèles de *Machine Learning***, pour basculer sur une modélisation plus dynamique du taux d'absentéisme : **l'utilisation d'un historique d'au moins 3 années était nécessaire.** De plus, les autres données 2017 étant complètes (consommation santé, caractéristiques des individus, données Open Data, …), l'utilisation de ces autres **données permet de vérifier la stationnarité des données sur 2018 et 2019 pour les travaux de projection**, et compenser les prédictions réalisées sur 2019. Enfin, les données de 2017 étant **moins impactantes lors des projections du modèle de série temporelle** (car un apprentissage moins long dans le temps par l'étude des données passées précédentes à court terme), celles-ci sont conservées.

Le lecteur restera donc vigilant lors de la lecture des résultats du modèle de *Machine Learning* : ce dernier pourrait présenter des **faiblesses lors de la prédiction du taux d'absentéisme, d'une part du fait du caractère statique de la modélisation, et d'autre part du fait du biais présent sur les données arrêt de travail 2017.**

1.3. Présentation des données pour cette étude

1.3.1 Données à disposition

Les données à disposition servant à l'étude de l'absentéisme proviennent d'un assureur chargé d'assurer les risques Santé et d'AT (arrêt de travail). Elles regroupent des informations sur plusieurs groupes d'entreprises **du secteur de l'industrie**. Les informations en lien avec les arrêts de travail ainsi que la démographie dans les entreprises sont issues **des données DSN complétées chaque mois** par l'ensemble des entreprises composant le portefeuille d'étude. Ces différentes déclarations ont permis à **l'assureur de retraiter les informations contenues dans chaque déclaration DSN**. De ce fait, les données reçues afin de mener l'étude de l'absentéisme ont été **préalablement traitées, anonymisées et construites sous forme de bases de données.**

Les bases transmises permettent ainsi d'avoir une vision d'ensemble de la situation des différents groupes d'entreprises assurés entre 2017 et 2019, par le biais d'informations diverses concernant :

- Les entreprises
- Les salariés
- Les frais de santé
- Les arrêts de travail

La diversité des types d'information est un élément crucial dans l'étude de l'absentéisme. Ce phénomène découlant de multiples causes possibles, l'acquisition d'un large éventail de données permet de comprendre davantage les évolutions de l'absentéisme au fil du temps et de préparer un pilotage ciblé de ce risque, dans la limite de la réglementation (cf. partie 1.3.5). Les diverses informations regroupées dans ces 4 bases permettent ainsi d'étudier l'absentéisme selon différents axes et de les confronter. Ces bases sont composées des informations suivantes :

- **La base « entreprises »,** formée de **408 lignes**, rassemble les informations relatives aux entreprises contenues dans le portefeuille d'étude. Les variables présentes sont :
  - Un identifiant entreprise différenciant ainsi les différentes entreprises présentes dans le portefeuille
  - Un identifiant de groupe caractérisant le groupe auquel les différentes entreprises sont rattachées
  - Le code postal correspondant à la localisation de l'entreprise
  - Le code NAF correspondant à l'activité exercée
  - Le détail des nombres de salariés travaillant dans l'entreprise en fonction de leur catégorie socio-professionnelle à savoir le nombre de cadres, d'ouvriers et d'ETAM présents

- **La base « salariés »,** formée de **47 309 lignes,** rassemble les données relatives aux salariés composant le portefeuille. Les variables présentes sont :
    - Un identifiant salarié permettant de différencier les individus
    - Un identifiant entreprise permettant de savoir à quelle entreprise l'individu est rattaché
    - Le sexe de l'individu
    - La date de naissance du salarié
    - Le code postal correspondant à la localisation de son domicile
    - La catégorie socio-professionnelle divisée en 3 catégories à savoir CADRE, ETAM (Employés, Techniciens, Agents de Maîtrise) et OUVRIER
    - La date d'ancienneté correspondant à la première fois où le salarié a été enregistré dans le portefeuille
    - Les dates d'entrée et de sortie dans l'entreprise correspondant donc à la période d'exposition
    - Les différents salaires pour les années 2017, 2018 et 2019
    - Le nombre d'enfant à la charge du salarié
    - La situation familiale
    - Variable dichotomique informant si le salarié est en temps partiel ou non
- **La base « frais de santé »,** formée de **1 087 793 lignes**, contient l'ensemble des actes de santé consommés par les salariés. Les informations présentes dans cette base sont les suivantes :
    - L'identifiant salarié de l'individu ayant consommé un acte
    - L'identifiant entreprise auquel l'individu est rattaché
    - La date de consommation de l'acte santé
    - Le libellé de l'acte consommé
    - Le montant des frais réels, du remboursement par la Sécurité Sociale et le remboursement de l'assurance complémentaire
- **La base des « arrêts de travail »,** formée de **34 260 lignes**, rassemble l'ensemble des arrêts de travail renseignés depuis la mise en place des DSN. Les données disponibles sont :

- L'identifiant salarié de l'individu ayant eu un arrêt de travail
- L'identifiant entreprise auquel il est rattaché
- Le motif de l'arrêt
- Les dates de début et de fin des arrêts correspondant donc à la durée totale de l'arrêt

A noter que le volume des informations renseignées dans cette dernière base est croissant avec le temps sur la période de 2017. En effet, **la DSN ayant été mise en place en 2017, une montée en charge des déclarations des arrêts de travail** est donc visible sur cette première année. Cette spécificité pourra être vue plus en détail lors des statistiques présentées dans la suite du mémoire.

Le traitement de ces bases de données doit être conforme avec les objectifs fixés dans le cadre de l'étude de l'absentéisme. Ceux-ci sont motivés par la compréhension et la modélisation du phénomène d'absentéisme au sein du portefeuille étudié.

1.3.2 Les approches envisagées pour la compréhension et la modélisation de l'absentéisme

La présentation des différents indicateurs de l'absentéisme a pu montrer que l'absentéisme est **un sujet aux multiples axes de réflexion**. Que ce soit sur la fréquence, la gravité, l'exposition ou encore le taux d'absentéisme, ces indicateurs dépendent également de différents paramètres. En effet, ils sont calculés sur un effectif précis d'individus et sur une période de temps définie. Or, la modification de l'un de ces deux paramètres entraîne une observation différente de ce phénomène. **L'absentéisme est donc fonction des individus étudiés et du temps**. Ainsi, différentes stratégies en fonction de ces deux paramètres peuvent être mises en place afin de l'étudier.

Tout d'abord, l'absentéisme peut être analysé de manière **individuelle ou collective**. Etudier l'absentéisme de manière individuelle revient à attribuer à **chaque individu anonymisé au préalable** un indicateur calculé sur les informations propres de chaque salarié. Cette manière d'apprécier l'absentéisme peut ainsi permettre, compte tenu de l'ensemble des résultats obtenus pour chaque individu, **une segmentation du portefeuille par profil d'individus**. **Une vision collective** de l'absentéisme revient à l'agrégation de l'ensemble des données de l'effectif étudié, afin d'en dégager un résultat définissant **le niveau global d'absentéisme**. Les indicateurs présentés permettent de gérer ces deux cas de figures en décidant d'agréger les données ou non. Cette réflexion permet d'affirmer que la meilleure façon de construire les bases de données revient donc à créer une unique ligne pour chaque individu présent dans le portefeuille. Les données pourront ainsi être étudiées individuellement pour dégager des profils de risque, et si besoin elles pourront être agrégées pour obtenir une vision collective.

L'absentéisme est également **une variable temporelle**. Généralement, l'absentéisme est étudié sur une période de temps précise. Cette manière d'étudier l'absentéisme revient donc à

prendre **une photographie du portefeuille entre deux instants** puis d'agréger les informations présentes durant cette période afin de calculer le niveau d'absentéisme. Cette façon de procéder est généralement la plus répandue. Les études statistiques sont faites d'une année sur l'autre (exemple : baromètre de l'absentéisme en France) [2,3], et les modélisations évaluent des niveaux d'absentéisme à un instant t. Pourtant, cette donnée possédant une dimension temporelle, l'étude de l'absentéisme peut être faite **de façon plus dynamique**. L'objectif ici est de **mettre à jour les informations** en lien avec l'absentéisme **au fil du temps** afin de comprendre et modéliser les variations d'un tel phénomène. Cette approche étant totalement différente de l'approche présentée précédemment, il est alors nécessaire de construire des bases de données distinctes adaptées à ces deux visions.

Le choix de la variable à utiliser afin de modéliser le niveau d'absentéisme reste le dernier paramètre à décider. Les différents indicateurs présentés lors de la première partie apportent tous un renseignement différent sur l'absentéisme. Pourtant, l'étude étant menée selon deux visions, à savoir une vision statique et une plus dynamique**, le choix du taux d'absentéisme comme variable modélisée** semble judicieux. En effet, cette variable peut être facilement mise à jour au cours du temps et combine à la fois l'exposition des salariés et le nombre de jours d'absence. Cette variable sera alors étudiée selon deux approches.

Une première approche par ***Machine Learning*** consiste à étudier l'absentéisme à l'aide de modèles utilisant le maximum de variables à disposition durant une période de temps précise. L'objectif sera alors de :

- **Comprendre et prédire le taux d'absentéisme annuel de chaque individu**
- Donner de **l'explicabilité** aux résultats
- Réaliser **une segmentation homogène des individus**

Chaque ligne de la base de données construite devra donc correspondre à un individu avec l'ensemble de ses caractéristiques sur la période d'observation. Les variables relatives aux caractéristiques de l'individu, de son entreprise, de ses consommations santé et des informations relatives à l'absentéisme permettront ainsi de répondre aux objectifs mentionnés précédemment.

Une deuxième approche par ***Forecasting*** consiste à prendre en compte la dimension temporelle de l'absentéisme. Il ne s'agit donc plus de prédire un taux d'absentéisme individuel sur une année, mais de prédire l'évolution du taux d'absentéisme pour un groupe d'individus à partir des données déjà observées dans le passé. Ces prédictions pourront s'appuyer sur d'autres données temporelles telles que les consommations de santé. Les objectifs sont donc de :

- **Prédire l'évolution du taux d'absentéisme pour un groupe d'individus** avec une certaine **fréquence** (annuelle, mensuelle, hebdomadaire, journalière…)
- Donner de **l'explicabilité à l'évolution de ce taux** à l'aide d'autres données temporelles comme la consommation santé

Une deuxième base de données devra donc être construite afin de répondre à ces objectifs : elle décrira pour chaque individu ses jours de présence et d'absence au travail jour après jour. Chaque ligne correspondant donc à la situation de chaque salarié au cours du temps.

1.3.3 Construction des deux bases de données

Dans la suite de ce mémoire, deux grandes méthodes seront présentées afin d'étudier l'absentéisme :

- Les **méthodes de *Machine Learning***
- Les **méthodes de *Forecasting*** et plus particulièrement des séries temporelles

Comme expliqué précédemment, deux bases de données doivent être construites afin de travailler sur ces deux axes. Toutefois même si ces deux bases de données n'auront pas les mêmes variables, elles devront néanmoins coïncider en décrivant les mêmes individus.

La construction des deux bases commence par celle permettant l'utilisation des séries temporelles. Une vision jour après jour de l'absentéisme doit donc être mise en place au travers de la création de cette base de données. En effet, il est nécessaire de pouvoir étudier les jours de présence et d'absence de chaque salarié afin d'étudier l'évolution du taux d'absentéisme au cours du temps.

Cette première base est alors créée à partir des bases initiales « salariés » et « arrêts de travail ». Chaque ligne correspond aux informations d'un salarié pendant **une certaine période d'exposition dans une entreprise avec une catégorie socio-professionnelle donnée**. Ce choix dans la création de la base, permet ainsi de pallier les problèmes de doublons et prend en compte l'évolution du salarié dans le temps. Cependant, par cette décision, un même salarié peut être comptabilisé plusieurs fois puisque celui-ci peut changer d'entreprise ou de catégorie socio-professionnelle. Néanmoins cette conséquence reste raisonnable (moins de 1% des lignes sont concernées) en faisant l'hypothèse que le changement d'entreprise ou de catégorie socio-professionnelle peut impacter le taux d'absentéisme. Chaque ligne correspond donc à un individu indépendant des autres lignes construites.

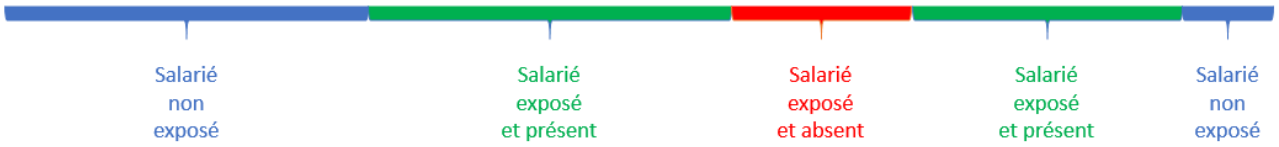
La base ainsi construite possède donc **42 534 lignes**, permettant d'ajouter les données nécessaires à la création de la base pour les séries temporelles. Les données arrêts de travail sont comprises entre le 01/01/2017 et le 31/12/2019. De ce fait pour chaque ligne de la nouvelle base, une variable correspondant à un jour compris dans la période d'observation est ajoutée afin de renseigner si l'individu est exposé et si oui ou non l'individu est absent. Les dates d'entrée et de sortie des individus dans l'entreprise proviennent de la base « salariés » et les dates de début et de fin des arrêts de la base « arrêts de travail ». Chaque variable correspondant à un jour de la période d'étude est renseignée ainsi :

- NA si aucune valeur n'est renseignée, correspondant au fait que le salarié n'est pas exposé au risque
- 0 si le salarié est exposé au risque et s'il est présent au travail

- 1 si le salarié est exposé au risque et s'il est en arrêt de travail

Figure 7 : Détail d'une ligne dans la base des séries temporelles

	01/01/2017	02/01/2017	...	14/07/2018	15/07/2018	16/07/2018	...	25/08/2018	26/08/2018	27/08/2018	28/09/2018	...	31/12/2018	01/01/2019
Salarié i	NA	NA	...	NA	0	0	...	0	1	1	0	...	0	NA



Cette base permettra ainsi dans la partie Forecasting de ce mémoire de voir l'évolution du taux d'absentéisme pour un groupe d'individus au cours du temps.

La deuxième base de données construite pour la partie *Machine Learning* s'appuie sur la première base créée, à savoir les 42 534 lignes mentionnées auparavant. Néanmoins celle-ci regroupe l'ensemble des informations obtenues sur les différents salariés du portefeuille, sans chercher à étudier une évolution temporelle du taux d'absentéisme jour après jour.

La base pour la partie Machine Learning regroupant l'ensemble des données disponibles sur les salariés reprend les informations contenues dans les bases « salariés », « entreprises », « frais de santé », « arrêts de travail » et la base précédemment construite.

- La base « salariés » permet de rapporter l'ensemble des caractéristiques des salariés comme présentées dans la partie précédente
- La base « entreprises » permet de rapporter des informations telles que le code NAF indiquant le secteur d'activité dans lequel l'entreprise évolue, la taille de l'entreprise au travers du nombre de salariés où l'individu évolue ainsi que le pourcentage d'ouvriers
- La base « frais de santé » permet de rapporter les frais de santé (montant de dépense engagée et montant de remboursement complémentaire) ainsi que le nombre d'actes consommés par famille d'acte, correspondant à un regroupement de plusieurs actes santé appartenant au même type de soins comme le dentaire, l'optique, les généralistes, etc…
- La base « arrêts de travail » permet de rapporter le nombre d'arrêts de travail effectués pour chaque salarié en fonction du motif d'arrêt renseigné
- La base précédemment créée permet de calculer le taux d'absentéisme pour chaque année et pour chaque individu

Figure 8 : Exemple d'une ligne dans la base de Machine Learning

Informations sur le salarié						Informations sur l'entreprise/établissment				Santé	Arrêt de travail		Variable CIBLE
IDENTIFIANT	AGE	SEXE	ANCIENNETE	CSP	...	IDENTIFIANT	Pourcentage OUVRIER	Region Entreprise	TAILLE ENTREPRISE	CONSO SANTE 2018	TAUX 2018	Nombre ARRET 2018	TAUX 2019
Salarié 1	45	M	1	Ouvrier	...	Entreprise 1	30%	Occitanie	300 salariés	***	2%	4	3%

Toutes ces informations ajoutées permettent alors d'obtenir une base permettant d'appliquer les algorithmes de *Machine Learning* pour chaque individu afin de prédire son taux d'absentéisme annuel de 2019.

1.3.4 Statistiques de synthèse sur les bases de données

L'étude de l'absentéisme appliquée à un portefeuille de salariés passe toujours par une phase de compréhension des données recueillies, ainsi que de la recherche des possibles axes de réflexions à adopter dans une logique de réduction de ce risque. Cette étape prend alors généralement **la forme d'un tableau de bord**, permettant à la fois de répertorier les statistiques en lien avec les salariés, mais également de confronter les données de l'absentéisme avec les différentes caractéristiques des individus ou de leur consommation en santé. L'objectif est de pouvoir donner une vision précise du portefeuille étudié, mais également d'en dégager une compréhension à différents niveaux de l'état du portefeuille en matière d'absentéisme.

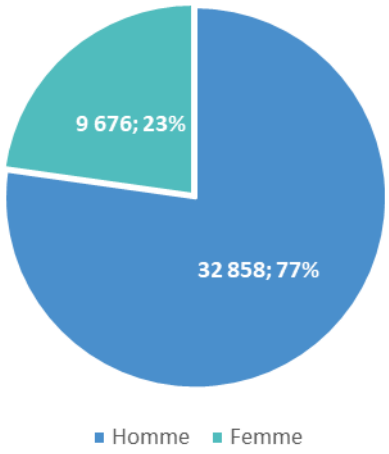
Comme montré précédemment, les bases de données construites détiennent différents types d'informations, à savoir des informations sur les salariés, sur leur consommation santé ainsi que sur leurs arrêts de travail. Une présentation des statistiques sur ces différents axes est nécessaire avant toute première modélisation. Ces données ayant un caractère temporel, les statistiques présentées prendront en compte **l'évolution du portefeuille** durant la période d'observation.

Les bases de données construites pour les méthodes de Machine Learning et de Forecasting contiennent toutes deux 42 534 lignes correspondant au nombre d'individus étudiés dans la suite de ce mémoire, répartis dans 404 entreprises au total, elles-mêmes réparties dans 10 grands groupes d'entreprises.

Une première étape consiste à regarder les statistiques sur les caractéristiques des individus composant le portefeuille avant d'étudier plus précisément les liens possibles avec l'absentéisme. Ces individus sont répartis dans 404 entreprises toutes en lien avec **le secteur de l'industrie**.

Les proportions d'hommes et de femmes sur l'ensemble de la période contenues dans le portefeuille sont les suivantes :

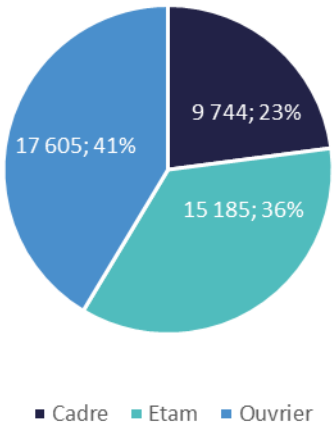
Figure 9 : Proportion d'hommes et de femmes dans le portefeuille



La proportion de femmes dans l'ensemble du portefeuille est **d'environ un quart**. Ce portefeuille est issu du secteur de l'industrie, un secteur majoritairement représenté par des hommes.

Au niveau des catégories socio-professionnelles (CSP), 3 catégories sont alors présentées, les Cadres, ETAM (employés, techniciens, agents de maîtrise) et Ouvriers. Sur toute la période d'observation, les proportions selon ces différentes catégories sont les suivantes :

Figure 10 : Proportion des CSP



Les catégories ETAM et Ouvrier sont donc les catégories les plus présentes dans le portefeuille à savoir 77%. Ces chiffres sont représentatifs du secteur de l'industrie.

Outre la catégorie socio-professionnelle, **les différents types d'activité** exercés par les salariés ont été renseignés dans les bases. Ces différentes activités réparties en **51 classes initialement ont été regroupées en 9 classes** afin de travailler sur un nombre restreint de

modalités. Ces différentes classes représentent des grands types d'activité existant dans le milieu de l'industrie, à savoir :

- Manufacture
- Technicien
- Transport
- Secrétariat et gestion de données
- Manutention
- Direction et hiérarchie
- Ingénierie
- Télécommunication
- Gestion d'infrastructure

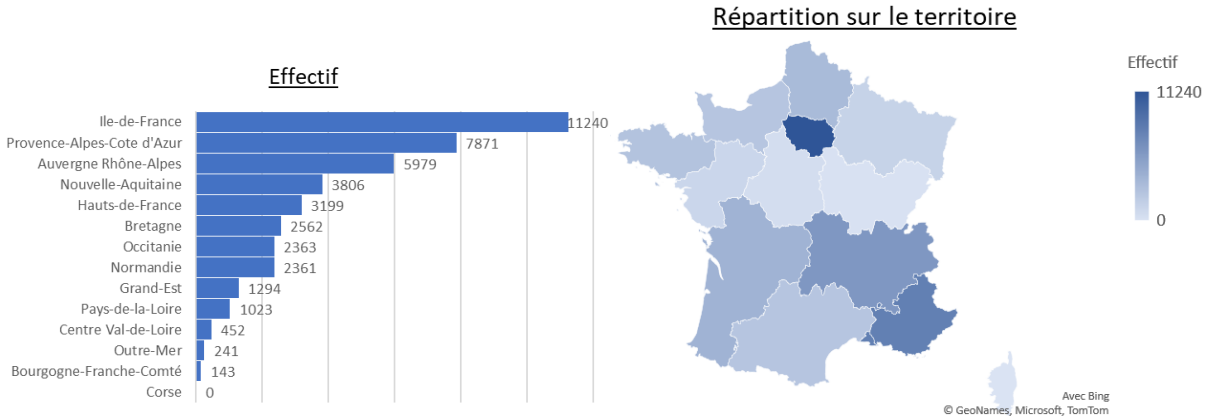
Sur la base complète, les effectifs, les CSP ainsi que la distinction entre hommes et femmes peuvent être répartis suivant ces types d'activité. Plus de **85% des effectifs occupent des postes nécessitant un travail physique** (Manufacture, Technicien, Manutention). Ces activités sont d'ailleurs majoritairement **composées d'Ouvriers** et par des hommes, avec une proportion de femmes inférieure à 20% dans ces secteurs. A l'inverse, **les secteurs composés de plus de 25% de femmes** sont caractérisés par des **proportions d'ETAM ou Cadres plus élevées** que celle des Ouvriers, avec des activités demandant une participation physique moins importante (Secrétariat et gestion informatique, Ingénierie, Direction et hiérarchie, Gestion d'infrastructure).

Tableau 4 : Répartition des effectifs, CSP et sexe en fonction du type d'activité

	%Effectif	%Ouvrier	%ETAM	%Cadre	%Homme	%Femme
Manufacture	73,27%	46%	31%	23%	81%	19%
Technicien	12,43%	47%	35%	18%	83%	17%
Gestion d'infrastructure	5,27%	6%	87%	7%	73%	27%
Secrétariat et gestion informatique	2,81%	26%	30%	44%	65%	35%
Ingénierie	2,51%	8%	70%	23%	75%	25%
Direction et hiérarchie	1,85%	6%	39%	55%	52%	48%
Télécommunication	1,38%	26%	48%	26%	79%	21%
Manutention	0,45%	60%	25%	15%	82%	18%
Transport	0,01%	20%	40%	40%	100%	0%

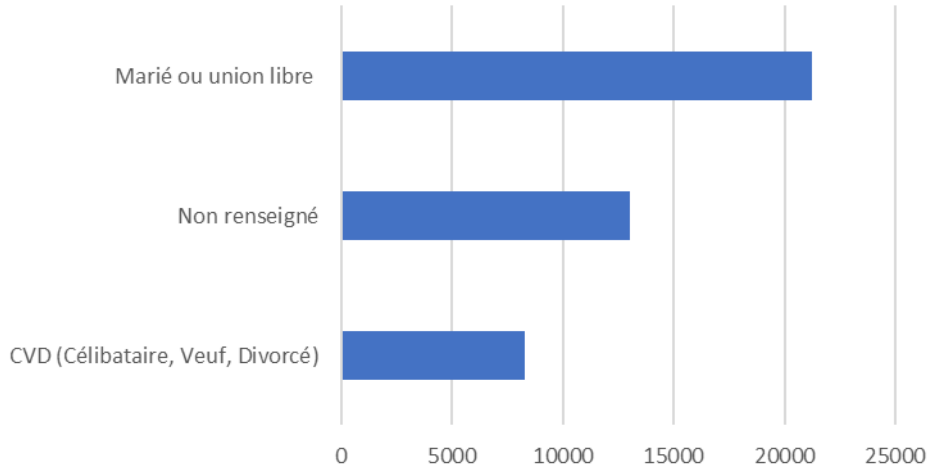
La répartition des salariés sur l'ensemble du territoire français est donnée ci-dessous. Les salariés sont **concentrés majoritairement en Île-de-France et en Provence-Alpes Côte d'Azur**. Pour rappel, la région Île-de-France est celle où le **taux d'absentéisme est le plus bas** ces dernières années. [2,3]

Figure 11 : Répartition des salariés sur le territoire français



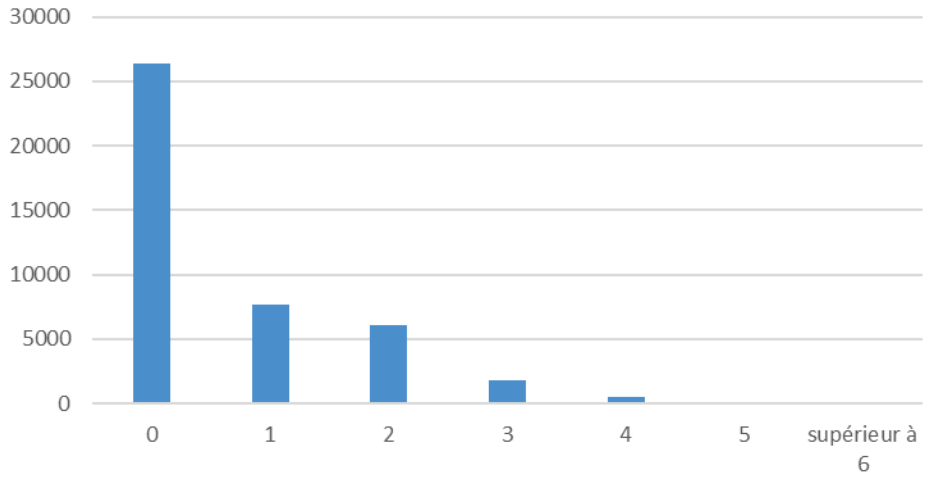
Les différentes situations de famille sont également renseignées. Les situations de famille les plus représentées sont les mariés et concubins. Cependant, 30% des individus de la base n'ont pas de situation familiale renseignée.

Figure 12 : Proportion des situations de famille



Le nombre d'enfants à charge complète les informations sur la situation de famille des salariés. Ainsi près de 40% des individus ont au moins un enfant à charge.

Figure 13 : Répartition du nombre d'enfants à charge



Ces premières statistiques sont des statistiques globales. Elles permettent d'avoir un aperçu rapide des données sur l'ensemble de la période d'observation. Cependant le portefeuille évolue au cours du temps avec une exposition au risque différente d'une année sur l'autre. Il est alors nécessaire de présenter l'évolution du portefeuille au cours de cette période.

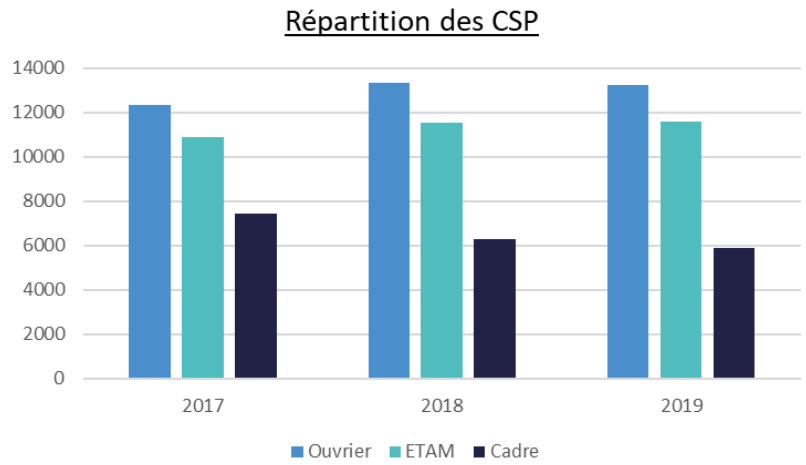
Sur la période d'observation, **les effectifs exposés** (ayant au moins un jour d'exposition dans l'entreprise) du portefeuille sont les suivants :

Tableau 5 : Répartition des salariés exposés en fonction du sexe et de l'année

	2017	2018	2019
Total	30 800	33 001	31 457
Homme	73,7%	74,5%	73,9%
Femme	26,3%	25,5%	26,1%

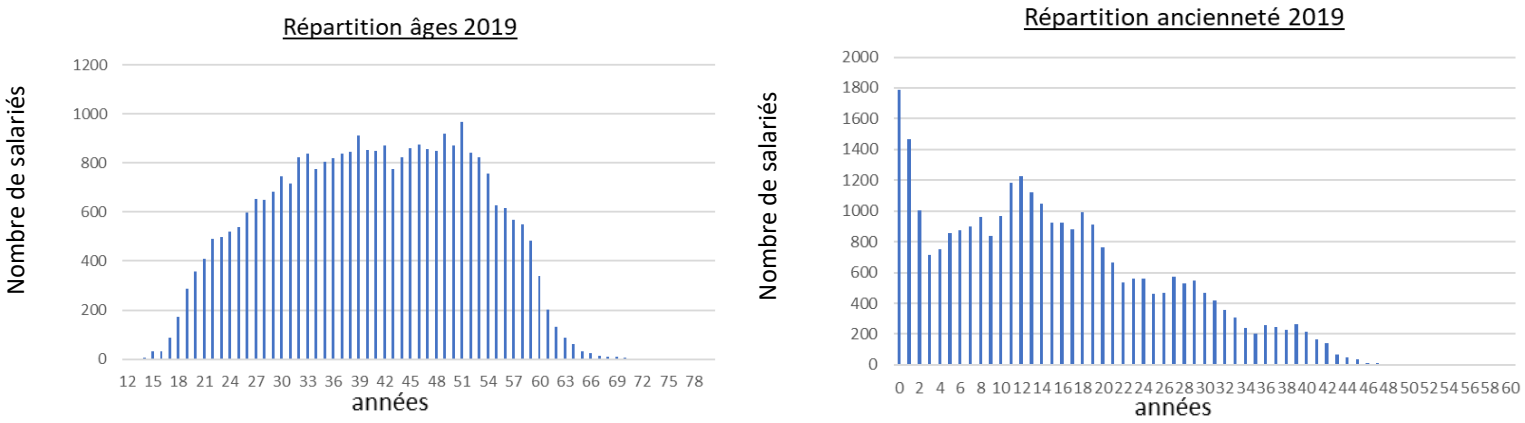
Le portefeuille reste donc stable d'une année sur l'autre, que ce soit au niveau du nombre de salariés ou de la répartition hommes/femmes. De même, l'évolution des catégories socio-professionnelles exposées sur ces trois années reste stable.

Figure 14 : Répartition des CSP au cours du temps



La répartition des âges et des anciennetés au cours des trois années reste également stable. La répartition des âges et anciennetés est alors présentée à l'aide des graphiques suivants sur 2019 :

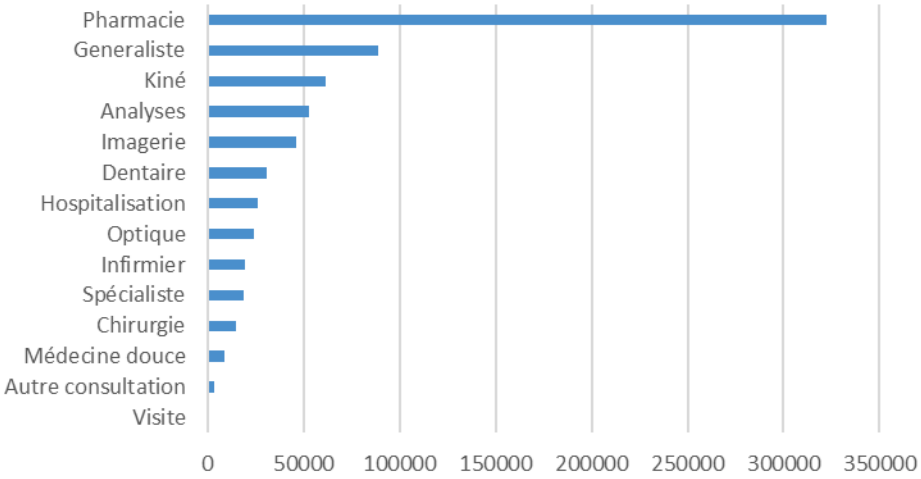
Figure 15 : Répartition des âges et anciennetés en 2019



Le portefeuille est en majorité composé d'individus entre 20 et 50 ans, avec un recrutement massif chaque année puisque le nombre d'individus ayant une ancienneté inférieure à un an est élevé, conséquence probable de contrats courts (ancienneté moyenne : 14 ans). Les individus ayant un âge inférieur à 18 ans correspondent aux apprentis du portefeuille. Le portefeuille étant stable au niveau des caractéristiques des individus, **l'étude porte à présent sur les actes de santé consommés** par les salariés sur la période de 2017 à 2019.

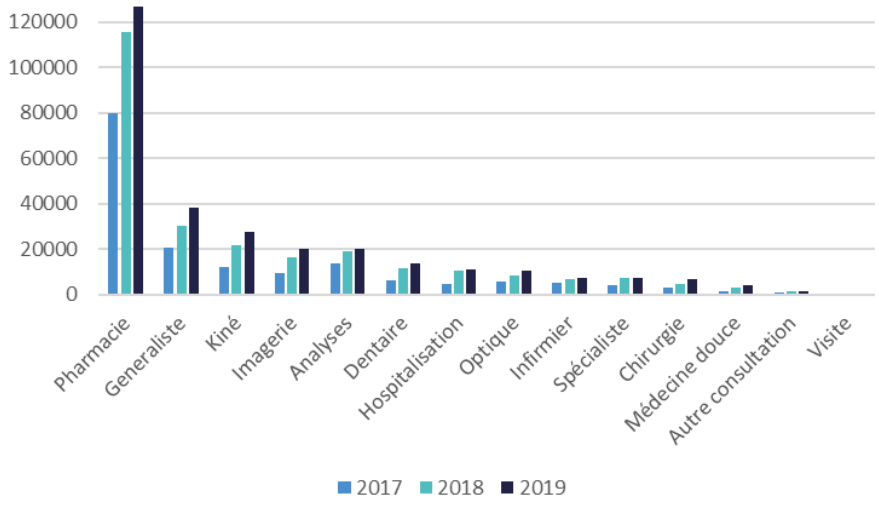
Les informations concernant la santé ont été rajoutées à partir de la base « frais de santé » mentionnée plus haut. Pour simplifier la lecture de l'ensemble des actes de soins et de leur nomenclature, ces actes ont été regroupés **en grands groupes d'actes** afin d'avoir une vision plus claire des consommations des salariés. Sur l'ensemble de la période d'étude la répartition du nombre d'actes de soins consommés au total pour chaque grand poste de soins est donnée par le graphique suivant :

Figure 16 : Nombre d'actes consommés par grand poste de soins



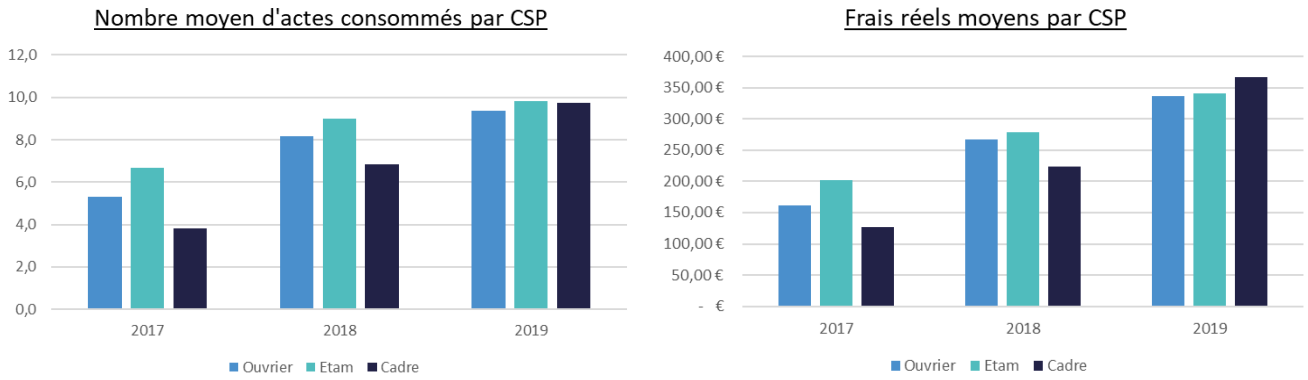
La pharmacie, les consultations de généralistes et les consultations de kinésithérapie sont les groupes d'actes les plus consommés durant cette période. L'évolution du nombre d'actes consommés chaque année montre **une augmentation de consommation en santé, et ce pour l'ensemble des postes, le nombre d'individus dans le portefeuille étant stable d'année en année** :

Figure 17 : Répartition du nombre d'actes consommés par grand poste de soins et par année



Cette constatation peut également être confirmée au travers du nombre moyen d'actes consommés et des frais réels moyens payés par CSP au cours du temps :

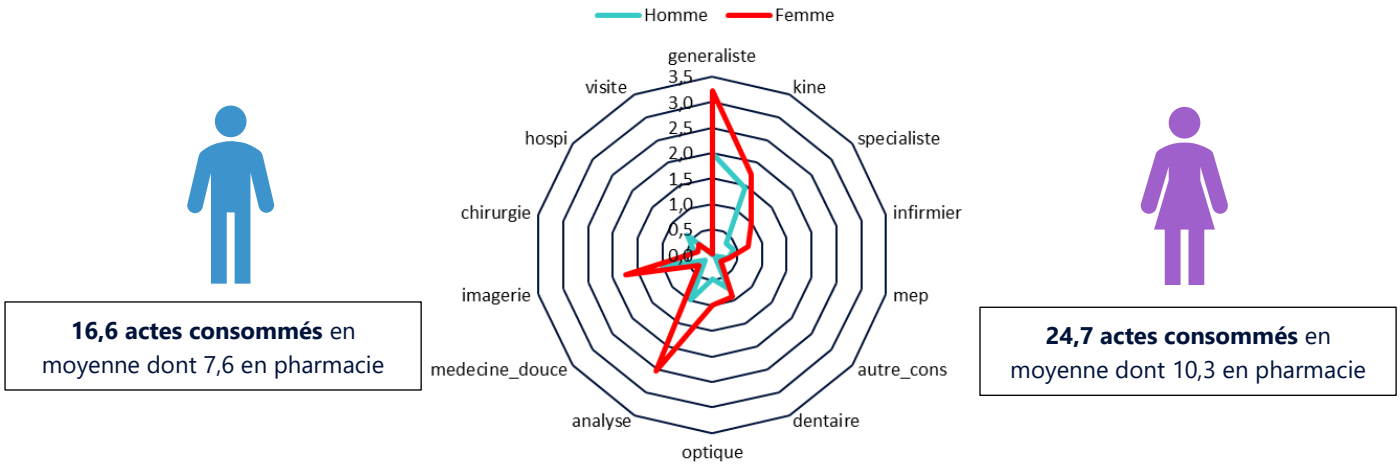
Figure 18 : Nombre d'actes consommés par CSP et frais réels moyens en santé par CSP



Le nombre d'actes et les frais réels des cadres semblent avoir drastiquement augmentés durant ces dernières années. Il sera alors intéressant de confronter cette augmentation de consommation de santé globale avec les données d'arrêts de travail pour voir si ces données sont corrélées ou non. (cf. partie 3)

La consommation santé peut également être décrite en fonction du sexe et de l'activité occupée par les individus. Les effectifs étant différents entre ces différents groupes, les statistiques portent donc sur le nombre moyen d'actes consommés en santé. **La consommation d'actes santé chez les femmes est en moyenne plus élevée que chez les hommes**. Avec une moyenne de 24,7 actes consommés dont 10,3 en pharmacie, cette consommation surpasse celle des hommes, qui est de 16,6 actes consommés en moyenne pour 7,6 actes en pharmacie. La répartition des actes consommés en moyenne pour les autres postes est présentée sur le graphique suivant :

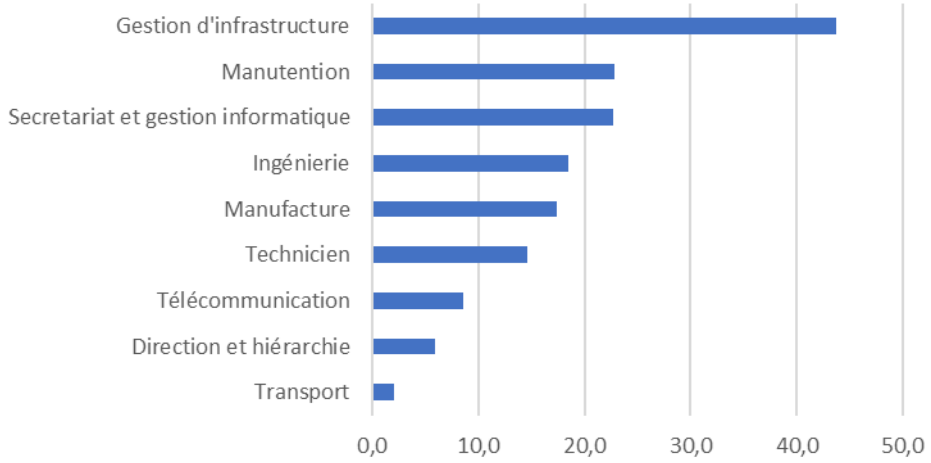
Figure 19 : Répartition des nombres d'actes consommés moyen pour chaque poste de santé et par sexe



Les femmes présentes dans ce portefeuille (âge moyen : 41 ans) consomment donc plus que les hommes (âge moyen : 40 ans) en quantité d'actes, quel que soit le poste de santé observé, excepté

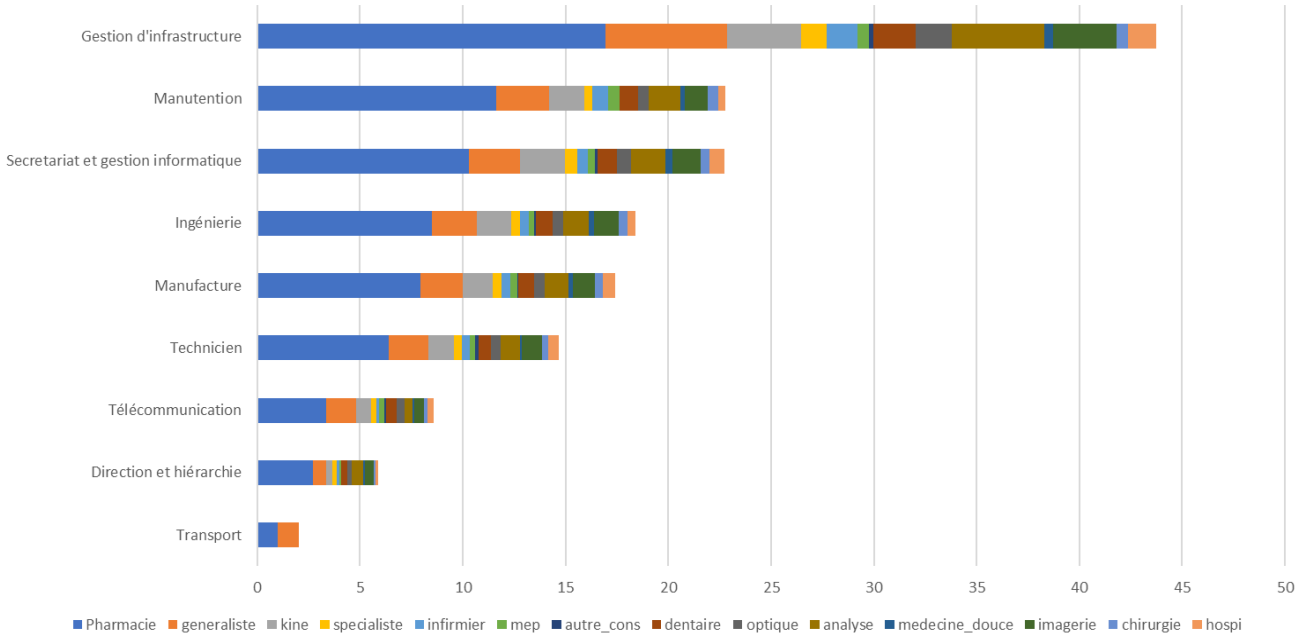
en chirurgie et hospitalisation. Les secteurs d'activité sont également un critère permettant de voir une répartition différente du nombre moyen d'actes consommés en santé. Les secteurs consommant le plus d'actes sont donc les métiers en lient avec la gestion d'infrastructures, la manutention et les métiers liés au secrétariat et à la gestion informatique.

Figure 20 : Nombre moyen d'actes consommés en santé en fonction de l'activité professionnelle



Ces répartitions peuvent être décomposées en fonction des postes santé. La plupart des actes consommés sont de la pharmacie pour l'ensemble des secteurs d'activité :

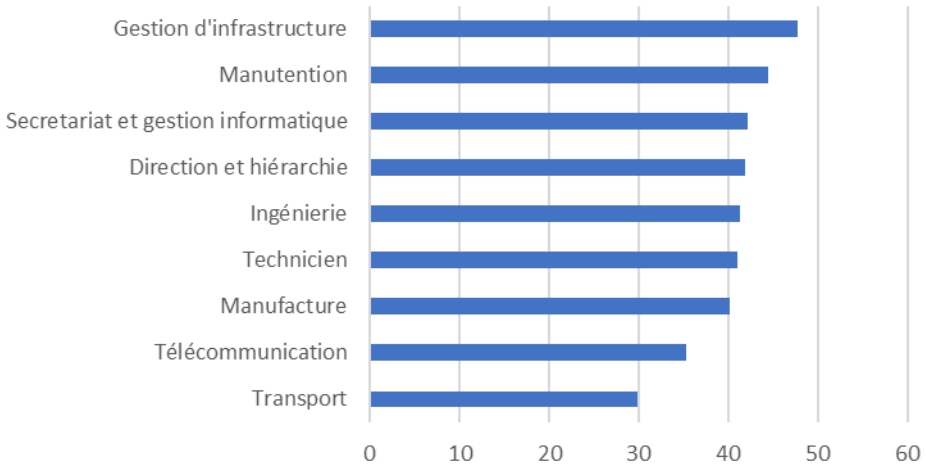
Figure 21 : Répartition des nombres moyens d'actes de santé en fonction de l'activité et des postes santé



A première vue, **les métiers réclamant une contribution physique de la part des salariés ne sont pas les secteurs consommant le plus d'actes de santé**, comme les techniciens et les

salariés travaillant dans les manufactures. En revanche, la manutention arrive à la deuxième place de ce classement avec plus de 20 actes consommés en moyenne par salarié. Cependant, il semble que ce soient **les activités avec des CSP telles que ETAM ou Cadres qui consomment le plus d'actes** de santé, comme la gestion d'infrastructures, le secrétariat et gestion informatique ou encore l'ingénierie. Premièrement, comme montré en amont, les femmes consomment plus que les hommes. Or, ce sont dans les secteurs d'activité regroupant le plus d'ETAM ou de Cadres que l'on retrouve le plus de femmes. De plus, ces catégories permettent généralement à ces salariés d'avoir de meilleures garanties en santé, leur permettant **un accès plus facile** aux différents postes de santé. Enfin, **l'étude de l'âge moyen** des salariés en fonction de ces différentes activités **montre une corrélation positive** entre ces deux données :

Figure 22 : Âge moyen en fonction de l'activité professionnelle



Les données en santé ont permis de montrer une évolution croissante du nombre de consommation en santé au cours du temps, en particulier une augmentation marquée chez les ETAM et Cadres. De plus, les profils de consommation en santé diffèrent en fonction du sexe, de l'âge, de la catégorie socio-professionnelle, ainsi que de l'activité professionnelle exécutée.

En dernier lieu, **les statistiques sur les arrêts de travail** des salariés composant le portefeuille. Ces informations ont été ajoutées à l'aide de la base « arrêts de travail ». Chaque arrêt est alors caractérisé par une date de début, une date de fin et un motif. Les arrêts de travail peuvent alors être étudiés selon le motif mais également leur durée ou le taux d'absentéisme de chaque individu sur une certaine période où l'individu est exposé.

L'étude sur les données d'arrêts de travail débute par la proportion d'absence au sein du portefeuille. Pour l'ensemble des salariés étudiés, **35% ont vécu au moins un arrêt de travail** entre 2017 et 2019. Le détail des proportions des individus ayant enregistré au moins un arrêt de travail pour chaque année en fonction du nombre de salariés présents cette même année est donné ci-dessous :

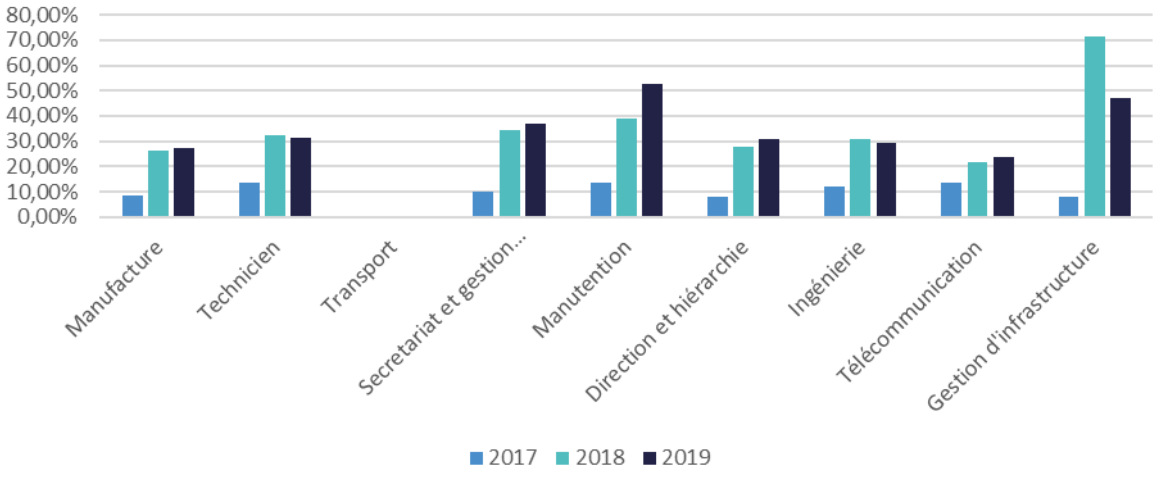
Tableau 6 : Proportion d'absence[5] pour chaque année et par sexe

	2017	2018	2019
Homme	10%	27%	29%
Femme	8%	38%	29%
Total	9%	30%	29%

Les femmes semblent plus touchées par les arrêts de travail que les hommes, au moins pour l'année de 2018, ce qui est en adéquation avec les différentes études sur l'absentéisme. Pour rappel, les statistiques sur les arrêts de 2017 sont faibles à cause de la mise en place progressive des DSN à partir de janvier 2017.

La proportion d'absence en fonction de l'activité professionnelle est un indicateur important. En effet, ces statistiques permettent de comprendre l'évolution du nombre de salariés touchés par un arrêt de travail au cours du temps. Les secteurs les plus touchés en termes d'exposition sont donc les activités de manutention, de secrétariat et gestion informatique ainsi que de gestion d'infrastructure. Le taux de proportion d'absence reste relativement stable pour l'ensemble des secteurs hormis pour la manutention et la gestion d'infrastructure où le taux d'exposition augmente pour le premier et diminue pour le second entre 2018 et 2019. L'activité de transport est la seule ayant échappée au phénomène d'absentéisme.

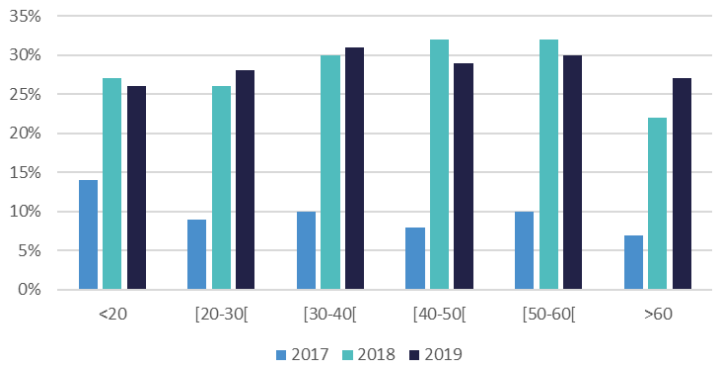
Figure 23 : Proportion d'absence en fonction de l'activité



[5] Proportion d'absence = $\frac{Nombre\ d'individus\ avec\ au\ moins\ un\ arrêt}{Nombre\ total\ d'individus}$

De même, la proportion d'individus ayant eu au moins un arrêt de travail peut être étudiée en fonction en fonction de l'âge. Les chiffres montrent une augmentation de l'exposition entre 2018 et 2019 pour les 20-40 ans ainsi que pour les plus de 60 ans. Néanmoins l'exposition reste la plus élevée chez les 30-60 ans.

Figure 24 : Proportion d'absence en fonction de l'âge et par année



Après avoir étudié la proportion d'absence, l'étude se concentre à présent sur les motifs d'arrêt. Les différents nombres d'arrêts en fonction du motif pour la période de 2017 à 2019 sont alors résumés dans le tableau suivant :

Tableau 7 : Nombre d'arrêts de travail par année et par motif

	2017	2018	2019	Proportion 2019
Maladie	2 996	14 635	12 677	86,8%
Congé suite à un accident de travail	223	658	816	5,6%
Paternité / accueil de l'enfant	136	584	631	4,3%
Temps partiel thérapeutique	-	49	186	1,3%
Maternité / adoption	60	123	160	1,1%
Congé suite à un accident de trajet	24	69	74	0,5%
Congé suite à une maladie professionnelle	41	57	57	0,4%
Femme enceinte dispensée de travail	-	-	-	0,0%
Total	3 480	16 175	14 601	

Le motif principal d'arrêt est la maladie représentant près **de 87% des arrêts**, suivi par les accidents de travail représentant 5,7% des arrêts en 2019. Cette distinction des arrêts peut également être traitée en lien avec les différents secteurs d'activité :

Tableau 8 : Répartition du nombre d'arrêts en fonction de l'activité

	2017	2018	2019	Total
Manufacture	2 203	9 622	9 884	21 709
Technicien	747	2 088	2 021	4 856
Gestion d'infrastructure	196	2 859	1 271	4 326
Secretariat et gestion informatique	102	557	632	1 291
Ingénierie	109	456	357	922
Direction et hiérarchie	61	278	282	621
Télécommunication	35	206	51	292
Manutention	28	111	104	243
Transport	-	-	-	-

Hormis le secteur de la manutention, l'ensemble des **secteurs regroupant le plus d'arrêts correspondent aux secteurs nécessitant un travail physique**, largement mené par des Ouvriers dans la manufacture et les métiers en lien avec les techniciens. Néanmoins, des secteurs possédant beaucoup d'ETAM et de Cadres comme la gestion d'infrastructure ainsi que le secrétariat et gestion informatique regroupent une partie non négligeable des arrêts de travail à savoir plus de 20%.

L'étude des données arrêts de travail se concentre ensuite sur la nature même des arrêts, à savoir les indicateurs présentés plus tôt dans ce mémoire : la fréquence, la gravité ainsi que le taux d'absentéisme. Par définition de la fréquence et de la gravité :

$$Fréquence \times Gravité = Taux\ d'absentéisme$$

Une première vision de ces éléments est proposée sur l'ensemble de l'historique des données entre 2017 et 2019 :

Tableau 9 : Fréquence, gravité et taux d'absentéisme sur l'ensemble des données

	Fréquence	Gravité	Taux d'absentéisme
Homme	0,11%	23,25	2,65%
Femme	0,12%	18,99	2,23%
Ensemble portefeuille	0,11%	22,03	2,53%

Le taux d'absentéisme dans le secteur de l'industrie est généralement meilleur que celui des autres secteurs d'activité comme par exemple celui du secteur tertiaire qui est généralement autour de 5%. Néanmoins, ces calculs prennent en compte également l'année de 2017, qui a connu une montée en charge des données DSN. Il est donc nécessaire **d'étudier ces éléments annuellement**.

Tableau 10 : Fréquence, gravité et taux d'absentéisme calculés annuellement

	2017			2018			2019		
	Freq	Grav	Tx	Freq	Grav	Tx	Freq	Grav	Tx
Homme	0,04%	27,32	1,06%	0,15%	21,34	3,14%	0,15%	24,11	3,70%
Femme	0,03%	34,34	0,94%	0,19%	12,42	2,41%	0,13%	25,78	3,32%
Ensemble portefeuille	0,04%	28,85	1,03%	0,16%	18,32	2,94%	0,15%	24,52	3,59%

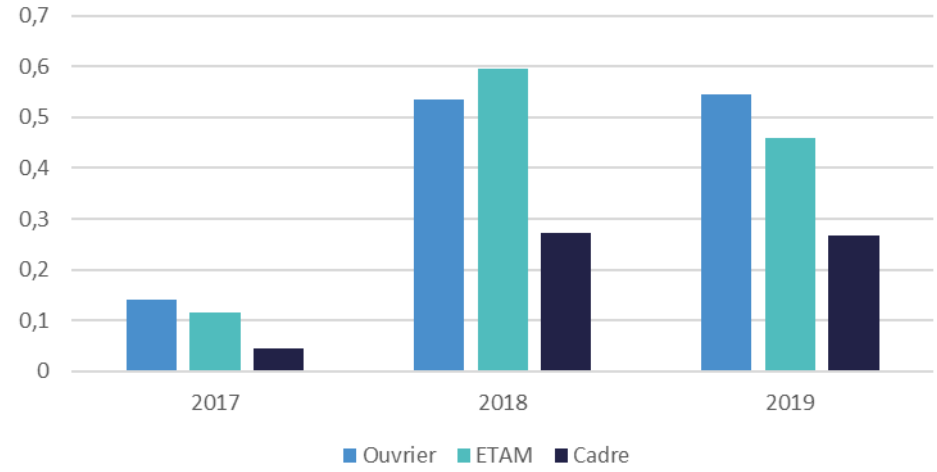
A partir de ce tableau, plusieurs points à mentionner :

- La fréquence en 2017 est faible par rapport aux données de 2018 et 2019 en raison de la montée en charge des données DSN
- **Une augmentation du taux d'absentéisme** sur l'ensemble du portefeuille entre 2018 et 2019
- **Une augmentation d'autant plus marquée chez les femmes avec une augmentation du taux d'absentéisme de 37% contre 17% chez les hommes** entre 2018 et 2019
- Une fréquence stable entre 2018 et 2019 chez les hommes avec une augmentation de la gravité, tandis que **la fréquence chez les femmes baisse mais avec une augmentation de la gravité de plus de 100%**
- Le taux d'absentéisme en 2019 se trouve entre 3 et 4%, un taux relativement bas par rapport aux études menées sur l'absentéisme

Ainsi, si le faible taux d'absentéisme pourrait indiquer un état de santé plutôt bon vis-à-vis de l'absentéisme, les indicateurs de la fréquence et de la gravité en revanche **indiquent une augmentation de la durée des arrêts de travail**. De plus, les distinctions entre hommes et femmes, montrent une **dégradation de l'absentéisme plus marquée chez les femmes**. Il est donc nécessaire d'étudier d'avantage les données des arrêts de travail sous d'autres angles.

Tout d'abord, le nombre moyen d'arrêts par CSP pour chaque année est présenté ci-dessous :

Figure 25 : Nombre moyens d'arrêts de travail d'un salarié en fonction de la CSP et de l'année



A partir de ce graphique, les ETAM en 2018 sont plus touchés par les arrêts de travail avec une amélioration en 2019, tandis que le risque reste stable entre 2018 et 2019 pour les ouvriers et les cadres. Pour rappel, les statistiques précédentes ont montré que les femmes dans ce portefeuille occupaient plutôt des postes en tant qu'ETAM ou de Cadre. De plus, le tableau précédent a montré que la fréquence des femmes en 2018 était plus importante que chez les hommes. Au niveau du nombre moyen de jours en arrêt de travail, une distinction est effectuée entre la moyenne sur l'ensemble des individus et sur ceux ayant eu au moins un arrêt de travail.

Figure 26 : Nombre moyen de jours en arrêt de travail par salarié, en fonction de la CSP et de l'année

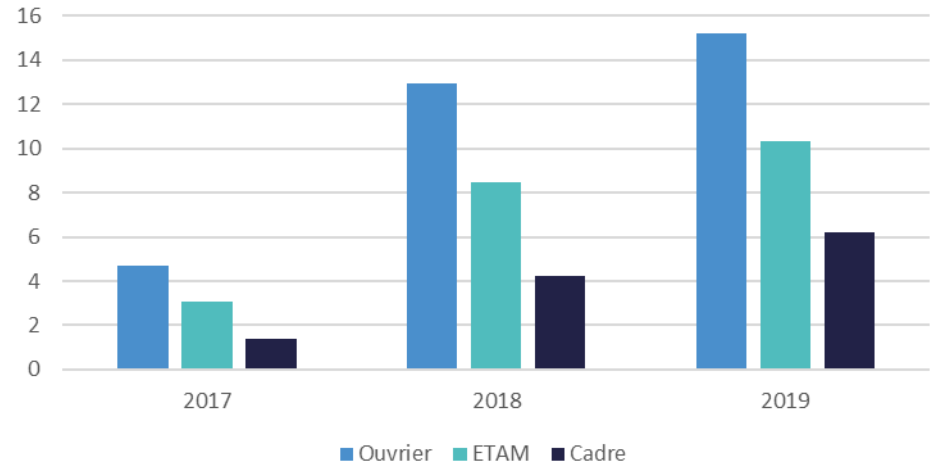
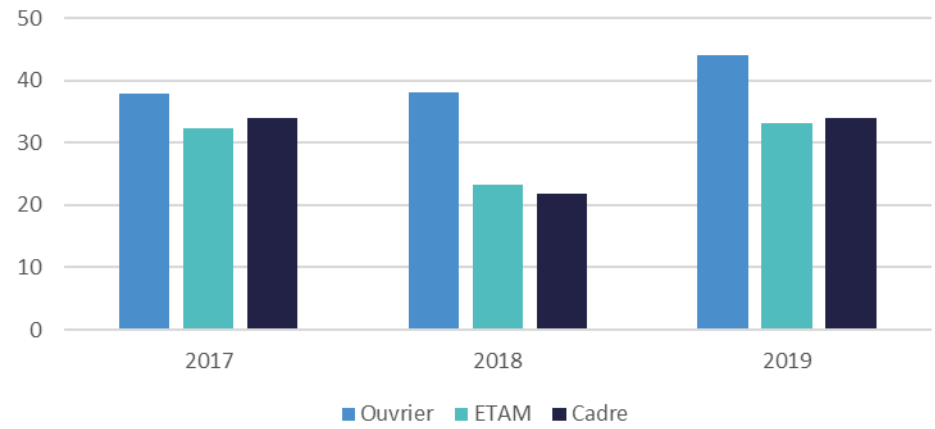
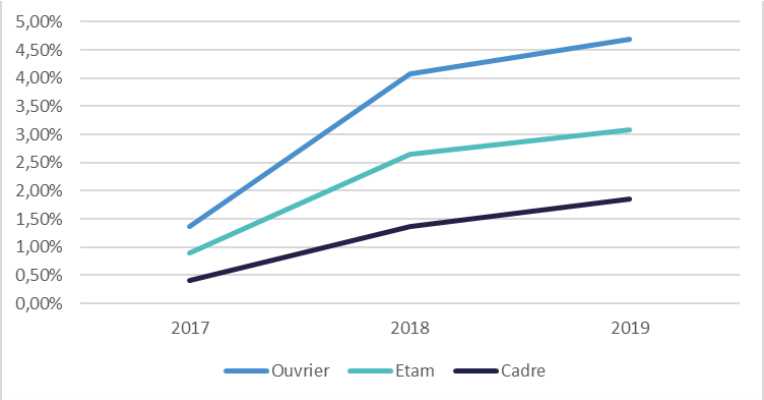


Figure 27 : Nombre moyen de jours en arrêt de travail par salarié, pour les individus ayant eu au moins un arrêt de travail en fonction de leur CSP et de l'année



Les salariés ayant le plus de nombre de jours d'arrêt sont les Ouvriers. Cependant, l'évolution entre 2018 et 2019 montre une augmentation plus importante chez les ETAM et les Cadres.

Figure 28 : Taux d'absentéisme en fonction de la CSP et de l'année



Le taux d'absentéisme est plus élevé chez les ouvriers qui sont souvent exposés à plus de risques que les catégories ETAM et Cadre. Une croissance du taux d'absentéisme au fur et à mesure des années est à noter, malgré le faible taux de 2017 qui est dû à la mise en place progressive des DSN durant cette année.

Des disparités peuvent également être visibles en fonction des régions :

Tableau 11 : Taux d'absentéisme en fonction de la région et de l'année

	2017	2018	2019
Grand Est	1,6%	5,4%	7,7%
Outre-Mer	1,7%	3,6%	6,9%
Auvergne-Rhone-Alpes	1,2%	4,0%	4,5%
Occitanie	0,9%	3,1%	4,3%
Bourgogne-Franche-Comte	1,9%	5,8%	3,9%
Hauts-de-France	1,4%	3,4%	3,9%
Ile-de-France	1,2%	3,1%	3,9%
Nouvelle-Aquitaine	1,3%	3,2%	3,7%
Provence-Alpes-Cote d'Azur	0,7%	2,4%	3,2%
Centre-Val de Loire	1,7%	3,9%	2,7%
Pays de la Loire	0,9%	1,9%	2,5%
Normandie	0,4%	0,9%	1,2%
Bretagne	0,2%	0,8%	0,9%

La variable région semble donc être une variable discriminante dans l'étude de l'absentéisme. La majorité des salariés étant concentrés en Île-de-France ou en Provence-Alpes-Côte d'Azur, le taux d'absentéisme est relativement faible. Néanmoins, des régions comme le Grand Est voit leur absentéisme se dégrader significativement entre 2018 et 2019.

Les différents secteurs d'activité de l'industrie sont également touchés de façon hétérogène.

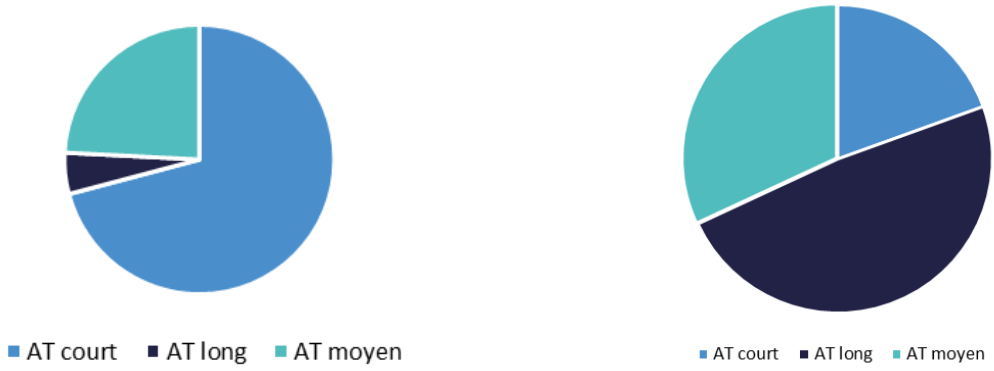
Tableau 12 : Fréquence, gravité et taux d'absentéisme en fonction de l'activité et de l'année

	2017			2018			2019		
	Freq	Grav	Tx	Freq	Grav	Tx	Freq	Grav	Tx
Manutention	0,06%	25,18	1,62%	0,27%	12,76	3,46%	0,26%	29,46	7,69%
Gestion d'infrastructure	0,03%	32,98	0,87%	0,39%	12,15	4,73%	0,17%	28,36	4,93%
Secrétariat et gestion informatique	0,04%	26,23	1,00%	0,19%	14,42	2,69%	0,20%	24,36	4,87%
Technicien	0,06%	23,40	1,37%	0,17%	18,75	3,15%	0,17%	21,07	3,68%
Manufacture	0,03%	31,10	0,98%	0,13%	20,96	2,78%	0,14%	25,09	3,40%
Direction et hiérarchie	0,03%	30,16	0,91%	0,14%	11,36	1,63%	0,21%	14,65	3,13%
Ingénierie	0,05%	22,82	1,07%	0,18%	15,02	2,65%	0,14%	22,88	3,12%
Télécommunication	0,12%	7,40	0,89%	0,25%	6,55	1,64%	0,16%	13,86	2,22%
Transport	0,00%	0,00	0,00%	0,00%	0,00	0,00%	0,00%	0,00	0,00%

Si la gravité a augmenté pour l'ensemble des activités, la fréquence elle, semble avoir augmenté pour les secteurs contenant une majorité d'Ouvriers, mais diminué pour les secteurs embauchant plus d'ETAM et de Cadres. C'est le cas pour la gestion d'infrastructure, secrétariat et gestion informatique ainsi que l'ingénierie. Il semble donc que **les individus tombent moins fréquemment en arrêt de travail. Cependant la durée de ces arrêts a augmenté**.

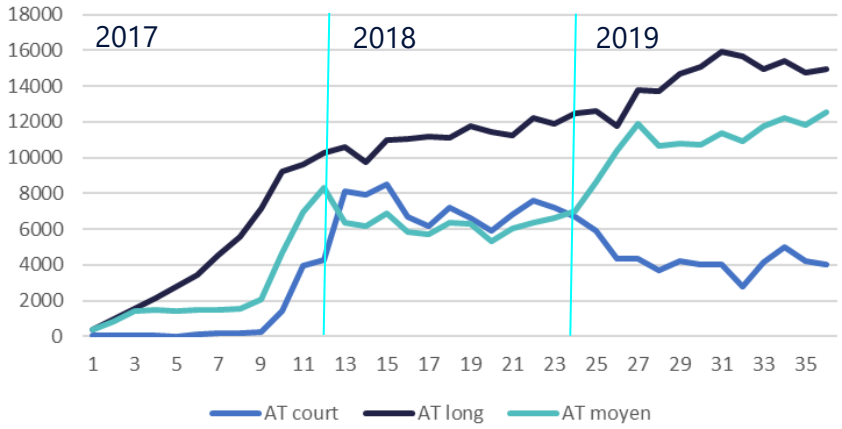
Pour s'en convaincre, on étudie la durée des différents arrêts rencontrés dans la base de données. Pour l'ensemble de ce mémoire, **les arrêts de travail de courte durée** correspondront à **une durée inférieure à 14 jours**, **les arrêts de travail de durée moyenne entre 14 et 90 jours** et **les arrêts longs de plus de 90 jours**. La distinction entre ces 3 types d'arrêts permet de montrer la différence entre le nombre de chaque type d'arrêt et le nombre moyen de jours pour chaque type d'arrêt :

Figure 29 : A gauche, répartition du nombre d'arrêt de travail en fonction de leur durée, à droite, répartition du nombre de jours d'arrêt en fonction du type d'arrêt



Les graphiques montrent donc que **la part des nombres de jours en arrêt de longue durée est conséquente malgré le peu de déclaration d'arrêts longs**. Or, la dégradation des différentes statistiques a permis de montrer que les salariés du portefeuille, en particulier chez les ETAM et les Cadres entraient moins souvent en arrêt de travail, cependant la durée de celui-ci a augmenté. Cette constatation peut être confirmée par le graphique suivant :

Figure 30 : Evolution mensuelle entre 2017 et 2019 du nombre de jours d'arrêt de travail en fonction du type d'arrêt



Tandis que les arrêts longs continuent d'augmenter, on remarque **une baisse du nombre de jours pour les arrêts courts, substitué par l'augmentation du nombre de jours d'arrêt pour arrêt de travail de durée moyenne.**

En conclusion, à partir des diverses statistiques présentées, il semble que ce portefeuille connaisse une dégradation de son absentéisme en termes de gravité des arrêts de travail. Le portefeuille étant resté stable en termes de démographie entre 2017 et 2019, les variables

explicatives à disposition pour expliquer cette dégradation peuvent être les données de santé, dont la consommation a augmenté entre 2017 et 2019, mais également les distinctions possibles entre salariés, à savoir leurs caractéristiques propres mais également leur environnement de travail.

Les différentes statistiques ayant été présentées, l'objectif dans la suite de ce mémoire sera de mettre en place une méthodologie basée sur les modèles de Machine Learning et de Forecasting afin de prédire et d'expliquer à partir de ces données l'absentéisme au sein de ce portefeuille.

1.3.5 Règlementation sur le traitement des données

Les différentes données présentées dans les parties précédentes sont des informations à caractère confidentiel. En effet le traitement de données de santé, d'arrêts de travail, d'informations personnelles sur les salariés sont des données qui doivent être traitées avec précaution et en respectant les règlementations en vigueur en termes de traitement et d'utilisation de données.

Le Règlement Général sur la Protection des Données (RGPD) responsabilise les organismes publics et privés qui traitent des données personnelles. Règlement européen du 25 mai 2018 poursuivant les prémices de la Loi française Informatique et Libertés de 1978, celui-ci s'applique à toute structure privée ou publique effectuant de la collecte et/ou du traitement de données, et ce quel que soit son secteur d'activité et sa taille. Le règlement s'applique à tous les organismes établis sur le territoire de l'Union Européenne, mais aussi à tout organisme implanté hors de l'UE mais dont l'activité cible directement des résidents européens ou toute structure qui traiterait ou collecterait des données personnelles pour le compte d'une autre entité. En ce sens, le travail de l'actuaire est alors directement impacté puisque le traitement de données personnelles fait partie intégrante du métier. Il est alors important de mettre en place le nécessaire afin de protéger les données dites « sensibles », d'utiliser uniquement les données utiles, d'anonymiser toutes données non nécessaires au cadre de l'étude et qui permettraient l'identification d'individus afin de respecter ces nouvelles directives. En particulier, chez les assureurs, la nomination d'un délégué à la protection des données (DPO) est obligatoire compte tenu de leur activité, celui-ci accompagnant la conformité en termes d'utilisation des données.

Les activités de l'actuaire en matière de traitement de données personnelles sont également régies par la Norme de Pratique relative à l'utilisation et la protection des données massives, des données personnelles et des données de santé à caractère personnel **(NPA 5).** Cette norme adoptée par l'Institut des actuaires le 16 novembre 2017 en tant que norme professionnelle de catégorie 3, a pour objectif de définir un cadre de travail dont l'actuaire doit respecter. En outre, l'actuaire doit avoir un comportement bienveillant envers les informations qu'il possède. Il doit se charger de protéger les données, de les anonymiser (en particulier pour les données de

santé), d'énoncer les hypothèses et les limites des modèles utilisés pour le traitement de ces données et de rendre compte des résultats.

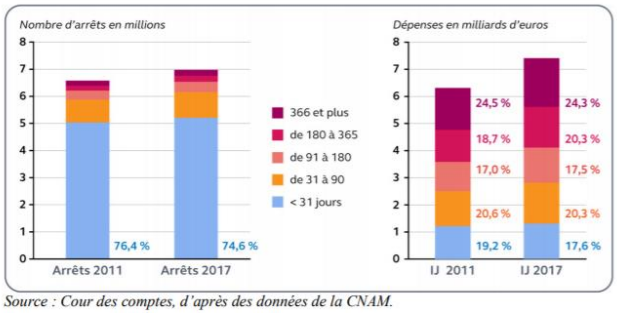
De ce fait, compte tenu des données à disposition afin d'étudier l'absentéisme sur un portefeuille appartenant au monde de l'industrie, les règlementations présentées ci-dessus doivent donc être respectées. Ainsi, l'ensemble des informations personnelles ou de santé ont été préalablement anonymisées afin de ne permettre aucune identification d'un individu. Pour rappel, il faut bien distinguer la technique de pseudonymisation et celle d'anonymisation complète. La technique de pseudonymisation consiste via une interface ou un fichier de codage à par exemple donner un identifiant à chaque individu à la place d'informations sensibles telles que l'adresse, le nom ou le prénom. Cependant cette technique est bien réversible par l'obtention de l'interface ou du fichier contenant les clés de jointure. La technique d'anonymisation elle consiste en revanche à supprimer l'ensemble des informations permettant de retrouver un individu. Dans ce processus d'anonymisation, deux aspects doivent donc être vérifiés, à savoir la suppression des données sensibles, mais également vérifier que des sous-segments du portefeuille ne peuvent pas être identifiés en raison du faible nombre d'individus le composant. Par exemple, un secteur d'activité comportant uniquement deux individus ne peut être conservé en raison du nombre trop faible d'individus représentant ce groupe, permettant facilement de remonter à leurs identités. Dès que cette opération d'anonymisation est réalisée, l'assureur peut conserver les données ad vitam eternam et des études statistiques peuvent alors être menées sans consentement de l'organisme partageant ces données.

Ainsi seules les données nécessaires sont conservées et les différentes hypothèses, limites des modèles et résultats seront présentés tout au long de ce mémoire afin de rendre un travail en accord avec les directives précédemment présentées.

Synthèse première partie : Contexte enjeux et cadre de l'étude réalisée

Le risque absentéisme au sein des entreprises est devenu depuis ces dernières années un phénomène préoccupant et ce pour différents acteurs. Outre l'impact direct de l'absentéisme sur la productivité des entreprises en lien avec l'absence des salariés, ce phénomène touche également les organismes couvrant ce risque, à savoir le système de protection sociale ainsi que les assureurs. Depuis ces dernières années, les indicateurs de l'absentéisme ont montré une amplification de ce phénomène. Le taux d'absentéisme, un indicateur global et composite de différentes informations sur l'absentéisme, n'a fait qu'augmenter ces dernières années.

$$Taux\ d'absentéisme = \frac{nombre\ de\ jours\ en\ arrêt}{nombre\ de\ jours\ d'exposition}$$

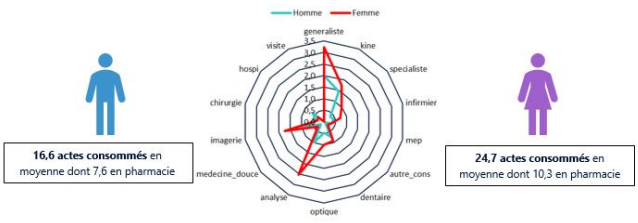


Source : Cour des comptes, d'après des données de la CNAM.

Le nombre d'arrêts, leurs durées et de ce fait leurs coûts ont augmenté malgré un taux de population active constant. Ce phénomène entraîne donc un coût de plus en plus conséquent à supporter pour les organismes prenant en charge ce risque.

L'étude de l'état de santé des entreprises face à ce risque mais également la mise en place de nouvelles méthodes de modélisations afin de comprendre les raisons d'un tel phénomène pour pouvoir le prédire et s'en prémunir sont les nouveaux enjeux pour l'ensemble des acteurs touchés par ce risque.

L'utilisation de bases données dans le respect des réglementations en vigueur concernant la protection des données, permet de réaliser les premières études statistiques. **Ces données proviennent du secteur de l'industrie sur une période de 3 ans entre 2017 et 2019, rassemblant des données sur les individus, les entreprises, leurs consommations santé et des données d'arrêt de travail, préalablement retraitées à partir des données DSN.**



La première étude statistique permet de montrer l'évolution du portefeuille étudié au cours des années d'historique, mais également de montrer des sous-segments en fonction de certaines caractéristiques, mettant en lumière des impacts différents en termes d'absentéisme. La consommation de prestation santé des individus est également étudiée, car celle-ci peut avoir un impact sur l'absentéisme. **L'effectif exposé étant relativement stable au cours de années entre 2017 et 2019, il est possible d'observer une augmentation du taux d'absentéisme ainsi que de la consommation santé. Les femmes consomment plus**

que les hommes et le niveau d'absentéisme est différent en fonction du secteur d'activité, de la catégorie socio-professionnelle mais également de la région. L'absentéisme est donc plus élevé chez les ouvriers. Une modification des arrêts de travail observés montre que les arrêts de courtes durée ont laissé place à des arrêts de plus longue durée entre 2018 et 2019 montrant une aggravation de l'absentéisme sur ce portefeuille.

À la suite de la présentation des données et de ce premier diagnostic, faisant état d'un absentéisme différent en fonction des caractéristiques des individus**, une partie modélisation a été réalisée afin de comprendre les liens existants entre absentéisme, caractéristiques des individus, des entreprises et consommation de santé.**

	2017			2018			2019		
	Freq	Grav	Tx	Freq	Grav	Tx	Freq	Grav	Tx
Manutention	0,06%	25,18	1,62%	0,27%	12,76	3,46%	0,26%	29,46	7,69%
Gestion d'infrastructure	0,03%	32,98	0,87%	0,39%	12,15	4,73%	0,17%	28,36	4,93%
Secrétariat et gestion informatique	0,04%	26,23	1,00%	0,19%	14,42	2,69%	0,20%	24,36	4,87%
Technicien	0,06%	23,40	1,37%	0,17%	18,75	3,15%	0,17%	21,07	3,68%
Manufacture	0,03%	31,10	0,98%	0,13%	20,96	2,78%	0,14%	25,09	3,40%
Direction et hiérarchie	0,03%	30,16	0,91%	0,14%	11,36	1,63%	0,21%	14,65	3,13%
Ingénierie	0,05%	22,82	1,07%	0,18%	15,02	2,65%	0,14%	22,88	3,12%
Télécommunication	0,12%	7,40	0,89%	0,25%	6,55	1,64%	0,16%	13,86	2,22%
Transport	0,00%	0,00	0,00%	0,00%	0,00	0,00%	0,00%	0,00	0,00%

Dans la deuxième partie, de nouvelles modélisations provenant des méthodes de Machine Learning sont présentées permettant de **proposer une segmentation du portefeuille en fonction des diverses variables présentées précédemment, mais également une première modélisation d'un indicateur de l'absentéisme : le taux d'absentéisme**. Cette modélisation cherchera à prédire le taux d'absentéisme permettant d'avoir une première estimation de cet indicateur à l'aide des renseignements sur les individus composant le portefeuille.

2. MODELISATION DU RISQUE ABSENTEISME PAR MACHINE LEARNING

Après l'étude des données à disposition afin d'avoir une première approche du risque absentéisme au sein des entreprise, cette nouvelle partie consiste à étudier la modélisation du risque absentéisme via de nouvelles **méthodes de Machine Learning**.

Comme vu précédemment, l'absentéisme peut être abordé selon différents angles. Que ce soit sur la fréquence, la gravité, la proportion d'absence ou directement le taux d'absentéisme, ces informations apportent des informations différentes sur le risque. Une modélisation du risque selon ces différentes données à l'aide de variables explicatives permettrait à la fois de **comprendre davantage la structure sous-jacente existante liée au risque** dans les différentes entreprises, mais également de donner **un outil de prédiction du risque**.

L'utilisation des méthodes de Machine Learning à la place des modélisations statistiques classiques, comme les GLM par exemple, repose sur le choix de modèles donnant **plus d'importance à l'apprentissage des données et à leurs prédictions non linéaires sans hypothèse sur la loi de distribution suivie par la variable cible**. En effet, l'absentéisme est un risque complexe, difficile à modéliser étant donné les nombreuses composantes qui peuvent influer sa survenance.

Un premier objectif est donc, à partir des méthodes de Machine Learning, de donner une segmentation des profils de risque afin **d'identifier des groupes d'individus remarquables, mais également de donner les variables qui ont une influence marquée sur ce risque**. **La construction de modèles de prédiction de l'absentéisme** dans un second temps permettra de faire le constat que des modèles ne prenant en compte que légèrement **la composante temporelle de l'absentéisme** ne permettent pas d'obtenir des résultats robustes, en particulier pour des observations du risque élevées.

2.1.1 Généralités sur les méthodes de Machine Learning

Les méthodes de *Machine Learning* ont considérablement modifié l'approche du traitement des données ces dernières décennies. Ces méthodes ne se focalisent plus sur la « modélisation » à proprement parler, mais se concentrent davantage sur **l'importance des données** en elles-mêmes.

Avec l'avènement du *Big Data*, les méthodes classiques de modélisation statistique peuvent rencontrer certains problèmes pouvant être résolus par ces nouvelles méthodes de *Machine Learning*. En effet, certains problèmes viennent mettre en cause la possibilité d'application des modèles classiques statistiques, basés sur deux hypothèses :

- **Le rapport des dimensions n et p est raisonnable**, à savoir, le nombre d'observations n doit être très grand par rapport au nombre de variables p. De plus, bien que n doive être le plus grand possible afin de pouvoir converger vers un résultat

asymptotique, le nombre de variables p lui ne doit pas être trop grand. En effet les estimateurs du maximum de vraisemblance conservent leur propriété de normalité asymptotique uniquement si $\frac{p^2}{n} \to 0$ lorsque $p, n \to \infty$. De ce fait, dès lors que $p > \sqrt{n}$, cette propriété n'est plus vérifiée.

- **Les hypothèses distributionnelles sont vérifiées**, les données ou les résidus doivent suivre des distributions appartenant à la famille exponentielle.

Si ces conditions ne sont pas vérifiées, d'autres méthodes peuvent alors venir concurrencer les méthodes de statistiques classiques tirées du modèle linéaire général.

Or, lors de cette étude, comme montré dans la première partie, les bases de données construites sont de grandes dimensions avec plus de 200 variables pour un petit peu plus de 42 000 observations. **La grande dimensionnalité des données empêche donc d'utiliser les méthodes de statistique classique** au risque d'obtenir des résultats aberrants.

Le choix d'une méthode de *Machine Learning* semble alors judicieux, permettant ainsi de traiter un nombre de données conséquent, **sans formuler d'hypothèses sur la distribution de la variable cible à prédire** et permettant également de mettre en lumière l'explicabilité des données en utilisant un modèle non linéaire.

Le principe du *Machine Learning* est, comme son nom l'indique, une méthode d'apprentissage artificiel. La méthode de *Machine Learning* comporte généralement deux phases :

- **Une phase dite d'apprentissage** permettant de calculer les paramètres du modèle sur une base d'observations (en général entre 50 et 70% de la base de données est utilisée à l'apprentissage du modèle)
- **Une phase dite de validation**, sur laquelle le modèle est testé afin de vérifier que les résultats prédits par le modèle entraîné soient cohérents avec les observations de la base[6]

Il existe différentes méthodes d'apprentissage, mais les principales sont **l'apprentissage supervisé et l'apprentissage non supervisé**. L'apprentissage non supervisé cherche généralement à trouver la structure sous-jacente des données, tandis que l'apprentissage supervisé cherche à classifier dans le cas d'une variable cible catégorielle et à faire une régression dans le cas d'une variable cible continue. Dans cette étude on se place dans un cadre supervisé.

L'utilisation des arbres de décision ainsi que des forêts aléatoires vont permettre de **réaliser une segmentation des individus en fonction des variables cibles de l'absentéisme, mais également de proposer des modélisations à partir de variables explicatives.** Une

[6] en général entre 30 et 50% des données sont exploitées. 20% des données peuvent parfois servir en tant que base de test afin d'étudier la robustesse du modèle

première phase consistant à expliquer ces nouvelles méthodes de *Machine Learning* convergeront ensuite vers la construction et l'interprétation des résultats obtenus.

2.1.2 Les méthodes par arbres

L'algorithme de *Random Forest* ou de « forêts aléatoires » fait partie de la famille des méthodes dites de « ***Bagging*** ». Diminutif de *« Bootstrap Aggregation »*, cette méthode combine les prévisions sorties à partir de plusieurs modèles afin de donner une prédiction unique pour chaque individu sur la base d'une opération. Généralement cette opération est une moyenne, une moyenne pondérée ou un vote majoritaire. L'avantage de l'utilisation de cette approche est de réduire la variance des prédictions, qui serait plus grande si la prédiction provenait d'un modèle unique.

La particularité du *Random Forest* réside dans la construction de plusieurs arbres de décision afin de moyenner les différentes prédictions de ces modèles.

Les arbres CART, les *weak learners* des Random Forest

Les arbres CART (*Classification And Regression Trees*) sont les modèles les plus simples dans l'ensemble des algorithmes de Machine Learning. Ils sont définis comme étant des *« weak learners »* puisque ce sont des modèles simples, donnant de meilleures prédictions que le hasard. L'objectif de cette partie est donc de présenter l'arbre CART et d'expliquer l'algorithme sous-jacent afin de comprendre par la suite l'algorithme du *Random Forest*.

L'intérêt de l'arbre CART est d'obtenir des groupes homogènes d'individus distincts en réalisant une segmentation des observations sur certains critères. Un arbre de décision est composé principalement de 3 éléments :

- **Une racine**, étant le point de départ de l'arbre, elle contient l'**ensemble de la population** à segmenter
- **Les branches**, représentant **les règles de division** qui permettent de segmenter la population sur des critères portant sur les variables
- **Les feuilles**, contenant les **sous-populations homogènes** à la suite de la segmentation des individus en fonction de leurs caractéristiques, fournissant également l'estimation de la variable cible

Pour expliquer la construction de l'arbre, quelques notations préliminaires sont introduites :

- $i \in [\![1, n]\!]$: numéro de l'individu
- $j \in [\![1, p]\!]$: numéro de la variable
- y_i : réponse observée de la variable cible Y pour l'individu numéro i
- $X_i = (X_{i1}, \dots, X_{ip})$: vecteur des données de l'individu i

- $l \in [\![1, L]\!]$: numéro des feuilles de l'arbre
- $E_n[Y]$ la moyenne empirique de Y
- ∇ l'ensemble des covariables, ∇_m le sous-ensemble associé au nœud m et $\nabla_{pa(m)}$ le sous-ensemble associé au nœud parent de m
- R_m la règle de segmentation relative au nœud m avec $R_m(x) = 1(x \in \nabla_m)$

Lors d'une régression, la valeur de la variable cible attendue est la suivante :

$$\pi_0(x) = E_0[Y|X = x].$$

En supposant que la régression suive une relation linéaire, l'estimateur de la variable cible se note :

$$\hat{\pi}(x) = \widehat{\beta_0} + x^T\hat{\beta}.$$

Puis les paramètres de la régression sont généralement estimés par Moindres Carrés Ordinaire (MCO). Cependant, les arbres CART sont des estimateurs différents puisque ceux-ci découpent l'espace des covariables successivement afin de créer un estimateur à chaque division de l'espace. Le calcul de l'ensemble des estimateurs potentiels n'est donc pas possible en termes de volume de calculs. Cependant pour contourner ce problème, un arbre dit maximal est construit, générant une suite d'estimateurs. L'arbre maximal est donc l'estimateur par morceaux le plus complexe de la suite d'estimateurs construits puisque celui-ci ne contient que des feuilles terminales avec un seul individu ou uniquement des individus ayant les mêmes caractéristiques.

Afin de construire l'arbre maximal, l'algorithme adopte un critère de division basé sur le critère d'homogénéité qui soit cohérent avec l'estimation de la valeur de la variable cible. L'estimation se fait toujours par MCO par la fonction de coût φ :

$$\pi_0(x) = arg\min_{\pi(x)} E_0[\varphi(Y, \pi(x))|X = x] \qquad \text{où} \qquad \varphi\big(Y, \pi(x)\big) = (Y - \pi(x))^2.$$

La différence ici réside dans l'estimation de $\pi_0(x)$ à chaque étape de la ramification de l'arbre.

L'enchaînement des étapes de la construction de l'arbre sont les suivantes :

- Toute la population est contenue dans la racine
- Recherche de la meilleure première segmentation donnant le meilleur gain d'homogénéité
- Segmentation de la population initiale en deux nœuds fils
- Réitération du processus précédent sur chaque nœud fils
- Réitération de l'ensemble du processus précédent jusqu'à obtenir un individu par feuille ou des feuilles contenant des individus ayant les mêmes caractéristiques et ne pouvant donc pas être segmentés davantage

Par construction, par le critère d'homogénéité lors de la segmentation, l'hétérogénéité diminue à chaque étape. L'arbre étant construit, il ne reste plus qu'à ajouter les valeurs des régressions

successives à chaque nœud ou feuille de l'arbre. L'arbre est alors associé à la fonction de régression suivante :

$$\hat{\pi}(x) = \sum_{m=1}^{M} \hat{\beta}_m^{tree} R_m(x) \qquad \text{avec} \qquad \hat{\beta}_m^{tree} = \begin{cases} E_n[Y|x \in \nabla_m], \, si \; m \neq racine \\ E_n[Y] \; sinon \end{cases}.$$

En régression classique cela revient à rechercher les coefficients tels que :

$$\hat{\beta}_m^{tree} = \arg\min_{\beta^{tree}} E_n[(Y - \sum \beta_m^{tree} R_m(x))^2].$$

De ce fait pour chaque feuille de l'arbre en sommant sur l'ensemble des nœuds :

$$\hat{\pi}^L(x) = \sum_{l=1}^{L} \hat{\gamma}_l R_l(x) \;\; \text{avec} \quad \hat{\gamma}_l = E_n[Y|x \in \nabla_l].$$

Ainsi, **un arbre est un estimateur par morceaux, où chaque morceau est une feuille dont la valeur est la moyenne empirique des valeurs de Y appartenant à cette feuille**.

Cependant l'arbre maximal n'est généralement pas le plus optimal. La construction de l'arbre maximal étant récursive, il est également possible de générer une suite d'estimateurs par morceaux depuis la racine. On note alors la suite $\{\Pi^K\}$ les sous-espaces telles que

$$\Pi^K = \{\pi^L(.) = \sum_{l=1}^{L} \gamma_l R_l(.) : L \leq K\}.$$

Pour K fixé, on cherche l'estimateur du sous-arbre appartenant à l'espace Π^K par :

$$\hat{\pi}^K(x) = \arg\min_{\gamma=(\gamma_{1,\dots,\gamma_L})} E_n[\varphi(Y, \pi^L(x))].$$

Cette suite permet de trouver des estimateurs par morceaux qui ne soient pas trop complexes et ainsi trouver le sous-arbre optimal en termes d'adéquation et de prévision. La sélection du modèle va donc se faire par le biais d'un paramètre α, nommé paramètre de complexité et le paramètre K, la dimension de l'estimateur ou le nombre de feuilles. L'arbitrage d'adéquation à vérifier est donc :

$$\Delta_\alpha(\hat{\pi}^K(x)) = E_n[\varphi(Y, \hat{\pi}^K(x))] + \alpha \frac{K}{n}.$$

De ce fait, pour α fixé, l'estimateur final optimise un critère coût-complexité :

$$\hat{\pi}_\alpha^K = \arg\min_{(\hat{\pi}^K)_{K=1,\dots,K(n)}} \Delta_\alpha(\hat{\pi}^K(x)).$$

A noter que si $\alpha = \infty$ alors l'estimateur est la racine de l'arbre, et pour $\alpha = 0$ l'arbre maximal. Pour obtenir ensuite le sous-arbre optimal parmi l'ensemble des α disponibles, une procédure d'élagage est réalisée puisqu'il n'est pas envisageable de parcourir l'ensemble des sous-arbres de l'arbre maximal. La procédure est la suivante :

- **Construire l'arbre maximal**
- **Considérer une première valeur de α** conduisant à prendre un sous-arbre optimal de l'arbre maximal

- A partir de ce sous-arbre optimal, **prendre une valeur de α plus grande consistant à remonter le nouvel arbre construit** précédemment
- Une suite de $\{\alpha_z\}$ est alors construite

$\hat{\alpha}$ est choisi pour obtenir l'arbre optimal tel que :

$$\hat{\pi}_{\hat{\alpha}}^{K}(x) = \arg \min_{(\hat{\pi}_{\alpha_z}^{K})_{\alpha=\alpha_1,\dots,\alpha_Z}} \Delta_{\alpha_z}\left(\hat{\pi}_{\alpha_z}^{K}(x)\right).$$

Généralement $\hat{\alpha}$ est choisi par la méthode de validation croisée afin de minimiser les erreurs de généralisation.

La création de ces arbres CART va alors permettre à la fois **de construire une segmentation des salariés du portefeuille d'étude, mais également de mettre en place une stratégie d'agrégation afin d'obtenir des résultats plus robustes, tout en réduisant la variance des prédictions au travers des forêts aléatoires.**

Le Bagging et l'algorithme de Random Forest

Comme expliqué en début de cette partie, l'objectif du *Random Forest* est de proposer un estimateur qui résulte de la moyenne des estimateurs par morceaux construits au-dessus. Cette moyennisation va alors permettre d'améliorer la robustesse de l'estimation de la valeur de la variable cible prédite.

Le taux d'absentéisme étant une variable cible continue, l'estimateur obtenu par le *Random Forest* pour l'individu i peut alors être exprimé en fonction des estimateurs CART par la relation suivante :

$$\hat{Y}_i^{RF} = \frac{1}{N}\sum_{i=1}^{N} \hat{Y}_i^{CART}.$$

Cependant, il faut définir la construction des différents arbres CART afin d'obtenir la moyenne des estimateurs. Les forêts aléatoires se basent sur deux principes, le Bagging et la randomisation. La construction de chaque arbre dans la forêt aléatoire suit alors le processus suivant :

- **Construction d'un échantillon *bootstrap*** de même taille que la base d'apprentissage
- Construction de l'arbre CART sur cet échantillon composé de **k variables** pour chaque individu. Un paramètre m est fixé avec $m < k$ (qui reste inchangé lors de la construction de chaque arbre) tel que :
  - Pour chaque nœud, **un tirage aléatoire de m variables sur les k variables** disponibles est effectué
  - **Recherche de la segmentation optimale** basée sur ces k variables
  - **Elagage de l'arbre** (3 stratégies disponibles)

- **Agrégation des arbres** montré dans l'équation précédente afin de construire l'estimateur de la forêt aléatoire

Les différentes méthodes d'élagage possibles pour chaque arbre sont alors les suivantes :

1. Construction de l'arbre maximal pour chaque échantillon bootstrapé, il n'y a donc pas d'élagage
2. Construction d'un arbre d'au plus q feuilles avec q fixé
3. Construction de l'arbre maximal puis élagage optimal par validations croisées

La méthode privilégiée reste la première méthode, assurant un bon compromis au niveau du volume de calculs à effectuer face à la qualité des prévisions qui ont un faible biais et une grande variance. Même si la 3ème méthode semblait être optimale, finalement le gain en fiabilité de prédiction n'est pas aussi grand comparé aux volumes des calculs supplémentaires à effectuer.

Les deux principaux hyperparamètres à fixer sont donc le nombre d'arbres construits dans la forêt et le nombre de variables m tirées aléatoirement parmi l'ensemble des variables disponibles à chaque construction d'arbre. Ces deux paramètres seront respectivement appelés *ntree* et *mtry* dans la suite de ce mémoire.

La méthode de *Random Forest* permet alors de mettre un algorithme simple en place, permettant la réduction de la variance de l'estimateur par les méthodes de *Bagging* et de randomisation qui **réduit également le problème des variables corrélées**. La méthode de *Random Forest* permettra dans un deuxième temps de dégager un classement robuste du pouvoir explicatif de chacune des variables dans la partie suivante de modélisation.

2.2. Construction des modèles et résultats

2.2.1 Introduction à la construction des modèles

L'objectif de cette partie est de mettre en place une méthodologie permettant de **segmenter les salariés du portefeuille** selon le risque absentéisme, puis de **proposer une modélisation du taux d'absentéisme afin d'identifier les variables influentes dans la prédiction du risque**.

Ces différents indicateurs sont calculés pour chaque salarié du portefeuille à partir des bases de *Machine Learning* et *Forecasting* afin de construire les modélisations. Bien que ces informations soient donc au niveau individuel, cette partie ne prétend pas donner une prédiction du niveau d'absentéisme individu par individu. En raison des données sensibles à disposition telles que les données de santé ou d'absentéisme, **l'objectif ici n'est pas d'individualiser les résultats mais bien au contraire de les mutualiser** afin d'obtenir des **visions collectives** à différents

niveaux du risque. Cependant, la construction de ces visions collectives passe par une agrégation des modélisations individuelles qui seront utilisées uniquement dans ce seul but.

Par la suite, deux types de modèles vont être construits :

- Un arbre CART afin de segmenter les individus en fonction de variables explicatives
- Un modèle de *Random Forest* afin de sortir les variables d'importance du modèle et étudier si la modélisation du risque absentéisme est possible

Pour ces différents modèles, la variable cible étudiée sera **le taux d'absentéisme**.

Pour ces différentes variables, il est important de préciser la fenêtre temporelle de calcul. En effet, celles-ci peuvent être agrégées de manière hebdomadaire, mensuelle, trimestrielle ou annuelle. Toutefois, afin de lisser les résultats en agrégeant les données sur une période assez longue, **les données seront calculées annuellement**. L'absentéisme étant une donnée qui peut être qualifiée de donnée haute fréquence sur de courtes périodes, le fait de travailler sur des périodes de longues durées permet ainsi de lisser le côté aléatoire de l'absence au travail. Cela permet d'obtenir des structures stables d'une période sur l'autre sans se soucier de la saisonnalité de telles données qui pourraient varier d'une période à l'autre sur une fenêtre temporelle plus petite.

Afin d'aborder ce problème de prédiction du risque absentéisme, un rapide rappel des données à disposition. La base *Machine Learning* a été construite dans la première partie de ce mémoire. Elle est composée de 42 534 lignes, chacune correspondant à un individu caractérisé par une situation unique, à savoir une catégorie socio-professionnelle et une entreprise où le salarié exerce son activité. Chaque ligne détient les informations sur les caractéristiques propres de l'individu, ses consommations de santé ainsi que ses arrêts de travail et ce pour les années 2017, 2018 et 2019. Au total, la base comporte ainsi 258 variables pour chaque individu.

La première étape consiste donc à regrouper les variables afin de pouvoir tester l'ajout progressif de ces groupes lors de la modélisation. Les variables servant à la modélisation sont réparties dans les différents groupes suivants :

- Caractéristiques des salariés (référence à l'âge, sexe, ancienneté…)
- Informations sur l'entreprise (référence à l'activité professionnelle, pourcentage d'ouvrier dans l'entreprise, taille de l'établissement…)
- Nombre d'actes santé consommés par grand type d'acte pour les 3 années
- Arrêts pour les 3 années (référence au nombre d'arrêts renseignés par motif)

Ces différents regroupements vont permettre de construire plus rapidement les modèles en sélectionnant les types de variables choisis afin de modéliser la variable cible.

Comme vu précédemment, les données utilisées comportent une composante temporelle. Il est alors nécessaire de définir la stratégie mise en place pour les deux modèles envisagés (arbre CART et forêt aléatoire) pour comprendre quelles variables explicatives seront utilisées lors de l'apprentissage mais également la validation des modèles.

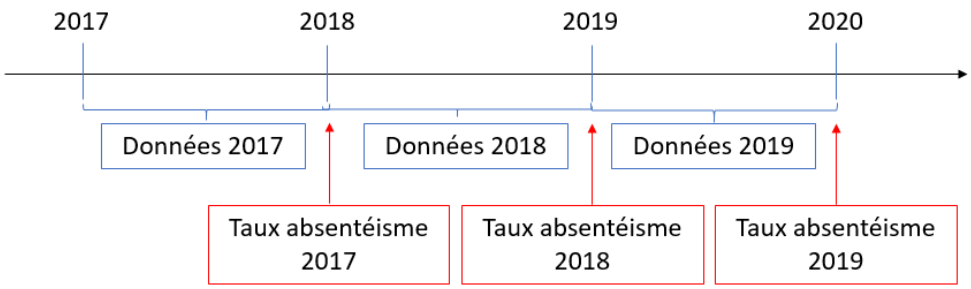
Pour la modélisation par arbre CART, **l'objectif ici est de segmenter les salariés en fonction de leur risque absentéisme** en utilisant des variables explicatives observées sur la même période que la variable cible utilisée. L'objectif ici est de pouvoir comprendre les caractéristiques de l'individu qui ont pu l'amener à être absent sur la période étudiée. Néanmoins, l'ajout de la variable cible calculée sur la période antérieure dans le modèle permettra d'étudier la possibilité que la connaissance de données antérieures à la période d'étude puisse améliorer la segmentation.

En revanche pour la modélisation par forêts aléatoires, deux stratégies peuvent être envisagées :

- **Une première modélisation** proche de la démarche de la partie précédente, vise à **se concentrer sur le calcul des variables d'importance à partir des données des salariés observées sur la même période** de calcul de la variable cible.
- **Une seconde modélisation cherchant à prédire, en fonction des données du passé, la variable cible dans le futur.** De ce fait, les données observées des salariés introduites dans le modèle sont antérieures au calcul de la variable cible.

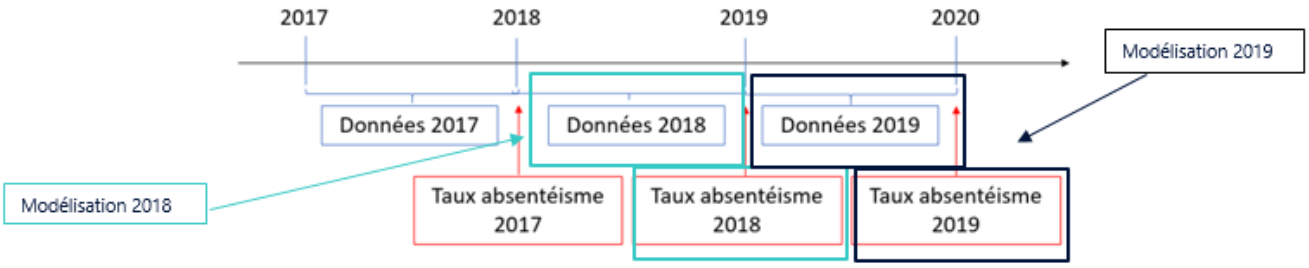
Pour montrer la différence majeure entre ces deux modélisations, une première illustration permet de comprendre le détail des périodes d'observation des données. Pour rappel, les bases de données permettent d'obtenir un historique de données entre 2017 et 2019. Les données en lien avec les salariés (caractéristiques salariés, caractéristiques entreprises, données de santé, données absentéisme) sont observables chaque année. En revanche, les variables cibles sont calculées en fin d'année, comme le montre l'illustration suivante pour le taux d'absentéisme par exemple.

Figure 31 : Détail de la temporalité des données



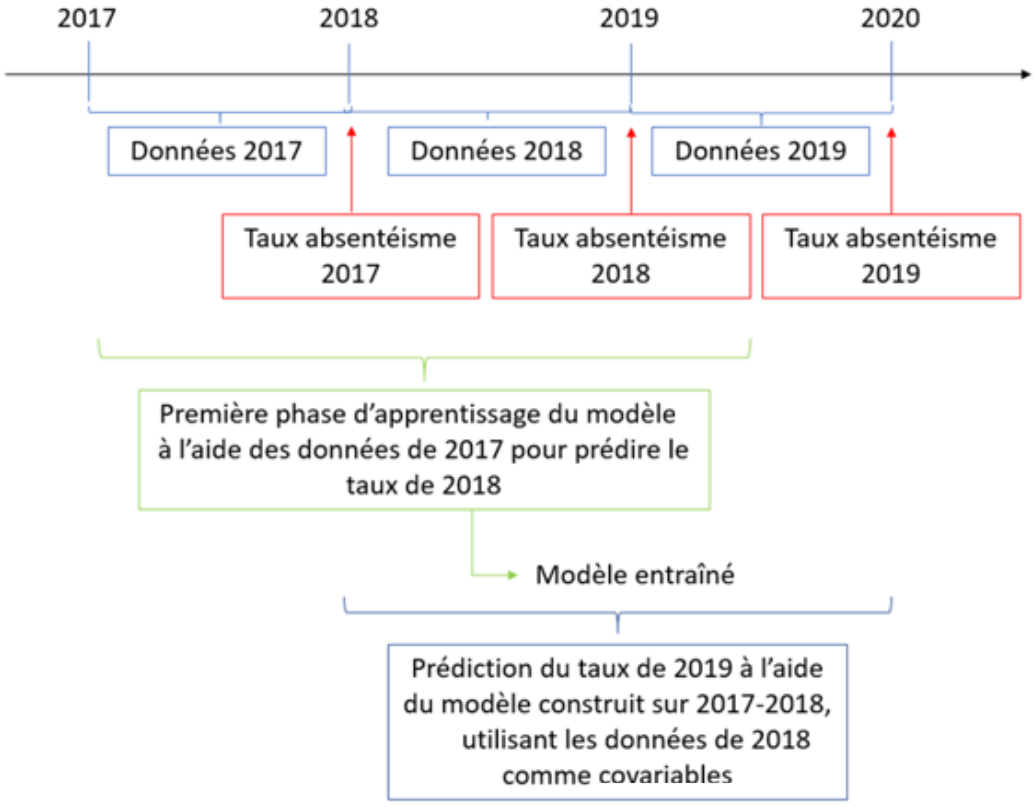
Le premier type de modélisation consiste donc à **expliquer la variable cible à partir des données observées cette même année afin de comprendre l'impact des diverses variables explicatives** dans le modèle et d'en dégager les variables d'importance.

Figure 32 : Premier type de modélisation pour les arbres CART et forêts aléatoires



En revanche, le deuxième type de modélisation pour les forêts aléatoires cherche à **réaliser une prédiction dans le futur de la variable cible du risque absentéisme à partir de données observées dans le passé.** Ainsi, réaliser une prédiction à l'année N, implique que les données des salariés N ne sont pas encore observables et de ce fait uniquement les données à l'année N-1 peuvent être introduites dans le modèle. La démarche de prédiction via les forêts aléatoires peut donc être résumé par l'illustration suivante :

Figure 33 : Second type de modélisation pour les forêts aléatoires (prédiction)



Enfin, dans le but de valider les modèles, deux métriques sont introduites dans la suite de ce mémoire :

- **L'erreur quadratique moyenne (RMSE)** donnée par

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

où y_i représente la valeur réelle observée de l'individu i pour la variable Y et $\hat{y}_i$ la valeur prédite. Ainsi plus la valeur du RMSE est faible, plus les prédictions sont précises.

- **Le Q^2**, qui au même titre du R^2 dans les modèles linéaires permet de donner **le pouvoir prédictif du modèle,** calculé comme suit :

$$Q^2 = 1 - \frac{RMSE^2}{Var(Y)}$$

Plus cette valeur est proche de 1 plus le modèle a un pouvoir prédictif important. Le Q^2 permet de **comparer objectivement des distributions différentes sur les bases d'apprentissage et de validation,** contrairement au RMSE.

2.2.2 Segmentation par arbre CART

Cette première modélisation par arbre CART est l'occasion de segmenter les salariés en différents **groupes de risque en lien avec l'intensité de l'absentéisme**. La segmentation va faire intervenir différents groupes de variables afin de comprendre la structure sous-jacente de l'absentéisme en fonction de ces différentes variables explicatives.

La segmentation par CART va donc porter sur le taux d'absentéisme. Cette modélisation sera menée à la fois sur 2018 et 2019. Afin de pouvoir comparer les différents groupes d'individus issus des arbres CART, il est nécessaire de donner **l'individu de référence** sur ce portefeuille, soit l'individu le plus représenté et exposé dans la base de données avec des caractéristiques spécifiques. La comparaison de cet individu de référence avec les groupes créés peut apporter des pistes de réflexion sur les différentes intensités d'absentéisme calculées. L'individu de référence est un homme marié de 40 ans, ayant environ 14 ans d'ancienneté, habitant en Ile-de-France et travaillant généralement dans le secteur de la manufacture dans des entreprises possédant en moyenne 258 salariés dont 42% d'ouvriers.

Différentes segmentations sont alors réalisées à l'aide de différents groupes de variables sur taux d'absentéisme comme variable cible, l'objectif étant de proposer différentes groupes de risque. Une première étape consiste à étudier le taux d'absentéisme de 2018 et 2019 selon les différents types de variables à disposition. Une première segmentation est alors proposée en fonction des caractéristiques propres des salariés, à savoir leurs caractéristiques personnelles ainsi que l'environnement dans lequel ils travaillent. Pour rappel, chaque feuille contient les informations suivantes : le taux d'absentéisme moyen ainsi que le pourcentage d'effectif contenu dans cette feuille. Les différentes feuilles sont ensuite regroupées en groupes en fonction du taux d'absentéisme moyen résultant de chaque feuille, avec en rouge un taux d'absentéisme élevé, en jaune modéré et en vert faible.

Que ce soit sur 2018 ou 2019, les variables catégorie socio-professionnelle ainsi que la localisation de l'entreprise sont des variables influentes dans la segmentation puisqu'elles se trouvent dans les premières étapes de segmentation dans l'arbre de décision. Pour rappel, les statistiques réalisées lors de la première partie de présentation des données montraient ces différences de taux d'absentéisme entre ces différentes modalités.

Figure 34 : Segmentation sur 2018 en fonction du taux d'absentéisme et des variables propres aux salariés

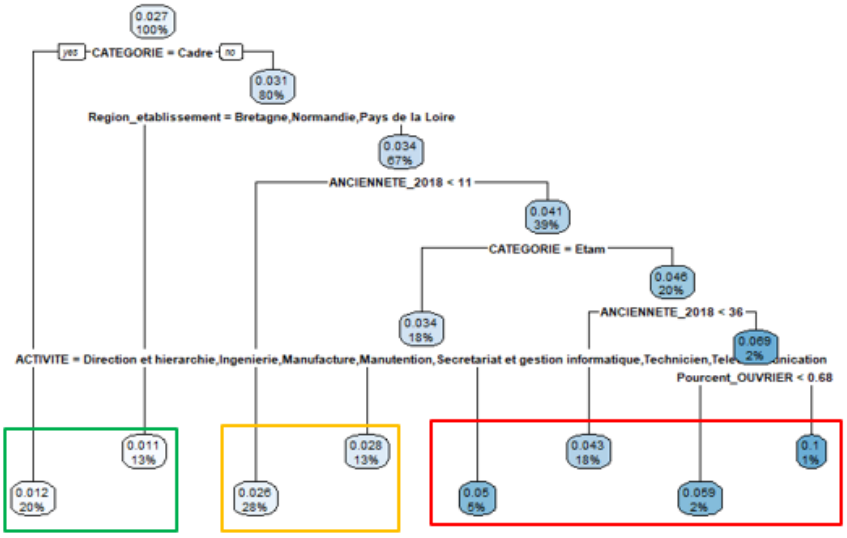
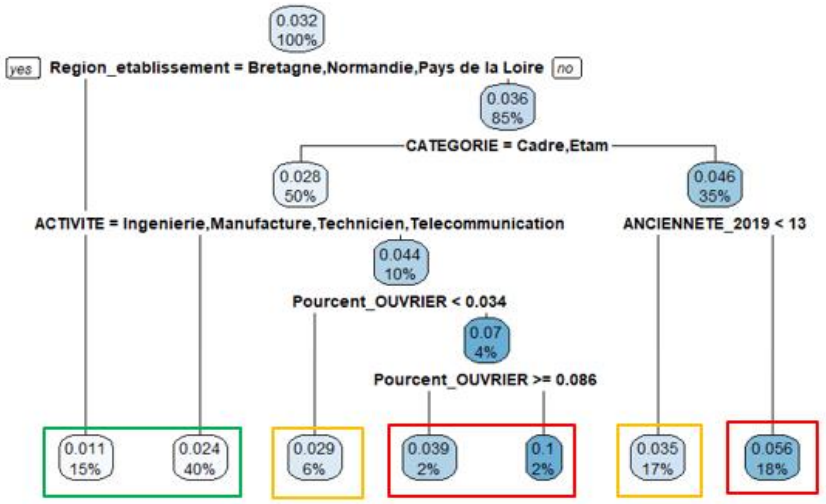


Figure 35 : Segmentation sur 2019 en fonction du taux d'absentéisme et des variables propres aux salariés



Une seconde segmentation utilisant les nombres d'actes consommés par grand poste en santé est ensuite réalisée. Que ce soit sur la modélisation de 2018 ou de 2019, les postes de soin généraliste, hospitalisation ainsi que kiné sont les trois premiers postes qui segmentent les salariés. En général, plus le nombre d'actes consommés dans l'ensemble des postes de soins santé est élevé, plus le taux d'absentéisme du salarié semble élevé, que ce soit sur 2018 ou 2019.

Figure 36 : Segmentation sur 2018 en fonction du taux d'absentéisme et de la consommation santé

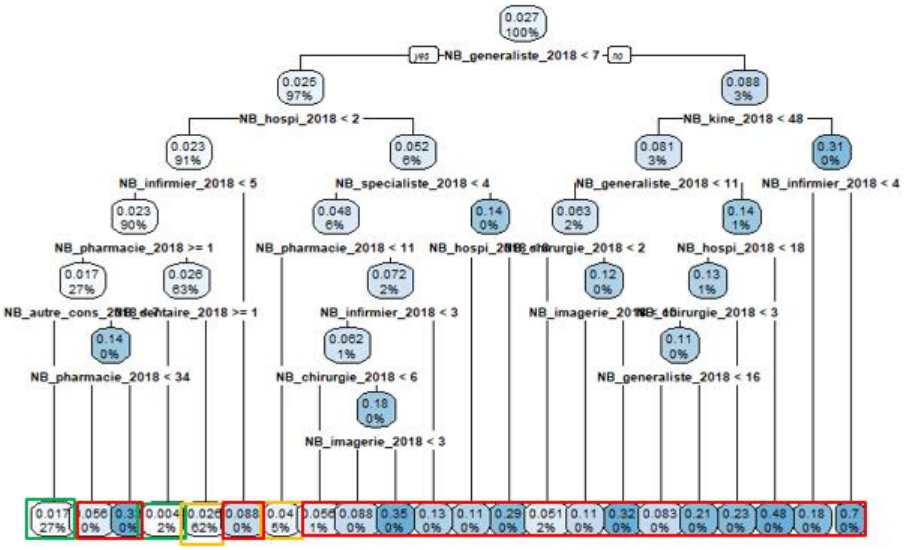
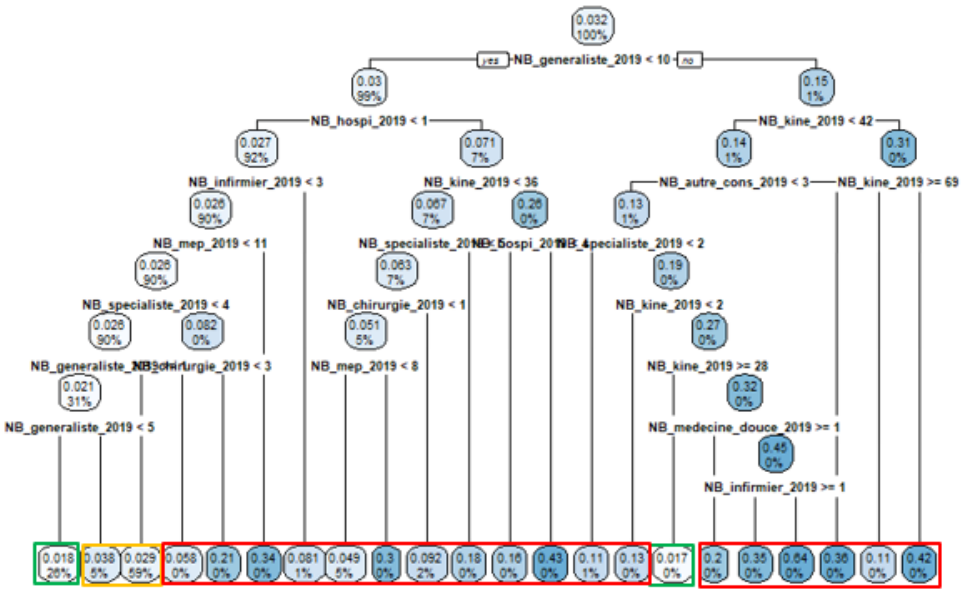


Figure 37 : Segmentation sur 2019 en fonction du taux d'absentéisme et de la consommation santé



Une dernière segmentation sur le taux d'absentéisme regroupant l'ensemble des variables à disposition permet d'étudier une segmentation prenant en compte l'interaction entre ces différentes variables. L'étude va également s'intéresser aux données rappelant l'absentéisme observé dans le passé. Néanmoins, l'année de 2017 ayant connu une montée en charge des données DSN, seul le taux d'absentéisme de l'année précédente est ajouté dans la modélisation de 2018, tandis que le nombre de jours d'absence en 2018 est également ajouté dans la modélisation de 2019.

Il semble que les données temporelles passées du risque absentéisme semblent importantes dans la modélisation du taux d'absentéisme puisque ces variables permettent de diminuer l'hétérogénéité des groupes dès les premières étapes de segmentation.

Figure 38 : Segmentation sur 2018 en fonction du taux d'absentéisme et de l'ensemble des variables

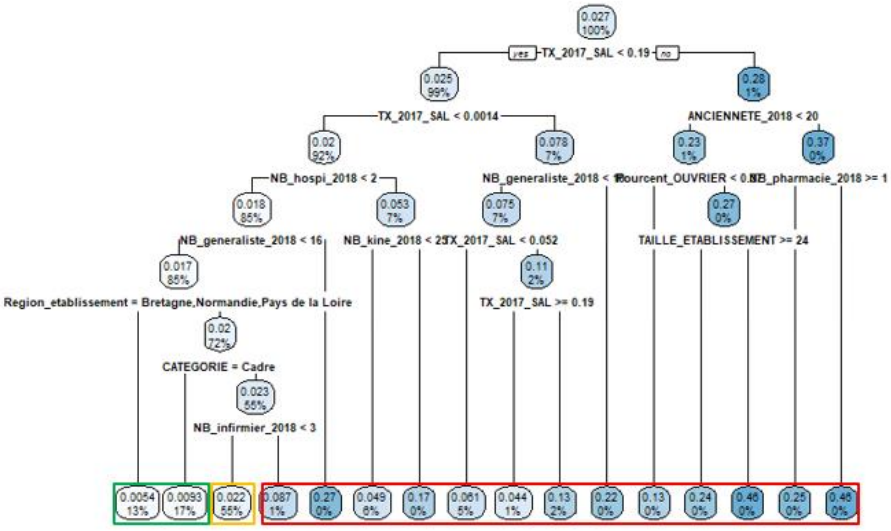
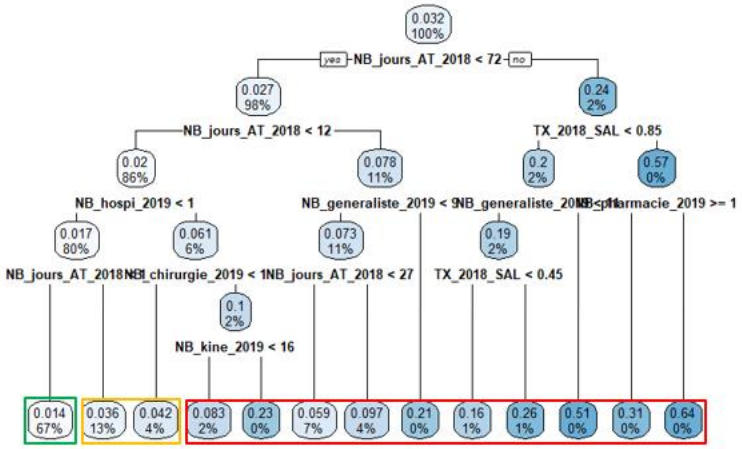


Figure 39 : Segmentation sur 2019 en fonction du taux d'absentéisme et de l'ensemble des variables



Synthèse segmentation des profils d'absentéisme

À la suite de ces diverses segmentations, quelques points sont remarquables. Certaines variables semblent avoir une influence plus importante que d'autres et ce au sein du même groupe de variable ou non. Au niveau des caractéristiques des salariés, **la catégorie socio-professionnelle, la localisation ainsi que l'ancienneté** jouent un rôle important dans la segmentation. Au niveau des variables relatives à la consommation santé, **trois grands postes de santé se dégagent**, des postes qui génèrent plus d'absences comme **l'hospitalisation, les consultations généralistes ou encore le poste kiné** synonyme de blessures physiques. Enfin, l'ajout de données d'absentéisme passées semble être un moyen fiable pour segmenter les individus. Les données passées de l'absentéisme semblent donc avoir une composante temporelle qui influe sur le niveau d'absentéisme présent.

Pour ces différentes segmentations, le niveau de risque est alors divisé en 3 sous-ensembles en fonction du niveau du taux d'absentéisme. Le niveau est défini comme étant « faible », « modéré » ou « fort » avec un seuil entre plus ou moins 20% par rapport au taux d'absentéisme moyen calculé sur l'ensemble de l'effectif. Au niveau de chaque groupe de risque ainsi construit, l'individu retenu possède les caractéristiques suivantes :

Tableau 13 : Caractéristiques des individus et de leur consommation santé en fonction de leur groupe de risque

	Sexe	Région établissement	Catégorie	Ancienneté	Age
Faible	Homme	Ile-de-France/Provence-Alpes Côte d'Azur	Ouvrier	14	40
Modéré	Homme	Ile-de-France/Auvergne-Rhône-Alpes	Ouvrier	15	41
Fort	Homme	Auvergne-Rhône-Alpes	Ouvrier	18	44

	NB_analyse	NB_généraliste	NB_hospi	NB_kiné	NB_pharmacie	Taux d'absentéisme N-1	NB jours d'absence N-1
Faible	0,45	0,93	0	0,56	3,08	0%	0
Modéré	0,97	1,7	1,1	1,08	5,6	1,70%	5
Fort	1,21	2,1	1,02	2,1	6,9	12%	42

A partir de ces résultats, il est possible de remarquer que les groupes de risque construits sont fonction principalement de l'âge, de l'ancienneté, du lieu géographique ainsi que des consommation santé et des données passées de l'absentéisme. Ces caractéristiques permettent ainsi de segmenter le portefeuille afin de donner la possibilité d'étudier plus en détail ces individus dans le temps.

2.2.3 Modélisation par forêts aléatoires

La partie précédente a permis de réaliser rapidement des segmentations du risque absentéisme en fonction de diverses variables. Cette segmentation permet de comprendre les différents niveaux d'absentéisme et d'étudier les salariés qui y sont affectés. Cependant, si les modèles par arbre CART sont des modèles simples, ceux-ci présentent des inconvénients qui

peuvent aboutir à de mauvais résultats. Tout d'abord, les résultats du modèle dépendent fortement de l'ordre des variables segmentant les données. Bien que ces variables arrivent dans un ordre permettant de diminuer l'hétérogénéité au sein des classes créées, certaines variables ayant un fort pouvoir prédictif peuvent ne pas apparaître au profit d'autres variables. Ces variables cachées n'interviennent donc pas dans la segmentation alors qu'elles auraient pu contribuer à la création d'un meilleur modèle. De plus, les arbres CART sont connus pour avoir une forte variabilité dans les résultats. Le modèle étant construit seulement sur une segmentation unique ne prenant en compte que certaines variables et sur un jeu de données précis, la robustesse du modèle est donc remise en question. Enfin, ce modèle ne permet pas de gérer le problème de forte corrélation des facteurs de risque.

Afin de répondre à ces limites, différentes techniques existent. Une de ces méthodes permettant d'améliorer à la fois les prédictions et la robustesse du modèle est **la méthode par bagging**. Comme expliqué précédemment, la méthode de bagging et en particulier les modèles de forêts aléatoires, par l'agrégation de différents modèles d'arbres de décision et par les principes de randomisation des variables ainsi que des échantillons bootstrappés des données vont ainsi contribuer à **la construction d'un modèle plus sophistiqué et robuste, diminuant la variance des résultats. Le problème des corrélations des facteurs de risque est quant à lui gérer à l'aide de la randomisation** des variables dans les multiples modèles créés.

Dans cette partie, les modèles de forêts aléatoires vont donc permettre de **modéliser le risque absentéisme d'une manière plus robuste.** Les deux types de construction de modèle présentés auparavant vont donc être utilisés afin d'étudier la possibilité de modéliser directement le risque absentéisme à partir des facteurs de risque observés la même année, ou de pouvoir faire une prédiction future du risque absentéisme à partir de données du passé. Pour rappel, l'objectif de cette modélisation consiste à donner **des résultats collectifs et non individualisés, et d'étudier les variables ayant un fort pouvoir prédictif sur le modèle**.

Modélisation du risque absentéisme

Pour ce premier type de modélisation par forêts aléatoires, l'objectif est de pouvoir **modéliser le risque absentéisme à partir de variables observées la même année** de calcul du taux d'absentéisme. De ce fait, au même titre que la modélisation par arbre CART, la finalité est de pouvoir **comprendre le pouvoir prédictif des variables** et de trouver celles qui jouent un rôle majeur dans l'identification des niveaux d'absentéisme.

Si la variable cible étudiée est calculée sur l'année N, cela implique que les données utilisées dans cette première modélisation sont également observées à l'année N. Les données à disposition sont donc les caractéristiques des individus, des entreprises, du nombre d'actes santé consommés ainsi que des motifs d'arrêts. De plus, il est également possible d'ajouter quelques variables provenant de l'année N-1 puisque dans la partie précédente, l'arbre CART a montré que le taux d'absentéisme passé ainsi que le nombre de jours en arrêt en N-1 sont des variables ayant une grande influence dans la segmentation.

Afin de pouvoir construire un modèle optimal, une recherche des hyperparamètres ainsi que des variables à ajouter dans le modèle est alors nécessaire. La base de données va être séparée en deux bases, une base d'apprentissage correspondant à 70% de la base initiale et une base de validation correspondant à 30%. L'optimisation du modèle de *Random Forest* est réalisée en trouvant les hyperparamètres *mtry* et *ntree* permettant de trouver le meilleur modèle selon une métrique, le RMSE dans cette étude.

Pour trouver les hyperparamètres optimaux utilisés dans le package R *randomForest*, la première étape consiste à trouver le paramètre *ntree* correspondant au nombre d'arbres construits dans la forêt aléatoire. Le modèle est alors entraîné sur la base d'apprentissage avec un *ntree* assez grand (ici 1 000) afin de représenter la courbe des erreurs *RMSE* en fonction du nombre d'arbres construits dans la forêt aléatoire. Le choix du *ntree* doit alors être réalisée afin de construire un modèle qui ne soit pas trop complexe en prenant un nombre d'arbres trop élevé, et tout en réduisant au maximum l'erreur RMSE. Ensuite, la recherche du mtry est réalisée à l'aide d'une recherche par quadrillage (ou *grid search*) avec les valeurs comprises entre le minimum de $\frac{n}{3}$ et $\sqrt{n}$ et le maximum entre ces deux valeurs, où n correspond au nombre de variables. Un dernier paramètre est alors optimisé, le nodesize correspondant au nombre minimum d'individus devant être présents dans chaque feuille composant l'arbre. A noter que ce paramètre peut être lié au surapprentissage. Ce paramètre est également déterminé à l'aide d'une recherche par quadrillage.

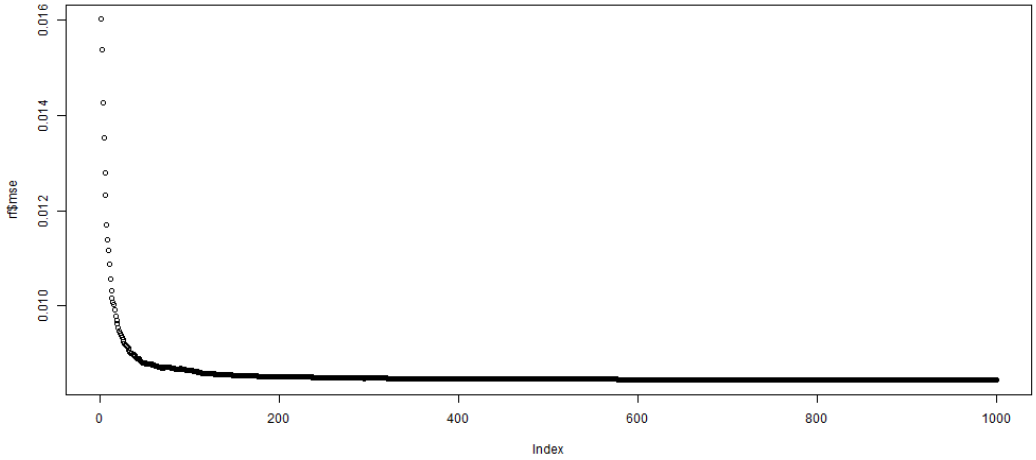
Ces trois paramètres sont alors optimisés pour chaque nouveau modèle, en particulier lorsque les variables retenues diffèrent entre ces derniers. Comme dit précédemment, les variables sont réparties en différents groupes apportant chacun une information supplémentaire sur les conditions du salarié. Plusieurs modèles sont alors construits pour identifier les groupes de variables devant être pris en compte pour obtenir le meilleur modèle à partir des critères RMSE et Q^2. Sur 2018 et 2019, les différents modèles construits font donc intervenir différents groupes de variables. Pour simplifier la notation, la variable à gauche du symbole « ~ » correspond à la variable cible, et les variables du signe correspond aux variables explicatives :

- **Modèle 1 : Taux d'absentéisme en N ~ caractéristiques des salariés et entreprises en N**
- **Modèle 2 : Taux d'absentéisme en N ~ caractéristiques des salariés et entreprises + données de santé en N**
- **Modèle 3 : Taux d'absentéisme en N ~ caractéristiques des salariés et entreprises + données de santé en N + motifs d'arrêt en N**
- **Modèle 4 : Taux d'absentéisme en N ~ caractéristiques des salariés et entreprises + données de santé en N + motifs d'arrêt en N + taux d'absentéisme et nombre de jours d'absence en N-1**

Les RMSE et Q^2 sont ensuite calculés après optimisation de chaque modèle au niveau des hyperparamètres afin de pouvoir les comparer. La démarche d'optimisation des hyperparamètres est présentée pour le modèle 4 pour l'année de 2019.

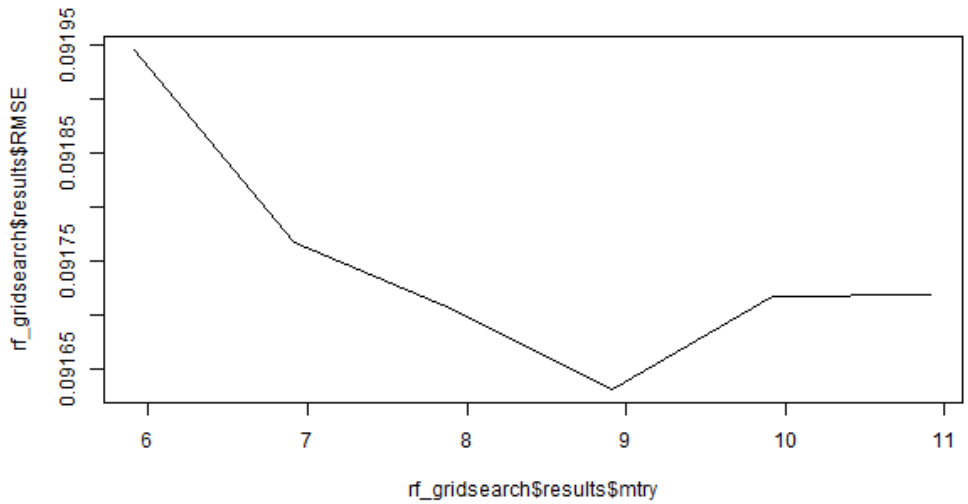
Après avoir ajouté les variables choisies au modèle 4 et avoir filtré les données sur les individus présents sur cette période, le paramètre *ntree* est déterminé en premier à l'aide de l'erreur RMSE tracée en fonction du nombre d'arbres constituant la forêt sur la base d'apprentissage :

Figure 40 : Erreur OOB en fonction du nombre d'arbres



Le paramètre *ntree* retenu est alors égal à 300 permettant ainsi de prendre un nombre d'arbres composant la forêt qui soit raisonnable et en abaissant également l'erreur RMSE. Ce paramètre étant fixé, le *mtry* est alors déterminé en cherchant à minimiser le RMSE du modèle comme présenté sur le graphique suivant :

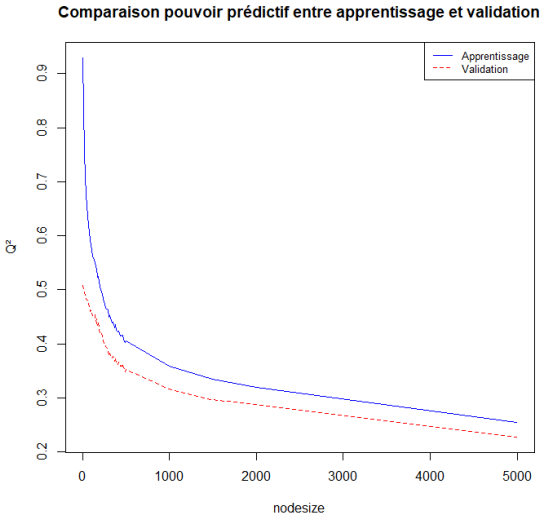
Figure 41 : Erreur du modèle en fonction du nombre de variables tirées aléatoirement

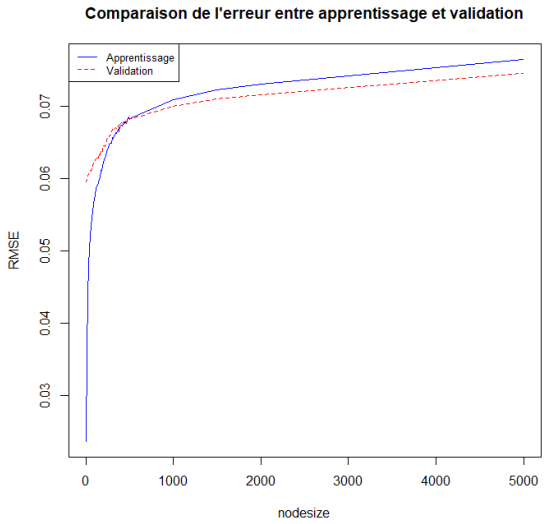


Le *mtry* retenu est égal à 9, correspondant au minimum local entre le minimum et le maximum du nombre de variables divisé par trois et de la racine de ce nombre. Au vu du nombre de variables encore présentes dans le modèle (à savoir 37), le fait de tirer aléatoirement 9 variables sur celles disponibles à chaque création d'arbre permettra ainsi de garantir la diminution des corrélations entre les arbres, donnant de meilleurs résultats.

Le *nodesize* correspond au nombre minimum d'individus qui doit être présent dans chaque feuille de l'arbre. Ainsi, plus le *nodesize* est petit, plus l'arbre sera grand puisque la segmentation doit continuer tant que le nombre minimum d'individus dans chaque feuille n'est pas atteint ou si l'ensemble des individus ont les mêmes caractéristiques et donc la segmentation est terminée. En se rappelant que l'estimation pour un individu revient à calculer la moyenne des valeurs observées pour l'ensemble des individus appartenant à cette même feuille, abaisser le *nodesize* revient alors à donner plus de crédibilité aux données grâce à une segmentation plus profonde. Cependant, le modèle peut vite tomber dans une configuration de surapprentissage.

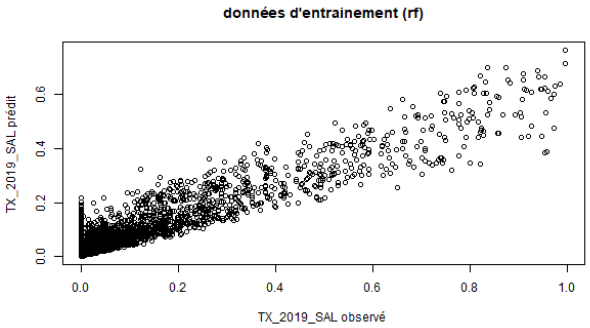
Afin de réduire cet effet, une étude du pouvoir prédictif et de l'erreur moyenne est nécessaire entre l'apprentissage et la validation afin de choisir un paramètre optimal. Les représentations suivantes permettent de comparer les métriques sur les bases d'apprentissage et de validation en fonction du *nodesize* :

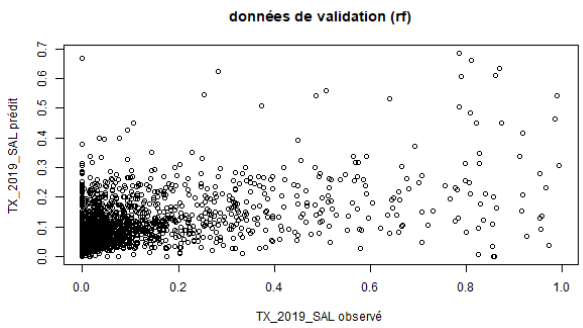




Que ce soit en fonction de l'erreur ou du pouvoir prédictif du modèle, il semble que les résultats de la base de validation se comportent de la même manière que ceux de la base d'apprentissage en fonction du *nodesize*. De ce fait, il semble que le modèle construit ne présente pas un phénomène de surapprentissage malgré le fait que le paramètre induise un apprentissage plus marqué sur la base d'apprentissage des données. Une faible valeur de *nodesize* est donc possible afin de garantir un pouvoir prédictif assez important et une erreur faible.

Après avoir calculé les hyperparamètres optimaux, il est ensuite possible de construire les modèles et de comparer les résultats de prédiction avec les valeurs observées sur les deux jeux de données, apprentissage et validation.





Ces différentes étapes ont été effectuées pour l'ensemble des modèles sur 2018 et 2019. Une comparaison est ensuite réalisée à l'aide des métriques présentées afin de voir l'apport des différents groupes de variables dans la modélisation. Pour rappel, les différentes modélisations sont les suivantes :

	Variable cible	Données individus et entreprises	Santé N	Motifs d'arrêt en N	Tx d'abs. et NB d'AT en N-1
Modèle 1	Tx d'abs. en N	x			
Modèle 2	Tx d'abs. en N	x	x		
Modèle 3	Tx d'abs. en N	x	x	x	
Modèle 4	Tx d'abs. en N	x	x	x	x

Tableau 14 : Comparatif des différents modèles de 2018 en fonction des groupes de variables insérés dans le modèle

	RMSE apprentissage	Q^2 apprentissage	RMSE validation	Q^2 validation
Modèle 1	3,6%	85,8%	8,8%	9,8%
Modèle 2	3,1%	85,4%	7,5%	13,7%
Modèle 3	2,4%	91,5%	6,4%	36,9%
Modèle 4	2,4%	92,9%	5,9%	51,0%

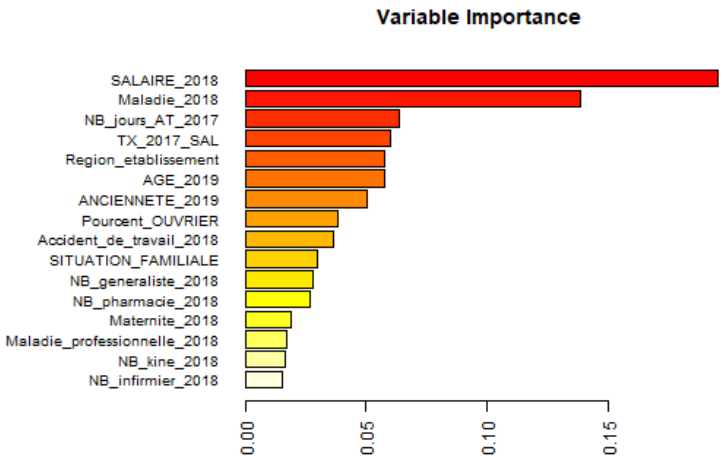
Tableau 15 : Comparatif des différents modèles de 2019 en fonction des groupes de variables insérés dans le modèle

	RMSE apprentissage	Q^2 apprentissage	RMSE validation	Q^2 validation
Modèle 1	4,5%	81,8%	9,7%	0,8%
Modèle 2	4,7%	80,3%	9,4%	8,0%
Modèle 3	3,7%	88,2%	8,5%	23,6%
Modèle 4	3,3%	90,9%	8,7%	38,3%

A partir des précédents tableaux, **l'apport de l'ensemble des groupes de variables semble justifié, abaissant les erreurs moyennes quadratiques et donnant un meilleur pouvoir prédictif** que ce soit sur les bases d'apprentissage ou de validation. Le meilleur modèle ayant été trouvé, il est temps d'analyser les variables d'importance du modèle 4 (c'est-à-dire avec l'ensemble des variables) de 2018 et 2019.

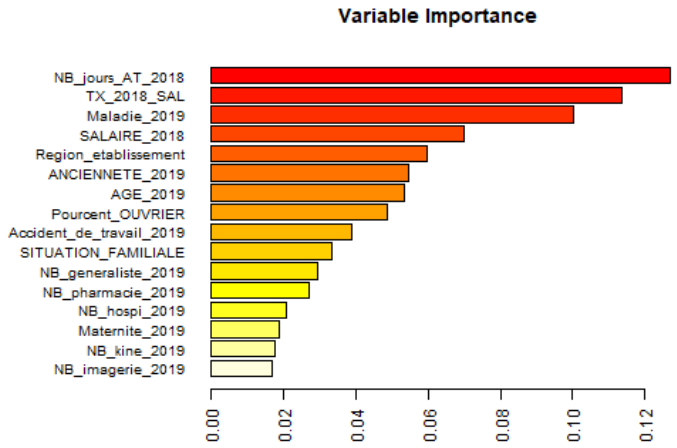
Les variables d'importance renseignent sur la contribution des variables sur l'ensemble du modèle. Plus leur contribution est élevée, plus ces variables sont influentes dans la prédiction du taux d'absentéisme. Les 10 premières variables d'importance pour le modèle de 2018 sont présentées ci-dessous.

Figure 42 : Variables d'importance pour le modèle de 2018



Pour le modèle de 2019, les 10 premières variables d'importance sont également présentées ci-dessous :

Figure 43 : Variables d'importance pour le modèle de 2019



A partir de ces résultats, il est à noter qu'entre 2018 et 2019 les variables possédant un fort pouvoir prédictif restent quasiment inchangées. Les premières variables correspondent à des **données du passé comme le taux d'absentéisme ou le nombre de jours en arrêt indiquant qu'un individu absent auparavant a plus de chance d'être absent dans le futur.** Ensuite, des informations sur le salarié tel que **le salaire, son âge, son ancienneté ainsi que sur la localisation de son travail** viennent complétées ces données. Enfin, **la fréquence de certains actes santé semblent jouer un rôle** également dans la prédiction du taux d'absentéisme, comme les consultations de généraliste, les hospitalisations, la kiné, ou encore la pharmacie, des variables déjà présentées lors de la segmentation par arbre CART.

Prédiction future du risque absentéisme

Les travaux précédents ont permis de calculer les variables d'importance à partir d'une modélisation du risque, c'est-à-dire en cherchant à identifier les variables ayant un fort pouvoir prédictif. Néanmoins, la modélisation du risque doit pouvoir servir à faire des prédictions futures afin de se prémunir d'un potentiel absentéisme futur.

Comme expliqué précédemment dans la section sur la construction des modèles, cette nouvelle modélisation doit prendre en compte uniquement les données du passé déjà observées. Ainsi, une prédiction du taux d'absentéisme annuel en 2019 ne peut pas prendre en compte des données de 2019, (ces données n'étant pas encore connues). Les seules données pouvant être utilisées sont donc les données des années précédentes à savoir en 2017 et 2018.

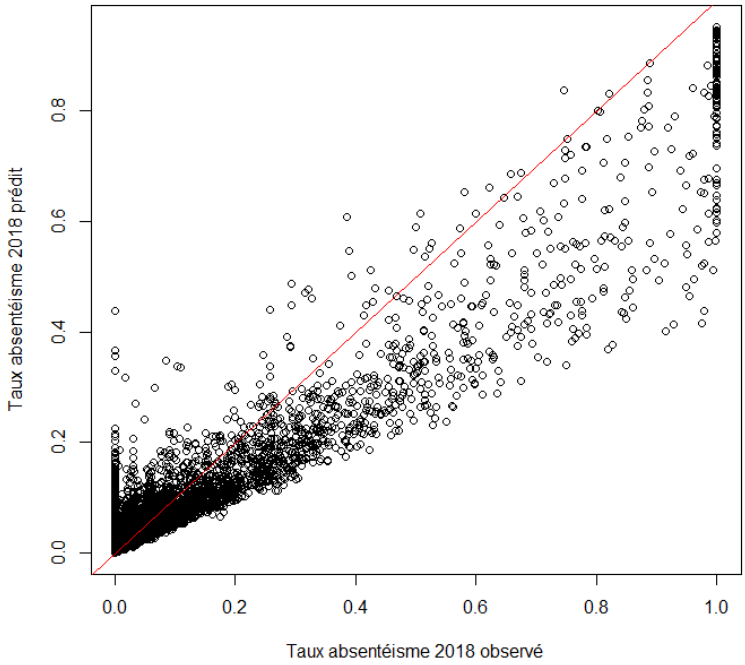
La notion de temporalité étant très importante dans la construction de ce nouveau modèle, celui-ci peut alors être qualifié de modèle dit *glissant*. En effet, l'objectif est de construire un modèle apprenant à prédire le taux d'absentéisme de 2018 à partir des données de 2017, pour ensuite faire une étape de validation sur la période suivante, à savoir une prédiction du taux d'absentéisme de 2019 à partir des données de 2018.

La modélisation peut donc être résumée comme suit :

- **Modèle : Taux d'absentéisme en N ~ caractéristiques des salariés et entreprises + données de santé en N-1 + motifs d'arrêt en N-1 + taux d'absentéisme et nombre de jours d'absence en N-1**

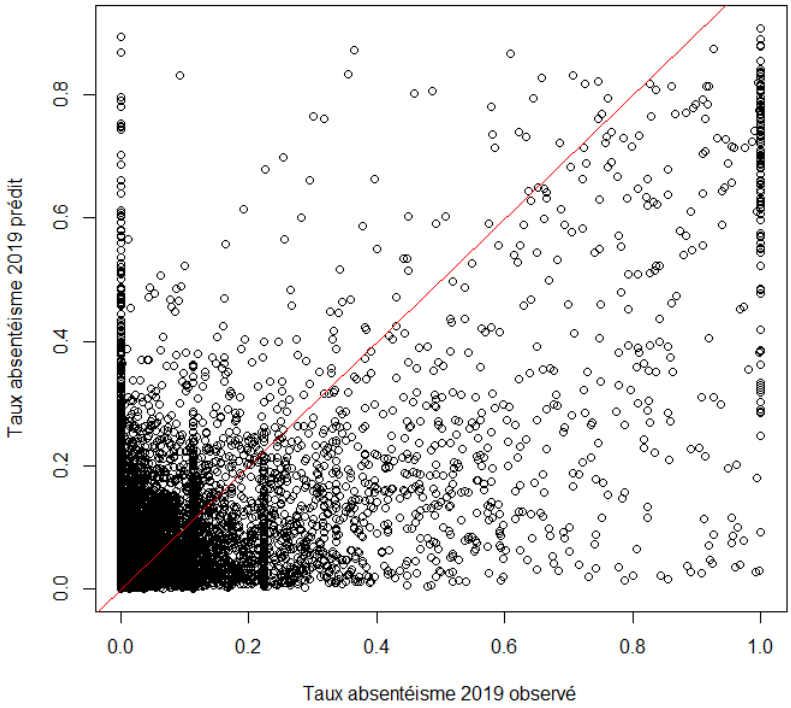
Comme lors de la modélisation du risque, la même démarche est appliquée, les hyperparamètres sont calculés de telles sortes à obtenir un modèle optimal. Le ntree et le mtry ont pour valeurs respectives 300 et 12. Le modèle (modèle A) apprend donc sur les données de 2017 pour prédire le taux de 2018. Sur cette phase d'apprentissage, l'adéquation obtenue entre les données observées et prédites est donnée par le graphique suivant :

Figure 44 : Confrontation taux d'absentéisme observé et prédit sur la base d'apprentissage



Puis la validation est réalisée en prenant comme base de validation les données de 2018 et comme variable cible le taux de 2019. L'adéquation entre les observations et les prédictions est donnée par le graphique suivant :

Figure 45 : Confrontation taux d'absentéisme observé et prédit sur la base de validation



La variabilité du taux d'absentéisme semble donc être difficilement prise en compte par le modèle de forêt aléatoire. Des sous-estimations au niveau des taux d'absentéisme élevés et des surestimations pour les faibles taux d'absentéisme sont constatées à la suite des prédictions. Les différentes métriques calculées sont les suivantes :

	RMSE apprentissage	Q^2 apprentissage	RMSE validation	Q^2 validation
Modèle A	8,4%	42,2%	9,8%	36,8%

Cependant pour rappel, le modèle initial a été entraîné sur les données de 2017 pour prédire le taux d'absentéisme de 2018 lors d'une première étape. Afin de comparer les résultats de modélisation, un découpage des données entre base d'apprentissage et base de validation est effectué lors de la première étape (modèle B). La validation est donc testée sur une partie des données de 2017 et du taux d'absentéisme de 2018. Les résultats sont les suivants en termes de métriques :

	RMSE apprentissage	Q^2 apprentissage	RMSE validation	Q^2 validation
Modèle B	8,4%	41,1%	8,6%	40,8%

Les deux modélisations ont des métriques proches. Cependant, le deuxième modèle lors de la validation sur l'année de 2018 détient un RMSE plus faible et un Q^2 plus élevé. On réitère l'expérience du découpage entre base d'apprentissage et de validation pour les données de 2018 afin de prédire le taux de 2019 (modèle C). Les résultats sont les suivants :

	RMSE apprentissage	Q^2 apprentissage	RMSE validation	Q^2 validation
Modèle C	9,0%	46,3%	9,1%	43,9%

Le modèle est mieux ajusté que celui appris sur 2017 (modèle A). Néanmoins comme présenté au début de cette partie sur la modélisation du risque, l'objectif n'est pas d'avoir un modèle de prédiction individuel mais bien collectif.

Afin de comparer les résultats d'un point de vue collectif, une hypothèse forte est retenue considérant que tous les individus dans la modélisation sont exposés tout au long des périodes d'observation. Ainsi les taux d'absentéisme individuels peuvent être agrégés par moyenne. Les résultats entre observations et prédictions pour ces trois derniers modèles sont regroupés dans le tableau suivant :

	Taux d'absentéisme moyen observé	Taux d'absentéisme moyen prédit
Modèle A	3,9%	5,6%
Modèle B	3,4%	3,6%
Modèle C	3,9%	4,1%

Ce résultat montre les limites du modèle dit glissant. En effet, pour les modèles apprenant d'une année sur l'autre en se calibrant sur certains taux d'absentéisme connus, la moyenne des prédictions sur la base de validation coïncident avec la moyenne des taux observés. Cependant, **dès lors que le modèle s'éloigne dans le temps des données ayant permis l'apprentissage du modèle, celui-ci n'a plus la capacité de donner une prédiction fiable, comme il est possible de voir sur le modèle A**.

2.2.4 Limites

Plusieurs raisons peuvent être à l'origine de cette surestimation du taux moyen d'absentéisme prédit ainsi que du mauvais calibrage du modèle sur les données de validation. Tout d'abord, **une hypothèse forte a été formulée pour calculer le taux d'absentéisme** moyen en affirmant que les individus sont présents sur l'ensemble de la période d'observation. En effet **l'exposition n'est pas prise directement en compte à l'intérieur du modèle.** Pourtant cette variable est un élément important lors de l'étude de l'absentéisme. D'ailleurs, lors du calcul des variables d'importance sur le modèle de 2018 en modélisation pure du risque, le salaire ressort comme étant une variable importante dans la prédiction. Cependant, cette variable prend en compte l'exposition des salariés. Ainsi de très faibles salaires n'expliquent pas la qualité de vie du salarié mais plutôt de son exposition sur l'année. Cette donnée de l'exposition pourrait être prise en compte dans d'autres modèles de *Machine Learning*, comme par exemple les *Distributed Random Forest* permettant d'incorporer cette donnée dans les fonctions de coûts.

Ensuite, les données d'apprentissage correspondent aux données de 2017. Or, la section précédente sur la modélisation pure du risque absentéisme a montré que **le taux d'absentéisme de l'année précédente avait une influence majeure** dans la prédiction du taux d'absentéisme. Cependant, comme expliqué dans la partie descriptive des données, les données de 2017 connaissent **une montée en charge des données DSN**. De ce fait les informations en lien avec **l'absentéisme ne sont pas aussi fiables que sur 2018 ou 2019**, ce qui peut pour avoir comme conséquence de surestimer les prédictions lorsque celles-ci utilisent des données d'absentéisme plus élevées comme en 2018 par exemple.

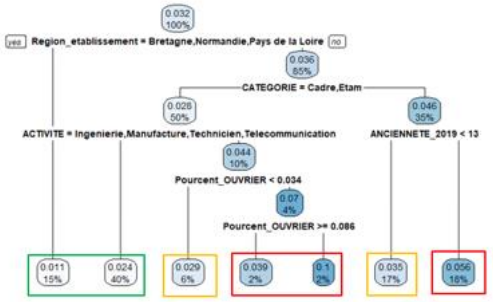
Enfin**, la temporalité est une composante qui n'est pas assez prise en compte** dans ce type de modèle. Lors de la prédiction du taux d'absentéisme à l'année N, aucune hypothèse n'est formulée sur la situation des salariés, que ce soit sur leurs caractéristiques, leur consommation santé ou leurs absences. De ce fait, la situation ayant pu changer entre 2017 et 2019, le modèle n'arrive plus à définir une bonne prédiction du taux d'absentéisme moyen.

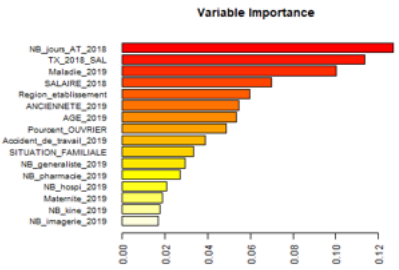
Ainsi, la nécessité d'utiliser de nouveaux modèles prenant en compte la temporalité, la saisonnalité et l'évolution des différents facteurs de risque doivent être construits. **L'utilisation des modèles de série temporelle afin d'explorer la piste d'une nouvelle modélisation étudiant la saisonnalité du risque absentéisme et des différents facteurs de risque est donc nécessaire et sera donc présentée dans la partie suivante**.

Synthèse deuxième partie : Modélisation du risque absentéisme par Machine Learning

L'utilisation des modèles de *Machine Learning* a permis de mettre en place une stratégie utilisant **un maximum de données** à disposition afin d'étudier, **modéliser et prédire le taux d'absentéisme annuel d'un groupe de salariés.**

La première partie consistant à segmenter le portefeuille à l'aide des arbres CART a permis d'identifier les premières variables segmentant les individus pour ensuite obtenir des premiers groupes d'individus ayant des profils d'absentéisme relativement homogènes. **Sur ces groupes qualifiés en profils faible, modéré ou fort d'absentéisme, l'âge, l'ancienneté, la consommation santé ainsi que l'absentéisme observé dans le passé sont des informations importantes**.





Après ce premier diagnostic sur **les profils d'absentéisme** qui pouvaient ressortir du portefeuille, les modèles de forêts aléatoires ont été construits en tentant de prédire **le taux d'absentéisme** de chaque individu sur l'année 2019 afin d'agréger ces résultats sous un taux d'absentéisme collectif. Après une recherche d'optimisation du modèle sur les paramètres mais également en ajoutant des groupes de variables, un premier test de recherche des variables les plus importantes a été effectué en utilisant les données de la même année que le taux d'absentéisme observé. Dans les variables les plus importantes sur la modélisation de l'indicateur, il est possible de retrouver les informations liées à l'absentéisme du passé, des informations sur le salarié ainsi que sur la consommation en santé comme les **actes de généraliste ou de kiné**.

Le réel enjeu par la suite reste **de prédire le taux d'absentéisme d'une année sur l'autre en fonction de données provenant de l'année précédente** puisque les données pour l'année N de prédiction ne sont pas encore observables.

Les modèles de Machine Learning n'ont pas l'avantage d'être assez flexibles avec les données temporelles et n'arrivent pas à donner de bonnes prédictions face aux observations. En effet, le modèle doit faire des prédictions sur des données évoluant au cours du temps. A noter également que la base de données ne dispose pas d'une profondeur d'historique suffisante pour faire des tests sur plusieurs modèles annuels successifs.

Le modèle de Machine Learning doit être entraîné à chaque nouvelle période de prédiction, et **les relations temporelles entre les variables comme l'évolution des frais de santé ou encore les arrêts de travail ne sont pas prises en compte.**

Une méthode permettant de pallier ce problème va donc faire l'objet de travaux complémentaires en troisième partie : l'utilisation des **séries temporelles. Cette nouvelle approche innovante aura pour objectif de modéliser le taux d'absentéisme en prenant en compte l'évolution de l'ensemble des variables au cours du temps.** La prédiction du taux d'absentéisme au cours du temps pourra ainsi permettre **un suivi plus accru du risque en fonction de variables explicatives** provenant des données de l'assureur**, ou encore de données open data ajoutant de l'information sur le contexte actuel.**

3. RISQUE ABSENTEISME ET MODELE TEMPOREL DYNAMIQUE

La partie précédente a permis de montrer **la rigidité des modèles de *Machine Learning* précédents face aux données temporelles**. Ces modèles ne permettent pas de prendre en compte les **interactions temporelles entre les différentes variables ainsi que l'évolution dynamique de celles-ci**. Les données à disposition pour construire le modèle ne tiennent alors compte que d'une image fixe du portefeuille à un instant précis. Les modèles de *Machine Learning* utilisés dans la partie précédente ont montré la difficulté à modéliser le risque. Les modèles de *Machine Learning* permettent de **réaliser des segmentations** sur les salariés en fonction de leurs caractéristiques et de leur lien avec l'absentéisme, et d'étudier le **pouvoir prédictif des variables explicatives**.

L'introduction de variables temporelles permettant d'expliquer **l'évolution du taux d'absentéisme au cours du temps** est donc l'enjeu de cette partie, en proposant **une approche plus dynamique de l'absentéisme**. Elle permettra ainsi de proposer une nouvelle vision et gestion du risque absentéisme à la fois pour l'assureur et l'employeur.

Si les modèles présentés dans la deuxième partie ne permettaient pas d'obtenir de bonnes projections du risque à des dates futures, ce nouveau cadre d'étude tente d'y parvenir en spécifiant certaines hypothèses qui permettront **une mise à jour du modèle plus flexible, tout en proposant une modélisation comportant deux forces : une modélisation de type Machine Learning à l'aide de variables explicatives, ainsi qu'une prise en compte de la temporalité à l'aide des modèles de séries temporelles.**

3.1. Les avantages a priori d'une étude par séries temporelles

3.1.1 Intégration de l'évolution temporelle des données

Les données à disposition pour ces travaux sont des données variant au cours du temps, telles que la consommation santé d'un individu et ses arrêts de travail. Comme expliqué dans les parties précédentes, ces données peuvent être étudiées de deux manières :

- En considérant une période d'étude fixe et en travaillant sur des variables agrégées
- En conservant l'ensemble des données pour chaque date de l'historique

Si le premier choix a été d'utiliser les modèles de Machine Learning afin d'étudier le taux d'absentéisme d'une année sur l'autre, l'étude a montré que **la modélisation ainsi que les prédictions dans le futur manquaient de robustesse**.

La deuxième méthode d'étude des données consistant à conserver l'évolution des facteurs de risque ainsi que celle de la variable cible relative à l'absentéisme tente de pallier ce problème. L'étude ne se concentre donc plus sur une modélisation ou une prédiction du risque pas à pas sur des périodes de temps fixes, mais bien sur **une étude plus dynamique des variables explicatives**

ainsi que du risque absentéisme. L'objectif n'est donc plus de prédire un unique taux d'absentéisme pour une certaine période à l'aide de données arrêtées, mais de chercher à **construire l'évolution future la plus probable de l'absentéisme au cours du temps**. Les différentes variables étant étudiées sur la base du temps, celles-ci peuvent être représentées à l'aide de leur représentation en séries temporelles.

L'introduction des séries temporelles permet entre autres d'ajouter une information évolutive du risque à une variable cible à modéliser : **le taux d'absentéisme**. Vu précédemment, le taux d'absentéisme détient une corrélation positive avec certaines variables variant elles aussi au court du temps comme la consommation santé ou les valeurs passées de ce taux. Etudier cette variable par série temporelle c'est également étudier ses valeurs passées et son évolution. Le taux d'absentéisme étant une variable regroupant plusieurs informations, pouvoir étudier son évolution avec une fréquence plus grande peut amener à comprendre des changements de tendance qui n'auraient pas pu être pris en compte lors d'une étude année après année comme dans la partie précédente.

3.1.2 Une modélisation plus flexible

La partie précédente sur l'utilisation des forêts aléatoires pour prédire le taux d'absentéisme a montré des faiblesses au niveau de la flexibilité du modèle. Bien que les modèles de forêts aléatoires permettent d'obtenir une modélisation du risque à partir d'autres facteurs explicatifs sur la même période d'observation, plus les prédictions sont éloignées des dernières observations moins les prédictions réalisées sont fiables.

Ce biais dans les prédictions futures du risque provient de l'apprentissage séquentiel montré dans la partie précédente. Afin de faire une prédiction pour l'année N sans avoir connaissance des données cette même année, le modèle doit être construit sur les données de l'année N-1 voire N-2. Or d'une année sur l'autre, les variables ayant été agrégées afin d'obtenir une vision annuelle des données explicatives, l'évolution de ces facteurs de risque n'est pas prise en compte. Une possibilité serait alors de pouvoir travailler sur des données agrégées sur une fréquence temporelle plus fine et de prendre en compte plusieurs dates d'observation afin de prédire une valeur future du risque. Cependant, les modèles de *Machine Learning* se basant sur un apprentissage empirique des données, des phases d'apprentissage et de validation sont nécessaires sur chaque nouvelle période de prédiction ce qui peut entraîner des lenteurs de mise en place des modèles et de calculs.

L'avantage des modèles de séries temporelles est que ceux-ci permettent un apprentissage et une prédiction des données en **prenant en compte l'évolution des données passées.** Ainsi, le modèle séquentiel qui avait été introduit pour les modèles de forêts aléatoires n'est pas reproduit ici, puisque l'apprentissage se fait directement sur l'ensemble des données passées. De plus, lors de la prédiction, une fenêtre de prédiction peut être plus ou moins large à l'inverse des autres modèles qui eux ne peuvent finalement donner qu'une seule prédiction à la fois.

La mise à jour de la modélisation est également un avantage dans le cadre des modèles de séries temporelles puisque celle-ci se fait uniquement en rajoutant les nouvelles données d'observation au fur et à mesure. Par opposition, les modèles séquentiels nécessitent un travail supplémentaire en ajoutant les nouvelles données d'observation mais également en déterminant les périodes d'apprentissage qui elles sont fixes.

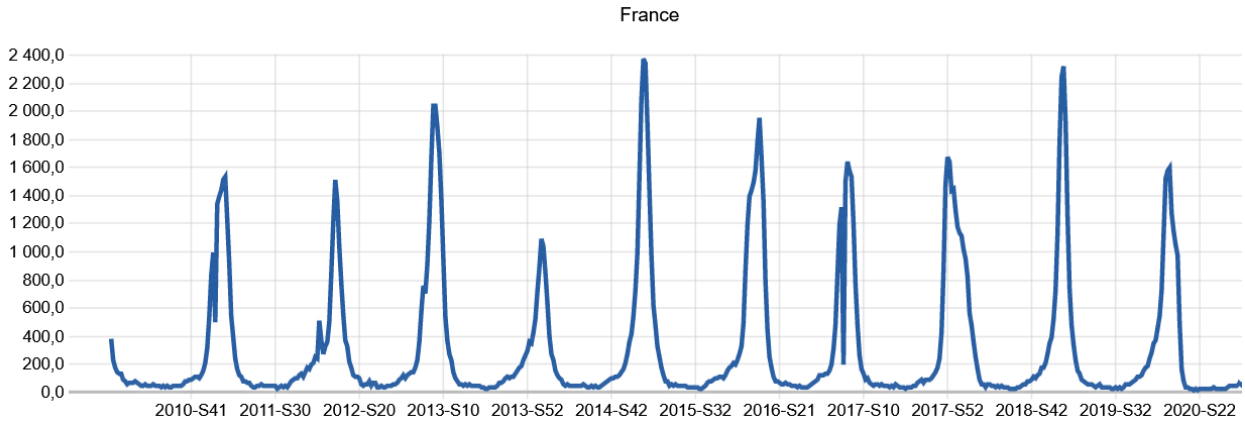
3.1.3 Utilisation de données temporelles exogènes

L'approche par séries temporelles permet d'ajouter certaines informations importantes à la modélisation du risque.

Avec la prise en compte de la dynamique des variables au cours du temps, il est alors possible d'ajouter des données variant au cours du temps. L'information introduite dans le modèle est alors plus complète comparativement à l'ajout de quelques informations décrivant le phénomène dans le cadre d'un modèle de Machine Learning. De nouvelles données exogènes au portefeuille d'étude peuvent être utilisées afin d'améliorer les prédictions du modèle.

Afin d'illustrer cette idée, il serait par exemple opportun de considérer l'évolution des épidémies de grippe sur l'année afin de prédire le taux d'absentéisme dans une entreprise. En effet, les épidémies sont des évènements exogènes à l'activité de l'entreprise. Cependant, ce phénomène peut expliquer une partie de l'absentéisme au sein d'une entreprise si plusieurs salariés sont touchés par la maladie.

Figure 46 : Taux d'actes médicaux SOS médecins pour la grippe en France (données de santé publique France)



En considérant les séries temporelles, l'information sur les diverses épidémies au cours des années précédentes est ainsi complète. La date de début, de fin, l'amplitude ainsi que l'évolution sont ainsi conservées lors de l'ajout de cette donnée dans le modèle. A l'inverse, lors de l'utilisation des modèles de forêts aléatoires, les informations de durée et d'amplitude auraient pu être prises en compte, mais les données relatives à l'évolution ainsi que les dates d'apparition auraient

demandé un travail supplémentaire pour l'intégration de ces données. Ainsi les modèles de séries temporelles permettent d'utiliser des données brutes sans retravailler la variable afin de l'intégrer à la modélisation.

3.2. Théorie des séries temporelles et des modèles

Afin de pouvoir étudier et construire les modèles de séries temporelles appliqués au risque absentéisme du portefeuille étudié, il est nécessaire d'introduire quelques notions théoriques.

3.2.1 Les séries temporelles

Dans un premier temps, d'un point de vue pragmatique, une série temporelle peut être définie par la phrase suivante :

Une série temporelle est une suite formée d'observations d'un phénomène régulièrement espacées au cours du temps

D'un point de vue mathématiques, afin d'introduire un peu de formalisme à l'étude de ces séries temporelles, celles-ci vont être considérées comme étant une réalisation particulière d'une famille de variables aléatoires $X = \{X_t\}_{t \in I}$ définies sur un espace de probabilité $(\Omega, A, \mathbb{P})$ où I est un ensemble de dates discrètes équiréparties. De ce fait, une série temporelle est la fonction telle que :

$$I \times \Omega \to \mathbb{R}$$

$$(t, \omega) \mapsto X_t(\omega) = x_t$$

Les enjeux derrière ces objets mathématiques sont multiples, que ce soit sur la description, la prévision ou encore les détections de rupture au cours du temps permettant de déterminer les instants où les paramètres de la série se modifient.

Au niveau de la décomposition de la série temporelle, il est supposé que la série temporelle puisse être décomposée en trois termes tel que :

$$X_t = m_t + S_t + Y_t$$

où :

- m_t est une tendance déterministe
- S_t est une saisonnalité déterministe
- Y_t est une perturbation aléatoire représentant l'erreur, de moyenne nulle et possédant soit une structure de corrélation non nulle et stable dans le temps soit possédant une structure non-stable mais qui peut tout de même être modélisé par un processus simple

Afin de pouvoir proposer une modélisation de séries temporelles sur les données, il est nécessaire de fixer des hypothèses sur la suite de variables aléatoires (X_t). En effet pour que la modélisation puisse être fiable et robuste, une hypothèse sur la stationnarité de la série doit être vérifiée.

Une série temporelle (X_t) est dite stationnaire strict si pour tous $(h,k)\,\epsilon\, I\times I, h\ et\ k\geq 1$ les vecteurs $(X_1,X_2,\dots,X_k)\ et\ (X_{1+h},X_{2+h},\dots,X_{k+h})$ ont même loi.

Ainsi la stationnarité stricte impose que les lois de tous les X_t soient les mêmes. Cependant, cette contrainte peut être trop contraignante lors de l'estimation des modèles ou dans le cas de prédictions. Une notion moins contraignante de stationnarité (dite du second ordre) est alors utilisée, se basant sur les fonctions moyennes et de covariance de la série temporelle (X_t).

Soit (X_t) telle que $\mathbb{E}[X_t^2]<\infty$ pour tout t. Les fonctions moyenne et covariance de (X_t) sont respectivement définie par :

$$\mu_X(t)=\mathbb{E}[X_t] \quad et \quad \gamma_X(r,s)=Cov(X_r,X_s)$$

(X_t) est dite stationnaire au second ordre si :

- $t \mapsto \mu_X(t)$ est indépendante de t,
- $t \mapsto \gamma_X(t,t+h)$ est indépendante de t, pour tout h

Dans ce cas-là, les fonctions d'autocovariance et d'autocorrélation sont définies pour la série (X_t)

$$\gamma_X(h)=Cov(X_t,X_{t+h}) \quad et \quad \rho_X(h)=\frac{\gamma_X(h)}{\gamma_X(0)}=Cor(X_t,X_{t+h})$$

Afin de pouvoir déterminer si la série est stationnaire différents tests sont possibles. Pour la suite de ce mémoire, le test utilisé pour vérifier la stationnarité des séries sera le test KPSS présenté en annexe.

Enfin la définition de régression linéaire théorique est introduite afin de pouvoir définir la notion d'autocorrélation partielle qui sera utile lors de l'identification des paramètres des modèles de séries temporelles. La régression linéaire théorique de X_t sur les données de $X_{t-1},\dots,X_{t-p}$ est la projection orthogonale dans $\mathcal{L}^2(\Omega,A,\mathbb{P})$ de X_t sur $H=Vect(X_{t-1},\dots,X_{t-p})$. Cette régression linéaire est alors notée : $EL(X_t|X_{t-1},\dots,X_{t-p})$

A partir de cette nouvelle définition, le coefficient d'autocorrélation partielle d'ordre k est défini par :

$$r(k)=Cor(\tilde{X}_t,\tilde{X}_{t-k})$$

Où

$$\tilde{X}_t=X_t-EL(X_t|X_{t-1},\dots,X_{t-k+1}) \quad et \quad \tilde{X}_{t-k}=X_{t-k}-EL(X_{t-k}|X_{t-1},\dots,X_{t-k+1})$$

En outre, l'autocorrélation partielle d'ordre k désigne la corrélation entre X_t et X_{t-k} lorsque l'influence des variables X_{t-k+i} avec $i < k$ a été prise en compte.

A partir de l'ensemble de ces définitions et notions, il est possible par la suite de présenter les différents processus permettant de modéliser une série temporelle.

3.2.2 Les modèles de séries temporelles

Les processus de séries temporelles les plus connus permettant de modéliser ces séries sont les processus AR, MA, ARMA, ARIMA et SARIMA. Ces différents modèles vont être brièvement présentés dans la suite de cette section, ainsi que le modèle utilisé pour la modélisation du risque absentéisme, le modèle ARIMAX qui est différent des derniers mentionnés dans sa construction.

Au préalable, l'opérateur retard L est présenté permettant de simplifier les notations lors de l'utilisation d'observations antérieures à l'instant t de la série :

$$L : (X_t)_{t\in\mathbb{Z}} \mapsto (Y_t)_{t\in\mathbb{Z}} \; tel \; que \; Y_t = LX_t = X_{t-1}$$

Cet opérateur peut être facilement appliqué aux polynômes qui seront utilisés par la suite lors de la présentation des différents processus :

Soit P un polynôme de la forme $P(z) = \sum_{k=0}^{p} a_k z^k$ avec $a_k \in \mathbb{R}\; et\; p \in \{\mathbb{N}; \infty\}$, le polynôme retard associé s'écrit donc :

$$P(L) = \sum_{k=0}^{p} a_k L^k$$

Enfin, la notion de bruit blanc est introduite :

- Un bruit blanc fort est une suite (ε_t) de variables indépendantes et identiquement distribuées (iid), centrées et de variance σ^2 finie
- Un bruit blanc faible est une suite (ε_t) de variables centrées, de variances constantes et non corrélées. Les caractéristiques d'un bruit blanc faible sont donc les suivantes :

$$\mathbb{E}[\varepsilon_t] = 0, \quad Var(\varepsilon_t) = \sigma^2, \quad Cov(\varepsilon_t, \varepsilon_{t'}) = 0 \; pour \; t \neq t'.$$

A partir des différentes notions introduites, il est possible de formaliser les différents processus pouvant modéliser une série temporelle.

Le processus MA(q)

Un processus stochastique stationnaire $(X_t)_{t\in\mathbb{Z}}$ est un MA(q) ou une moyenne mobile d'ordre q s'il satisfait l'équation suivante :

$$X_t = \mu + \Theta(L)\varepsilon_t = \mu + \sum_{k=0}^{q} \varepsilon_{t-k}\theta_k$$

Où Θ est un polynôme de degré q, les θ_k sont des nombres réels (avec $\theta_q \neq 0$ et $\theta_0 = 1$) et $(\varepsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc faible de variance σ_ε^2.

Le processus AR(p)

Un processus stochastique $(X_t)_{t\in\mathbb{Z}}$ est un AR(p) ou un processus autorégressif d'ordre p s'il est stationnaire et s'il satisfait l'équation suivante :

$$\Phi(L)X_t = \mu + \varepsilon_t \text{ ou encore } X_t = \mu + \sum_{k=1}^{p} X_{t-k}\varphi_k + \varepsilon_t$$

Où Φ est un polynôme de degré p, les φ_k sont des nombres réels (avec $\varphi_p \neq 0$) et $(\varepsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc faible de variance σ_ε^2.

Ces deux premiers processus possèdent certaines propriétés permettant de les calibrer. Les deux propriétés principales permettant d'identifier les ordres q et p respectifs des processus MA et AR sont les suivantes :

- Pour un processus MA(q), les autocorrélations $\big(\rho_X(h)\big)$ sont nulles pour $|h| > q$
- Pour un processus AR(p), les autocorrélations partielles $\big(r(h)\big)$ sont nulles pour $|h| > p$

Certaines séries temporelles peuvent également être représentées à l'aide d'un processus possédant une composante autorégressive et d'une moyenne mobile. Ces nouveaux processus sont présentés par la suite.

Le processus ARMA(p,q)

Un processus stochastique stationnaire $(X_t)_{t\in\mathbb{Z}}$ est un ARMA(p,q) s'il satisfait l'équation suivante :

$$\Phi(L)X_t = \mu + \Theta(L)\varepsilon_t$$

Ou encore

$$X_t = \mu + \varepsilon_t + \sum_{k=1}^{p} X_{t-k}\varphi_k + \sum_{j=1}^{q} \varepsilon_{t-j}\theta_j$$

Où Φ et Θ sont des polynômes de degrés respectifs p et q, les φ_k et les θ_j sont des nombres réels (avec $\varphi_p \neq 0 \; et \; \theta_q \neq 0$) et $(\varepsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc faible de variance σ_ε^2.

Cependant, certaines séries temporelles peuvent ne pas être stationnaire (vérifié par un test KPSS par exemple), et dans ce cas-là un autre processus doit être utilisé afin de différencier la série jusqu'à obtenir une série répondant au critère de stationnarité. Ce nouveau processus est le processus ARIMA(p,d,q).

Le processus ARIMA(p,d,q)

Un processus stochastique $(X_t)_{t\geq -p-d}$ est un processus ARIMA(p,d,q) s'il satisfait l'équation suivante :

$$\Phi(L)(1-L)^d X_t = \mu + \Theta(L)\varepsilon_t$$

De ce fait, une différenciation d'ordre d est effectuée sur l'ensemble des termes autorégressifs de la série, permettant ainsi de modéliser cette série dans un cadre vérifiant la stationnarité. Ainsi, si $(1-L)^d X_t$ est asymptotiquement équivalent à un processus stochastique stationnaire et $(1-L)^{d-1} X_t$ ne l'est pas alors le processus stochastique discret (X_t) est dit intégré d'ordre d.

D'autres processus appelés SARIMA permettent de prendre en compte une certaine saisonnalité des données lorsque celle-ci est pressentie à l'aide d'une représentation graphique des données ou encore par étude des autocorrélations. **Cependant l'usage de ces processus nécessitent un historique de données suffisamment conséquent pour pouvoir les appliquer.** Comme montré dans la prochaine partie, l'absentéisme détient une saisonnalité qui est plus ou moins réplicable chaque année. Néanmoins dans le cadre de ces travaux et n'ayant un historique que de 3 ans, l'utilisation de ces processus n'est pas recommandée dans ce contexte.

Ces différents processus montrés ci-dessus présentent des modélisations basées uniquement sur les données passées de la série temporelle. Or, les travaux sur les modèles de *Machine Learning* ont permis de montrer l'importance de certaines variables dans la modélisation du risque absentéisme. L'objectif serait donc de pouvoir ajouter dans ces modèles de séries temporelles des données permettant à la fois de prendre en compte l'évolution du risque par le modèle de série temporelle mais également d'ajuster ces résultats à l'aide de variables explicatives. Cet objectif peut être atteint à l'aide d'un nouveau processus nommé ARIMAX. [19,20]

Le modèle ARIMAX(p,d,q)

A la différence des autres processus présentés précédemment, le modèle ARIMAX est avant tout un modèle de régression linéaire, dont l'erreur va être modélisé à partir d'un processus ARIMA présenté ci-dessus. La régression va donc faire intervenir les différentes variables exogènes introduites dans le modèle qui seront appelées à présent régresseurs. Ces régresseurs sont donc eux-mêmes des variables qui varient au cours du temps et qui peuvent donc également être représentées en tant que série temporelle.

Soient $(Y_t)_{t\in\mathbb{Z}}$ la série temporelle cible à modéliser, et $(X_{k,t})_{k\in[\![1;n]\!];t\in\mathbb{Z}}$ les régresseurs utilisés lors de la régression linéaire. Alors pour tout t, chaque observation de la variable cible va être modélisée par :

$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_n x_{n,t} + u_t$$

Avec u_t suivant un processus ARIMA(p,d,q). Le problème revient donc à un problème de moindres carrés généralisés puisque les erreurs de la régression linéaire sont de ce fait corrélées. **Cette nouvelle approche va donc permettre de prendre en compte des variables explicatives tout en conservant le caractère temporel des variables à l'aide de l'étude des corrélations des résidus**.

3.2.3 Forecasting, méthode de prédiction

Après avoir identifié le processus permettant de modéliser une série temporelle l'objectif est de pouvoir réaliser des prédictions des futures valeurs de la série et ainsi étudier l'évolution future de la variable. Cette approche appelée *Forecasting* se base alors sur les différents paramètres trouvés lors de la calibration du modèle et des valeurs prises par les variables au cours du temps.

Le modèle utilisé dans la suite étant le modèle ARIMAX, il est nécessaire d'expliquer la manière dont les valeurs sont prédites. Comme vu précédemment, le modèle ARIMAX possède deux composantes, **une composante comportant une régression linéaire, et une partie comportant les erreurs du premier modèle qui suivent un processus ARIMA**. De ce fait, lors de la projection des prédictions dans le futur, celles-ci résultent donc de la somme de la prédiction de la moyenne et de la prédiction de l'erreur.

La prévision de Y_{t+1} notée $\hat{y}_{t+1}$est donc fonction des valeurs passées de $(Y_i)_{i\leq t}$ des observations des variables explicatives $(x_{1,1}, \dots, x_{n,1}), \dots, (x_{1,t}, \dots, x_{n,t})$ mais également des hypothèses réalisées en t+1 pour les régresseurs. De ce fait la prévision est donnée par :

$$\hat{y}_{t+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{1,t+1} + \cdots + \hat{\beta}_n x_{n,t+1} + \hat{u}_{t+1}$$

La prédiction de la moyenne étant déterminée par les régresseurs au temps t+1, la prédiction du processus est quant à elle basée sur la régression linéaire théorique des données passées de la série, prenant en compte les paramètres du processus, à savoir p, d et q. Cette étude du *Forecasting* sur le risque absentéisme sera alors étudiée lors de l'application de ces méthodes sur les données du portefeuille étudié.

3.2.4 Métrique et validation

Afin de pouvoir valider de la justesse des modèles de séries temporelles construits, différents tests existent. En outre, les tests de Ljung-Box ou encore les tests de normalité de Jarque-Bera permettent de valider l'absence d'autocorrélation des résidus ainsi que la blancheur des résidus.

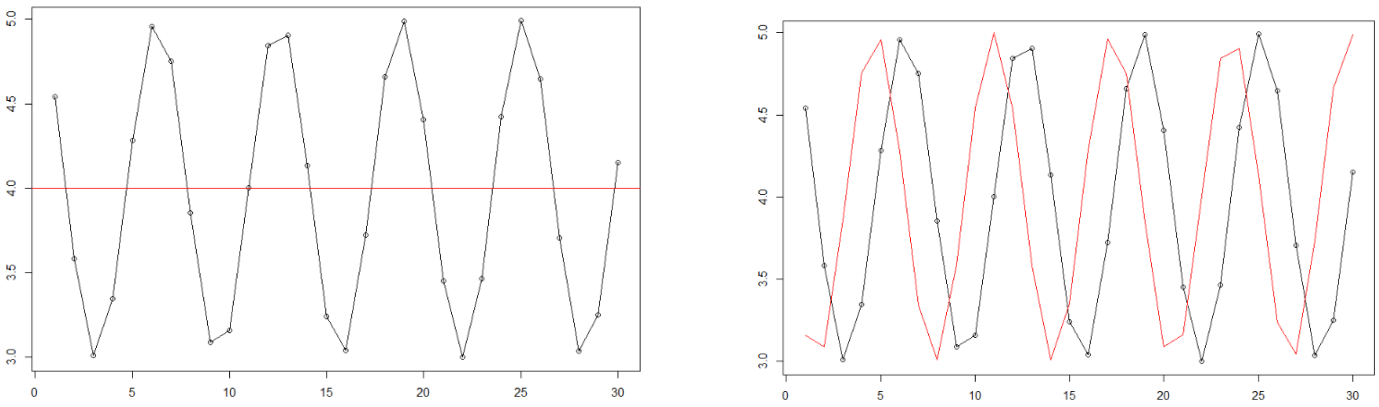
Cependant, ces tests sont utilisés lors de la modélisation de la série temporelle par un certain processus. La suite des travaux en revanche se penche sur la validation des prédictions obtenues par *Forecasting* de la série. A l'instar des métriques définies lors de la partie sur les

modèles de *Machine Learning*, des métriques doivent être choisies dans le cadre des modèles de séries temporelles afin de comparer les différents résultats de prédictions.

A la différence de l'étude de la justesse des modèles en *Machine Learning* où le RMSE et le Q^2 avaient été présentés, l'étude de la justesse des prévisions des modèles de séries temporelles doit prendre en compte l'évolution de la série. En particulier, lors de l'étude du taux d'absentéisme, l'objectif est de pouvoir prédire des augmentations ou des diminutions de l'indicateur du risque. De ce fait, si le RMSE permet encore de vérifier l'ajustement global des prédictions sur les valeurs observées, cette métrique ne permet pas de capter suffisamment la justesse de la variabilité des prédictions vis-à-vis des observations.

Par exemple, sur les deux graphiques suivants, une même série est comparée à deux propositions de prédictions. Dans le premier cas, la prédiction de la série en rouge revient à prendre la moyenne de la série tandis que la deuxième proposition consiste à proposer une série décalée dans le temps par rapport aux valeurs observées réelles. Au niveau de la justesse de prédiction, il serait préférable de considérer la deuxième prédiction puisque celle-ci capte assez fidèlement les périodes de croissance et de décroissance de la série malgré un léger décalage. Pourtant, ce léger décalage peut causer une valeur de RMSE plus grande que lors du calcul de première prédiction puisque celle-ci correspond à la valeur moyenne de la série. Ainsi une nouvelle mesure permettant de mesurer la similarité des valeurs observées et prédites doit être introduite.

Figure 47 : Exemple de prédictions menant à un mauvais choix de prévision en fonction du calcul du RMSE



Cette métrique est alors le « Dynamic Time Warping » (DTW) qui correspond à la distance de déformation temporelle. Cette métrique se base sur la notion que le temps est élastique et non pas linéaire. Elle permet ainsi de contrer le problème montré précédemment dans la comparaison de séries temporelles rencontrant un déphasage. Le principe réside donc dans l'étude des correspondances de sous-séquences ayant des caractéristiques similaires entre deux séries, même si ces sous-séquences ne coïncident pas sur le même intervalle de temps.

Ainsi, pour deux séries temporelles U et V de dimensions respectives m et n, l'objectif est de passer en revue l'ensemble des valeurs des séries jusqu'à obtenir une correspondance

maximale entre ces deux séries. Une matrice de distance de dimension $(n \times m)$. Une première distance est initialisée en comparant les valeurs des premières observations des deux séries :

$$d_{cum}(1,1) = |u_1 - v_1|$$

Les valeurs de la première ligne et de la première colonne de la matrice de distance sont également initialisées par les formules suivantes :

$$d_{cum}(1,j) = |u_j - v_1| + d_{cum}(1,j-1)\ pour\ j > 1$$

$$d_{cum}(i,1) = |u_1 - v_i| + d_{cum}(i-1,j)\ pour\ i > 1$$

Pour les autres valeurs telles que $i + j > 2$ on obtient :

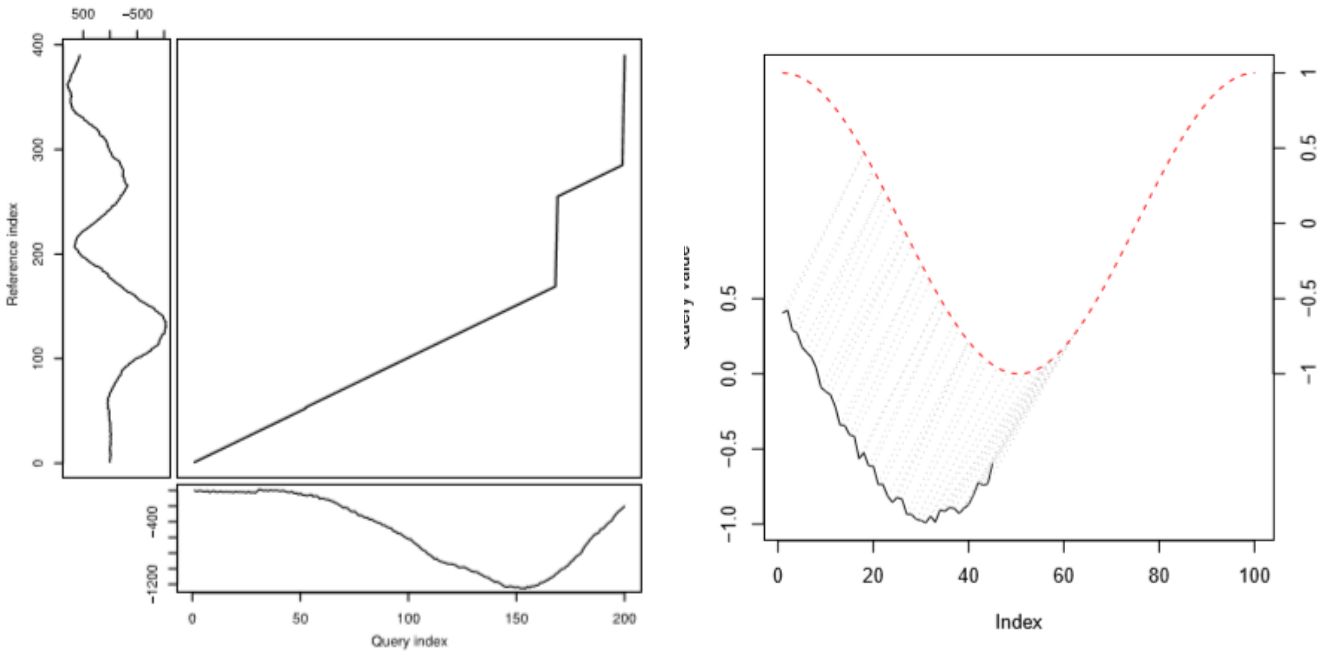
$$d_{cum}(i,j) = d_{ED}(u_i, v_j) + min\{d_{cum}(i-1,j-1), d_{cum}(i-1,j), d_{cum}(i,j-1)\}$$

Où d_{ED} correspond à la distance euclidienne. Afin de trouver la bonne correspondance entre l'ensemble des observations de U et de V, il est nécessaire de trouver le chemin W permettant de minimiser la distance cumulée. Ce chemin va donc être composé de K points avec $\max(n,m) \leq K \leq n + m - 1$ où chaque k représente un couple d'observations de U et V soit (u_i, v_j). La distance DTW optimale est donc donnée par :

$$d_{DTW}(U,V) = min \sqrt{\sum_{k=1}^{K} d_{ED}(k)}$$

La métrique permettant de mesurer la similarité existante entre deux séries, cette métrique peut être représentée de deux manières. Une première représentation basée directement sur la matrice de distance construite auparavant, où la courbe sur le graphique de gauche suivant permet de voir l'alignement des deux séries point par point. Puis une autre représentation directement effectuée sur les deux séries observées en ajoutant les liens de similarité entre celles-ci sur le graphique de droite suivant :

Figure 48 : représentation de la métrique liée à la similarité entre deux séries



D'autres mesures de similarité issues de la métrique DTW permettent d'obtenir un meilleur calcul de similarité entre deux séries temporelles. Ces métriques sont par exemples la « Derivative Dynamic Time Warping (DDTW) » prenant en compte la forme des séries temporelles à l'aide des dérivées, ou encore l'« Adaptive Feature Based Dynamic Time Warping (AFBDTW) » permettant de prendre en compte un critère local et global sur le calcul de la similarité. Ces deux autres métriques sont présentées en annexe. [Annexe 1]

Ainsi pour la suite, lors de l'application des modèles de séries temporelles ainsi que des prévisions effectuées dans le futur, les métriques RMSE ainsi que DTW seront utilisées afin de pouvoir comparer les prévisions et les observations sur différents modèles possibles.

Ces diverses notions théoriques sur les séries temporelles ayant été introduites, une partie sur l'application des modèles de séries temporelles sur les données du portefeuille va pouvoir être présentée.

3.3. Prédiction du taux d'absentéisme par série temporelle

3.3.1 Préparation des données

La première partie de ce mémoire a montré que l'étude de l'absentéisme est devenue un réel enjeu pour les assureurs, la Sécurité Sociale ou encore les dirigeants d'entreprise. Que ce soit sur la compréhension du risque, sur la réduction des coûts ou encore sur le management des salariés en fonction du risque, tous ces aspects doivent être pris en compte. Si l'étude statistique de la première partie ainsi que l'étude par *Machine Learning* dans la seconde partie ont permis d'avoir une certaine approche du risque absentéisme, l'étude et la prédiction du risque dans le futur peuvent également aider à anticiper sa survenance. De ce fait, cette partie sera consacrée à

la mise en place d'une modélisation par séries temporelles afin **de prédire sur une fenêtre de temps future les valeurs probables prises par le taux d'absentéisme en fonction d'autres variables explicatives.**

Afin de construire le modèle et de projeter le risque absentéisme dans le futur, les données utilisées proviennent donc des bases de données construites initialement. La base de données utilisée est donc la deuxième base qui avait été présentée dans la première partie de ce mémoire lors de la description des données (cf. partie 1.3.3). Cette base reprend les informations de présence et d'absence de chaque salarié et ce pour chaque jour sur la période d'étude de 2017 à 2019. Pour rappel chaque ligne représentant un salarié se trouve sous la forme suivante :

Figure 49 : Structure de la base de données utilisée pour le Forecasting

De la même manière que dans le cadre des modèles par *Machine Learning*, les résultats sur l'étude de l'absentéisme par séries temporelles ne cherchent pas à obtenir une modélisation ou des prédictions individuelles. L'enjeu réside dans l'étude de l'évolution du risque pour un groupe d'individus. Cette vision permet ainsi de lisser les résultats d'absentéisme et de proposer une solution d'anticipation du risque au niveau d'un groupe d'individus pour les assureurs ainsi que les chefs d'entreprises. Ceux-ci doivent souvent gérer un portefeuille composé de plusieurs entreprises dans le cadre des assureurs afin d'avoir un ratio de sinistralité faible sur l'ensemble du groupe, et pour les chefs d'entreprise ceux-ci doivent gérer leurs équipes et la production au sein de l'entreprise en fonction de l'ensemble du personnel à disposition. De ce fait la vision agrégée de l'absentéisme semble justifiée afin d'obtenir une vision globale du risque et d'agir en conséquence.

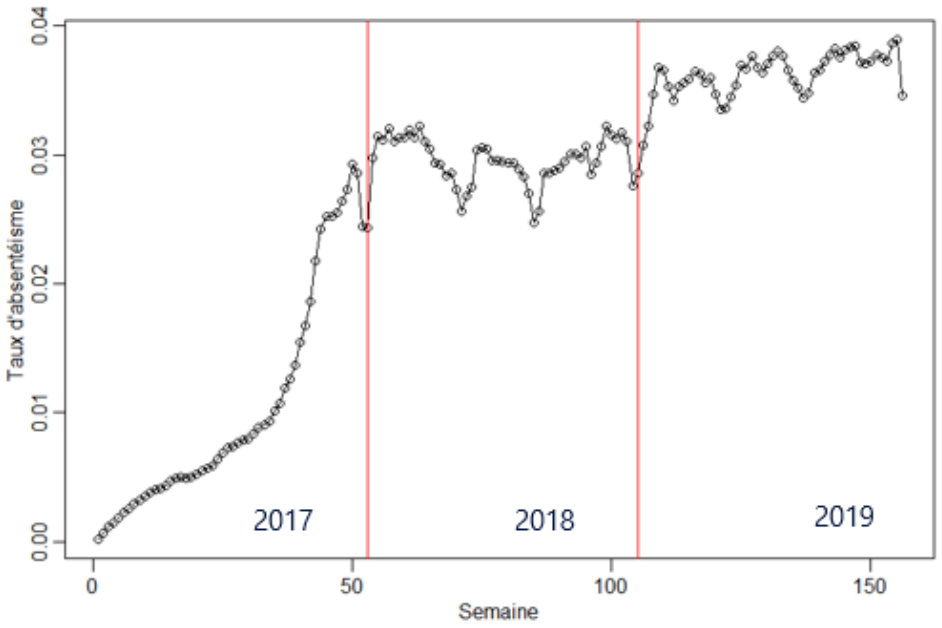
Pour rester dans le même cadre de la seconde partie, **la variable cible relative au risque absentéisme sera le taux d'absentéisme calculé sur un groupe d'individus**. Au niveau de la fréquence d'observations des différents taux d'absentéisme au cours du temps, ceux-ci peuvent être calculés annuellement, mensuellement ou de façon hebdomadaire. Néanmoins, l'historique des données est seulement de 3 ans entre 2017 et 2019. Afin de garder un maximum de données pour pouvoir construire un modèle de séries temporelles, les taux d'absentéisme seront calculés pour chaque semaine composant les 3 années d'historique, malgré la montée en charge des données DSN sur 2017 biaisant ainsi les données d'absentéisme sur cette période.

Une première étape va donc consister à construire une modélisation du taux d'absentéisme par modèle ARIMAX sur l'ensemble du portefeuille. Cette modélisation permettra

de réaliser **une prévision du taux d'absentéisme sur une fenêtre de temps future** permettant ainsi d'anticiper l'évolution probable moyenne du risque. Comme expliqué dans la partie théorique, des régresseurs doivent être introduits afin de pouvoir construire le modèle. Différentes informations peuvent servir à la modélisation du taux d'absentéisme. Ces variables doivent nécessairement être des variables comportant une évolution au cours du temps. Ainsi, les données de consommation santé sont des variables pouvant être ajoutées au modèle. A l'instar du taux d'absentéisme, la consommation santé est alors calculée hebdomadairement en sommant les nombres d'actes consommés pour chaque grand poste de santé.

Les taux d'absentéisme sur l'ensemble du portefeuille étant calculés semaine après semaine entre 2017 et 2019, la représentation du taux d'absentéisme sur l'ensemble du portefeuille et de l'historique de données peut être présentée :

Figure 50 : Taux d'absentéisme hebdomadaire sur l'ensemble du portefeuille entre 2017 et 2019



Sur l'ensemble du portefeuille, le taux d'absentéisme connaît une augmentation due à la montée en charge des données DSN sur l'année de 2017. Ensuite, ce taux se stabilise sur l'année de 2018 sur une valeur moyenne inférieure à 3%, puis se restabilise en 2019 sur un taux moyen supérieur à 3%. La construction du modèle va consister à réaliser un apprentissage sur les données de 2017 et 2018 pour pouvoir valider les prédictions faites sur l'année de 2019. La prise en compte des données de 2017 malgré leur caractère atypique est principalement due au faible historique de données disponibles. De plus, les données de santé quant à elles sont disponibles sur 2017 et ne connaissent pas de comportement atypique dû à la montée en charge de la DSN.

Les différents régresseurs utilisés sont donc **les nombres d'actes consommés par grands postes.** Cependant d'autres régresseurs sont ajoutés lors de la modélisation afin de pouvoir construire une prédiction plus juste de la réalité des données observées sur 2019. Au vu

de la représentation des taux d'absentéisme sur l'ensemble du portefeuille, des régresseurs exogènes peuvent être introduits afin d'ajuster les prédictions :

- **Le taux d'absentéisme de l'année précédente**
- **Le taux d'absentéisme global en France**
- **Les épidémies de grippe et gastro**
- **Les périodes de vacances scolaires**

Ces différentes données exogènes peuvent à priori expliquer certaines évolutions du taux d'absentéisme au cours du temps. Les modèles de *Machine Learning* ont effectivement permis de montrer que le taux d'absentéisme de l'année précédente pouvait avoir un pouvoir prédictif important.

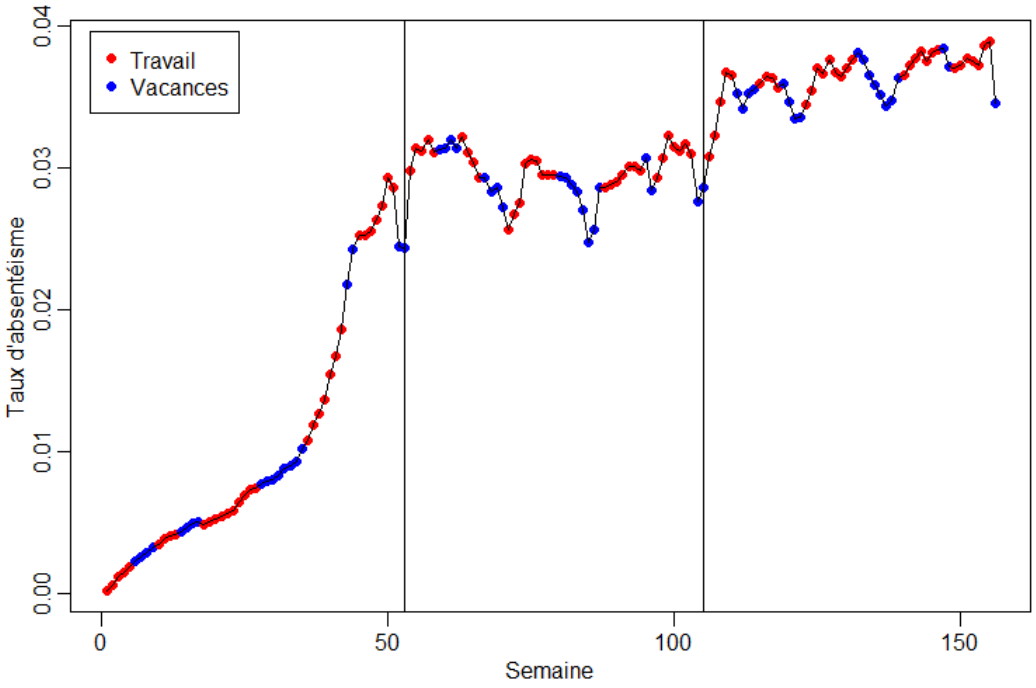
L'introduction du taux d'absentéisme de l'année précédente permet ainsi de vérifier si ce résultat se généralise également sur les modèles de séries temporelles. De plus, la série à modéliser n'ayant pas un historique de données important, la modélisation par des modèles de type SARIMA introduisant une saisonnalité n'est pas possible. En effet, il faudrait indiquer au modèle que la série a une fréquence de 52 semaines. N'ayant que deux années d'observation pour l'apprentissage, ceci n'est pas possible, d'où une autre justification de l'introduction de cette donnée pour prendre en compte la valeur de l'observation une année plus tôt.

Le taux d'absentéisme global en France tous secteurs d'activité confondus a augmenté ces dernières années. Il est possible de remarquer sur la figure précédente que le taux d'absentéisme connaît une forte augmentation entre 2018 et 2019. Afin de pouvoir ajuster les prédictions de 2019 à l'aide de cette augmentation notable sur l'ensemble des secteurs d'activité, le taux global d'absentéisme en France est un nouveau régresseur ajouté au modèle. Les valeurs du taux de 2017, 2018 et 2019 utilisées sont respectivement de 4,59%, 4,72% et 5,10%.

Si l'absentéisme peut être provoqué par les conditions de travail, celui-ci peut également être provoqué par des évènements exogènes, telles que les épidémies saisonnières de grippe ou de gastro. Un pic est notable sur les premières semaines de 2018 et 2019. Ces épidémies étant saisonnières, il est alors judicieux d'ajouter des données relatives à ces épidémies en tant que régresseurs. Les données utilisées pour ajouter cette composante proviennent du site de santé publique France et concernent les taux de passages aux urgences pour ces deux épidémies.

Enfin, le taux d'absentéisme sur l'ensemble du portefeuille connaît des phases d'augmentations et de diminutions assez régulières dans le temps. Ces périodes de diminutions semblent correspondre aux vacances scolaires comme l'indique le graphique suivant présentant le taux d'absentéisme en fonction des périodes de vacances scolaires et hors vacances scolaires. L'introduction de ce régresseur permet ainsi de différencier ces deux types de période et de ramener une saisonnalité qui pourrait ne pas être prise en compte par le modèle ARIMAX sans composante saisonnière.

Figure 51 : Evolution du taux d'absentéisme sur l'ensemble du portefeuille en fonction des vacances scolaires

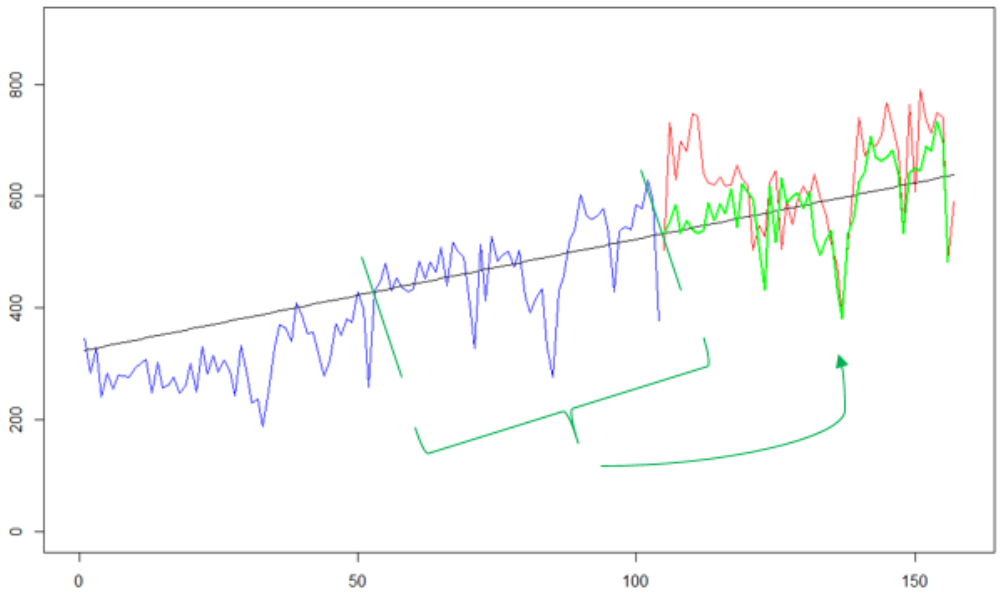


Les régresseurs utilisés lors de la modélisation ayant été présentés, une dernière étape consiste à expliquer quelles données vont être introduites lors de la prédiction du taux. En effet la modélisation intervenant sur les années 2017-2018 où les données sont connues, les prédictions faites sur 2019 supposent que les régresseurs utilisés auparavant ne sont pas encore observés durant cette période. Ainsi, des hypothèses sur les données utilisées sur la fenêtre de prédiction doivent être formulées afin de projeter le risque. Les hypothèses utilisées sont les suivantes :

- La tendance globale des données observées est maintenue sur les données utilisées lors de la prédiction
- L'évolution des données observées est reportée sur la fenêtre de prédiction sur des périodes de temps correspondantes

En outre, deux étapes sont nécessaires pour construire les régresseurs qui interviennent lors du calcul des prévisions du risque. Pour capter la tendance des données passées, une régression est effectuée puis projetée sur la fenêtre de prédiction. Puis à partir de cette régression, les évolutions passées pour la même période de temps d'une année sur l'autre sont translatées. Un exemple est présenté sur le graphique suivant où une régression linéaire a été effectuée sur les données de consultations de généralistes sur 2017-2018 puis projetée sur 2019 afin d'ajouter l'évolution des données en 2018. La comparaison des données réelles observées sur 2019 en rouge avec celles translatées en vert montre un bon ajustement.

Figure 52 : Méthode de construction des régresseurs pour la prédiction, exemple sur le poste consultations de généralistes



Ces étapes ont été effectuées sur l'ensemble des régresseurs afin de pouvoir construire des prédictions sur 2019 suite à l'apprentissage des données sur 2017 et 2018. Il aurait été également possible de construite un modèle de type ARIMA sur les régresseurs afin de les projeter. Néanmoins, au vu des résultats obtenus par la méthode détaillée plus tôt, le choix a été de ne pas complexifier la construction du modèle, surtout au vu du nombre de données observables dans le passé.

3.3.2 Application du Forecasting sur l'ensemble du portefeuille

A partir des données de 2017 et 2018 sur l'évolution du taux d'absentéisme ainsi que les divers régresseurs choisis pour modéliser cet indicateur, un modèle ARIMAX est mis en place pour prédire l'évolution future de ce taux sur l'ensemble des salariés. La construction du modèle commence par une première phase consistant à valider l'hypothèse de stationnarité des différentes données introduites dans le modèle. Le test KPSS est alors utilisé afin de tester la stationnarité des différentes séries temporelles introduites dans la modélisation, où :

$$\begin{cases} H_0 : La\ série\ temporelle\ est\ stationnaire \\ H_1 : La\ série\ temporelle\ n'est\ pas\ stationnaire \end{cases}$$

En faisant le test sur les données de 2017 à 2018, le test KPSS renvoie des *p-values* inférieures à 0,01 pour quasiment l'ensemble des données (sauf pour les 3 derniers régresseurs exogènes présentés précédemment possédant une structure stable au cours du temps, à savoir les épidémies saisonnières, les périodes de vacances scolaires), l'hypothèse nulle est alors rejetée et donc les séries ne sont pas stationnaires. Il est alors nécessaire de différencier les régresseurs ainsi que la série des taux d'absentéisme. Après une première différenciation, le test KPSS renvoie des *p-values* supérieures à 0,1 pour l'ensemble des séries. L'hypothèse nulle n'est pas rejetée, les séries temporelles sont considérées comme étant stationnaires. Un modèle linéaire est alors effectué

entre le taux d'absentéisme et les différents régresseurs introduits et différenciés pour ensuite étudier les résidus du modèle.

Les graphiques d'autocorrélation et d'autocorrélation partielle sur la série des résidus du modèle précédent sont donc étudiés pour déterminer les paramètres du modèle ARIMAX.

Figure 53 : Graphique des autocorrélations

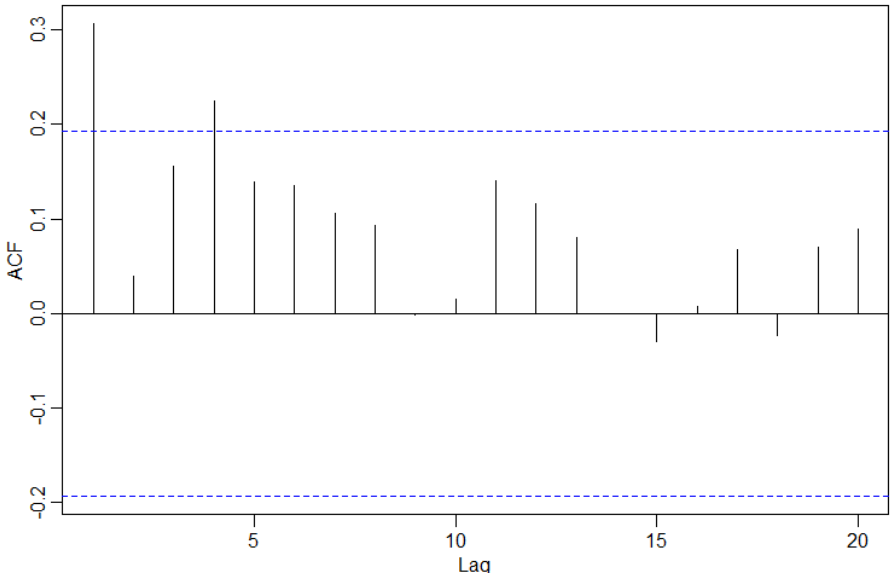
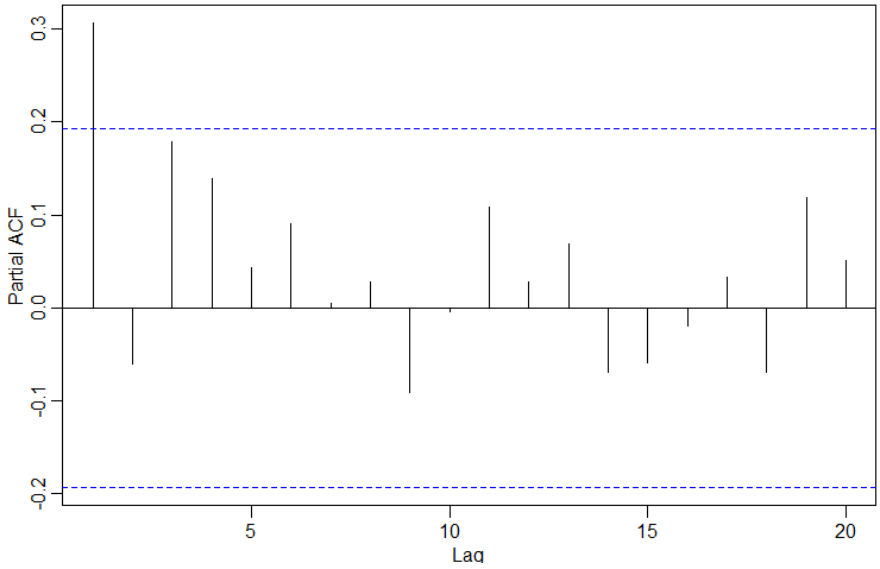


Figure 54 : Graphique des autocorrélations partielles



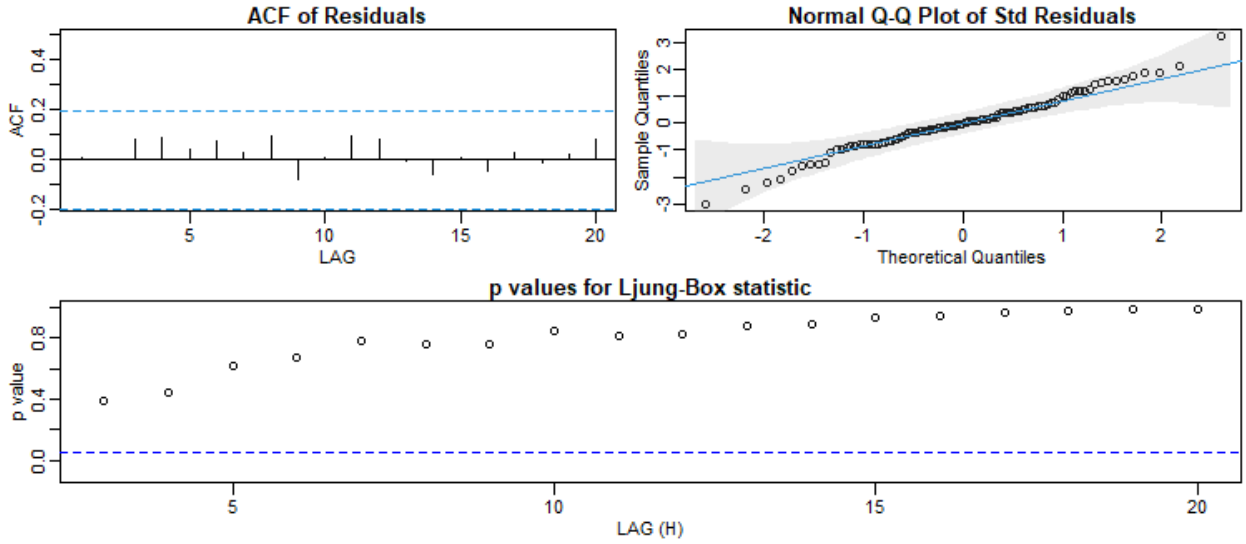
Déterminer les ordres p et q du modèle revient donc à chercher respectivement les autocorrélations et les autocorrélations partielles statistiquement différentes de zéro. D'après les graphiques, les autocorrélations sont nulles à partir du 4ème décalage de la série et du 1er pour les

autocorrélations partielles. Une première proposition consiste donc à tester le modèle ARIMAX(1,1,4). Cependant, en testant les estimations avec un test de Student, les résultats obtenus sont les suivants :

	AR1	MA1	MA2	MA3	MA4
t-statistique	-2,3190	0,1673	-3,7667	-1,2728	1,3790
p-value	0,0204	0,8671	0,0002	0,2031	0,1679

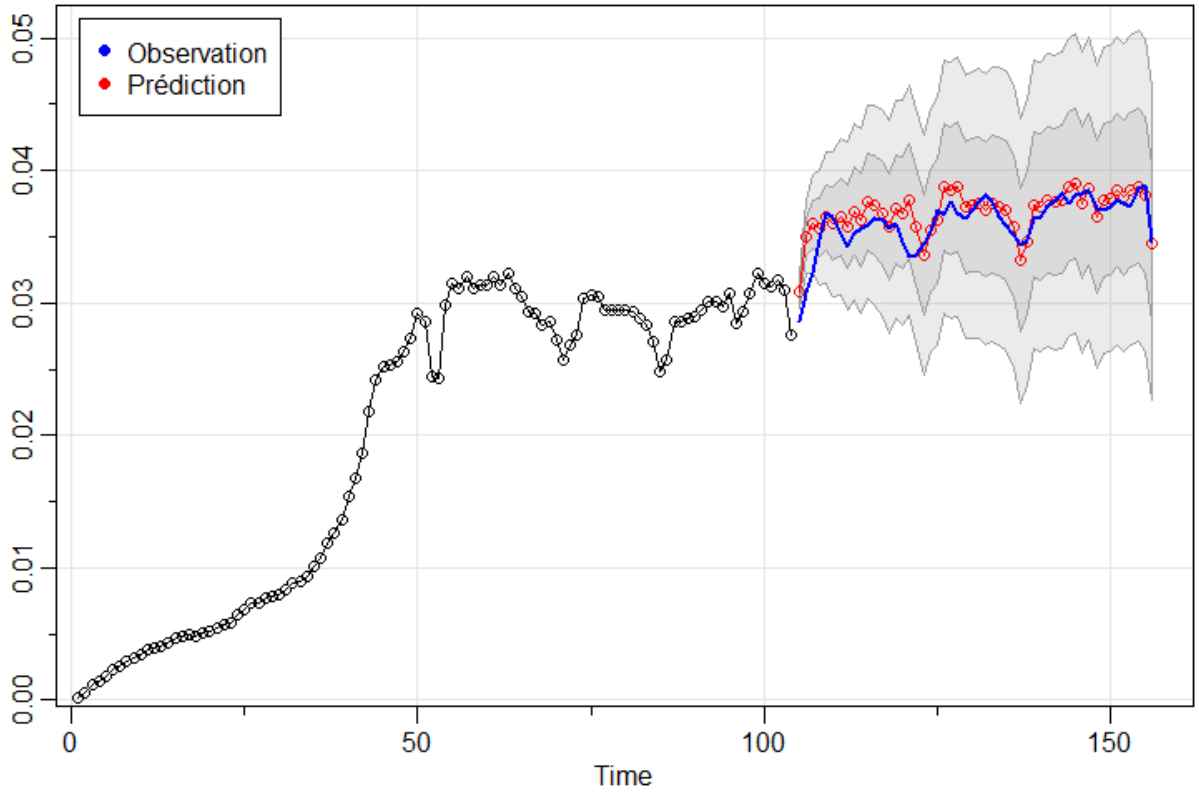
Les tests statistiques ne sont pas significatifs pour les coefficients de la composante en moyenne mobile. Le processus consiste donc à retirer progressivement les termes de plus haut degré jusqu'à obtenir des coefficients significatifs. Au final, le meilleur modèle est obtenu en utilisant un ARIMAX(1,1,1). Les paramètres ayant été définis, il est nécessaire de statuer sur la blancheur des résidus ainsi que sur la non-présence d'autocorrélation. Deux tests sont donc utilisés, le test de Ljung-Box pour statuer sur l'absence d'autocorrélation des résidus et le test de Jarque-Bera pour la blancheur des résidus. Les deux p-values étant supérieures à 0,05, les hypothèses nulles ne sont donc pas rejetées ce qui confirme statistiquement l'absence d'autocorrélation ainsi que de la blancheur des résidus. Ces éléments peuvent être résumé par les graphiques suivant présentant les autocorrélations des résidus, le *Q-Q plot* ainsi que les *p-values* du test de Ljung-Box pour les 20 premiers décalages :

Figure 55 : Vérifications des hypothèses d'absence d'autocorrélation et de blancheur des résidus



Ces vérifications ayant été faites, il est possible de projeter le taux d'absentéisme sur une fenêtre de prédiction future. L'exercice va donc consister à prédire l'évolution du taux d'absentéisme sur 2019 et de comparer les prédictions aux observations faites sur cette période. La prédiction moyenne ainsi que les intervalles de confiance à 80 et 95% sont construits lors d'une première étape sur l'ensemble de l'année de 2019. A partir du modèle construit auparavant, l'évolution moyenne prédite sur 2019 est la suivante :

Figure 56 : Prédiction du taux d'absentéisme sur 2019



L'évolution du taux d'absentéisme prédite de manière hebdomadaire sur 2019 semble donc avoir un bon ajustement avec les observations, les observations se trouvant dans les intervalles de confiance à 80 et 95%. Si les modèles de *Machine Learning* présentés permettait d'avoir uniquement une prédiction moyenne annuelle, cette nouvelle approche présentée ci-dessus permet en revanche d'apprécier l'évolution du taux d'absentéisme avec ses fluctuations synonymes d'augmentation ou de diminution d'absences dans les entreprises.

Ces modèles présentent également l'avantage d'être mis à jour au cours du temps facilement. Plus le temps passe plus de nouvelles données sont observables et peuvent donc être ajoutées au modèle pour mettre à jour les prédictions. Il suffit de recalculer les paramètres du modèle comme fait précédemment pour la prédiction sur l'année entière et de projeter les nouveaux taux d'absentéisme sur une nouvelle fenêtre de prédiction. Une mise à jour est donc envisagée en considérant qu'après chaque trimestre de nouvelles observations sont obtenues. Trois nouvelles prédictions sont ainsi faites à savoir sur les 3 derniers trimestres, sur le second semestre ainsi que sur le dernier trimestre de 2019. En utilisant la même démarche présentée pour la prédiction sur l'année entière, les modèles calculés sont respectivement des ARIMAX(1,1,0), ARIMAX(1,1,1) et ARIMAX(2,1,0). Les résultats des 4 modèles de prédiction sont ainsi présentés graphiquement pour voir l'évolution de la mise à jour du modèle :

Figure 57 : Prédiction sur l'année 2019

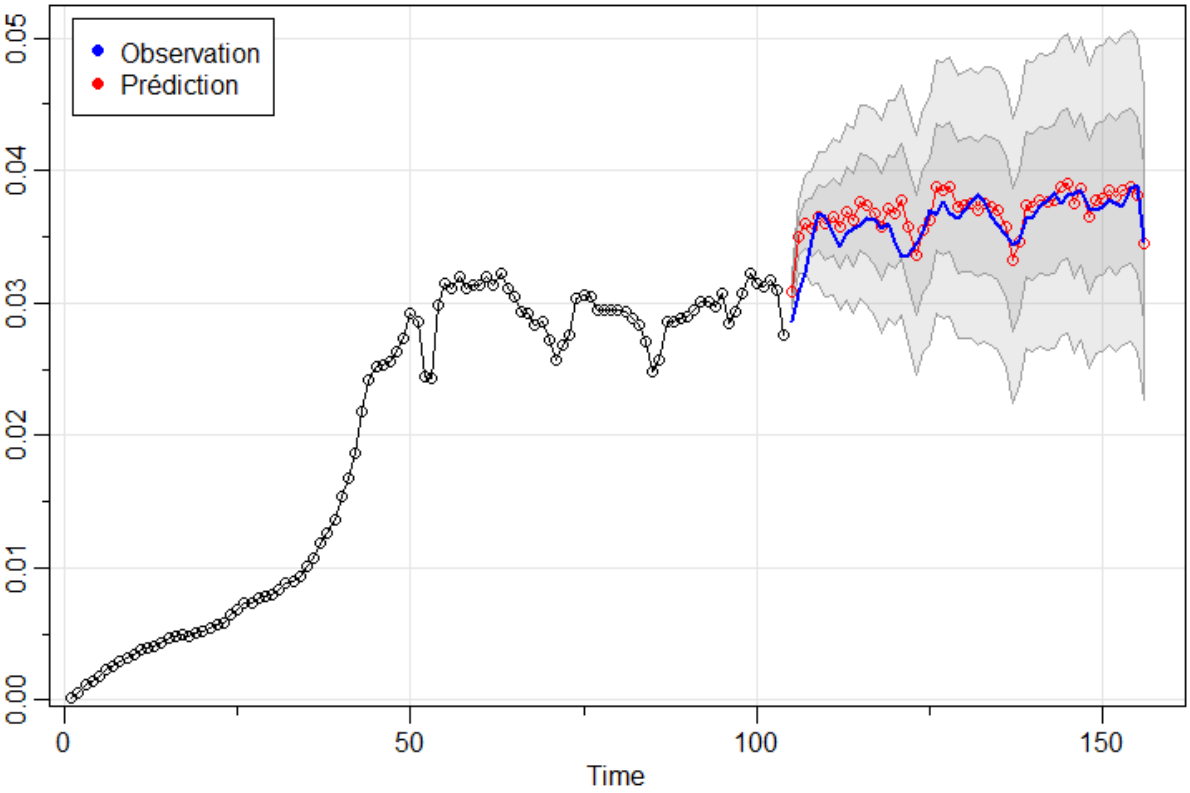


Figure 58 : Prédiction sur les 3 derniers trimestres de 2019

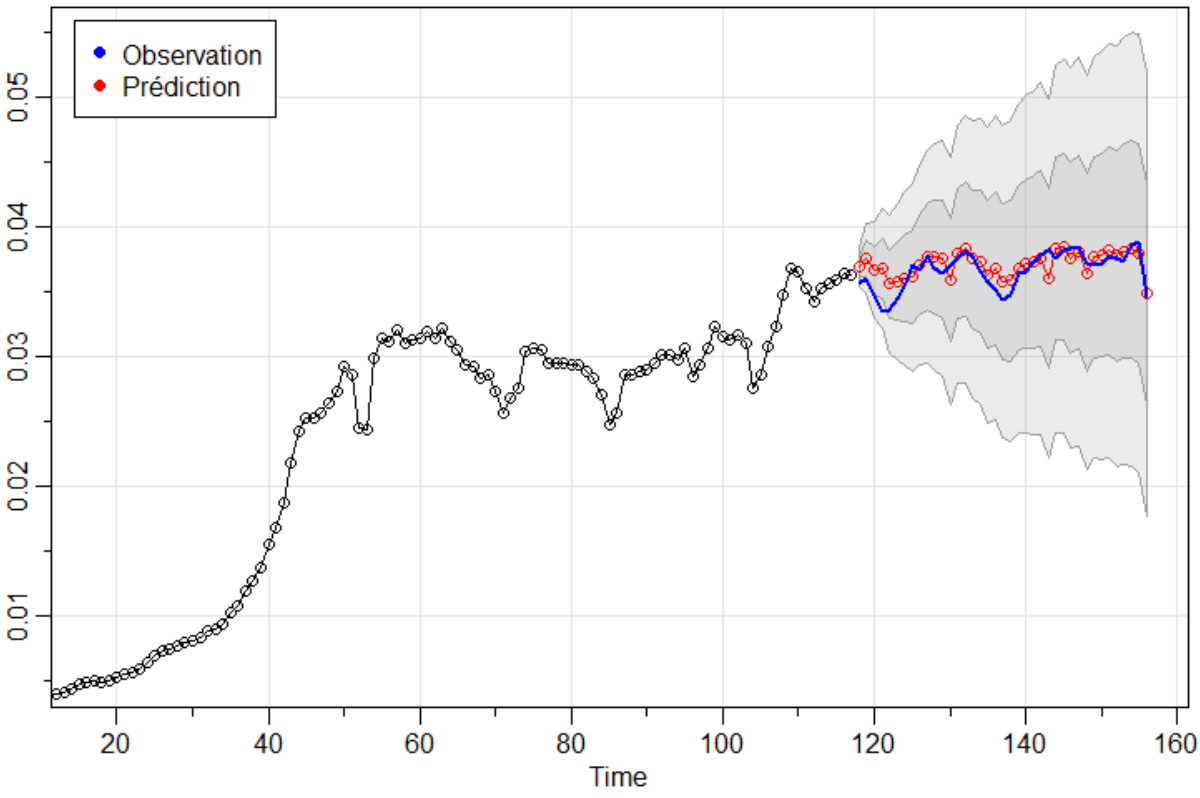


Figure 59 : Prédiction sur le second semestre de 2019

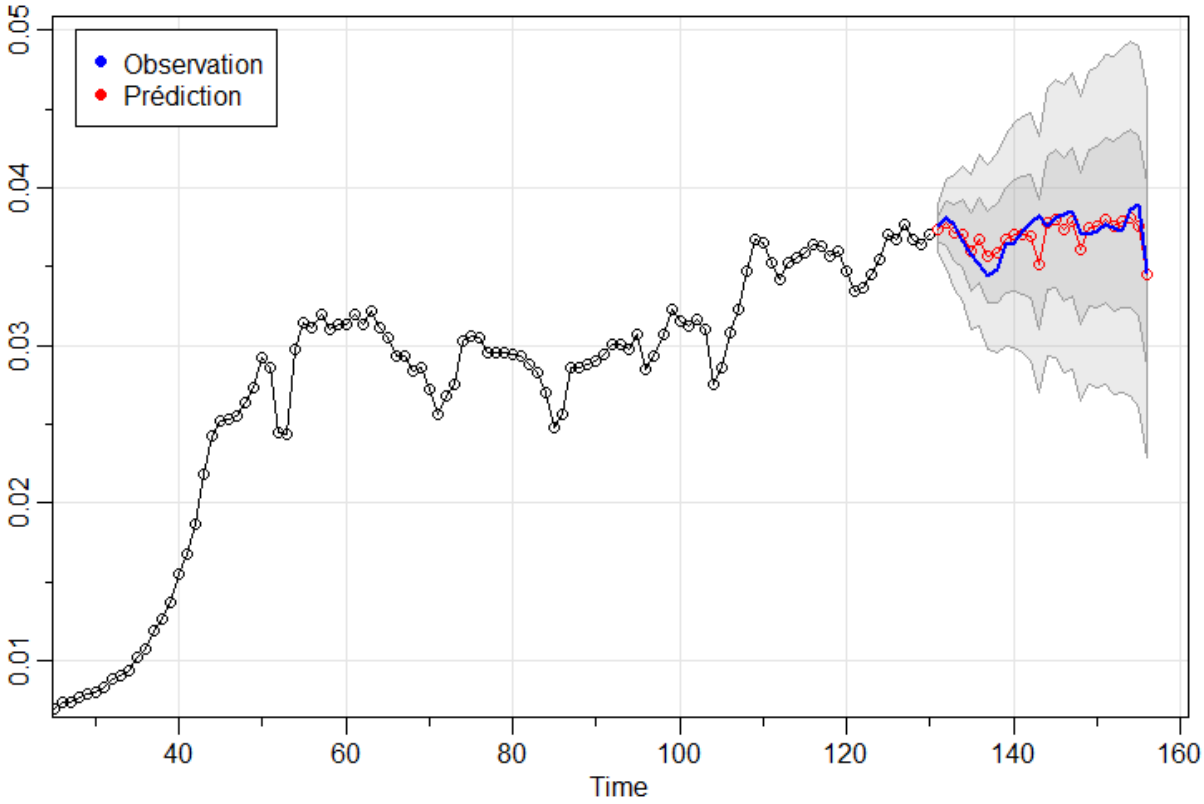
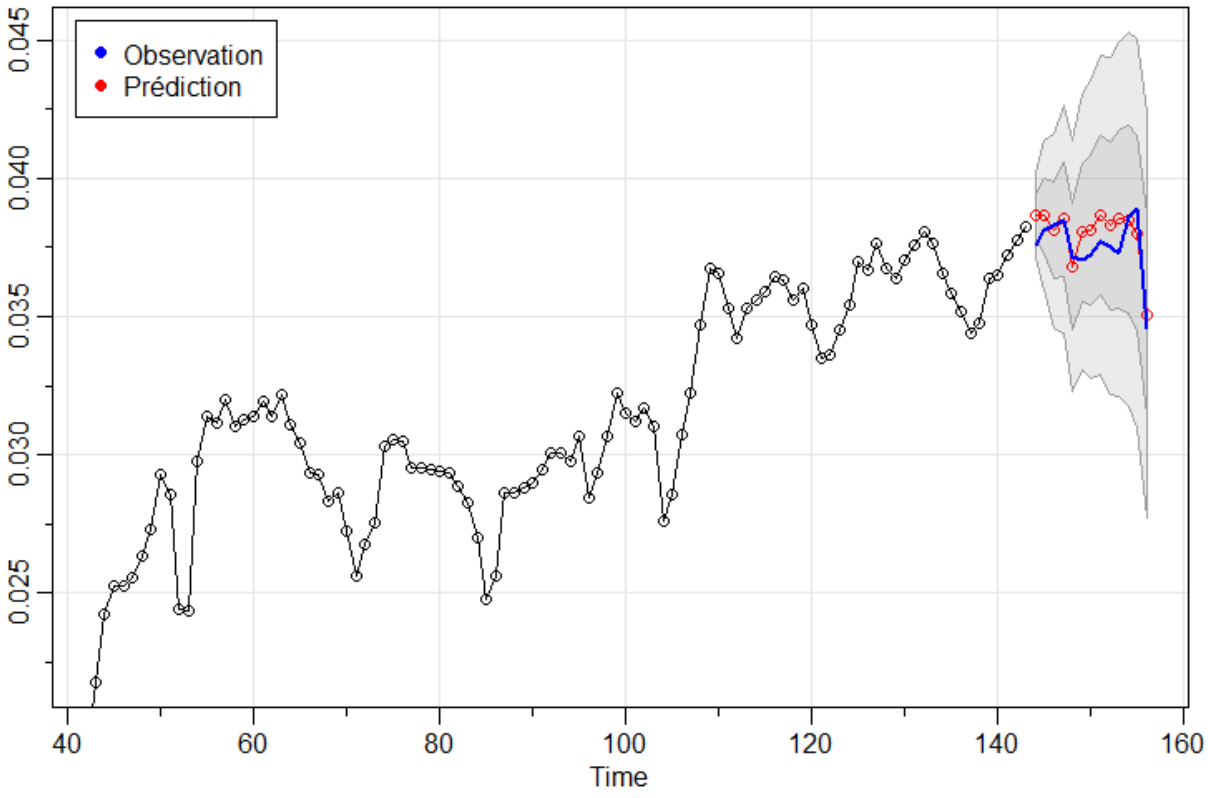


Figure 60 : Prédiction sur le dernier trimestre de 2019



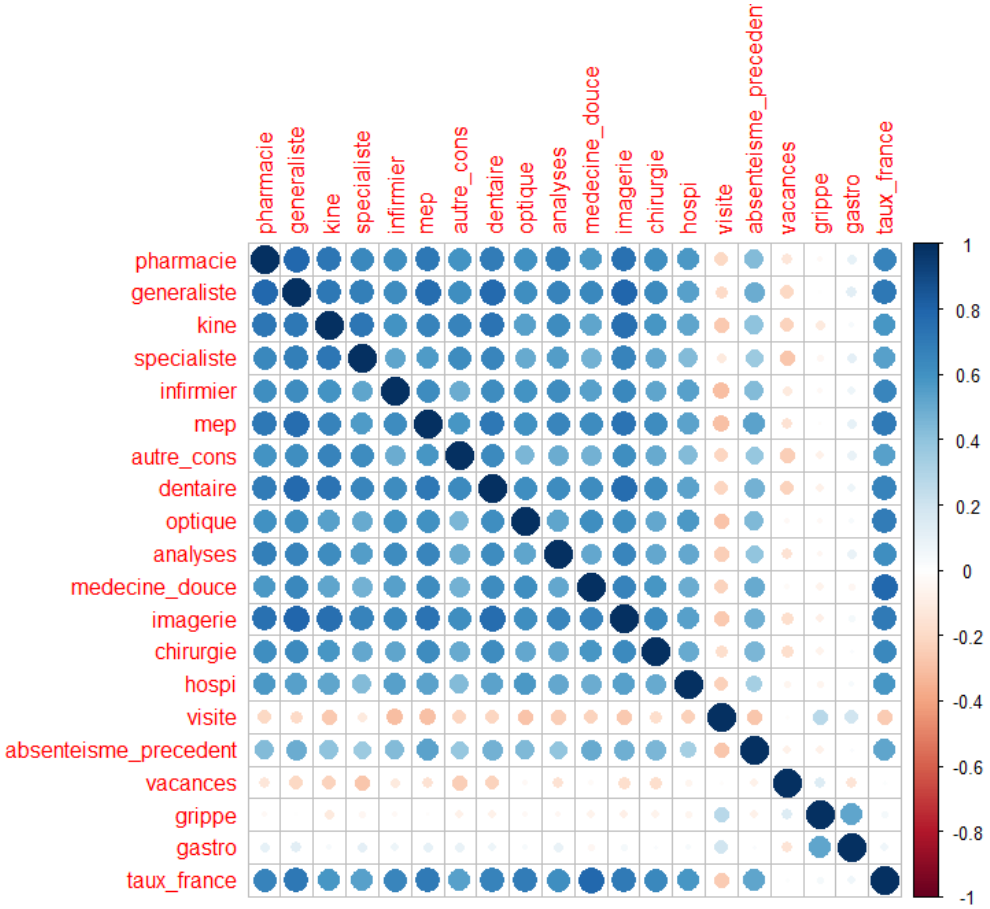
La mise à jour du modèle permet de projeter le taux d'absentéisme en fonction des nouvelles données observables. Les divers graphiques ont permis de montrer que l'adéquation avec les observations étaient de bonne qualité et l'évolution du taux d'absentéisme était bien prise en compte.

3.3.3 Explicabilité des résultats et amélioration du modèle

Multicolinéarité

Ces diverses modélisations utilisent un nombre important de régresseurs. La majorité de ces régresseurs ont la particularité d'être des données de consommation santé. Une forte corrélation existe donc entre ces différentes variables, montrée par le corrélogramme suivant :

Figure 61 : Corrélogramme des différentes variables



Cette forte corrélation peut engendrer un problème de multicolinéarité dans l'estimation des coefficients obtenus par la régression effectuée précédemment. Cette multicolinéarité entraîne une augmentation de la variance des coefficients, les rendant plus instables et difficiles à interpréter. En outre, les possibles problèmes rencontrés peuvent être les suivants :

- Du fait d'une trop grande variance estimée par le modèle, les coefficients peuvent sembler non significatifs malgré une relation significative existante entre la variable prédite et une variable explicative
- Les coefficients estimés peuvent fortement varier d'un échantillon à un autre
- La suppression de l'une des variables dans le modèle peut significativement modifier les coefficients calculés pour les autres variables

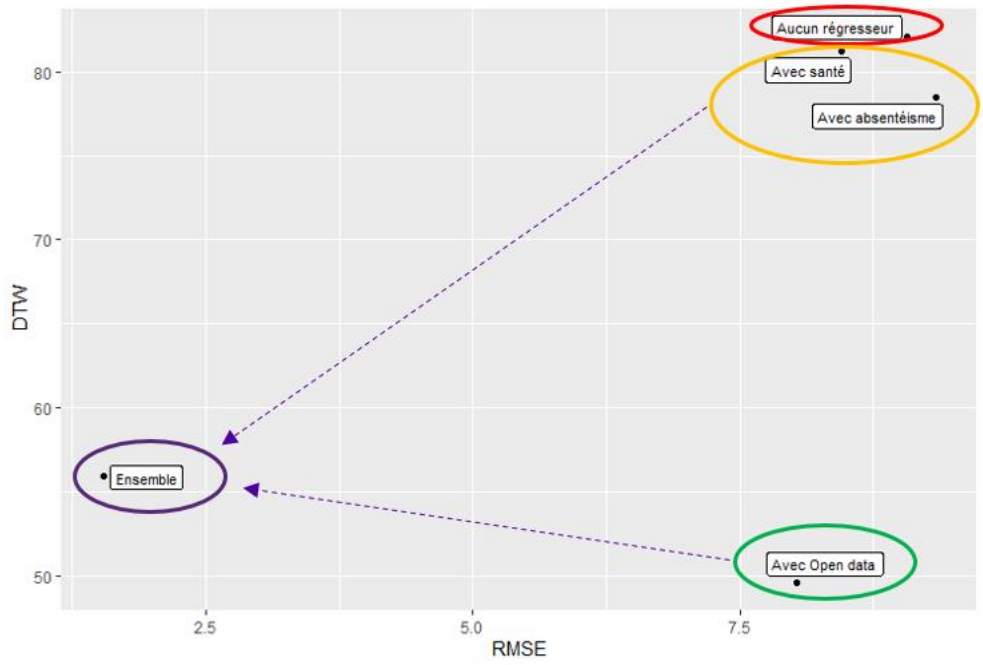
Après calcul du *variance inflation factor* (VIF), certaines variables possèdent des VIF supérieurs à 10 indiquant la présence de multicolinéarité dans le modèle. Il faut donc rester vigilant quant à l'interprétation des résultats du modèle au travers des coefficients. Cependant, la multicolinéarité n'a aucune incidence sur l'adéquation de l'ajustement, ni sur la qualité de la prévision.

Quantification de l'apport de chaque régresseur

Une approche permettant de quantifier l'effet de chaque régresseur ou groupe de régresseurs sur les prédictions effectuées est alors mise en place à l'aide des métriques présentées précédemment. A partir du modèle initial, chaque régresseur ou groupe de régresseurs va être enlevé du modèle afin de mesurer l'impact sur la prédiction à l'aide des deux métriques. Il s'agit ainsi d'un problème d'optimisation d'un front de Pareto en cherchant à minimiser à la fois le RMSE de la prédiction et le DTW à la suite du retrait d'une ou plusieurs variables.

Une première étape consiste à étudier l'impact des variables par groupes de régresseurs. Comme présenté plus tôt dans ce mémoire, les différents régresseurs peuvent être classés en différents types : santé, absentéisme ou *open data*. L'ajout d'un de ces groupes ou de l'ensemble des régresseurs permet d'apprécier le gain sur la justesse de la prédiction effectuée par le modèle. Les différents résultats en fonction des métriques sont alors tracés sur le graphique suivant, renseignant pour chaque point l'apport des régresseurs dans la justesse des prédictions.

Figure 62 : Comparaison des modèles par ajout de groupes de régresseurs



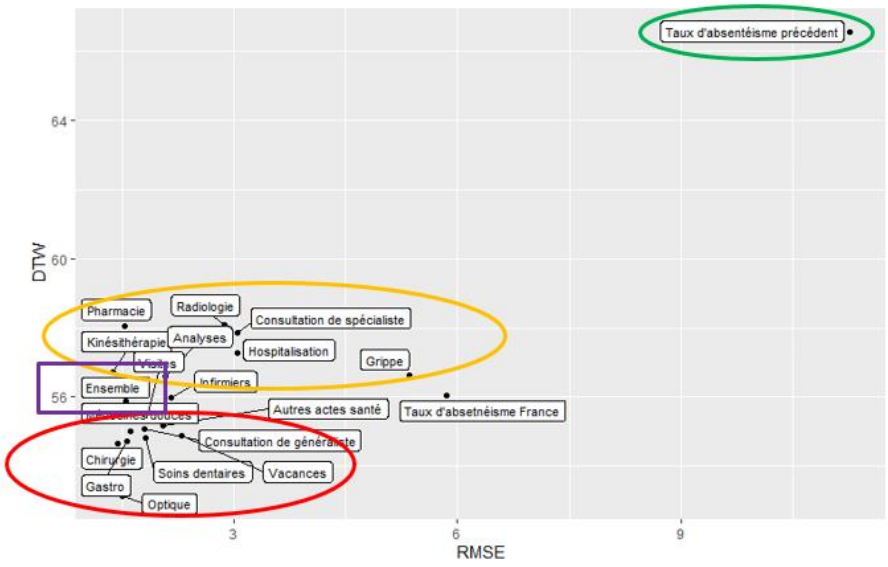
Cette première comparaison des modèles par ajout de groupes de régresseurs permet d'observer l'évolution des prédictions faites sur une année entre un modèle sans régresseur (et de ce fait un modèle uniquement de type ARIMA) et le modèle complet noté « Ensemble » où celui-ci contient l'ensemble des régresseurs. Les premières conclusions pouvant être tirées sont les suivantes :

- Ajouter des régresseurs dans la modélisation permet **d'améliorer les prédictions du taux d'absentéisme dans le futur**
- **Les régresseurs liés aux données santé et à l'absentéisme permettent d'améliorer légèrement le modèle**
- **Les données Open Data diminue considérablement la mesure du DTW**. Ces données semblent donc apporter de **l'information dans la modélisation en termes de saisonnalité de la série**. En effet, les vacances scolaires ainsi que les maladies saisonnières introduites dans le modèle peuvent potentiellement améliorer les prédictions d'augmentation ou de diminution de l'absentéisme sur certaines périodes
- **L'ajout simultané de l'ensemble des régresseurs semble donner un modèle optimal, diminuant à la fois le RMSE et le DTW des prédictions contre les observations**

L'ajout des différents groupes de régresseurs permet dans un premier temps d'étudier l'amélioration des prédictions. Cependant, l'étude de l'impact de chaque régresseur et de leur contribution reste un sujet important pour comprendre quelles sont les variables les plus influentes permettant de moduler au mieux la prédiction du taux d'absentéisme. La démarche consiste à partir du modèle comprenant l'ensemble des régresseurs (noté « Ensemble ») et

d'enlever un à un les régresseurs pour évaluer chaque impact, à l'instar d'un indice de Sobol ou de Shapley mais calculé empiriquement.

Figure 63 : Comparaison des modèles par retrait de chaque régresseur



Au travers de ce dernier graphique, il est possible d'observer l'impact de chaque régresseur. Les diverses conclusions qui peuvent être émises sont les suivantes :

- **Le taux d'absentéisme observé dans le passé a un fort impact sur l'amélioration de la prédiction**. En outre, le retrait de cette variable fait augmenter le RMSE ainsi que le DTW. De ce fait, dans ce cas précis où la série est assez stable au cours du temps, **les données passées du taux d'absentéisme sont prédictives du taux futur**
- Les effets du retrait des données de santé et des Open Data sont assez proches au vu du graphique. Cette agglomération autour du modèle central comportant l'ensemble des régresseurs indique la forte corrélation existante entre celles-ci. Cependant deux groupes se distinguent :
  - Un premier groupe de régresseurs se situant au-dessus du modèle central, signifiant que leur impact contribue à **une meilleure prédiction dans le futur du taux d'absentéisme. Ces régresseurs correspondent majoritairement à des données liées à de l'absentéisme, telles que l'hospitalisation, la kiné couplée à la radiologie, les consultations de spécialiste ou encore la grippe**
  - Le second groupe se situe en-dessous du modèle central, ce qui a pour conséquence **d'améliorer le modèle lorsque ces régresseurs sont retirés de la modélisation. Ces données sont moins liées à l'absentéisme, comme par exemple les données de consommation sur l'optique, les soins dentaires ou**

encore la médecine douce. Cependant, d'autres variables plus en lien avec l'absentéisme à priori semblent non significatives dans l'amélioration du modèle comme les données sur la gastro, les vacances scolaires ou encore la chirurgie. Néanmoins, les données étant corrélées et la présence de multicolinéarité étant probable, d'autres variables peuvent détenir la même influence que celles-ci. Typiquement, la grippe ou encore l'hospitalisation peuvent détenir la même information que gastro ou chirurgie

Cette manière de calculer l'impact des variables empiriquement permet d'avoir un premier regard sur l'explicabilité de celles-ci. Elle permet également de donner une possible amélioration de la prédiction en retirant les régresseurs faisant augmenter les métriques RMSE et DTW. Par exemple, à l'aide du graphique précédent, on s'aperçoit qu'en enlevant les régresseurs gastro, chirurgie, optique et soins dentaires les résultats sur les métriques par rapport au modèle central sont les suivants :

	Modèle central	Modèle retravaillé
RMSE	1,54	1,36
DTW	55,89	53,65

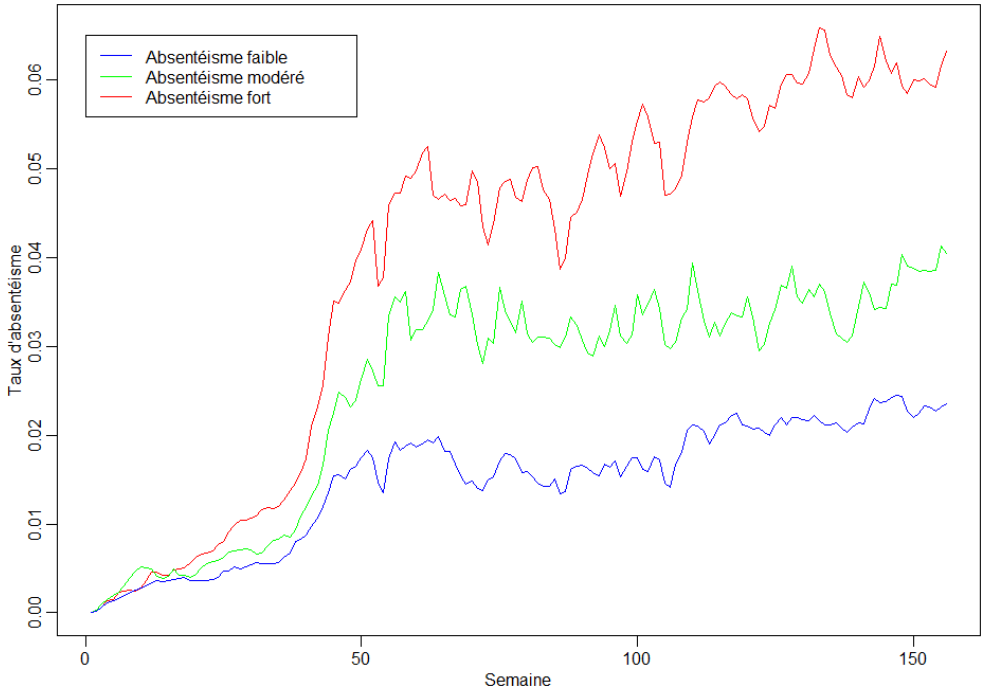
3.4. Détection de changement de tendance

Les travaux réalisés dans la partie précédente ont permis de modéliser et projeter le taux d'absentéisme de façon dynamique en fonction des divers régresseurs. L'adéquation des observations avec les prédictions effectuées, malgré un historique des données restreint, provient en majorité de **la régularité de la série au cours du temps, en termes de tendance ou de saisonnalité. Le modèle ARIMAX du fait de la régularité des régresseurs, parvient à proposer une projection du risque ayant quasiment la même tendance que les observations passées.**

Cependant, cette régularité sur la série a été obtenue en se concentrant sur l'ensemble du portefeuille. Cette mutualisation des données sur l'ensemble du portefeuille permet ainsi d'obtenir une certaine régularité. Or l'étude de l'absentéisme sur des sous-segments du portefeuille peut entraîner une observation de fluctuations du taux d'absentéisme plus importantes qui auraient pour conséquence une mauvaise modélisation du risque et ainsi réaliser de mauvaises projections dans le futur.

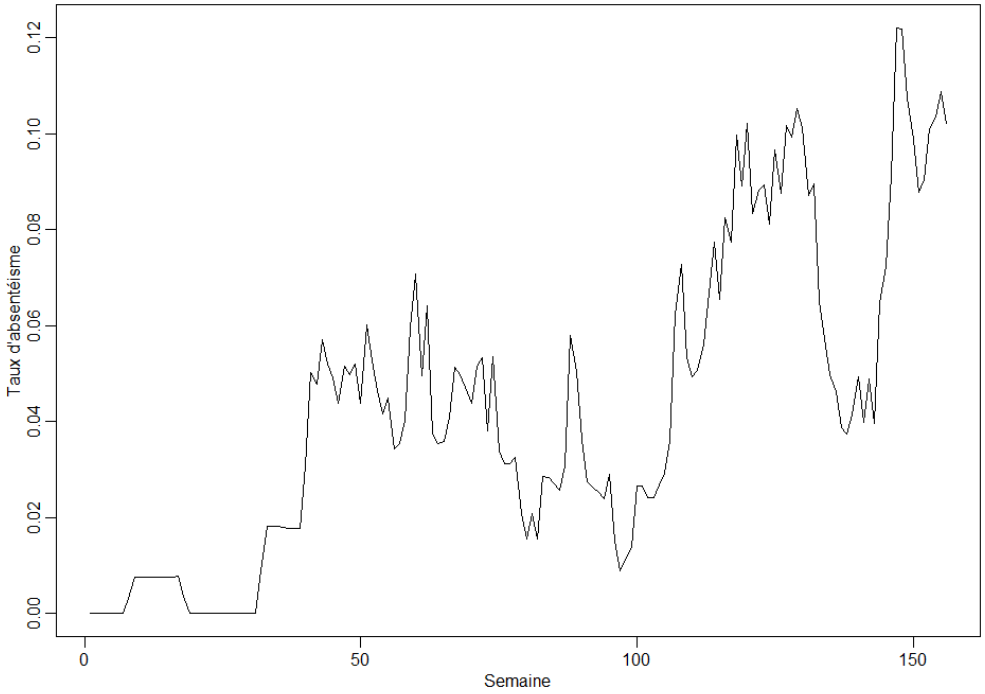
La régularité du taux d'absentéisme sur des sous-segments peut être obtenue à l'aide d'une segmentation cherchant à classifier les individus en fonction de leur profil d'absentéisme présentée dans la partie 2.2.2. En examinant l'évolution du taux d'absentéisme sur les profils faible, modéré et fort de l'absentéisme en 2019, il est possible de voir que les séries temporelles du taux d'absentéisme obtenus pour chaque groupe restent stables, comme pour le taux d'absentéisme étudié sur l'ensemble du portefeuille.

Figure 64 : Evolution du taux d'absentéisme entre 2017 et 2019 pour les 3 profils d'absentéisme



En revanche, étudier l'absentéisme sur des sous-segments du portefeuille basés sur des critères exogènes à l'absentéisme peut conduire à étudier des profils d'absentéisme hétérogènes, et de ce fait entraîner des fluctuations du taux d'absentéisme trop importantes pour pouvoir utiliser la méthode de modélisation présentée précédemment. Par exemple, étudier le taux d'absentéisme en fonction d'un type d'activité présent dans l'entreprise confirme ces craintes comme le montre le graphique suivant :

Figure 65 : Taux d'absentéisme entre 2017 et 2019 pour l'activité de Manutention



Appliquer une méthodologie similaire de modélisation sur des données présentant cette configuration pourrait entraîner une mauvaise prédiction du risque, le taux d'absentéisme ayant fortement augmenté entre 2018 et 2019. Il se peut donc que la période d'entrainement du modèle sur l'historique passé des données ne puisse plus convenir à une projection fidèle du risque en raison de ce changement soudain de tendance ou de stabilité du modèle.

Pour résoudre cette problématique, une nouvelle statistique est introduite afin de **détecter statistiquement les ruptures de modèle dans des systèmes dynamiques** tels que ceux-ci, faisant intervenir également des interactions entre la variable cible modélisée et les variables explicatives introduites. Cette étape de vérification de la stabilité au cours du temps des modélisations effectuées peut ainsi permettre de répondre à deux objectifs issus de la méthodologie appliquée précédemment :

- Vérifier la stabilité du lien entre régresseurs et variable cible, permettant de donner plus de crédibilité aux prévisions effectuées
- Etudier les nouvelles données observées au fur et à mesure du temps pour établir un potentiel changement de tendance impactant les futures prédictions

Cette vérification va porter sur le modèle linéaire construit entre les différents régresseurs et la variable cible. Pour rappel, le modèle s'écrit :

$$y_i = x_i^T \beta_i + u_i \quad , (i = 1, \dots, n)$$

Où les y_i correspondent aux observations de la variables cible aux temps i (ici le taux d'absentéisme au cours du temps), x_i le vecteur des observations au temps i des k régresseurs présents, β_i les coefficients du modèle et u_i les résidus.

Proposer un test étudiant la possibilité que le modèle n'ait pas une structure stable au cours du temps revient à poser l'hypothèse nulle suivante :

$$H_0 :\ \beta_i = \beta_0\ , (i = 1, \dots, n)$$

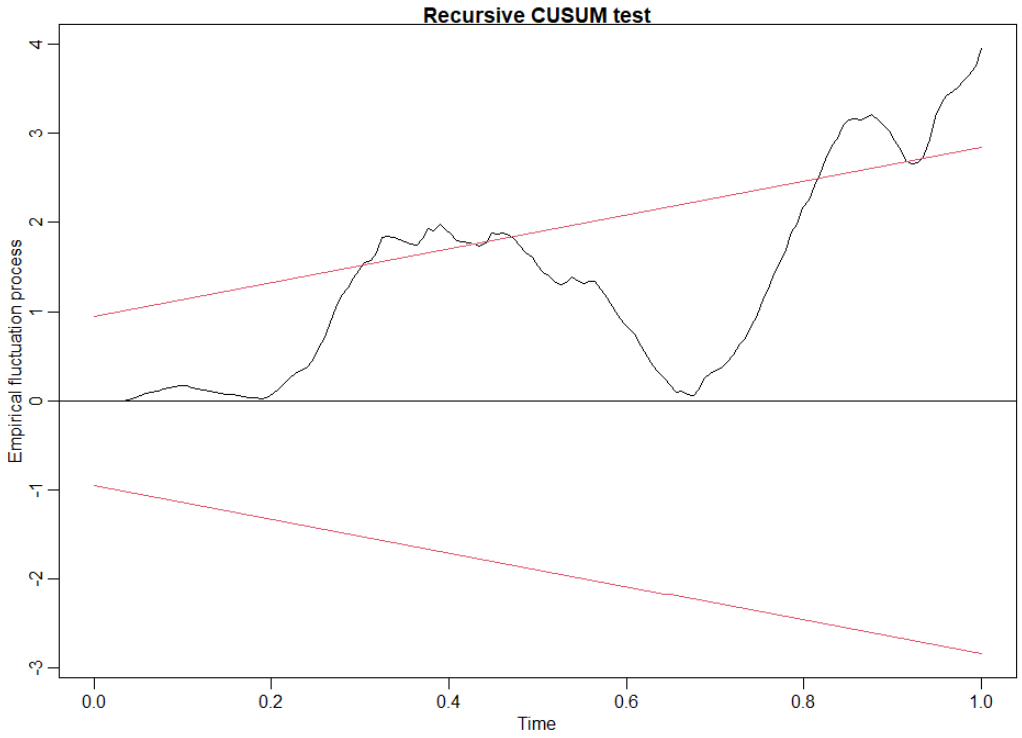
Afin de valider ou d'infirmer cette hypothèse sous une certaine probabilité α, un test de fluctuation généralisé (Generalized fluctuation test) est mis en place. Le principe même de ce test est de construire un modèle entre la variable cible et les différentes variables explicatives, puis de calculer empiriquement et pas à pas les possibles fluctuations existantes, soit sur les résidus du modèle, soit directement sur les estimations des β. Ce processus itératif se nomme *Empirical fluctuation process* (EFP), et peut être construit à partir de processus différents.

Pour la suite, les fluctuations étudiées seront principalement basées sur les résidus et non sur les estimations des β. En effet, comme expliqué précédemment, la forte corrélation et la possible multicolinéarité existante entre les variables explicatives peuvent biaiser les estimations de ces derniers. Ainsi, étudier les fluctuations des résidus reste une méthode plus fiable de capter de potentiels changements de structures dans le modèle.

Deux processus peuvent être appliqués, celui utilisant la statistique CUSUM étudiant la somme cumulée des résidus standardisés, ou la statistique MOSUM étant plus dynamique, utilisant une partie des résidus standardisés. De même, les résidus empiriques utilisés dans le calcul de ces deux statistiques peuvent soit provenir des moindres carrés ordinaires, soit être calculés récursivement. L'ensemble de ces informations contenus dans ce paragraphes se trouvent en annexe pour des explications complémentaires. [Annexe 2]

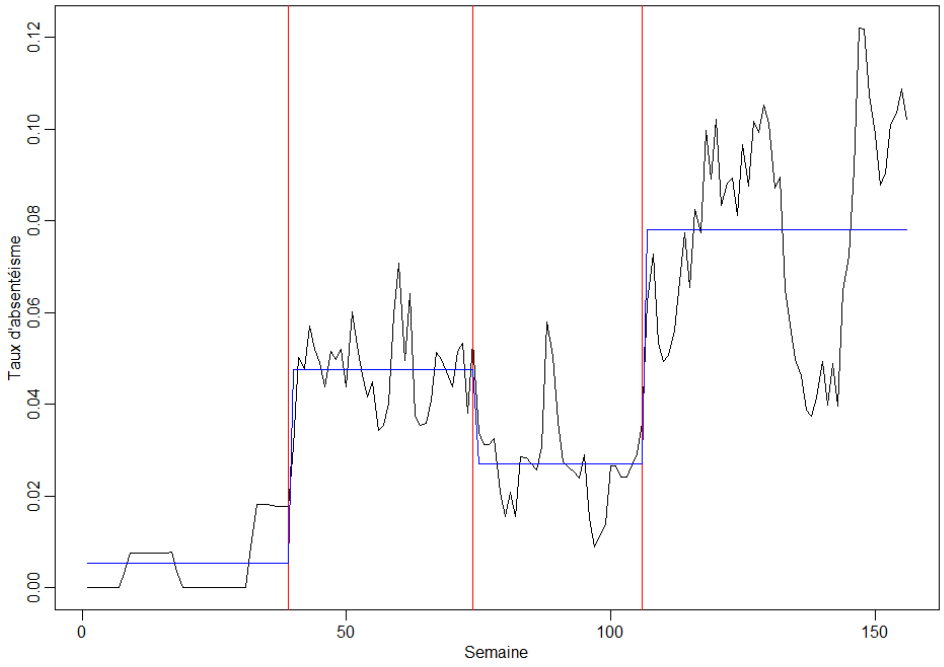
A l'aide de cette nouvelle méthodologie, il est alors possible **d'estimer de potentiels changements de structure dans le modèle, mais également de déterminer les dates de rupture dans cette dynamique.** En reprenant l'exemple du taux d'absentéisme calculé pour les salariés travaillant dans le secteur de la manutention et en calculant les régresseurs pour ce segment, il est possible d'appliquer cette méthodologie de rupture. Le processus utilisé est basé sur la statistique du CUSUM sur des résidus dits récursifs. Sur le graphique suivant, la courbe de la statistique calculé au cours du temps coupe l'intervalle en rouge à plusieurs reprises. L'hypothèse H_0 de stabilité du modèle peut donc être rejetée avec une probabilité de 95%.

Figure 66 : Test de stabilité du modèle entre le taux d'absentéisme et l'ensemble des régresseurs pour les salariés dans la manutention



Les dates de ruptures peuvent ensuite être calculées afin d'observer les périodes où le modèle reste relativement stable au cours du temps. D'après les résultats obtenus et montrés sur le graphique suivant, 3 temps de ruptures sont observables, avec des paliers de taux d'absentéisme moyens bien identifiables :

Figure 67 : Périodes de stabilité et de ruptures du modèles en fonction du taux d'absentéisme



Cette nouvelle approche permet ainsi de s'interroger sur la raison de ces diverses ruptures dans la stabilité du taux d'absentéisme. **Un travail d'investigation plus approfondi** peut alors être mis en place pour **comprendre davantage les raisons d'un changement brutal** (à la hausse ou à la baisse) de cette variable cible. Ici par exemple, l'étude portait sur les salariés travaillant dans le secteur de la manutention. **L'apport de nouvelles données autres que celles présentées jusqu'ici avec une meilleure explicabilité de la dynamicité du taux d'absentéisme pourraient permettre d'améliorer la stabilité du modèle, et ainsi permettre d'effectuer des prédictions plus robustes en termes d'adéquations avec les futures observations.** Sur des sous-segments basés sur l'activité professionnelle par exemple, l'apport de données RH comme la satisfaction au travail, la charge de travail ou encore les méthodes de management pourraient apporter plus d'explicabilité aux changements brutaux de taux d'absentéisme aux différents points de ruptures.

3.5. Intérêt dans le contexte actuel (COVID-19)

Depuis fin 2019, la pandémie de COVID-19 est venue altérer le monde dans lequel nous vivions avec de forts impacts financiers, économiques, sanitaires et sociaux. Les périodes d'incertitudes liées à l'avancée du virus sur le territoire ainsi que les nombreuses restrictions mises en place pour contrer ce phénomène ont notablement touché les entreprises. Avec l'instauration du chômage partiel, des confinements successifs ainsi que des couvre-feux, la vie des salariés a été chamboulée, entraînant par la même occasion un contexte de travail sans précédent.

Outre les cas avérés de personnes infectées par le nouveau virus, d'autres causes entrent en jeu dans l'évolution de l'absentéisme durant cette période particulière. Selon certaines études, l'application du chômage partiel ou du télétravail ont eu des effets négatifs sur la santé individuelle ainsi que sur la santé au travail, pouvant favoriser l'apparition de troubles responsables d'arrêts maladie. Durant les périodes de fortes restrictions sanitaires, les principales raisons d'une dégradation de la santé des salariés sont les suivantes :

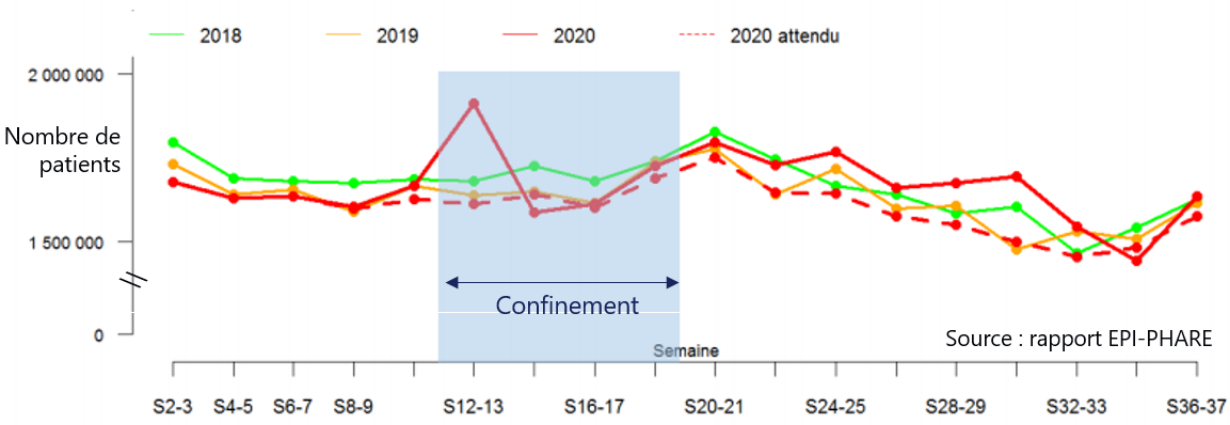
- Une baisse de la pratique physique pouvant entraîner des troubles musculosquelettiques
- Une dégradation du sommeil en raison du stress
- Une dégradation de l'hygiène alimentaire

Ces exemples de dégradations sur la santé physique des individus parmi tant d'autres, ont un impact pouvant directement mener à un arrêt de travail. Cependant, il ne faut pas négliger l'aspect psychologique jouant fortement sur le niveau d'absentéisme. En effet les suspicions de cas d'infections, de cas contact, les restrictions au travers des confinements et couvre-feux amenant à un isolement des individus peuvent également avoir des impacts dramatiques.

Pour preuve, l'étude menée par EPI-PHARE a pu mettre en lumière le lien entre la délivrance d'ordonnances de médicaments anxiolytiques et les périodes de confinement. Ainsi, comme le montre le graphique suivant, durant la période de confinement de 2020, le nombre de

délivrances d'anxiolytiques a fortement augmenté. Il est également possible de voir que ce phénomène s'est prolongé au-delà de cette période, le nombre de délivrances observées étant supérieur aux estimations attendues en 2020.

Figure 68 : Lien entre le nombre de délivrances d'anxiolytiques et la période de confinement



Nombre par quinzaine des délivrances sur ordonnance de médicaments anxiolytiques

Ce nouveau contexte impacte donc la santé des individus, un facteur majeur dans l'augmentation de l'absentéisme. A partir du graphique précédent, il est aisé de comprendre d'une part que le contexte actuel est bien différent de celui avant 2019, mais également qu'en raison des variants, des nouvelles restrictions mises en place ou levées en fonction de l'avancée du virus, ce contexte change rapidement d'une période à l'autre.

De ce fait, l'étude de l'absentéisme doit s'adapter à ces nouveaux changements. Une étude ou une modélisation dite statique comme montrée dans la deuxième partie de ce mémoire n'est donc plus un choix judicieux. Une analyse ou une modélisation plus dynamique au travers par exemple d'un taux d'absentéisme comme présentée dans la troisième partie de ce mémoire permet de prendre en compte des changements de régimes à l'aide de variables explicatives évoluant au cours du temps. Cette nouvelle modélisation par séries temporelles pourra ainsi permettre de prendre par exemple en compte les périodes de confinement, le taux de vaccination au cours du temps ainsi que l'apparition de nouveaux variants en plus des variables déjà présentées dans ce mémoire.

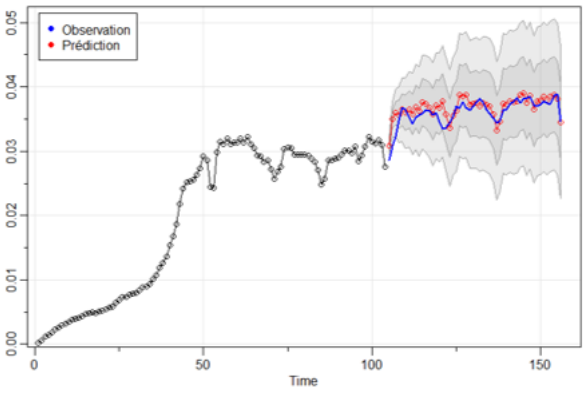
Néanmoins, ces périodes de transitions entre restrictions renforcées, levée de ces dernières et l'apparition de nouveaux variants ne permet pas à priori d'avoir des modèles stables au cours du temps. Ainsi, la présentation dans la partie précédente d'une méthode de détection de rupture de stabilité des modèles et de détection des dates de rupture, pourraient permettre d'identifier des points de rupture en lien avec des fluctuations brusques de certaines variables. Des périodes correspondant à des configurations quasiment similaires pourraient ainsi être étudiées ou comparées.

Ces pistes de réflexions pourraient emmener à de nouvelles études prenant en compte ce nouveau contexte venant choqué les anciennes modélisations d'avant 2019. De plus, l'intégration

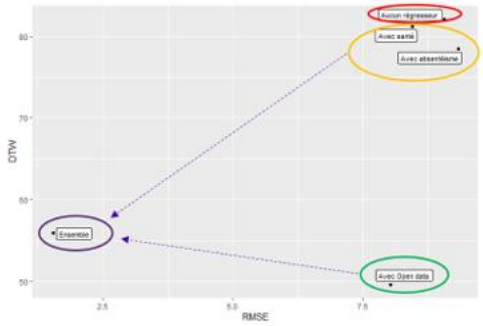
de nouvelles variables et de nouveaux liens de causes à effet entre la consommation santé avec le contexte actuel pourraient être davantage investigués afin de mieux comprendre l'évolution de l'absentéisme durant ces périodes de restrictions et de doutes pesant sur la population, et en particulier sur les salariés d'une entreprise.

Synthèse troisième partie : Risque absentéisme et modèle temporel dynamique

L'introduction ainsi que l'utilisation des séries temporelles permet dans cette partie d'étudier et de modéliser le risque absentéisme sous un nouvel angle. Comme vu lors de la seconde partie de ce mémoire sur les modélisations par Random Forest, les prédictions du taux d'absentéisme d'une période sur l'autre n'étaient pas satisfaisantes. En cause, la temporalité et l'évolution des variables dans la modélisation n'étaient pas prises en compte.

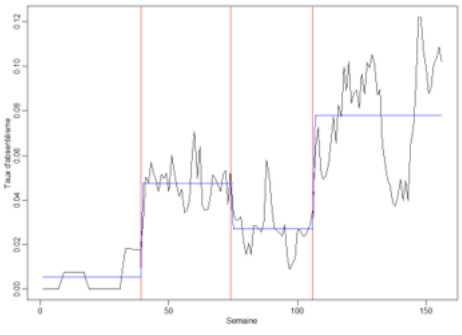


L'objectif principal consiste à étudier les fluctuations de l'indicateur du taux d'absentéisme au cours du temps en travaillant sur des modélisations utilisant la **théorie des séries temporelles**. Par l'ajout de variables explicatives appelées **régresseurs** venant moduler la prédiction, celles-ci apportent de l'information supplémentaire sur une possible augmentation ou diminution du risque. L'ajout de ces données répond à un double enjeu à savoir **corriger les prédictions de l'absentéisme** à l'aide de données corrélées à cette variable cible, mais également **donner de l'explicabilité aux prédictions effectuées.**



Afin de mesurer la pertinence des modèles testés ainsi que l'impact de chaque régresseur, deux métriques sont alors construites, à savoir **le RMSE ainsi que le Dynamic Time Warping (DTW)**, celle-ci apportant de l'information sur l'adéquation de la forme de la série. A l'aide de ces deux métriques, la recherche de la meilleure modélisation passe par le retrait ou l'ajout de certains régresseurs. Le meilleur modèle est alors obtenu par optimisation d'un problème de **front de Pareto**, où l'on cherche à minimiser à la fois les deux métriques.

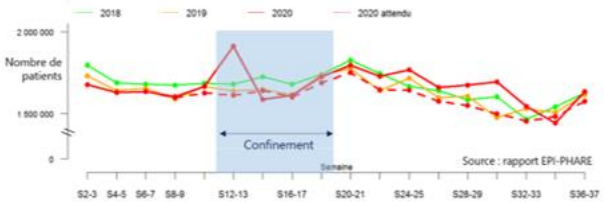
Cependant ces nouvelles modélisations et projections du taux d'absentéisme sur des fenêtres de prédictions différentes ne sont fiables que sur des séries elles-mêmes **stables au cours du temps**. La recherche de séries stables de l'absentéisme peut passer par la reprise des résultats de **la segmentation des profils d'absentéisme.**



En revanche sur des séries ayant de trop grandes fluctuations au cours du temps, il peut être intéressant

d'étudier la série en cherchant à calculer **les points de rupture** dans la modélisation et les dates auxquelles ces ruptures se produisent. Elles peuvent être **significatives d'un changement de régime soit de la variable cible soit des variables explicatives**. Ces points de rupture permettent alors d'avertir **une potentielle augmentation ou diminution brusque du risque**.

En ces temps de pandémie, où l'absentéisme peut devenir un risque instable et plus difficile à modéliser, cette modélisation par série temporelle mais également la recherche de points de ruptures peut aider à la détection de changements de régimes et à réfléchir sur les causes de tels changements, comme la levée de certaines restrictions ou encore l'arrivée de nouveaux variants.



Conclusion

Depuis ces dernières années, **le risque absentéisme a connu une augmentation significative** entraînant par la même occasion des coûts non négligeables pour les entreprises et les organismes couvrant ce risque. Le suivi de l'absentéisme auprès des entreprises est par conséquent devenu un sujet incontournable.

Le nouvel apport des données DSN (déclaration sociale nominative) transmises mensuellement permet aujourd'hui aux assureurs d'obtenir davantage d'informations sur les arrêts de travail. Auparavant les données relatives aux arrêts de travail n'étaient connues qu'après la période de franchise dépassée. Aujourd'hui avec ces nouvelles informations, l'assureur peut avoir une vision complète de l'ensemble des arrêts, du début jusqu'à leur fin, et par conséquent peut **apporter une véritable expertise sur le niveau et le suivi de l'absentéisme d'une entreprise.**

Au travers de ces travaux menés sur des données provenant des données DSN d'une entreprise du secteur de l'industrie et réexploitées par une autre entité, l'étude de l'absentéisme a été réalisée sous différents angles. Après une première étude statistique permettant de montrer que le niveau d'absentéisme touche différemment les individus en fonction de leurs caractéristiques, la mise en place de **modèles de Machine Learning et de Forecasting permet d'approfondir cette étude.**

Une première partie consistant à **segmenter le portefeuille à l'aide des arbres CART en fonction de variables explicatives permet dans un premier temps de classer les individus en groupes relativement homogène d'absentéisme afin d'identifier des profils d'absentéisme sur lesquels des actions de prévention peuvent être envisagées**. L'étude des caractéristiques de ces profils d'intensité faible, modéré ou fort d'absentéisme montre qu'ils sont fonction de l'âge de l'ancienneté, mais également des absences passées du salarié.

Ces observations ont été confirmées par **les modèles de Random Forest, plus robustes que les arbres CART en termes de modélisation. Les variables d'importance ont ainsi pu montrer le lien étroit entre l'absentéisme mesuré à un instant t et les absences passées du salarié**. Par la suite, la mise en place d'une modélisation du taux d'absentéisme en fonction des caractéristiques des individus, de leur consommation santé ainsi que de leurs absences passées a permis de donner une prédiction du taux d'absentéisme sur le groupe de salariés après une agrégation des résultats individuels de prédiction. Cependant, les résultats obtenus montraient une sous-estimation des taux d'absentéisme élevés observés ainsi que d'une surestimation des taux d'absentéisme faibles observés. De plus, la mise à jour du modèle, et la prédiction des taux d'absentéisme sur des périodes successives semblent de mauvaise qualité en raison du **caractère statique de cette modélisation par Random Forest qui ne prend pas assez en compte la temporalité des données.**

Pour remédier à ce problème, **une modélisation à l'aide des séries temporelles (modèle ARIMAX) a été mise en place afin de se concentrer sur les fluctuations du taux**

d'absentéisme au cours du temps. L'ajout de variables explicatives dans la modélisation nommées régresseurs, a permis de moduler les prédictions effectuées au cours du temps afin d'obtenir des prévisions proches des données observées. Ce nouveau type de modélisation permet une mise à jour plus aisée des prédictions au cours du temps en fonction des données récoltées jour après jour, montrant son apport plus dynamique comparée à une modélisation plus statique par Random Forest par exemple.

Une approche empirique de la contribution de chaque régresseur sur la prédiction a été effectuée grâce à la construction de deux métriques, à savoir le RMSE et le Dynamic Time Warping (DTW), permettant de se placer dans un problème d'optimisation d'un front de Pareto en cherchant à minimiser les deux métriques. Par ajout ou retrait des différents régresseurs, il est alors possible d'identifier les variables apportant une modulation plus juste sur les prédictions effectuées.

Néanmoins, ces modélisations du taux d'absentéisme par série temporelles utilisant des régresseurs comportent des faiblesses. **La grande corrélation existante entre l'absentéisme et les données de santé par exemple ne permet pas d'obtenir des contributions comme pourraient le proposer les modèles GLM. De plus, ces modélisations sont efficaces uniquement sur des données assez stables au cours du temps.**

Pour contourner ce dernier aspect, il est possible **d'étudier les points de rupture de la modélisation au cours du temps à l'aide de la statistique CUSUM. Cette dernière permet d'avertir le modélisateur d'un changement de régime sur la série temporelle de la variable cible ou sur celles des variables explicatives.** Le calcul des dates de rupture peut ainsi permettre d'identifier des périodes de temps correspondantes à l'arrivée d'une tierce cause non encore introduite dans la modélisation.

En ces temps d'instabilité liés à l'apparition de la pandémie de COVID-19, les causes de l'absentéisme deviennent plus nombreuses. En cause, l'état psychologique des individus, les diverses restrictions gouvernementales ou encore les conditions de travail dégradées peuvent renforcer le risque d'absentéisme. L'ajout de nouvelles données Open Data telles que les périodes de confinement ou de couvre-feu, l'avancée de la vaccination ou encore l'arrivée de nouveaux variants couplé à une étude de la consommation santé des individus pendant cette période pourraient aboutir à une meilleure modélisation du risque absentéisme durant ces périodes d'instabilité.

Les enjeux de la prédiction de l'absentéisme sont multiples :

- **Pour l'entreprise : maîtrise des arrêts, de sa productivité, de son organisation**
- **Pour l'assureur : maîtrise de la sinistralité et de son S/P en arrêt de travail**

Les travaux de ce mémoire ont permis de créer des indicateurs de profil d'absentéisme et de prédiction de l'absentéisme utilisables à la fois pour les assureurs et les entreprises. Ces problématiques utilisant des données mises à jour régulièrement (données DSN, consommation

santé, Open Data), il faut aussi s'interroger sur l'industrialisation du processus et l'adaptation de la modélisation aux différentes entreprises étudiées à l'avenir.

Bibliographie

[1] Anact-Aract. (2015). *Dix questions sur l'absentéisme*.

[2] Ayming. (2019). *11ème Baromètre de l'Absentéisme et de l'Engagement Ayming*. Etude quantitative.

[3] Ayming, (2020). *12ème Baromètre de l'Absentéisme et de l'Engagement Ayming*. Etude quantitative.

[4] Statista, (2019). « https://fr.statista.com/infographie/11000/absenteisme-au-travail-en-france/ ».

[5] Institut Sapiens, (2018) Laurent Cappelletti & Henri Savall, « Le coût caché de l'absentéisme au travail : 108 milliards € », « https://www.institutsapiens.fr/le-cout-cache-de-labsenteisme-au-travail-108-milliards-e-2/ »

[6] Cour des comptes (2019), « Chapitre 3, les indemnités journalières : des dépenses croissantes pour le risque maladie, une nécessaire maîtrise des arrêts de travail », « https://www.ccomptes.fr/system/files/2019-10/RALFSS-2019-03-indemnites-journalieres.pdf »

[7] Genuer, R., Poggi, J.-M., & Tuleau, C. (2008). *Random Forests: some methodological insights.* INRIA, Institut de Recherche en Informatique et en Automatique.

[8] Bouville, G. (2014). *La progression de l'absentéisme : nouveaux comportements des salariés ou nouvelles contraintes organisationnelles ?*

[9] Institut des Actuaires. (2017). *Norme de Pratique relative à l'utilisation et la protection des données massives, des données personnelles et des données de santé à caractère personnel-NPA 5.*

[10] Larousse, Récupéré sur « https://www.larousse.fr/dictionnaires/francais/absent%C3%A9isme/261 »

[11] Sécurité Sociale. (2020). Récupéré sur « https://www.securite-sociale.fr/la-secu-cest-quoi/organisation/les-branches »

[12] Toni, G. (2009). *Computing and Visualizing Dyamic Time Warping Alignments in R : The dtw Package*. Journal of Statistical Software.

[13] EPI-PHARE (2021), "Covid-19 : usage des médicaments de ville en France », « Rapport 6 : Point de situation au 25 avril 2021 »

[14] Santé Publique France, https://geodes.santepubliquefrance.fr/#view=map2&c=indicator

[15] Hala Najmeddine, Frédéric Suard, Arnaud Jay, Philippe Marechal, Marié Sylvain. Mesures de similarité pour l'aide à l'analyse des données énergétiques de bâtiments. RFIA 2012 (Reconnaissance des Formes et Intelligence Artificielle), Jan 2012, Lyon, France. pp.978-2-9539515-2-3. hal-00656546

[16] Zeileis, A.,Leisch, F.,Hornik, K.andKleiber, C.(2002).strucchange: An R package for testing for structural change in linear regression models .Journal of Statistical Software,71–38. URL http://www.jstatsoft.org/v07/i02/.

[17] Bruce H. Andrews, Matthew D. Dean, Robert Swain, Caroline Cole, University of Southern Maine (2013), "Building ARIMA and ARIMAX Models for Predicting Long-Term Disability Benefit Application Rates in the Public/ Private Sectors" Sponsored by Society of Actuaries.

[18] Yves Aragon, "Séries temporelles avec R, Méthodes et cas", Livre, Edition Springer

[19] CNIL. (s.d.). *RGPD : par où commencer*. Récupéré sur https://www.cnil.fr

ANNEXES

Annexe 1 : Les distances de similarités entre séries temporelles

Le principe de la distance DTW a déjà été présenté au cours de ce mémoire. Cependant, la réelle mesure appliquée dans les calculs présentés, est basée sur d'autres variantes, plus efficaces et plus robustes dans le calcul de la similarité existante entre deux séries temporelles.

La « *Derivative Dynamic Time Warping* » (DDTW) prend davantage en compte la forme des séries temporelles au travers de la dérivée première. En effet, si dans le cas de la DTW le terme de distance utilisé dans le calcul était la distance euclidienne, ici c'est l'estimation de la dérivée qui est préférée :

$$d_x(v_i) = \frac{(v_i - v_{i-1}) + (v_{i+1} - v_{i-1})/2}{2}$$

Néanmoins, étudier uniquement les valeurs de la série ou les dérivés de celle-ci est encore trop réducteur, puisque l'aspect global de la série n'est pas pris en compte. La nouvelle mesure « *Adaptive Feature Based Dynamic Time Warping* » (AFBDTW), qui est celle utilisée dans le mémoire, calcule à la fois l'aspect local ainsi que global de la série temporelle.

Deux nouvelles fonctions sont alors ajoutées pour prendre les deux aspects en compte :

$$f_{local}(i) = (u_i - u_{i-1}, u_i - u_{i+1})$$

$$f_{global}(i) = \left(u_i - \frac{1}{i-1}\sum_{k=1}^{i-1} u_k, u_i - \frac{1}{m-i}\sum_{k=i+1}^{m} u_k \right)$$

La distance entre u_i et v_j est alors définie par :

$$dist(u_i, v_j) = W_1 . dist_{local}(u_i, v_j) + W_2 . dist_{global}(u_i, v_j)$$

Avec

$$dist_{local}(u_i, v_j) = \left|(f_{local}(u_i))_1 - (f_{local}(v_j))_1\right| + \left|(f_{local}(u_i))_2 - (f_{local}(v_j))_2\right|$$

$$dist_{global}(u_i, v_j) = \left|(f_{global}(u_i))_1 - (f_{global}(v_j))_1\right| + \left|(f_{global}(u_i))_2 - (f_{global}(v_j))_2\right|$$

$$W_1 + W_2 = 1{,}0 \leq W_1 \leq 1{,}0 \leq W_2 \leq 1$$

Cette nouvelle mesure est alors plus adaptée dans le calcul de la similarité entre séries temporelles prenant davantage en compte des effets de décalages temporels. C'est cette même mesure qui est alors utilisée pour ce mémoire, avec une répartition des poids entre local et global de 50%. La distance d_{cum} reste toutefois la même que présentée dans le corps du texte.

Annexe 2 : Détection de rupture de modèle

La détection de rupture au cours du temps sur les modèles linéaires présentée dans le corps du mémoire se base sur l'étude des résidus du modèle. Pour rappel, le modèle linéaire s'écrit comme :

$$y_i = x_i^T \beta_i + u_i, \qquad (i = 1, \dots, n)$$

Où pour chaque temps i, y_i est l'observation de la variable cible en fonction des observations des variables explicatives x_i et des coefficients de régression β_i.

On note pour la suite $\hat{\beta}^{(i)}$ l'estimation des coefficients de régression par la méthode des moindres carrés ordinaire (MCO ou OLS) basée sur les observations jusqu'au temps i. Ainsi, $\hat{\beta}^{(n)}$ correspond à l'estimation commune des coefficients de régression du modèle linéaire sur l'ensemble des données disponibles. De même, on peut noter $X^{(i)}$ la matrice de régression des observations jusqu'au temps i.

A partir du modèle précédent, les résidus par MCO sont les suivants :

$$\hat{u}_i = y_i - x_i^T \hat{\beta}^{(n)}$$

Cependant dans la recherche de rupture de structure pas à pas, il est également possible de construire les résidus dits récursifs suivants :

$$\tilde{u}_i = \frac{y_i - x_i^T \hat{\beta}^{(i-1)}}{\sqrt{1 + x_i^T \left(X^{(i-1)^T} X^{(i-1)} \right)^{-1} x_i}} \qquad (i = k+1, \dots, n)$$

Ces résidus construits sont de moyenne nulle et de variance σ^2 sous l'hypothèse nulle d'absence de changement de structure. L'estimation empirique de cette variance est donnée par :

$$\tilde{\sigma}^2 = \frac{1}{n-k} \sum_{i=k+1}^{n} (\tilde{u}_i - \bar{u})^2$$

Le processus de détection des points de rupture dans le modèle peut alors être effectué à l'aide de la statistique CUSUM. On considère alors la somme cumulée des résidus récursifs standardisés :

$$W_n(t) = \frac{1}{\tilde{\sigma}\sqrt{\delta}} \sum_{i=k+1}^{k+\lfloor t\delta \rfloor} \tilde{u}_i \qquad (0 \leq t \leq 1)$$

Où $\delta = n - k$ et $\lfloor t\delta \rfloor$ est la partie entière de $t\delta$. Sous l'hypothèse nulle d'une structure stable dans le modèle, le processus limite des fluctuations empiriques du processus $W_n(t)$ est un mouvement Brownien standard $W(t)$. Dans le cas où l'hypothèse nulle est rejetée, cela signifie qu'au moins un temps t_0 correspond à un changement de structure. Ainsi, les résidus récursifs ont une moyenne nulle jusqu'au temps t_0 puis cette moyenne dévie de la moyenne nulle.

Une autre alternative possible est d'évaluer les résidus sur une fenêtre de temps dynamique, et donc de ne plus considérer l'ensemble des résidus jusqu'à un instant t. Ainsi, pour une fenêtre de temps de durée h, la statistique MOSUM est construite comme :

$$M_n(t|h) = \frac{1}{\tilde{\sigma}\sqrt{\delta}} \sum_{i=k+\lfloor N_\delta t \rfloor+1}^{k+\lfloor N_\delta t \rfloor+\lfloor \delta h \rfloor} \tilde{u}_i = W_n\left(\frac{\lfloor N_\delta t \rfloor + \lfloor \delta h \rfloor}{\delta}\right) - W_n\left(\frac{\lfloor N_\delta t \rfloor}{\delta}\right) \qquad (0 \le t \le 1-h)$$

Où $N_\delta = (\delta - \lfloor \delta h \rfloor)/(1-h)$.

Comme pour la statistique CUSUM, le processus limite est un mouvement Brownien. Pour tout changement de structure dans le modèle, le processus déviera alors de la moyenne nulle.