Mémoire présenté le :
pour l'obtention du diplôme
de Statisticien Mention Actuariat
et l'admission à l'Institut des Actuaires

Par : Monsieur Guillaume STASINSKI

Impact des mesures prises contre le Covid-19 sur les dépenses santé en France métropolitaine en 2020 à l'aide de la base Open DAMIR

Confidentialité : NON

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

Membres présents du jury de la filière : Signature :	Entreprise Nom : Galea & Associés Signature : GALEA & Associés 25 rue de Choiseul 75002 PARIS Tél. 01 43 22 11 11 R.C.S. Paris - 492 379 839
	Directeur de mémoire en entreprise Nom : Sergio OROZCO Signature :
Membres présents du jury de l'Institut des Actuaires : 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 :

Résumé

Dans un contexte sanitaire marqué par une pandémie d'ampleur et de conséquences inédites, l'économie française s'est retrouvée mise sous quarantaine. Le secteur de la santé a particulièrement été touché par cette crise : des cabinets de professionnels de santé ont momentanément fermé et des opérations hospitalières ont dû être reportées.

L'objectif de ce mémoire est d'étudier l'impact des mesures prises contre le Covid-19 sur les dépenses santé en France métropolitaine en 2020 et de projeter les dépenses attendues sur l'année 2021. Cette étude portera sur les cinq postes de consommation suivants : *soins de ville courants*, *pharmacie*, *dentaire*, *optique* et *prothèses auditives*.

L'utilisation de la base Open DAMIR de 2015 à 2020 fait partie intégrante de ce mémoire. Cet historique conséquent permet d'analyser et de mieux comprendre les impacts des mesures prises contre le Covid-19. L'utilisation d'une telle base de données permet également d'avoir la vision de l'Assurance Maladie sur la dépense et non une vision assurantielle qui se serait avérée être trop spécifique pour cette étude.

L'utilisation de la puissance du langage de programmation Python et de ces nombreuses librairies permet de rendre cette étude facilement reproductible à l'avenir.

Finalement, afin d'appréhender au mieux l'impact du Covid-19 sur la dépense connue en 2020 et attendue sur 2021, l'utilisation de plusieurs techniques d'analyse et de prédiction de séries temporelles sont utilisées dans ce mémoire. De plus, l'introduction de ***Prophet*** - nouvelle technique de prédiction de séries temporelles développée par Facebook – permet de mieux capturer les changements de tendance engendrés par un tel évènement.

Mots clés *: Covid-19, Assurance Maladie, Open DAMIR, séries temporelles, Python, Prophet, Dask*

Summary

In a health context marked by a pandemic of unprecedented scope and consequences, the French economy has been placed under quarantine. The health sector was particularly affected by this crisis: health professionals' offices were temporarily closed, and hospital operations had to be postponed.

The objective of this paper is to study the impact of the measures taken against Covid-19 on healthcare spending in metropolitan France in 2020 and to project the expected spending for the year 2021. This study will focus on the following five consumption items: routine city care, pharmacy, dental, optical and hearing aids.

The use of the Open DAMIR database from 2015 to 2020 is an integral part of this study. This consistent history allows us to analyze and better understand the impacts of the measures taken against Covid-19. The use of such a database also allows us to have the Medicare vision of the expenditure and not an insurance vision which could be too specific in our case.

The use of the power of the Python programming language and its numerous libraries makes this study easily reproducible in the future.

Finally, to better understand the impact of the Covid-19 on the spending known in 2020 and expected in 2021, the use of several analysis and prediction techniques of time series are used in this thesis. In addition, the introduction of ***Prophet*** - a new time series prediction technique developed by Facebook - allows us to better capture the changes in trend caused by such an event.

Keywords: Covid-19, Health Insurance, Open DAMIR, time series, Python, Prophet, Dask

Note de synthèse

Introduction

L'année 2020 restera dans la mémoire générale comme l'année ayant vu l'émergence d'une nouvelle maladie – le Covid-19 – dont l'entièreté des répercussions mettra du temps à être totalement appréhendée et mesurée. La rapide propagation de la maladie à travers le globe – rapidement qualifiée de pandémie par l'Organisation Mondiale de la Santé (OMS) – a pris de très nombreux États de court : des décisions radicales ont été par conséquent prises. C'est notamment le cas de la France qui a imposé un premier confinement de sa population à l'échelle nationale le 17 mars 2020.

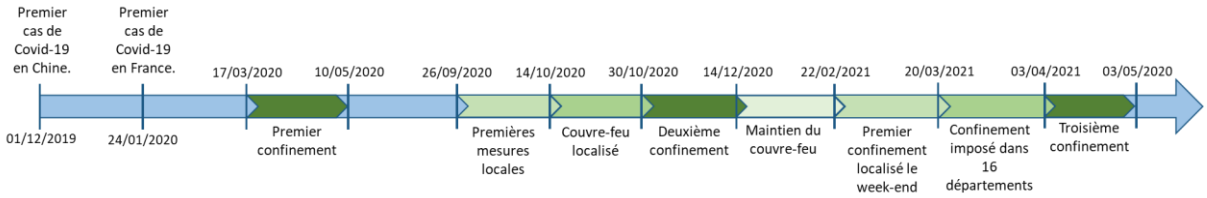
Cet arrêt brutal de l'économie et de la « vie d'avant » a touché de très nombreux secteurs, y compris celui de la santé. En effet, en plus des images montrant les hôpitaux français saturés dans certaines régions obligeant alors le transfert de patients entre régions et à l'étranger, de nombreux soins ont dû être reportés ou annulés.

Ce mémoire consiste à étudier l'impact des mesures prises contre le Covid-19 sur les dépenses santé en France métropolitaine en 2020 et à effectuer une projection des dépenses sur l'année 2021. Nous concentrons nos travaux sur les cinq postes de consommation suivants : *soins de ville courants*, *pharmacie*, *dentaire*, *optique* et *prothèses auditives*.

Tout d'abord, nous nous attardons plus en détail sur le contexte particulier de l'année 2020 et sur les indicateurs mis en place afin de mesurer la circulation de la pandémie dans le pays. Ensuite, nous nous intéressons à la base Open DAMIR ainsi qu'aux techniques et retraitements mis en place pour permettre sa bonne utilisation. Puis, nous explicitons les deux modèles de séries temporelles traditionnellement utilisés et projetons les dépenses attendues en 2020 à partir des données d'apprentissage de 2015 à 2019. Enfin, nous introduisons un nouveau modèle d'analyse et de prédiction de séries temporelles complété par l'introduction d'un régresseur central. Nous évaluons la pertinence de ce nouveau modèle en projetant les dépenses sur 2020. Pour finir, nous projetons les dépenses sur 2021 et concluons en conséquence.

2020 : une année sous tension

Le Covid-19 aura profondément impacté la société française. Bien que toutes les mesures aient été prises dans le but de ralentir la propagation et de protéger au mieux la population, celles-ci ont eu pour effet de mettre un cran d'arrêt à l'économie et à de nombreux pans de la vie.



Chronologie des mesures prises dans le cadre de la gestion du Covid-19 en France du début de la pandémie à la fin du troisième confinement

Afin de mesurer l'évolution de la pandémie en France, de nombreux indicateurs de suivi permettant de suivre de façon quasi journalière la circulation virale du virus, la tension hospitalière ainsi que la mortalité ont été mis en place. L'indicateur que nous avons retenu dans ce mémoire est celui de la tension en réanimation et non un indicateur lié à la circulation virale du virus. Cet indicateur se calcule de la manière suivante :

$$\frac{Nombre\ de\ places\ occupées\ en\ réanimation\ pour\ cause\ de\ Covid}{Nombre\ de\ lits\ disponibles\ en\ réanimation}$$

Malgré son intérêt manifeste, cet indicateur souffre d'imperfection. Tout d'abord, le nombre de lits disponibles ne prend pas en compte la création de nouvelles places, ce qui a été le cas au plus fort de la crise du Covid. De plus, au numérateur, il est également compris les malades qui sont admis en unité de soins intensifs et en unité de surveillance continue. Ainsi, cet indicateur a tendance à surestimer la situation réelle en réanimation.

Cette fermeture administrative de l'économie aura également eu un impact indéniable sur la consommation santé des français. De nombreux actes ont dû être reportés entraînant alors un retard des traitements. En France, d'après une étude réalisée par la Fédération Hospitalière de France – la FHF, porte-voix des hôpitaux publics – sur l'année 2020, 2,3 millions de séjours prévus n'ont pas pu être honorés dont 1,4 million de séjours en médecine et 900 000 séjours de chirurgie. Comparé à 2019, cela représente une baisse de 12% et de 15% respectivement. Ces chiffres sont particulièrement édifiants lorsqu'ils sont ramenés à la période du premier confinement. Nous pouvons voir que :

- La chirurgie en hospitalisation a baissé de 58 % ;
- La chirurgie en ambulatoire a décru de 80 % ;
- Les coloscopies de diagnostic se sont effondrées de 87 % ;
- Les transplantations rénales ont chuté de 80 % ;

Ces chiffres s'entendent bien évidemment comparés à 2019 sur une période équivalente et promettent potentiellement une bombe à retardement qui pourra avoir plusieurs impacts pour l'avenir : une mortalité accrue et une hausse du coût des traitements pour prise en charge tardive pouvant alors entraîner une hausse de la dépendance.

Afin d'étudier plus en détails l'impact du Covid-19 sur la dépense Santé en France, nous exploitons la base Open DAMIR.

Source des données utilisées : la base Open DAMIR

La base Open DAMIR (Dépenses de l'Assurance Maladie Inter Régimes) est une extraction du SNIIRAM (Système Nationale d'Information Inter-régimes). Cette base de données est une extraction anonymisée de l'ensemble des remboursements effectués par l'Assurance Maladie, à l'exception d'une grande partie des dépenses hospitalières. Les données de la base sont disponibles au format CSV et regroupées selon les mois de règlement. Afin d'obtenir les données des mois de règlement de l'année N, il faut attendre la mise à jour annuelle de la base, qui s'effectue généralement lors du second trimestre de l'année $N+1$. Le choix a été fait de récupérer les années de règlement de 2015 à 2020. De cette base de données, nous avons fait le choix de retenir les variables présentées dans le tableau ci-dessous.

Variable de la base	Explication de la variable
AGE_BEN_SNDS	Tranche d'âge du bénéficiaire au moment des soins
BEN_RES_REG	Région de résidence du bénéficiaire
BEN_SEX_COD	Sexe du bénéficiaire
PRS_NAT	Nature de la prestation
FLX_ANN_MOI	Année et mois de règlement
SOI_ANN	Année de soins
SOI_MOI	Mois de soins
PRS_PAI_MNT	Montant de la dépense

Trois variables supplémentaires ont également été construites : *Poste*, *REG_ANN*, *REG_MOI*. La variable *Poste* découle de la variable *PRS_NAT*. Un travail d'identification du type d'acte et du poste de consommation a été effectué sur chacune des prestations. Le choix a été fait de ne conserver que les postes de consommations : ***soins de ville courants***, ***pharmacie***, ***optique***, ***dentaire*** et ***prothèses auditives***. De plus, sur un faible nombre de lignes, l'âge, le sexe et la région sont indiqués comme étant « *inconnu* ». Leur poids étant considéré négligeable en termes de prestations, ces lignes ont été supprimées et ne sont donc pas considérées dans l'étude. De même, l'étude porte uniquement sur la France métropolitaine. Les lignes avec la modalité « *Régions et Départements d'outre-mer* » ont été retirées.

De par la taille des bases Open DAMIR (un fichier mensuel a un poids 5 giga-octets), il est impensable de vouloir utiliser des méthodes et outils traditionnels pour leur traitement. Ainsi, afin d'extraire les données qui nous intéressent, nous avons utilisé le langage de programmation *Python* ainsi que la librairie *Dask*. Cette librairie a été pensée pour la manipulation des bases de données semblables à celle de l'Open DAMIR. En parallèle et afin d'accélérer les temps d'extractions et de manipulation, nous avons fait appel aux serveurs de Google à travers l'utilisation du service *Google Colaboratory*.

A partir d'une première étude sur les cadences de règlement de 2015 à 2019, nous avons observé qu'au bout de trois et douze mois après la date de survenance, nous connaissons plus de 95 % et de 99 % de la dépense à l'ultime respectivement. Nous considérons alors que les survenances de 2015 à 2019 sont développées à l'ultime en provenance des données Open DAMIR.

Cependant pour l'année de soins 2020, nous avons remarqué que la dépense n'était pas complète, notamment pour les mois de novembre et décembre, ayant uniquement deux et un mois de développement. Il a donc été nécessaire d'estimer la dépense à l'ultime pour cette année de soins.

À l'aide d'un test rétroactif de validité, nous faisons le choix de considérer que les coefficients de développement sur l'année 2020 sont similaires à ceux observés sur l'année 2019. Nous avons donc décidé d'effectuer une liquidation par méthode de Chain Ladder sur les cinq triangles par poste de consommation de l'année de soins 2020 à horizon de douze mois, afin d'obtenir des prestations à l'ultime. Nous obtenons ainsi une vision au 31/12/2021 (à l'ultime) de l'année de soins 2020.

Les analyses que nous effectuons par la suite sont donc effectuées à partir d'une base de données reconstruite comme suit :

- Une vision réelle (en provenance des bases Open DAMIR) au 31/12/2020 pour les années de survenance de 2015 à 2019 ;
- Une vision estimée (via Chain Ladder) au 31/12/2021 de la dépense pour l'année de survenance de 2020.

Projection de la dépense sur l'année 2020 et comparaison avec le réel

Etant données la structure et la disponibilité des données Open DAMIR, nous avons intuitivement décidé d'utiliser le concept des séries temporelles. A partir de cette première idée, nous cherchons à déterminer s'il existe un modèle de séries temporelles sous-jacent à nos données sur la période avant 2020. Nous cherchons notamment à identifier toute tendance ou saisonnalité dans les séries temporelles étudiées.

Pour ce faire, nous faisons appel aux deux modèles d'analyse et de prédiction de séries temporelles les plus connus : $ARIMA$ (*Autoregressive Integrated Moving Average*) et $SARIMA$ (Seasonal

Autoregressive Integrated Moving Average). Nous rappelons les équations de ces deux modèles ci-dessous ($\forall t \geq 0$) :

- Le modèle $ARIMA$: $\phi_p(L)(1-L)^d y_t = \mu + \theta_q(L)\epsilon_t$
- Le modèle $SARIMA$: $\phi_p(L)(1-L)^d * \Phi_P(L^m)(1-L^m)^D * Y_t = \mu + \theta_q(L) * \Theta_Q(L^m) * \epsilon_t$

Où :

- ϵ_t, un bruit blanc faible. Il s'agit d'une suite de variables aléatoires réelles identiquement distribuées non corrélées entre eux. ($\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$) ;
- m est l'opérateur de saisonnalité ($m = 12$ pour des données mensuelles) ;
- $\phi_p(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$;
- $\theta_q(L) = 1 + \theta_1 L - \cdots + \theta_q L^q$;
- Les polynômes saisonniers $\Phi_P(L^m)$ et $\Theta_Q(L^m)$ sont définis comme suit :
  - $\Phi_P(L^m) = 1 - \Phi_1 L^m - \cdots - \phi_P L^{Pm}$;
  - $\Theta_Q(L^m) = 1 + \Theta_1 L^m - \cdots + \Theta_Q L^{Qm}$.
- Toutes les racines de l'équation caractéristique associée à ϕ, θ, Φ et Θ sont de module inférieur à 1 ;

Nous résumons, dans le tableau ci-dessous, les modèles ARIMA et SARIMA retenus après analyses sur nos cinq postes de consommation.

Poste de consommation	Modèle $ARIMA$ retenu	Modèle $SARIMA$ retenu
Soins de ville courants	$ARIMA(11,0,2)$	$SARIMA(0,0,2)(0,1,0)_{12}$
Pharmacie	$ARIMA(11,0,1)$	$SARIMA(2,0,0)(1,1,0)_{12}$
Dentaire	$ARIMA(11,0,1)$	$SARIMA(1,0,1)(0,1,0)_{12}$
Optique	$ARIMA(11,0,0)$	$SARIMA(0,0,1)(3,1,0)_{12}$
Prothèses auditives	$ARIMA(11,0,1)$	$SARIMA(1,0,1)(1,1,0)_{12}$

Vu le nombre important de terme autorégressif sur les modèles $ARIMA$ de nos cinq postes de consommation, nous pouvons facilement en conclure qu'il existe une forte saisonnalité dans la dépense et que l'application du modèle SARIMA était indispensable.

Une fois nos modèles établis, nous pouvons effectuer des projections sur l'année 2020. Il en ressort des conclusions similaires pour nos cinq postes de consommation :

- Sur la période n°1 (pré-confinement), la dépense estimée par nos deux modèles est très proche de la dépense observée ;
- Sur la période n°2 (premier confinement), nos modèles ne captent pas la chute des dépenses. Il était en effet impensable de prédire la survenance d'un évènement exceptionnel tel que celui du Covid-19 à partir de l'historique ;
- Sur la période n°3 (entre deux confinements), la dépense attendue est très proche de la dépense réelle. Nous n'observons pas de rattrapage évident de la consommation et retrouvons des niveaux de consommation équivalents au pré-covid ;
- Sur la période n°4 (deuxième confinement), les prédictions des modèles retenues sont à nouveau éloignées du réel. Il semblerait qu'un rattrapage conséquent est observé sur cette période. Il faut cependant noter que cette période est celle où il y a le plus d'incertitude : en effet sur cette période nous ne connaissons pas l'intégralité de la dépense effective à partir des données de la base Open DAMIR.

Introduction d'un nouveau modèle de séries temporelles : Prophet développé par Facebook

Comme nous avons pu le voir, les modèles traditionnels d'analyse et de prédiction des séries temporelles se sont avérés être incapables de prédire la baisse de la consommation de la deuxième période puis la hausse (rattrapage des dépenses) sur la quatrième période. Ceci était entièrement prévisible dû au caractère atypique de 2020.

Afin d'améliorer nos prédictions et dans l'objectif de projeter les dépenses sur l'année 2021 au mieux possible, nous avons souhaité explorer d'autres approches. Nous avons utilisé ***Prophet,*** un modèle mathématique développé par les ingénieurs de Facebook facilement exploitable sous Python. L'équation du modèle retenu est la suivante :

$$y(t) = g(t) + s(t) + \boldsymbol{r(t)} + \epsilon_t$$

Où :

- $g(t)$ représente la tendance de la série temporelle ($g(t) = (k + a(t)\delta^T)t + (m + a(t)\gamma^T)$) ;
- $s(t)$ représente la saisonnalité de la série temporelle ($s(t) = \sum_{n=1}^{N} \left[a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right)\right]$;
- $r(t)$ représente les régresseurs. C'est-à-dire des données externes à la série temporelle ajoutées par l'utilisateur qui pourraient avoir une influence sur celle-ci avec $r(t) = \sum_{n=1}^{N} \alpha_i * W_{i,t}$;
- ϵ_t représente le terme d'erreur, tout ce qui n'a pas pu être correctement expliqué par un des quatre composantes du modèle. La seule hypothèse sur ce terme d'erreur est qu'il est normalement distribué.

La principale différence avec les modèles cités précédemment est l'introduction des régresseurs. Il s'agit de séries temporelles externes permettant de capter des informations complémentaires à la série étudiée. Notre objectif étant de pouvoir capter les spécificités de l'année 2020 que la seule série des dépenses en santé ne pourrait pas identifier.

Prophet a été conçu avec un objectif en tête : simplifier l'analyse toute en conservant la qualité de prédiction des séries temporelles. Prophet cherche ainsi à s'adresser aux spécialistes d'un domaine (dans notre cas, celui de la santé) et évite ainsi de générer un modèle de type « boite noire » qu'un utilisateur serait incapable de comprendre d'un simple coup d'œil. Prophet se base ainsi sur le concept de la modélisation **« analyst-in-the-loop »**, c'est-à-dire une modélisation combinant l'analyse statistique objective traditionnelle avec une analyse subjective tirant sa force de l'expérience acquise sur le terrain.

Afin d'améliorer nos projections sur 2020 et comme cela a été brièvement introduit lors de la présentation de l'équation, nous allons également introduire un régresseur central $\Re$ construit à partir des régresseurs primaires suivants :

- Indicateur du climat des affaires - Tous secteurs - France métropolitaine ($ConfAffaire$) ;
- Indice de chiffre d'affaires - Ensemble du Commerce ($ChiffreAffCom$) ;
- Taux d'occupation dans l'hôtellerie - France métropolitaine ($OccupHotel$) ;
- Démographie - Nombre de décès - France métropolitaine ($Deces$) ;
- Données hospitalières relatives à l'épidémie de COVID-19 où nous nous intéresserons principalement aux individus en réanimation ($ReaCovid$).

Nos cinq régresseurs ont été normalisés entre 0 et 1 à l'aide de l'équation suivante :

$$r'(t) = \frac{r(t) - min(r)}{max(r) - min(r)}$$

Avec :

- r'(t), représentant le régresseur normalisé en t ;
- $r(t)$, représentant la valeur du régresseur en t ;
- $min(r)$, représentant la valeur minimum de la série chronologique observée ;
- $max(r)$, représentant la valeur maximum de la série chronologique observée.

L'équation du régresseur central définie est la suivante :

$$\Re = \frac{ConfAffaire - Deces + OccupHotel + ChiffreAffCom - ReaCovid}{5}$$

Nous avons ensuite effectué une double application du modèle $Prophet$:

- L'application du modèle sans régresseur pour le poste de consommation Pharmacie ;
- L'application du modèle avec le régresseur central $\Re$ pour les quatre autres postes de consommation.

Nous en avons également profité pour comparer les résultats du modèle Prophet avec ceux obtenus avec le modèle $SARIMA$. Les résultats sont résumés dans le tableau ci-dessous (les valeurs présentées sont en millions d'euros) :

	Dépense au 31/12/2020	**Prédiction SARIMA**	**Prédiction Prophet**	**Erreur SARIMA**	**Erreur Prophet**	**% erreur SARIMA**	**% erreur Prophet**
Soins de ville courants	**51 289**	49 345	51 494	**1 944**	**- 205**	**3,9 %**	**- 0,4 %**
Pharmacie	**35 756**	34 950	35 242	**806**	**515**	**2,3 %**	**1,5 %**
Dentaire	**10 741**	11 423	11 903	**- 682**	**- 1 162**	**- 6,0 %**	**- 9,8 %**
Optique	**6 162**	8 606	6 957	**- 2 444**	**- 794**	**- 28,4 %**	**- 11,4 %**
Prothèses auditives	**2 298**	2 552	2 307	**- 254**	**- 8**	**- 9,9 %**	**- 0,3 %**

Suite à l'application du modèle $Prophet$, nous pouvons tirer nos premières conclusions :

- Le modèle $Prophet$ a été bien plus pertinent pour prédire la consommation que le modèle SARIMA – à l'exception du poste *Dentaire* ;
- L'application de notre régresseur a permis de légèrement gagner en précision sur la période n°2 mais les prédictions ne captent pas entièrement l'effet Covid ;
- La prédiction du modèle $Prophet$ appréciée à l'année s'est révélée être très proche de ce qui était attendu (à l'exception du poste *Dentaire*).

Dans un second temps, nous avons effectué des projections de dépenses sur l'année 2021 à partir des modèles Prophet retenus. Nous avons dû estimer les régresseurs primaires au 31/12/2021 puis nous avons construit notre régresseur central sur le même schéma défini précédemment. Les résultats de nos projections sur l'année 2021 sont présentés ci-dessous (les valeurs présentées sont en millions d'euros) :

Mois de survenance	Soins de ville courants	Pharmacie (sans régresseur)	Dentaire	Optique	Prothèses auditives	Total
Janvier	4 920,56	3 175,46	989,31	563,73	216,87	**9 865,93**
Février	4 353,48	2 751,32	925,00	622,37	194,02	**8 846,20**
Mars	4 935,28	3 108,36	1 109,60	693,09	230,39	**10 076,71**
Avril	4 499,85	2 892,21	967,83	642,10	206,81	**9 208,80**
Mai	4 698,39	2 921,01	1 005,05	612,35	206,27	**9 443,06**
Juin	4 667,74	2 974,66	1 116,30	654,21	217,13	**9 630,04**
Juillet	4 111,45	2 911,81	993,42	632,15	186,84	**8 835,68**
Août	3 568,46	2 717,55	535,60	510,17	131,36	**7 463,14**
Septembre	4 562,96	2 918,97	1 008,47	605,26	192,69	**9 288,35**
Octobre	4 710,93	3 076,35	1 086,33	632,96	218,32	**9 724,89**
Novembre	4 698,35	2 999,01	1 068,46	619,99	225,46	**9 611,27**
Décembre	4 259,96	2 995,08	1 014,00	753,46	223,15	**9 245,66**
Total	**53 987,40**	**35 441,80**	**11 819,36**	**7 541,85**	**2 449,32**	**111 239,73**

Conclusion

Il est indéniable que la pandémie a eu un impact sur la dépense santé en 2020 et que ses conséquences continueront à se faire sentir dans les années à venir. Les modèles que nous avons ainsi pu appliquer nous l'ont parfaitement illustré : une baisse très importante de la consommation pendant les périodes de confinement suivie d'une hausse assez conséquente sur la quatrième période (que nous pouvons attribuer à un rattrapage).

Le modèle $Prophet$ présente réellement l'avantage d'être plus compréhensible que les modèles traditionnels d'analyse et de prédiction de séries temporelles. La facilité avec laquelle nous pouvons l'utiliser le rend ainsi bien plus accessible.

Cependant, même si l'application du modèle $Prophet$ nous a permis de légèrement mieux appréhender la consommation de la deuxième et quatrième période, il ne s'est pas avéré aussi efficace qu'anticipé. Ainsi et comme la base Open DAMIR nous le permet, il serait judiciable d'affiner davantage nos analyses selon la région et la tranche d'âge de l'assuré. Une étude plus poussée sur les régresseurs pourrait également être nécessaire si une telle approche est retenue. Enfin, une étude de *backtesting*, comparant les prédictions retenues et des valeurs réelles constatées pour l'année de survenance 2021 permettrait de valider notre approche.

Executive Summary

Introduction

2020 will be remembered as the year of the emergence of a new disease - Covid-19 - whose full impact will take time to be fully understood and measured. The rapid spread of the disease across the globe - quickly qualified as a pandemic by the World Health Organization (WHO) - took many states by surprise: radical decisions were therefore taken. This is notably the case of France, which imposed a first nationwide containment of its population on March 17, 2020.

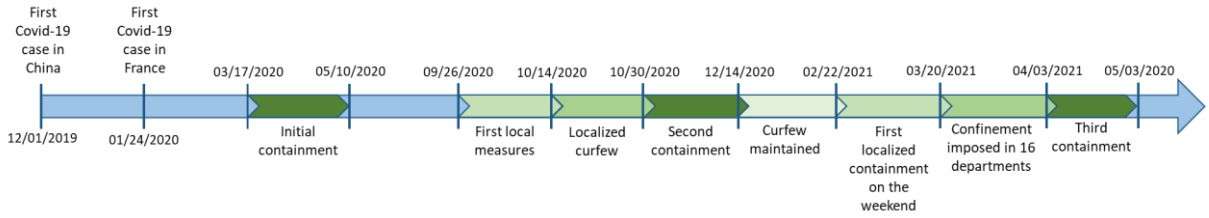
This sudden stop of the economy and of the "life before" has affected many sectors, including the health sector. Indeed, in addition to the images showing French hospitals saturated in some regions, forcing the transfer of patients between regions and abroad, many treatments had to be postponed or cancelled.

This brief consists of studying the impact of the measures taken against Covid-19 on healthcare spending in metropolitan France in 2020 and projecting spending to 2021. We focus our work on the following five consumption items: routine city care, pharmacy, dental, optical and hearing aids.

First, we focus in more detail on the specific context of the year 2020 and on the indicators put in place to measure the circulation of the pandemic in the country. Then, we focus on the Open DAMIR database and the techniques and reprocessing implemented to allow its proper use. Then, we explain the two time series models traditionally used and project the expected expenditures in 2020 from the learning data from 2015 to 2019. Finally, we introduce a new time series analysis and prediction model completed by the introduction of a central regressor. We evaluate the relevance of this new model by projecting expenditures to 2020. Finally, we project expenditures to 2021 and conclude accordingly.

2020: a year under pressure

Covid-19 has had a profound impact on French society. Although all measures were taken in order to slow down the spread and to protect the population as much as possible, they had the effect of bringing the economy and many aspects of life to a halt.



Evolution of the measures taken within the framework of the management of Covid-19 in France from the beginning of the pandemic to the end of the third containment

In order to measure the evolution of the pandemic in France, many monitoring indicators have been set up and allow the viral circulation of the virus, hospital stress and mortality to be followed on an almost daily basis. The indicator that we have chosen for this report is that of the tension in the intensive care unit and not an indicator linked to the viral circulation of the virus. This indicator is calculated as follows:

$$\frac{Number\ of\ occupied\ ICU\ beds\ due\ to\ Covid}{Number\ of\ beds\ available\ in\ intensive\ care}$$

Despite its obvious interest, this indicator suffers from imperfections. First, the number of available beds does not take into account the creation of new places, which was the case at the height of the

Covid crisis. In addition, the numerator also includes patients who are admitted to intensive care units and continuous care units. Thus, this indicator tends to overestimate the real situation in the ICU.

This administrative closure of the economy will also have had an undeniable impact on the health consumption of the French. Many procedures had to be postponed, resulting in a delay of treatments. In France, according to a study conducted by the French Hospital Federation - the FHF, the voice of public hospitals - on the year 2020, 2.3 million planned stays could not be honored including 1.4 million stays in medicine and 900,000 surgical stays. Compared to 2019, this represents a 12% and 15% decrease, respectively. These numbers are particularly edifying when taken back to the period of the first containment. We can see that:

- Inpatient surgery decreased by 58%;
- Outpatient surgery decreased by 80%;
- Diagnostic colonoscopies collapsed by 87%;
- Kidney transplants have dropped by 80%.

These figures are obviously compared to 2019 over an equivalent period and potentially promise a time bomb that may have several impacts for the future: increased mortality and increased cost of treatment for delayed management that may then lead to increased dependency.

In order to study in more detail the impact of Covid-19 on healthcare expenditure in France, we use the Open DAMIR database.

Source of the data used: the Open DAMIR database

The Open DAMIR database (Dépenses de l'Assurance Maladie Inter Régimes) is an extraction of the SNIIRAM (Système Nationale d'Information Inter-régimes). This database is an anonymized extraction of all reimbursements made by the Health Insurance, with the exception of a large part of hospital expenditures. The data in the database is available in CSV format and is grouped according to the month of payment. In order to obtain the data for the months of payment of year N, it is necessary to wait for the annual update of the database, which generally takes place during the second quarter of year N+1. The choice was made to retrieve the settlement years from 2015 to 2020. From this database, we chose to retain the variables presented in the table below.

Variable de la base	Explication de la variable
AGE_BEN_SNDS	Age range of beneficiary at time of care
BEN_RES_REG	Region of residence of the beneficiary
BEN_SEX_COD	Gender of the beneficiary
PRS_NAT	Type of benefit
FLX_ANN_MOI	Year and month of payment
SOI_ANN	Year of care
SOI_MOI	Month of care
PRS_PAI_MNT	Amount of expense

Three additional variables were also constructed: Item, REG_ANN, REG_MOI. The item variable is derived from the PRS_NAT variable. The type of procedure and consumption item were identified for each service. The choice was made to keep only the consumption items: routine city care, pharmacy, optical, dental and hearing aids. In addition, on a small number of lines, age, sex and region are indicated as "unknown". As their weight is considered negligible in terms of benefits, these lines have been removed and are therefore not considered in the study. Similarly, the study only covers metropolitan France. The lines with the modality "Overseas Regions and Departments" have been removed.

Because of the size of the Open DAMIR databases (a monthly file has a weight of 5 gigabytes), it is unthinkable to use traditional methods and tools. Thus, to extract the data, we are interested in, we used the Python programming language and the Dask library. This library has been designed to handle databases like the Open DAMIR one. In parallel and to accelerate the extraction and manipulation times, we used the Google servers using the Google Colaboratory service.

From a first study on the settlement rates from 2015 to 2019, we observed that at the end of three and twelve months after the occurrence date, we know more than 95% and 99% of the expenditure at the ultimate respectively. We then consider the 2015 through 2019 occurrences to be developed to ultimate from the Open DAMIR data.

However, for the 2020 year of care, we noticed that the expenditure was not complete, especially for the months of November and December, having only two and one months of development. Therefore, it was necessary to estimate the ultimate expenditure for this care year.

Using a backtest of validity, we make the choice to consider that the development coefficients on the year 2020 are like those observed on the year 2019. We therefore decided to perform a Chain Ladder liquidation on the five triangles by consumption item of the 2020 year of care with a 12-month horizon, to obtain ultimate benefits. This gives us a 12/31/2021 (ultimate) view of the 2020 year of care.

The analyses we perform next are therefore based on a reconstructed database as follows:

> An actual view (from the Open DAMIR databases) as of 12/31/2020 for the 2015 through 2019 years of occurrence;
> An estimated view (via Chain Ladder) as of 12/31/2021 of the expenditure for the 2020 occurrence year.

Projection of spending for the year 2020 as of December 31, 2019

Due to the structure and availability of the Open DAMIR data, we intuitively decided to use the time series concept. From this initial idea, we seek to determine if there is a time series pattern underlying our data over the period before 2020. We seek to identify any trends or seasonality in the time series studied.

To do this, we use the two most well-known models for the analysis and prediction of time series: ARIMA (Autoregressive Integrated Moving Average) and SARIMA (Seasonal Autoregressive Integrated Moving Average). We recall the equations of these two models below ($\forall t \geq 0$):

- $ARIMA$ model: $\phi_p(L)(1-L)^d y_t = \mu + \theta_q(L)\epsilon_t$
- $SARIMA$ model: $\phi_p(L)(1-L)^d * \Phi_P(L^m)(1-L^m)^D * Y_t = \mu + \theta_q(L) * \Theta_Q(L^m) * \epsilon_t$

With:

- ϵ_t, a weak white noise. It is a sequence of identically distributed real random variables that are not correlated with each other. ($\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$) ;
- m is the seasonality operator ($m = 12$ for monthly data);
- $\phi_p(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$;
- $\theta_q(L) = 1 + \theta_1 L - \cdots + \theta_q L^q$;
- The seasonal polynomials $\Phi_P(L^m)$ et $\Theta_Q(L^m)$ are defined as follow:
  - $\Phi_P(L^m) = 1 - \Phi_1 L^m - \cdots - \phi_P L^{Pm}$;
  - $\Theta_Q(L^m) = 1 + \Theta_1 L^m - \cdots + \Theta_Q L^{Qm}$.

- All the roots of the characteristic equation associated to ϕ, θ, Φ et Θ are of modulus lower than 1;

We summarize, in the table below, the ARIMA and SARIMA models obtained on our five consumption segments.

Consumption segment	Model $ARIMA$ retained	Model $SARIMA$ retained
Routine city care	$ARIMA(11,0,2)$	$SARIMA(0,0,2)(0,1,0)_{12}$
Pharmacy	$ARIMA(11,0,1)$	$SARIMA(2,0,0)(1,1,0)_{12}$
Dental	$ARIMA(11,0,1)$	$SARIMA(1,0,1)(0,1,0)_{12}$
Optical	$ARIMA(11,0,0)$	$SARIMA(0,0,1)(3,1,0)_{12}$
Hearing aids	$ARIMA(11,0,1)$	$SARIMA(1,0,1)(1,1,0)_{12}$

From the large number of autoregressive terms on the ARIMA models of our five consumption items, we can easily conclude that there is a strong seasonality in spending and that the application of the SARIMA model was essential.

Once our models are established, we can make projections to the year 2020. Similar conclusions emerge for our five consumption segments:

- For period n°1, the expenditure estimated by our two models is very close to the observed expenditure;
- For period n°2, our two models are wrong. It was indeed unthinkable to imagine the occurrence of an exceptional event such as Covid-19;
- In period 3, the expected expenditure is very close to the actual expenditure. We do not observe any obvious catch-up in consumption;
- In period 4, our models are wrong again. It would seem that a substantial catch-up is observed in this period. It should be noted, however, that this period is the most uncertain of all, since it is mainly this period for which we do not know the entirety of the actual expenditure from the Open DAMIR database.

Introduction of a new time series model: Prophet developed by Facebook

As we have seen, traditional time series analysis and prediction models proved to be unable to predict the decrease in consumption in the second period and the increase in consumption in the fourth period. This was entirely predictable. However, in order to improve our predictions and with the goal of projecting spending to the year 2021 as efficiently as possible, we wanted to explore other approaches and found Prophet, a mathematical model developed by Facebook engineers that can be easily applied on Python. The equation of the chosen model is as follows:

$$y(t) = g(t) + s(t) + \boldsymbol{r(t)} + \epsilon_t$$

With:

- $g(t)$ represents the trend of the time series ($g(t) = (k + a(t)\delta^T)t + (m + a(t)\gamma^T)$);
- $s(t)$ represents the seasonality of the time series ($s(t) = \sum_{n=1}^{N} \left[a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right)\right]$;
- $r(t)$ represents the regressors. That is to say, the data external to the time series added by the user that could have an influence on it ($r(t) = \sum_{n=1}^{N} \alpha_i * W_{i,t}$) ;
- ϵ_t represents the error term, everything that could not be correctly explained by one of the four components of the model. The only assumption about this error term is that it is normally distributed.

Prophet was designed with one objective in mind: to simplify the analysis and prediction of time series while maintaining quality. Prophet thus seeks to address specialists in a given field (in our case, health) and thus avoids generating a « black box » model that a user would be unable to understand at a glance. Prophet is based on the concept of « **analyst-in-the-loop** » modeling, i.e., modeling that combines traditional objective statistical analysis with a subjective analysis that draws its strength from experience acquired in the field.

In order to improve our projections to 2020, and as briefly introduced in the presentation of the equation, we will also introduce a central regressor $\Re$ constructed from the following primary regressors:

- Business climate indicator - All sectors - Metropolitan France ($ConfAffaire$);
- Sales index - All trade ($ChiffreAffCom$);
- Occupancy rate in hotels - Metropolitan France ($OccupHotel$);
- Demographics - Number of deaths - Metropolitan France ($Deces$);
- Hospital data related to the COVID-19 epidemic where we will focus primarily on individuals in intensive care. ($ReaCovid$)

Our five regressors were normalized between 0 and 1 using the following equation:

$$r = \frac{r(t) - min(r)}{max(r) - min(r)}$$

With:

- r, representing the regressor;
- $r(t)$, representing the value of the regressor in t ;
- $min(r)$, representing the minimum value of the observed time series;
- $max(r)$, representing the maximum value of the observed time series.

The central regressor equation defined is as follows:

$$\Re = \frac{ConfAffaire + Deces + OccupHotel + ChiffreAffCom - ReaCovid}{5}$$

We then performed a double application of the Prophet model:

- The application of the model without regressor for the consumption segment *Pharmacy*;
- The application of the model with the central regressor for the four other consumption segments.

We also took the opportunity to compare the results of the Prophet model with those obtained with the SARIMA model. The results are summarized in the table below (in millions of euros):

	Expenditure as of 12/31/2020	**SARIMA Prediction**	**Prophet Prediction**	**SARIMA error**	**Prophet error**	**% SARIMA error**	**% Prophet error**
Routine city care	**51 289**	49 345	51 494	**1 944**	**- 205**	**3,9 %**	**- 0,4 %**
Pharmacy	**35 756**	34 950	35 242	**806**	**515**	**2,3 %**	**1,5 %**
Dental	**10 741**	11 423	11 903	**- 682**	**- 1 162**	**- 6,0 %**	**- 9,8 %**
Optical	**6 162**	8 606	6 957	**- 2 444**	**- 794**	**- 28,4 %**	**- 11,4 %**
Hearing aids	**2 298**	2 552	2 307	**- 254**	**- 8**	**- 9,9 %**	**- 0,3 %**

From this table, we can draw our first conclusions from the application of this new model:

- $Prophet$ was much more relevant in predicting consumption than the $SARIMA$ model - with the exception of the Dental segment;
- The application of our regressor allowed us to gain slightly in accuracy on period n°2 but the results remain below the initial expectation;
- The prediction of the Prophet model, assessed on a yearly basis, turned out to be very close to what was expected as we can see in the table below (except for the Dental segment).

In a second step, to make our predictions for the year 2021, we estimated the evolution of our primary regressors during the year and then we built our central regressor on the same scheme defined previously. The results of our projections for the year 2021 are presented below (in millions of euros):

	Routine city care	Pharmacy (without regressor)	Dental	Optical	Hearing aids	Total
January	4 920,56	3 175,46	989,31	563,73	216,87	**9 865,93**
February	4 353,48	2 751,32	925,00	622,37	194,02	**8 846,20**
March	4 935,28	3 108,36	1 109,60	693,09	230,39	**10 076,71**
April	4 499,85	2 892,21	967,83	642,10	206,81	**9 208,80**
May	4 698,39	2 921,01	1 005,05	612,35	206,27	**9 443,06**
June	4 667,74	2 974,66	1 116,30	654,21	217,13	**9 630,04**
July	4 111,45	2 911,81	993,42	632,15	186,84	**8 835,68**
August	3 568,46	2 717,55	535,60	510,17	131,36	**7 463,14**
September	4 562,96	2 918,97	1 008,47	605,26	192,69	**9 288,35**
October	4 710,93	3 076,35	1 086,33	632,96	218,32	**9 724,89**
November	4 698,35	2 999,01	1 068,46	619,99	225,46	**9 611,27**
December	4 259,96	2 995,08	1 014,00	753,46	223,15	**9 245,66**
Total	**53 987,40**	**35 441,80**	**11 819,36**	**7 541,85**	**2 449,32**	**111 239,73**

Conclusion

Thus, it is undeniable that the pandemic has had an impact on health expenditure in 2020, the consequences of which will continue to be felt in the years to come. The models we have been able to apply have illustrated this perfectly: a very significant drop in consumption followed by a fairly substantial increase in the fourth period (which we can attribute to a catch-up).

$Prophet$ really has the advantage of being more understandable than traditional time series analysis and prediction models. The ease with which we can use it makes it much more accessible.

However, even if the application of the $Prophet$ model allowed us to slightly better understand the consumption of the second and fourth periods, it did not prove to be as efficient as anticipating. Thus, as the Open DAMIR database allows us to do, it would be advisable to further refine our analyses by region and age group of the insured. Further study of regressors might also be necessary if such an approach is retained.

Remerciements

Avant toute chose, je tiens à adresser mes remerciements à Sergio OROZCO pour son soutien, ses conseils, sa disponibilité et ses diverses relectures lors de la rédaction de ce mémoire. Ses remarques pertinentes et idées m'ont été très profitables. J'ai ainsi pu avancer efficacement et réaliser ce mémoire dans les meilleures conditions possibles.

Je tiens, par ailleurs, à exprimer ma gratitude pour mon tuteur académique – Nicolas BOUSQUET – pour le suivi de ce mémoire et les précieux conseils qu'il a pus me fournir.

Je remercie également Norbert GAUTRON, Léonard FONTAINE, ainsi que Julien BOUDOT de m'avoir fait confiance en me permettant d'effectuer mon stage et mon alternance au sein du cabinet GALEA & Associés.

Je tiens également à remercier l'ensemble des consultants du cabinet pour leur accueil et leur bonne humeur, me permettant de m'intégrer au sein des équipes et de passer une alternance agréable en leur compagnie.

Enfin, un grand merci à tous mes proches qui m'ont soutenu tout au long de la réalisation de mon mémoire et sans qui cette épreuve aurait été plus difficile.

Table des matières

I. Introduction 21

II. Protection sociale et Assurance Maladie en France 22

A. Présentation de la Sécurité Sociale 22
1. Les branches de la Sécurité Sociale 22
2. La gestion de la Sécurité Sociale 23
3. Le financement de la Sécurité Sociale 23
4. Les comptes de la Sécurité Sociale en 2019 24
5. Le déficit de la sécurité sociale 24

B. L'Assurance Maladie obligatoire et les régimes complémentaires 26
1. L'Assurance Maladie Obligatoire 26
2. L'Assurance Maladie Complémentaire 27

C. La réforme 100 % Santé 31
1. L'identification d'une source d'inégalité : le renoncement aux soins 31
2. Les tenants et aboutissants de la réforme 32
3. Les premiers impacts chiffrés de la réforme 33

III. Présentation du contexte politico-économique : 2020 et le COVID 34

A. Le contexte particulier de 2020 34
1. Contexte sanitaire 34
2. La pandémie et la France 35

B. Indicateurs de suivi du Covid-19 37
1. La circulation virale 37
2. La tension hospitalière 38
3. La mortalité 39

C. Impact sur la consommation de santé 44
1. Le renoncement aux soins 44
2. Rattrapage de la consommation santé après le premier confinement 46

IV. Présentation des données : la base Open DAMIR 47

A. Le SNIIRAM 47
1. Auteurs 47
2. Objectifs 47
3. Construction de la base 48
4. Contrôle par la CNIL 49

B. Open DAMIR 50
1. Champ d'étude de la base Open DAMIR 50
2. Réforme et impact 50
3. Limites de la base Open DAMIR 50
4. Les variables retenues pour l'étude 51
5. Manipulation de la base Open DAMIR 52

C. Statistiques descriptives sur les données de la base OPEN DAMIR sur les années étudiées 55
1. Évolution globale de la consommation 55
2. Évolution par poste de consommation 57

D. Estimation de la dépense à l'ultime pour l'année de soins 2020 64
1. Présentation de la méthode déterministe de Chain Ladder 64

2. Application du test rétroactif dit « backtesting » de validité 68
3. Étude sur les cadences des règlements 71
4. Application : estimation de la charge ultime de la survenance 2020 72

V. La prédiction des dépenses de l'année 2020 73

A. Présentation des modèles et concepts utilisés 73
1. Introduction aux séries temporelles 73
2. Les concepts fondamentaux 73

B. Application des deux méthodes de prédiction de séries temporelles 84
1. Application d'ARIMA et de SARIMA : exemple sur le poste *Soins de ville courants* 85
2. Conclusion sur nos modèles 90

C. Prédictions sur 2020 91
1. Présentation des résultats obtenus après application de nos méthodes 91
2. Limites de cette première approche 102

VI. Mise en place d'un nouveau modèle : Prendre en compte le caractère exceptionnel de 2020 103

A. Présentation du modèle mathématique de Prophet 103
1. Équation du modèle 103
2. Ajustement du modèle 107
3. Les spécificités de Prophet 107

B. Présentation des régresseurs 108
1. Nos régresseurs primaires 108
2. Manipulation et transformation de nos régresseurs 113

C. Implémentation du modèle sous Python 115
1. Présentation de la fonction Prophet 115
2. L'implémentation 116

D. Projections des dépenses sur 2021 128
1. Présentation 128
2. Visualisation de la projection 129

VII. Conclusion 131

VIII. Références 132

1. Bibliographie 132
2. Sitographie 133

IX. Annexe 134

A. Annexe 1 : vérification des hypothèses de Chain Ladder sur les autres postes de consommation 134

B. Annexe 2 : test rétroactif de validité sur les autres postes de consommation 138

C. Annexe 3 : étude sur les cadences de règlements pour les autres postes de consommation 140

D. Annexe 4 : application des modèles ARIMA & SARIMA sur les autres postes de consommation 144

E. Annexe 5 : algorithme d'optimisation BFGS 156

F. Annexe 6 : prédictions du modèle Prophet par mois de soin 157

I. Introduction

L'année 2020 restera dans la mémoire générale comme l'année ayant vu l'émergence d'une nouvelle maladie – le Covid-19 – dont l'entièreté des répercussions mettra du temps à être totalement appréhendée et mesurée. La rapide propagation de la maladie à travers le globe – rapidement qualifiée de pandémie par l'Organisation Mondiale de la Santé (OMS) – a pris de très nombreux Etats de court : des décisions radicales ont été par conséquent prises. C'est notamment le cas de la France qui a imposé un premier confinement de sa population à l'échelle nationale le 17 mars 2020.

Cet arrêt brutal de l'économie et de la « vie d'avant » a touché de très nombreux secteurs, y compris celui de la santé. En effet, en plus des images montrant les hôpitaux français saturés dans certaines régions françaises obligeant alors le transfert de patients entre régions et à l'étranger, de nombreux soins ont dû être reportés ou annulés.

Ce mémoire consiste à mesurer l'impact des mesures prises contre le Covid-19 sur les dépenses santé en France métropolitaine en 2020 et à effectuer une projection des dépenses sur l'année 2021.

Pour ce faire, nous exploitons la base Open DAMIR de 2015 à 2020 et concentrons nos travaux sur les cinq postes de consommation suivants : *soins de ville courants*, *pharmacie*, *dentaire*, *optique* et *prothèses auditives*. Après un retraitement indispensable de cette base de données disponible par mois de règlement, nous effectuerons une analyse statistique sur chaque poste de consommation afin d'identifier clairement toute tendance ou saisonnalité dans la dépense avant la pandémie. Dans un second temps, nous ferons appel à des techniques traditionnelles d'analyse et de prédiction de séries temporelles. Finalement, afin d'affiner nos prédictions, nous introduisons un nouveau modèle de prédiction de séries temporelles développé par les ingénieurs de Facebook : ***Prophet***.

II. Protection sociale et Assurance Maladie en France

A. Présentation de la Sécurité Sociale

La Sécurité Sociale est un système de protection sociale obligatoire ayant pour vocation de protéger l'ensemble de la population. Instituée en 1945 à la sortie de la seconde guerre mondiale par le gouvernement de Gaulle puis régulièrement réformée depuis, la Sécurité Sociale représente l'un des piliers fondateurs de la France contemporaine. Son objectif est double :

- Le premier objectif consiste à assister chaque individu lorsqu'il se retrouve confronté à un risque de la vie : maladie, invalidité, décès, accident du travail, chômage, maladie professionnelle, vieillesse, etc. Elle est universelle : tous les individus sont couverts sans prise en compte de la situation financière, de la situation familiale ou de l'état de santé.
- Son second objectif, plus subtil mais tout aussi important, est celui de la prévention. La finalité est d'agir en amont des maladies et de ses complications en incitant les assurés à se faire dépister. Une détection précoce est ainsi synonyme pour l'assuré d'un traitement moins lourd, de séquelles moindres et d'une augmentation de la durée de vie. Pour la Sécurité Sociale, cela permet de réduire les coûts associés au traitement lourd de santé. Cette volonté de prévention passe également par une analyse de terrain et par l'identification des « risques de demain » liés à la santé comme nous le verrons par la suite.

La Sécurité Sociale est composée de divers régimes dont le principal, le régime de base, couvre environ 80 % de la population. Il faut également considérer des régimes spéciaux couvrant une partie spécifique de la population telle que les salariés de la fonction publique ou du secteur agricole par exemple.

La Sécurité Sociale est un modèle mixte mélangeant :

- Le modèle *bismarckien* : les assurés sont couverts s'ils se sont assurés pour un risque via le versement de cotisations. La prestation est proportionnelle au salaire ;
- Le modèle *beveridgien* : les assurés sont couverts de façon universelle. Pour accéder à la couverture, il n'est pas nécessaire d'avoir participé à son financement. La prestation est forfaitaire et identique pour tous.

1. Les branches de la Sécurité Sociale

Les divers domaines d'intervention et catégories de risques couverts par la Sécurité Sociale ont été divisés en six branches par le code de la Sécurité Sociale. Ces six branches cherchent à couvrir chacune une catégorie de risque bien spécifique :

- La branche maladie (maladie, maternité, invalidité et décès) ;
- La branche accidents du travail et maladies professionnelles (AT/MP) ;
- La branche vieillesse et veuvage (retraite) ;
- La branche famille (allocations familiales, handicap, logement, RSA) ;
- La branche recouvrement (récupération des cotisations et contributions sociales) ;
- La branche autonomie.

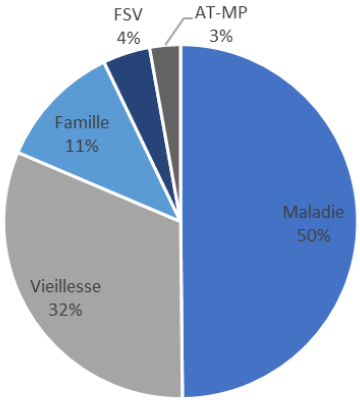


Figure II-1 – Répartition en 2019 des dépenses par branche

La sixième et dernière branche a été instaurée en ce début d'année 2021 et est dédiée à l'autonomie des personnes âgées et handicapées. Cette nouvelle branche illustre la volonté de prévention de la Sécurité Sociale en raison d'une population française de plus en plus vieillissante et sédentarisée. D'après l'INSEE, 1 200 000 personnes seront dépendantes en 2040, contre 800 000 en 2006.

2. La gestion de la Sécurité Sociale

Chaque branche de la sécurité sociale se voit gérer par une caisse à l'échelle nationale :

- La caisse nationale de l'assurance maladie (CNAM) qui gère les branches maladie et AT/MP ;
- La caisse nationale des allocations familiales (CNAF) qui gère la branche famille ;
- La caisse nationale d'assurance vieillesse (CNAV) qui gère la branche vieillesse ;
- L'URSSAF qui gère la branche recouvrement ;
- La caisse nationale de solidarité pour l'autonomie (CNSA).

3. Le financement de la Sécurité Sociale

Comme la Sécurité Sociale est un régime obligatoire, n'effectuant pas de distinction entre les personnes couvertes, son financement repose sur le principe de solidarité nationale. Chaque individu participe ainsi à son financement en fonction de ses moyens. En 2018, nous pouvons distinguer les six sources de contributions suivantes :

- **Cotisations sociales :** cotisations et contributions versées par les travailleurs et les entreprises représentant 54 % des ressources totales ;
- **Contribution sociale généralisée (CSG) :** contribution obligatoire supplémentaire prélevée sur les revenus d'activité et de remplacement, les revenus liés au patrimoine et au placement ainsi que les gains et mises liés au jeu. Cette contribution participe à hauteur de 26 % des ressources de la Sécurité Sociale ;
- **Impôts, taxes et autres contributions sociales :** avec par exemple les taxes sur les produits nuisant à la santé (tabac, alcool …). Ce poste représente 11 % des ressources de la Sécurité Sociale ;
- **Transfert net :** à hauteur de 5 %. Il s'agit principalement de refacturations entre branches au titre de la prise en charge de prestations ou de cotisations ;
- **Contribution de l'état :** à hauteur de 2 % et qui sert à financer les dépenses liées à la solidarité telles que les fonds de solidarité vieillesse ;
- **Autres produits** à hauteur de 2 %.

4. Les comptes de la Sécurité Sociale en 2019

Les recettes et dépenses de la Sécurité Sociale en 2019 sont résumées dans le tableau ci-dessous.

Branche du régime générale	Produit (en M€)	Charge (en M€)	Résultat (en M€)
Maladie	215 182	216 648	- 1 466
Vieillesse	135 717	137 125	- 1 408
Famille	51 401	49 877	+ 1 525
AT-MP	13 214	12 239	+ 975
Fonds de Solidarité Vieillesse	17 214	18 767	- 1 553
Total	432 728	434 656	- 1 928

Figure II-2 – Résultat de la Sécurité Sociale en 2019

Il est intéressant de noter que les dépenses représentent 17 % du PIB français de l'année 2019.

5. Le déficit de la sécurité sociale

Au début des années 2000, le solde de la Sécurité Sociale est légèrement bénéficiaire grâce aux réformes des années précédentes : la réforme des retraites de 1993 permettant le redressement des comptes de la branche vieillesse, ainsi que les réformes de 1996 touchant les branches maladie et famille. Cette situation s'explique également par l'environnement macroéconomique favorable à la France à cette époque.

Cependant depuis 2002, les comptes de la Sécurité Sociale n'ont jamais retrouvé un état d'équilibre. La crise financière de 2008 est venue accroître ce déficit avec un point d'orgue en 2010 où la perte nette de la Sécurité Sociale s'est élevée à 28 milliards d'euros.

Les années suivantes ont permis d'observer un retour de cette accalmie et un déficit qui s'est réduit progressivement grâce à une augmentation des recettes et des dépenses mieux maîtrisées.

Malgré une amélioration progressive de la situation, la Sécurité Sociale ne sera pas épargnée par les conséquences de la pandémie du Covid-19 de 2020 : le déficit attendu pour cette année-là semblerait être l'un des plus importants jamais enregistrés par le régime. Avec un déficit qui atteindrait 38,9 milliards d'euros (dont 3 milliards de provisions pour des cotisations encore recouvrables), le déficit de 2020 serait équivalent au déficit cumulé des six dernières années (de 2014 à 2019).

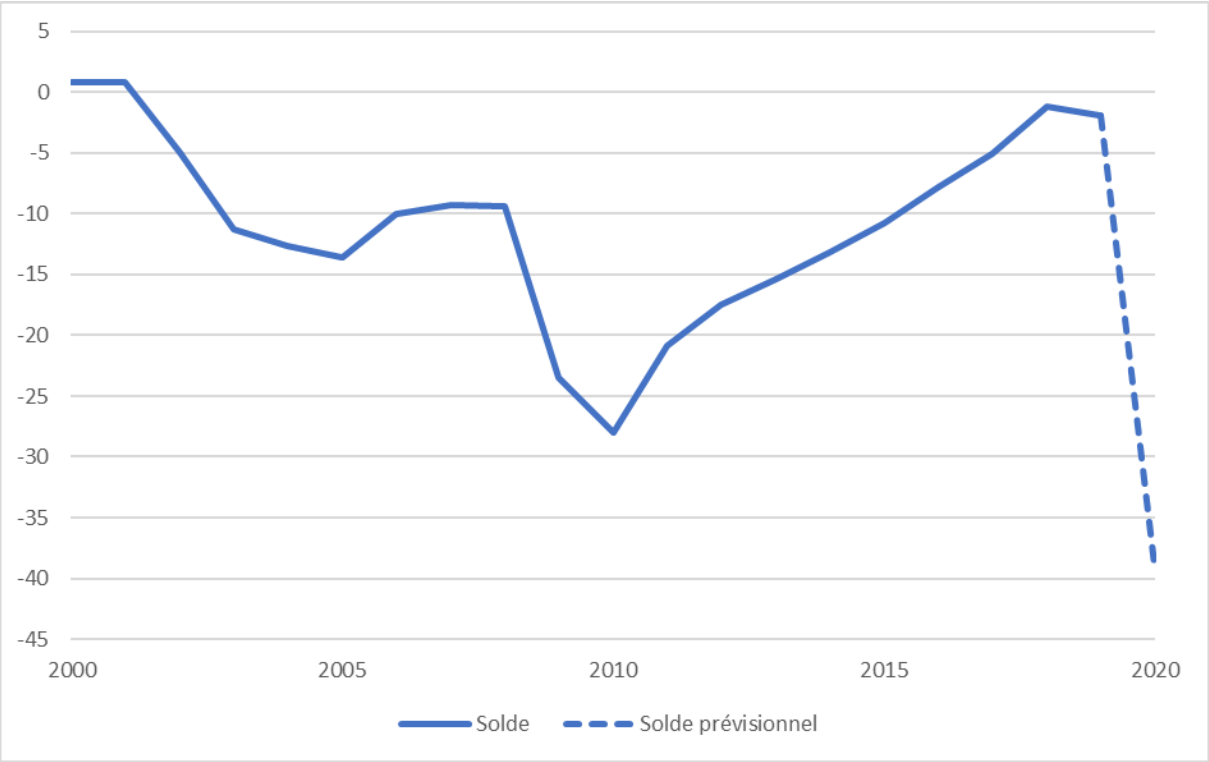


Figure II-3 – Solde du régime général depuis 2000 (en M€)

Les causes du déficit sont multiples et peuvent être regroupées en deux catégories : les causes conjoncturelles et structurelles.

a) Causes conjoncturelles du déficit

Les recettes de la Sécurité Sociale sont intimement liées à l'état de l'économie française puisque dès que la croissance est ralentie, la masse salariale totale se réduit influant directement sur le taux de chômage. Un taux de chômage élevé entraîne logiquement une diminution des cotisations sociales et une augmentation des dépenses liées à l'indemnisation du chômage et des aides sociales. Cela a pour effet de creuser davantage le déficit.

b) Causes structurelles du déficit

Avec une augmentation de la durée moyenne de vie et l'apparition de maladies spécifiques aux personnes âgées, des soins de plus en plus pointus et coûteux sont dispensés. De plus, une augmentation du niveau de vie incite les populations à faire plus attention à leur santé et donc à faire appel à des soins considérés auparavant comme « de confort ». Les progrès médicaux ont un double impact sur le déficit dont les effets sont inversés puisque d'un côté, l'apparition de nouveaux traitements et technologies est une nouvelle source de dépense supplémentaire et de l'autre, les progrès médicaux peuvent rendre les traitements et soins plus efficaces et donc contribuer à réduire les dépenses de santé, participant donc à réduire le déficit.

Après avoir parcouru brièvement la protection sociale en France, nous allons nous intéresser plus particulièrement à la CNAM, caisse en charge des branches *maladie* et *AT-MP* de la Sécurité Sociale.

B. L'Assurance Maladie obligatoire et les régimes complémentaires

1. L'Assurance Maladie Obligatoire

La CNAM représente la plus grosse caisse de la Sécurité Sociale étant donné sa mission principale : rembourser les frais de santé de la population. Elle mène également une politique de gestion du risque ayant pour but d'améliorer la santé publique, de renforcer l'efficacité du système de soins et de maîtriser l'évolution des dépenses de santé.

L'Assurance Maladie distribue deux catégories de prestations :

- Les prestations en nature qui rassemblent les prestations relatives à la santé. Elle rembourse par exemple les consultations, les soins dentaires ou les prothèses auditives ;
- Les prestations en espèces qui regroupent les prestations relatives à la prévoyance. Nous pouvons y retrouver les indemnités journalières versées suite à un accident du travail ou une maladie professionnelle ainsi que les prestations versées dans le cadre d'une invalidité ou d'un décès.

Dans ce mémoire, nous traiterons la branche *maladie* de l'Assurance Maladie, en laissant de côté la branche *AT-MP*.

Le remboursement des frais de santé de la branche *maladie* correspond donc à des prestations en nature et regroupe :

- Les honoraires de médecins et auxiliaires médicaux ;
- Les analyses et examens de laboratoire ;
- Les médicaments ;
- L'hospitalisation ;
- Les frais de transport ;
- La cure thermale ;
- Les « autres » frais médicaux : optique, appareillages, prothèses, ….

L'Assurance Maladie va venir rembourser une partie de la dépense selon une base de remboursement prédéfinie étant amenée à évoluer au fil du temps comme nous le verrons par la suite.

L'exemple traditionnellement utilisé pour illustrer l'intervention de l'Assurance Maladie est la consultation chez un généraliste conventionné exerçant en secteur 1[1]. L'assuré connait le montant de cette prestation à l'avance, elle coûte 25 €. Cette dépense peut être découpée comme suit :

- L'Assurance Maladie prend en charge 70% de la dépense effectué selon un tarif conventionné auquel il faut retirer une participation forfaitaire obligatoire2 de 1€. L'Assurance Maladie rembourse ainsi 70% * 25€ - 1€ = 16,5 €.
- L'assuré a donc un reste à charge de 8,5 €.

Ce reste à charge peut être pris en charge par une assurance maladie complémentaire qui vient – comme son nom l'indique – en complément des remboursements du système de base.

[1] Le secteur 1 correspond aux tarifs de base fixés par l'Assurance Maladie. Les tarifs sont fixés par convention et connus à l'avance.

[2] Certaines catégories de la population n'auront pas de participation forfaitaire à régler. C'est le cas pour les individus de moins de 18 ans, les femmes après leur 6ème mois de grossesse et les bénéficiaires de la Complémentaire santé solidaire ou de l'aide médicale de l'État.

2. L'Assurance Maladie Complémentaire

En complément du régime de base, tout individu a la possibilité de souscrire une complémentaire santé qui viendra compléter les remboursements de soins effectués par la Sécurité Sociale. Obligatoire (pour une très grande majorité des salariés du privé par exemple) ou facultative (pour les retraités par exemple), les complémentaires santé répondent au principe de non-enrichissement et visent à prendre en charge de façon partielle ou totale les dépenses de santé qui n'ont pas été couvertes par le régime général. Ce reste à charge pour l'assuré est appelée *ticket modérateur*. Les complémentaires santé se reposent ainsi sur la liberté tarifaire pratiquée par certains practiciens ou sur certains produits et actes. Les complémentaires santé peuvent également intervenir sur des garanties dites « de confort » tels qu'une visite chez un diététicien.

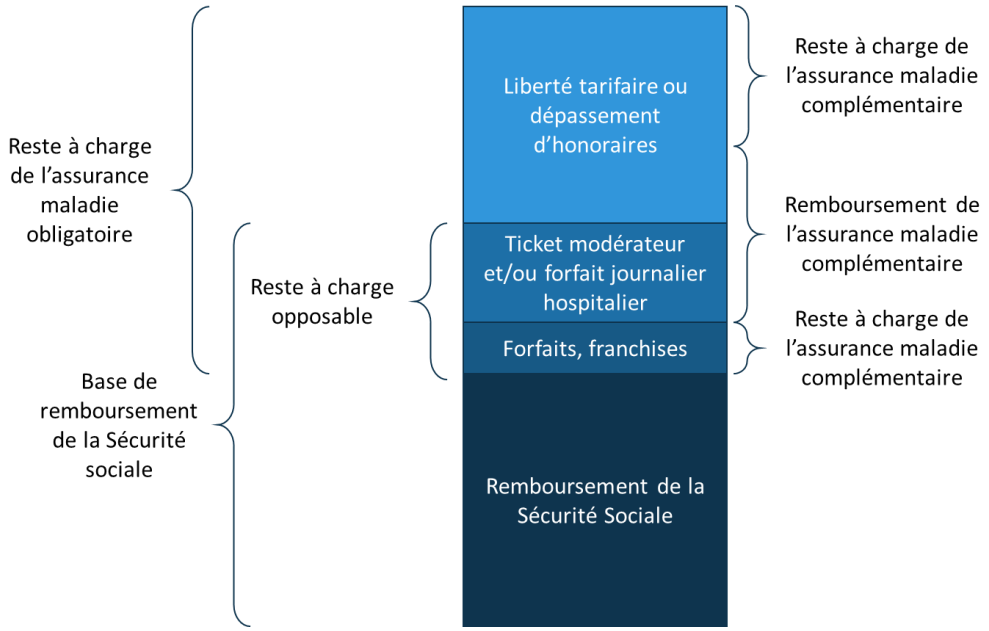


Figure II-4 – Exemple de découpage d'une dépense du panier de soins remboursable pour un soin de ville courant

Reprenons notre exemple précédent d'une consultation chez un médecin généraliste conventionné exerçant en secteur 1. La dépense peut être précisée comme suit :

- La participation forfaitaire s'élève à 1 € et est à la charge de l'assuré ;
- L'Assurance Maladie prend en charge 70% de la dépense effectué selon un tarif conventionné auquel il faut retirer la participation forfaitaire. Ainsi, L'Assurance Maladie rembourse 70% * 25€ - 1€ = 16,5 €.
- Le ticket modérateur correspond à ce qui reste à charge de l'assuré, après déduction des deux découpages précédents. Le ticket modérateur s'élève donc à 7,5 €. C'est sur cette partie-ci que les complémentaires santé peuvent intervenir.

Pour la consultation d'un médecin généraliste conventionné exerçant en secteur 2[3], la dépense santé peut s'élever pour l'assuré à 30€, par exemple. Cette dépense peut être découpée comme suit :

- La participation forfaitaire s'élève à 1€ ;

[3] A l'inverse du secteur 1, les tarifs du secteur 2 sont fixés à la discrétion du généraliste, c'est ce qu'on appelle les honoraires libres. C'est dans ce cadre-là que nous parlerons alors de dépassement d'honoraires.

- L'Assurance Maladie prend en charge 70% de la dépense effectuée selon un tarif conventionné fixé à 23€. Comme pour le secteur 1, la participation forfaitaire est fixée à 1 €. Ainsi, l'Assurance Maladie va rembourser 70% * 23 € - 1 € = 15,1 €.
- Le ticket modérateur est de 23 € - (15,1 € + 1 €) = 6,9 €
- Le dépassement d'honoraire est de 7 €.

La complémentaire santé sera ainsi amenée à rembourser une partie ou la totalité des 13,9 € encore à charge de l'assuré.

Les acteurs du marché de la complémentaire santé

Il existe trois typologies d'acteurs qui se partagent le marché de l'assurance santé. Il s'agit des mutuelles, des sociétés d'assurances et des institutions de prévoyance. Chaque acteur est régi par une réglementation qui lui est propre : le Code de la Mutualité, le Code des Assurances et le Code de la Sécurité Sociale respectivement. Alors que le but d'une société d'assurance est de générer du profit, les mutuelles et institutions de prévoyance ont un but non lucratif puisqu'elles sont soit contrôlées par leurs adhérents dans le cadre des mutuelles, soit administrées paritairement[4] pour les institutions de prévoyance.

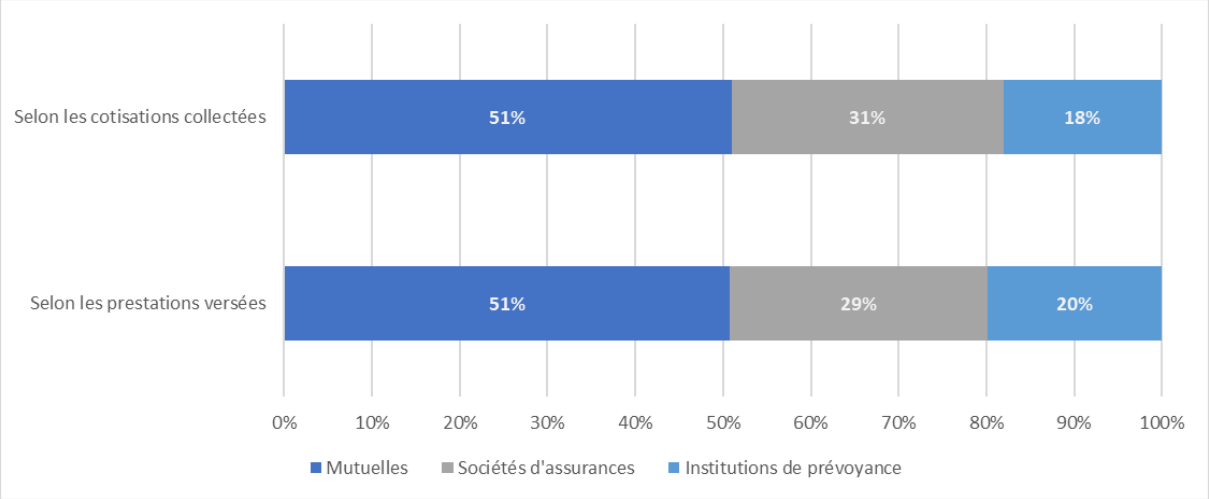


Figure II-5 – Part des marchés des acteurs de la complémentaire santé

Le graphique, ci-dessus, nous permet d'observer la répartition de ces trois acteurs sur le marché selon les cotisations collectés et les prestations versées.

Sur les dernières années, nous pouvons observer que les sociétés d'assurances occupent une place de plus en plus importante dans ce marché au détriment des deux autres familles d'acteurs. En effet, en 2012, les prestations et cotisations des sociétés d'assurances ne représentaient qu'environ 20% du marché.

Par ailleurs, il est à remarquer l'apparition de nouveaux acteurs dans le monde de la complémentaire santé tels que les banques, la grande distribution et les néo-assureurs.

Les mutuelles gardent une position privilégiée sur ce marché étant donné que l'Assurance Maladie représente le cœur de leur activité. En outre, si nous nous intéressons au champ des risques sociaux (la Santé, la Retraite, le Décès et l'Incapacité Invalidité) couverts par ces trois acteurs, nous pouvons

[4] Être administré de façon paritaire impose une cogestion de l'institution de prévoyance par un nombre égal de représentants des salariés et de représentants des entreprises.

remarquer que l'activité principale des mutuelles est de couvrir les risques liés à la Santé, à l'inverse des sociétés d'assurances qui restent assez généralistes.

Le graphique ci-dessous, représente la répartition des cotisations collectées en 2016 sur les risques Santé, Retraite, Décès et Incapacité / Invalidité par les trois familles d'acteurs citées précédemment.

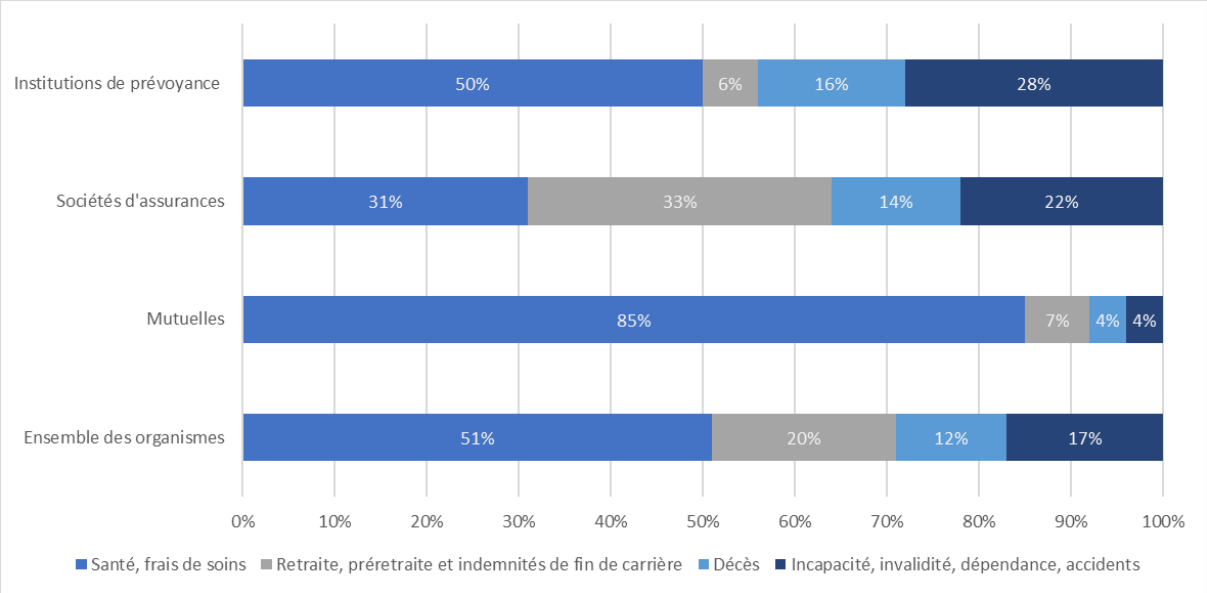


Figure II-6 – Répartition des cotisations collectées en 2016 par famille d'organisme assurantiel

Il est également possible d'observer une concentration des acteurs sur ce marché qui s'explique en partie par l'émergence de nouvelles exigences réglementaires incitant les organismes à se réorganiser. Cette réorganisation s'est – de ce fait – traduite par des « fusions-absorptions ».

En 2017, nous dénombrons 346 mutuelles, 103 sociétés d'assurances et 25 institutions de prévoyance, soit un total de 474 organismes pratiquant une activité d'assurance santé. Ces chiffres sont à mettre en parallèle avec ceux de 2001 puisqu'il existait alors 1 702 organismes, environ trois fois plus d'entités. Ce sont principalement les mutuelles qui ont été concernées par ces « fusions-absorptions » puisque leur nombre a été divisé par quatre et celui des institutions de prévoyance par deux.

En observant la répartition de la dépense courante de santé entre financeurs en 2015, nous observons que 79 % de la dépense courante sont pris en charge par le régime obligatoire, 14 % par les régimes complémentaires et 7 % reste à la charge des ménages.

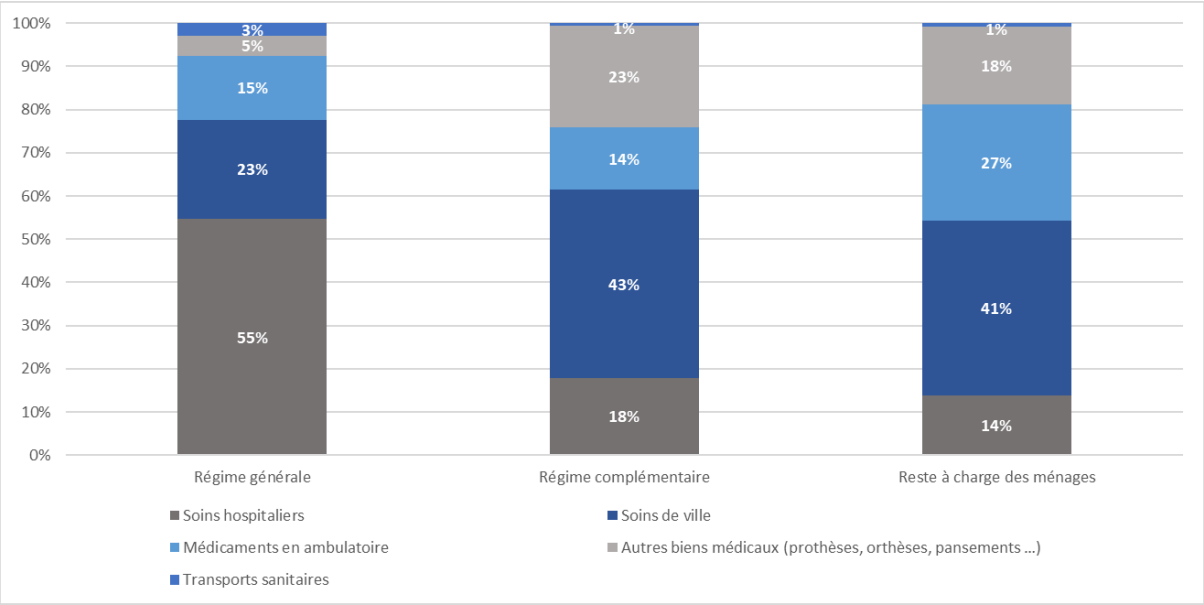


Figure II-7 – Poids des différents postes de consommation en 2019

Cependant, la sous-répartition de la dépense – c'est-à-dire selon les postes de consommation – n'est pas similaire entre les régimes comme nous pouvons le voir dans le graphique ci-dessus. Nous constatons que 55% des dépenses de santé du régime général sont liées aux soins hospitaliers, alors que l'hospitalisation représente 18% pour le régime complémentaire. Dans le sens inverse, les soins de ville représentent 23% des dépenses du régime général, alors qu'ils représentent 43% du régime complémentaire.

Les contrats de santé responsable

Le contrat responsable est entré en vigueur au 1er janvier 2006. Pour qu'un contrat de santé soit considéré responsable, il doit respecter un cahier des charges, régi par des obligations réglementaires, comprenant notamment des niveaux de prises en charge, des garanties planchers et des plafonds de remboursement applicables à certains postes de soins. Le principe fondateur d'un contrat responsable est le même que celui à l'origine de l'assurance : la *solidarité*. Pour y adhérer, il n'y a pas besoin de remplir un questionnaire médical et les cotisations sont indépendantes de l'état de santé de l'assuré.

Malgré l'existence des complémentaires santé, certains actes de soins présentent un reste à charge conséquent, incitant certains assurés au renoncement des soins.

C. La réforme 100 % Santé

La réforme « 100 % Santé », appelée anciennement « RAC 0 » a pour but d'offrir – dans le cadre d'un contrat d'assurance maladie complémentaire, individuel ou collectif et selon certaines conditions – le remboursement total des soins. L'objectif de cette réforme peut se résumer à l'aide de la phrase suivante : « bien voir, bien entendre et soigner son hygiène bucco-dentaire ». Cette réforme vise avant tout à lutter contre le phénomène de renonciation aux soins.

1. L'identification d'une source d'inégalité : le renoncement aux soins

Un sondage BVA pour la fondation April[5] d'avril 2018 nous permet ainsi d'avoir un aperçu du niveau de renonciation aux soins des français avant l'application de la réforme 100% Santé. A l'échelle nationale, 3 français sur 4 déclaraient avoir déjà renoncé au moins une fois à se faire soigner. Lorsqu'on s'intéresse plus particulièrement aux raisons de cette renonciation, le « manque de moyen financier » concerne un tiers des répondants.

La renonciation d'accès aux soins pour des raisons financières concernerait ainsi en majorité les populations les plus jeunes (18 – 34 ans) et les personnes sans complémentaire santé. Il en ressort également que 84% d'entre eux y ont renoncé à cause d'un reste à charge trop important et 73% à cause de l'impossibilité d'avancer les frais. Il a également été observé que l'accès aux soins est plus difficile pour les salariés d'entreprises de tailles moyennes du fait d'une complémentaire santé insuffisante. Ainsi, sur les répondants couverts par une complémentaire santé, 25% ont été obligés de renoncer à des soins car la complémentaire ne remboursait pas suffisamment et cela concernait en majorité les plus bas revenus.

Parmi les soins médicaux auxquels les répondants ont renoncé, la renonciation à une prothèse dentaire concerne 28% des répondants, celle concernant un équipement optique (lunettes, lentilles) concerne 25% des répondants et celle relative à un appareillage auditif concerne 5% des répondants.

Ce sondage ne nous permet pas d'identifier en détail les raisons liées à la renonciation d'accès à une prothèse auditive, dentaire ou à un équipement d'optique. Le site du gouvernement[6] nous permet d'obtenir plus de détails sur ce point en particulier : « *parmi les 20% de Français aux revenus les plus bas, près d'un sur cinq d'entre eux renonce à s'équiper en optique, et près d'un sur trois renonce à des soins dentaires, pour des raisons financières.* ».

Les postes de consommation concernés

Cette réforme concerne exclusivement trois postes de consommation bien distincts et dont le reste à charge pour les ménages représentait une partie particulièrement élevée :

- L'optique (reste à charge moyen de 22% sur l'offre optique) ;
- Le dentaire (reste à charge moyen de 43% sur les prothèses dentaires) ;
- Les aides auditives (reste à charge moyen de 53% sur les aides auditives et seul un français sur trois nécessitant un appareillage est appareillé).

La réforme cherche ainsi à prévenir le renoncement aux soins pour ces trois postes en proposant un ensemble de soins de qualité. Cette réforme participe par la même occasion, à éviter l'apparition de complications graves causée par un renoncement de se faire soigner. Du point de vue des ménages, la réforme permet d'effacer le reste à charge après une intervention combinée du régime général et du régime complémentaire de l'assuré. Les contrats de santé dits responsables ont ainsi eu l'obligation

[5] Les français et le renoncement aux soins – Un sondage BVA pour la fondation April

[6] « 100 % santé » : remboursement intégral dans les domaines de l'optique, de l'audiologie et du dentaire.

de proposer des paniers de remboursement « 100% Santé » pour les trois postes de consommation concernés par la réforme.

2. Les tenants et aboutissants de la réforme

La mise en œuvre de la réforme « 100% Santé » a ainsi été implémentée progressivement et s'est achevée en ce début d'année 2021. Suite à la mise en œuvre de cette réforme, les consommateurs ont désormais accès à des soins dentaires, optiques et auditives avec un reste à charge nul. Cette situation se matérialise par le plafonnement des prix devant être pratiqués par les professionnels de santé pour les appareillages et soins de base. L'assuré a tout de même la possibilité de choisir un équipement plus cher – dont les prix restent libres – mais il doit alors s'acquitter le cas échéant d'un reste à charge.

a) Optique

Depuis le 1er janvier 2020, au sein du panier « 100% Santé » en optique, le coût de l'équipement de correction visuel se voit ainsi plafonné en fonction du type de correction :

- Pour des verres unifocaux, le coût de la paire de verres est compris entre 65€ et 235€ ;
- Pour des verres progressifs, le coût de la paire de verres est compris entre 150€ et 340€.

La monture est, quant à elle, plafonnée à hauteur de 30€. L'équipement reste de qualité puisque par exemple, les verres ont l'obligation d'avoir reçu un traitement anti-reflet.

b) Dentaire

Depuis le 1er janvier 2021, au sein du panier « 100% Santé » en dentaire, la réforme concerne un large ensemble de prothèses fixes ou mobiles, aux matériaux et qualité esthétique divers dépendant directement de la localisation de la dent (selon le caractère visible ou non de celle-ci). Ainsi, le prix d'une couronne en céramique est plafonné à 500€ et celui d'un bridge en céramique à 1 465€ pour citer quelques exemples.

c) Aide auditive

Le reste à charge assumé par l'assuré, dans le cadre de l'aide auditive, est le plus important des trois postes de consommation. Depuis le 1er janvier 2021, au sein du panier « 100% Santé » en aides auditives, l'ensemble des appareils d'assistance à l'ouïe est ainsi concerné : contour d'oreille classique, contour à écouteur déporté et intra-auriculaire. Le prix des aides auditives est ainsi plafonné à hauteur de 950€ par oreille pour les personnes âgées de 20 ans et plus et est plafonné à hauteur de 1 400€ pour les moins de 20 ans et les patients atteints de cécité.

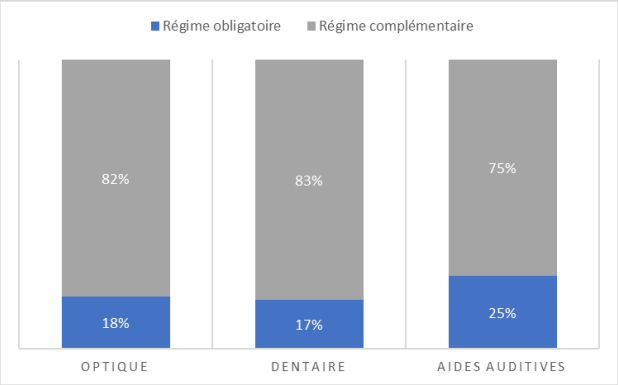


Figure II-8 – Répartition de la prise en charge du panier "100% Santé" entre le régime obligatoire et le régime complémentaire

3. Les premiers impacts chiffrés de la réforme

L'objectif de la réforme « 100% Santé » est donc clairement établi : proposer une offre de soin en dentaire, en optique et en audition qui soit de qualité et sans reste à charge pour l'assuré tout en répondant le mieux possible à ses besoins. Il s'agit d'une réforme nécessitant un investissement très important, chiffré à 1 milliard d'euros à l'horizon 2023 pour le régime obligatoire et les régimes complémentaires[7].

En première approche, nous pouvons imaginer que la consommation sur ces postes augmente nécessairement suite à la réforme. En effet, avec un reste à charge nul, les français disposant d'une complémentaire santé seraient ainsi moins susceptibles de renoncer à ces soins pour des raisons financières.

Un communiqué de presse[8] de Carte Blanche Partenaire, plateforme de services santé, datant de février 2021 nous permet d'avoir un premier aperçu des impacts de cette réforme mais uniquement sur les postes dentaire et optique puisque le remboursement intégral des aides auditives a débuté en janvier 2021. Lorsque nous nous intéressons à la part du panier 100% Santé dans les actes ou équipements pris en charge au sein du réseau Carte Blanche Partenaire en 2020, nous pouvons voir que le panier 100% Santé représente 53% des actes ou équipements en dentaire et 7,2% des actes ou équipements en optique.

Il faut également noter qu'en optique, les consommateurs ont préféré combiner une monture 100% Santé avec des verres du panier libre ou inversement. Ce choix de combiner « Panier 100% Santé » et « Panier Libre » représente 56% des actes et entraîne donc un reste à charge pour le consommateur. Le groupe a également constaté, qu'au sein de son réseau, « ***le reste à charge des bénéficiaires ayant opté pour un achat dans le panier libre a augmenté de 8 millions d'€, passant de 119 millions d'€ en 2019 à 127 millions d'€*** **en 2020***, à cause de la baisse du plafond de remboursement de la monture* [dans le cadre d'un contrat responsable]*, celui-ci passant de 150 € à 100 €* ».

En dentaire, en 2020, le panier 100% Santé a été très sollicité par les assurés pour les dents visibles puisqu'il représente 87% des prises en charge. Pour les dents non-visibles, le taux de recours au panier 100% Santé s'élève à seulement 16%. Cette différence importante s'explique par l'existence d'offres distinctes entre les dents visibles et non-visibles. Pour les dents visibles, le panier 100% Santé intègre des matériaux esthétiquement proches de la couleur naturelle des dents (céramique, zircone) alors que pour les dents non-visibles, le panier 100% Santé ne propose que du métal, couleur qui ne se fond pas naturellement dans la mâchoire. L'esthétisme a donc une forte influence sur le choix du panier chez les consommateurs.

Connue de tous, l'année 2020 a été marquée par une pandémie et des confinements successifs ayant pu empêcher ou retarder l'accès aux soins. Les chiffres de l'année 2020 ne sont pas forcément les plus pertinents surtout pour une réforme aussi récente que la réforme « 100 % Santé » : les études sur les impacts de la réforme méritent d'être continuées avant d'énoncer des conclusions.

[7] « 100% Santé, des soins pour tous, 100% pris en charge – Dossier de presse du mois de juin 2018 »
[8] « Réforme 100% Santé : lorsque le reste-à-charge augmente ! »

III. Présentation du contexte politico-économique : 2020 et le COVID

A. Le contexte particulier de 2020

1. Contexte sanitaire

L'année 2020 restera dans les mémoires comme étant une année sombre, l'année où le monde a été touché de façon indifférenciée par une nouvelle maladie, le Covid-19. Elle fera de l'ombre à l'élection présidentielle américaine de novembre, aux terribles incendies contre lesquels l'Australie se battait depuis août 2019 ou encore à la crise diplomatique entre les États-Unis d'Amérique et l'Iran suite à l'assassinat d'un général iranien.

L'impact inédit de cette épidémie se mesure dans un premier temps par la couverture médiatique qui lui a été consacrée. En France, les sujets liés au Covid-19 ont ainsi représenté 51 % de l'offre d'information globale et 47 % de la durée totale des journaux télévisés sur l'année 2020 d'après l'Institut National de l'Audiovisuel. C'est la première fois qu'un sujet lié à la santé est autant mis en avant par les médias. C'est également la première fois qu'un ministre, et en l'occurrence le ministre de la Santé, passe plus de temps à l'antenne que le Président de la République ou le Premier Ministre. L'année 2020 est donc définitivement placée sous le signe de la Santé.

Même si « le quand et le où » de la pandémie restent encore inconnus à ce jour, les origines du COVID-19 se situeraient dans un marché de fruit de mer à Wuhan, situé dans la région de Hubei en Chine. Les premières personnes infectées l'auraient été entre le 17 novembre et le 1er décembre de l'année 2019, probablement suite à la consommation d'une chauve-souris elle-même contaminée par la maladie.

La maladie se répand rapidement et l'Organisation Mondiale de la Santé déclare *l'état d'urgence de santé publique de portée internationale* le 30 janvier 2020. Malgré cette première alerte, l'OMS revoit rapidement sa position puisque le 11 mars 2020, l'épidémie de coronavirus est qualifiée de pandémie. Un grand nombre de pays ont, à ce moment-là, déjà des cas de Covid-19 sur leur territoire. Rapidement dépassée par la situation sanitaire engendrée par ce virus, de nombreux pays imposent une restriction des déplacements, une fermeture des frontières, l'interdiction des rassemblements et des évènements culturels et sportifs. Ces restrictions seront souvent suivies d'un confinement généralisé sur l'ensemble de leur territoire, forçant alors les populations à rester chez eux le plus possible afin de freiner la propagation du virus.

En plus de l'impact sanitaire de ce virus, les réponses apportées par les divers gouvernements – certes nécessaires – ont entraîné d'importantes perturbations sociales et économiques. Elles ont notamment généré la plus grande récession mondiale depuis la Grande Dépression des années 1930.

Les réactions, souvent prises dans la panique, ont provoqué des déficiences d'approvisionnements généralisées exacerbées par des achats de panique, des pénuries alimentaires engendrées par des perturbations agricoles ainsi qu'une baisse drastique de la consommation de biens considérés comme non essentiels. Cette crise a cependant bien profité à certains acteurs, notamment les géants du Web américain comme Netflix et Amazon. Cette baisse de la consommation a également permis d'observer une diminution des émissions de polluants et de gaz à effet de serre dans des proportions, ceci-dit, toutes relatives puisque rapidement « rattrapées » une fois les confinements terminés.

De malade asymptomatique à malade gravement atteint, le Covid-19 ne touche pas de façon uniforme les individus. Il n'existe pas de « schéma prédéfini » des symptômes de la maladie. Cependant, les symptômes les plus fréquents sont similaires à ceux d'une grippe : fièvre, maux de tête, courbatures, fatigue, toux voire difficultés respiratoires pour les cas les plus graves. Chaque catégorie de la population ne réagit également pas de la même façon. Les complications les plus graves concernent ainsi principalement les personnes âgées et les plus fragiles, présentant des pathologies préexistantes.

Un traitement précis et uniquement dédié au traitement du COVID n'aura jamais été clairement identifié avant l'arrivée des vaccins et aura été surtout sujet à controverse. Comme la plupart des cas de COVID-19 sont bénins, similaires à la grippe, les traitements conseillés cherchent tout d'abord à soulager les principaux symptômes énumérés ci-dessous. Il est ainsi conseillé de prendre du paracétamol mais également de s'assurer de sa bonne hygiène personnelle et d'avoir une alimentation saine.

Fort heureusement, les travaux autour de vaccins se mettent rapidement en place et dès le 21 décembre 2020, la communauté européenne autorise la mise sur le marché du vaccin de Pfizer-BioNTech. Les premiers vaccins seront administrés le 27 décembre, 6 jours après la mise sur le marché. D'autres vaccins seront également autorisés le mois suivant tel que le vaccin de Moderna le 6 janvier 2021, le vaccin d'AstraZeneca le 29 janvier 2021 et le vaccin de Johnson & Johnson (Janssen) le 11 mars 2021.

2. La pandémie et la France

La France n'a malheureusement pas été épargnée par la maladie puisqu'en date du 8 juin 2021, nous recensons en France près de 5,7 millions de cas et près de 110 000 morts depuis le début la pandémie. Sur une note plus positive, près de 28 millions de personnes ont au moins reçu une dose de vaccin, tous vaccins confondus, nous rapprochant d'une immunité collective, indispensable pour retrouver une vie « d'avant crise ».

Les premiers cas de coronavirus identifiés en France datent du 24 janvier 2020 à Bordeaux, même si une étude rétroactive sur des échantillons montre qu'un patient avait probablement déjà contracté le virus aux alentours du 27 décembre 2019, soit près d'un mois avant le premier cas officiellement confirmé. Il s'ensuit la première admission à l'hôpital d'un touriste chinois le 28 janvier 2020 qui décèdera 2 semaines après faisant de ce touriste le premier décès dû au Covid-19 en France et également le premier décès dû au Covid-19 hors d'Asie.

Au vu des signes inquiétants remontés par les hôpitaux de l'est de la France et de l'Île-de-France indiquant que l'épidémie devient incontrôlable, des mesures radicales sont prises au début du mois de mars. Le 12 mars, le Président de la République ordonne la fermeture de tous les établissements d'enseignement à partir du lundi 16 mars et ce jusqu'à nouvel ordre. Le 14 mars, le Premier ministre annonce la fermeture de tous les établissements publics considérées comme « non-essentiels » : seuls les pharmacies, banques, magasins alimentaires, stations-services, bureaux de tabac et de presse sont autorisés à rester ouverts. En parallèle, le ministère de la santé demande la déprogrammation des interventions chirurgicales non urgentes. Finalement, le lundi 16 mars la décision de confiner l'intégralité du pays à compter du lendemain, le mardi 17 mars, est prise. Un confinement généralisé est alors imposé et seuls les trajets nécessaires sont autorisés, tout contrevenant s'exposant à une amende de 135 €, voir 3 750 € et de la prison en cas de récidive. Initialement prévu sur uniquement deux semaines après sa mise en place, le confinement sera prolongé à deux reprises pour finalement se terminer le 10 mai au soir, soit 1 mois et 28 jours après son instauration.

Après une période d'accalmie de la circulation du virus pendant l'été, la rentrée scolaire de septembre voit les principaux indicateurs commencer à se détériorer et des décisions gouvernementales sont donc prises en conséquence.

Le gouvernement commence par procéder à l'application de mesures locales comme celle du 26 septembre qui impose la fermeture des bars et restaurants pendant 2 semaines dans les zones en alerte maximale, c'est-à-dire les zones où le taux d'incidence[9] est supérieur à 250 cas pour 100 000

[9] Le taux d'incidence est évoqué plus longuement dans la partie Taux d'incidence.

habitants, à 100 cas pour 100 000 personnes âgées et les zones où les services de réanimation régionaux accueillent 30 % ou plus de patients atteints par le Covid-19. Dans les zones en alerte renforcée, c'est-à-dire lorsque le taux d'incidence est supérieur à 150 cas pour 100 000 habitants et à 50 cas pour 100 000 personnes âgées, les rassemblements de plus de 10 personnes sont interdits et les bars ont l'obligation de fermer leurs portes à partir de 22h. Les gymnases, les salles de sport, des fêtes et polyvalents sont également fermés. Pour cette deuxième période de l'année, les décisions gouvernementales ne sont plus au niveau national, mais elles suivent une logique régionale.

Comme la situation ne s'améliore pas, le 14 octobre est décidé par le Président de la République un couvre-feu de 21h à 6h. Les conditions d'applications de ce couvre-feu à l'échelle du département sont similaires à celles retenues dans le cadre des zones en alerte maximale. Finalement comme la situation sanitaire continue à empirer, un nouveau confinement est décrété à date du 30 octobre. Ce deuxième confinement se terminera le 14 décembre au soir, soit d'une durée de 1 mois et 16 jours.

Suite à la levée de ce deuxième confinement, un couvre-feu est mis en place entre 20h et 6h à l'échelle métropolitaine sauf le soir du 24 décembre où les déplacements entre 20h et 6h sont tolérés. Afin de prévenir d'un troisième confinement, le gouvernement modifiera les horaires du couvre-feu le 16 janvier 2021 : celui-ci est désormais effectif dès 18h.

Le 22 février 2021, le gouvernement décide d'utiliser une nouvelle approche : un confinement le week-end pour les zones les plus touchées sur le modèle de celui de mars 2020. Le littoral des Alpes-Maritimes et la métropole dunkerquoise seront ainsi les premières zones concernées pour les week-ends du 26 février et du 5 mars.

Cependant, ces mesures ne s'avèrent pas suffisantes et le 20 mars 2021, le gouvernement remet en place un confinement dans seize départements. Une attestation est cette fois-ci nécessaire pour les déplacements au-delà de 10 km de son lieu de résidence. Les déplacements non indispensables dans les autres régions sont également interdits et les commerces non-essentiels sont de nouveau fermés. Le couvre-feu est maintenu mais retardé d'une heure. Il est désormais en vigueur de 19h à 6h.

Néanmoins, la situation sanitaire continue à se dégrader. Un troisième confinement national est alors imposé dès le 3 avril 2021. Les mesures du confinement local du 20 mars s'appliquent désormais à l'ensemble de la France métropolitaine.

Le nouveau confinement sera partiellement modifié un mois plus tard, le 3 mai 2021, et des points d'étapes pour son assouplissement sont décrétés :

- Le 19 mai 2021, le couvre-feu passe de 19h00 à 21h00. Les magasins, les musées, les terrasses, les lieux de sports en plein air et les cinémas pourront rouvrir ;
- Le 9 juin 2021, le couvre-feu est de nouveau repoussé et démarre désormais à 23h. Les cafés et restaurants peuvent de nouveau servir des clients en intérieur avec une limite de 6 convives par table ;
- La fin du couvre-feu, initialement prévue le 30 juin 2021, est finalement avancée de 10 jours, soit le 20 juin 2021.

B. Indicateurs de suivi du Covid-19

Suite à l'ampleur de la pandémie, de nombreux indicateurs de suivi ont été mis en place afin de suivre la progression de la maladie plus finement. Nous avons déjà abordé succinctement un de ces indicateurs lorsque nous avons abordé les zones en alerte maximale et les zones en alerte renforcée : le nombre de nouveaux cas pour 100 000 habitants. Cette partie va nous permettre d'expliciter cet indicateur-ci ainsi que les principaux autres utilisés.

Nous pouvons distinguer trois grandes familles d'indicateurs :

- La circulation virale ;
- La tension hospitalière ;
- La mortalité.

1. La circulation virale

Afin de mesurer la circulation du virus dans le pays, certains indicateurs-clés ont rapidement été mis en place.

a) Nouveaux cas confirmés quotidiens

L'un des premiers indicateurs mis en place et facile à comprendre est celui du nombre de nouveaux cas confirmés quotidiens qui indique le nombre de personnes testées positives au Covid-19 sur la journée. Cet indicateur permet de mesurer – en partie – l'évolution de la circulation du virus en France.

Cependant, cet indicateur possède aussi des limites notamment pour évoluer l'ampleur de l'infection lors de la première vague car il est très sensible à la disponibilité des tests. Ainsi, lors de la première vague où les tests étaient réservés aux cas les plus graves, cet indicateur indiquait que seulement 4 500 personnes étaient infectées par jour. A l'inverse, lors du pic de la seconde vague, cet indicateur indiquait dix fois plus de personnes infectées. Même si la disponibilité des tests s'est progressivement améliorée entre les deux vagues, il n'est pas judicieux de conserver ce seul indicateur afin d'évaluer l'ampleur de la pandémie.

b) Taux d'incidence

Intimement lié au nombre de nouveaux cas confirmés dont il dépend, le taux d'incidence représente le nombre de personnes positives au Covid-19 rapporté à une population de 100 000 habitants. Le taux d'incidence présente exactement les mêmes limites que l'indicateur de nouveaux cas confirmés. Cet indicateur est calculé de la manière suivante :

$$\frac{100\,000 * nombre\ de\ cas\ positif}{Population}$$

Cet indicateur est notamment utilisé pour comparer des zones géographiques entre elles.

Ces deux premiers indicateurs sont calculés quotidiennement.

c) Taux de reproduction effectif

Le taux de reproduction effectif, familièrement appelé « R zéro », représente le taux de reproduction du virus, c'est-à-dire le nombre de personnes qui seront infectées par un individu malade. Cet indicateur se base sur la méthode de Cori et se calcule comme suit :

$$R_{effectif} = \frac{Admission\ Urgence_{7}}{Admission\ Urgence_{7|7}}$$

Il s'agit du rapport entre le nombre d'admission aux urgences sur les sept derniers jours et le nombre d'admission aux urgences sur les sept derniers jours il y a 7 jours. Cet indicateur n'est donc pas statique

et traduit l'évolution de la situation par rapport à la semaine précédente. Il permet de prendre du recul sur les deux taux décrits précédemment et de donner une idée plus précise et concrète de l'évolution de la pandémie dans le pays.

Ainsi, nous pouvons établir trois seuils à partir de cet indicateur.

- $R_{effectif} < 1$, l'épidémie régresse ;
- $R_{effectif} = 1$, l'épidémie stagne ;
- $R_{effectif} > 1$, l'épidémie progresse.

Finalement, même si les indicateurs de la circulation virale restent intéressants à observer, il est nécessaire de les compléter avec notamment la famille d'indicateurs liée à la tension hospitalière.

2. La tension hospitalière

Il s'agit de plusieurs indicateurs de terrain, dépendant directement de la situation réelle dans les hôpitaux ; cette famille d'indicateurs reflète au mieux la situation réelle de la circulation du virus.

a) La tension en réanimation

Le premier indicateur qui nous intéresse est celui de la tension en réanimation. Il représente la pression de l'épidémie sur le système hospitalier. Il est calculé de la manière suivante :

$$\frac{Nombre\ de\ places\ occupées\ en\ réanimation\ pour\ cause\ de\ Covid}{Nombre\ de\ lits\ disponibles\ en\ réanimation}$$

Malgré son intérêt manifeste, cet indicateur souffre d'imperfection. Tout d'abord, le nombre de lits disponibles ne prend pas en compte la création de nouvelles places, ce qui a été le cas au plus fort de la crise du Covid-19. De plus, le numérateur intègre également les malades qui sont admis en unité de soins intensifs et en unité de surveillance continue. Ainsi, cet indicateur a tendance à surestimer la situation réelle en réanimation.

Cependant, cet indicateur reste bien plus pertinent que le taux d'incidence comme nous pouvons le voir dans la figure ci-dessous.

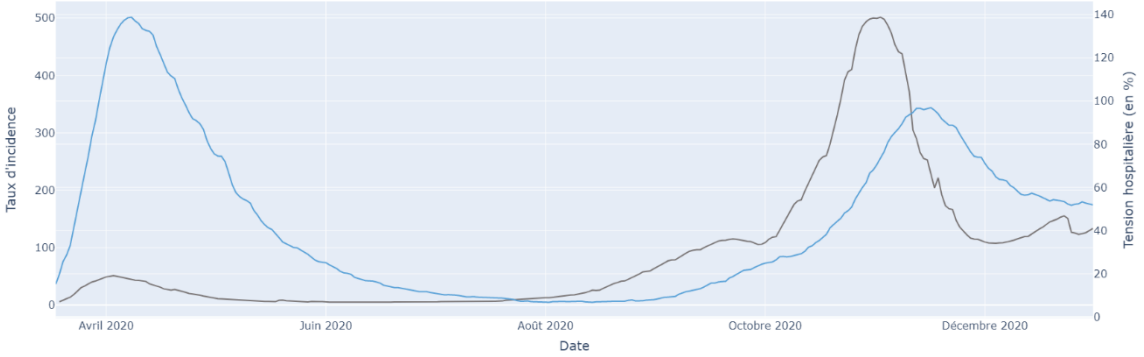


Figure III-1– Comparaison entre le taux d'incidence (courbe grise) et la tension en réanimation (courbe bleue) du Covid-19 en 2020.

Le taux d'incidence reflète particulièrement mal l'ampleur de la pandémie sur le territoire en mars 2020, alors que la tension en réanimation est de 100 % le 30 mars 2020, le taux d'incidence lui est de 44 cas pour 100 000 habitants. A l'inverse, suite à l'amélioration progressive de la disponibilité des tests, le 15 novembre, la tension en réanimation est de 96,5 % pour un taux d'incidence de 253 cas

pour 100 000 habitants. Nous pouvons également remarquer que le pic du taux d'incidence précède d'une dizaine de jour le pic observé en tension en réanimation.

b) Nouvelle hospitalisation et entrées en réanimation

Cet indicateur reflète le nombre de personnes ayant été diagnostiquées positives au Covid-19 et qui ont été admises en hospitalisation ou en réanimation sur les dernières 24 heures. Cet indicateur permet de donner une tendance simple de la situation sanitaire sur le plan hospitalier. Si cet indicateur est en baisse alors l'épidémie circule moins et inversement. Il doit cependant être pris avec suffisamment de recul puisqu'à cause d'un délai d'incubation et d'aggravation du Covid-19 assez important, il reflète la situation d'il y a quinze jours.

3. La mortalité

Possible conséquence malheureuse d'une infection au Covid-19, la mortalité liée à cette infection permet également de suivre la circulation du virus au sein du pays, même si certaines limites doivent être considérées. Tout d'abord, l'approche face à cette maladie a évolué et a été modifiée au fil du temps grâce à l'expérience acquise par le corps médical. Nous pouvons par exemple noter que les soins prodigués à l'hôpital se sont améliorés. En effet, au début de la pandémie, les patients gravement atteints étaient endormis et placés sous respirateur artificiel. Désormais, ils sont mis sous oxygénothérapie, un dispositif bien plus léger pour le système de soins et pour le patient évitant un coma artificiel.

La méthodologie de calcul des décès diffère également puisque trois systèmes différents sont invités à coexister :

- Le décompte à l'hôpital qui regroupe les décès à l'hôpital ;
- Le décompte dans les établissements médico-sociaux qui regroupe les décès dans les EPHAD notamment ;
- Le décompte par les certificats de décès qui regroupe les décès à domicile.

Tout d'abord, comme de nombreuses personnes atteintes par le Covid-19 sont souvent porteuses d'autres pathologies, il est difficile de considérer que le décès est uniquement dû au Covid-19 : il a souvent entrainé l'aggravation des symptômes et la détérioration de la santé de l'individu jusqu'au décès.

De plus, le décompte par certificats de décès nécessite une analyse préalable afin d'en identifier la cause, ce qui implique alors un décalage conséquent entre le décès et l'identification de la cause, rendant difficile sa considération dans les statistiques.

Le décompte des décès en EPHAD a également été mis de côté au début de l'épidémie et n'a pas été mis à jour aussi fréquemment que celui des décès à l'hôpital (les mardis et vendredis de chaque semaine).

A cause de ces facteurs, une étude plus poussée sur la mortalité réelle induite par le Covid-19 sera nécessaire une fois toutes les données remontées et disponibles pour analyse. Nous nous intéresserons à l'étude de la mortalité plus en détail dans le paragraphe suivant.

Impact sur la mortalité : courte étude comparative entre 2018, 2019 et 2020

Dans cette partie, nous allons étudier l'impact du Covid-19 sur la mortalité en France en comparant les années 2018 et 2019 avec l'année 2020. Cette analyse exclut les données de l'année 2021 car incomplètes au moment de l'étude. Les données exploitées sont les données en provenance de l'INSEE sur le nombre de décès quotidiens.

L'année 2020 a ainsi été une année bien plus meurtrière que les deux années précédentes :

- En 2018, nous dénombrons 609 937 décès ;
- En 2019, nous dénombrons 613 409 décès ;
- En 2020, nous dénombrons 668 759 décès, soit une surmortalité de 9 % et de 10% comparé à 2019 et 2018 respectivement.

Cependant, toutes les catégories de la population n'ont pas été concernées de façon uniforme par cette hausse de la mortalité comme nous pouvons l'observer sur le graphique ci-dessous.

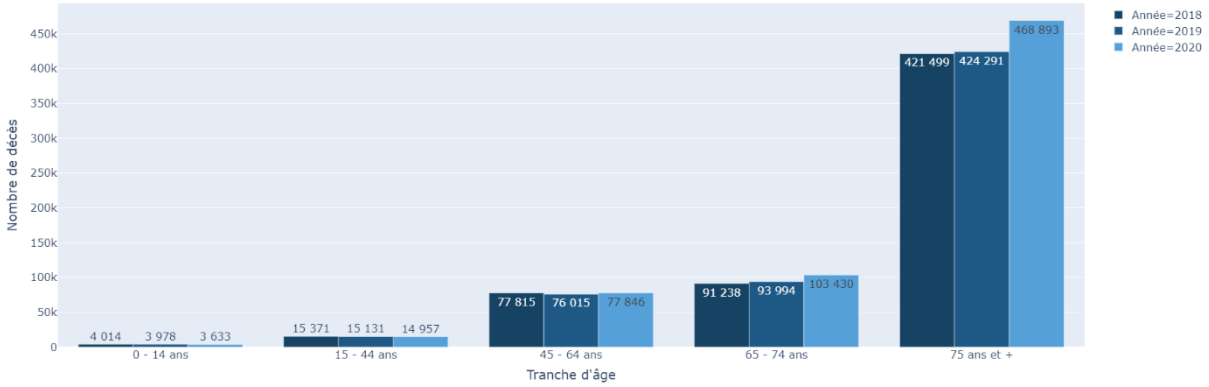


Figure III-2 – Graphique à barres du nombre de décès par tranche d'âge selon l'année de décès

Ainsi, cette hausse de la mortalité a principalement concerné les tranches d'âges les plus élevées. La catégorie des *75 ans et +* a été particulièrement touchée avec une surmortalité observée de 11%. La catégorie des *65 – 74 ans* a subi une hausse de 10%. Les autres tranches d'âges n'ont pas connu de changement significatif à part la tranche d'âge des *0 – 14 ans* qui a connu une sous-mortalité de 9% comparé à 2019.

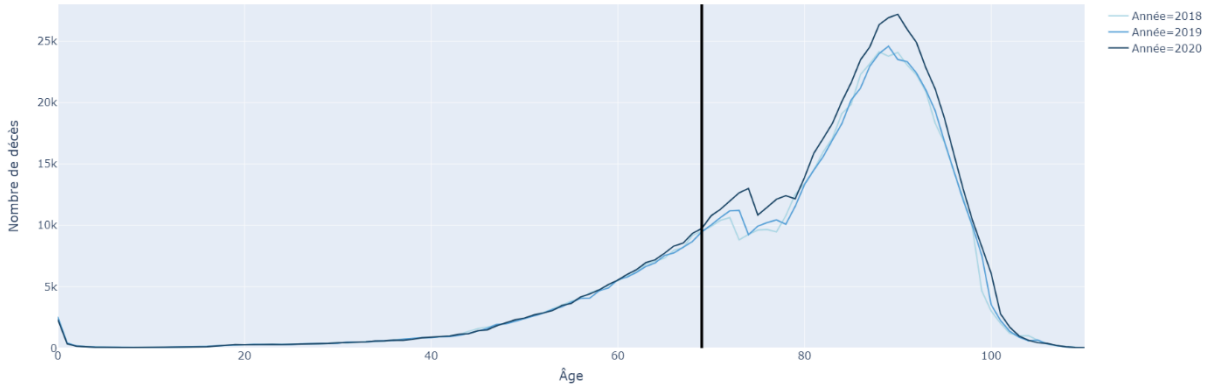


Figure III-3 – Évolution de la mortalité selon l'âge pour les années 2018, 2019 et 2020

Sur ce deuxième graphique, nous pouvons voir que la surmortalité commence à s'observer dès 69 ans (représenté par une ligne verticale dans la figure ci-dessus). Le détail de la surmortalité par sexe et par tranche d'âge est présenté ci-dessous :

Tranche d'âge	**Surmortalité des hommes en 2020**	**Surmortalité des femmes en 2020**
0 - 14 ans	- 12 %	- 5 %
15 - 44 ans	- 1 %	- 1 %
45 - 64 ans	+ 3%	+ 1 %
65 - 74 ans	+ 11%	+ 9 %
75 ans et +	+ 12%	+ 9 %

Figure III-4 – Comparaison de la surmortalité des hommes et des femmes entre 2019 et 2020

Lorsque nous étudions l'évolution de la mortalité en fonction du sexe et de la tranche d'âge entre 2019 et 2020, nous observons plus particulièrement une hausse de la mortalité pour les hommes âgés de plus de 65 ans. Les femmes de plus de 65 ans sont également concernées par cette hausse de la mortalité mais dans des proportions relativement moindres. Une étude plus approfondie sur les causes de décès serait nécessaire afin de pouvoir conclure que les hommes ont plus de risque de décéder du Covid-19 que les femmes.

Dans un même temps, nous observons une nette diminution de la mortalité chez les individus de moins de 44 ans – qu'importe le sexe – et plus particulièrement chez les hommes de moins de 14 ans où la sous-mortalité est de 12 %. Les confinements et couvre-feux successifs les ont moins exposés aux accidents[10].

Malgré tout, ce résultat est à prendre avec du recul. En effet, un enfant de moins de 14 ans a très peu de risques de décéder. Toute variation à la hausse ou à la baisse peut vite entraîner de fortes disparités entre les années. Ainsi, bien que réjouissante, cette variation n'est pas forcément aussi significative : elle ne représente finalement que 267 morts en moins au niveau national alors que la hausse de 12 % pour les hommes de plus de 75 ans et de 9 % pour les femmes de plus de 75 ans se traduit par une hausse des décès de 22 467 et de 22 135 respectivement.

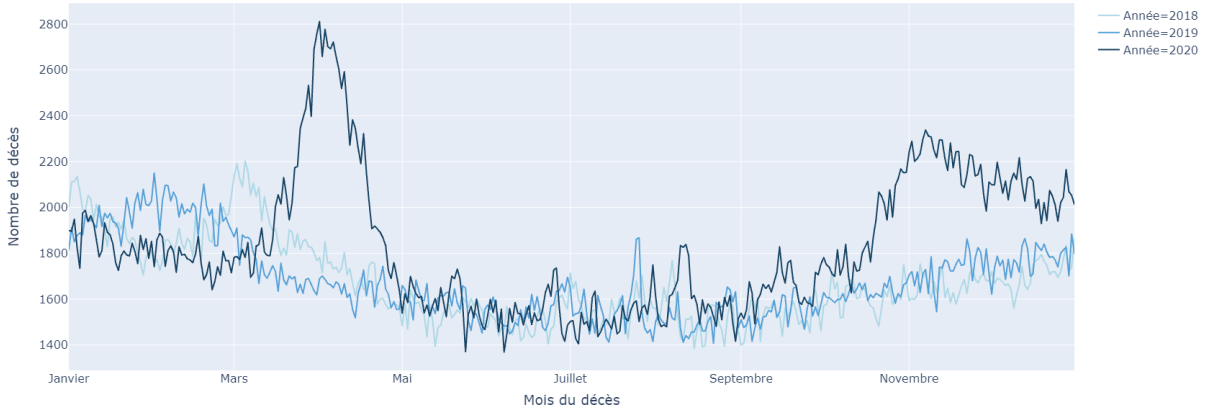


Figure III-5 – Évolution de la mortalité totale au cours des années 2018, 2019 et 2020

10 https://www.franceinter.fr/societe/covid-19-en-2020-le-virus-a-tue-les-plus-de-65-ans-les-confinements-ont-protege-les-jeunes-des-accidents

La courbe de la figure précédente nous permet de clairement identifier les périodes de surmortalité en lien avec les périodes de confinement en 2020. Le premier pic de surmortalité a ainsi lieu au moment du premier confinement, du 16 mars au 1er mai avant de retrouver un rythme similaire à celui de 2018 et 2019. Le deuxième pic de mortalité a lieu bien plus tôt, aux alentours du 10 octobre, soit 3 semaines avant le début du deuxième confinement. Sans être aussi extrême en termes de décès par jour que le premier pic, le second pic de mortalité observé semble s'installer dans le temps. Ainsi, du 10 octobre à la fin de l'année 2020, nous dénombrons en moyenne 2 080 morts par jour, comparé aux 1 720 morts par jour en 2019 et aux 1 670 morts par jour en 2018 sur la même période.

Cette hausse de la mortalité est cependant inégale entre les régions. Ainsi, quand nous observons la mortalité par région en France Métropolitaine sans faire de distinction au niveau de l'âge ou du sexe, de nettes différences commencent à apparaître. Cela est résumé dans le tableau ci-dessous.

Région	Surmortalité sur la période 1er janvier - 31 juin 2020	Surmortalité sur la période 1er juillet - 31 décembre 2020	Surmortalité annuelle
Auvergne-Rhône-Alpes	+ 4%	**+ 26%**	+ 15%
Bourgogne-Franche-Comté	+ 6%	**+ 18%**	+ 12%
Bretagne	0 %	+ 2%	+ 1%
Centre-Val de Loire	+ 3%	+ 9%	+ 6%
Corse	+ 1%	+ 10%	+ 6%
Grand Est	**+ 17%**	+ 10%	+ 13%
Hauts-de-France	+ 7%	+ 14%	+ 11%
Ile-de-France	**+ 27%**	+ 9%	+ 18%
Normandie	+ 1%	+ 10%	+ 6%
Nouvelle-Aquitaine	- 3%	+ 7%	+ 2%
Occitanie	- 1%	+ 11%	+ 5%
Pays de la Loire	+ 2%	+ 6%	+ 4%
Provence-Alpes-Côte d'Azur	+ 1%	+ 15%	+ 8%

Figure III-6 - Étude de la surmortalité par région entre 2019 et 2020

Ainsi, les régions d'Île-de-France et Grand Est sont particulièrement touchées lors du premier confinement avec des taux de surmortalité de + 27 % et + 17 % respectivement. Cependant, la situation change rapidement et lors du deuxième confinement, les régions d'Auvergne-Rhône-Alpes et Bourgogne-Franche-Comté sont largement plus touchées avec des taux de surmortalité de + 26 % et de + 18 % respectivement. Finalement, seule la Bretagne semble avoir été épargnée par la pandémie puisque son taux de surmortalité annuel s'élève à seulement + 1 %, cette hausse restant dans un intervalle d'erreur raisonnable.

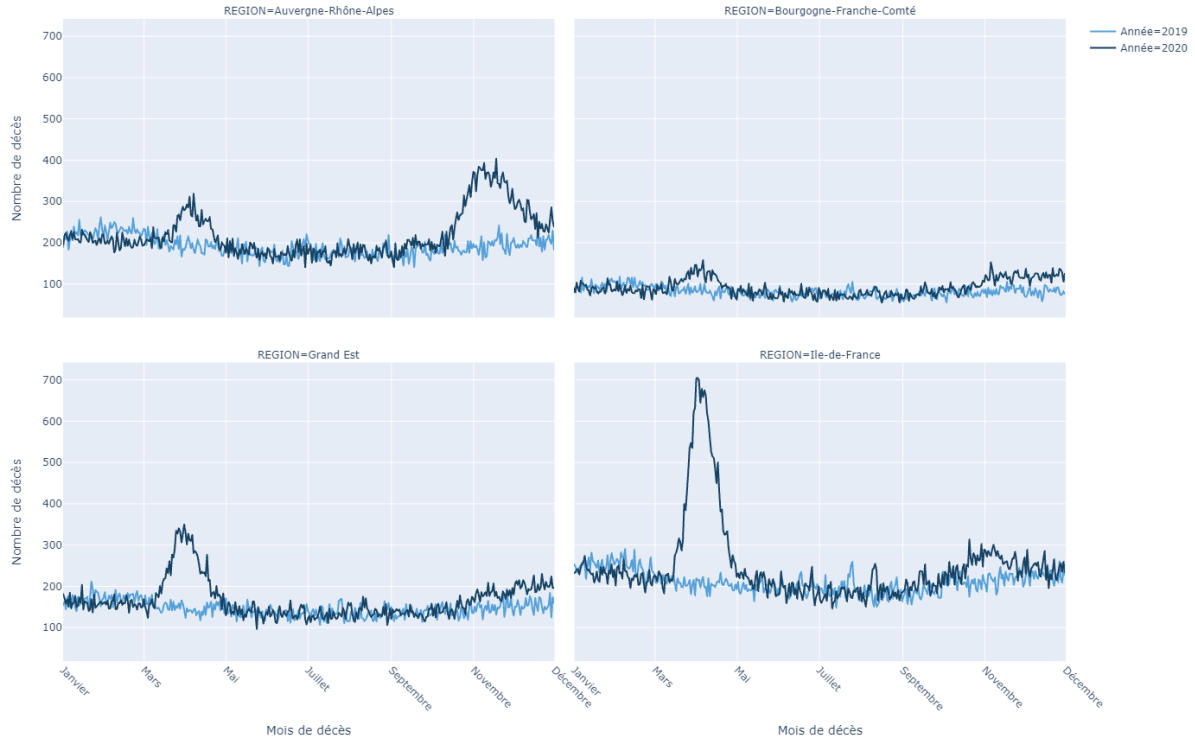


Figure III-7 – Évolution de la mortalité selon certaines régions entre 2019 et 2020

Ainsi, les pics de surmortalité évoqués précédemment s'identifient rapidement à l'aide de la figure précédente où les régions d'Auvergne-Rhône-Alpes, Bourgogne-Franche-Comté, Grand Est et Île-de-France ont été représentées. La région d'Île-de-France a connu une surmortalité importante sur le mois d'avril 2020 comparé à avril 2019 (+ 32 % de surmortalité entre les deux périodes).

C. Impact sur la consommation de santé

1. Le renoncement aux soins

La crise du coronavirus aura eu un impact indéniable sur la consommation Santé des français. La peur d'être contaminé, la peur de déranger les professionnels médicaux et la fermeture de certains cabinets et établissements de santé particulièrement lors du premier confinement font qu'une partie de la population renonce à des soins. Les conséquences de tels renoncements pourront être désastreuses et seront mesurées dans les années à venir.

a) Le report d'actes

Les hôpitaux ont connu des vagues successives de saturation de leur service à cause de l'afflux de patients atteints par le Covid-19, dont le traitement était considéré comme prioritaire. Cette politique assumée a pour effet de déprogrammer une partie des actes chirurgicaux prévus afin de libérer des lits et donc de la place.

D'après une étude du *COVIDSurg Collaborative* à l'échelle mondiale, près de 28,4 millions d'opérations ont été annulées ou reportées pendant les douze semaines du premier pic épidémique de 2020 (de mars à mai environ). La plupart des opérations concernées sont des opérations bégnines (25 638 922 opérations annulées ou reportées pour 28 404 603 opérations prévues, soit 90,2% de taux d'annulation) mais ces annulations ont aussi concerné les opérations pour cancer. Sur les 6 162 311 opérations prévues, 2 324 070 millions ont été annulées, soit un taux d'annulation de 37,7%.

L'étude considère que pour résorber le retard accumulé sur ces douze semaines et sous l'hypothèse que le volume normal d'opérations chirurgicales augmenterait de 20 % après la pandémie, il faudrait environ 45 semaines pour rattraper ce retard, soit plus de 10 mois. Bien évidemment, d'autres vagues ont été observées par la suite, rendant caduque la possibilité de résorber le retard accumulé en 10 mois.

En France, d'après une étude réalisée par la Fédération Hospitalière de France – la FHF, porte-voix des hôpitaux publics – sur l'année 2020, 2,3 millions de séjours prévus n'ont pas pu être honorés dont 1,4 million de séjours en médecine et 900 000 séjours de chirurgie. Comparé à 2019, cela représente une baisse de 12% et de 15% respectivement. Ces chiffres sont particulièrement édifiants lorsqu'ils sont ramenés à la période du premier confinement. Nous pouvons voir que :

- La chirurgie en hospitalisation a baissé de 58 % ;
- La chirurgie en ambulatoire a décru de 80 % ;
- Les coloscopies de diagnostic se sont effondrées de 87 % ;
- Les transplantations rénales ont chuté de 80 % ;

Ces chiffres s'entendent bien évidemment comparé à 2019 sur une période équivalente et promettent potentiellement une bombe à retardement qui pourra avoir plusieurs impacts pour l'avenir : une mortalité accrue et une hausse du coût des traitements pour prise en charge tardive pouvant alors entraîner une hausse de la dépendance.

b) Le retard des traitements

En plus des soins nécessaires mais non prodigués quand cela était nécessaire, les diagnostics tardifs sont synonymes de traitements plus lourds, plus coûteux pour les services publics et pour le patient. Ils participent également à réduire les chances de survie.

L'impact du décalage des soins prodigués à des malades du cancer a été étudié par des chercheurs canadiens et britanniques. Leur analyse a porté sur 34 études permettant de regrouper au total plus de 1,2 million de patients et concernent dix-sept traitements différents. De ces dix-sept traitements,

sept concernent des traitements pour les cancers suivants : vessie, sein, côlon, rectum, poumon, col de l'utérus, tête et cou (ces cancers concernent 44 % de tous les cancers déclarés dans le monde).

La conclusion est froidement cohérente et se résume très simplement : un retard de traitement de quatre semaines diminue les chances de survie du patient qu'importe le type de cancer. Les conclusions de l'étude pour les patients devant bénéficier d'une chirurgie sont résumées dans le tableau ci-dessous :

Type de cancer	Surmortalité en cas de retard de traitement de quatre semaines
Vessie	+ 6 %
Sein	+ 8 %
Colon	+ 6 %
Tête et cou	+ 6 %
Poumon	+ 6 %

Figure III-8 – Pourcentage de surmortalité identifiée en cas de retard de traitement sur certains types de cancer

De plus, même un simple délai de deux semaines augmente le risque de décès. C'est le cas notamment pour le cancer du sein où le risque de décès augmente alors de 4 %.

Cependant, le plus difficile à appréhender sont les complications résultantes d'un dépistage tardif (et par conséquent un traitement tardif). Cela représentera à terme un coût et une charge économique et mentale pour les patients et pour la société en général qui ne pourront être mesurés que dans les années à venir.

Il faut ajouter à cela les problèmes de santé suite à une contamination au Covid-19, appelés communément les patients « Covids longs » et souffrant de migraines, de douleurs corporelles, de frissons, de fièvre, d'essoufflements, de brouillard cérébral, de tachycardie, d'insomnies et d'étourdissements bien longtemps après leur contamination. Une prise de recul est nécessaire afin de mesurer les impacts et la durée des « Covids longs » (5% des personnes ayant été infectées présentent encore au moins un symptôme 6 mois après selon le Ministère des Solidarités et de la Santé). Il est également trop tôt pour savoir s'il s'agit de séquelles irréversibles nécessitant alors un traitement à vie.

2. Rattrapage de la consommation santé après le premier confinement

Comme de nombreuses composantes de l'économie, la consommation en Santé a subi une baisse certaine mais irrégulière dans le temps, entre le premier et le deuxième confinement.

La consommation au gré des confinements : mesurer l'impact du premier confinement

Le cabinet de courtage GEREP a publié[11], le 10 novembre 2020, ses « principales conclusions de l'impact "consommation" sur la complémentaire santé » du premier confinement. L'étude compare ainsi le montant moyen remboursé par bénéficiaire par semaine entre 2019 et 2020. Elle permet de concrétiser l'impact du confinement sur les dépenses. Les trois périodes suivantes sont prises en considération :

	Echelle de temps	Commentaire
Période n°1	Du 19 janvier au 14 mars	Pré-confinement
Période n°2	Du 15 mars au 16 mai	Confinement
Periode n°3	Du 17 mai au 12 septembre	Post-confinement

Figure III-9 – Présentation des périodes considérées dans l'étude

Lorsque nous nous intéressons à toutes les prestations sans faire de distinction, nous observons qu'une légère chute de consommation de 4% était déjà observée sur la première période. Ensuite, sur la période de confinement, un recul de la consommation de 63% est observé et finalement, sur la troisième période de post-confinement, il n'a été observé qu'une hausse de 1% de la consommation, bien loin des taux nécessaires pour parler de « rattrapage de la consommation ».

De plus, cette chute de la consommation est à mettre en contraste en fonction de l'acte observé. Le secteur de l'optique est particulièrement concerné puisqu'une baisse de 35 % est observable sur la première période, baisse qui est *« certainement lié à la mise en œuvre du 100 % Santé* ». Une chute vertigineuse de 85 % de la consommation est ensuite observée sur la seconde période sans que la troisième période ne vienne rattraper le retard accumulé puisque nous pouvons également y remarquer une diminution de 13 % de la consommation. Il faut encore prendre en compte l'impact de la réforme « 100 % Santé » qui vient dissimuler un potentiel réajustement de la consommation.

Contrairement à l'optique, le courtier GEREP a considéré que le dentaire « *sera l'un des rares postes pour lequel nous aurons un rattrapage quasi-complet* ». Sur la première période, une surconsommation de l'ordre de 2 % était observable avant une chute vertigineuse de 85 % lors de la période de confinement suivi par une reprise assez nette de la consommation puisque la surconsommation s'élève à 27 % sur la troisième période.

Le troisième et dernier poste étudié est celui des consultations chez les généralistes et spécialistes. Ce poste est intéressant pour deux raisons : il ne subit pas l'impact de la réforme 100 % Santé et les consultations peuvent être réalisées à distance, par téléconsultation. Ainsi, une légère hausse de 1 % de la consommation est observée sur la 1ère période, suivi d'une baisse de 53 % lors du confinement. Un retour à « la normale » est observé par la suite puisque la baisse n'est plus que de 8 % sur la troisième période.

[11] Les conclusions complètes de l'étude sont disponibles ici.

IV. Présentation des données : la base Open DAMIR

A. Le SNIIRAM

1. Auteurs

Le Système National d'Information Inter-régimes de l'Assurance Maladie abrégé en SNIIRAM a vu le jour en 1999 par la loi de financement de la sécurité sociale pour 1998 (Article L. 161-28-1). Il s'agit d'un entrepôt de données regroupant anonymement les données collectées par :

- L'ensemble des organismes gérant un régime de base d'assurance maladie (régime général, régime agricole, régime des indépendants …) ;
- L'Agence technique de l'information sur l'hospitalisation (Atih) qui récupère les informations relatives à l'activité hospitalière.

La gouvernance du SNIIRAM est assurée par trois gestionnaires :

- L'État ;
- Les régimes d'Assurance Maladie obligatoire ;
- Les professionnels de santé libéraux représentés par l'UNPS.

La gestion du SNIIRAM est assurée par la Caisse nationale de l'Assurance Maladie des travailleurs salariées (CnamTS).

2. Objectifs

La création d'un tel système répond à trois objectifs :

- Connaître les dépenses de l'ensemble des régimes d'assurance maladie selon des critères prédéfinis : zones géographiques (similaires aux grandes régions administratives), nature de dépense, catégorie des professionnels responsables des dépenses ainsi que par professionnel ou établissement ayant prodigué le soin ;
- Pouvoir communiquer aux prestataires de soins des informations pertinentes sur leur activité, leurs recettes et leurs prescriptions ;
- Etudier les impacts de la mise en œuvre de nouvelles politiques de santé publique permettant de contribuer à une meilleure gestion de l'Assurance Maladie ainsi qu'à une amélioration de la qualité des soins.

Cela a également permis d'informatiser, de structurer et de sauvegarder les données venant d'autres systèmes de santé tels que la SNCF, la Mutualité Sociale Agricole ou le Régime Social des Indépendants.

Cet entrepôt de données, sur près de 70 millions d'habitants, a permis de réaliser de nombreuses études autour de la santé, démontrant tout l'intérêt de l'existence d'un tel système de données :

- Sur les parcours de soins des patients ;
- Sur les modalités de prise en charge du cancer ;
- Une étude médico-économique afin d'estimer le coût des traitements du diabète[12].

[12] Il est possible de consulter les études en cliquant sur le lien suivant : ameli.fr

3. Construction de la base

Le SNIIRAM repose sur le regroupement de données sur :

- **Les patients :**
  - Informations individuelles : âge, sexe, commune et département de résidence, date de décès.
  - Informations sur sa prise en charge : bénéfice de la couverture maladie universelle complémentaire (CMU-C), diagnostic de l'affection longue durée (pour la prise en charge en charge à 100%).
- **La consommation de soins en ville :**
  - Toutes les prestations remboursées avec le codage détaillé de la prestation (actes médicaux, biologie, dispositifs médicaux, médicaments) ;
  - Tous les indicateurs de montants (présenté au remboursement, base de remboursement et montant remboursé) ;
  - Détail par date de soins et date de remboursement.
- **La consommation de soins en établissement,** nous pouvons distinguer :
  - Les séjours facturés directement à l'Assurance Maladie : cliniques privées et une partie du secteur médico-social handicap ;
  - Les séjours du PMSI (Programme de Médicalisation des Systèmes d'Information) pour l'ensemble des établissements sanitaires publics ou privés, soins de suite et de réadaptation, hospitalisation à domicile et psychiatrie pour les disciplines médecine-chirurgie-obstétrique.
- **L'offre de soins**, nous pouvons distinguer :
  - La spécialité du prescripteur ainsi que la spécialité ou catégorie de l'exécutant ;
  - Le lieu d'exécution de l'acte ;
  - Le département de prescription et d'exécution ;
  - Le statut conventionnel (libéral) et le statut juridique (établissement).
- **Les pathologies traitées**.

Le SNIIRAM est composé d'une quinzaine de bases de données, proposant 3 grands thèmes d'analyse :

- Les dépenses ;
- L'offre de soins (professionnels de santé et établissements) ;
- Les bénéficiaires et leurs parcours (bénéficiaires / professionnels de santé / actes)

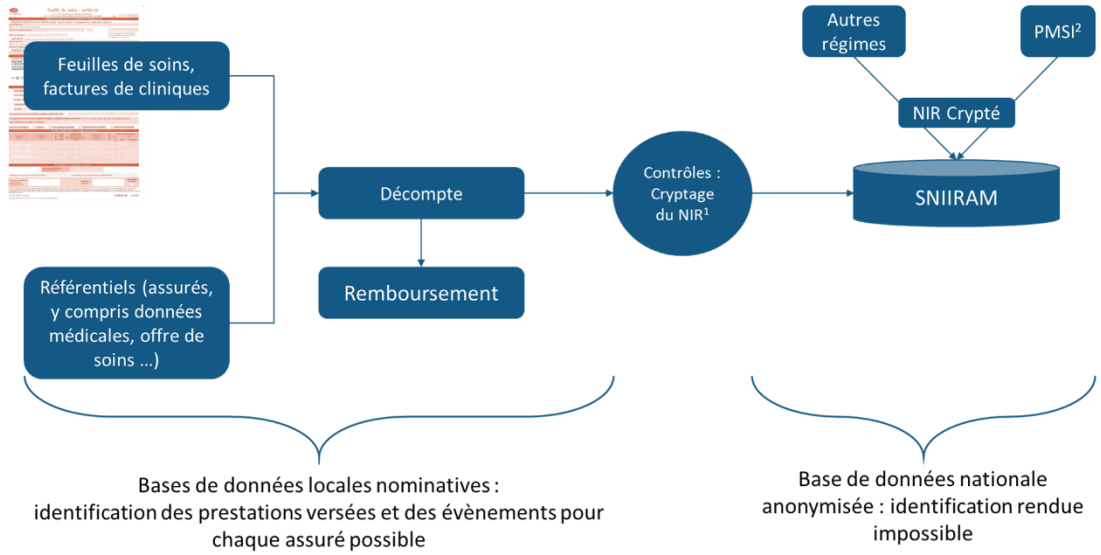
Nous pouvons identifier trois grands niveaux d'accès :

- Le niveau 1 correspond aux données agrégées des bénéficiaires et d'offreurs de soins. Les données sont anonymes et peuvent être conservées indéfiniment. Les données sont accessibles facilement sur les sites du Gouvernement ;
- Le niveau 2 correspond aux données agrégées des bénéficiaires mais individualisées au niveau de l'offre de soins. Les données ont une durée de conservation maximale de 10 ans. Les données sont accessibles à tous excepté l'identification des professionnels de santé ;
- Le niveau 3 correspond aux données individualisées des bénéficiaires et d'offreurs de soins. Les données ont une durée de conservation de 3 ans auquel il faut ajouter l'année en cours si les données sont exhaustives. Pour les données PMSI, cette conservation est de 10 ans et de

20 ans pour les échantillons EGB[13]. Les données sont accessibles à une liste limitée d'organismes et son accès est très contrôlé.

Pour donner une idée de l'ampleur de ce système, quelques chiffres sont disponibles sur le site du SNIIRAM. Nous pouvons ainsi lire « [qu'] *en 2015, 8,9 milliards de feuilles de soins ont été gérées et anonymisées pour alimenter l'entrepôt SNIIRAM* », que « *20 milliards de lignes de prestations* [sont] *disponibles* [pour] *une capacité de stockage égale à 600 téraoctets* ».

Le schéma ci-dessous illustre le processus de création des bases SNIIRAM :



1. NIR : Numéro d'Inscription au Répertoire des personnes physiques. Il s'agit du code alphanumérique permettant d'identifier de façon unique une personne dans le répertoire national d'identification des personnes physique gérée par l'INSEE.
2. PMSI : Programme de Médicalisation des systèmes d'Information

Figure IV-1 – Processus de construction de la base SNIIRAM

4. Contrôle par la CNIL

Même si la gestion de la base est assurée par un organisme public, celle-ci reste encadrée et régulièrement contrôlée par la Commission Nationale de l'Information et des Libertés (CNIL). L'organisme assure ainsi un contrôle permanent sur l'accès au SNIIRAM avec notamment un traçage des accès et des requêtes utilisateurs.

De plus, le NIR des individus est anonymisé avant d'être stocké sur cet entrepôt de données. Les requêtes concernant moins de dix individus et le croisement de données sensibles sont également interdits.

La CNIL a ainsi mis en demeure la CnamTS en 2018 à cause de « *nombreux manquements à la sécurité des données à caractères personnel* ». La CNIL a ainsi relevé que les « *insuffisances en termes de sécurité concern*[aient], *notamment, la pseudonymisation des données, les procédures de sauvegarde, l'accès aux données par les utilisateurs du* [SNIIRAM] *et par des prestataires, la sécurité des postes de travail des utilisateurs du* [SNIIRAM]*, les extractions de données individuelles du* [SNIIRAM] *ainsi que la mise à disposition d'extractions de données agrégées du* [SNIIRAM] *aux partenaires* ».

[13] EGB : Échantillon Généraliste de Bénéficiaires. Il s'agit d'un échantillon permanent représentatif de la population protégée par l'assurance maladie, qu'il ait perçu ou non des remboursements de soins.

B. Open DAMIR

1. Champ d'étude de la base Open DAMIR

La base Open DAMIR est une extraction spécifique du SNIIRAM et regroupe l'ensemble des remboursements effectués par l'Assurance Maladie tous régimes confondus depuis 2009. Cette base de données appartient au premier niveau d'accès. Les données sont agrégées et anonymisées. Les données de la base Open DAMIR sont disponibles au format CSV et regroupées selon les mois de règlement. Afin d'obtenir les données des mois de l'année N, il faut attendre la mise à jour annuelle de la base, qui s'effectue généralement lors du second trimestre de l'année $N+1$.

Chaque ligne correspond à un agrégat d'individu regroupé selon des critères bien précis. Chaque ligne est décrite par 55 variables dont :

- 6 axes d'analyses :
  - Période ;
  - Prestation ;
  - Organisme de prise en charge ;
  - Bénéficiaire des soins ;
  - Professionnel de santé exécutant ;
  - Professionnel de santé prescripteur.
- 4 indicateurs de montant :
  - Total de la dépense ;
  - Base de remboursement ;
  - Montant remboursé ;
  - Dépassement.
- 3 indicateurs de volume :
  - Dénombrement ;
  - Quantité ;
  - Coefficient.

2. Réforme et impact

La base Open DAMIR est restée très similaire dans sa structure depuis sa création en 2009. Une mise à jour a eu lieu en 2015 afin de prendre en compte les nouvelles régions administratives, ce qui a permis de légèrement gagner en précision puisque les axes géographiques sont passés de 9 à 13 zones géographiques distinctes.

Dans le cadre du présent mémoire, nous avons décidé de conserver les années de règlement de 2015 à 2020, soit un historique de consommation de 6 ans, afin de conserver des régions homogènes sur notre période d'étude.

3. Limites de la base Open DAMIR

Malgré la grande exhaustivité de ces bases, celles-ci présentent quelques défauts. Ces défauts ont notamment été identifiés dans des mémoires précédents et également par l'un des groupes de travail de l'Institut des Actuaires. Il est bon de les rappeler afin de pouvoir établir clairement les hypothèses retenues pour la suite de cette étude.

En plus de la difficulté liée à la manipulation d'une telle base de données comme abordée précédemment, nous pouvons également noter que la base Open DAMIR fait l'impasse sur une très grande majorité des dépenses du secteur hospitalier. Loin d'être une limite rédhibitoire, il est impératif de l'avoir en tête afin d'éviter d'être surpris lors de l'analyse du fait du faible poids du poste

hospitalisation dans la dépense totale. Les dépenses en hospitalisation, étant insuffisamment exhaustives, elles ne seront donc pas intégrées dans notre étude.

4. Les variables retenues pour l'étude

Pour la suite de l'étude, les variables retenues sont les suivantes :

Variable de la base	Explication de la variable
AGE_BEN_SNDS	Tranche d'âge du bénéficiaire au moment des soins
BEN_RES_REG	Région de résidence du bénéficiaire
BEN_SEX_COD	Sexe du bénéficiaire
PRS_NAT	Nature de la prestation
FLX_ANN_MOI	Année et mois de règlement
SOI_ANN	Année de soins
SOI_MOI	Mois de soins
PRS_PAI_MNT	Montant de la dépense

Figure IV-2 – Présentation des variables retenues dans la base Open DAMIR pour la suite de l'étude

Trois autres variables ont été créées : « Poste », « REG_ANN » et « REG_MOI ».

Les variables REG_ANN ET REG_MOI ont été calculées à partir de la variable *FLX_ANN_MOI*, qui regroupe l'année et le mois de règlement. Cette variable présente ses modalités sous la forme AAAAMM. Une simple fonction sous Python nous a ainsi permis de dissocier le mois et l'année au sein de deux variables distinctes.

La variable *Poste* a été construite à partir de la variable *PRS_NAT*, qui regroupe les modalités expliquant la nature de la prestation. Un travail préalable d'identification du type d'acte et du poste de consommation correspondant a ainsi été effectué, ce qui a permis d'associer à chaque prestation un type d'acte et le poste de consommation qui lui est associé.

Chaque prestation peut ainsi appartenir à l'un des postes de consommation suivant :

- Soins de ville courants ;
- Pharmacie ;
- Dentaire ;
- Optique ;
- Prothèses auditives ;
- Hospitalisation ;
- Autres.

Ce premier travail d'analyse nous permet ainsi de conserver uniquement les postes de consommation que nous souhaitons étudier. C'est pourquoi, nous nous intéresserons uniquement à la dépense santé globale selon les postes de consommations : **soins de ville courants**, **pharmacie**, **optique**, **dentaire** et **prothèses auditives**.

De plus, sur un faible nombre de lignes, l'âge, le sexe et la région sont indiqués comme étant « *inconnu* ». Leur poids étant considéré négligeable en termes de prestations, ces lignes ont été supprimées et ne sont donc pas prises en compte dans l'étude. De même, l'étude porte uniquement sur la France métropolitaine. Les lignes avec la modalité « *Régions et Départements d'outre-mer* » ont été retirées.

5. Manipulation de la base Open DAMIR

De par sa structure et au regard de la population concernée par ces bases, les bases de données Open DAMIR représentent donc un agrégat de données conséquent. Ainsi, chaque mois de règlement représente environ cinq giga-octets de données. Il est donc impensable de vouloir manipuler cette base de données à l'aide d'outils traditionnels sur une configuration d'ordinateur portable standard. La taille de cette base de données représente ainsi l'une de ses principales limites et peut représenter un frein à son exploitation.

Nous allons donc spécifiquement nous intéresser aux outils et techniques qui ont été mis en place afin de pouvoir manipuler cette base de données et d'extraire les données pertinentes pour les parties suivantes.

a) Outils choisis

La base Open DAMIR que nous souhaitons étudier peut-être stockée dans l'espace de stockage de la machine mais ne peut pas être stockée dans la mémoire RAM de l'ordinateur. Cette principale contrainte nous a donc invité à chercher de nouvelles techniques décrites par la suite.

(a) Puissance de la librairie « *Dask* » sous Python

L'outil traditionnel de manipulation de bases de données sous Python est la librairie *Pandas*. Cependant, de par son fonctionnement et sa conception, son utilisation pour manipuler les bases Open DAMIR s'est avérée insuffisante. Il a donc été nécessaire de faire appel à une autre librairie, toujours sous Python. Cette librairie a été spécifiquement développée pour manipuler les bases de données de taille moyenne.

Pour mieux comprendre le fonctionnement de cette librairie, nous allons commencer par comprendre succinctement le fonctionnement de la librairie *Pandas*[14]. Lorsque nous indiquons à la librairie de lire un fichier, *Pandas* va le parcourir et enregistrer son intégralité dans la mémoire RAM de l'ordinateur. Lors de l'application d'opération complexe, tous les traitements seront réalisés au sein de cette même mémoire à l'aide d'un seul cœur du processeur. Ces traitements seront exécutés selon l'ordre précisé par l'utilisateur.

A l'opposé de *Pandas*, *Dask*[15] va essayer de réduire au maximum l'utilisation de la mémoire RAM. Tout d'abord, le fichier brut n'est jamais stocké dans la mémoire RAM. Lorsque nous indiquons à Dask de lire un fichier, il conserve en mémoire uniquement son chemin d'accès et son architecture (le nom des colonnes et le type de variables, obtenus à partir d'un échantillon de la base). Ensuite, la librairie incite l'utilisateur à écrire toutes les opérations qu'il souhaite réaliser sur cette base de données. *Dask* ne commence les calculs qu'au moment où l'utilisateur indiquera un « *.compute()* » lors de la dernière opération. Dask subdivise le travail sur tous les cœurs du processeur afin d'en accélérer le traitement. Ce processus est appelé parallélisation ou distribution des calculs. Les résultats obtenus suite à ce traitement sont ensuite entreposés temporairement dans le stockage de la machine. Cette approche permet d'améliorer les temps de calculs puisque Dask tente d'optimiser les traitements en les effectuant dans l'ordre le plus optimale possible.

[14] La description proposée n'est pas propre à la librairie *Pandas*. Il s'agit du fonctionnement traditionnel effectué par les outils d'analyse des petites bases de données et que nous pouvons retrouver sous R ou des tableurs.
[15] La documentation de Dask est disponible ici.

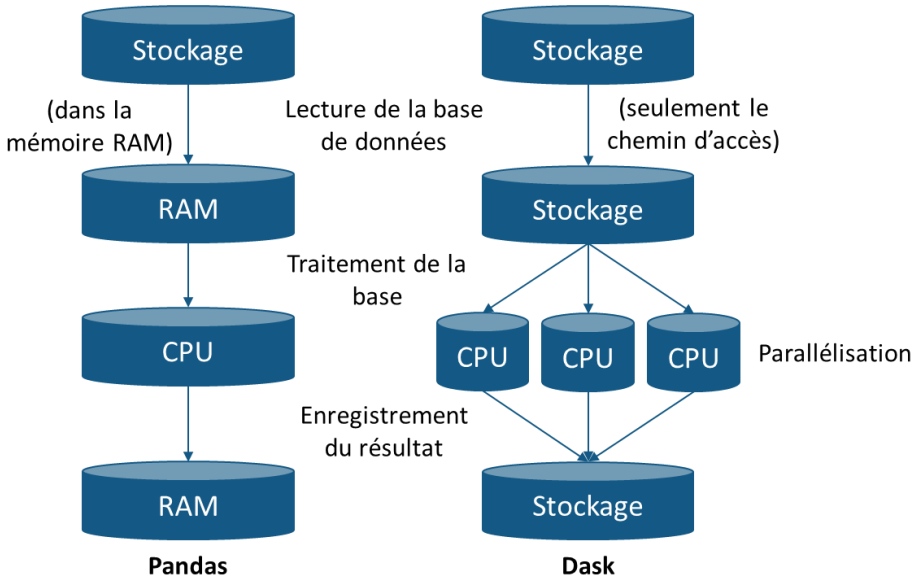


Figure IV-3 – Observation des différences au niveau du processus de manipulation de bases de données entre Pandas à gauche et Dask à droite

Ainsi, *Dask* possède deux avantages en comparaison avec *Pandas* : la mémoire RAM est moins sollicitée et les calculs sont parallélisés, permettant de manipuler les bases de données OPEN DAMIR dans des temps raisonnables.

Même si le fonctionnement de Dask en arrière-plan diffère de Pandas, l'architecture des fonctions de cette librairie sont très similaires à celles sous Pandas, mis à part l'ordre spécifique de démarrage des calculs via la ligne de code « *.compute()* » .

(b) Google Colab

Google propose l'accès à ses serveurs via le service Google Colaboratory[16], souvent raccourci en Colab et permettant à n'importe quel utilisateur – sous réserve de posséder une adresse mail Google – d'écrire, d'exécuter et de partager du code Python dans son navigateur. Colab repose sur l'architecture du Notebook Jupyter et a donc un fonctionnement similaire. La seule différence étant que le code écrit s'exécute sur les serveurs de Google et non sur l'ordinateur local.

L'utilisation des serveurs Google présente également l'avantage d'avoir accès à une machine plus puissante. Comme l'exécution se réalise à distance, il est toujours possible d'utiliser l'ordinateur local sans ralentissement pour d'autres tâches annexes. De plus, cette configuration est exclusivement dédiée au fonctionnement du Notebook, contrairement à un ordinateur local qui doit également faire tourner des services annexes pour fonctionner correctement. Ainsi, nous avons accès à quasiment 100% de la puissance de cette machine virtuelle.

Dans le cadre de ce mémoire, Google Colab est utilisé dans le seul but de manipuler des bases de données. Etant donné que les données Open DAMIR sont déjà anonymisées et à libre disposition, nous ne procédons pas à des éventuels partages de données confidentielles. De plus, les bases de données et les différents notebooks développés, peuvent être téléchargés pour une utilisation locale.

[16] Google Colaboratory est accessible gratuitement ici.

b) Les étapes d'amincissement de la base Open DAMIR

Pour obtenir une base de données regroupant les informations qui ont été utilisées dans ce mémoire, nous avons commencé par importer chaque mois de consommation étudié dans Google Drive, correspondant à 72 mois de règlement de 2015 à 2020. A l'aide de *Dask*, nous avons ensuite traité chaque mois individuellement afin de conserver les colonnes précédemment citées[17] et de regrouper les lignes qui nous intéressent pour la suite de notre analyse. Nous avons ensuite concaténé chaque mois de règlement au sein de son année de règlement correspondant avant de pouvoir concaténer chaque année de règlement au sein d'un seul et même fichier. Ce premier travail d'analyse nous a ainsi permis de réduire nos trois cent soixante giga-octets de données en un fichier unique d'environ trois giga-octets, optimisé au format « parquet »[18] et sur lequel les manipulations deviennent bien plus aisées.

[17] Les colonnes retenues sont : AGE_BEN_SNDS, BEN_RES_REG, BEN_SEX_COD, PRS_NAT, FLX_ANN_MOI, SOI_ANN, SOI_MOI et PRS_PAI_MNT

[18] Le format parquet est un format orienté manipulation de bases de données. Il présente l'avantage d'être open source et de fournir des schémas efficaces de compression et d'encodage des données avec des performances améliorées pour traiter des données complexes en masse. Source : Wikipedia.

C. Statistiques descriptives sur les données de la base OPEN DAMIR sur les années étudiées

Nos statistiques descriptives porteront donc sur les six années de consommation disponibles (de 2015 à 2020) et se feront du plus général au plus spécifique. Elles portent sur les données connues et communiquées en date du 31 décembre 2020. Afin d'effectuer nos statistiques descriptives sur une base commune, nous considérons uniquement les prestations effectuées et payées lors de l'année N, sans distinction de date de survenance.

1. Évolution globale de la consommation

Pour commencer, regardons l'évolution de la consommation de 2015 à 2020 selon les cinq postes de consommations précisées précédemment.

	2015	2016	2017	2018	2019	2020	Evolution 2019 / 2020
Soins de ville courants	41,86	44,21	45,86	47,25	48,78	47,78	**- 2,1 %**
Pharmacie	32,76	32,89	32,93	33,27	33,79	34,34	**1,7 %**
Dentaire	9,69	9,99	10,15	10,42	10,85	10,17	**- 6,2 %**
Optique	5,66	5,77	5,73	6,04	6,53	5,80	**- 11,1 %**
Prothèses auditives	1,72	1,84	1,93	2,04	2,10	2,05	**- 2,1 %**
Total	**91,69**	**94,70**	**96,61**	**99,02**	**102,05**	**100,16**	**- 1,9 %**

Figure IV-4 – Consommation annuelle par poste de consommation entre 2015 et 2020 (en milliards d'euros)

Nous pouvons remarquer – sur les années étudiées – que l'année 2020 est la première année où la dépense globale est en baisse en comparaison avec l'année précédente. Certes relativement faible au global, cette baisse se doit d'être observée en détail afin d'identifier toute tendance de fond.

Nous pouvons notamment remarquer que le poste *Optique* a été particulièrement touché par cette baisse de la consommation : la dépense globale a chuté de près de 11% en un an. A l'inverse, le poste Pharmacie a lui légèrement augmenté de 1,64 %.

Les consommations annuelles par tranche d'âge sont présentées ci-dessous.

	2015	2016	2017	2018	2019	2020	Evolution 2019/2020
0-19 ANS	9,72	10,18	10,35	10,77	11,21	10,24	**- 8,7 %**
20 - 29 ANS	5,36	5,56	5,57	5,77	6,02	6,01	**- 0,2 %**
30 - 39 ANS	7,98	8,33	8,49	8,71	8,97	8,70	**- 3,0 %**
40 - 49 ANS	10,85	10,95	10,95	10,98	11,09	10,64	**- 4,1 %**
50 - 59 ANS	14,32	14,55	14,73	14,89	15,07	14,56	**- 3,4 %**
60 - 69 ANS	16,26	16,92	17,08	17,25	17,45	17,15	**- 1,7 %**
70 - 79 ANS	12,99	13,54	14,33	15,15	16,11	16,57	**2,9 %**
80 ANS ET +	14,21	14,68	15,11	15,50	16,13	16,30	**1,1 %**

Figure IV-5 – Consommation annuelle par tranche d'âge entre 2015 et 2020 (en milliards d'euros)

Lorsque nous nous intéressons à l'évolution de la consommation par tranche d'âge, nous pouvons observer une baisse sur la majorité des tranches d'âges et notamment celle des 0-19 ans : la dépense globale de cette tranche d'âge a baissé de 8,7%. A l'inverse, les tranches d'âges les plus élevées ont vu leur consommation globale augmenter.

La vision de l'évolution de la consommation globale par sexe de 2015 à 2020 ne vient pas perturber notre analyse pour le moment. Qu'importe le sexe, la consommation est en baisse et vient conforter notre tendance de fond.

	2015	2016	2017	2018	2019	2020	Evolution 2019/2020
FEMININ	51,20	52,85	53,88	55,13	56,80	55,43	**- 2,4 %**
MASCULIN	40,50	41,85	42,73	43,90	45,25	44,73	**- 1,2 %**

Figure IV-6 – Consommation annuelle par sexe entre 2015 et 2020 (en milliards d'euros)

Lorsque nous nous intéressons à cette même consommation en fonction de la région, nous retrouvons les mêmes conclusions que celles établies précédemment : cependant, certaines régions ont subi une baisse plus importante.

	2015	2016	2017	2018	2019	2020	Evolution 2019/2020
Aquitaine-Limousin-Poitou-Charentes	8,52	8,81	9,02	9,24	9,51	9,37	**- 1,5 %**
Auvergne-Rhône-Alpes	10,41	10,84	11,14	11,49	11,80	11,51	**- 2,5 %**
Bourgogne-Franche-Comté	3,70	3,83	3,89	4,00	4,09	3,98	**- 2,8 %**
Bretagne	4,22	4,35	4,44	4,56	4,68	4,62	**- 1,3 %**
Centre-Val de Loire	3,31	3,36	3,45	3,55	3,64	3,55	**- 2,5 %**
Grand Est	9,46	9,83	9,94	10,17	10,44	10,05	**- 3,7 %**
Hauts-de-France - Nord-Pas-de-Calais-Picardie	8,79	9,09	9,23	9,38	9,71	9,57	**- 1,5 %**
Ile-de-France	15,73	16,24	16,52	16,93	17,58	17,25	**- 1,9 %**
Languedoc-Roussillon-Midi-Pyrénées	9,19	9,48	9,68	9,91	10,24	10,09	**- 1,4 %**
Normandie	4,28	4,38	4,46	4,58	4,69	4,66	**- 0,8 %**
Pays de la Loire	4,38	4,50	4,62	4,76	4,90	4,87	**- 0,6 %**
Provence-Alpes-Côte d'Azur et Corse	9,70	10,00	10,22	10,45	10,76	10,64	**- 1,1 %**

Figure IV-7 – Consommation annuelle par région entre 2015 et 2020 (en milliards d'euros)

C'est notamment le cas pour la région *Grand Est* qui a vu sa consommation globale chuter de 3,6% entre 2019 et 2020. À l'inverse, les régions *Normandie* et *Pays de la Loire* ont observé les plus faibles baisses de consommation globale.

2. Évolution par poste de consommation

Nous allons maintenant nous intéresser à l'évolution des dépenses par poste de consommation. Comme nous avons pu le voir dans la figure IV-4, l'évolution de la consommation entre 2019 et 2020 n'est pas uniforme selon le poste de consommation.

Cette seconde partie sera l'occasion d'analyser la consommation globale en fonction du mois de consommation. Nous allons également définir quatre périodes afin d'effectuer une comparaison plus fine entre 2019 et 2020 :

- Période n°1 : de janvier à février correspondant au pré-confinement ;
- Période n°2 : de mars à mai correspondant au premier confinement ;
- Période n°3 : de juin à septembre correspondant à la période entre les deux confinements ;
- Période n°4 : d'octobre à décembre correspondant au deuxième confinement.

Chaque période jouera un rôle dans notre analyse. La période n°1 représente la période pré-Covid. Elle sera utilisée comme base de comparaison par rapport à l'année précédente. La période n°2 coïncide avec la période du premier confinement. La période n°3 évoque la période d'accalmie, celle où la vie a retrouvé un cours normal. La période n°4 correspond à la période du deuxième confinement. A partir de la période n°2, nous pourrons identifier les impacts de la crise sanitaire sur les dépenses en santé.

	2019	2020	Evolution 2019/2020
Période n°1	24 695,23	25 297,22	**2,4%**
Période n°2	37 425,99	27 644,34	**- 26,1%**
Période n°3	44 978,20	48 762,97	**8,4%**
Période n°4	30 882,49	33 332,11	**7,9%**
Total	**137 981,91**	**135 036,64**	**- 2,1%**

Figure IV-8 – Consommation globale entre 2019 et 2020 sur les périodes prédéfinis (en millions d'euros)

Au global, nous pouvons déjà dire que le premier confinement a eu un réel impact sur la consommation santé des français, qu'importe le poste de consommation auquel nous nous intéressons. A l'inverse, le deuxième confinement ne semble pas avoir influé négativement sur cette consommation. Cela est en raccord avec la stratégie gouvernementale mise en place au fur et à mesure que les appréhensions et interrogations liées à ce nouveau virus se sont dissipées.

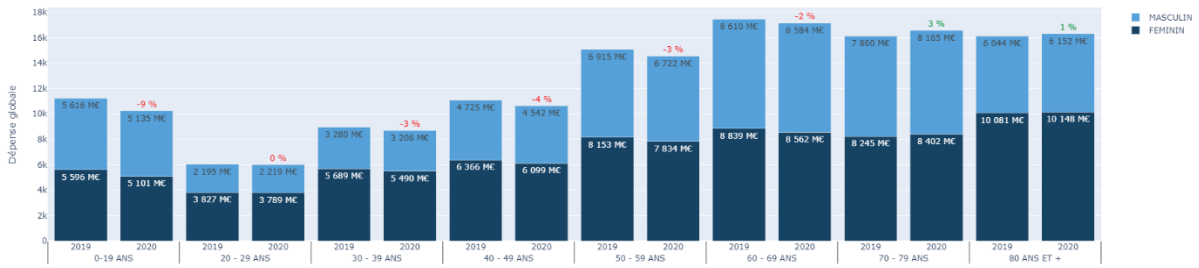


Figure IV-9 – Consommation par tranche d'âge et sexe entre 2019 et 2020 (en millions d'euros)

Le graphique ci-dessus représente les dépenses en 2019 et 2020 par tranche d'âge et par sexe, il permet d'étudier l'évolution de la consommation totale (tous postes de consommation confondus).

A la suite un exemple de lecture : en prenant la tranche d'âge *40 – 49 ans*, nous constatons que la consommation des femmes en 2019 s'élève à 6 366 M€ et à 6 099 M€ en 2020. La consommation des hommes s'élève quant à elle, à 4 725 M€ en 2019 et 4 542 M€ en 2020. La consommation de cette tranche d'âge sans distinction de sexe a donc baissé d'environ 4%. Nous retrouvons bien le pourcentage indiqué pour cette tranche d'âge dans la figure IV-5.

Pour la suite, chaque sous-partie montre les évolutions des postes de consommation étudiés. Chacune est constituée de deux graphiques et d'un tableau illustrant principalement l'évolution de la consommation entre 2019 et 2020. Nous mettrons particulièrement en avant dans nos graphiques les années de règlement 2019 et 2020, en bleu et en rouge respectivement.

a) Soins de ville courants

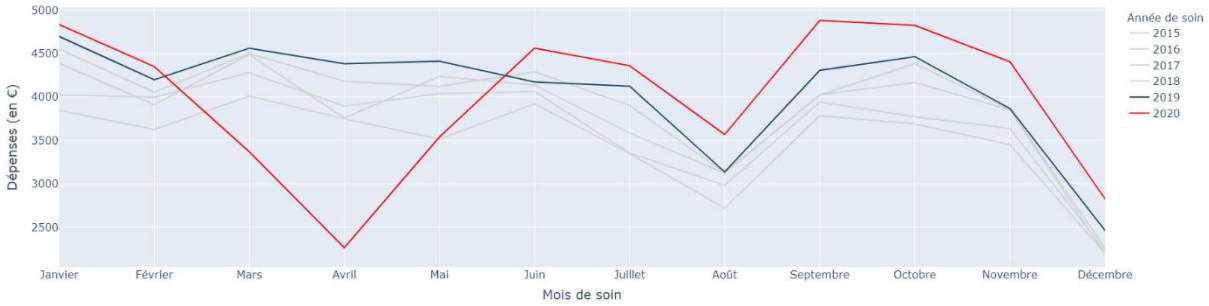


Figure IV-10 – Consommation du poste Soins de ville courants selon le mois de règlement (en millions d'euros)

	2019	2020	Evolution 2019/2020
Période n°1	8 895,06	9 184,07	**+ 3,3 %**
Période n°2	13 355,10	9 170,39	**- 31,3 %**
Période n°3	15 739,89	17 369,44	**+ 10,4 %**
Période n°4	10 794,25	12 057,46	**+ 11,7 %**
Total	**48 784,30**	**47 781,35**	**- 2,1 %**

Figure IV-11 – Consommation du poste Soins de ville courants selon les périodes prédéfinies (en millions d'euros)

Le poste de consommation *Soins de ville courants* a vu sa consommation globale se réduire légèrement entre 2019 et 2020 mais cette baisse n'a pas été uniforme sur l'année comme nous pouvons le voir sur les deux figures précédentes En légère hausse sur la première période, la consommation s'est rétractée de plus de 30 % sur la deuxième période. Nous observons également un rattrapage de la consommation sur la troisième et quatrième période.

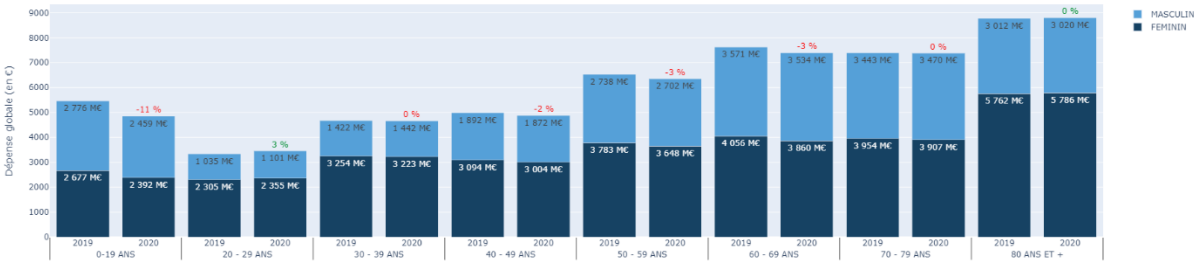


Figure IV-12 – Consommation du poste Soins de ville courants entre 2019 et 2020 selon la tranche d'âge et le sexe (en millions d'euros)

Sur toute l'année, c'est la tranche d'âge des *0 – 19 ans* qui a subi la plus forte des contractions : la consommation en *Soins de ville courants* s'est réduite de 11 %. A l'opposé, la consommation des *70 ans et +* s'est équilibrée sur l'année tandis que celle des *20 – 29 ans* a légèrement augmenté.

b) Pharmacie

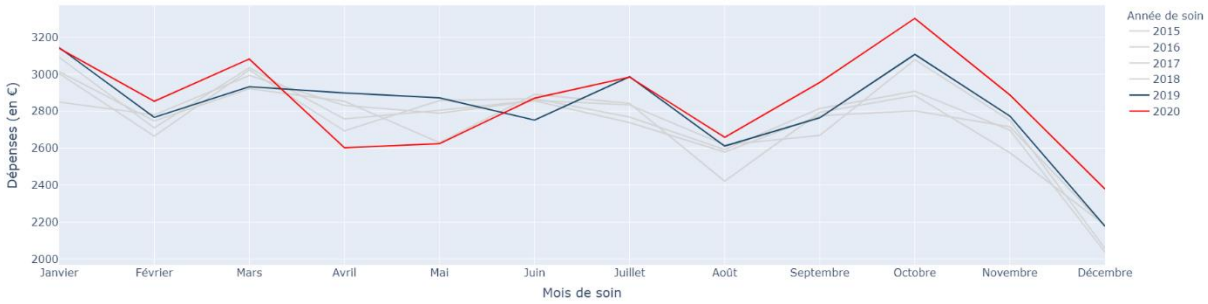


Figure IV-13 – Consommation du poste Pharmacie selon le mois de règlement (en millions d'euros)

	2019	2020	Evolution 2019/2020
Période n°1	5 911,44	5 993,95	**+ 1,4 %**
Période n°2	8 703,65	8 309,29	**- 4,5 %**
Période n°3	11 115,57	11 470,46	**+ 3,2 %**
Période n°4	8 059,13	8 569,46	**+ 6,3 %**
Total	**33 789,78**	**34 343,16**	**+ 1,6 %**

Figure IV-14 – Consommation du poste Pharmacie selon les périodes prédéfinies (en millions d'euros)

Le poste de consommation *Pharmacie* est le seul de notre liste ayant augmenté entre 2019 et 2020. Cette hausse n'est pas uniforme sur l'année : la baisse de la consommation sur la deuxième période est finalement toute relative puisque les pharmacies n'ont pas été obligés de fermer lors des confinements. Nous pouvons supposer que la baisse observée serait due à une population qui a cherché à se protéger du virus en limitant au maximum ses déplacements et contacts avec l'extérieur. Dans un second temps, nous pouvons voir que la consommation était en hausse sur les deux dernières périodes. Nous pouvons supposer que ce rattrapage est dû en partie à la mise sur le marché de test PCR et antigénique venant gonfler la consommation.

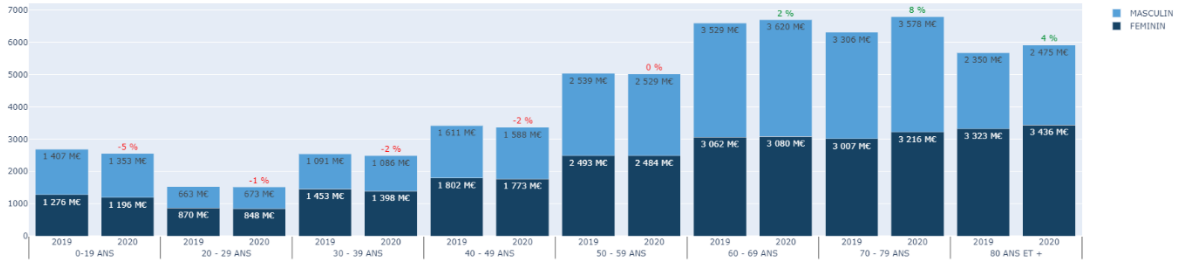


Figure IV-15 – Consommation du poste Pharmacie entre 2019 et 2020 selon la tranche d'âge et le sexe (en millions d'euros)

Sur toute l'année, les personnes de moins de 59 ans ont diminué leur consommation en pharmacie, alors que ceux ayant plus de 60 ans ont consommé plus de pharmacie en 2020.

c) *Dentaire*

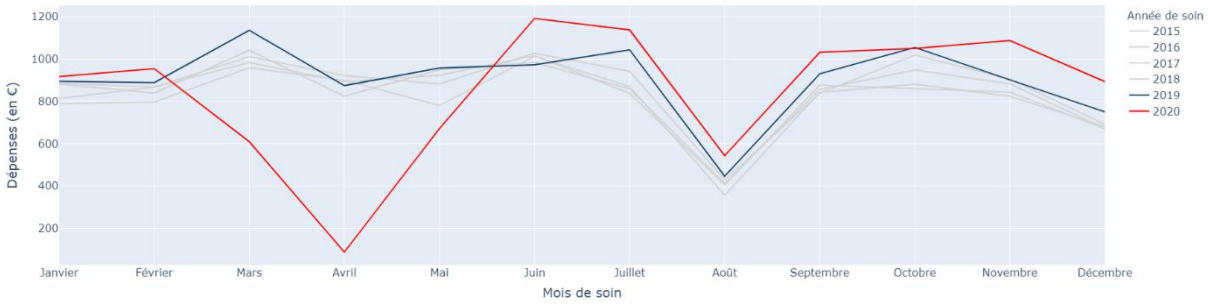


Figure IV-16 – Consommation du poste Dentaire selon le mois de règlement (en millions euros)

	2019	2020	Evolution 2019/2020
Période n°1	1 782,98	1 871,00	**+ 4,9 %**
Période n°2	2 967,52	1 367,72	**- 53,9 %**
Période n°3	3 392,87	3 904,66	**+ 15,1 %**
Période n°4	2 706,20	3 030,83	**+ 12,0 %**
Total	**11 245,31**	**10 574,34**	**- 6,2 %**

Figure IV-17 – Consommation du poste Dentaire selon les périodes prédéfinies (en millions d'euros)

La dépense du poste de consommation *Dentaire* a été particulièrement sensible à la crise du Covid-19 comme nous pouvons le voir dans les deux figures précédentes. La consommation fut très basse en avril 2020, en cohérence avec la fermeture des cabinets dentaires du premier confinement. Au global, nous pouvons voir que la consommation était en hausse au début de l'année et qu'un rattrapage certain s'est effectué après la deuxième période.

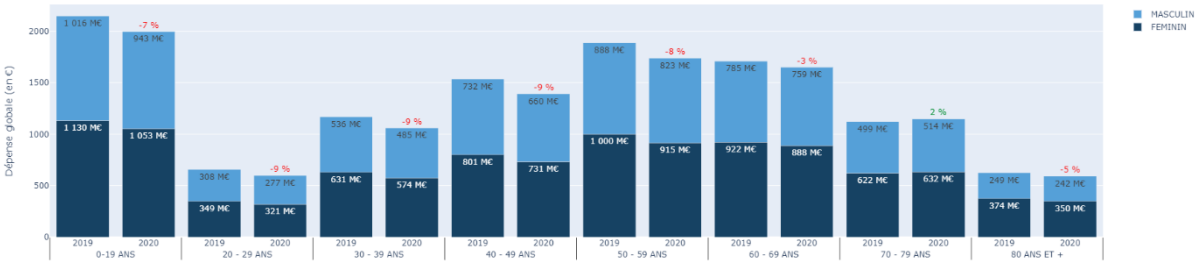


Figure IV-18 – Consommation du poste Dentaire entre 2019 et 2020 selon la tranche d'âge et le sexe (en millions d'euros)

Sur toute l'année, toutes les tranches d'âge, sauf la tranche d'âge 70 – 79, présente une diminution de dépense en dentaire.

d) Optique

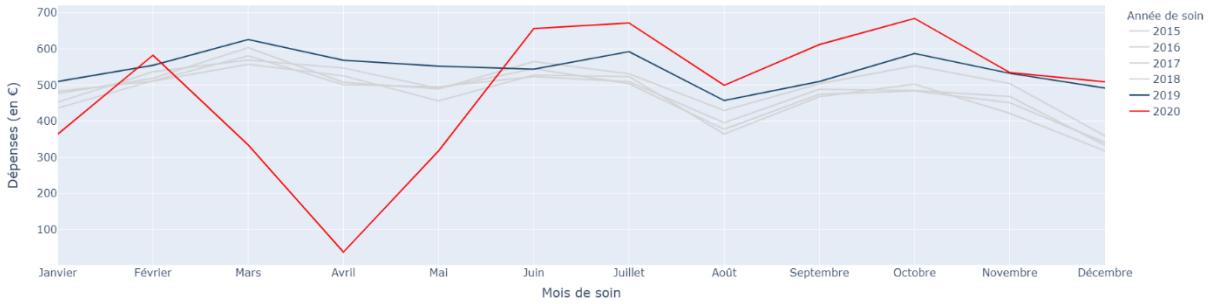


Figure IV-19 – Consommation du poste Optique selon le mois de règlement (en millions d'euros)

	2019	2020	Evolution 2019/2020
Période n°1	1 064,38	947,07	**- 11,0 %**
Période n°2	1 747,41	689,78	**- 60,5 %**
Période n°3	2 103,33	2 439,74	**+ 15,9 %**
Période n°4	1 611,52	1 727,84	**+ 7,2 %**
Total	**6 526,63**	**5 804,43**	**- 11,1 %**

Figure IV-20 – Evolution de la consommation du poste de consommation Optique selon les périodes prédéfinies (en millions d'euros)

De ces premiers graphiques et résultats, il semblerait que le poste de consommation *Optique* a été le plus sensible à la crise du Covid-19. La dépense en 2020 est en baisse de **11 %** comparée à l'année 2019. Cette baisse n'est cependant pas similaire sur l'année. Sur la première période, nous observons ainsi une baisse de **11 %**. Cette baisse est certainement liée à la mise en place de la réforme 100% Santé et des problèmes qui ont émergé dans l'application du tiers payant. Comme dans le poste Dentaire, nous retrouvons un rattrapage de la consommation sur la troisième et quatrième période.

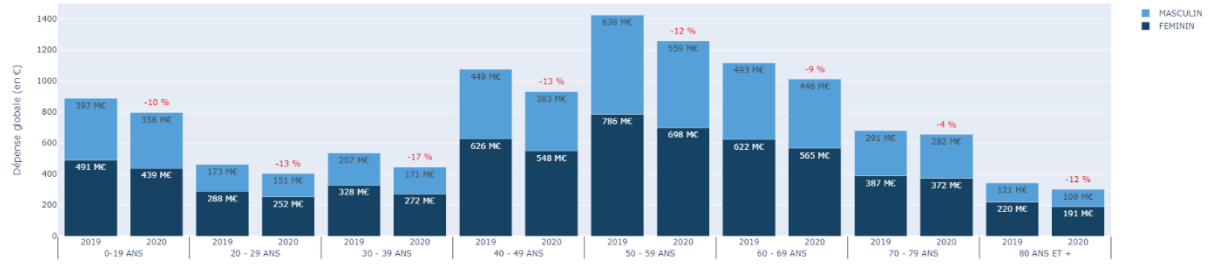


Figure IV-21 – Consommation du poste Optique entre 2019 et 2020 selon la tranche d'âge et le sexe (en millions d'euros)

Sur toute l'année et de façon assez uniforme, la consommation est en baisse pour toutes les tranches d'âges.

e) *Prothèses auditives*

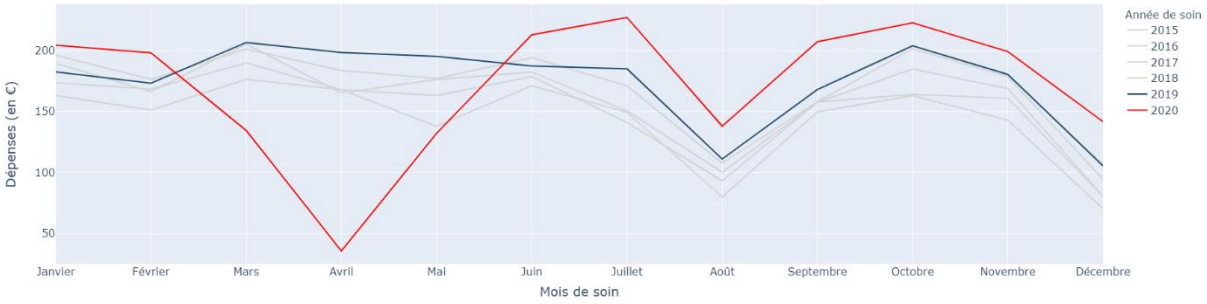


Figure IV-22 – Consommation du poste de Prothèses auditives selon le mois de règlement (en millions d'euros)

	2019	2020	Evolution 2019/2020
Période n°1	355,75	402,50	**+ 13,1%**
Période n°2	600,23	302,35	**- 49,6%**
Période n°3	651,50	785,25	**+ 20,5%**
Période n°4	490,01	563,58	**+ 15,0%**
Total	**2 267,52**	**2 053,68**	**- 2,7%**

Figure IV-23 – Consommation du poste Prothèses auditives selon les périodes prédéfinies (en millions d'euros)

Dernier poste de consommation étudié dans cette partie, les observations pour le poste *Prothèses auditives* sont similaires à celles qui ont été faites pour le poste *Dentaire*. Une hausse de la consommation sur la première période suivie d'une baisse lors de la deuxième à cause du confinement. Un rattrapage a lieu ensuite sur les deux périodes suivantes.

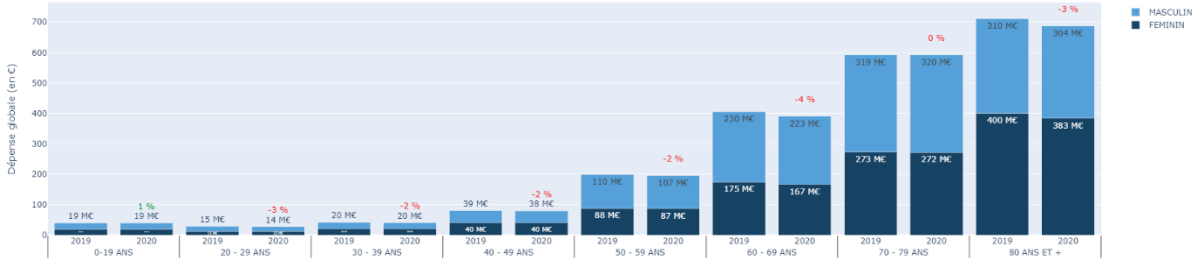


Figure IV-24 – Consommation du poste Prothèses auditives entre 2019 et 2020 selon la tranche d'âge et le sexe (en millions d'euros)

Sur toute l'année, nous observons une baisse globale de la dépense sauf pour la tranche d'âge des 0-19 ans où une légère hausse est observable.

D. Estimation de la dépense à l'ultime pour l'année de soins 2020

Afin de pouvoir exploiter la totalité de la consommation disponible en fonction de la date de soin, nous sommes obligés d'estimer la consommation à l'ultime de l'année de soins 2020. Cela correspond à « prédire » la consommation totale à laquelle nous nous attendons pour cette année à l'aide de l'historique des données disponibles.

Nous utiliserons pour se faire la méthode la plus utilisée et la plus connue : la méthode de Chain Ladder. Nous estimerons nos coefficients de passage à l'aide de la consommation connue entre 2015 et 2019 puis nous utiliserons la consommation connue en 2020 afin d'effectuer un test rétroactif (dit « backtesting ») de validité. Ce test nous permettra de valider ou non la pertinence de l'utilisation des coefficients obtenus à partir de notre historique.

Comme la méthode de Chain Ladder est une technique représentant une partie mineure de ce mémoire, celle-ci sera présentée brièvement. Nous définirons les concepts basiques et les hypothèses indispensables à vérifier afin de pouvoir appliquer cette méthode.

1. Présentation de la méthode déterministe de Chain Ladder

La méthode de Chain Ladder est une méthode déterministe, simple à appliquer et dont l'approche est directe. Son concept est facile à comprendre et repose sur un principe simple : nous utilisons les développements de paiements passés afin de provisionner la dépense future. La méthode de Chain Ladder repose sur des triangles de charges cumulés.

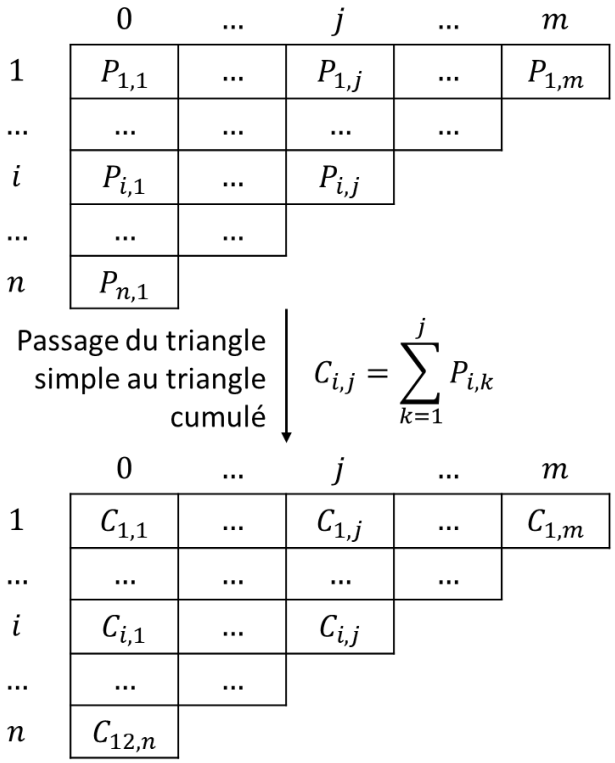


Figure IV-25 – Passage d'un triangle de paiements simples en un triangle de paiements cumulés

Avec :

- n, le dernier mois de survenance et m, le dernier mois de développement disponible ;
- i, le mois de survenance de la consommation, $i \in \{1, \dots, n\}$;
- j, le mois de développement, $j \in \{1, \dots, m\}$;
- $P_{i,j}$, la consommation associée au mois de survenance i et au mois de développement j ;

- $C_{i,j}$, la consommation cumulée comme définie ci-dessus.

Une fois notre triangle de paiements cumulés établi, nous pouvons définir les facteurs de développement individuel $f_{i,j}$, obtenus à partir de la formule suivante.

$$f_{i,j} = \frac{C_{i,j+1}}{C_{i,j}}$$

Nous souhaitons connaître le facteur de passage d'un mois de règlement à l'autre. Ce facteur $\hat{f}_j$ sera estimé comme suit :

$$\hat{f}_j = \frac{\sum_{i=0}^{m-j+1} C_{i,j+1}}{\sum_{i=0}^{m-j+1} C_{i,j}}$$

Par ces facteurs, nous pouvons ensuite définir les montants suivants :

- Les charges ultimes par mois de survenance $\hat{C}_{i,m} = C_{i,m-i} * \prod_{j=m-i}^{m-1} \hat{f}_j$;
- Les provisions par mois de survenance $\hat{R}_i = \hat{C}_{i,m} - C_{i,m-i}$;
- Les provisions totales $\hat{R} = \sum_{i=1}^{n} \hat{R}_i$.

Afin d'appliquer la méthode de Chain Ladder sur nos données, il est nécessaire de vérifier quelques hypothèses :

- Les mois de survenance des soins sont indépendants entre eux ;
- Les mois de développement sont les variables explicatives du comportement des dépenses futures.
- Les facteurs de développement sont indépendants de l'année d'origine i.

Nous effectuons notre étude sur les cinq postes de consommation précédemment définis. Nous devons donc vérifier nos hypothèses cinq fois. Comme nous cherchons à établir la dépense à l'ultime de la consommation de 2020, nous avons donc soixante mois de survenance (de janvier 2015 à décembre 2019) pour soixante-mois de règlement.

Nous effectuerons donc notre analyse sur l'année de survenance 2019 (année la plus récente) et vérifierons dans un premier temps les hypothèses sur le poste de consommation *Optique*. Pour les autres postes, la vérification des hypothèses d'application de la méthode Chain Ladder sera présentée en annexe

a) Vérification des hypothèses : exemple d'application sur le poste de consommation Optique

(a) L'alignement des couples $(C_{i,j}, C_{i,j+1})$

Soit j fixé, nous supposons l'existence d'un coefficient $f_{i,j}$ tel que $C_{i,j+1} = f_{i,j} * C_{i,j}$. Nous nous attendons donc à ce que les couples $(C_{i,j}, C_{i,j+1})$ soient alignés sur une droite passant par l'origine. Nous projetterons nos couples dans un premier temps puis récupèrerons dans un tableau les $R^2 ajusté$ pour la droite de régression associée au mois de consommation de 2019.

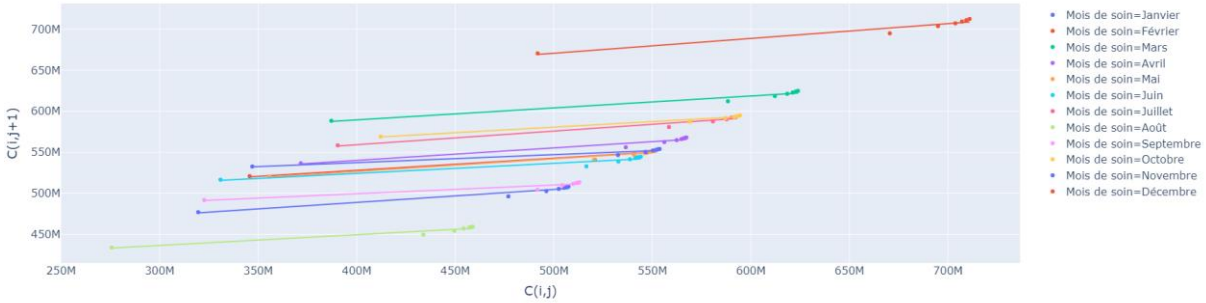


Figure IV-26 – Représentation graphique des couples $(C_{i,j}, C_{i,j+1})$ en 2019 pour le poste de consommation Optique

Mois	$R^2 ajusté$
Janvier	94,26 %
Février	94,06 %
Mars	95,27 %
Avril	94,12 %
Mai	94,64 %
Juin	89,65 %
Juillet	95,53 %
Août	92,55 %
Septembre	89,54 %
Octobre	96,89 %
Novembre	92,74 %
Décembre	93,58 %

Figure IV-27 – Tableau des R^2 ajusté de chaque droite de la figure IV-26

Les résultats obtenus de la projection des couples $(C_{i,j}, C_{i,j+1})$ et résumés dans les deux figures précédentes sont satisfaisants. Nous pouvons observer que toutes les droites semblent parallèles : les pentes de chaque courbe sont sensiblement similaires. Le $R^2 ajusté$ est également satisfaisant. Cette première hypothèse est donc vérifiée : les points sont bien alignés.

(b) L'examen du triangle de développement

A partir des facteurs de développement individuel $f_{i,j}$, nous pouvons réaliser le triangle de développement ci-dessous. Nous cherchons à vérifier que les facteurs $f_{i,j}$ sont sensiblement constants. Nous établissons notre tableau pour $j \in \{0, \dots, 5\}$.

	0	1	2	3	4	5
$\hat{f}_j$	147,48 %	103,53 %	101,06 %	100,44 %	100,25 %	100,16 %
$\mathbb{E}[f_{i,j}]$	148,28 %	103,52 %	101,06 %	100,44 %	100,25 %	100,16 %
$Var[f_{i,j}]$	0,419 %	0,003 %	0,000 %	0,000 %	0,000 %	0,000 %

Figure IV-28 – Tableau de comparaison entre les facteurs de développement individuel et les facteurs de passage pour les 6 premiers mois de développement sur les survenances de 2019

Nous constatons donc que l'espérance des facteurs de développement individuel est très proche du facteur de passage donné par la méthode de Chain Ladder. Nous pouvons également remarquer que la variance converge très rapidement, dès le deuxième mois de règlement. A partir du deuxième mois de développement, nous avons donc des facteurs de développement individuel assez homogènes comme nous pouvons le constater dans les figures ci-dessous.

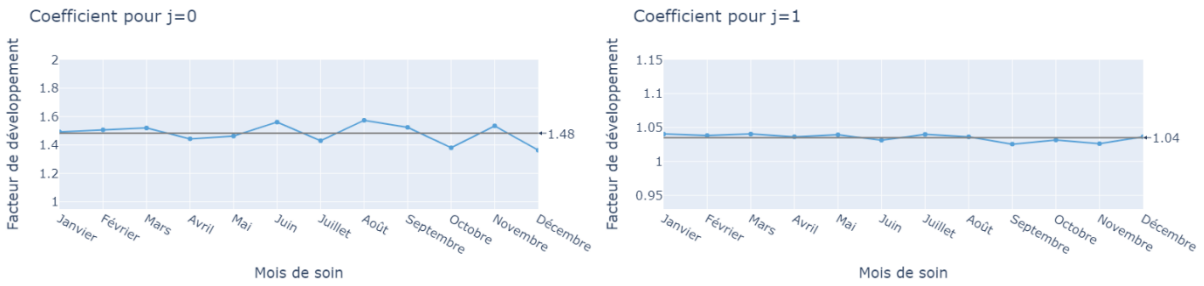


Figure IV-29 – Coefficients individuels du premier et deuxième mois de développement selon les mois de survenance de 2019

De ce que nous pouvons observer, la volatilité des coefficients individuels est plutôt contenue dès le deuxième mois de règlement. Nos coefficients individuels convergent assez rapidement vers le facteur de passage. Nos conditions d'utilisation de la méthode Chain-Ladder sont ainsi bien vérifiées.

L'analyse au global pour tous les postes de consommation s'avère satisfaisante. l'analyse sur les autres postes de consommation est disponible en annexe 1.

2. Application du test rétroactif dit « backtesting » de validité

a) Exemple d'application sur le poste de consommation soins de ville courants

Afin d'effectuer notre test rétroactif de validité, nous effectuons la comparaison entre les données connues en 2020 et les estimations faites à partir de la méthode Chain Ladder sur l'année 2020. Un test rétroactif de validité consiste à partir de la consommation initiale lors du premier mois de survenance du soin, de développer la dépense via les facteurs de passage et calculer la charge jusqu'au dernier mois de règlement connu.

Nous allons comparer deux méthodes retenues pour la détermination des coefficients de passage :

- Méthode A : En prenant en considération l'historique complet de 2015 à 2019 (soit 60 mois d'historique)
- Méthode B : En prenant uniquement l'année 2019 (soit les 12 mois les plus récents).

Nous commençons par déterminer les coefficients de passage entre chaque mois de soin. Le tableau ci-dessous résume les coefficients de passage simple et les coefficients de passage cumulés selon le décalage entre le mois de survenance des soins et le mois de vision de cette consommation. Nous allons nous intéresser aux coefficients cumulés pour la suite.

Un coefficient cumulé correspond au facteur à appliquer à notre dépense connue au mois d'origine.

Décalage du règlement	Coefficient de passage	Coefficient cumulé
1 mois	159,63%	159,63%
2 mois	107,99%	172,38%
3 mois	102,42%	176,54%
4 mois	101,04%	178,38%
5 mois	100,55%	179,36%
6 mois	100,36%	180,00%
7 mois	100,25%	180,44%
8 mois	100,18%	180,77%
9 mois	100,14%	181,02%
10 mois	100,11%	181,21%
11 mois	100,09%	181,38%

Figure IV-30 – Coefficients de passage pour le poste de consommation soins de ville courants selon la méthode B

Par exemple, pour le mois de janvier 2020, nous avons onze mois de dépense connue. Nous allons donc appliquer un coefficient de **181,38 %** à la consommation connue le même mois afin d'obtenir la charge approximée au 31/12/2020 selon la méthode de Chain Ladder. Ensuite, nous effectuerons une comparaison entre cette charge estimée et la dépense réellement observée en fin d'année.

Les résultats sont résumés dans le tableau ci-dessous :

Mois de survenance i	Vision au 31/12/2020				
	Dépense réelle observée	Dépense projetée (méthode A)	Dépense projetée (méthode B)	**Différence d'estimation (méthode A)**	**Différence d'estimation (méthode B)**
Janvier	4 833,24	4 851,98	4 840,51	**18,74**	**7,27**
Février	4 350,83	4 322,03	4 311,91	**- 28,80**	**- 38,92**
Mars	3 368,69	3 607,35	3 599,10	**238,66**	**230,41**
Avril	2 264,15	2 101,08	2 096,42	**- 163,07**	**- 167,73**
Mai	3 537,55	3 121,22	3 114,64	**- 416,33**	**- 422,90**
Juin	4 561,32	4 618,10	4 609,17	**56,78**	**47,85**
Juillet	4 359,69	4 506,91	4 500,34	**147,23**	**140,65**
Août	3 567,46	3 163,25	3 161,10	**- 404,20**	**- 406,35**
Septembre	4 880,97	4 929,68	4 930,70	**48,70**	**49,73**
Octobre	4 824,34	4 834,37	4 842,28	**10,03**	**17,94**
Novembre	4 405,41	4 258,27	4 266,97	**- 147,14**	**- 138,43**
Décembre	2 827,71	N/A	N/A	**N/A**	**N/A**
Total	**44 953,64**	**44 314,24**	**44 273,15**	**- 1,4 %**	**- 1,5 %**

Figure IV-31 – Test de « backtesting » sur le poste de consommation soins de ville courants (en millions d'euros)

L'avant-dernière colonne représente la différence entre la consommation projetée au 31 décembre 2020 selon la méthode A et la dépense réelle connue à cette même date. La dernière colonne représente la différence entre la consommation projetée au 31 décembre 2020 selon la méthode B et la dépense réelle connue à cette même date.

Sur le même tableau, le vert représente une surestimation de la charge au 31 décembre 2020 tandis que le rouge représente une sous-estimation.

La ligne « Total » représente la somme de tous les mois de survenance à l'exception du mois de décembre comme nous ne pouvons pas appliquer la méthode de *backtesting* sur ce mois de survenance. Pour la différence d'estimation, nous présentons sur la ligne total l'écart relatif entre le montant estimé selon la méthode et le montant réellement attendu.

A partir de cette même figure, nous pouvons tirer des conclusions sur l'application de notre test rétroactif de validité. Nous pouvons voir que l'estimation de la charge connue selon l'historique 2015 à 2019 est légèrement plus pertinente pour le poste Soins de ville courants. Cependant, l'écart obtenu entre les deux méthodes semble peu important pour facilement conclure que la méthode A est plus adéquate que la méthode B. Nous effectuons donc – dans un second temps – ce test rétroactif pour les autres postes de consommation. L'analyse sur les autres postes de consommation est disponible en annexe 2.

b) Conclusion du test rétroactif de validité

Nous synthétisons les résultats du « backtest » dans le tableau ci-dessous :

Poste	Vision au 31/12/2020				
	Dépense réelle observée	Dépense projetée (méthode A)	Dépense projetée (méthode B)	**Différence d''estimation (méthode A)**	**Différence d'estimation (méthode B)**
Soins de ville courants	44 953,64	44 314,24	44 273,15	**- 1,4 %**	**- 1,5 %**
Pharmacie	31 964,29	31 726,57	31 798,93	**- 0,7 %**	**- 0,5 %**
Dentaire	9 281,04	9 344,14	9 224,70	**0,7 %**	**- 0,6 %**
Optique	5 295,58	5 828,50	5 127,51	**10,1 %**	**- 3,2 %**
Prothèses auditives	1 911,91	2 277,26	2 041,84	**19,1 %**	**6,8 %**

Figure IV-32 – Synthèse des tests rétroactifs de validité sur tous les postes de consommation étudiés (en millions d'euros)

Le tableau ci-dessus nous permet de constater qu'aucune des deux approches ne semble plus pertinente sur les postes *Soins de ville courants*, *Pharmacie* et *Dentaire*. Cependant, considérer uniquement l'année 2019 permet de drastiquement réduire l'écart de la différence d'estimation pour les postes *Optique* et *Prothèses Auditives* comparé à l'estimation de la charge utilisant les données de 2015 à 2019.

La **méthode B** nous semble par conséquent plus pertinente que la méthode A. De plus, nous faisons l'hypothèse que l'estimation des dépenses de l'année N peut être expliquée en utilisant uniquement les cadences des dépenses de l'année précédente. En effet, des récents changements réglementaires ou de traitement au niveau de la gestion de la Sécurité Sociale auront un impact plus important sur l'année N : restreindre la profondeur de l'historique à uniquement la dernière année complète connue nous semble plus approprié afin de projeter les dépenses de l'année N.

3. Étude sur les cadences des règlements

Les bases Open DAMIR étant publiquement mises à disposition une fois par an, nous ne possédons pas une vision récente des dépenses. Ainsi, par exemple, pour le mois de survenance de décembre 2020, nous ne possédons que la vision au 31/12/2020. Il est donc nécessaire pour la suite du mémoire d'estimer des dépenses à l'ultime pour ces mois.

Afin de répondre à cette problématique, nous étudions les cadences des règlements.

Exemple d'application : *Soins de ville courants*

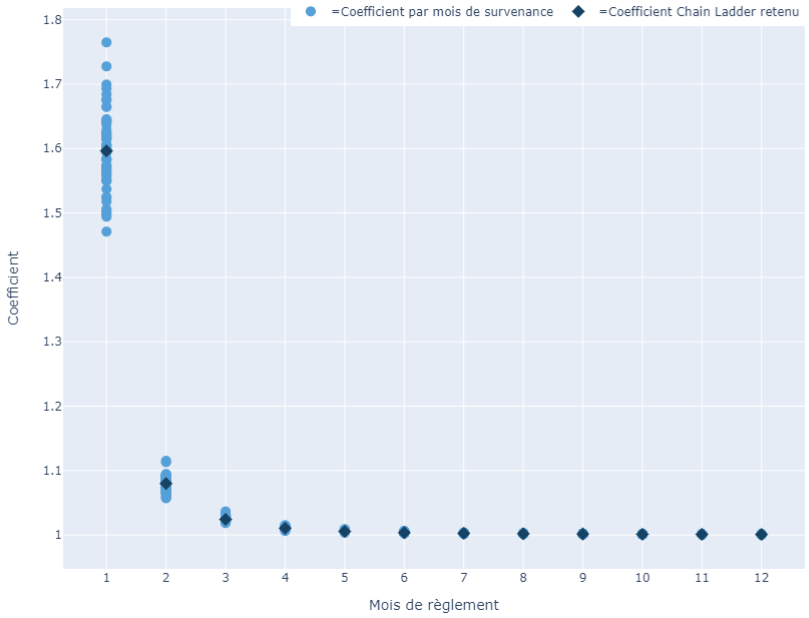


Figure IV-33 – Coefficients de passage sur les 12 premiers mois de règlement pour le poste Soins de ville courants

Nous pouvons constater, à l'aide de la figure ci-dessus, que les coefficients de passage d'un mois de règlement à l'autre convergent très rapidement vers 1.

Mois de survenance	Dépenses 3 mois après	Dépenses 12 mois après	Dépense réelle au 31/12/2020	% 3 mois après	% 12 mois après
Janvier 2015	3 711,37	3 852,36	3 875,00	95,8 %	99,4 %
Janvier 2016	3 893,02	4 026,74	4 046,55	96,2 %	99,5 %
Janvier 2017	4 255,51	4 394,99	4 423,52	96,2 %	99,4 %
Janvier 2018	4 407,26	4 555,63	4 608,34	95,6 %	98,9 %
Janvier 2019	4 552,62	4 701,92	4 772,59	95,4 %	98,5 %

Figure IV-34 – Vision de la consommation 3 et 12 mois après la dépense pour le poste Soins de ville courants

Nous pouvons voir que 12 mois après la date de survenance, nous constatons plus de 98 % de la dépense comparée à la dépense réellement connue au 31/12/2020. Nous retenons donc que la charge ultime est connue au maximum douze mois après la date de survenance.

L'analyses sur les autres postes de consommation est disponible en annexe 3.

4. Application : estimation de la charge ultime de la survenance 2020

Nous avons donc une base de soins complète pour les survenances 2015 à 2019 en provenance des données Open DAMIR. Ces années de survenance sont considérées développées à l'ultime. Cependant, la survenance 2020 n'est pas totalement développée. Nous projetons donc la charge ultime par poste de consommation par la méthode de Chain Ladder présentée auparavant.

Nous appliquons donc les coefficients de développement mensuel obtenus à partir de la dernière année complète d'historique et nous développons sur 12 mois les dépenses de la survenance 2020 afin d'obtenir une vision au 31/12/2021.

Les dépenses à l'ultime pour l'année de survenance 2020 sont présentées ci-dessous :

Mois de survenance 2020	Poste de consommation				
	Soins de ville courants	Pharmacie	Dentaire	Optique	Prothèses auditives
Janvier	4 836,88	3 141,98	917,45	364,67	204,50
Février	4 358,14	2 856,72	955,82	583,19	198,35
Mars	3 378,04	3 088,53	610,82	334,71	134,77
Avril	2 273,57	2 609,83	87,56	38,02	35,82
Mai	3 558,66	2 634,42	675,44	318,79	132,70
Juin	4 599,90	2 886,20	1 202,88	659,34	214,14
Juillet	4 412,22	3 007,16	1 152,96	676,02	228,97
Août	3 630,14	2 686,84	553,02	503,91	139,71
Septembre	5 018,58	3 001,88	1 058,56	620,11	211,48
Octobre	5 080,17	3 388,37	1 095,36	700,79	232,65
Novembre	5 009,46	3 100,65	1 180,07	566,70	230,02
Décembre	5 132,83	3 353,61	1 251,23	795,54	335,09

Figure IV-35 – Synthèse de la charge ultime estimée pour l'année de survenance 2020 pour tous les postes de consommation étudiées

Les analyses dans les parties suivantes seront effectuées à partir de la base de données suivante :

- Une vision complète de la dépense sur 12 mois après pour les années de survenance de 2015 à 2019 ;
- Une vision estimée de la dépense pour l'année de survenance de 2020.

V. La prédiction des dépenses de l'année 2020

A. Présentation des modèles et concepts utilisés

Nous commençons par l'explication des méthodes utilisées par la suite. Nous rappelons que la base Open DAMIR que nous avons souhaitée étudier propose un découpage des dépenses par mois de règlement. Nous avons été amenés à manipuler cette base de données afin d'étudier et visualiser la dépense suivant la date de soin.

De plus, comme nous souhaitons étudier l'impact du Covid-19 sur les dépenses, nous avons intuitivement décidé d'utiliser le concept des séries temporelles. A partir de cette première idée, nous chercherons à déterminer s'il existe un modèle de séries temporelles sous-jacent à nos données sur la période avant 2020. Nous montrerons en quoi l'année 2020 avec le Covid-19 a été une année atypique qui peut ne pas être captée à partir du modèle.

1. Introduction aux séries temporelles

Une série temporelle est une séquence d'observation quantitative prise successivement dans le temps. Une série temporelle introduit donc un ordre explicite et logique entre les observations. Cette forte dimension temporelle correspond au squelette de la série et représente sa principale source d'information. Changer l'ordre des observations entraîne la prise en considération d'une nouvelle série temporelle.

Avant de s'intéresser de plus près aux deux modèles de prédiction de séries temporelles que nous avons utilisés, nous allons commencer par définir les concepts nécessaires à l'analyse de ces séries.

2. Les concepts fondamentaux

a) La structure d'une série temporelle

L'analyse des séries temporelles repose sur les observations faites à des moments antérieurs, appelés temps de retard ou *lags*. Ce que nous cherchons à prédire sont les observations à des moments postérieurs. Nous pouvons donc définir :

- $T-n$: correspond au temps antérieur ou lag du point de référence. Il s'agit des données qui vont venir alimenter le modèle et qui servent de référence dans la prédiction.
- T : le point de référence ;
- $T+n$: correspond au temps postérieur ou prévisionnel. Il s'agit de l'information inconnue à la date de référence et qui est extrapolée suite à l'analyse de la série temporelle.

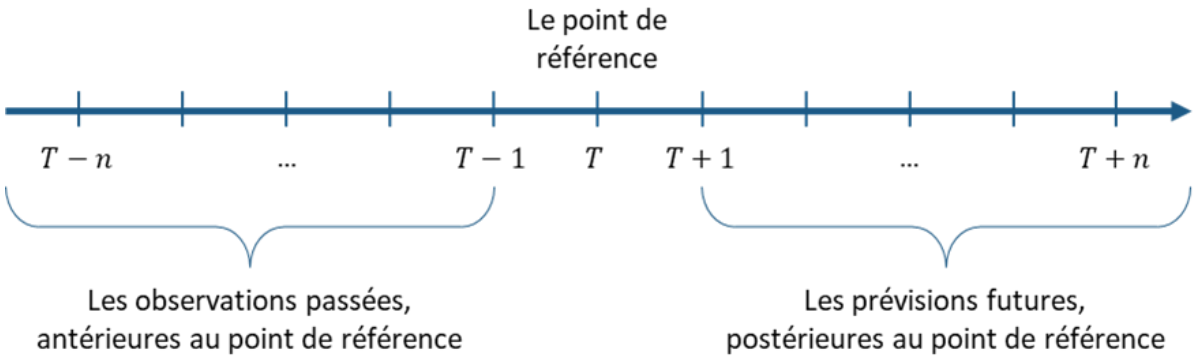


Figure V-1 – Structure d'une série temporelle

b) Les principales composantes d'une série temporelle

Une série temporelle peut être répartie en quatre composantes distinctes :

- **Un niveau :** C'est la valeur de base de la série. Cela correspond à la moyenne de la série.
- **Une tendance :** Il s'agit du comportement de la série dans le temps. La tendance décrit une augmentation ou diminution constatée au fil du temps.
- **Une saisonnalité :** Il s'agit des schémas ou cycles répétitifs dans le comportement de la série dans le temps.
- **Un bruit :** Cela correspond à tout ce que le modèle n'est pas en mesure d'expliquer. Le bruit peut être vu comme étant la différence entre la valeur réelle et la valeur générée par le modèle.

Toutes les séries temporelles ont un niveau et un bruit. La tendance et la saisonnalité sont facultatives. Cependant, la tendance et la saisonnalité représentent la plus grande source d'information dans l'analyse et la prédiction. Ces deux composantes permettent d'affiner les prédictions et d'obtenir des résultats de prédiction satisfaisants.

Ces composantes se combinent soit de manière additive, soit de manière multiplicative afin de former la série temporelle.

- **Le modèle additif**

Le modèle additif correspond à la somme des composantes :

$$y(t) = Niveau + Tendance + Saisonnalité + Bruit$$

Un modèle additif est linéaire, cela signifie que les changements dans le temps se font toujours dans les mêmes proportions : la tendance est une ligne droite et la série temporelle a la même fréquence (largeur des cycles) et la même amplitude (hauteur des cycles) que la saisonnalité.

- **Le modèle multiplicatif**

Le modèle multiplicatif correspond au produit des composantes :

$$y(t) = Niveau * Tendance * Saisonnalité + Bruit$$

Un modèle multiplicatif est non-linéaire. La série temporelle présente des amplitudes de plus en plus importantes.

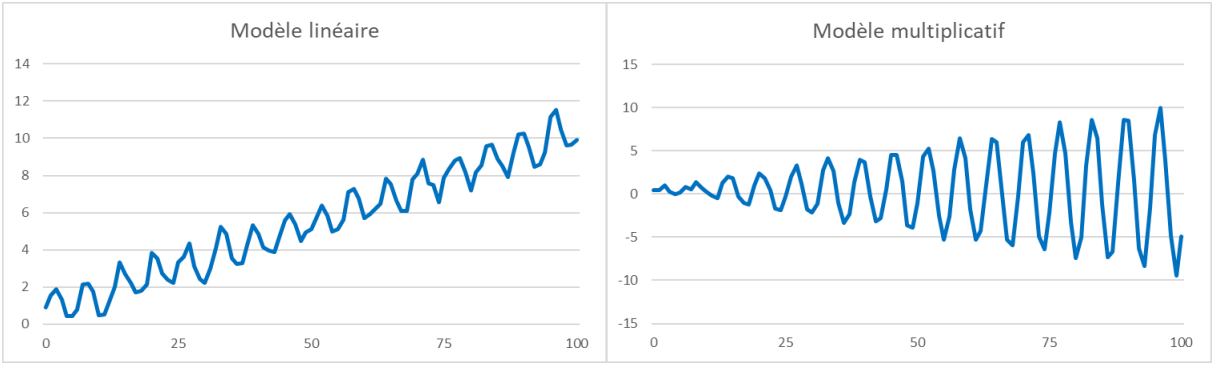


Figure V-2 – Visualisation des différences entre un modèle linéaire (à gauche) et d'un modèle multiplicatif (à droite)

Afin de pouvoir effectuer nos prédictions, nous supposons que les informations contenues dans Y_{t-1},…, Y_{t-n} fournissent les informations sur les valeurs futures de la série. Nous utilisons pour cela, le concept de stationnarité.

Comme une série temporelle possède une structure imposant un ordre aux observations, cela implique donc des hypothèses sur la série. Nous supposons notamment que les observations de la série sont consistantes sur la période étudiée. Autrement dit, la moyenne et la variance sont supposées constantes dans le temps.

Si la série n'est pas stationnaire alors nous ne pouvons pas l'utiliser pour prédire le futur. Il est donc nécessaire de rendre stationnaire la série avant de s'occuper de la construction des modèles. Pour rendre une série temporelle stationnaire, il est nécessaire de traiter la composante *Tendance* et la composante *Saisonnalité* de notre série. Une série temporelle est stationnaire au second ordre si ses moments d'ordre un et d'ordre deux ne dépendent plus du temps.

c) Le prétraitement des données

(a) La tendance

Comme évoqué précédemment, une tendance est une augmentation ou une diminution à long terme du niveau de la série temporelle. Nous pouvons distinguer deux types de tendance :

- La tendance déterministe : il s'agit d'une tendance qui augmente ou diminue de façon constante.
- La tendance stochastique : il s'agit d'une tendance qui augmente ou diminue de façon instable.

De plus, une tendance peut être globale ou locale. Une tendance globale s'applique à l'ensemble de la série alors qu'une tendance locale s'applique à une sous-partie de la série. Ainsi, une tendance déterministe globale sera bien plus facile à identifier à l'aide des outils traditionnels disponibles sous Python.

(b) La saisonnalité

Pour rappel, la saisonnalité est un schéma qui est amené à se répéter au fil du temps. Si ce schéma se répète à la même fréquence, alors il est considéré saisonnier. Autrement, nous le définissons simplement comme un cycle. **Il sera intéressant – lors de la construction de nos modèles – de remarquer la présence d'une saisonnalité : les dépenses mensuelles peuvent se ressembler d'une période à une autre, nous déterminons l'existence d'un cycle de 12 mois.**

(c) Contrôler le résultat

Une fois les composantes *Tendance* et *Saisonnalité* retirées de la série temporelle, il est nécessaire d'effectuer des contrôles pour s'assurer de la bonne stationnarité de la série temporelle. Plusieurs contrôles sont possibles afin de vérifier cette hypothèse :

- Contrôle visuel : cela consiste à observer la courbe de la série temporelle et à identifier visuellement si la tendance et/ou la saisonnalité est toujours présente. Cette méthode est très subjective ;
- Statistiques sommaires : comme la série est supposée stationnaire, il est alors possible de vérifier que la moyenne et la variance sont bien constantes dans le temps. Cela consiste à sélectionner des partitions aléatoires de la série et à y vérifier ces conditions ;
- Tests statistiques : la meilleure façon de vérifier l'hypothèse de stationnarité de la série temporelle consiste à effectuer un test statistique de stationnarité. Nous expliciterons ici deux tests : le test augmenté de Dickey Fuller (ADF) dont l'hypothèse nulle est la non-stationnarité

et le test Kwiatkowski–Phillips–Schmidt–Shin (KPSS) dont l'hypothèse nulle est la stationnarité. Ces deux tests, ayant une hypothèse nulle opposée, permettent de compléter notre analyse.

(i) Le test augmenté de Dickey Fuller – le test ADF

Le test augmenté de Dickey Fuller est un test statistique appelé *test de racine unitaire*. L'objectif du test est de vérifier à quel point la série temporelle est définie par une tendance. Les hypothèses de ce test sont :

- Hypothèse nulle H_0 : la série temporelle a une racine unitaire et est donc non-stationnaire. La structure de la série est dépendante du temps.
- Hypothèse alternative H_1 : nous pouvons considérer que la série temporelle n'a pas de racine unitaire et est donc stationnaire. Il n'existe donc pas de structure dépendante du temps dans la série.

Pour effectuer un test statistique, nous utilisons la $p - value$ du test et nous la comparons à un certain seuil significatif α. L'interprétation est la suivante :

- $p - value > \alpha$: dans ce cas, nous ne sommes pas en mesure de rejeter l'hypothèse nulle H_0. Pour autant, nous ne pouvons pas l'accepter. Un autre test peut s'avérer nécessaire ;
- $p - value \leq \alpha$: dans ce cas, nous sommes en mesure de rejeter l'hypothèse nulle et d'accepter l'hypothèse alternative avec un risque proportionnel à la $p - value$ d'avoir tort. Selon le test ADF, nous pouvons considérer que la série temporelle est stationnaire.

(ii) Le test KPSS

En complément du test augmenté de Dickey Fuller, il sera intéressant d'utiliser le test KPSS qui permettra de confirmer le résultat comme les hypothèses de ce test sont l'inverse du test précédent :

- Hypothèse nulle H_0 : la série temporelle est stationnaire. La structure de la série est donc indépendante du temps.
- Hypothèse alternative H_1 : la série temporelle a une racine unitaire et n'est donc pas stationnaire. Il existe une structure dépendante du temps dans la série.

Comme pour le test ADF, nous utilisons la $p - value$ du test et la comparons à un certain seuil significatif α. L'interprétation est la suivante :

- $p - value > \alpha$: dans ce cas, nous ne sommes pas en mesure de rejeter l'hypothèse nulle H_0. Pour autant, nous ne pouvons pas l'accepter. Un autre test peut s'avérer nécessaire ;

$p - value \leq \alpha$: dans ce cas, nous sommes en mesure de rejeter l'hypothèse nulle et d'accepter l'hypothèse alternative avec un risque proportionnel à la $p - value$ d'avoir tort. Dans le cas du test KPSS, nous pouvons considérer que la série temporelle n'est pas stationnaire.

d) Les modèles de séries temporelles

Une fois que les prétraitements sur les séries temporelles ont été appliqués, en vérifiant la stationnarité, nous sommes en mesure d'appliquer les modèles de prédiction de séries temporelles. Dans cette première partie, nous nous sommes intéressés à deux de ces modèles.

Nous commençons par nous intéresser au modèle le plus connu pour prédire les séries temporelles : « *Autoregressive Integrated Moving Average* » ou ARIMA.

(a) Le modèle ARIMA

Un modèle ARIMA s'écrit comme $ARIMA(p, d, q)$. Chaque paramètre de ce modèle apporte une information complémentaire et indispensable à son bon fonctionnement :

- AR (*Auto-Régression liée au terme* p) : le modèle considère qu'il existe une relation de dépendance entre l'observation t et un certain nombre d'observations *p* décalées ;
- I (*Intégré lié au terme* d) : le nombre de différenciation nécessaire afin de rendre la série temporelle stationnaire. Les observations seront ainsi différenciés *d* fois ;
- MA (*Moyenne Mobile lié au terme* q) : le modèle considère qu'il existe une relation de dépendance entre l'observation et les erreurs résiduelles d'un modèle de moyenne mobile appliqué aux observations décalées. Le terme *q* représente finalement l'ordre de la moyenne mobile.

Cela nous permet d'établir la définition suivante :

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

$$\phi_p(L)(1 - L)^d y_t = \mu + \theta_q(L)\epsilon_t, \forall t \geq 0$$

Avec :

- ϵ_t, un bruit blanc faible. Il s'agit d'une suite de variables aléatoires réelles identiquement distribuées et non corrélées entre elles. ($\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$) ;
- $\phi_p(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$, fonction de l'auto-régression ;
- $\theta_q(L) = 1 + \theta_1 L - \cdots + \theta_q L^q$, fonction de la moyenne mobile ;
- Toutes les racines de l'équation caractéristique associée à ϕ sont de module inférieur à 1 ;
- Toutes les racines de l'équation caractéristique associée à θ sont de module inférieur à 1 ;
- L correspond à l'opérateur retard (le *lag*) que nous avons évoqué en introduction. Sa formule est la suivante : $L^k y_t = y_{t-k}$.

Quelques cas particuliers sont à noter :

- Dans le cas où *d = 0*, nous nous retrouvons dans le cas des modèles *ARMA* classique, c'est-à-dire lorsque la série temporelle a déjà été rendue stationnaire. Un modèle ARIMA est donc la généralisation d'un modèle ARMA ;
- Dans le cas où *p = 0* et *d = 0*, nous sommes dans le cas d'un modèle *MA* ;
- Dans le cas où *d = 0* et *q = 0*, nous sommes dans le cas d'un modèle *AR*.

(i) Sélection de p et q

Afin de choisir *p* et *q*, il existe deux graphiques « de diagnostic » qui aident à les choisir de façon visuelle.

Le premier graphique, l'ACF ou fonction d'autocorrélation d'ordre *k*, aide à choisir l'ordre du terme *p* de l'auto-régression. Le graphique obtenu par l'ACF résume la corrélation d'une observation avec ces valeurs de décalage jusqu'à l'ordre k. L'abscisse représente le coefficient de retard et l'axe des ordonnées représente le coefficient de corrélation (entre 1 et -1). Le graphique compare ainsi la corrélation entre Y_t avec $Y_{t-1}, \dots, Y_{t-k}$.

Le second, le PACF ou fonction d'autocorrélation partielle d'ordre *k*, aide à choisir l'ordre du terme *q* de la moyenne mobile. Ce graphique représente la corrélation d'une série temporelle après avoir éliminé l'effet de toute corrélation due à des retards plus courts que l'ordre *k*. Le graphique obtenu présente des axes similaires à celui de l'ACF.

L'ordre *p* et *q* sont ainsi choisis tant que la corrélation entre les termes est significative. Cela signifie que :

- Le modèle est un modèle autorégressif (AR) si le PACF a une coupure nette après le premier terme et que l'ACF continue après le premier terme.
- Le modèle est un modèle de moyenne mobile (MA) si l'ACF a une coupure nette après le premier terme et que le PACF continue après le premier terme.
- Le modèle est un mélange des deux (ARIMA) lorsque l'ACF et le PACF « se traînent » tous les deux.

Pour la suite, nous serons amenés à utiliser une librairie Python qui optimise ce processus. Cette librairie est évoquée à la page suivante.

(ii) Évaluer le modèle

Afin d'évaluer la qualité et la pertinence du modèle choisi, il est intéressant de diviser la série en deux. La première partie de la série correspond aux données d'entraînement du modèle tandis que la deuxième partie de la série correspond aux données de test du modèle. Sur cette deuxième partie, nous pourrons mesurer le caractère approprié du modèle. En occurrence, nous utiliserons l'historique 2015 – 2019 pour prédire 2020.

Afin d'évaluer la pertinence du modèle produit, nous définissons par la suite quelques indicateurs. Nous faisons le choix d'utiliser deux indicateurs simples, faciles à comprendre.

Le premier indicateur que nous allons mettre en place est celui de l'erreur de prédiction qui représente la simple différence entre la valeur prédite et la valeur attendue. Une somme nulle de cet indicateur ne signifie pas forcément que le modèle obtenu est parfait mais permet de dire que le modèle a bien capturé le niveau de la série sur la période étudiée.

$$Erreur_i = Prédiction_i - Valeur\ attendue_i = \hat{y}_i - y_i = \epsilon_i$$

Nous serons également invités à utiliser la moyenne de l'erreur absolue définie de la façon suivante :

$$MEA = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|\epsilon_i|$$

Cet indicateur d'erreur nous indique le bruit moyen en valeur absolue que le modèle n'a pas été capable de capturer, qu'il soit positif ou négatif.

(iii) L'implémentation dans Python

Afin d'implémenter le modèle $ARIMA$ dans Python, nous allons faire appel à la librairie $pmdarima$ et à l'une de ces fonctions associées : $auto\_arima([\dots])$.

Comme son nom l'indique, cette fonction cherche de façon automatique les paramètres associés au modèle ARIMA. La fonction teste une multitude de valeurs pour p, d et q afin de trouver la combinaison optimale : celle qui colle le mieux à notre jeu de données d'entraînement et qui, en retour, nous permettra d'obtenir une prédiction satisfaisante sur la période de test.

En utilisant cette fonction, les deux graphiques de diagnostic évoqués précédemment ne sont pas analysés. Cependant, le travail préalable d'analyse et de transformation est bien mené : les cinq séries temporelles des dépenses, l'un par poste de consommation, sont rendus stationnaires avant de les communiquer à la fonction. Finalement, cette fonction nous permet d'optimiser le choix des paramètres p, d et q.

(a) Choix des paramètres optimaux

La fonction $auto\_arima([\dots])$ est amenée à générer un grand nombre de modèles à l'aide du jeu de données test que nous fournissons en entrée. Il est donc nécessaire de faire le tri parmi les modèles afin de sélectionner celui qui est le plus pertinent sur nos données d'entraînement.

Il existe plusieurs critères de qualité d'un modèle afin de sélectionner le modèle le plus pertinent : *AIC*, *BIC*, *HQIC,* parmi d'autres. Cependant, l'échantillon de données étant de petite taille (60 observations entre janvier 2015 et décembre 2019 avant toute transformation), nous sommes donc amenés à utiliser comme mesure de qualité du modèle le critère d'information d'Akaike corrigé – *l'AICc*. Le meilleur modèle est celui qui minimise ce critère. La formule de ce critère est la suivante :

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

Avec :

- $AIC = 2k - \ln(L)$
- k, le nombre de paramètres à estimer du modèle ;
- L, le maximum de la fonction de vraisemblance du modèle ;
- n, la taille de l'échantillon.

L'*AICc* cherche à représenter un compromis entre la qualité de la prédiction et la complexité du modèle. En effet, il peut être intéressant d'avoir un modèle très complexe puisqu'intuitivement, nous pouvons nous dire que plus un modèle est complexe, alors plus la prédiction produite « collera à la réalité ». En même temps, un modèle complexe entraîne un risque de sur ajustement : le modèle sur ajusté prend en compte le bruit aléatoire et toutes les légères déviations pouvant exister. Ce sur ajustement empêche alors le modèle d'effectuer des prédictions fiables.

En règle générale, un modèle doit être un équilibre entre le biais de prédiction et la parcimonie du nombre de paramètres.

(b) Vérification du bruit « blanc »

Pour vérifier que le résidu du modèle est bien un bruit blanc, nous serons amenés à vérifier plusieurs hypothèses :

- Il n'existe pas de corrélation entre les résidus ;
- La moyenne est nulle et la variance constante ;
- Les résidus suivent une loi normale, i.e. qu'ils sont normalement distribués.

Corrélation des résidus

Afin de vérifier l'existence ou l'absence de corrélation entre nos résidus, nous sommes amenés à utiliser le test Q de *Ljung-Box* dont les hypothèses sont les suivantes :

- Hypothèse nulle H_0 : les données sont indépendamment distribuées.
- Hypothèse alternative H_1 : les données ne sont pas indépendamment distribuées. Elles sont donc corrélées.

La statistique du test est la suivante :

$$Q = n(n+2)\sum_{k=1}^{m}\frac{\hat{\rho}_k^2}{n-k}$$

Avec :

- n, la taille de l'échantillon.
- $\hat{\rho}_k^2$, l'autocorrélation de l'échantillon au décalage k.
- m, le nombre de décalage testé.

Sous l'hypothèse nulle H_0, la statistique du test suit asymptotiquement une $\chi^2_{(h)}$ (loi du « *khi-deux* »). L'hypothèse nulle est rejetée lorsque $Q > \chi^2_{1-\alpha,h}$ où $\chi^2_{1-\alpha,h}$ est la valeur de la table de distribution de la loi du khi-deux avec h degrés de liberté et de seuil de signification α. Le nombre de degrés de liberté s'obtient comme suit : $h = m - p - q$ avec p, le terme autorégressif et q, la moyenne mobile du modèle ARIMA ou SARIMA.

Comme pour les tests précédemment, nous utilisons la $p - value$ du test et la comparons à un certain seuil significatif α. L'interprétation est la suivante :

- $p - value > \alpha$: dans ce cas, nous ne sommes pas en mesure de rejeter l'hypothèse nulle H_0. Nous pouvons admettre qu'il n'existe pas de corrélation entre nos résidus.
- $p - value \leq \alpha$: dans ce cas, nous sommes en mesure de rejeter l'hypothèse nulle et d'accepter l'hypothèse alternative avec un risque proportionnel à la $p - value$ d'avoir tort. Dans notre cas, nous pouvons dire que nos données ne sont pas indépendamment distribuées.

Étude des résidus

Afin de confirmer que les résidus de notre modèle suivent une loi normale, nous allons utiliser le test de Jarque-Bera. Il est cependant important de noter que ce test peut manquer de pertinence – tout en étant le plus performant – du fait de la faible taille de notre échantillon de test[19]. Les hypothèses de ce test sont :

- Hypothèse nulle H_0 : le *kurtosis* et le « *skewness* » sont les mêmes que ceux d'une loi normale. Nous pouvons en déduire que nos données suivent approximativement une loi normale ;
- Hypothèse alternative H_1 : le *kurtosis* et le « *skewness* » diffèrent de ceux d'une loi normale. Nous pouvons en déduire que nos données ne suivent pas une loi normale.

La statistique du test est la suivante :

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right)$$

Avec :

- n, la taille de l'échantillon ;
- S, le coefficient d'asymétrie ou « $Skewness$ » de l'échantillon. Ce coefficient se calcule comme suit :

$$S = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

- K, le coefficient d'aplatissement ou $Kurtosis$ de l'échantillon. Ce coefficient se calcule comme suit :

$$K = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{2}}$$

- Sous l'hypothèse nulle H_0, nous nous attendons à ce que $K = 3$ et $S = 0$.

Sous l'hypothèse nulle H_0, la statistique du test suit asymptotiquement une $\chi^2_{(2)}$. L'hypothèse nulle est rejetée lorsque $JB > \chi^2_{1-\alpha,2}$ où $\chi^2_{1-\alpha,2}$ est la valeur de la table de distribution de la loi du khi-deux avec 2 degrés de liberté et de seuil de signification α.

Comme pour les tests précédemment, nous utilisons la $p-value$ du test et la comparons à un certain seuil significatif α. L'interprétation est la suivante :

- $p - value > \alpha$: dans ce cas, nous ne sommes pas en mesure de rejeter l'hypothèse nulle H_0. Nous pouvons admettre que nos résidus suivent une loi normale.
- $p - value \leq \alpha$: dans ce cas, nous sommes en mesure de rejeter l'hypothèse nulle et d'accepter l'hypothèse alternative avec un risque proportionnel à la $p - value$ d'avoir tort. Dans notre cas, nos résidus ne suivent pas une loi normale.

[19] Si vous souhaitez plus de détail, cela a été abordé ici.

Étude de la variance

Pour finir, nous allons chercher la nature de la variance de nos résidus :

- Si la variance est constante, alors nous parlons d'homoscédasticité ;
- A l'inverse, si celle-ci évolue dans le temps, nous parlons alors d'hétéroscédasticité.

Dans notre cas, nous cherchons à obtenir une variance homoscédastique. Pour ce faire, nous appliquerons le test de Breusch-Pagan dont les hypothèses sont :

- Hypothèse nulle H_0 : la variance est constante. Les résidus sont donc homoscédastiques ;
- Hypothèse alternative H_1 : la variance évolue au fil du temps. Nous pouvons en déduire que les résidus de notre modèle sont hétéroscédastiques.

La statistique du test est la suivante :

$$BP = n * R^2$$

Avec :

- n, la taille de l'échantillon ;
- R^2, le coefficient de détermination de la régression des résidus du modèle élevés au carré.

Sous l'hypothèse nulle H_0, la statistique du test suit asymptotiquement une $\chi^2_{(h)}$ (loi du « *khi-deux* »). L'hypothèse nulle est rejetée lorsque $Q > \chi^2_{1-\alpha,h}$ où $\chi^2_{1-\alpha,h}$ est la valeur de la table de distribution de la loi du khi-deux avec h degrés de liberté et de seuil de signification α. h correspond au nombre de variables indépendantes et correspond à $h = p + q$.

Comme précédemment, nous utilisons la $p - value$ du test et la comparons à un certain seuil significatif α. L'interprétation est la suivante :

- $p - value > \alpha$: dans ce cas, nous ne sommes pas en mesure de rejeter l'hypothèse nulle H_0. Nous pouvons admettre que la variance de nos résidus est homoscédastique, celle-ci est constante au fil du temps.
- $p - value \leq \alpha$: dans ce cas, nous sommes en mesure de rejeter l'hypothèse nulle et d'accepter l'hypothèse alternative avec un risque proportionnel à la $p - value$ d'avoir tort. Dans notre cas, la variance de nos résidus change au fil du temps.

(b) Le modèle SARIMA

Afin d'aller plus loin dans notre étude et comme une visualisation de nos jeux de données dans la partie III a pu nous le confirmer, il sera intéressant de rajouter au modèle $ARIMA$ le concept de saisonnalité, nous permettant alors d'affiner notre prédiction. Pour cela, nous allons utiliser le modèle $Seasonal\ ARIMA - SARIMA$.

Un processus stochastique $(Y_t)_{t\geq p-d}$ est un modèle $SARIMA(p,d,q)(P,D,Q)_m$ s'il satisfait l'équation suivante :

$$\phi_p(L)(1-L)^d * \Phi_P(L^m)(1-L^m)^D * y_t = \mu + \theta_q(L) * \Theta_Q(L^m) * \epsilon_t, \forall t \geq 0$$

Avec :

- ϵ_t, un bruit blanc faible. Il s'agit d'une suite de variables aléatoires réelles identiquement distribuées non corrélées entre eux. ($\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$) ;
- m est l'opérateur de saisonnalité ($m = 12$ pour des données mensuelles, $m = 4$ pour des données trimestrielles)
- $\phi_p(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$;
- $\theta_q(L) = 1 + \theta_1 L + \cdots + \theta_q L^q$;
- La série temporelle non stationnaire y_t est transformé en processus stationnaire $\hat{y}_t$ selon l'équation suivante : $\hat{y}_t = (1-L)^d(1-L^m)^D y_t$. d correspond à l'ordre de différenciation et D correspond à l'ordre de différenciation saisonnier ;
- Les polynômes saisonniers $\Phi_P(L^m)$ et $\Theta_Q(L^m)$ sont définis comme suit :
  - $\Phi_P(L^m) = 1 - \Phi_1 L^m - \cdots - \phi_P L^{Pm}$;
  - $\Theta_Q(L^m) = 1 + \Theta_1 L^m - \cdots + \Theta_Q L^{Qm}$.
- Toutes les racines de l'équation caractéristique associée à ϕ sont de module inférieur à 1 ;
- Toutes les racines de l'équation caractéristique associée à θ sont de module inférieur à 1 ;
- Toutes les racines de l'équation caractéristique associée à Φ sont de module inférieur à 1 ;
- Toutes les racines de l'équation caractéristique associée à Θ sont de module inférieur à 1 ;

L'implémentation dans Python

Comme pour le modèle $ARIMA$, nous utilisons la même fonction : $auto\_arima([\ldots])$ en lui précisant de rechercher spécifiquement les paramètres liés à la saisonnalité. De plus, nous utiliserons également l'*AICc* comme critère de sélection du modèle. Nous vérifierons également que nos résidus sont bien du bruit blanc à l'aide des mêmes tests.

B. Application des deux méthodes de prédiction de séries temporelles

Nous souhaitons prédire la dépense de l'année 2020 en date de décembre 2019. Pour ce faire, nous allons utiliser les bases de données produites à la fin de la partie IV. Nous avons donc une vision par mois de soin complète pour les survenances de 2015 à 2019 et une vision par mois de soin estimée pour l'année de survenance 2020.

Ainsi, nous allons diviser nos cinq bases de données, l'une par poste de consommation, selon le même schéma suivant :

- Les dépenses historiques de 2015 à 2019 serviront de données d'entraînement pour nos deux modèles ;
- Les dépenses développées à l'ultime de 2020 serviront de données de test pour nos deux modèles.

Il est donc important de noter que **nous entraînons nos modèles sur des données historiques (2015-2019) et comparons les prédictions obtenues à des données calculées (2020)**.

De plus, au vu des deux modèles appliqués et du positionnement temporel que nous prenons, nous considérons l'hypothèse forte suivant : **la consommation de 2020 suit un schéma similaire à ce qui a été observée précédemment**. Cette hypothèse est a priori fausse au regard du caractère spécial de l'année 2020. Les prédictions produites auraient pu ainsi être vues comme étant la consommation attendue si la pandémie du Covid-19 n'avait jamais eu lieu et qu'aucune réforme liée aux postes de consommation n'avait été adoptée. Cette première approche est donc une approche « naïve ».

Nous rappelons que nos méthodes seront appliquées sur les postes de consommation suivants :

- Soins de ville courants ;
- Pharmacie ;
- Dentaire ;
- Optique ;
- Prothèses auditives

Le poste *Soins de ville courants* servira d'exemple : la présentation des modèles et des hypothèses à vérifier sera donc plus fine. Cette présentation sera raccourcie pour les autres postes de consommation sauf en cas d'informations importantes à remarquer.

Lors de l'application des divers tests, nous fixons le seuil significatif de nos tests suivant : $\alpha = 5\ \%$.

1. Application d'ARIMA et de SARIMA : exemple sur le poste *Soins de ville courants*

Le poste de consommation *Soins de ville courants* peut être décomposé de la façon suivante via des statistiques par mois et année de dépense :

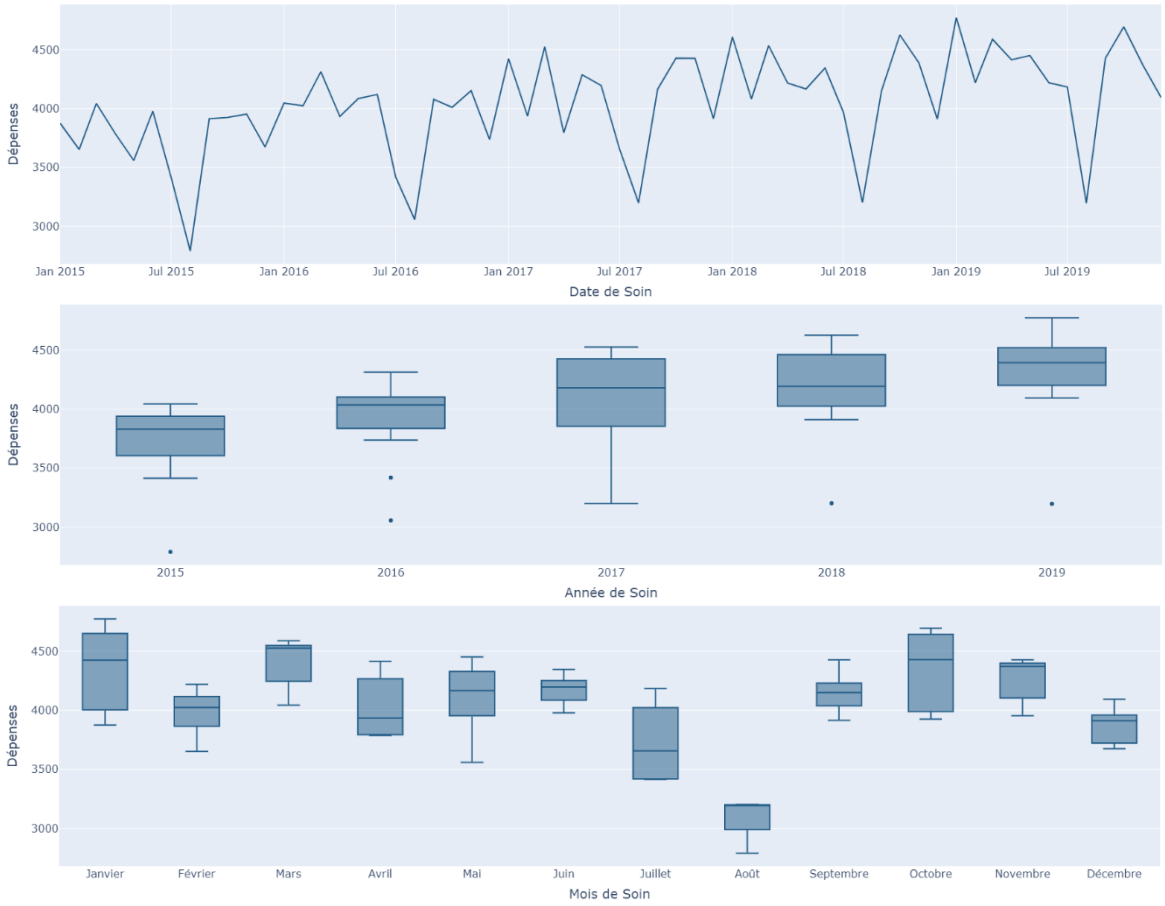


Figure V-3 – Décomposition de la série temporelle du poste de consommation Soins de ville courants (en millions d'euros)

a) La stationnarité de notre série temporelle

Les modèles $ARIMA$ et $SARIMA$ intègrent dans leur équation le concept de différenciation via la variable d. Cependant, après analyse et étude de ces deux modèles, nous avons jugé plus prudent et pertinent de fournir à ces deux modèles une série temporelle déjà stationnaire.

Nous allons donc vérifier que notre série temporelle est stationnaire à l'aide des tests ADF et KPSS. Sur la série sans aucun pré-traitement, nous obtenons les p-values suivantes :

	Test ADF	**Test KPSS**	**Conclusion du test**
P-value associée	0,377	0,026	Rejet de la stationnarité

Figure V-4 – Test de stationnarité sur la série temporelle Soins de ville courants avant prétraitement

Comme nous pouvons le voir, la p-value de ces deux tests complémentaires nous permet d'affirmer que la série n'est pas stationnaire. Pour la rendre stationnaire, nous allons procéder par différentiation d'ordre un. Cela consiste à appliquer un lag d'ordre un à l'intégralité de notre série.

Nous obtenons alors les p-values suivantes qui nous permettent de conclure que la série temporelle obtenue est bien stationnaire.

	Test ADF	Test KPSS	Conclusion du test
P-value associée	0,001	0,1	Série stationnaire

Figure V-5 – Test de stationnarité sur la série temporelle Soins de ville courants après prétraitement

Après avoir rendu stationnaire notre série, nous nous retrouvons avec un jeu de 59 données qui servira de basee d'apprentissage aux modèles $ARIMA$ et $SARIMA$.

b) Le modèle ARIMA

SARIMAX Results

Dep. Variable:	y	No. Observations:	59
Model:	SARIMAX(11, 0, 2)	Log Likelihood	-1198.571
		AIC	2427.143
		BIC	2458.306
		HQIC	2439.308

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	1.412e+08	1.01e-09	1.4e+17	0.000	1.41e+08	1.41e+08
ar.L1	-0.8774	0.093	-9.449	0.000	-1.059	-0.695
ar.L2	-0.9648	0.064	-15.189	0.000	-1.089	-0.840
ar.L3	-0.9308	0.069	-13.524	0.000	-1.066	-0.796
ar.L4	-0.8726	0.089	-9.819	0.000	-1.047	-0.698
ar.L5	-0.9670	0.066	-14.667	0.000	-1.096	-0.838
ar.L6	-0.8644	0.098	-8.789	0.000	-1.057	-0.672
ar.L7	-0.9357	0.057	-16.407	0.000	-1.048	-0.824
ar.L8	-0.8868	0.081	-10.987	0.000	-1.045	-0.729
ar.L9	-0.9171	0.068	-13.556	0.000	-1.050	-0.785
ar.L10	-0.9377	0.051	-18.481	0.000	-1.037	-0.838
ar.L11	-0.8539	0.091	-9.394	0.000	-1.032	-0.676
ma.L1	-0.4148	0.217	-1.909	0.056	-0.841	0.011
ma.L2	0.3800	0.191	1.989	0.047	0.006	0.754
sigma2	1.884e+16	3.76e-18	5.01e+33	0.000	1.88e+16	1.88e+16

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	0.99
Prob(Q):	0.93	Prob(JB):	0.61
Heteroskedasticity (H):	0.70	Skew:	0.15
Prob(H) (two-sided):	0.43	Kurtosis:	2.44

Figure V-6 – Modèle ARIMA pour le poste de consommation Soins de ville courants

Après application de la fonction $auto\_arima([\dots])$ sur notre jeu de données stationnaire, nous obtenons le modèle $ARIMA(11,0,2)$ résumé dans la figure ci-dessus.

Nous pouvons remarquer que quasiment tous les termes sont significatifs (Chaque terme a une valeur dans la colonne $P > |z|$ inférieur à 0,05 sauf le premier terme de moyenne mobile. Un modèle $ARIMA(11,0,1)$ – non présenté ici – a cependant généré beaucoup trop de termes non significatifs pour être conservé). L'équation du modèle est donc la suivante :

$$\hat{y}_t = 1{,}41 * 10^8 - 0{,}88 * \hat{y}_{t-1} - 0{,}96 * \hat{y}_{t-2} - 0{,}93 * \hat{y}_{t-3} - 0{,}87 * \hat{y}_{t-4} - 0{,}97 * \hat{y}_{t-5} - 0{,}86 * \hat{y}_{t-6} - 0{,}94 * \hat{y}_{t-7} - 0{,}89 * \hat{y}_{t-8} - 0{,}92 * \hat{y}_{t-9} - 0{,}94 * \hat{y}_{t-10} - 0{,}85 * \hat{y}_{t-11} + \epsilon_t - 0{,}41 * \epsilon_{t-1} + 0{,}38 * \epsilon_{t-2}$$

Avec $\hat{y}_t = y_t - y_{t-1}$.

Nous pouvons déjà remarquer que notre modèle $ARIMA$ tente de capter une saisonnalité. La profondeur de l'auto-régression semble en effet un peu excessive au vu de la taille de notre échantillon. L'application du modèle $SARIMA$ dans un second temps viendra confirmer cette hypothèse.

Nous devons également vérifier certaines hypothèses sur les résidus de notre modèle à l'aide des tests évoqués précédemment.

- **Corrélation des résidus :** nous pouvons voir que nos résidus ne sont pas corrélés entre eux. En effet, la p-value du test de *Q de Ljung-Box* (encadré en bleu dans la figure V-6), qui est de 0,93, est bien plus élevée que notre seuil significatif de rejet. Nous ne pouvons donc pas rejeter l'hypothèse nulle.
- **Étude des résidus :** nous vérifions que nos résidus suivent bien une loi normale à l'aide du test de *Jarque-Bera* dont la p-value vaut 0,61 (encadré en vert dans la figure V-6). Nous ne pouvons donc pas rejeter l'hypothèse nulle et pouvons admettre que les résidus suivent bien une loi normale.
- **Variance constante :** nous vérifions que la variance de nos résidus est constante au fil du temps à l'aide du test de *Breusch-Pagan*. La p-value de ce test est de 0,43 (encadré en jaune sur la figure V-6) et nous permet d'accepter avec prudence l'homoscédasticité de la variance de nos résidus.
- **Moyenne de nos résidus :** enfin, nous calculons la moyenne des résidus de notre modèle. Nous trouvons une valeur de $-25{,}65 * 10^6$.

Ces tests nous permettent d'affirmer théoriquement que les résidus sont du bruit blanc. Cependant dans la pratique, comme la moyenne de celui-ci n'est pas complètement nulle, nous devrons compenser en ajoutant ce montant à la projection de données. Cela nous permet d'admettre avec quelques réserves que nos résidus sont bien un bruit blanc.

Nous pouvons également effectuer une analyse visuelle de ces résidus à l'aide de la fonction $plot\_diagnostic()$ associée à la librairie $pmdarima$. L'analyse visuelle se compose de quatre graphiques dont :

- En haut à gauche : ce graphique nous permet d'observer la courbe des résidus. Nous pouvons ainsi identifier s'il existe une composante saisonnière dans nos résidus.
- En haut à droite : ce graphique projette la densité de probabilité de nos résidus (*KDE*) par rapport à la densité de probabilité d'une loi normale $\mathcal{N}(0,1)$.
- En bas à gauche : le graphique « Q-Q normal » nous permet de voir si la distribution ordonnée des résidus (en bleu) suit la tendance linéaire de la distribution d'une loi normale. Tout écart significatif signifierait que la distribution est asymétrique. (La valeur du $skew$ serait importante).
- En bas à droite : ce graphique est un corrélogramme. Il nous permet de voir s'il existe une corrélation entre les résidus. Toute autocorrélation impliquerait qu'il existe un sous-modèle dans nos résidus qui n'est pas expliqué par le modèle originel.

Finalement, la figure ci-dessous vient compléter l'analyse précédemment effectuée sur nos résidus. Dans notre cas, son analyse visuelle nous permet de confirmer notre intuition : les résidus de notre modèle suivent bien un bruit blanc. La moyenne de ce bruit blanc est cependant différente de zéro et doit donc être compensée.

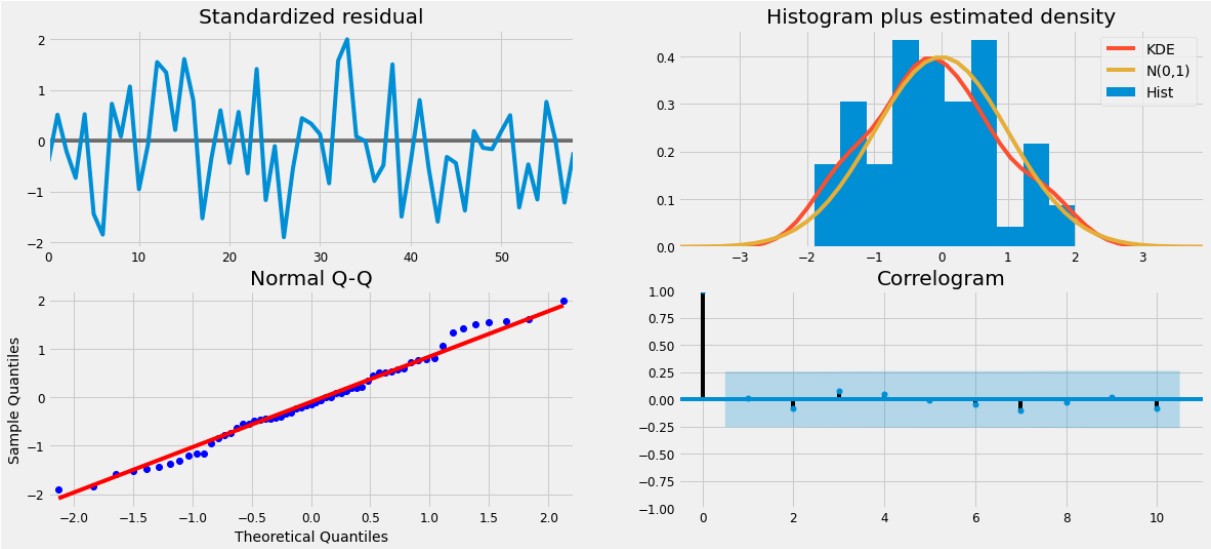


Figure V-7 – Analyse des résidus du modèle ARIMA du poste de consommation Soins de ville courants

c) Le modèle SARIMA

SARIMAX Results

Dep. Variable:	y	No. Observations:	59
Model:	SARIMAX(0, 0, 2)x(0, 1, [], 12)	Log Likelihood	-951.878
		AIC	1911.755
		BIC	1919.156
		HQIC	1914.540

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	-7.899e+06	3.61e-11	-2.19e+17	0.000	-7.9e+06	-7.9e+06
ma.L1	-1.0881	0.087	-12.465	0.000	-1.259	-0.917
ma.L2	0.6177	0.073	8.467	0.000	0.475	0.761
sigma2	2.02e+16	1.64e-19	1.23e+35	0.000	2.02e+16	2.02e+16

Ljung-Box (L1) (Q):	0.03	Jarque-Bera (JB):	0.92
Prob(Q):	0.87	Prob(JB):	0.63
Heteroskedasticity (H):	0.79	Skew:	-0.14
Prob(H) (two-sided):	0.65	Kurtosis:	2.37

Figure V-8 – Modèle SARIMA pour le poste de consommation Soins de ville courants

Dans un second temps, nous allons considérer qu'il existe un fort concept de saisonnalité dans nos données par l'application d'un modèle $SARIMA$. La figure ci-dessus résume ainsi les principales caractéristiques du modèle $SARIMA(0,0,2)(0,1,0)_{12}$ obtenu.

L'équation de ce modèle est la suivante :

$$\hat{y}_t = -7{,}9 * 10^6 + \hat{y}_{t-12} + \epsilon_t - 1{,}09 * \epsilon_{t-1} + 0{,}62 * \epsilon_{t-2}$$

Nous pouvons ainsi voir que l'application d'un modèle $SARIMA$ est indispensable afin de prendre en compte la forte saisonnalité présente dans notre jeu de données. La saisonnalité du modèle $SARIMA$ a ainsi remplacé tous les termes autorégressifs du modèle $ARIMA$.

Comme précédemment, nous vérifions que nos résidus suivent bien un bruit blanc. Les divers statistiques de test nous permettent de dire que les résidus ne sont pas corrélés entre eux, qu'ils suivent bien une loi normale et que la variance est constante. La moyenne de ces résidus est de $18{,}84 * 10^6$. Ainsi, nous admettons que nos résidus sont bien un bruit blanc.

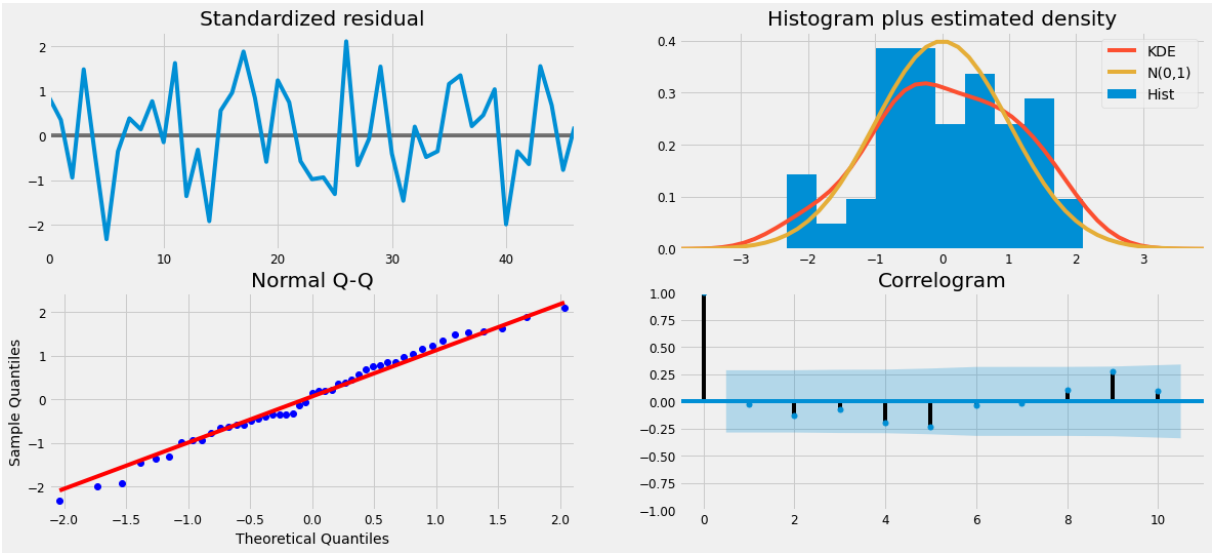


Figure V-9 – Analyse des résidus du modèle SARIMA du poste de consommation Soins de ville courants

L'application des modèles ARIMA et SARIMA sur les autres postes de consommation est disponible en annexe 4.

2. Conclusion sur nos modèles

Poste de consommation	Modèle $ARIMA$ retenu	Modèle $SARIMA$ retenu
Soins de ville courants	$ARIMA(11,0,2)$	$SARIMA(0,0,2)(0,1,0)_{12}$
Pharmacie	$ARIMA(11,0,1)$	$SARIMA(2,0,0)(1,1,0)_{12}$
Dentaire	$ARIMA(11,0,1)$	$SARIMA(1,0,1)(0,1,0)_{12}$
Optique	$ARIMA(11,0,0)$	$SARIMA(0,0,1)(3,1,0)_{12}$
Prothèses auditives	$ARIMA(11,0,1)$	$SARIMA(1,0,1)(1,1,0)_{12}$

Figure V-10 – Synthèse des modèles ARIMA et SARIMA retenus sur les cinq postes de consommation

A l'aide du tableau de synthèse ci-dessus, nous pouvons voir que tous les modèles $ARIMA$ présentent un terme d'auto régression élevé de 11 : tous nos postes de consommation présentent donc une saisonnalité. Nous pouvons également voir que le modèle ARIMA le plus fréquemment utilisé est le modèle $ARIMA(11,0,1)$ et concerne le poste de consommation *Pharmacie*, *Dentaire* et P*rothèses auditives*.

A l'inverse, il n'existe pas de modèle $SARIMA$ « type ». Nous pouvons seulement noter que tous les modèles $SARIMA$ retenus présentent un ordre de différenciation saisonnier, ce qui confirme notre intuition sur la saisonnalité de la dépense.

C. Prédictions sur 2020

1. Présentation des résultats obtenus après application de nos méthodes

Une fois les modèles $ARIMA$ et $SARIMA$ générés pour nos cinq postes de consommation, nous les utilisons afin d'obtenir la prévision des dépenses sur l'année 2020. Nous comparons nos projections à la consommation réelle estimée de 2020. Nous rappelons que les dépenses 2020 ont été estimées à l'aide de la méthode de Chain Ladder en considérant que la cadence de règlement de 2020 est similaire à celle qui a été observée en 2019. De plus, nous rappelons que l'erreur de la projection est calculée selon la formule suivante :

$$Erreur_i = Valeur\ attendue_i - Prédiction_i$$

L'interprétation est la suivante :

- Une valeur égale à zéro signifie que notre modèle a parfaitement projeté la consommation observée en i ;
- Une valeur supérieure à zéro signifie que notre modèle a sous-estimé la consommation observée en i ;
- Une valeur inférieure à zéro signifie que notre modèle a surestimé la consommation observée en i.

Nous effectuerons également nos comparaisons selon les périodes définies lors de la présentation des données. Nous rappelons ici les périodes de comparaison définies précédemment :

- Période n°1 : de janvier à février correspondant au pré-confinement ;
- Période n°2 : de mars à mai correspondant au premier confinement ;
- Période n°3 : de juin à septembre correspondant à la période entre les deux confinements ;
- Période n°4 : d'octobre à décembre correspondant au deuxième confinement.

Dans les tableaux proposés, une erreur en vert signifie que le modèle a sous-estimé la consommation tandis qu'une erreur en rouge signifie que le modèle a surestimé la consommation.

Chaque poste de consommation est donc analysé comme suit :

- Un tableau regroupant la valeur attendue, les prédictions du modèle ARIMA et SARIMA ainsi que la différence de cette projection par mois de soin sur l'année 2020 ;
- Une projection graphique de la valeur attendue et des prédictions du modèle afin de pouvoir réaliser une comparaison visuelle des différences observées dans le tableau ;
- Un deuxième tableau regroupant les résultats obtenus sur les périodes prédéfinis ;
- Une analyse textuelle de ces prédictions selon les périodes définies précédemment.

a) Soins de ville courants

Après avoir généré le modèle $ARIMA$ et $SARIMA$ pour le poste soins de ville courants à l'aide de notre base d'entraînement, nous sommes en mesure de projeter, dans un premier temps, la dépense attendue par mois de soin pour l'année 2020 puis de comparer cette projection à la dépense réellement observée. Les résultats obtenus sont résumés dans les trois figures suivantes

Figure V-11 – Prédictions des modèles ARIMA et SARIMA sur l'année 2020 pour le poste Soins de ville courants (en millions d'euros)

Nous pouvons ainsi voir que les deux modèles ont très bien appréhendé la dépense du premier mois de l'année mais qu'ils sous-estiment la dépense du mois de février. Comme prévu, nos modèles n'ont pas pu prévoir l'effondrement de la consommation sur la deuxième période. En revanche, La consommation observée sur la troisième période correspond à peu près à ce que nos modèles ont projeté. Finalement, la surconsommation qui semble se dessiner en fin d'année n'a pas pu être appréhendée par nos modèles.

Mois de soin de 2020	Dépense estimée à l'ultime	Prédiction ARIMA	Prédiction SARIMA	Erreur ARIMA	Erreur SARIMA	% erreur ARIMA	% erreur SARIMA
Janvier	4 837	4 847	4 867	**- 10**	**- 30**	**- 0,19 %**	**- 0,63 %**
Février	4 358	3 566	3 568	**792**	**790**	**22,20 %**	**22,14 %**
Mars	3 378	4 397	4 472	**- 1 018**	**- 1 094**	**- 23,17 %**	**- 24,47 %**
Avril	2 276	3 944	3 929	**- 1 670**	**- 1 655**	**- 42,34 %**	**- 42,13 %**
Mai	3 559	4 128	4 140	**- 569**	**- 581**	**- 13,79 %**	**- 14,04 %**
Juin	4 560	3 784	3 872	**816**	**727**	**21,56 %**	**18,78 %**
Juillet	4 412	4 026	4 067	**386**	**345**	**9,59 %**	**8,48 %**
Août	3 630	3 196	3 119	**434**	**512**	**13,58 %**	**16,40 %**
Septembre	5 019	5 164	5 332	**- 145**	**- 313**	**- 2,81 %**	**- 5,88 %**
Octobre	5 080	4 377	4 371	**703**	**709**	**16,07 %**	**16,22 %**
Novembre	5 009	3 735	3 781	**1 274**	**1 228**	**34,11 %**	**32,48 %**
Décembre	5 133	3 778	3 825	**1 356**	**1 308**	**35,85 %**	**34,19 %**
Total	**51 289**	**48 941**	**49 345**	**2 348**	**1 944**	**4,80 %**	**3,94 %**

Figure V-12 – Résultat de la prédiction du modèle ARIMA et SARIMA pour le poste de consommation soins de ville courants apprécié par mois de soin (en millions d'euros)

Le tableau ci-dessous regroupe les dépenses attendues et projetées selon les périodes prédéfinies :

	Dépense estimée à l'ultime	Prédiction ARIMA	Prédiction SARIMA	Erreur ARIMA	Erreur SARIMA	% erreur ARIMA	% erreur SARIMA
Période n°1	9 195	8 413	8 436	**782**	**759**	**9,30 %**	**9,00 %**
Période n°2	9 210	12 468	12 541	**- 3 258**	**- 3 331**	**- 26,13 %**	**- 26,56 %**
Période n°3	17 661	16 170	16 391	**1 491**	**1 270**	**9,22 %**	**7,75 %**
Période n°4	15 222	11 890	11 977	**3 332**	**3 245**	**28,02 %**	**27,09 %**
Total	**51 289**	**48 941**	**49 345**	**2 348**	**1 944**	**4,80 %**	**3,94 %**

Figure V-13 – Prédictions 2020 pour le poste de consommation soins de ville courants selon les périodes prédéfinies (en millions d'euros)

En analysant par périodes, il est important de remarquer que l'écart entre la prédiction et l'estimé de la deuxième période est du même ordre de grandeur que celui constaté dans la quatrième période. Nous pouvons intuitivement dire que la consommation n'ayant pas été effectuée lors du premier confinement a été « rattrapée » en fin d'année. Cependant, une telle conclusion est à prendre avec du recul dans la mesure où nous comparons notre projection à une dépense estimée et non à une dépense réelle.

b) *Pharmacie*

Le poste *Pharmacie* est le poste dont la projection est la plus satisfaisante. Le graphe ci-dessous montre les prédictions ARIMA – SARIMA et les valeurs estimées.

Figure V-14 – Prédictions des modèles ARIMA et SARIMA sur l'année 2020 pour le poste Pharmacie (en millions d'euros)

Cela peut notamment s'expliquer par le fait que les pharmacies sont en grande majorité restées ouvertes tout au long de l'année contrairement aux opticiens – par exemple – qui ont été contraints de fermer.

Mois de soin de 2020	Dépense estimée à l'ultime	Prédiction ARIMA	Prédiction SARIMA	Erreur ARIMA	Erreur SARIMA	% erreur ARIMA	% erreur SARIMA
Janvier	3 142	3 089	3 223	**53**	**- 81**	**1,70 %**	**- 2,50 %**
Février	2 857	2 598	2 503	**258**	**354**	**9,95 %**	**14,14 %**
Mars	3 089	2 984	3 161	**105**	**- 72**	**3,52 %**	**- 2,29 %**
Avril	2 610	2 916	2 768	**- 306**	**- 158**	**- 10,51 %**	**- 5,72 %**
Mai	2 634	2 784	2 912	**- 149**	**- 278**	**- 5,37 %**	**- 9,53 %**
Juin	2 886	2 825	2 874	**61**	**12**	**2,15 %**	**0,43 %**
Juillet	3 007	3 112	3 040	**- 105**	**- 33**	**- 3,36 %**	**- 1,08 %**
Août	2 687	2 595	2 605	**912**	**82**	**3,53 %**	**3,19 %**
Septembre	3 002	2 997	3 029	**5**	**- 27**	**0,18 %**	**- 0,90 %**
Octobre	3 388	3 187	3 309	**202**	**80**	**6,33 %**	**2,41 %**
Novembre	3101	2 714	2 689	**387**	**411**	**14,26 %**	**15,30 %**
Décembre	3 354	2 832	2 838	**522**	**515**	**18,41 %**	**18,16 %**
Total	**35 756**	**34 633**	**34 950**	**1 123**	**806**	**3,24 %**	**2,31 %**

Figure V-15 – Résultat de la prédiction du modèle ARIMA et SARIMA pour le poste de consommation Pharmacie apprécié par mois de soin (en millions d'euros)

Nous pouvons ainsi voir que sur la première période, nos deux modèles arrivent à capter la tendance globale de la consommation. La consommation observée lors de la deuxième période reste au sein des intervalles de confiance de nos modèles. Néanmoins, nous pouvons observer une légère baisse de la consommation en avril et en mai : lors du premier confinement de la pandémie du Covid-19, la peur d'être contaminé par le virus était très forte, incitant alors les individus à rester chez eux au maximum. La consommation observée sur la troisième période est très proche de celle projetée par nos modèles.

	Dépense estimée à l'ultime	Prédiction ARIMA	Prédiction SARIMA	Erreur ARIMA	Erreur SARIMA	% erreur ARIMA	% erreur SARIMA
Période n°1	5 999	5 688	5 726	**311**	**273**	**5,47 %**	**4,77 %**
Période n°2	8 333	8 684	8 841	**- 351**	**- 508**	**- 4,04 %**	**- 5,75 %**
Période n°3	11 582	11 529	11 548	**53**	**34**	**0,46 %**	**0,30 %**
Période n°4	9 843	8 732	8 836	**1 110**	**1 006**	**12,71 %**	**11,39 %**
Total	**35 756**	**34 633**	**34 950**	**1 123**	**806**	**3,24 %**	**2,31 %**

Figure V-16 – Prédictions 2020 pour le poste de consommation Pharmacie selon les périodes prédéfinies (en millions d'euros)

Finalement, les modèles sous-estiment de façon importante la consommation sur la quatrième période. Plusieurs facteurs peuvent rentrer en jeu pour expliquer cette surconsommation, l'une des explications étant la mise à disposition des tests de dépistage contre le Covid-19.

En fin de compte, la surconsommation que nous observons en fin d'année nous invite à nous demander si cette surconsommation était passagère ou si, au contraire, elle continuera sur l'année 2021 avec la mise en place du troisième confinement. La mise à disposition de vaccins contre le Covid-19 en France au début de l'année 2021 viendra, dans tous les cas, créer une consommation « anormale » à ce que nous aurions dû observer si la pandémie du Covid-19 n'avait jamais eu lieu.

c) *Dentaire*

En nous intéressant de plus près au poste *Dentaire*, nous remarquons rapidement que nos deux modèles ont très bien projeté la consommation sur la première période de l'année, les résultats montrés dans le graphe ci-dessous.

Figure V-17 – Projection des prédictions des modèles ARIMA et SARIMA sur l'année 2020 pour le poste Dentaire (en millions d'euros)

L'erreur cumulée sur ces deux premiers mois de l'année est négligeable. De ce premier résultat, nous pouvons admettre que la consommation projetée par les modèles que nous observons sur le reste de l'année 2020 correspond à peu près à ce que nous aurions dû observer en l'absence de pandémie du Covid-19.

Mois de soin de 2020	Dépense estimée à l'ultime	Prédiction ARIMA	Prédiction SARIMA	Erreur ARIMA	Erreur SARIMA	% erreur ARIMA	% erreur SARIMA
Janvier	917	910	920	**7**	**- 2**	**0,80 %**	**- 0,23 %**
Février	956	946	960	**10**	**- 4**	**1,08 %**	**- 0,43 %**
Mars	611	1 205	1 189	**- 594**	**- 578**	**- 49,32 %**	**- 48,62 %**
Avril	88	691	694	**- 604**	**- 606**	**- 87,33 %**	**- 87,38 %**
Mai	675	1 017	1 030	**- 341**	**- 355**	**- 33,58 %**	**- 34,44 %**
Juin	1 203	982	968	**221**	**235**	**22,48 %**	**24,29 %**
Juillet	1 153	995	1 023	**158**	**130**	**15,85 %**	**12,67 %**
Août	553	405	348	**149**	**205**	**36,72 %**	**59,09 %**
Septembre	1 059	1 403	1 446	**- 344**	**- 388**	**- 24,54 %**	**- 26,89 %**
Octobre	1 095	1 103	1 088	**- 7**	**7**	**- 0,67 %**	**0,69 %**
Novembre	1 180	817	825	**363**	**355**	**44,46 %**	**43,01 %**
Décembre	1 251	926	933	**325**	**318**	**35,06 %**	**34,12 %**
Total	**10 741**	**11 400**	**11 423**	**- 658**	**- 682**	**- 5,78 %**	**- 5,97 %**

Figure V-18 – Résultat de la prédiction du modèle ARIMA et SARIMA pour le poste de consommation Dentaire apprécié par mois de soin (en millions d'euros)

Sans surprise, la prédiction de la dépense de nos modèles sur la deuxième période surestime largement la consommation réellement observée, les cabinets dentaires ayant dû fermer lors du premier confinement. La consommation observée sur la troisième période atteint les niveaux de prédiction des modèles, même si nous observons une légère surconsommation de juin à août et une importante sous-consommation en septembre.

Finalement, nos modèles sous-estiment largement la consommation observée sur la quatrième période. En effet, il semblerait qu'un léger rattrapage de la consommation non effectuée sur la

deuxième période a eu lieu lors de cette quatrième période, sans que ce rattrapage ne puisse compenser l'intégralité des soins non effectués lors de la deuxième période. Nous pouvons finalement nous demander si un rattrapage complet de la dépense non effectuée lors de la deuxième période aura lieu en 2021.

	Dépense estimée à l'ultime	Prédiction ARIMA	Prédiction SARIMA	Erreur ARIMA	Erreur SARIMA	% erreur ARIMA	% erreur SARIMA
Période n°1	1 873	1 856	1 8780	**18**	**- 6**	**0,95 %**	**- 0,33 %**
Période n°2	1 374	2 913	2 913	**- 1 539**	**- 1 539**	**- 52,84 %**	**- 52,84 %**
Période n°3	3 967	3 786	3 785	**183**	**183**	**4,83 %**	**4,83 %**
Période n°4	3 527	2 846	2 845	**681**	**681**	**23,91 %**	**23,94 %**
Total	**10 741**	**11 400**	**11 423**	**- 658**	**- 682**	**- 5,78 %**	**- 5,97 %**

Figure V-19 – Résultat de la prédiction du modèle ARIMA et SARIMA pour le poste de consommation Dentaire apprécié selon les périodes prédéfinies (en millions d'euros)

d) Optique

Le poste *Optique* représente le poste de consommation où les deux modèles retenus ont systématiquement surestimé la consommation réellement observée, y compris sur la quatrième période.

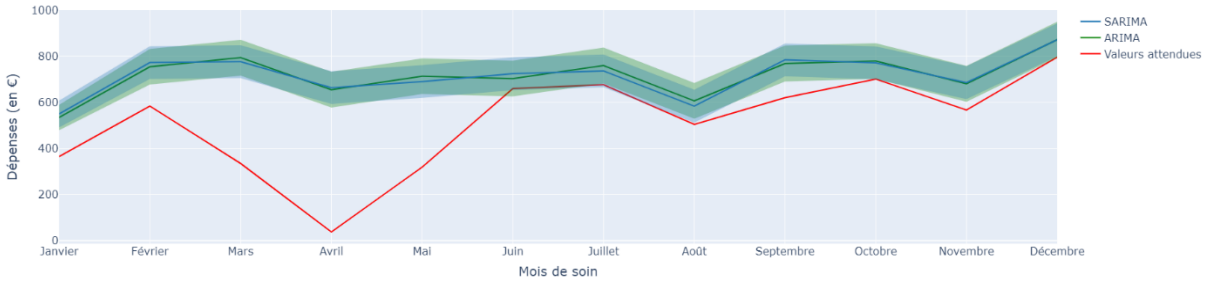


Figure V-20 – Projection des prédictions des modèles ARIMA et SARIMA sur l'année 2020 pour le poste Optique (en millions d'euros)

Les prédictions sont particulièrement éloignées sur la première période où l'erreur de projection sur ces deux premiers mois est la plus importante de l'année (en retirant l'erreur de projection engendrée par la pandémie du Covid-19). Cette importante erreur de projection peut notamment être due aux problèmes qui ont émergé suite à la mise en place de la réforme 100% Santé et du tiers payant. En effet, l'écart entre la prédiction et les valeurs attendues en janvier ne peut pas être expliqué par la pandémie.

Mois de soin de 2020	Dépense estimée à l'ultime	Prédiction ARIMA	Prédiction SARIMA	Erreur ARIMA	Erreur SARIMA	% erreur ARIMA	% erreur SARIMA
Janvier	365	534	550	**- 169**	**- 185**	**- 31,72 %**	**- 33,64 %**
Février	583	754	772	**- 171**	**- 189**	**- 22,64 %**	**- 24,48 %**
Mars	335	793	776	**- 459**	**- 442**	**- 57,81 %**	**- 56,88 %**
Avril	38	655	663	**- 617**	**- 625**	**- 94,19 %**	**- 94,27 %**
Mai	319	713	690	**- 394**	**- 371**	**- 55,29 %**	**- 53,79 %**
Juin	659	703	724	**- 44**	**- 65**	**- 6,22 %**	**- 8,94 %**
Juillet	676	759	736	**- 83**	**- 60**	**- 10,96 %**	**8,16 %**
Août	504	606	583	**- 102**	**- 79**	**- 16,87 %**	**- 13,59 %**
Septembre	620	768	784	**- 148**	**- 164**	**- 19,26 %**	**- 20,89 %**
Octobre	701	779	771	**- 78**	**- 70**	**- 10,01 %**	**- 9,13 %**
Novembre	567	680	685	**- 113**	**- 119**	**- 16,64 %**	**- 17,30 %**
Décembre	796	872	871	**- 76**	**- 75**	**- 8,72 %**	**- 8,66 %**
Total	**6 162**	**8 616**	**8 606**	**- 2 454**	**- 2 444**	**- 28,48 %**	**- 28,40 %**

Figure V-21 – Résultat de la prédiction du modèle ARIMA et SARIMA pour le poste de consommation Optique apprécié par mois de soin (en millions d'euros)

Comme observé sur le poste *Dentaire*, la consommation en *Optique* sur la deuxième période s'effondre. En effet, les opticiens n'ont pas pu rester ouverts lors de ce premier confinement entraînant mécaniquement une baisse de la consommation.

	Dépense estimée à l'ultime	Prédiction ARIMA	Prédiction SARIMA	Erreur ARIMA	Erreur SARIMA	% erreur ARIMA	% erreur SARIMA
Période n°1	948	1 288	1 323	**- 340**	**- 374**	**- 26,41 %**	**- 28,29 %**
Période n°2	692	2 161	2 130	**- 1 470**	**- 1 438**	**- 68,00 %**	**- 67,53 %**
Période n°3	2 459	2 837	2 827	**- 377**	**- 368**	**- 13,30 %**	**- 13,01 %**
Période n°4	2 063	2 330	2 327	**- 267**	**- 264**	**- 11,46 %**	**- 11,36 %**
Total	**6 162**	**8 616**	**8 606**	**- 2 454**	**- 2 444**	**- 28,48 %**	**- 28,40 %**

Figure V-22 – Résultat de la prédiction du modèle ARIMA et SARIMA pour le poste de consommation Optique apprécié selon les périodes prédéfinies (en millions d'euros)

Enfin, nous observons sur les deux dernières périodes de l'année, que la consommation réelle semble rejoindre la consommation prédite sans jamais la dépasser. Ainsi, nous pouvons supposer que la consommation non effectuée lors de la période du premier confinement a été rattrapée en partie par la suite. De plus, il semblerait que la réforme 100% Santé soit venue baisser le coût du panier moyen en optique. Nous nous retrouvons ainsi avec deux phénomènes influant fortement sur la consommation et dont il est difficile de mesurer l'impact avec précision. Comme précédemment, il semblerait que malgré la pandémie de Covid-19, la réforme 100% Santé soit venue bousculer l'hypothèse centrale de nos modèles : la consommation de 2020 est difficile à prévoir si nous utilisons l'historique de la consommation passée. Nous observons ainsi un changement de tendance certain.

e) *Prothèses auditives*

Pour le poste de consommation Prothèses auditives, nous retrouvons des tendances de résultat assez similaire à ce que nous avons pu observer pour le poste *Dentaire*.

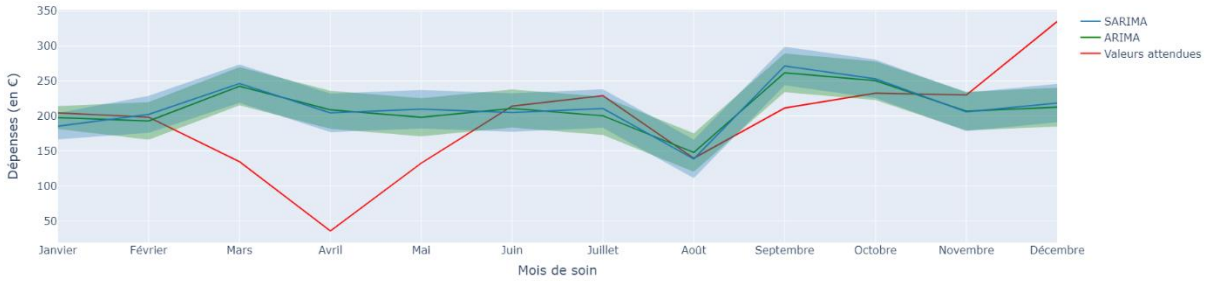


Figure V-23 – Projection des prédictions des modèles ARIMA et SARIMA sur l'année 2020 pour le poste Prothèses auditives (en millions d'euros)

En effet, nos modèles sont très satisfaisants sur la première période : la consommation projetée par les deux modèles est très proche de la consommation réellement observée. Une conclusion similaire peut en être tirée : la consommation projetée par les modèles sur l'année 2020 correspond à peu près à ce que nous aurions dû observer en l'absence de pandémie de Covid-19.

Mois de soin de 2020	Dépense estimée à l'ultime	Prédiction ARIMA	Prédiction SARIMA	Erreur ARIMA	Erreur SARIMA	% erreur ARIMA	% erreur SARIMA
Janvier	205	198	185	**7**	**19**	**3,38 %**	**10,26 %**
Février	198	193	202	**5**	**- 4**	**2,79 %**	**- 1,99 %**
Mars	135	243	246	**- 108**	**- 112**	**- 44,45 %**	**- 45,29 %**
Avril	36	209	204	**- 17**	**- 169**	**- 82,86 %**	**- 82,48 %**
Mai	133	198	210	**- 66**	**- 77**	**- 33,05 %**	**- 36,75 %**
Juin	214	211	205	**3**	**9**	**1,64 %**	**4,51 %**
Juillet	229	200	211	**29**	**18**	**14,33 %**	**8,56 %**
Août	140	148	139	**- 8**	**1**	**- 5,56 %**	**0,77 %**
Septembre	211	262	272	**- 50**	**- 60**	**- 19,21 %**	**- 22,12 %**
Octobre	233	250	253	**- 18**	**- 20**	**- 7,09 %**	**- 8,08 %**
Novembre	230	207	206	**23**	**24**	**11,14 %**	**11,69 %**
Décembre	335	213	219	**122**	**117**	**57,55 %**	**53,32 %**
Total	**2 298**	**2 531**	**2 552**	**- 233**	**- 254**	**- 9,21 %**	**- 9,94 %**

Figure V-24 – Résultat de la prédiction du modèle ARIMA et SARIMA pour le poste de consommation Prothèses auditives apprécié par mois de soin (en millions d'euros)

	Dépense estimée à l'ultime	Prédiction ARIMA	Prédiction SARIMA	Erreur ARIMA	Erreur SARIMA	% erreur ARIMA	% erreur SARIMA
Période n°1	403	391	388	**12**	**15**	**3,09 %**	**3,87 %**
Période n°2	303	650	661	**- 346**	**- 357**	**- 53,32 %**	**- 54,09 %**
Période n°3	794	821	826	**- 26**	**- 32**	**- 3,21 %**	**- 3,82 %**
Période n°4	798	670	678	**128**	**120**	**19,06 %**	**17,73 %**
Total	**2 298**	**2 531**	**2 552**	**- 233**	**- 254**	**- 9,21 %**	**- 9,94 %**

Figure V-25 – Résultat de la prédiction du modèle ARIMA et SARIMA pour le poste de consommation Prothèses auditives apprécié selon les périodes prédéfinies (en millions d'euros)

Sur la deuxième période, les conclusions sont similaires à ce que nous avons pu voir précédemment : la consommation en santé a drastiquement chuté et cela n'a pas été prévu par nos modèles comme nous pouvions nous y attendre.

La consommation réelle observée sur la troisième période rejoint globalement la projection de nos modèles sur cette même période.

Enfin, nous pouvons observer un début de rattrapage de la consommation sur la quatrième période, sans que ce rattrapage ne puisse égaler la dépense non effectuée sur la deuxième période. Ce tte conclusion est comme précédemment à prendre avec du recul dans la mesure où la consommation observée sur la quatrième période correspond principalement à une consommation estimée et non à une consommation réelle.

Néanmoins, nous pouvons nous demander si le poste *Prothèses auditives* sera également concerné par un changement de tendance de la dépense en 2021, similaire à ce que nous avons pu observer pour le poste *Optique*. Est-ce que la hausse de la consommation observée en fin d'année 2020 va continuer sur l'année 2021 ou bien est-ce que la réforme 100% Santé viendra baisser le coût du panier moyen et par conséquent la consommation globale ? Nous prendrons ce point en compte dans la partie suivante.

f) Conclusion des prédictions sur 2020

Au global, il en ressort que les modèles ARIMA et SARIMA projettent convenablement la consommation observée sur la première période. Il est donc possible de projeter la consommation future en fonction des données passées. Sur la seconde période, il est évident que nos modèles ne captent pas la baisse des dépenses liées au confinement. Les modèles n'ont pas été construits pour prendre en considération l'impact du Covid-19 sur la consommation en santé.
Nos modèles sont relativement pertinents sur la troisième période : la consommation estimée est assez proche des niveaux des prédictions des modèles, sans toutefois identifier un évident rattrapage de la consommation. De façon générale, nos modèles sous-estiment la dépense appréciée à l'ultime sur la quatrième période de l'année : un rattrapage plus important est observé en fin d'année.

2. Limites de cette première approche

Il était a priori évident que l'approche développée n'était pas suffisante pour prédire correctement la dépense de 2020 à cause du Covid-19. Nous avons ainsi pu voir que le fait de considérer que la consommation de 2020 allait être dans la continuité des années précédentes n'est pas une hypothèse satisfaisante : les prédictions réalisées sur la consommation de 2020 à l'aide du modèle ARIMA et SARIMA ne peuvent pas prendre en compte le caractère atypique et totalement imprévisible de l'année 2020.

Cependant, cette première approche nous a permis de constater que nos modèles sont suffisamment satisfaisants sur la période pré-covid puisque, comme nous avons pu le voir, seules les prédictions sur le poste *Optique* ont été sensiblement fausses sur la première période.

Cette première approche nous a permis de facilement identifier les deux impacts que nous devrons impérativement prendre en compte dans la partie suivante afin d'affiner nos prédictions :

- L'impact de la réforme 100 % Santé sur certains postes de consommation ;
- L'impact du COVID 19 relativement homogène sur tous les postes de consommation en santé ;

Ces limites s'opposent. Une réforme telle que celle du 100 % Santé est une réforme prévisible et dont les effets peuvent être étudiés à l'avance. Même de façon incomplète, il est possible de mesurer les impacts d'une telle réforme en amont et d'estimer les tendances futures de la dépense.

A l'inverse, un évènement tel que celui que nous avons vécu depuis mars 2020 et ses conséquences étaient des phénomènes difficilement prévisibles. Les impacts se mesurent donc une fois l'évènement passé et lorsque les retombées commencent à être identifiées. Comme nous avons pu l'évoquer lors de la deuxième partie de ce mémoire, il est encore difficile de mesurer l'impact qu'aura eu un tel évènement sur le dépistage des cancers et sur le retard des traitements.

Il faut également rappeler que nous comparons nos projections à des montants de consommation approximés. La méthode de Chain Ladder retenue pourrait notamment faire l'objet d'une analyse plus approfondie afin de comparer nos projections avec les dépenses estimées les plus exactes possibles.

Nous sommes donc amenés à prendre en compte ces limites. Dans la dernière partie, une nouvelle approche sera étudiée. De nouvelles prédictions de séries temporelles obtenues à partir du modèle *Prophet*, développé par des ingénieurs de Facebook, seront étudiées.

VI. Mise en place d'un nouveau modèle : Prendre en compte le caractère exceptionnel de 2020

A. Présentation du modèle mathématique de Prophet

L'année 2020, de par son caractère particulier, est venu casser les dynamiques observées les années passées. Il est donc nécessaire de prendre en compte les spécificités qui découlent de cette année.

Pour se faire nous allons faire appel à Prophet[20], un outil de prédiction des séries temporelles qui est le fruit du travail de l'équipe de Data Science de Facebook. Ce modèle mathématique a fait l'objet d'un article scientifique[21] que nous présenterons plus en détail par la suite. Nous allons étudier ce modèle et notamment voir ce qui le rend différent des modèles ARIMA et SARIMA utilisés dans la partie précédente.

Notre objectif reste le même : nous cherchons à projeter la dépense liée aux cinq postes de consommation étudiés en se basant sur le concept des séries temporelles.

Nous avons fait le choix d'utiliser Prophet car ce modèle semble avoir été pensé pour ce genre de situation où nous observons des « changements de tendance » et où les données utilisées présentent une forte saisonnalité.

Prophet est défini – par ses créateurs Sean J. Taylor et Benjamin Letham – comme étant un « *modèle de régression modulable avec des paramètres interprétables qui peuvent être ajustés intuitivement par des analystes ayant une connaissance du domaine des séries temporelles* ». C'est ce que nous allons voir maintenant.

1. Équation du modèle

Le modèle mathématique de Prophet – dans sa forme la plus simple – correspond à une décomposition additive similaire à ce que nous avons pu voir dans la partie V. A. 2. b – le modèle additif et que nous rappelons : $y(t) = Niveau + Tendance + Saisonnalité + Bruit$.

Prophet adapte cette décomposition en mettant de côté le niveau et en ajoutant deux composantes supplémentaires. L'équation du modèle Prophet se décompose donc en quatre composantes principales auquel il faut rajouter le terme d'erreur. L'équation de ce modèle additif est la suivante :

$$y(t) = g(t) + s(t) + h(t) + r(t) + \epsilon_t$$

Avec :

- $g(t)$ représente la **tendance** de la série temporelle, modélisant tous les changements qui ne sont pas périodiques ;
- $s(t)$ représente la **saisonnalité**, les changements périodiques que nous pouvons observer dans la série temporelle ;
- $h(t)$ représente l'effet que peuvent avoir **des vacances ou des évènements inattendus** qui se produisent sur des périodes irrégulières d'une année sur l'autre ;
- $r(t)$ représente les **régresseurs**. C'est-à-dire les données externes à la série temporelle ajoutées par l'utilisateur qui pourraient avoir une influence sur celle-ci ;

[20] https://facebook.github.io/prophet/
[21] https://peerj.com/preprints/3190/

- ϵ_t représente le **terme d'erreur**, tout ce qui n'a pas pu être correctement expliqué par un des quatre composants du modèle, la seule hypothèse sur ce terme d'erreur étant sa distribution normale.

Chaque composante de l'équation peut être décomposée en une équation supplémentaire comme nous allons le voir.

Nous considérons pour la suite que nous étudions la série temporelle $Y = [y_1, \dots, y_n]$

a) La tendance

La tendance de la série peut être de deux formes :

- Un modèle de croissance par saturation ;
- Un modèle linéaire par morceaux.

Nous allons principalement nous intéresser au modèle linéaire par morceaux comme il s'agit du modèle que nous mettrons en place par la suite.

Le modèle linéaire par morceaux

L'équation du modèle linéaire par morceaux est la suivante :

$$g(t) = (k + a(t)\delta^T)t + (m + a(t)\gamma^T)$$

Il s'agit d'un modèle de régression linéaire simple et dont la pente de la droite peut être amenée à changer au cours du temps. Ainsi, Prophet a la possibilité d'adapter la tendance aux changements observables dans la série temporelle. Dans le modèle, nous pouvons distinguer deux parties (une en rouge, l'autre en bleu).

(a) La pente de l'équation

k représente la pente de l'équation. Nous pouvons le voir comme étant la vitesse avec laquelle le modèle croît au fil du temps. La vitesse de croissance n'est pas constante au fil du temps, d'où l'introduction d'un paramètre supplémentaire $a(t)\delta^T$. Ce paramètre nous permet de définir les points d'inflexion où la pente va être modifiée.

Considérons qu'il y a S points d'inflexion au temps $s_{j,j=1,\dots,S}$. À chacun de ces temps, nous définissons le vecteur d'ajustement de taux $\delta \in \mathbb{R}^S$ où δ_j représente le changement de taux se produisant en s_j. La pente de l'équation, à tout instant *t*, est donc $k + \sum_{j:t>s_j} \delta_j$. Les auteurs ont fait le choix de rendre cette approche plus compréhensible en définissant un vecteur $a(t) \in \{0,1\}^S$ tel que

$$a_j(t) = \begin{cases} 1 \; si \; t \geq s_j \\ 0 \; si \; t < s_j \end{cases}$$

Nous retombons alors sur l'équation générale de la pente de l'équation définie précédemment $(k + a(t)\delta^T)t$.

(b) Le décalage de l'équation

A partir de ce que nous avons vu dans la partie précédente, l'équation de la tendance passe par l'origine, ce qui n'est pas forcément le cas. Nous introduisons alors la variable m qui correspond à la valeur d'origine de notre modèle. Lors de chaque changement dans la tendance, il est nécessaire d'introduire un terme supplémentaire $a(t)\gamma^T$, afin de conserver la continuité du modèle linéaire. Nous définissons donc un nouveau vecteur d'ajustement de taux $\gamma \in \mathbb{R}^S$ où $\gamma_j = -s_j\delta_j$.

Nous définissons, a priori, que les points de changements δ_j suivent une loi de $Laplace(0, \tau)$, sauf indication contraire de l'utilisateur. Le paramètre τ contrôle ainsi la flexibilité du modèle à modifier sa tendance et à s'adapter à tout changement. Foncièrement, plus le τ est élevé, plus le modèle présentera des points d'inflexion. Afin d'éviter du sur apprentissage, l'utilisateur doit donc choisir un τ peu élevé.

(c) L'estimation des paramètres

Le modèle identifie tout d'abord le coefficient directeur k et l'ordonnée à l'origine m de la tendance. L'estimation de ces paramètres est simple et est effectuée comme suit :

- $k = y_n - y_1$
- $m = y_1$

Si l'utilisateur ne précise pas de point d'inflexion, le modèle considère de nombreux points d'inflexion (un point en début de chaque mois par exemple lorsque le jeu de données s'étale sur plusieurs années) et pose comme hypothèse que les vecteurs d'ajustements δ_j suivent une loi de $Laplace(0, \tau)$. De base, τ est supposé égal à 0,05 mais l'utilisateur a la possibilité de le modifier. De plus, plus τ tend vers 0, plus le modèle se rapproche d'une simple droite linéaire. A l'inverse, plus τ augmente, plus la tendance va se rapprocher de la série temporelle d'origine, nous exposant alors à un risque de surapprentissage.

b) La saisonnalité

Comme nous avons notamment pu le voir dans la *partie IV*, certains postes de consommation présentent une forte saisonnalité, ce qui explique notamment pourquoi nous avons fait le choix d'utiliser un modèle $SARIMA$ en complément du modèle $ARIMA$.

Dans notre cas, nous pouvons facilement identifier l'effet induit par les vacances d'été sur la baisse de la consommation. Les modèles de saisonnalité sont donc des fonctions périodiques de t. Afin de modéliser la saisonnalité, le modèle s'appuie sur les séries de Fourier, ce qui permet alors de fournir un modèle flexible aux effets périodiques.

Nous définissons tout d'abord la variable P, qui représente la période de notre jeu de données. Comme nous souhaitons identifier la saisonnalité annuelle, nous devons fixer $P = 365{,}25$.

L'équation de saisonnalité se définit comme suit :

$$s(t) = \sum_{n=1}^{N} \left[a_n \cos\left(\frac{2\pi n t}{P}\right) + b_n \sin\left(\frac{2\pi n t}{P}\right) \right]$$

L'estimation des paramètres du modèle

A partir de l'équation définie précédemment, nous pouvons voir qu'il est donc nécessaire d'évaluer $2 * N$ paramètres à partir de la matrice de saisonnalité suivante :

$$X_t = \left[\cos\left(\frac{2\pi(1)t}{P}\right), \sin\left(\frac{2\pi(1)t}{P}\right), \dots, \dots, \cos\left(\frac{2\pi(N)t}{P}\right), \sin\left(\frac{2\pi(N)t}{P}\right) \right]$$

La saisonnalité est ensuite définie à l'aide de l'équation suivante : $s(t) = \beta * X(t)$ où nous considérons – a priori – que β suit une loi normale centrée en 0 et de variance σ^2. La variance est fixée à $\sigma^2 = 10$ au lancement du modèle, ce qui permet d'imposer un lissage préalable sur la saisonnalité.

Selon les tests qui ont pu être effectués par les personnes à l'origine de ce modèle, il en est ressorti que pour des données présentant une saisonnalité annuelle, fixer $N = 10$ permet d'obtenir une saisonnalité efficace tout en évitant le risque de sur-apprentissage sur les données d'entraînement.

c) Les vacances et évènements inhabituels

Enfin, la troisième composante de l'équation représente tous les évènements qui ne peuvent pas être modélisés à l'aide de la composante saisonnalité. Il s'agit des évènements ayant lieu dans l'année mais dont la date exacte est amenée à changer. C'est le cas en France du jour férié lié à la Fête de Pâques, pouvant avoir lieu entre le 22 mars et le 25 avril, par opposition à la Fête Nationale ayant lieu tous les ans le 14 juillet.

Pour chaque évènement $i, i = 1, \dots L$ dont la date est invitée à évoluer au fil du temps, nous posons D_i, la date de cet évènement. Nous considérons par la suite une fonction indicatrice afin d'identifier si la date t correspond à la date de l'évènement D_i. Nous attribuons par la suite à chaque évènement inhabituel un paramètre κ_i correspondant au changement dans la prévision.

L'identification des paramètres κ_i s'effectue à travers la génération d'une matrice de régresseur $Z(t)$ telle que :

$$Z(t) = [\mathbb{1}(t \in D_1), \dots, \mathbb{1}(t \in D_L)]$$

Puis, à considérer que $h(t) = \kappa * Z(t)$. Comme pour la saisonnalité, nous considérons a priori que $\kappa \sim \mathcal{N}(0, \nu^2)$

Nous ne serons pas amenés à utiliser cette troisième composante. En effet, comme nous avons pu l'observer précédemment, la saisonnalité du mois d'août est suffisamment marquée sans besoin de le repréciser.

d) L'ajout de régresseur

Le modèle Prophet permet d'intégrer des variables externes pouvant avoir un quelconque effet sur la série temporelle. C'est notamment ce qui nous permettra de prendre en compte les effets du Covid-19 sur les dépenses des postes étudiés comme nous le verrons dans un second temps lors de l'exploration de ces régresseurs.

Même si la tendance et la saisonnalité représentent une source d'information précieuse, il peut être intéressant de considérer l'influence de variables extérieures à la série temporelle au sein du modèle : c'est ce que nous appelons les régresseurs. Un régresseur pourra ainsi venir expliquer – en partie – tout changement dans le comportement de la série temporelle.

L'implémentation des régresseurs au sein du modèle Prophet est similaire à celui que nous avons pu voir pour les vacances et évènements inhabituelles.

L'équation des régresseurs se définit comme suit :

$$r(t) = \sum_{i=1}^{N} \alpha_i * W_{i,t}$$

A partir d'une matrice de régresseurs fournie par l'utilisateur, il sera nécessaire d'estimer les N paramètres α.

$$W_t = [W_{1,t}, \dots, W_{N,t}]$$

La régression est ensuite définie à l'aide de l'équation suivante : $r(t) = \alpha * W(t)$ où nous considérons – a priori – que α suit une loi normale centrée en 0 et de variance σ^2. Comme pour la saisonnalité, nous avons $\sigma^2 = 10$, permettant d'imposer un lissage préalable des régresseurs sur le modèle.

2. Ajustement du modèle

Une fois l'équation du modèle établie a priori, il est nécessaire de l'ajuster en fonction des données d'entrées. Pour se faire, Prophet fait appel à l'algorithme d'optimisation Broyden-Fletcher-Goldgarb-Shanno dans sa version optimisée (L-BFGS) afin de trouver une estimation a posteriori. L'explication du processus d'optimisation est disponible en annexe 5.

3. Les spécificités de Prophet

Par son approche, Prophet cherche à « vulgariser » les séries temporelles et à faciliter la réalisation de prédiction sur celles-ci. Ainsi, Prophet cherche à s'adresser aux spécialistes de tous domaines, y compris à ceux qui n'ont aucune connaissance sur les séries temporelles. Cette vulgarisation ne dispense pas pour autant l'utilisateur d'effectuer une analyse statistique au préalable. En effet, ce modèle, comme tous les autres, est perfectible. Il y a donc toujours le risque de passer à côté d'informations cruciales.

Prophet permet à l'utilisateur d'introduire des variables externes au modèle, cela permet d'affiner les prédictions. Prophet se base ainsi sur le concept de la modélisation **« analyst-in-the-loop »**, c'est-à-dire une modélisation combinant l'analyse statistique objective traditionnelle avec une analyse subjective tirant sa force de l'expérience acquise par des études passées.

B. Présentation des régresseurs

1. Nos régresseurs primaires

Afin d'affiner nos futures projections sur nos postes de consommation, nous allons mettre en place des régresseurs. En introduisant des régresseurs, notre objectif est d'essayer de prendre en compte l'impact du Covid-19 sur les dépenses en 2020 et de prédire – au mieux – les dépenses attendues sur l'année 2021. Nous faisons donc l'hypothèse que les régresseurs peuvent décrire la dépense observée et observable sur nos postes de consommation. L'application de ces régresseurs sur l'année 2020 dans un premier temps nous permettra de vérifier la pertinence d'une telle hypothèse puisque nous serons amenés à appliquer – dans un second temps – notre modèle Prophet sur l'année 2021 pour les cinq postes de consommation.

Après un travail initial d'analyse et d'identification des séries temporelles pouvant avoir un lien avec nos cinq postes de consommation, nous avons identifié une liste de régresseurs primaires intéressant à exploiter :

- Indice d'inflation sous-jacente - Base 2015 - Glissement annuel - Ensemble des ménages - France métropolitaine - Ensemble[22] ;
- Indice CVS-CJO de la production industrielle (base 100 en 2015) - Industrie manufacturière (NAF rév. 2, niveau A10, poste CZ)[23] ;
- Indicateur du climat des affaires - Tous secteurs - France métropolitaine[24] ;
- Indice de chiffre d'affaires - Ensemble du Commerce[25] ;
- Indice des prix à la consommation - Base 2015 - Ensemble des ménages - France - Ensemble[26]
- La consommation en biens des ménages[27] ;
- Taux d'occupation dans l'hôtellerie - France métropolitaine[28] ;
- Démographie - Nombre de décès - France métropolitaine[29] ;
- Démographie - Nombre de naissances vivantes - France métropolitaine[30] ;
- Données hospitalières relatives à l'épidémie de COVID-19 où nous nous intéresserons principalement aux individus en réanimation[31].

D'une façon intuitive, on peut affirmer que les dix régresseurs retenus ont été impactés directement par le Covid-19. Les changements observés en 2020 s'avèrent être une conséquence de la pandémie et non une cause. Après nos recherches et de façon assez logique, il n'existe pas de séries temporelles étant à l'origine de cette maladie. Nous allons rapidement présenter les régresseurs définis ci-dessus. Tous les indicateurs, à l'exception du nombre de personnes en réanimation, ont été obtenus à partir des données de l'INSEE.

[22] La série est consultable sur le site de l'INSEE en cliquant ici.
[23] La série est consultable sur le site de l'INSEE en cliquant ici.
[24] La série est consultable sur le site de l'INSEE en cliquant ici.
[25] La série est consultable sur le site de l'INSEE en cliquant ici.
[26] La série est consultable sur le site de l'INSEE en cliquant ici.
[27] La série est consultable sur le site de l'INSEE en cliquant ici.
[28] La série est consultable sur le site de l'INSEE en cliquant ici.
[29] La série est consultable sur le site de l'INSEE en cliquant ici.
[30] La série est consultable sur le site de l'INSEE en cliquant ici.
[31] La série est consultable sur le site data.gouv.fr en cliquant ici et en téléchargeant la série s'appelant *donnees-hospitalieres-covid19-AAAA-mm-dd-HHhMM.csv*.

a) Inflation

Le graphe ci-dessous nous permet de visualiser l'évolution du taux de l'inflation en France (selon une base 100 en 2015) de janvier 2015 à décembre 2020.

Figure VI-1 – Évolution mensuelle de l'inflation en France entre le 1er janvier 2015 et le 31 décembre 2020

b) Indice de la production industrielle

Le graphe ci-dessous nous permet de visualiser l'évolution de l'indice de production industrielle en France (selon une base 2015) de janvier 2015 à décembre 2020.

Figure VI-2 – Évolution mensuelle de l'indice de production industrielle en France entre le 1er janvier 2015 et le 31 décembre 2020

c) Indicateur du climat des affaires

Il s'agit d'une « *enquête mensuelle de conjoncture dans l'industrie* [qui] *a pour objet de transcrire l'opinion des industriels sur leur activité récente et sur leurs perspectives d'activité* ». Cet indicateur permet « *de fournir des informations précoces sur l'évolution de l'activité, la demande et les capacités de production dans l'industrie, à des fins de diagnostic conjoncturel et de prévision à court terme de la conjoncture industrielle, tant sur le plan national qu'européen.* »

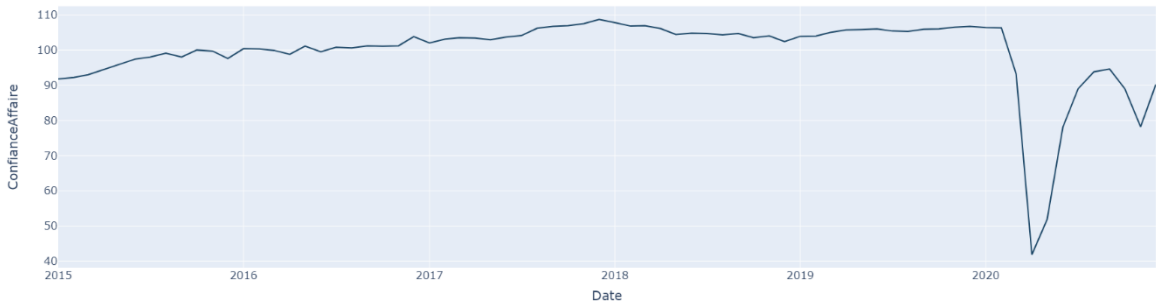


Figure VI-3 – Évolution mensuelle de l'indicateur du climat des affaires entre le 1er janvier 2015 et le 31 décembre 2020

d) Indice de chiffre d'affaires pour l'ensemble du commerce

Le graphe ci-dessous nous permet de visualiser l'évolution d'indice du chiffre d'affaires des commerces de janvier 2015 à décembre 2020.

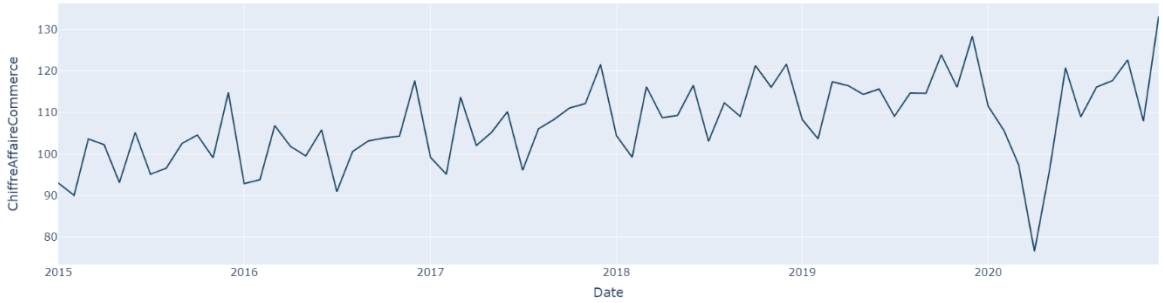


Figure VI-4 – Évolution mensuelle de l'indice du chiffre d'affaires pour l'ensemble du commerce entre le 1er janvier 2015 et le 31 décembre 2020

e) Indice des prix à la consommation

Le graphe ci-dessous nous permet de visualiser l'évolution de l'indice des prix à la consommation (base 100 en 2015) de janvier 2015 à décembre 2020.

Figure VI-5 – Évolution mensuelle de l'indice des prix à la consommation entre le 1er janvier 2015 et le 31 décembre 2020

f) Consommation des ménages en bien

Le graphe ci-dessous nous permet de visualiser l'évolution de la consommation des ménages de janvier 2015 à décembre 2020.

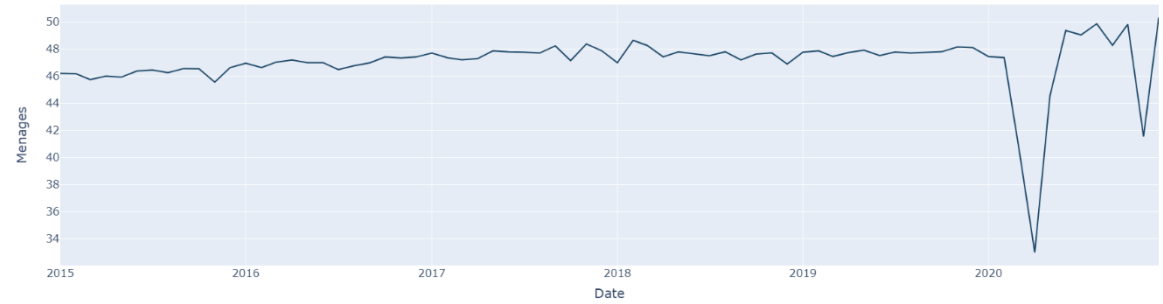


Figure VI-6 – Évolution mensuelle de la consommation des ménages entre le 1er janvier 2015 et le 31 décembre 2020 (en milliards d'euros)

g) Taux d'occupation dans l'hôtellerie

Le graphe ci-dessous nous permet d'apprécier le taux d'occupation dans l'hôtellerie en France métropolitaine. Cette série présente une forte saisonnalité comme nous pouvons le voir dans la figure ci-dessous.

Figure VI-7 – Évolution mensuelle du taux d'occupation dans l'hôtellerie entre le 1er janvier 2015 et le 31 décembre 2020

h) Nombre de décès mensuel

Cette information fait partie des indicateurs standards liés à la démographie et permet de suivre le nombre mensuel de décès au sein de la population française. Dans la figure ci-dessous, nous pouvons ainsi voir les divers pics de mortalité en France sur les cinq dernières années.

Figure VI-8 – Évolution mensuelle du nombre de décès entre le 1er janvier 2015 et le 31 décembre 2020

i) Nombre de naissances vivante mensuelle

Cette information fait partie des indicateurs standards liés à la démographie et permet de suivre le nombre mensuel de naissances. Dans la figure ci-dessous, nous pouvons notamment observer une forte saisonnalité dans les naissances.

Figure VI-9 – Évolution mensuelle du nombre de naissance vivante entre le 1er janvier 2015 et le 31 décembre 2020

j) Nombre de personnes en réanimation pour cause de Covid-19

La série chronologique a été obtenue sur le site du gouvernement au sein des indicateurs liés au Covid-19. Cette série est disponible uniquement depuis mars 2020, début de la pandémie en France. Le nombre de personnes en réanimation pour cause de Covid-19 est donc inexistant avant ce mois-ci. De plus, cet indicateur a l'avantage d'être constant dans le temps. En effet, comme nous avons pu le voir dans la partie précédente, cet indicateur est indépendant du nombre de test et permet donc de refléter au mieux la situation épidémique réelle dans le pays.

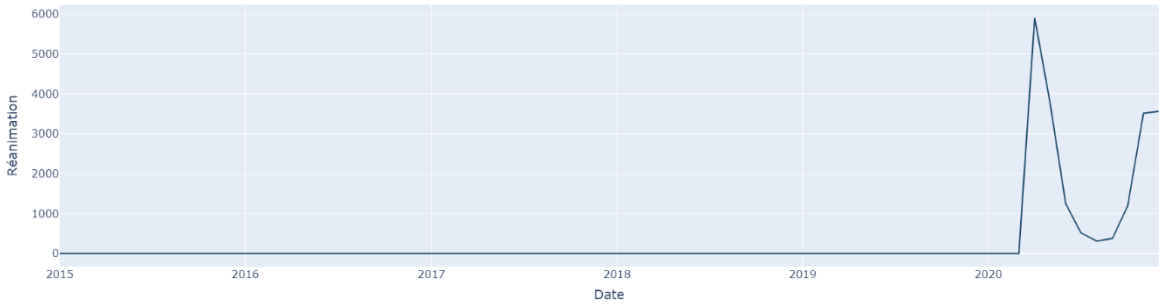


Figure VI-10 – Évolution mensuelle du nombre de personnes en réanimation atteintes de Covid-19 entre le 1er janvier 2015 et le 31 décembre 2020

2. Manipulation et transformation de nos régresseurs

a) Un prétraitement indispensable

Comme cela peut se remarquer lors de la présentation de régresseurs primaires, ceux-ci ne possèdent pas la même échelle : taux, valeurs brutes …. Il est donc nécessaire de normaliser chacun de nos régresseurs avant de les manipuler. Pour se faire, nous allons les normaliser entre 0 et 1 à l'aide de l'équation suivante :

$$r'(t) = \frac{r(t) - min(r)}{max(r) - min(r)}$$

Avec :

- $r'(t)$, représentant le régresseur normalisé en t ;
- $r(t)$, représentant la valeur du régresseur en t ;
- $min(r)$, représentant la valeur minimum de la série chronologique observée ;
- $max(r)$, représentant la valeur maximum de la série chronologique observée.

Une fois ce prétraitement de nos régresseurs effectués, nous obtenons des séries temporelles comparables, sur lesquelles il nous est possible d'y appliquer des traitements similaires.

b) Une approche individuelle

Nous avons commencé par tester ces régresseurs normalisés de façon individuelle sur nos cinq postes de consommation afin d'identifier la pertinence de chacun. Cette première étape nous a permis d'identifier que même si un régresseur est intéressant à utiliser pour le poste *Dentaire* par exemple, il ne le sera pas forcément pour le poste *Optique*. Nous avons donc dû adapter notre approche en conséquence. En effet, notre objectif est de conserver les mêmes régresseurs pour tous les postes de consommation de l'étude en considérant que les régresseurs retenus ont le même impact pour chaque poste. De même, nous essayons d'avoir une bonne estimation possible sur l'année au global, mais nous acceptons de ne pas saisir les spécificités mensuelles. Au regard de cette hypothèse, nous acceptons de ne pas avoir une prédiction correcte sur la deuxième période afin d'éviter le sur apprentissage de nos modèles. Dans l'idéal, il aurait été judicieux d'intégrer des séries spécifiques (coût moyen des appareillages dentaire pour le poste *Dentaire* par exemple) mais le suivi mensuel de ces informations n'est pas disponible.

c) Une approche par classe

L'approche individuelle n'a pas toutefois été satisfaisante : nous avons fait le choix de créer de mixes d'indicateurs primaires en classes regroupant plusieurs séries chronologiques. Ainsi, nous avons défini quatre grandes classes, qui sont une combinaison linéaire simple des régresseurs primaires appartenant :

- Une **classe macroéconomique** regroupant l'inflation, la production industrielle, l'indice de confiance du monde des affaires ainsi que l'indice de chiffre d'affaires pour l'ensemble du commerce ;
- Une **classe microéconomique** regroupant l'indice des prix à la consommation, la consommation en biens des ménages et le taux d'occupation dans l'hôtellerie ;
- Une **classe démographique** regroupant les naissances et les décès dans la population française ;
- Une **classe liée au Covid** regroupant uniquement le nombre de personnes en réanimation.

Après tests, nous avons constaté que cette approche n'était pas plus pertinente que l'approche individuelle. Nous avons donc fait le choix d'effectuer des combinaisons linéaires de régresseurs appartenant à des classes différentes.

d) Création du régresseur central $\mathfrak{R}$

Les divers tests que nous avons pu effectuer nous ont permis d'en éliminer de nombreux régresseurs primaires qui se sont avérés redondants. Comme cela a été précisé lors de la présentation du régresseur lié au Covid-19, nous n'avons pas de données disponibles avant mars 2020. Il est donc impossible de l'utiliser lors de l'entraînement de nos modèles car Prophet ne peut pas prédire un impact futur si celui-ci n'a pas été identifié dans le passé. Ce régresseur doit donc être pris en compte autrement, c'est ce que nous faisons par l'introduction de notre régresseur central $\mathfrak{R}$ que nous définissons comme suit :

$$\mathfrak{R} = \frac{ConfAffaire - Deces + OccupHotel + ChiffreAffCom - ReaCovid}{5}$$

Nous pouvons visualiser l'évolution de ce nouveau régresseur dans la figure ci-dessous.

Figure VI-11 – Évolution mensuelle du régresseur central entre le 1er janvier 2015 et le 31 décembre 2020

C. Implémentation du modèle sous Python

1. Présentation de la fonction Prophet

Avant de pouvoir appliquer la fonction Prophet sous Python, certains ajustements doivent être effectués sur notre jeu de données afin de le normaliser selon les standards de cette librairie. Cette standardisation est relativement simple puisque Prophet a besoin de recevoir – au minimum – un tableau de données contenant deux colonnes :

- La première colonne reflète l'échelle de temps de notre jeu de données. Celle-ci doit nécessairement s'appeler *ds* (pour *datestamp* ou horodatage en français) ;
- La seconde colonne représente la variable que nous étudions et que nous souhaitons projeter sur une période future. La colonne doit nécessairement s'appeler *y* ;
- Des colonnes supplémentaires, appelées régresseurs, peuvent être intégrées au tableau de données mais celles-ci sont facultatives. Nous les verrons plus en détails par la suite.

Une fois notre jeu de données standardisé, nous pouvons y appliquer le modèle Prophet.

Afin de comprendre ce qui se fera par la suite, nous allons commencer par expliciter la fonction ainsi que les paramètres associés. Nous pourrons ainsi plus facilement comprendre ce qui se fera par la suite tout en nous concentrant exclusivement sur l'application des modèles.

Initialement, la fonction se présente comme suit :

$$Prophet(growth = 'linear', changepoints = None, n\_changepoints = 25, changepoint\_range = 0{,}8, yearly\_seasonality = 'auto', weekly\_seasonality = 'auto', daily\_seasonality = 'auto', holidays = None, seasonality\_mode = 'additive', seasonality\_prior\_scale = 10, holidays\_prior\_scale = 10, changepoint\_prior\_scale = 0.05, mcmc\_samples = 0, interval\_width = 0.8, uncertainty\_samples = 1000, stan\_backend = None)$$

Paramètres	Explication	Valeurs associées
growth	Pour indiquer au modèle si la tendance est linéaire ou logistique (s'il s'agit d'un modèle de croissance par saturation ou d'un modèle linéaire par morceaux).	Par défaut, linéaire. Comme vu précédemment, option retenue dans nos modèles.
changepoints	Pour indiquer au modèle s'il doit prendre en compte certaines dates de changement dans la tendance de la série.	Par défaut, aucun. Le modèle décide alors de lui-même. Sinon, il est possible de lui préciser une série de date.
n_changepoint	Pour indiquer au modèle le nombre de changements à inclure.	Par défaut, 25. L'utilisateur reste libre de choisir ce qu'il souhaite.
changepoint_range	Pour indiquer au modèle la proportion de données historiques dans laquelle il peut chercher des modifications de tendance.	Par défaut 80%. L'utilisateur reste libre de choisir ce qu'il souhaite.
yearly_saisonality	Pour indiquer au modèle s'il doit chercher une saisonnalité annuelle.	Par défaut, automatique. L'utilisateur peut également préciser si celle-ci existe ou non.
weekly_saisonality	Pour indiquer au modèle s'il doit chercher une saisonnalité hebdomadaire.	
daily_saisonality	Pour indiquer au modèle s'il doit chercher une saisonnalité journalière.	
holidays	Pour préciser au modèle s'il doit prendre en compte des évènements spécifiques pouvant influer sur la série temporelle.	Par défaut, aucun mais l'utilisateur a la possibilité d'insérer un tableau de données précisant ces évènements spéciaux.
seasonality_mode	Pour préciser au modèle si la saisonnalité est additive ou multiplicative dans le modèle.	Par défaut, additive.
seasonality_prior_scale	Pour préciser au modèle l'intensité de la saisonnalité dans le modèle.	Par défaut, 10. Augmenter cette valeur permet de prendre en considération des fluctuations plus importantes. A l'inverse, baisser cette valeur réduit l'effet de la fluctuation.
holidays_prior_scale	Pour préciser au modèle l'intensité des évènements spéciaux dans le modèle.	

changepoint_prior_scale	Pour préciser au modèle la flexibilité avec laquelle il doit sélectionner automatiquement les points de changement.	Par défaut, 0.05. Accroitre cette valeur pour augmenter le nombre de points et la baisser pour diminuer le nombre de points.
mcmc_samples	Pour indiquer au modèle s'il doit effectuer une inférence bayésienne complète selon la valeur précisée.	Par défaut 0, le modèle utilise alors le maximum a posteriori pour l'estimation. Sinon, une inférence bayésienne est effectuée selon le nombre précisé.
interval_width	Pour indiquer au modèle l'amplitude des intervalles de confiance.	Par défaut, 80%. Nous le fixerons à 95 %.
Uncertainty_samples	Pour indiquer au modèle le nombre de tirages à utiliser afin d'estimer les intervalles de confiance.	Par défaut, 1000. Cette valeur est modifiable à la discrétion de l'utilisateur.
stan_backend	Ce qui permet l'application du processus d'inférence bayésienne.	Par défaut, aucun. Prophet cherchera le modèle le plus optimal. Nous ne serons pas amenés à modifier cette option.

Figure VI-12 – Présentation des paramètres de la fonction Prophet sous Python

2. L'implémentation

Nous allons considérer deux cas dans l'implémentation de Prophet :

- L'application du **modèle sans régresseur** pour le poste de consommation Pharmacie ;
- L'application du **modèle avec le régresseur central $\mathfrak{R}$** pour les quatre autres postes de consommation.

Comme pour la partie IV, nous divisons nos cinq bases de données selon un schéma identique que nous rappelons ici :

- Les dépenses historiques de 2015 à 2019 serviront de données d'entraînement pour notre modèle ;
- Les dépenses développées à l'ultime de 2020 serviront de données de test pour notre modèle.

Les hypothèses sur les résidus du modèle ne sont pas aussi fortes que pour les modèles ARIMA et SARIMA. Ainsi, nous vérifierons uniquement que les résidus sont bien distribués.

Dans une même sous-partie, nous appliquons le modèle Prophet à notre jeu de données et nous projetons la consommation attendue par ce même modèle. Afin de mesurer la pertinence du modèle Prophet nouvellement créé, nous comparerons la projection avec les valeurs attendues sur l'année 2020 ainsi qu'avec les prédictions du modèle SARIMA sur cette même année.

Dans notre projection graphique et dans un but de lisibilité, nous faisons le choix d'ignorer l'intervalle de confiance du modèle SARIMA mais nous le conservons pour le modèle Prophet.

Lors de la comparaison entre le modèle Prophet et le modèle SARIMA, nous n'y mettons pas les tableaux de comparaison de la projection par mois de soin mais uniquement par période. Les tableaux par mois de survenance sont disponibles en annexe 6.

a) Application sur le poste Pharmacie sans régresseur

Comme nous avons pu le voir dans la partie IV, les modèles ARIMA et SARIMA ont été plutôt pertinents pour prédire la consommation attendue en 2020 sur le poste *Pharmacie*. Nous faisons donc le choix de ne pas introduire le régresseur central dans ce modèle et appliquons donc le modèle Prophet dans sa forme la plus simple.

(a) Application du modèle Prophet

Nous générons donc un modèle Prophet dont l'équation est de la forme : $y(t) = g(t) + s(t) + \epsilon_t$, avec *g* la tendance, *s* la saisonnalité et ϵ l'erreur.

Une fois le modèle Prophet appliqué sur les dépenses historiques, nous pouvons visualiser les résultats.

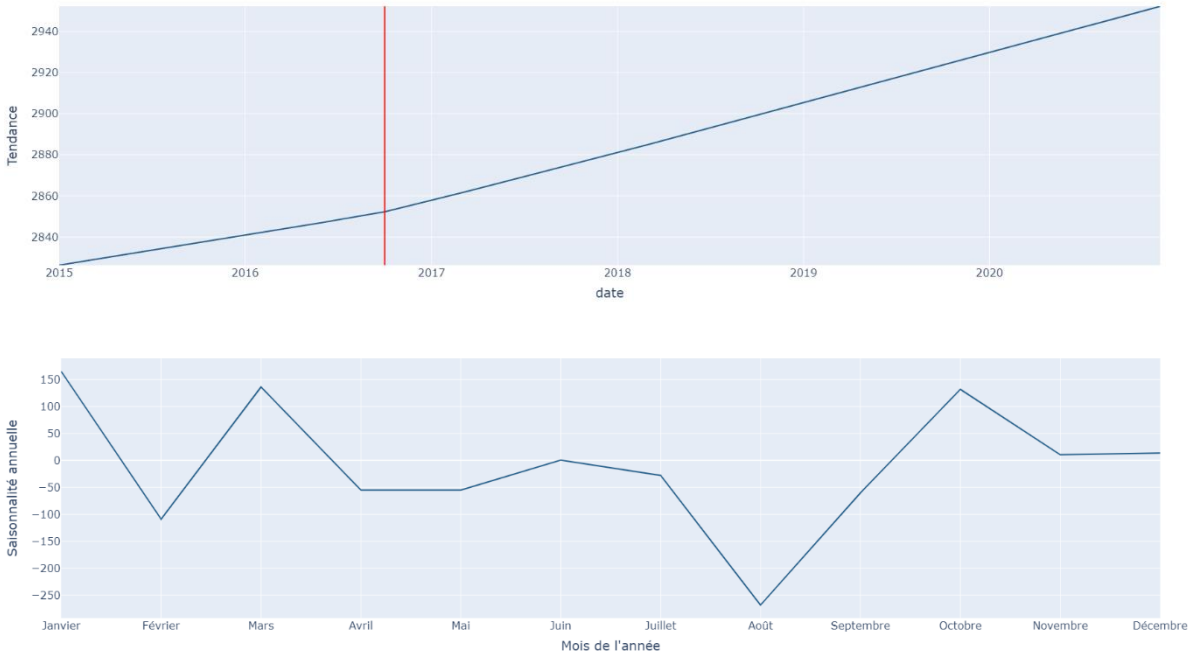


Figure VI-13– Tendance (en haut) et saisonnalité (en bas) du modèle Prophet pour le poste Pharmacie (en millions d'euros)

Sur la figure ci-dessus, le premier graphique représente la tendance identifiée par le modèle Prophet et nous permet de remarquer la présence d'un point d'inflexion en octobre 2016 (mis en évidence par le trait vertical rouge) : la pente de la tendance est plus importante après cette date. Le second graphique nous permet d'observer la saisonnalité identifiée par le modèle Prophet. Nous pouvons notamment remarquer que la forme de la courbe est similaire à ce que nous avons pu obtenir sur le poste *Pharmacie* dans la partie précédente. Ainsi, nous observons trois pics de consommation dans l'année : en janvier, mars et octobre et une baisse marquée de la consommation en août.

Quant aux résidus, nous pouvons les visualiser dans la figure ci-dessous et conclure que nos résidus semblent normalement distribués ($\epsilon_t \sim \mathrm{N}(4{,}55 * 10^2, 5{,}78 * 10^7)$).

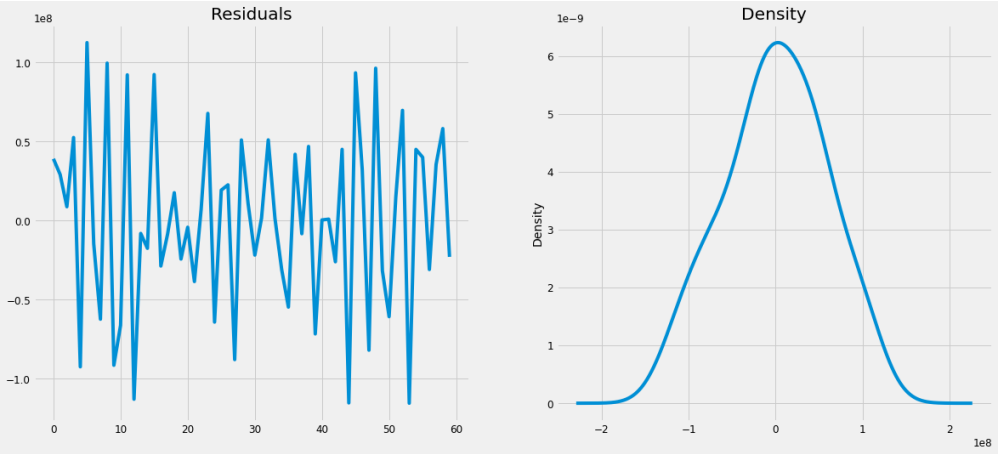


Figure VI-14 – Résidus du modèle Prophet pour le poste Pharmacie

(b) Projection sur l'année de consommation 2020

Dans un second temps, nous projetons sur 2020 la consommation pour le poste *Pharmacie* et nous pouvons visualiser – dans la figure ci-dessous – le résultat produit par le modèle Prophet.

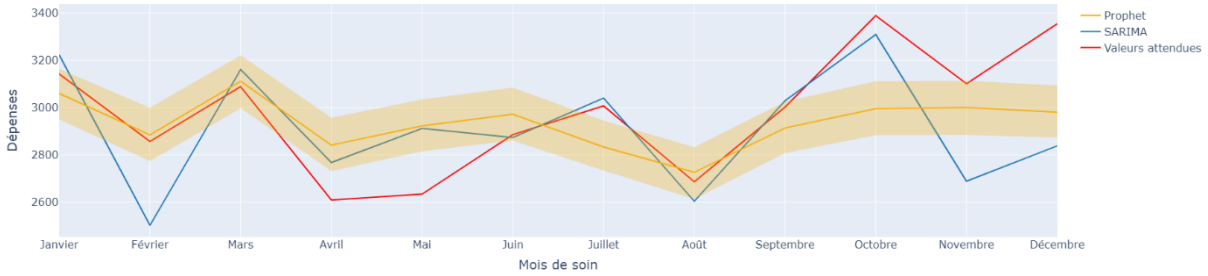


Figure VI-15 – Prédictions sur 2020 du modèle Prophet et SARIMA pour le poste de consommation Pharmacie (en millions d'euros)

Le modèle Prophet est très performant sur la première période où l'erreur est quasiment inexistante. Cependant, le modèle Prophet est moins efficace sur la deuxième et troisième période par rapport au modèle SARIMA. Sur le cumulé de l'année, le modèle Prophet prédit mieux la consommation que le modèle SARIMA.

	Observée	Prédiction SARIMA	Prédiction Prophet	Erreur SARIMA	Erreur Prophet	% erreur SARIMA	% erreur Prophet
Période n°1	5 999	5 726	5 944	**273**	**55**	**4,77 %**	**0,93 %**
Période n°2	8 333	8 841	8 875	**- 508**	**- 542**	**- 5,75 %**	**-6,11 %**
Période n°3	11 582	11 548	11 447	**34**	**135**	**0,30 %**	**1,18 %**
Période n°4	9 843	8 836	8 976	**1 006**	**867**	**11,39 %**	**9,66 %**
Total	**35 756**	**34 950**	**35 242**	**806**	**515**	**2,31 %**	**1,46 %**

Figure VI-16 – Prédictions sur 2020 du modèle Prophet et SARIMA pour le poste de consommation Pharmacie selon les périodes prédéfinies (en millions d'euros)

b) *Application du modèle Prophet à l'aide du régresseur central* $\mathfrak{R}$

Dans un second temps, nous appliquons notre modèle Prophet sur les quatre autres postes de consommation et y ajoutons le régresseur que nous avons créé afin d'affiner nos prédictions.

Nous générons donc quatre modèles Prophet dont l'équation est de la forme : $y(t) = g(t) + s(t) + \boldsymbol{r(t)} + \epsilon_t$, avec *g* la tendance, *s* la saisonnalité, *r* le régresseur et ϵ l'erreur.

(a) Soins de ville courants

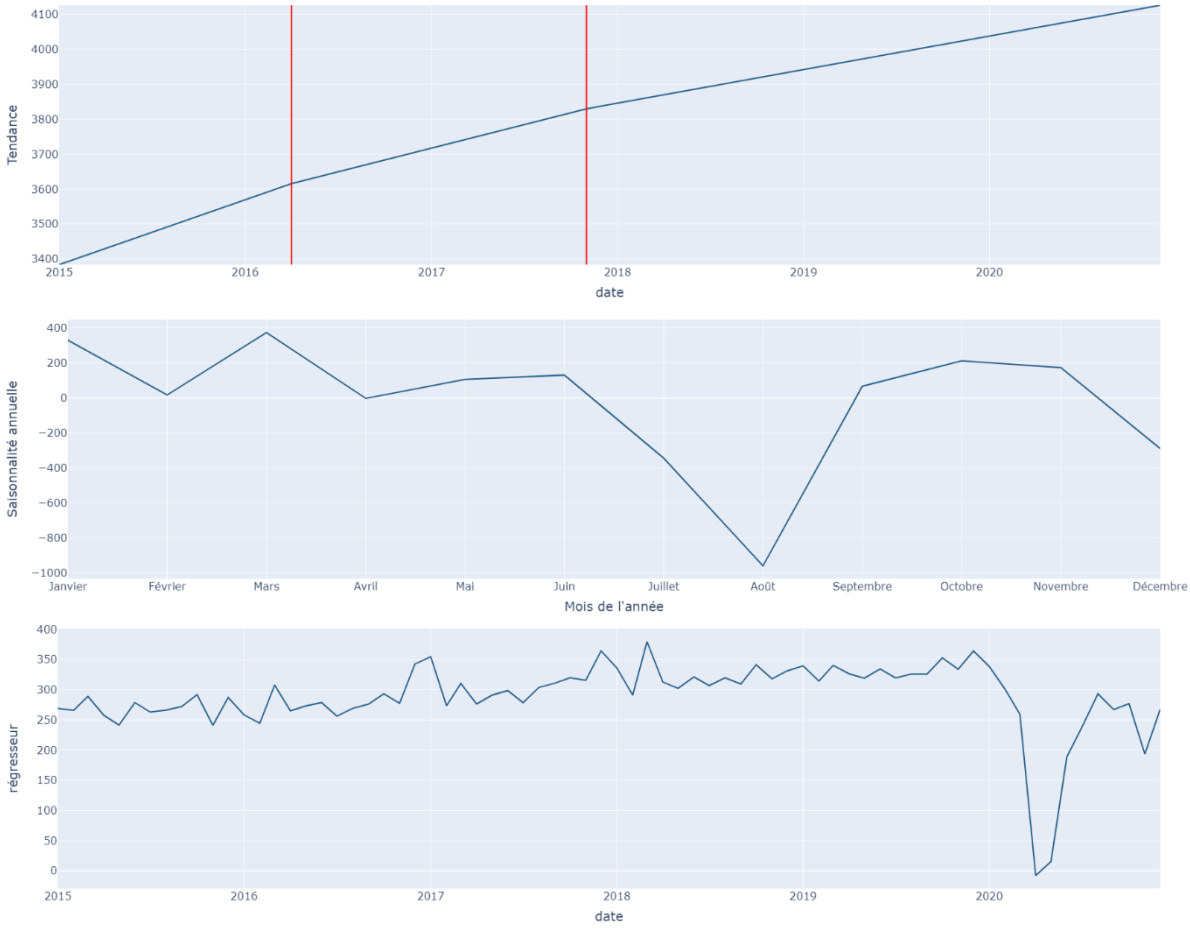


Figure VI-17 –Tendance (en haut), saisonnalité (au milieu) et régresseur (en bas) du modèle Prophet pour le poste Soins de ville courants (en millions d'euros)

A l'aide de la figure ci-dessus, nous pouvons visualiser les trois composantes du modèle Prophet sur le poste de consommation *soins de ville courants*. Dans un premier temps, nous pouvons identifier deux changements dans la tendance : en avril 2016 et novembre 2017 (les deux traits verticaux en rouge dans le graphique) : la croissance dans la consommation s'est réduite lors de ces deux changements de tendance.

Le dernier graphique représente l'impact du régresseur et est une copie conforme du graphique B.6 modulo un facteur multiplicatif. Ainsi, si nous regardons l'impact du régresseur en décembre 2019, nous pouvons voir qu'il est venu augmenter la dépense de 364,15 millions d'euros. La valeur brute de ce régresseur à cette même date est de 0,595. Cela signifie donc que le régresseur supplémentaire est de la forme $\boldsymbol{r(t)} = \boldsymbol{612,01} * r_t$ (avec r_t correspondant à la valeur brute du régresseur en date t)

Quant aux résidus, nous pouvons les visualiser dans la figure ci-dessous et conclure que nos résidus semblent normalement distribués ($\epsilon_t \sim \mathrm{N}(7{,}49 * 10^2, 8{,}6 * 10^7)$).

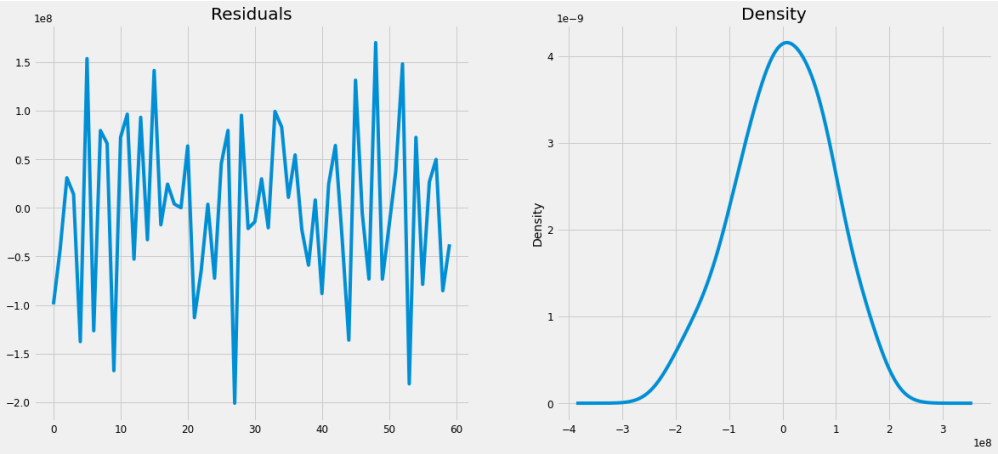


Figure VI-18 – Résidus du modèle Prophet pour le poste Soins de ville courants

Dans un second temps, nous projetons sur 2020 la consommation pour le poste *Soins de ville courants*.

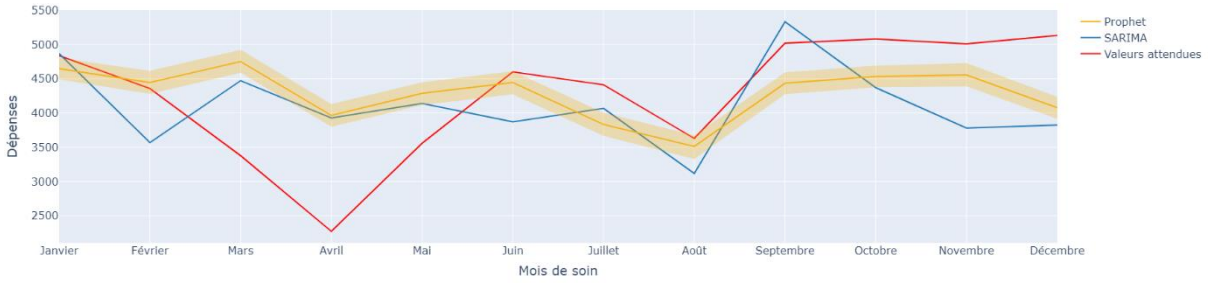


Figure VI-19 – Prédictions sur 2020 du modèle Prophet et SARIMA pour le poste de consommation Soins de ville courants (en millions d'euros)

On relève les mêmes points que dans le poste Pharmacie, modélisé sans régresseurs : sur la première période, le modèle Prophet est meilleur que le modèle SARIMA. Cependant, SARIMA est meilleur sur la deuxième et troisième période. Malgré tout, l'application du modèle Prophet avec régresseur nous permet d'avoir des meilleures estimations sur le cumulé de l'année.

	Observée	Prédiction SARIMA	Prédiction Prophet	Erreur SARIMA	Erreur Prophet	% erreur SARIMA	% erreur Prophet
Période n°1	9 195	8 436	9 095	**759**	**100**	**9,00 %**	**1,1 %**
Période n°2	9 210	12 541	13 003	**- 3 331**	**- 3 792**	**- 26,56 %**	**- 29,16 %**
Période n°3	17 661	16 391	16 228	**1 270**	**1 433**	**7,75 %**	**8,83 %**
Période n°4	15 222	11 977	13 168	**3 245**	**2 054**	**27,09 %**	**15,6 %**
Total	**51 289**	**49 345**	**51 494**	**1 944**	**- 205**	**3,94 %**	**- 0,4 %**

Figure VI-20 – Prédictions sur 2020 du modèle Prophet et SARIMA pour le poste de consommation Soins de ville courants selon les périodes prédéfinies (en millions d'euros)

(b) Dentaire

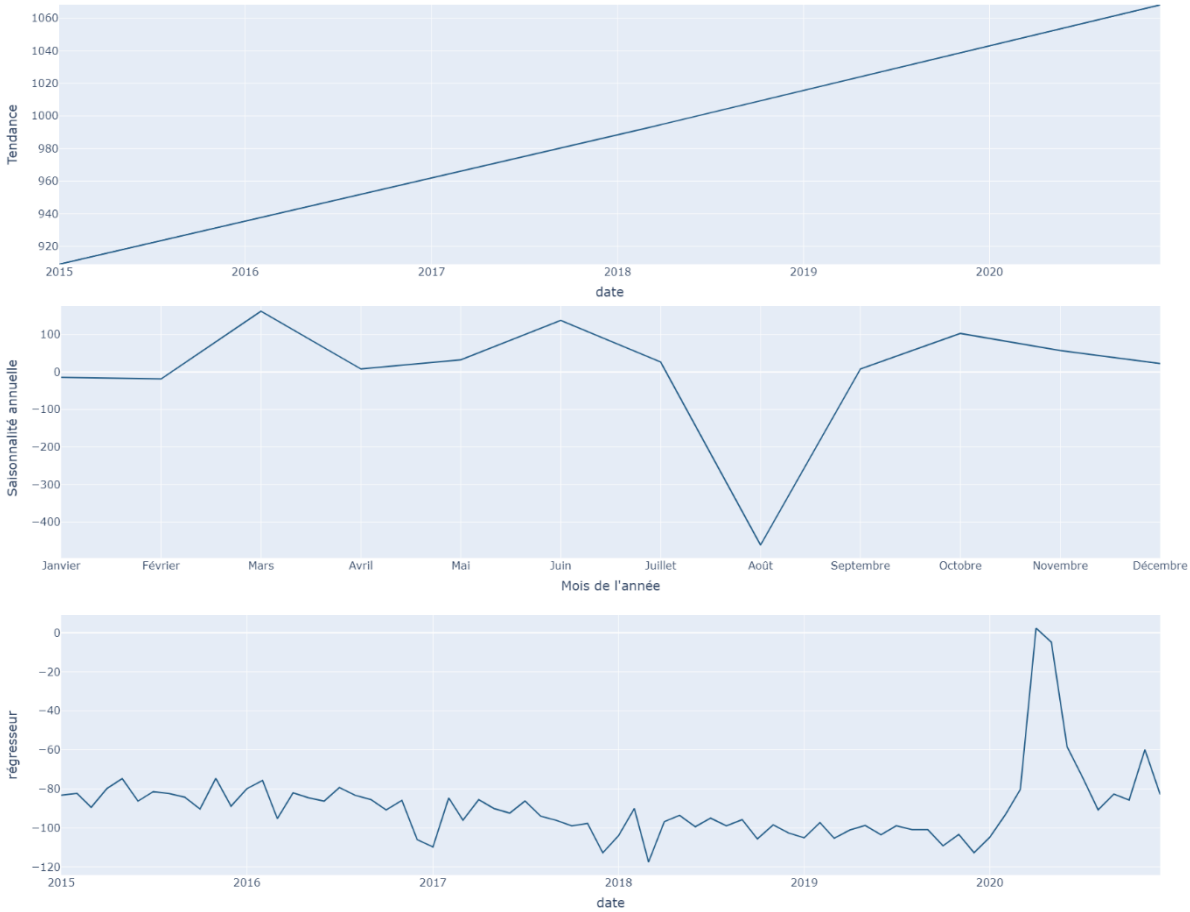


Figure VI-21 – Tendance (en haut), saisonnalité (au milieu) et régresseur (en bas) du modèle Prophet pour le poste Dentaire (en millions d'euros)

A l'aide de la figure ci-dessus, nous pouvons visualiser les trois composantes du modèle Prophet sur le poste de consommation *dentaire*. Nous pouvons notamment voir que la tendance est restée la même entre janvier 2015 et décembre 2019.

Le dernier graphique représente l'impact du régresseur. Ainsi, si nous regardons l'impact du régresseur en décembre 2019, nous pouvons voir qu'il est venu diminuer la dépense de 112,72 millions d'euros. La valeur brute de ce régresseur à cette même date est de 0,595. Cela signifie donc que le régresseur supplémentaire est de la forme $\boldsymbol{r(t) = -189,45 * r_t}$ (avec r_t correspondant à la valeur brute du régresseur en date t)

Quant aux résidus, nous pouvons les visualiser dans la figure ci-dessous et conclure que nos résidus semblent normalement distribués ($\epsilon_t \sim \mathrm{N}(6{,}15 * 10^3, 3 * 10^7)$).

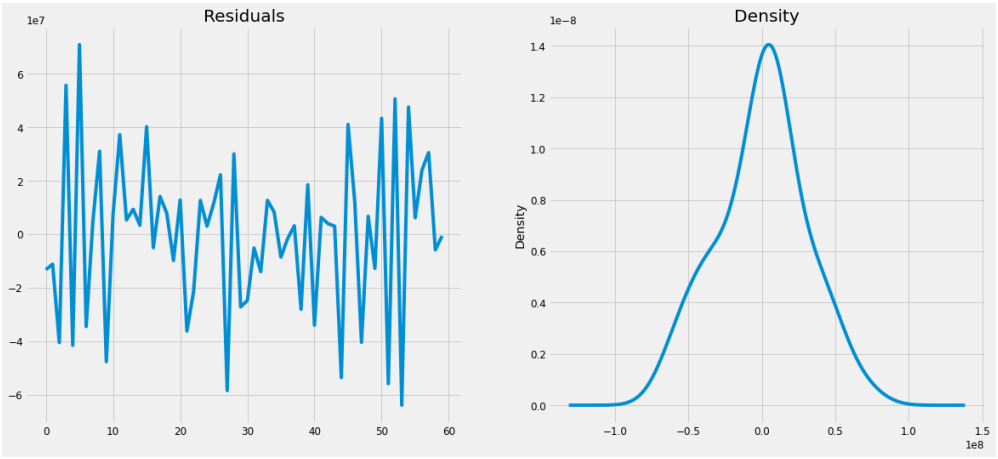


Figure VI-22 – Résidus du modèle Prophet pour le poste Dentaire

Dans un second temps, nous projetons sur 2020 la consommation pour le poste *Dentaire*.

Figure VI-23 – Prédictions sur 2020 du modèle Prophet et SARIMA pour le poste de consommation Dentaire (en millions d'euros)

L'application du modèle Prophet sur le poste *Dentaire* s'est avérée être décevant. Nous obtenons des résultats bien moins pertinents que le modèle SARIMA au global et sur trois des quatre périodes que nous étudions. Finalement, l'application de notre régresseur nous a uniquement permis de gagner en pertinence sur la quatrième période, qui est, elle-même, une approximation de la consommation réelle. Le modèle Prophet pour le poste Dentaire est ainsi très loin d'être pertinent.

	Observée	Prédiction SARIMA	Prédiction Prophet	Erreur SARIMA	Erreur Prophet	% erreur SARIMA	% erreur Prophet
Période n°1	1 873	1 878	1 848	**- 6**	**25**	**- 0,33 %**	**1,35 %**
Période n°2	1 374	2 913	3 286	**- 1 539**	**- 1 912**	**- 52,84 %**	**- 58,19 %**
Période n°3	3 967	3 785	3 631	**183**	**337**	**4,83 %**	**9,28 %**
Période n°4	3 527	2 845	3 138	**681**	**388**	**23,94 %**	**12,36 %**
Total	**10 741**	**11 423**	**11 903**	**- 682**	**- 1 162**	**- 5,97 %**	**- 9,76 %**

Figure VI-24 – Prédictions 2020 pour le poste de consommation Dentaire selon les périodes prédéfinies (en millions d'euros)

(c) Optique

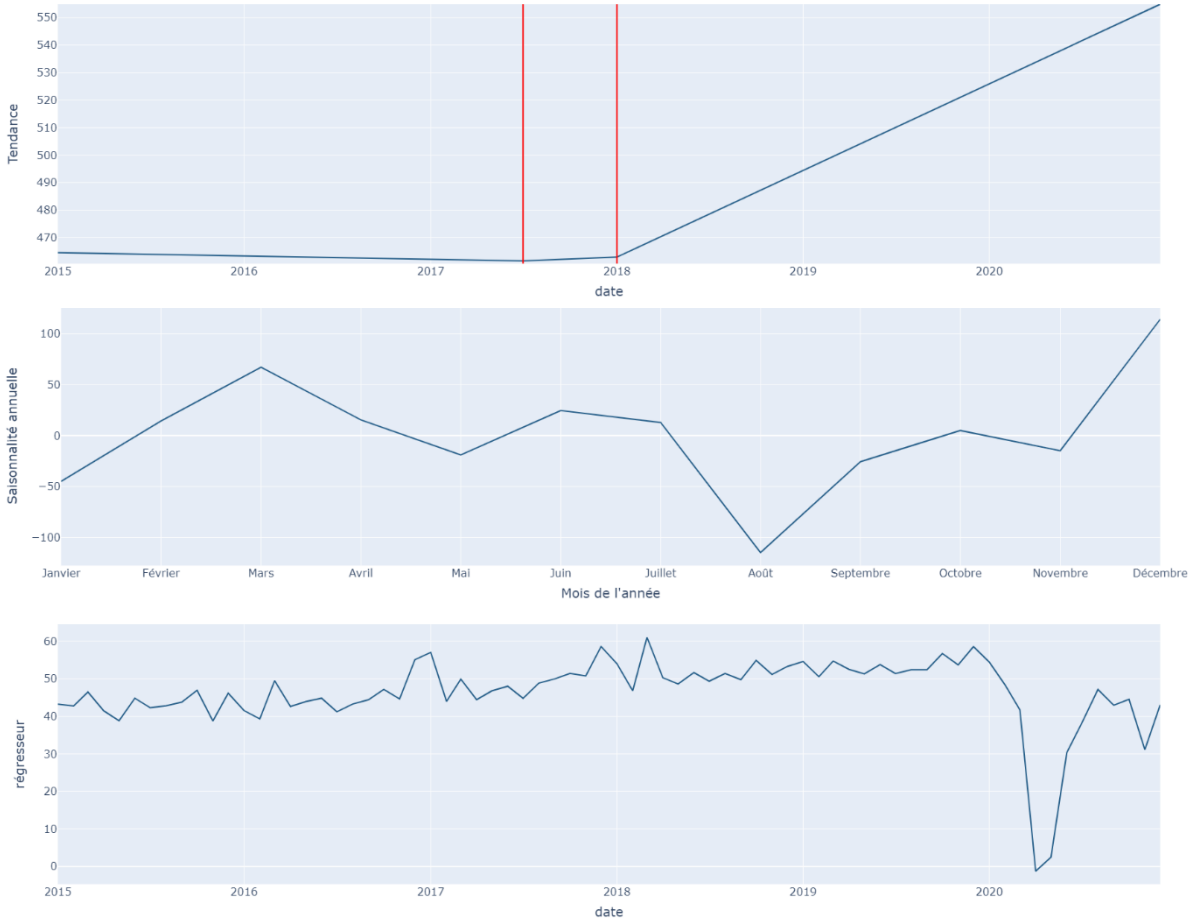


Figure VI-25 – Tendance (en haut), saisonnalité (au milieu) et régresseur (en bas) du modèle Prophet pour le poste Optique (en millions d'euros)

A l'aide de la figure ci-dessus, nous pouvons visualiser les trois composantes du modèle Prophet sur le poste de consommation *Optique*. Dans un premier temps, nous pouvons identifier deux changements dans la tendance d'intensité différente : en juillet 2017 et en janvier 2018 (les deux traits verticaux en rouge) : la croissance de la consommation a légèrement augmenté lors du premier changement et a connu une hausse importante sur le second.

Le dernier graphique représente l'impact du régresseur. Ainsi, si nous regardons l'impact du régresseur en décembre 2019, nous pouvons voir qu'il est venu augmenter la dépense de 58,56 millions d'euros. La valeur brute de ce régresseur à cette même date est de 0,595. Cela signifie donc que le régresseur supplémentaire est de la forme $\boldsymbol{r(t) = 98,42 * r_t}$ (avec r_t correspondant à la valeur brute du régresseur en date t)

Quant aux résidus, nous pouvons les visualiser dans la figure ci-dessous et conclure que nos résidus semblent normalement distribués ($\epsilon_t \sim \mathrm{N}(-1{,}19 * 10^2, 1{,}72 * 10^7)$).

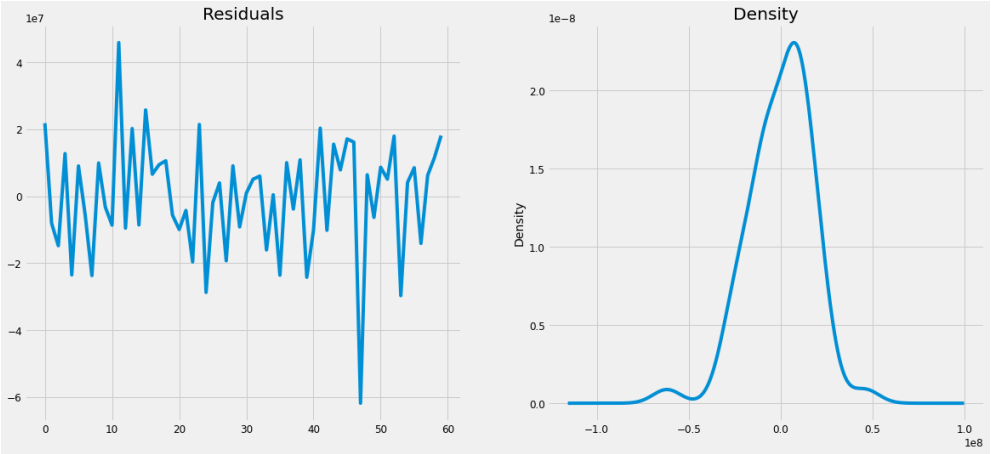


Figure VI-26 – Résidus du modèle Prophet pour le poste Optique

Dans un second temps, nous projetons sur 2020 la consommation pour le poste *Optique*.

Figure VI-27 – Prédictions sur 2020 du modèle Prophet et SARIMA pour le poste de consommation Optique (en millions d'euros)

L'application du modèle Prophet sur le poste *Optique* s'est ainsi révélé bien plus intéressant que d'appliquer le modèle SARIMA. Nous obtenons ainsi des résultats plus pertinents au global et sur les quatre périodes que nous avons définies.

	Observée	Prédiction SARIMA	Prédiction Prophet	Erreur SARIMA	Erreur Prophet	% erreur SARIMA	% erreur Prophet
Période n°1	948	1 323	1 134	**- 374**	**- 186**	**- 28,29 %**	**- 16,4 %**
Période n°2	692	2 130	1 713	**- 1 438**	**- 1 021**	**- 67,53 %**	**- 59,6 %**
Période n°3	2 459	2 827	2 230	**- 368**	**229**	**- 13,01 %**	**10,27 %**
Période n°4	2 063	2 327	1 880	**- 264**	**184**	**- 11,36 %**	**9,79 %**
Total	**6 162**	**8 606**	**6 957**	**- 2 444**	**- 794**	**- 28,40 %**	**- 11,41 %**

Figure VI-28 – Prédictions sur 2020 du modèle Prophet et SARIMA pour le poste de consommation Optique selon les périodes prédéfinies (en millions d'euros)

(d) Prothèses auditives

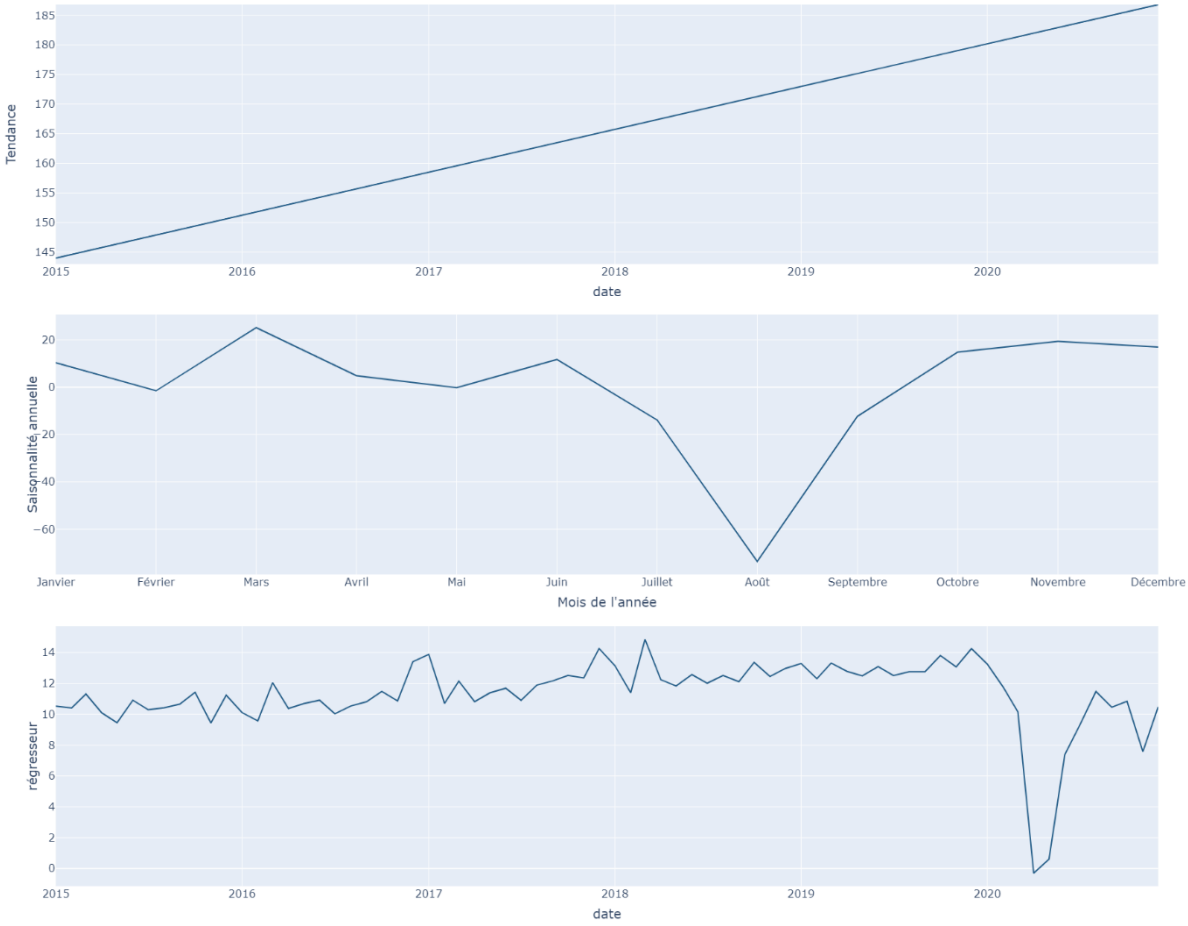


Figure VI-29 – Tendance (en haut), saisonnalité (au milieu) et régresseur (en bas) du modèle Prophet pour le poste Prothèses auditives (en millions d'euros)

A l'aide de la figure ci-dessus, nous pouvons visualiser les trois composantes du modèle Prophet sur le poste de consommation *Prothèses auditives*. Nous pouvons notamment voir que la tendance est croissante de façon continue entre janvier 2015 et décembre 2019.

Le dernier graphique représente l'impact du régresseur. Ainsi, si nous regardons l'impact du régresseur en décembre 2019, nous pouvons voir qu'il est venu diminuer la dépense de 14,25 millions d'euros. La valeur brute de ce régresseur à cette même date est de 0,595. Cela signifie donc que le régresseur supplémentaire est de la forme $\boldsymbol{r(t) = 23,95 * r_t}$ (avec r_t correspondant à la valeur brute du régresseur en date t)

Quant aux résidus, nous pouvons les visualiser dans la figure ci-dessous et conclure que nos résidus semblent normalement distribués ($\epsilon_t \sim \mathrm{N}(-3{,}77 * 10^1, 5{,}4 * 10^6)$).

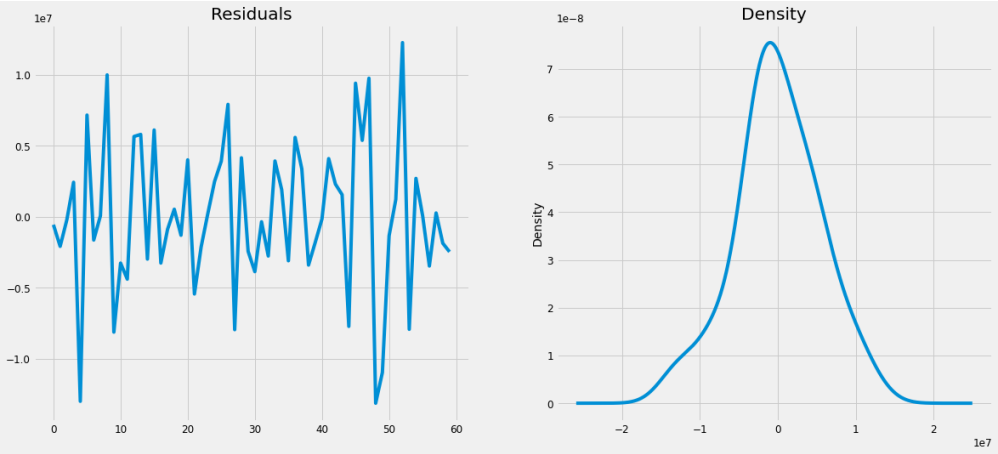


Figure VI-30 – Résidus du modèle Prophet pour le poste Prothèses auditives

Dans un second temps, nous projetons sur 2020 la consommation pour le poste *Prothèses auditives*.

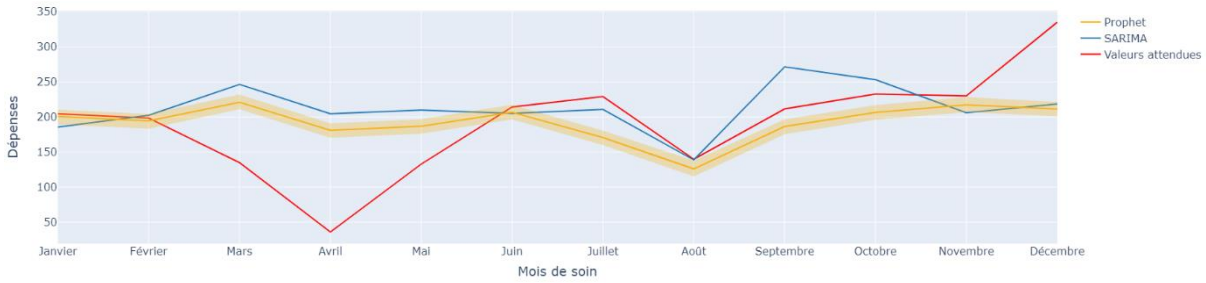


Figure VI-31 – Prédictions sur 2020 du modèle Prophet et SARIMA pour le poste de consommation Prothèses auditives (en millions d'euros)

Au global, le modèle Prophet s'en est mieux sorti pour prédire la consommation que le modèle SARIMA ainsi que sur les deux premières périodes. Le modèle SARIMA est cependant plus pertinent sur les deux dernières périodes. Finalement, les erreurs de prédictions sur l'année du modèle Prophet s'équilibrent puisque nous retrouvons une erreur quasi nulle pour ce modèle.

	Observée	Prédiction SARIMA	Prédiction Prophet	Erreur SARIMA	Erreur Prophet	% erreur SARIMA	% erreur Prophet
Période n°1	403	388	395	**15**	**8**	**3,87 %**	**2,03 %**
Période n°2	303	661	589	**- 357**	**- 285**	**- 54,09 %**	**- 48,39 %**
Période n°3	794	826	689	**- 32**	**105**	**- 3,82 %**	**15,24 %**
Période n°4	798	678	634	**120**	**164**	**17,73 %**	**25,87 %**
Total	**2 298**	**2 552**	**2 307**	**- 254**	**- 8**	**- 9,94 %**	**- 0,35 %**

Figure VI-32 – Prédictions sur 2020 du modèle Prophet et SARIMA pour le poste de consommation Prothèses auditives selon les périodes prédéfinies (en millions d'euros)

c) *Conclusion des prédictions sur 2020 à l'aide du modèle Prophet*

Pour donner suite à l'application du modèle Prophet sur nos cinq postes de consommation et à la comparaison avec le modèle SARIMA, nous pouvons en tirer nos premières conclusions :

- $Prophet$ a été bien plus pertinent pour prédire la consommation que le modèle SARIMA – à l'exception du poste *Dentaire* ;
- L'application de notre régresseur a permis de légèrement gagner en précision sur la période n°2 mais les prédictions ne captent pas entièrement l'effet Covid ;
- La prédiction du modèle $Prophet$ appréciée à l'année s'est révélée être très proche de ce qui était attendu (à l'exception du poste *Dentaire*). Ci-dessous un tableau synthétisant les prédictions au global.

	Dépense au 31/12/2020	Prédiction SARIMA	Prédiction Prophet	Erreur SARIMA	Erreur Prophet	% erreur SARIMA	% erreur Prophet
Soins de ville courants	**51 289**	49 345	51 494	**1 944**	**- 205**	**3,9 %**	**- 0,4 %**
Pharmacie	**35 756**	34 950	35 242	**806**	**515**	**2,3 %**	**1,5 %**
Dentaire	**10 741**	11 423	11 903	**- 682**	**- 1 162**	**- 6 %**	**- 9,8 %**
Optique	**6 162**	8 606	6 957	**- 2 444**	**- 794**	**- 28,4 %**	**- 11,4 %**
Prothèses auditives	**2 298**	2 552	2 307	**- 254**	**- 8**	**- 9,9 %**	**- 0,3 %**

Figure VI-33 – Résumé des résultats obtenus à l'aide du modèle Prophet et SARIMA sur les cinq postes de consommation (en millions d'euros)

D. Projections des dépenses sur 2021

1. Présentation

Afin de pouvoir prédire sur 2021, nous devons commencer par estimer les valeurs de nos régresseurs jusqu'en décembre 2021. Les séries n'étant disponibles que jusqu'au mois de juin, deux choix s'offrent à nous : l'avis d'expert ou l'utilisation d'une méthode de prédiction de séries temporelles.

Une étude approfondie sur chacun des régresseurs serait nécessaire mais, dans un objectif de présenter des résultats concis, nous faisons le choix d'utiliser $Prophet$ afin de projeter nos régresseurs sur la partie inconnue de l'année 2021. Une fois nos cinq régresseurs primaires estimés, nous pouvons calculer notre régresseur central selon la même formule précédemment définie. Une visualisation est disponible dans le graphique ci-dessous.

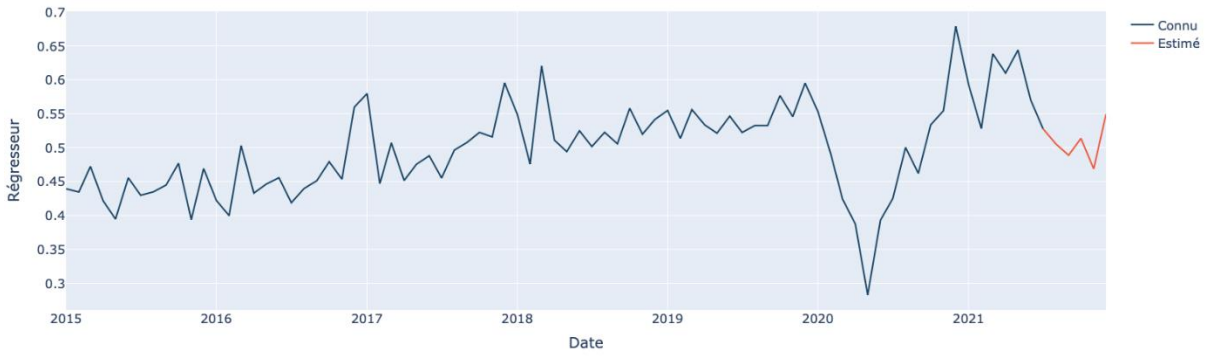


Figure VI-34 – Régresseur central de 2015 à 2021 (en bleu le régresseur central connu, en rouge son estimation)

Une fois notre régresseur central calculé sur 2021, nous projetons la dépense sur chacun des postes de consommation sur cette même année. Nous repartons bien du modèle Prophet retenu pour chacun de nos postes. Nous résumons, dans le tableau ci-dessous, les résultats obtenus pour l'année 2021 sur nos cinq postes de consommation.

	Soins de ville courants	Pharmacie (sans régresseur)	Dentaire	Optique	Prothèses auditives	Total
Janvier	4 921	3 175	989	564	217	**9 866**
Février	4 353	2 751	925	622	194	**8 846**
Mars	4 935	3 108	1 110	693	230	**10 077**
Avril	4 500	2 892	968	642	207	**9 209**
Mai	4 698	2 921	1 005	612	206	**9 443**
Juin	4 668	2 975	1 116	654	217	**9 630**
Juillet	4 111	2 912	993	632	187	**8 836**
Août	3 568	2 718	536	510	131	**7 463**
Septembre	4 563	2 919	1 008	605	193	**9 288**
Octobre	4 711	3 076	1 086	633	218	**9 725**
Novembre	4 698	2 999	1 068	620	225	**9 611**
Décembre	4 260	2 995	1 014	753	223	**9 246**
Total	**53 986**	**35 442**	**11 819**	**7 542**	**2 449**	**111 240**

Figure VI-35 – Projections du modèle Prophet sur l'année 2021 par poste de consommation (en millions d'euros)

Il est impératif de noter que ces résultats sont critiquables, notamment en pharmacie. Nous ne prenons pas en compte le surplus des dépenses attendu en 2021 sur ce poste. En effet, les tests de dépistage du Covid-19 et les vaccins n'étant pas pris en considération dans les données avant 2021.

2. Visualisation de la projection

a) Soins de ville courants

Figure VI-36 – Projection du modèle Prophet en 2021 sur le poste de consommation Soins de ville courants (en millions d'euros)

b) Pharmacie

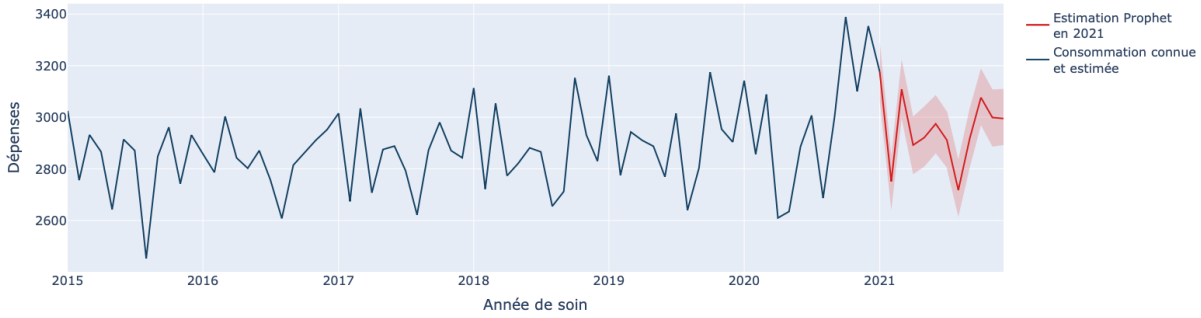


Figure VI-37 – Projection du modèle Prophet en 2021 sur le poste de consommation Soins de ville courants (en millions d'euros)

c) Dentaire

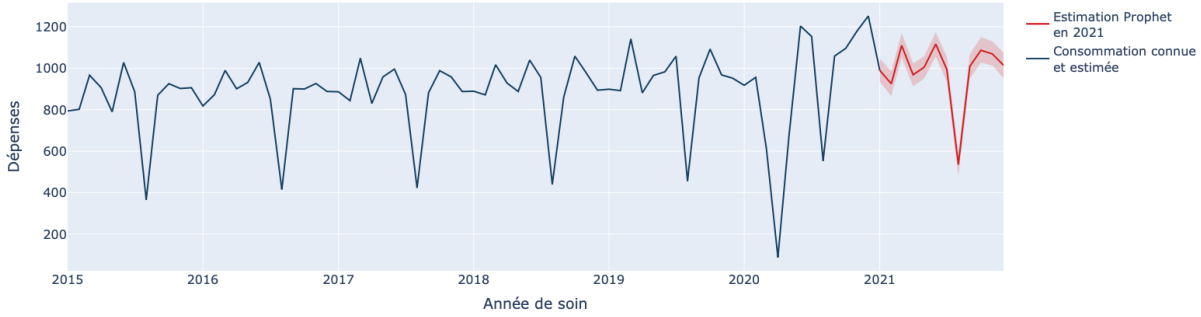


Figure VI-38 – Projection du modèle Prophet en 2021 sur le poste de consommation Soins de ville courants (en millions d'euros)

d) Optique

Figure VI-39 – Projection du modèle Prophet en 2021 sur le poste de consommation Soins de ville courants (en millions d'euros)

e) Prothèses auditives

Figure VI-40 – Projection du modèle Prophet en 2021 sur le poste de consommation Soins de ville courants (en millions d'euros)

VII. Conclusion

Afin d'obtenir une vision globale de la dépense en santé, nous avons fait appel et avons manipulé la base Open DAMIR de 2015 à 2020. Nos modèles ont été entraînés sur les données mensuelles de 2015 à 2019 et les données de l'année 2020 ont servi de données de validation.

Cette vision intégrale s'est avérée être riche en enseignement :

- Tous les postes de consommation ont été fortement impactés par le premier confinement (à l'exception du poste *pharmacie* qui a subi une baisse toute relative) ;
- Sur la période suivant le premier confinement (de juin à septembre), aucun rattrapage clair de la consommation ne peut être observé. Ce rattrapage semble plutôt s'effectuer sur le dernier trimestre de l'année ;
- À la suite de l'application de nos modèles, il semblerait que nous pouvons commencer à observer l'impact de la réforme 100 % Santé en Optique sur l'année 2020, ceci sur la première période de l'année pré-Covid.

Nous avons également abordé l'analyse et la projection de séries temporelles de deux façons :

- La première approche consiste à faire appel à des modèles canoniques dans l'actuariat, avec l'application des modèles $ARIMA$ et $SARIMA$. Il s'agit d'une première approche que nous pouvons considérer comme étant une approche de marché assez bien explorée ;
- La seconde approche est plus adaptée au monde de l'entreprise et permet d'intégrer des informations extérieures afin d'affiner l'étude des séries temporelles. A travers l'application du modèle $Prophet$ développé par les ingénieurs de Facebook, nous avons pu mettre en application une approche plus simple à comprendre par une majorité : les composantes du modèle sont facilement compréhensibles et manipulables. Comme nous avons pu le voir, ce modèle a été pensé pour être utilisé par les spécialistes des domaines souhaitant obtenir des projections sur leur secteur d'activité.

Comme nous pouvons nous y attendre, la première approche s'est avérée incapable de prendre en compte la baisse de la consommation entraînée par le premier confinement et la hausse observée sur le quatrième trimestre. Afin de tenter d'améliorer nos projections, l'application du modèle $Prophet$ couplée à l'introduction de régresseurs externes s'est avérée être légèrement plus pertinente que l'application des modèles ARIMA et SARIMA – à l'exception du poste de consommation *dentaire*.

Ainsi, certaines pistes d'amélioration pourraient être envisagées :

- La base Open DAMIR nous permet d'obtenir une vision plus fine. Nous pourrions ainsi envisager d'effectuer notre analyse selon la région, la tranche d'âge et le sexe de l'assuré ;
- Une étude plus fine par poste de consommation avec notamment l'introduction de régresseur spécifique peut également se montrer être une voie d'exploration ;
- Il serait également plus pertinent d'effectuer l'analyse à partir de données réelles et non de données estimées. Comme la base Open DAMIR n'est mise à jour qu'une seule fois par an, une vision à l'ultime de la dépense par mois de consommation sur l'année 2020 ne sera disponible qu'en juin 2022.

L'impact des mesures prises contre le Covid-19 est indéniable en 2020 et les conséquences de ces mesures continueront à se faire sentir dans les années à venir : les retards de traitements et de détection des maladies pourraient par la suite entraîner une « nouvelle crise de santé ». Cependant, cette crise serait bien plus étalée dans le temps et ses effets seront captés lors des exercices à venir.

VIII. Références

1. Bibliographie

Algorithms for Optimization [Livre] / aut. Kochenderfer Mykel J. et Wheeler Tim A.. - [s.l.] : Massachusetts Institute of Technology, 2019.

Analyse de données SANTE pour la FRANCE à partir de la base Open DAMIR [Rapport] / aut. Cabinet Moeglin. - 2017.

Analyse sur le risque croissant en assurance santé [Rapport] / aut. Courbon Jérémie. - 2017.

Apport des Open Data pour évaluer les impacts de la réforme 100% Santé [Rapport] / aut. Garnier Alban. - 2020.

Approche de calcul individuel des provisions de la RC - Automobile et application à la réassurance [Rapport] / aut. Orozco Penaloza Sergio. - 2016.

Base de données du SNIIRAM : l'ouverture de la boîte de Pandore [Rapport] / aut. Fautrel Bruno. - 2018.

Comment les nouvelles bases de données santé en Open Data peuvent-elles être source d'inspiration pour l'assurance santé de demain. [Rapport] / aut. Quennelle Pascale et Raymond Marc. - 2015.

Dossier de presse : 100 % santé - Des soins pour tous, 100 % pris en charge [Rapport] / aut. Ministère des solidarités et de la santé.

Etude d'impact de la Covid-19 sur les frais de Santé et de Prévoyance [Rapport] / aut. Mercer. - 2020.

Etude sur le système de financement de la santé [Rapport] / aut. Commission Santé de l'Institut des Actuaires. - 2015.

Forecasting and assessing Risk of Individual Electricity Peaks [Livre] / aut. Jacob Maria, Neves Claudia et Vukadinovic Greetham Danica. - [s.l.] : Springer Open, 2020.

Forecasting at Scale [Rapport] / aut. Taylor Sean J. et Letham Benjamin. - 2017.

La complémentaire santé : acteurs, bénéficiaires, garanties [Rapport] / aut. Direction de la recherche, des études, de l'évaluation et des statistiques. - 2019.

Le renoncement aux soins pour raisons financières : une approché économétrique [Rapport] / aut. Després Caroline [et al.]. - 2011.

Le système national d'information interrégimes de l'Assurance Maladie - SNIIRAM [Rapport] / aut. Direction de la Stratégie, des Etudes et des statistiques. - 2015.

L'épidémie de Covid-19 a eu un impact relativement faible sur la mortalité en France [Rapport] / aut. Toubiana Laurent [et al.]. - 2020.

Les dépenses de santé des français en 2020 [Rapport] / aut. Vespieren. - 2021.

Les dépenses de santé en 2019 : résultats des comptes de la santé [Rapport] / aut. Direction de la recherche, des études, de l'évaluation et des statistiques. - 2020.

Les données personnelles de santé gérées par l'Assurance Maladie : une utilisation à développer, une sécurité à renforcer [Rapport] / aut. Cour des comptes. - 2016.

Les Français et le renoncement aux soins [Rapport] / aut. Fondation April. - 2018.

Les résultats de la sécurité sociale en 2019 : l'interruption d'une longue séquence de retour à l'équilibre [Rapport] / aut. Cour des comptes. - 2020.

Méthodologies de test de la racine unitaire [Rapport] / aut. Cem Ertur. - [s.l.] : Laboratoire d'analyse et de techniques économiques (LATEC), 1998.

Réforme 100 % Santé : lorsque le reste-à-charge augmente ! [Rapport] / aut. Carte Blanche Partenaires. - 2021.

Renoncement aux soins pour raisons financières [Rapport] / aut. Direction de la recherche, des études, de l'évaluation et des statistiques. - 2015.

Small sample power of tests of normality when the alternative is an alpha-stable distribution [Rapport] / aut. Frain John C.. - 2006.

Time Series Analysis [Rapport] / aut. Wintenberger Olivier. - 2019.

2. Sitographie

Argus de l'assurance :

- https://www.argusdelassurance.com/assurance-de-personnes/sante/covid-19-une-baisse-des-depenses-de-sante-plus-limitee-qu-attendu-en-2020.178034
- https://www.argusdelassurance.com/assurance-de-personnes/sante/covid-19-un-rattrapage-des-depenses-de-sante-qui-reste-hypothetique.172354
- https://www.argusdelassurance.com/assurance-de-personnes/sante/covid-19-l-impact-de-la-pandemie-mesure-par-la-secu.164061

Assurance Maladie :

- https://assurance-maladie.ameli.fr/etudes-et-donnees/donnees/bases-de-donnees-open-data/liste-bases-de-donnees-open-data

Cour des comptes :

- https://www.ccomptes.fr/fr/documents/50123
- https://www.ccomptes.fr/fr/documents/52078

Data Gouv :

- https://www.data.gouv.fr/fr/datasets/open-damir-base-complete-sur-les-depenses-dassurance-maladie-inter-regimes/

Direction de la recherche, des études, de l'évaluation et des statistiques

- https://drees2-sgsocialgouv.opendatasoft.com/explore/dataset/306_les-comptes-de-la-sante/information/
- https://drees.solidarites-sante.gouv.fr/communique-de-presse/communique-de-presse/en-2019-3-102-euros-de-depenses-de-sante-par-habitant
- https://drees.solidarites-sante.gouv.fr/article/les-travaux-de-la-drees-lies-la-crise-sanitaire-de-la-covid-19

GitLab :

- https://gitlab.com/healthdatahub/open-data/open-damir

Légifrance :

- https://www.legifrance.gouv.fr/cnil/id/CNILTEXT000036641332/

Sécurité Sociale :

- https://www.securite-sociale.fr/files/live/sites/SSFR/files/medias/CCSS/2020/RAPPORT%20CCSS%20JUIN%202020.pdf

IX. Annexe

A. Annexe 1 : vérification des hypothèses de Chain Ladder sur les autres postes de consommation

1. Soins de ville courants

a) L'alignement des couples $(C_{i,j}, C_{i,j+1})$

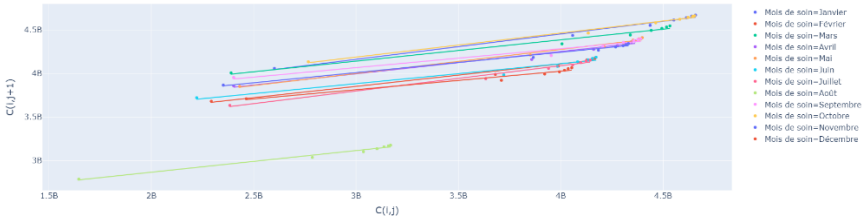


Figure IX-1 – Représentation graphique des couples $(C_{i,j}, C_{i,j+1})$ en 2019 pour le poste de consommation Soins de ville courants

Mois	$R^2 ajusté$
Janvier	98,72 %
Février	97,98 %
Mars	97,49 %
Avril	97,08 %
Mai	98,75 %
Juin	94,77 %
Juillet	98,18 %
Août	98,03 %
Septembre	94,47 %
Octobre	98,83 %
Novembre	98,73 %
Décembre	95,52 %

Figure IX-2 – Tableau des R^2 ajusté de chaque droite de la figure IX-1

b) L'examen du triangle de développement

	0	**1**	**2**	**3**	**4**	**5**
$\hat{f}_j$	159,63 %	107,99 %	102,42 %	101,04 %	100,55 %	100,36 %
$\mathbb{E}[f_{i,j}]$	160,12 %	108 %	102,41 %	101,04 %	100,55 %	100,36 %
$Var[f_{i,j}]$	0,432 %	0,012 %	0,001 %	0,000 %	0,000 %	0,000 %

Figure IX-3 – Tableau de comparaison entre les facteurs de développement individuel et les facteurs de passage pour les 6 premiers mois de développement sur les survenances de 2019

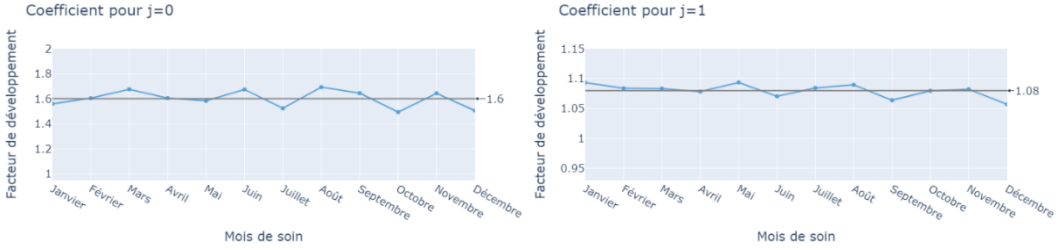


Figure IX-4 – Coefficients individuels du premier et deuxième mois de développement selon les mois de survenance de 2019

2. Pharmacie

a) L'alignement des couples $(C_{i,j}, C_{i,j+1})$

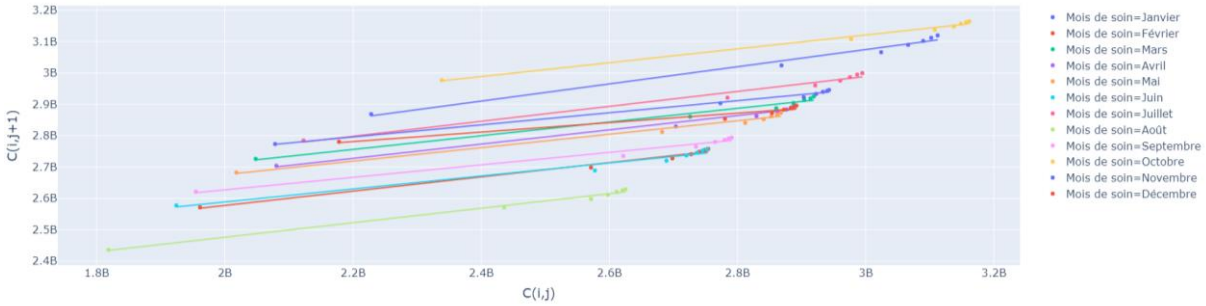


Figure IX-5 – Représentation graphique des couples $(C_{i,j}, C_{i,j+1})$ en 2019 pour le poste de consommation Pharmacie

Mois	$R^2 ajusté$
Janvier	93,98 %
Février	98,64 %
Mars	97,97 %
Avril	98,25 %
Mai	98,3 %
Juin	96,09 %
Juillet	97,81 %
Août	98,89 %
Septembre	97,41 %
Octobre	98,9 %
Novembre	99,14 %
Décembre	94,63 %

Figure IX-6 – Tableau des R^2 ajusté de chaque droite de la figure IX-5

b) L'examen du triangle de développement

	0	**1**	**2**	**3**	**4**	**5**
$\hat{f}_j$	131,32 %	104,63 %	101,04 %	100,50 %	100,29 %	100,2 %
$\mathbb{E}[f_{i,j}]$	131,45 %	104,64 %	101,04 %	100,50 %	100,29 %	100,2 %
$Var[f_{i,j}]$	0,058 %	0,005 %	0,001 %	0,000 %	0,000 %	0,000 %

Figure IX-7 – Tableau de comparaison entre les facteurs de développement individuel et les facteurs de passage pour les 6 premiers mois de développement sur les survenances de 2019

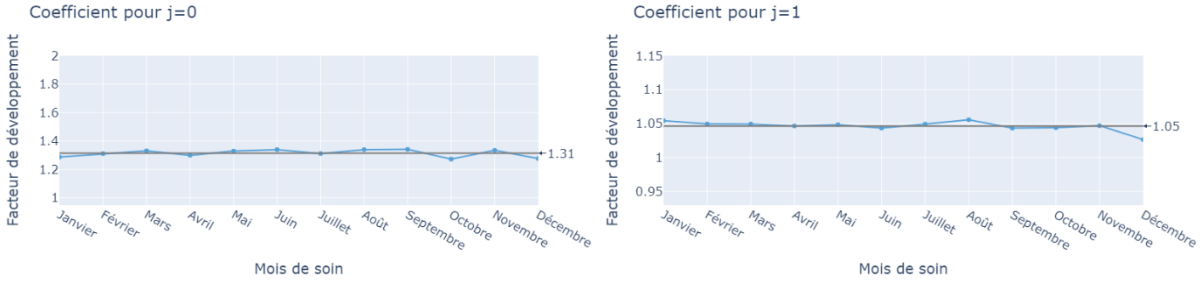


Figure IX-8 – Coefficients individuels du premier et deuxième mois de développement selon les mois de survenance de 2019

3. Dentaire

a) L'alignement des couples $(C_{i,j}, C_{i,j+1})$

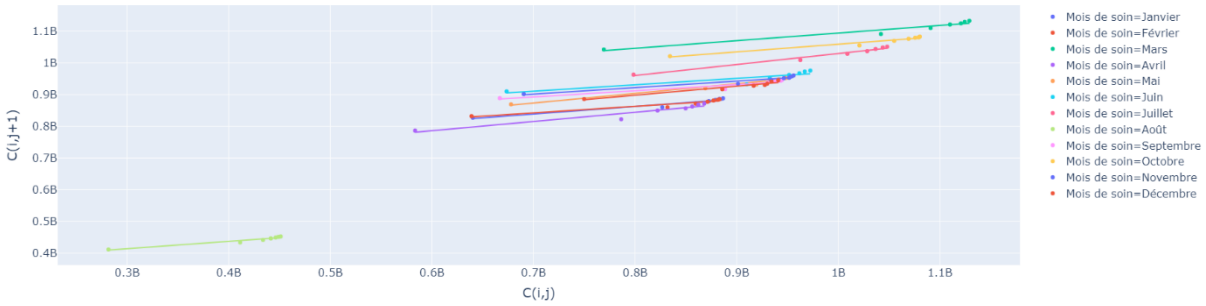


Figure IX-9 – Représentation graphique des couples $(C_{i,j}, C_{i,j+1})$ en 2019 pour le poste de consommation Dentaire

Mois	R^2 ajusté
Janvier	93,76 %
Février	90,94 %
Mars	94,02 %
Avril	90,41 %
Mai	96,08 %
Juin	77,42 %
Juillet	97,39 %
Août	93,12 %
Septembre	88,72 %
Octobre	95,11 %
Novembre	93,35 %
Décembre	93,05 %

Figure IX-10 – Tableau des R^2 ajusté de chaque droite de la figure IX-9

b) L'examen du triangle de développement

	0	**1**	**2**	**3**	**4**	**5**
$\hat{f}_j$	129,12 %	104 %	101,68 %	100,77 %	100,47 %	100,42 %
$\mathbb{E}[f_{i,j}]$	130,31 %	104,07 %	101,68 %	100,78 %	10048 %	100,43 %
$Var[f_{i,j}]$	0,532 %	0,009 %	0,003 %	0,001 %	0,000 %	0,000 %

Figure IX-11 – Tableau de comparaison entre les facteurs de développement individuel et les facteurs de passage pour les 6 premiers mois de développement sur les survenances de 2019

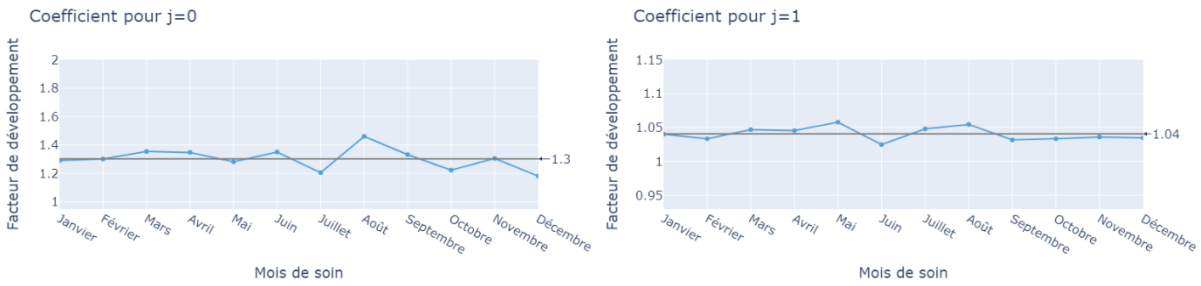


Figure IX-12 – Coefficients individuels du premier et deuxième mois de développement selon les mois de survenance de 2019

4. Prothèses auditives

a) L'alignement des couples $(C_{i,j}, C_{i,j+1})$

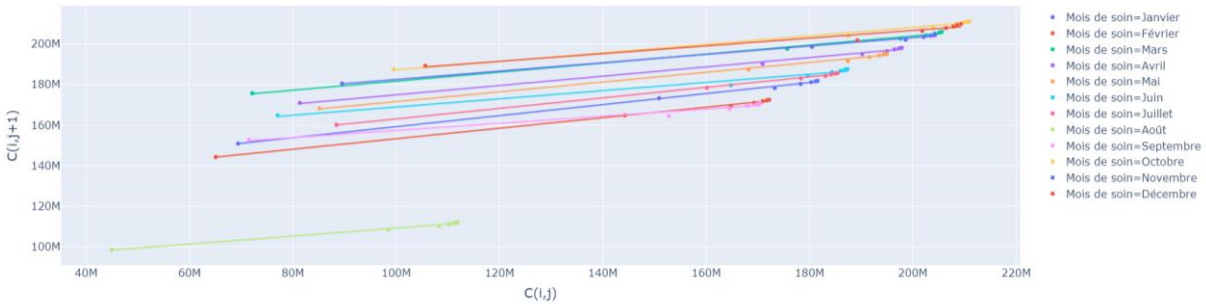


Figure IX-13 – Représentation graphique des couples $(C_{i,j}, C_{i,j+1})$ en 2019 pour le poste de consommation Prothèses auditives

Mois	R^2 ajusté
Janvier	99,82 %
Février	99,8 %
Mars	99,55 %
Avril	99,47 %
Mai	99,44 %
Juin	97,55 %
Juillet	99,84 %
Août	98,95 %
Septembre	96,64 %
Octobre	98,91 %
Novembre	99,65 %
Décembre	95,6 %

Figure IX-14 – Tableau des R^2 ajusté de chaque droite de la figure IX-13

b) L'examen du triangle de développement

	0	1	2	3	4	5
$\hat{f}_j$	204,63 %	110,56 %	102,37 %	100,82 %	100,40 %	100,25 %
$\mathbb{E}[f_{i,j}]$	207,27 %	110,65 %	102,36 %	100,83 %	100,41 %	100,24 %
$Var[f_{i,j}]$	3,182 %	0,056 %	0,001 %	0,000 %	0,000 %	0,000 %

Figure IX-15 – Tableau de comparaison entre les facteurs de développement individuel et les facteurs de passage pour les 6 premiers mois de développement sur les survenances de 2019

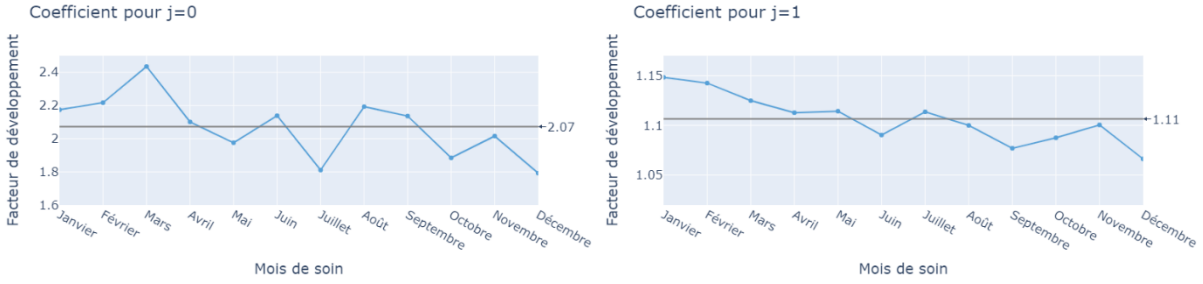


Figure IX-16 – Coefficients individuels du premier et deuxième mois de développement selon les mois de survenance de 2019

B. Annexe 2 : test rétroactif de validité sur les autres postes de consommation

1. Pharmacie

Mois de survenance i	Vision au 31/12/2020				
	Dépense réelle observée	Dépense projetée (méthode A)	Dépense projetée (méthode B)	**Différence d'estimation (méthode A)**	**Différence d'estimation (méthode B)**
Janvier	3 140,46	3 109,97	3 115,47	**- 30,49**	**- 24,98**
Février	2 853,50	2 773,68	2 778,38	**- 79,82**	**- 75,12**
Mars	3 082,67	3 284,76	3 290,14	**202,08**	**207,46**
Avril	2 602,54	2 540,03	2 544,10	**- 62,51**	**- 58,44**
Mai	2 624,08	2 449,09	2 452,91	**- 174,99**	**- 171,17**
Juin	2 870,52	2 930,49	2 935,10	**59,97**	**64,58**
Juillet	2 984,78	3 032,97	3 038,17	**48,19**	**53,39**
Août	2 659,08	2 542,15	2 547,69	**- 116,93**	**- 111,39**
Septembre	2 956,08	2 983,10	2 991,94	**27,02**	**35,86**
Octobre	3 302,32	3 267,47	3 279,49	**- 34,85**	**- 22,83**
Novembre	2 888,27	2 812,88	2 825,54	**- 75,39**	**- 62,73**
Décembre	2 378,87	N/A	N/A	**N/A**	**N/A**
Total	**31 964,29**	**31 726,57**	**31 798,93**	**- 0,7 %**	**- 0,5 %**

Figure IX-17 – Application du test rétroactif de validité sur le poste de consommation Pharmacie (en millions d'euros)

2. Dentaire

Mois de survenance i	Vision au 31/12/2020				
	Dépense réelle observée	Dépense projetée (méthode A)	Dépense projetée (méthode B)	**Différence d'estimation (méthode A)**	**Différence d'estimation (méthode B)**
Janvier	916,72	921,43	909,22	**4,71**	**- 7,51**
Février	954,28	985,15	972,08	**30,87**	**17,80**
Mars	609,15	642,00	633,43	**32,85**	**24,28**
Avril	87,20	19,20	18,94	**- 68,00**	**- 68,26**
Mai	671,37	565,80	558,22	**- 105,57**	**- 113,15**
Juin	1 192,00	1 245,48	1 228,73	**53,48**	**36,72**
Juillet	1 137,76	1 240,68	1 224,03	**102,91**	**86,26**
Août	543,16	480,18	473,94	**- 62,98**	**- 69,23**
Septembre	1 031,73	1 054,56	1 041,67	**22,83**	**9,94**
Octobre	1 050,00	1 117,95	1 104,43	**67,95**	**54,43**
Novembre	1 087,67	1 071,72	1 060,02	**- 15,95**	**- 27,64**
Décembre	893,16	N/A	N/A	**N/A**	**N/A**
Total	**9 281,04**	**9 344,14**	**9 224,70**	**0,7 %**	**- 0,6 %**

Figure IX-18 – Application du test rétroactif de validité sur le poste de consommation Dentaire (en millions d'euros)

3. Optique

Mois de survenance i	Vision au 31/12/2020				
	Dépense réelle observée	Dépense projetée (méthode A)	Dépense projetée (méthode B)	**Différence d'estimation (méthode A)**	**Différence d'estimation (méthode B)**
Janvier	364,50	326,61	286,01	**- 37,89**	**- 78,49**
Février	582,58	603,76	528,78	**21,18**	**- 53,80**
Mars	334,14	427,19	374,20	**93,05**	**40,06**
Avril	37,93	32,40	28,38	**- 5,53**	**- 9,55**
Mai	317,71	237,42	208,07	**- 80,29**	**- 109,64**
Juin	656,32	767,27	672,68	**110,95**	**16,36**
Juillet	671,83	793,02	695,70	**121,19**	**23,86**
Août	499,54	542,08	476,01	**42,54**	**- 23,52**
Septembre	612,05	734,73	646,32	**122,68**	**34,27**
Octobre	684,40	777,85	686,61	**93,45**	**2,21**
Novembre	534,59	586,16	524,75	**51,57**	**- 9,83**
Décembre	508,85	N/A	N/A	**N/A**	**N/A**
Total	**5 295,58**	**5 828,50**	**5 127,51**	**10,1 %**	**- 3,2 %**

Figure IX-19 – Application du test rétroactif de validité sur le poste de consommation Optique (en millions d'euros)

4. Prothèses auditives

Mois de survenance i	Vision au 31/12/2020				
	Dépense réelle observée	Dépense projetée (méthode A)	Dépense projetée (méthode B)	**Différence d''estimation (méthode A)**	**Différence d'estimation (méthode B)**
Janvier	204,39	223,03	198,90	**18,64**	**- 5,49**
Février	198,11	216,01	192,64	**17,90**	**- 5,47**
Mars	134,50	163,54	145,85	**29,05**	**11,35**
Avril	35,71	26,53	23,66	**- 9,18**	**-12,05**
Mai	132,15	133,15	118,77	**1,00**	**- 13,37**
Juin	212,91	269,36	240,36	**56,45**	**27,45**
Juillet	227,10	302,33	269,94	**75,22**	**42,83**
Août	138,02	159,71	142,77	**21,69**	**4,75**
Septembre	207,22	254,42	227,96	**47,21**	**20,74**
Octobre	222,67	286,80	258,29	**64,12**	**35,62**
Novembre	199,14	242,39	222,69	**43,25**	**23,56**
Décembre	141,77	N/A	N/A	**N/A**	**N/A**
Total	**1 911,91**	**2 277,26**	**2 041,84**	**19,1 %**	**6,8 %**

Figure IX-20 – Application du test rétroactif de validité sur le poste de consommation Prothèses auditives (en millions d'euros)

C. Annexe 3 : étude sur les cadences de règlements pour les autres postes de consommation

1. Pharmacie

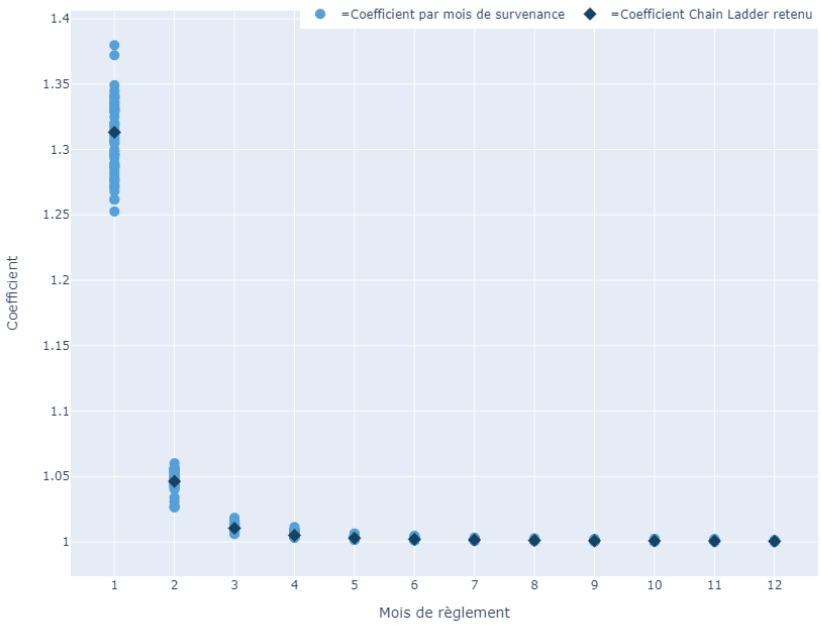


Figure IX-21 – Coefficients de passage sur les 12 premiers mois de règlement pour le poste Pharmacie

Mois de survenance	Dépenses 3 mois après	Dépenses 12 mois après	Dépense réelle au 31/12/2020	% 3 mois après	% 12 mois après
Janvier 2015	2 948,84	3 016,73	3 024,73	97,5 %	99,7 %
Janvier 2016	2 794,82	2 852,32	2 857,73	97,8 %	99,8 %
Janvier 2017	2 955,86	3 007,12	3 014,88	98,0 %	99,7 %
Janvier 2018	3 017,31	3 095,22	3 112,86	96,9 %	99,4 %
Janvier 2019	3 066,19	3 148,58	3 161,06	97,0 %	99,6 %

Figure IX-22 – Vision de la consommation 3 et 12 mois après la dépense pour le poste Pharmacie

2. Dentaire

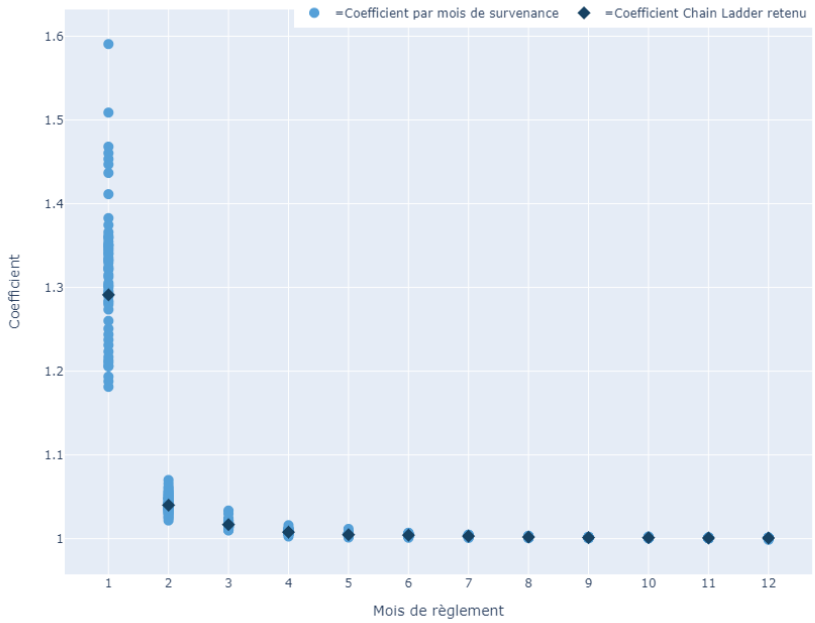


Figure IX-23 – Coefficients de passage sur les 12 premiers mois de règlement pour le poste Dentaire

Mois de survenance	Dépenses 3 mois après	Dépenses 12 mois après	Dépense réelle au 31/12/2020	% 3 mois après	% 12 mois après
Janvier 2015	758,42	790,14	793,86	95,5 %	99,5 %
Janvier 2016	792,04	814,77	817,83	96,8 %	99,6 %
Janvier 2017	861,26	882,91	886,18	97,2 %	99,6 %
Janvier 2018	863,63	885,90	888,84	97,2 %	99,7 %
Janvier 2019	872,61	895,10	898,68	97,1 %	99,6 %

Figure IX-24 – Vision de la consommation 3 et 12 mois après la dépense pour le poste Dentaire

3. Optique

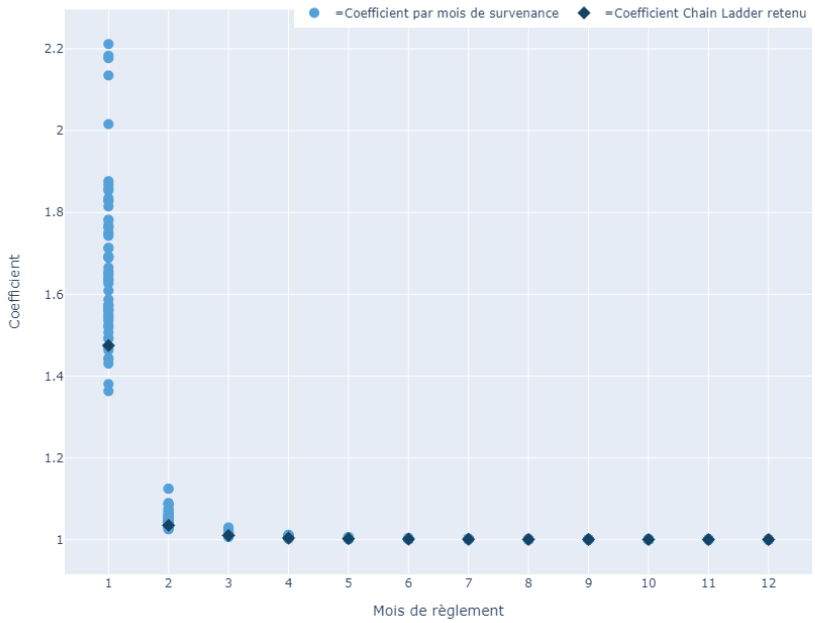


Figure IX-25 – Coefficients de passage sur les 12 premiers mois de règlement pour le poste Optique

Mois de survenance	Dépenses 3 mois après	Dépenses 12 mois après	Dépense réelle au 31/12/2020	% 3 mois après	% 12 mois après
Janvier 2015	469,15	483,35	485,41	96,7 %	99,6 %
Janvier 2016	442,77	453,27	455,22	97,3 %	99,6 %
Janvier 2017	429,12	436,71	438,18	97,9 %	99,7 %
Janvier 2018	469,97	477,15	478,72	98,2 %	99,7 %
Janvier 2019	502,53	510,14	511,20	98,3 %	99,8 %

Figure IX-26 – Vision de la consommation 3 et 12 mois après la dépense pour le poste Optique

4. Prothèses auditives

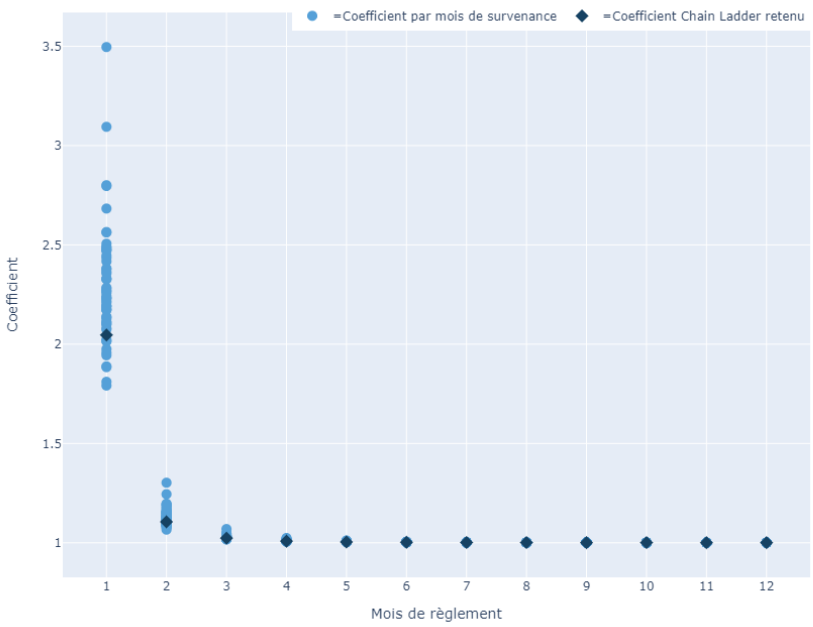


Figure IX-27 – Coefficients de passage sur les 12 premiers mois de règlement pour le poste Prothèses auditives

Mois de survenance	Dépenses 3 mois après	Dépenses 12 mois après	Dépense réelle au 31/12/2020	% 3 mois après	% 12 mois après
Janvier 2015	156,74	163,39	163,83	95,7 %	99,7 %
Janvier 2016	168,75	173,69	174,24	96,8 %	99,7 %
Janvier 2017	184,35	189,55	190,12	97,0 %	99,7 %
Janvier 2018	191,50	196,25	197,06	97,2 %	99,6 %
Janvier 2019	178,29	182,63	183,02	97,4 %	99,8 %

Figure IX-28 – Vision de la consommation 3 et 12 mois après la dépense pour le poste Prothèses auditives

D. Annexe 4 : application des modèles ARIMA & SARIMA sur les autres postes de consommation

1. Pharmacie

Le poste de consommation *Pharmacie* peut être décomposé de la façon suivante via des statistiques par mois et année de dépense :

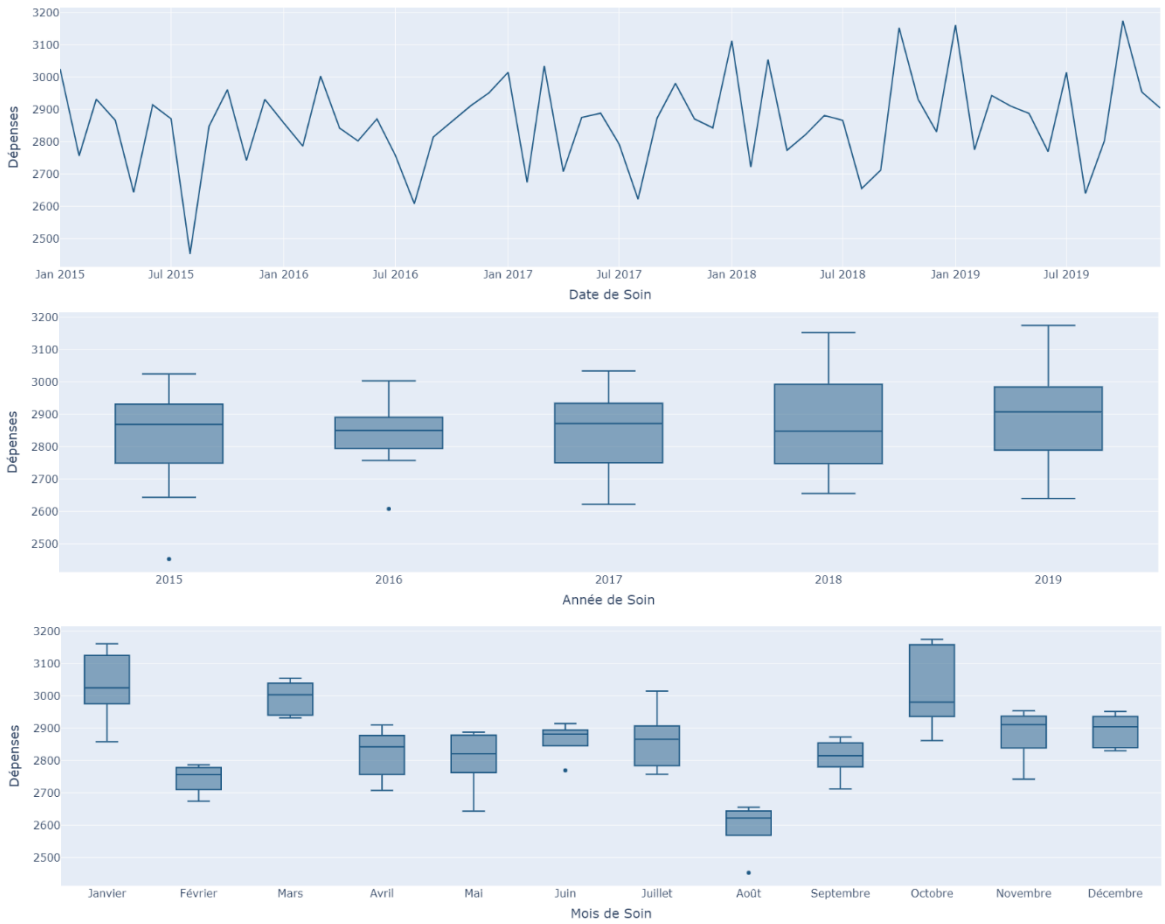


Figure IX-29 – Décomposition de la série temporelle du poste de consommation Pharmacie (en millions d'euros)

a) La stationnarité de notre série temporelle

Comme pour l'analyse du poste de consommation *Soins de ville courants*, nous devons vérifier que notre série temporelle est bien stationnaire. Nous effectuons donc les tests ADF et KPSS sur notre série sans aucun pré-traitement et résumons les résultats obtenus dans le tableau ci-dessous.

	Test ADF	**Test KPSS**	**Conclusion**
P-value associée	0,93	0,036	Rejet de la stationnarité

Figure IX-30 – Test de stationnarité sur la série temporelle Pharmacie avant prétraitement

A l'aide de ces deux tests, nous pouvons en conclure que notre série n'est pas stationnaire, il est donc nécessaire de la rendre stationnaire. Pour cela, nous allons effectuer une différenciation d'ordre un. Nous effectuons de nouveau le test ADF et KPSS sur ce nouveau jeu de données et pouvons en conclure que la série temporelle et cette fois-ci bien stationnaire.

	Test ADF	Test KPSS	Conclusion
P-value associée	0,000	0,1	Série stationnaire

Figure IX-31 – Test de stationnarité sur la série temporelle Pharmacie après prétraitement

b) Le modèle ARIMA

SARIMAX Results

Dep. Variable:	y	No. Observations:	59
Model:	SARIMAX(11, 0, 1)	Log Likelihood	-1156.818
		AIC	2341.636
		BIC	2370.721
		HQIC	2352.990

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	2.112e+07	2.59e-08	8.16e+14	0.000	2.11e+07	2.11e+07
ar.L1	-0.8742	0.219	-3.986	0.000	-1.304	-0.444
ar.L2	-0.9034	0.147	-6.153	0.000	-1.191	-0.616
ar.L3	-0.8832	0.192	-4.600	0.000	-1.259	-0.507
ar.L4	-0.9170	0.185	-4.948	0.000	-1.280	-0.554
ar.L5	-0.9624	0.147	-6.538	0.000	-1.251	-0.674
ar.L6	-0.9029	0.185	-4.888	0.000	-1.265	-0.541
ar.L7	-0.9768	0.146	-6.684	0.000	-1.263	-0.690
ar.L8	-0.9428	0.202	-4.666	0.000	-1.339	-0.547
ar.L9	-0.8246	0.175	-4.712	0.000	-1.168	-0.482
ar.L10	-0.7802	0.199	-3.920	0.000	-1.170	-0.390
ar.L11	-0.7341	0.139	-5.268	0.000	-1.007	-0.461
ma.L1	-0.7743	0.245	-3.155	0.002	-1.255	-0.293
sigma2	6.428e+15	1.32e-17	4.88e+32	0.000	6.43e+15	6.43e+15

Ljung-Box (L1) (Q):	0.26	Jarque-Bera (JB):	5.11
Prob(Q):	0.61	Prob(JB):	0.08
Heteroskedasticity (H):	0.83	Skew:	-0.63
Prob(H) (two-sided):	0.69	Kurtosis:	3.69

Figure IX-32 – Modèle ARIMA pour le poste de consommation Pharmacie

Après application de la fonction $auto\_arima([\dots])$ sur notre jeu de données stationnaire, nous obtenons le modèle $ARIMA(11,0,1)$ résumé dans la figure ci-dessus. Nous pouvons remarquer que tous les termes sont significatifs. L'équation du modèle est la suivante :

$$\hat{y}_t = 2{,}11 * 10^7 - 0{,}87 * \hat{y}_{t-1} - 0{,}9 * \hat{y}_{t-2} - 0{,}88 * \hat{y}_{t-3} - 0{,}92 * \hat{y}_{t-4} - 0{,}96 * \hat{y}_{t-5} - 0{,}9 * \hat{y}_{t-6} - 0{,}98 * \hat{y}_{t-7} - 0{,}94 * \hat{y}_{t-8} - 0{,}82 * \hat{y}_{t-9} - 0{,}78 * \hat{y}_{t-10} - 0{,}73 * \hat{y}_{t-11} + \epsilon_t - 0{,}77 * \epsilon_{t-1}$$

L'analyse des résidus du modèle ne nous permet pas d'affirmer facilement qu'il s'agit d'un bruit blanc. En effet, un doute peut subsister sur la normalité de données au vu de la faible valeur de la p-value. Cependant, à l'aide du graphique en haut à droite de la figure ci-dessous, nous pouvons dire que nos résidus semblent être gaussiens. Ainsi, nous admettons que les résidus de notre modèle sont bien un bruit blanc de moyenne $-\ 20{,}63 *\ 10^6$.

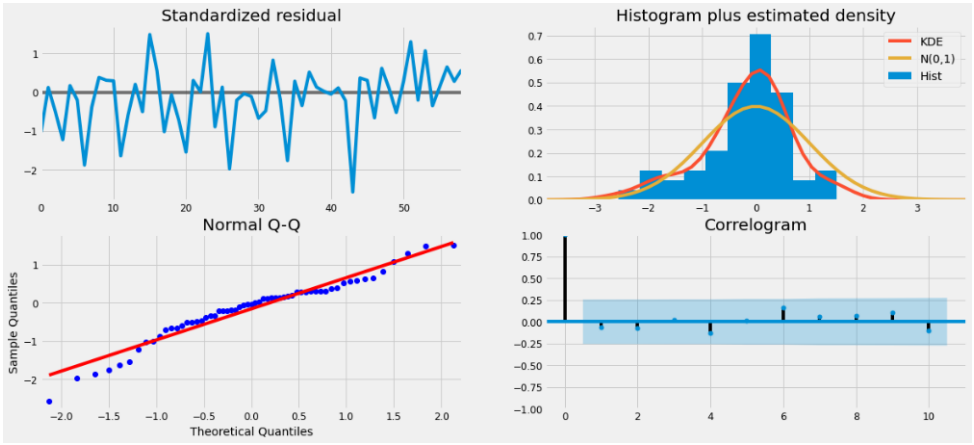


Figure IX-33 – Analyse des résidus du modèle ARIMA du poste de consommation Pharmacie

c) Le modèle SARIMA

SARIMAX Results

Dep. Variable:	y	No. Observations:	59
Model:	SARIMAX(2, 0, 0)x(1, 1, 0, 12)	Log Likelihood	-928.155
		AIC	1866.310
		BIC	1875.560
		HQIC	1869.791

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	2.231e+06	8.8e-10	2.54e+15	0.000	2.23e+06	2.23e+06
ar.L1	-0.9644	0.157	-6.155	0.000	-1.272	-0.657
ar.L2	-0.5546	0.149	-3.723	0.000	-0.846	-0.263
ar.S.L12	-0.4284	0.156	-2.746	0.006	-0.734	-0.123
sigma2	9.323e+15	3.74e-18	2.49e+33	0.000	9.32e+15	9.32e+15

Ljung-Box (L1) (Q):	0.65	Jarque-Bera (JB):	0.68
Prob(Q):	0.42	Prob(JB):	0.71
Heteroskedasticity (H):	0.59	Skew:	-0.17
Prob(H) (two-sided):	0.30	Kurtosis:	2.52

Figure IX-34 – Modèle SARIMA pour le poste de consommation Pharmacie

Nous avons pu remarquer que les termes autorégressifs du modèle ARIMA pour le poste de consommation *Pharmacie* essaient de capter une saisonnalité dans notre jeu de données. Cela nous invite donc à appliquer un modèle $SARIMA$ dont les résultats sont résumés dans la figure ci-dessus. Le modèle obtenu est un modèle $SARIMA(2,0,0)(1,1,0)_{12}$ dont l'équation est :

$$\hat{y}_t = 2{,}31 * 10^6 - 0{,}96 * \hat{y}_{t-1} - 0{,}55 * \hat{y}_{t-2} + 0{,}57 * \hat{y}_{t-12} + 0{,}55 * \hat{y}_{t-13} + 0{,}31 * \hat{y}_{t-14} + 0{,}43 * \hat{y}_{t-24} + 0{,}41 * \hat{y}_{t-25} + 0{,}24 * \hat{y}_{t-26} + \epsilon_t$$

L'analyse des résidus du modèle nous fait admettre que nos résidus sont un bruit blanc de moyenne $4{,}41 * 10^6$.

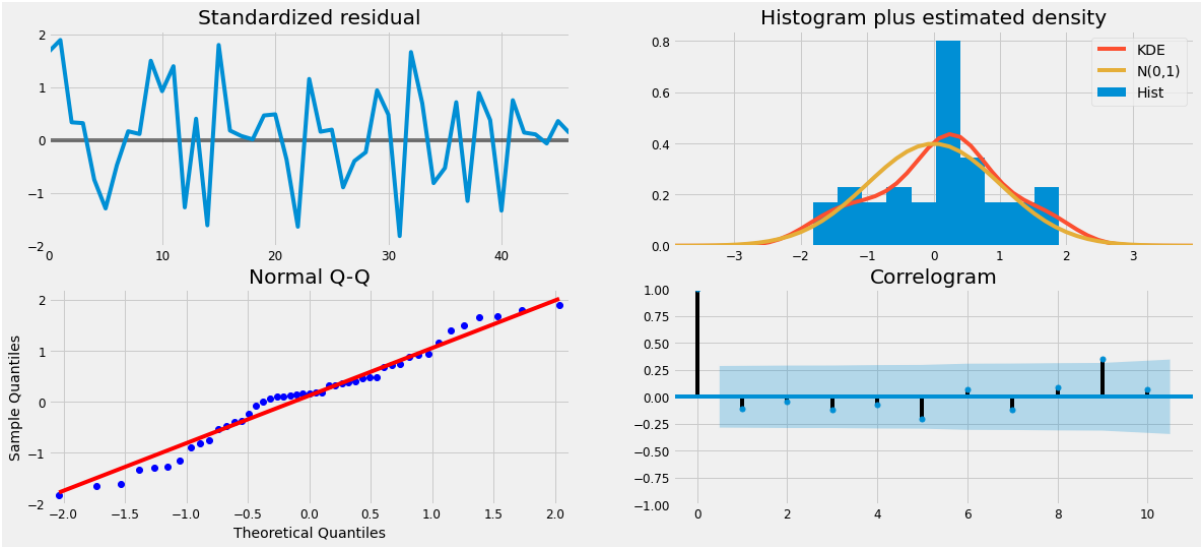


Figure IX-35 – Analyse des résidus du modèle SARIMA du poste de consommation Pharmacie

2. Dentaire

Le poste de consommation *Dentaire* peut être décomposé de la façon suivante via des statistiques par mois et année de dépense :

Figure IX-36 – Décomposition de la série temporelle du poste de consommation Dentaire (en millions d'euros)

a) La stationnarité de notre série temporelle

Comme lors de l'analyse de la stationnarité des séries temporelles des deux postes de consommation précédents, nous effectuons donc le test ADF et KPSS sur notre série sans aucun pré-traitement et résumons les résultats obtenus dans le tableau ci-dessous.

	Test ADF	Test KPSS	Conclusion
P-value associée	0,724	0,021	Rejet de la stationnarité

Figure IX-37 – Test de stationnarité sur la série temporelle Dentaire avant prétraitement

A l'aide de ces deux tests, nous pouvons en conclure que notre série n'est pas stationnaire, il est donc nécessaire de transformer la série en la différenciant à l'ordre 1. Nous effectuons de nouveau le test ADF et KPSS sur ce nouveau jeu de données et pouvons en conclure que la série temporelle est cette fois-ci bien stationnaire.

	Test ADF	Test KPSS	Conclusion
P-value associée	0,000	0,1	Série stationnaire

Figure IX-38 – Test de stationnarité sur la série temporelle Dentaire après prétraitement

b) *Le modèle ARIMA*

SARIMAX Results

Dep. Variable:	y	No. Observations:	59
Model:	SARIMAX(11, 0, 1)	Log Likelihood	-1136.460
		AIC	2300.921
		BIC	2330.006
		HQIC	2312.275

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	2.314e+07	2.4e-09	9.64e+15	0.000	2.31e+07	2.31e+07
ar.L1	-0.9691	0.071	-13.576	0.000	-1.109	-0.829
ar.L2	-1.0123	0.062	-16.220	0.000	-1.135	-0.890
ar.L3	-0.9895	0.061	-16.170	0.000	-1.109	-0.870
ar.L4	-0.9928	0.087	-11.473	0.000	-1.162	-0.823
ar.L5	-0.9860	0.077	-12.750	0.000	-1.138	-0.834
ar.L6	-0.9666	0.082	-11.859	0.000	-1.126	-0.807
ar.L7	-0.9924	0.063	-15.731	0.000	-1.116	-0.869
ar.L8	-0.9728	0.062	-15.801	0.000	-1.093	-0.852
ar.L9	-0.9787	0.059	-16.523	0.000	-1.095	-0.863
ar.L10	-0.9677	0.043	-22.474	0.000	-1.052	-0.883
ar.L11	-0.9393	0.057	-16.396	0.000	-1.052	-0.827
ma.L1	-0.4081	0.216	-1.886	0.059	-0.832	0.016
sigma2	2.234e+15	4.78e-18	4.67e+32	0.000	2.23e+15	2.23e+15

Ljung-Box (L1) (Q):	1.02	Jarque-Bera (JB):	1.01
Prob(Q):	0.31	Prob(JB):	0.60
Heteroskedasticity (H):	1.02	Skew:	-0.11
Prob(H) (two-sided):	0.97	Kurtosis:	3.60

Figure IX-39 – Modèle ARIMA pour le poste de consommation Dentaire

Le meilleur modèle pour notre jeu de données est un modèle $ARIMA(11,0,1)$. L'équation du modèle est la suivante :

$$\hat{y}_t = 2{,}31 * 10^7 - 0{,}97 * \hat{y}_{t-1} - 1{,}01 * \hat{y}_{t-2} - 0{,}99 * \hat{y}_{t-3} - 0{,}99 * \hat{y}_{t-4} - 0{,}99 * \hat{y}_{t-5} - 0{,}97 * \hat{y}_{t-6} - 0{,}99 * \hat{y}_{t-7} - 0{,}97 * \hat{y}_{t-8} - 0{,}98 * \hat{y}_{t-9} - 0{,}97 * \hat{y}_{t-10} - 0{,}94 * \hat{y}_{t-11} + \epsilon_t - 0{,}41 * \epsilon_{t-1}$$

Comme précédemment, nous admettons que nos résidus sont du bruit blanc de moyenne $-2{,}95 * 10^6$.

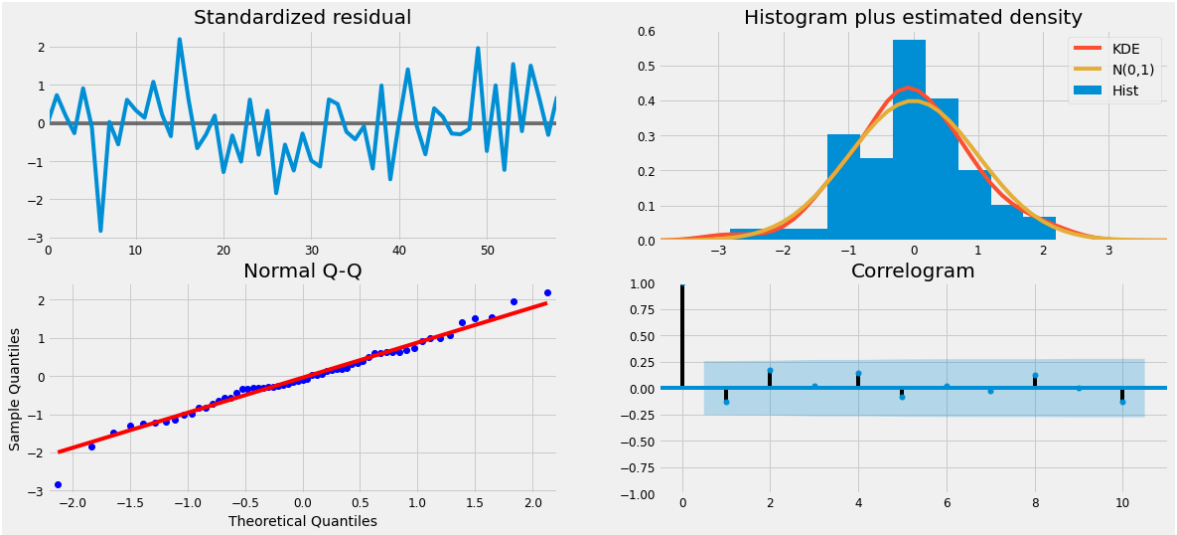


Figure IX-40 – Analyse des résidus du modèle ARIMA du poste de consommation Dentaire

c) *Le modèle SARIMA*

SARIMAX Results

Dep. Variable:	y	No. Observations:	59
Model:	SARIMAX(1, 0, 1)x(0, 1, [], 12)	Log Likelihood	-895.425
		AIC	1798.849
		BIC	1806.250
		HQIC	1801.634

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	4.229e+05	1.34e+06	0.316	0.752	-2.2e+06	3.05e+06
ar.L1	-0.5219	0.145	-3.597	0.000	-0.806	-0.238
ma.L1	-0.8222	0.113	-7.307	0.000	-1.043	-0.602
sigma2	2.108e+15	0.000	8.71e+18	0.000	2.11e+15	2.11e+15

Ljung-Box (L1) (Q):	0.32	Jarque-Bera (JB):	0.58
Prob(Q):	0.57	Prob(JB):	0.75
Heteroskedasticity (H):	0.70	Skew:	0.27
Prob(H) (two-sided):	0.49	Kurtosis:	2.91

Figure IX-41 – Modèle SARIMA pour le poste de consommation Dentaire

Comme lors de l'application du modèle ARIMA pour les modèles précédents, les termes autorégressifs tentent de capter une saisonnalité dans notre jeu de données. Nous appliquons donc un modèle $SARIMA(1,0,1)(0,1,0)_{12}$ à nos données. Son équation est la suivante :

$$\hat{y}_t = 4{,}23 * 10^5 - 0{,}52 * \hat{y}_{t-1} + \hat{y}_{t-12} + 0{,}52 * \hat{y}_{t-13} + \epsilon_t - 0{,}82 * \epsilon_{t-1}$$

Nous admettons que nos résidus sont du bruit blanc de moyenne $11 * 10^6$.

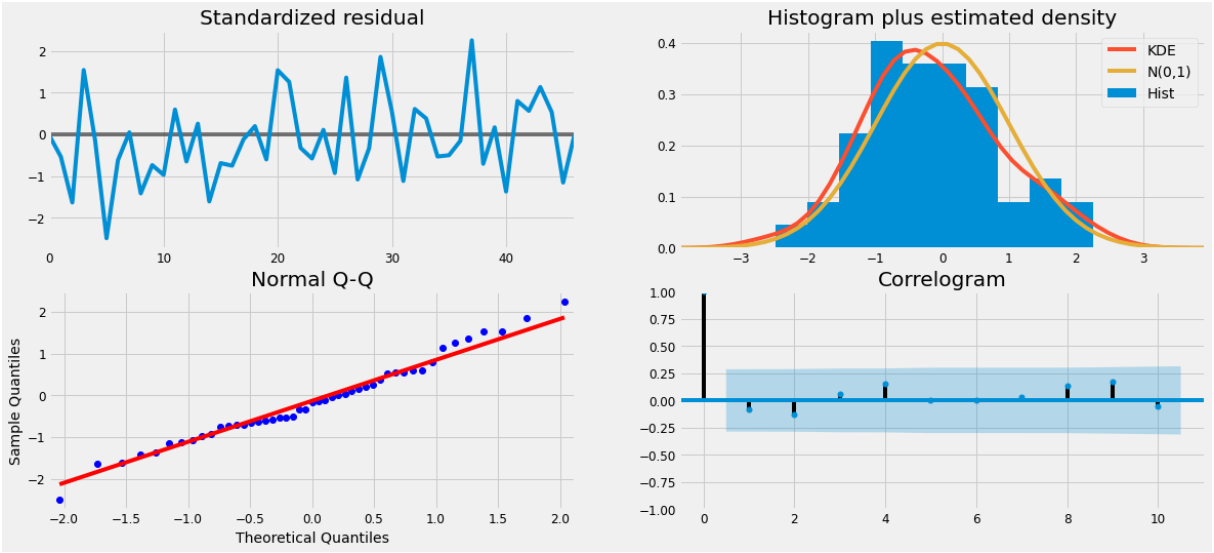


Figure IX-42 – Analyse des résidus du modèle SARIMA du poste de consommation Dentaire

3. Optique

Le poste de consommation *Optique* peut être décomposé de la façon suivante via des statistiques par mois et année de dépense :

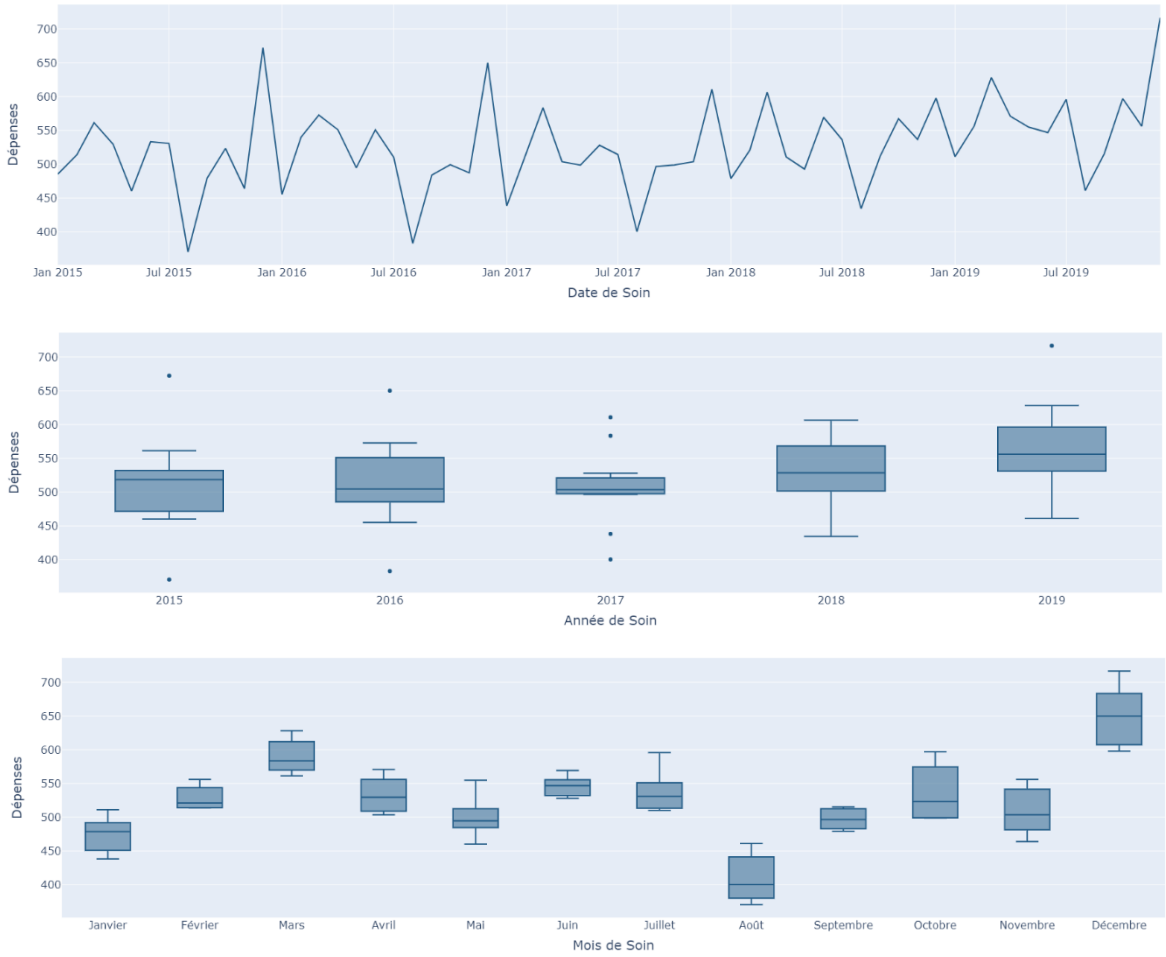


Figure IX-43 – Décomposition de la série temporelle du poste de consommation Optique (en millions d'euros)

(a)

La stationnarité de notre série temporelle

Comme pour l'analyse des postes de consommation précédents, nous effectuons le test ADF et KPSS sur notre série sans aucun pré-traitement afin de vérifier si celle-ci est stationnaire et résumons les résultats obtenus dans le tableau ci-dessous.

	Test ADF	Test KPSS	Conclusion
P-value associée	0,995	0,035	Rejet de la stationnarité

Figure IX-44 – Test de stationnarité sur la série temporelle Optique avant prétraitement

A l'aide de ces deux tests, nous pouvons en conclure que notre série n'est pas stationnaire.

Nous allons effectuer une différenciation d'ordre 1 sur celle-ci et effectuons de nouveau le test ADF et KPSS. Nous pouvons en conclure que la série temporelle est cette fois-ci bien stationnaire.

	Test ADF	Test KPSS	Conclusion
P-value associée	0,0001	0,094	Série stationnaire

Figure IX-45 – Test de stationnarité sur la série temporelle Optique après prétraitement

a) Le modèle ARIMA

SARIMAX Results

Dep. Variable:	y	No. Observations:	59
Model:	SARIMAX(11, 0, 0)	Log Likelihood	-1107.046
		AIC	2240.092
		BIC	2267.100
		HQIC	2250.635

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	1.285e+07	5.89e-09	2.18e+15	0.000	1.29e+07	1.29e+07
ar.L1	-0.9716	0.071	-13.777	0.000	-1.110	-0.833
ar.L2	-0.9486	0.092	-10.362	0.000	-1.128	-0.769
ar.L3	-0.8903	0.137	-6.515	0.000	-1.158	-0.622
ar.L4	-0.9267	0.138	-6.694	0.000	-1.198	-0.655
ar.L5	-0.8689	0.158	-5.482	0.000	-1.179	-0.558
ar.L6	-0.8820	0.147	-5.997	0.000	-1.170	-0.594
ar.L7	-0.8994	0.120	-7.521	0.000	-1.134	-0.665
ar.L8	-0.9140	0.121	-7.582	0.000	-1.150	-0.678
ar.L9	-0.8628	0.101	-8.512	0.000	-1.062	-0.664
ar.L10	-0.8698	0.128	-6.788	0.000	-1.121	-0.619
ar.L11	-0.8684	0.077	-11.248	0.000	-1.020	-0.717
sigma2	7.973e+14	5.61e-17	1.42e+31	0.000	7.97e+14	7.97e+14

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	11.90
Prob(Q):	0.93	Prob(JB):	0.00
Heteroskedasticity (H):	1.74	Skew:	0.81
Prob(H) (two-sided):	0.22	Kurtosis:	4.49

Figure IX-46 – Modèle ARIMA pour le poste de consommation Optique

Le meilleur modèle pour notre jeu de données est un modèle $ARIMA(11,0,0)$. Nous pouvons remarquer que dans ce modèle, un paramètre sur deux des termes autorégressifs n'est pas pertinent. L'application d'un modèle SARIMA s'avère donc nécessaire. L'équation du modèle est la suivante :

$$\hat{y}_t = 1{,}29 * 10^7 - 0{,}97 * \hat{y}_{t-1} - 0{,}95 * \hat{y}_{t-2} - 0{,}89 * \hat{y}_{t-3} - 0{,}93 * \hat{y}_{t-4} - 0{,}87 * \hat{y}_{t-5} - 0{,}88 * \hat{y}_{t-6} - 0{,}9 * \hat{y}_{t-7} - 0{,}91 * \hat{y}_{t-8} - 0{,}86 * \hat{y}_{t-9} - 0{,}87 * \hat{y}_{t-10} - 0{,}87 * \hat{y}_{t-11} + \epsilon_t$$

Une analyse de nos résidus ne nous assure pas qu'il s'agit bien d'un bruit blanc. Nous ne pouvons pas conclure sur la normalité de nos résidus. L'analyse visuelle du graphique en haut à droite de la figure ci-dessous semble nous indiquer que nos résidus suivent bien une loi normale de moyenne $1,25 * 10^6$. Nous ne concluons donc pas avec certitude que nos résidus sont du bruit blanc.

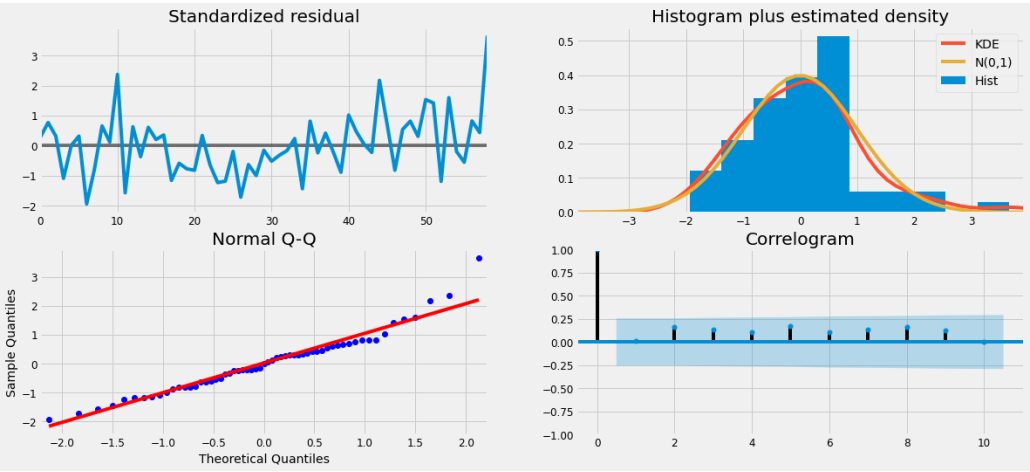


Figure IX-47 – Analyse des résidus du modèle SARIMA du poste de consommation Optique

b) *Le modèle SARIMA*

SARIMAX Results

Dep. Variable:	y	No. Observations:	59
Model:	SARIMAX(0, 0, 1)x(3, 1, [], 12)	Log Likelihood	-867.555
		AIC	1747.111
		BIC	1758.212
		HQIC	1751.288

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	2.836e+06	2.24e+06	1.268	0.205	-1.55e+06	7.22e+06
ma.L1	-0.7148	0.247	-2.896	0.004	-1.199	-0.231
ar.S.L12	-0.3339	0.166	-2.012	0.044	-0.659	-0.009
ar.S.L24	-0.3546	0.170	-2.086	0.037	-0.688	-0.021
ar.S.L36	-0.2374	0.118	-2.019	0.043	-0.468	-0.007
sigma2	8.701e+14	0.009	9.85e+16	0.000	8.7e+14	8.7e+14

Ljung-Box (L1) (Q):	3.37	Jarque-Bera (JB):	0.64
Prob(Q):	0.07	Prob(JB):	0.73
Heteroskedasticity (H):	1.38	Skew:	-0.17
Prob(H) (two-sided):	0.53	Kurtosis:	3.46

Figure IX-48 – Sortie du modèle SARIMA du poste de consommation Optique

Comme lors de l'application du modèle ARIMA pour les modèles précédemment étudiés, les termes autorégressifs tentent de capter une saisonnalité dans notre jeu de données. Nous appliquons donc un modèle $SARIMA(0,0,1)(3,1,0)_{12}$ à nos données. Son équation est la suivante :

$$\hat{y}_t = 2{,}84 * 10^6 + 0{,}67 * \hat{y}_{t-12} - 0{,}02 * \hat{y}_{t-24} + 0{,}09 * \hat{y}_{t-36} + 0{,}24 * \hat{y}_{t-48} + \epsilon_t - 0{,}71 * \epsilon_{t-1}$$

Les divers tests nous assurent que nos résidus sont bien du bruit blanc mais il sera nécessaire de compenser la moyenne de $5{,}35 * 10^6$ de nos résidus.

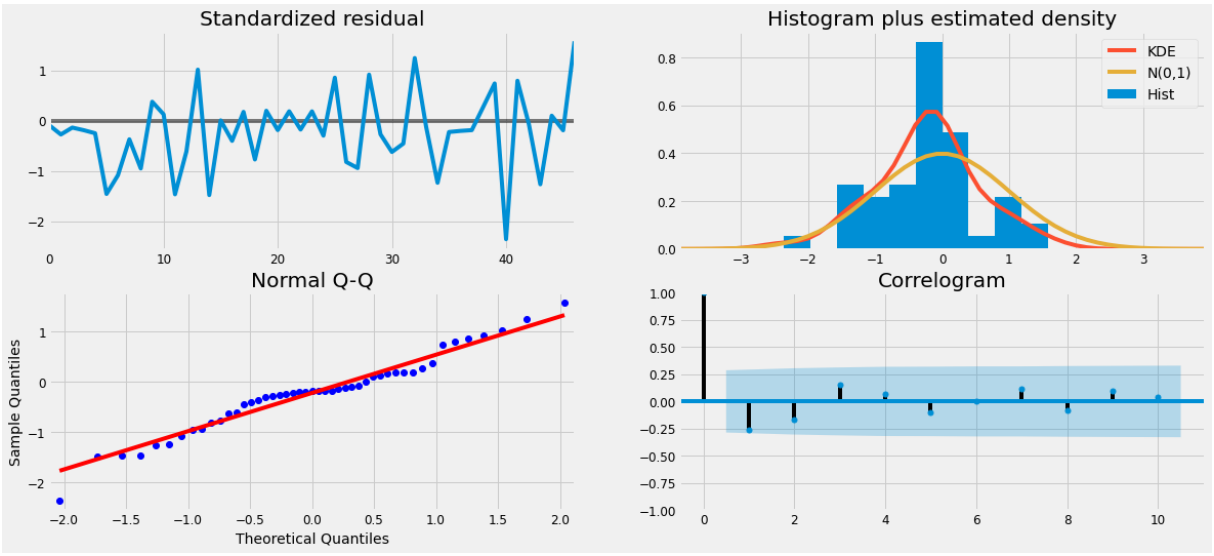


Figure IX-49 – Analyse des résidus du modèle SARIMA du poste de consommation Optique

4. Prothèses auditives

Le poste de consommation *Prothèses auditives* peut être décomposé de la façon suivante via des statistiques par mois et année de dépense :

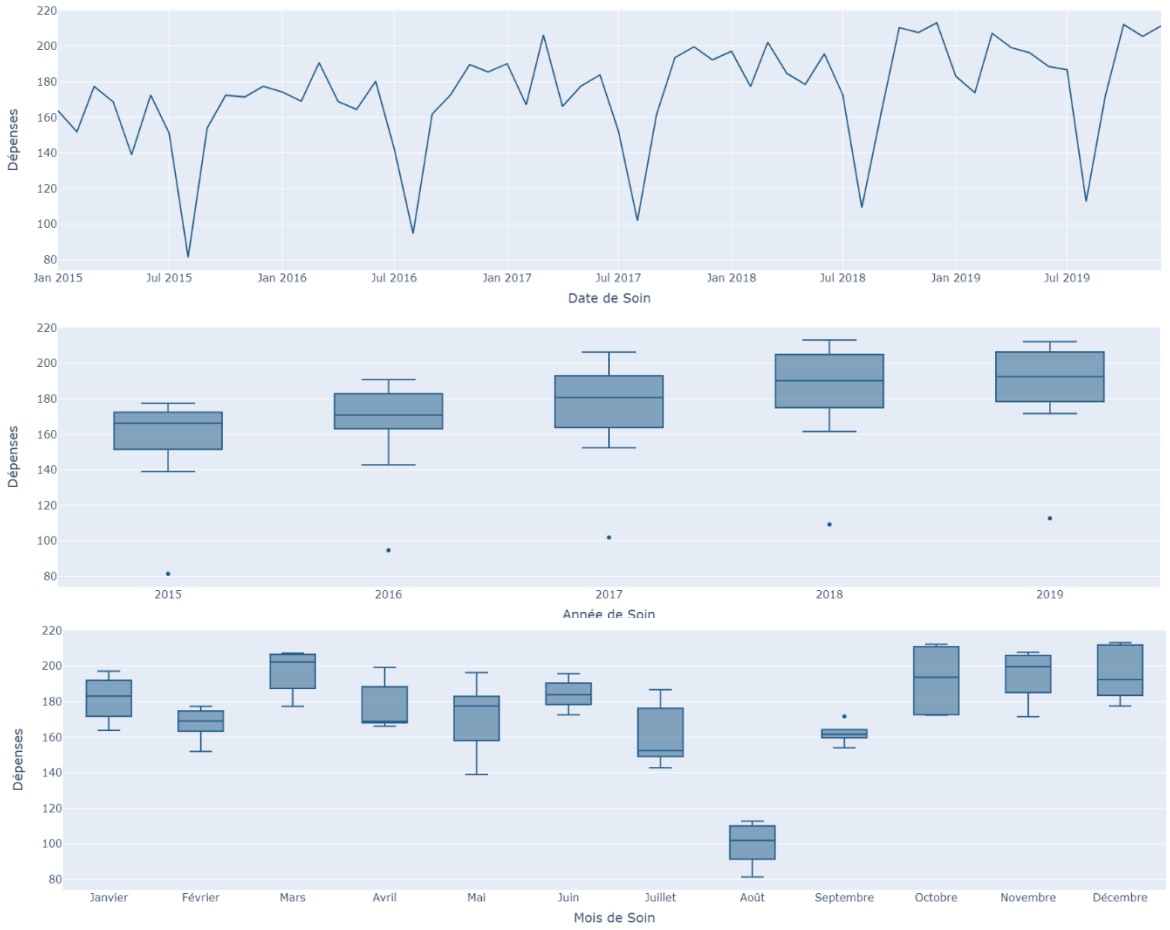


Figure IX-50 – Décomposition de la série temporelle du poste de consommation Prothèses auditives (en millions d'euros)

a) La stationnarité de notre série temporelle

Comme pour l'analyse des postes de consommation précédents, nous allons vérifier la stationnarité de notre série sans aucun pré-traitement et résumons les résultats obtenus dans le tableau ci-dessous.

	Test ADF	Test KPSS	Conclusion du test
P-value associée	0,53	0,022	Rejet de la stationnarité

Figure IX-51 – Test de stationnarité sur la série temporelle Prothèses auditives avant prétraitement

A l'aide de ces deux tests, nous pouvons en conclure que notre série n'est pas stationnaire, il est donc nécessaire d'effectuer une différenciation d'ordre 1 et d'effectuer à nouveau les tests ADF et KPSS sur ce nouveau jeu de données.

	Test ADF	Test KPSS	Conclusion du test
P-value associée	0,000	0,07	Série stationnaire

Figure IX-52 – Test de stationnarité sur la série temporelle Prothèses auditives après prétraitement

Au vu des résultats obtenus, nous pouvons en conclure que la série temporelle est bien stationnaire.

b) Le modèle ARIMA

SARIMAX Results

Dep. Variable:	y	No. Observations:	59
Model:	SARIMAX(11, 0, 1)	Log Likelihood	-1034.319
		AIC	2096.637
		BIC	2125.723
		HQIC	2107.991

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	7.351e+06	4.37e-08	1.68e+14	0.000	7.35e+06	7.35e+06
ar.L1	-0.8829	0.095	-9.333	0.000	-1.068	-0.697
ar.L2	-0.9285	0.072	-12.925	0.000	-1.069	-0.788
ar.L3	-0.9373	0.076	-12.299	0.000	-1.087	-0.788
ar.L4	-0.8931	0.089	-10.066	0.000	-1.067	-0.719
ar.L5	-0.9537	0.060	-15.872	0.000	-1.071	-0.836
ar.L6	-0.8715	0.092	-9.514	0.000	-1.051	-0.692
ar.L7	-0.9337	0.078	-11.942	0.000	-1.087	-0.780
ar.L8	-0.9350	0.077	-12.173	0.000	-1.086	-0.784
ar.L9	-0.8873	0.092	-9.680	0.000	-1.067	-0.708
ar.L10	-0.9540	0.059	-16.154	0.000	-1.070	-0.838
ar.L11	-0.8423	0.089	-9.455	0.000	-1.017	-0.668
ma.L1	-0.4391	0.199	-2.209	0.027	-0.829	-0.050
sigma2	6.846e+13	3.63e-16	1.89e+29	0.000	6.85e+13	6.85e+13

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	0.77
Prob(Q):	0.92	Prob(JB):	0.68
Heteroskedasticity (H):	1.25	Skew:	-0.19
Prob(H) (two-sided):	0.62	Kurtosis:	2.59

Figure IX-53 – Modèle ARIMA pour le poste de consommation Prothèses auditives

Le meilleur modèle pour notre jeu de données est un modèle $ARIMA(11,0,1)$. Nous pouvons remarquer que tous les termes sont pertinents. L'équation du modèle est la suivante :

$$\hat{y}_t = 7{,}35 * 10^7 - 0{,}88 * \hat{y}_{t-1} - 0{,}92 * \hat{y}_{t-2} - 0{,}94 * \hat{y}_{t-3} - 0{,}89 * \hat{y}_{t-4} - 0{,}95 * \hat{y}_{t-5} - 0{,}87 * \hat{y}_{t-6} - 0{,}93 * \hat{y}_{t-7} - 0{,}94 * \hat{y}_{t-8} - 0{,}89 * \hat{y}_{t-9} - 0{,}95 * \hat{y}_{t-10} - 0{,}84 * \hat{y}_{t-11} + \epsilon_t - 0{,}44 * \epsilon_{t-1}$$

Comme précédemment, nous admettons que nos résidus sont bien un bruit blanc de moyenne $-0{,}91 * 10^6$.

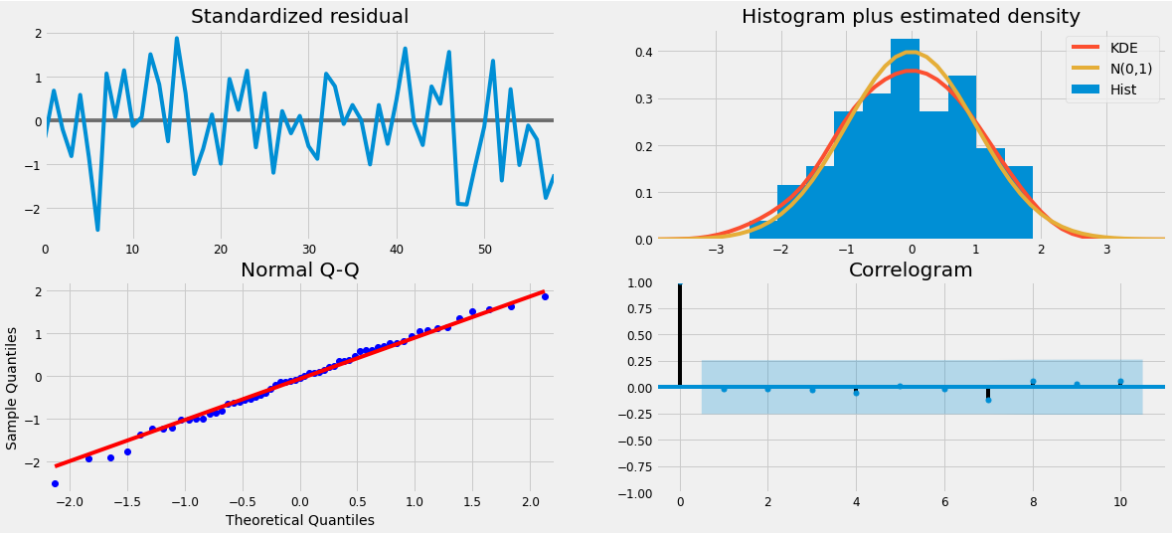


Figure IX-54 – Analyse des résidus du modèle ARIMA du poste de consommation Prothèses auditives

c) Le modèle SARIMA

SARIMAX Results

Dep. Variable:	y		No. Observations:	59	
Model:	SARIMAX(1, 0, 1)x(0, 1, [], 12)		Log Likelihood	-822.595	
Date:	Tue, 06 Jul 2021		AIC	1653.189	
			BIC	1660.590	
			HQIC	1655.974	

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	-6.308e+05	6.79e+05	-0.928	0.353	-1.96e+06	7.01e+05
ar.L1	-0.3772	0.147	-2.559	0.011	-0.666	-0.088
ma.L1	-0.6096	0.104	-5.871	0.000	-0.813	-0.406
sigma2	9.172e+13	0.001	1.27e+17	0.000	9.17e+13	9.17e+13

Ljung-Box (L1) (Q):	0.51	Jarque-Bera (JB):	1.47
Prob(Q):	0.48	Prob(JB):	0.48
Heteroskedasticity (H):	1.18	Skew:	-0.41
Prob(H) (two-sided):	0.74	Kurtosis:	2.75

Figure IX-55 – Sortie du modèle SARIMA du poste de consommation Prothèses auditives

Comme lors de l'application du modèle ARIMA pour les modèles précédents, les termes autorégressifs tentent de capter une saisonnalité dans notre jeu de données. Nous appliquons donc un modèle $SARIMA(1,0,1)(1,1,0)_{12}$ à nos données. Son équation est la suivante :

$$\hat{y}_t = -6{,}31 * 10^5 - 0{,}38 * \hat{y}_{t-1} + \hat{y}_{t-12} + 0{,}38 * \hat{y}_{t-13} + \epsilon_t - 0{,}61 * \epsilon_{t-1}$$

L'analyse des résidus du modèle nous confirme qu'il s'agit bien d'un bruit blanc mais une moyenne non nulle ne nous permet pas de conclure avec certitude. Nous admettons donc que nos résidus suivent un bruit blanc de moyenne $1{,}72 * 10^6$.

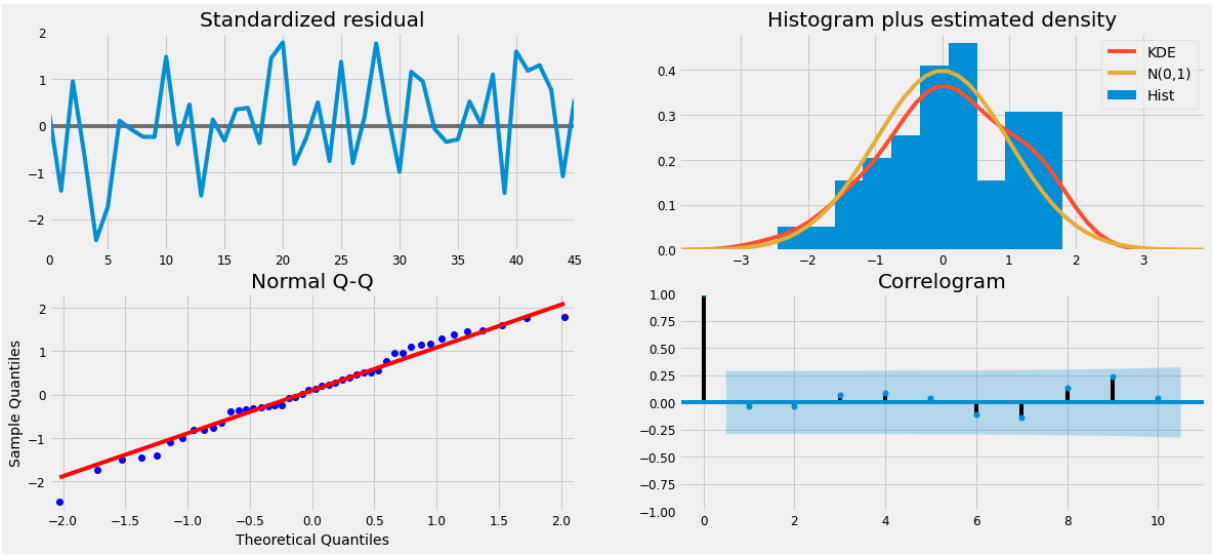


Figure IX-56 – Analyse des résidus du modèle SARIMA du poste de consommation Prothèses auditives

E. Annexe 5 : algorithme d'optimisation BFGS

L'algorithme d'optimisation BFGS appartient à une famille d'optimisation appelée méthodes Quasi-Newton. Il s'agit d'un algorithme de recherche local, destiné aux problèmes d'optimisation avec un seul optima. Afin d'optimiser la fonction a priori, le modèle fait appel à la dérivée du second ordre.

L'objectif de l'optimisation est de trouver les valeurs permettant de minimiser une fonction. Connaître comment la sortie de la fonction évolue en fonction des valeurs d'entrées nous indique dans quelle direction se déplacer afin d'améliorer la minimisation de notre fonction.

La dérivée d'une fonction $f(x)$ sur une seule variable x correspond à la vitesse de changement de la valeur de f en x. A partir de cette première approche, nous pouvons définir le gradient qui correspond à une généralisation de la dérivée mais appliqué à une fonction multivariée, c'est-à-dire possédant plusieurs variables d'entrées. Le gradient regroupe ainsi n dérivées partielles[32] avec n représentant le nombre de variables d'entrées. Nous pouvons ainsi définir le gradient comme étant le vecteur des dérivées premières de la fonction :

$$\nabla f(x) = \left[\frac{\partial f(x)}{\partial x_1}, \dots, \frac{\partial f(x)}{\partial x_n}\right]$$

A partir du gradient, nous pouvons définir la hessienne qui correspond à la dérivée du gradient, autrement dit, la dérivée au second ordre de la fonction d'entrée. La hessienne est ainsi la matrice des dérivées secondes de la fonction :

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_n \partial x_n} \end{bmatrix}$$

Alors que les méthodes d'optimisation faisant appel au gradient permettent de déterminer dans quelle direction se diriger afin d'optimiser la fonction, les méthodes d'optimisation du second ordre -comme la méthode de Newton – permettent également de connaître la taille du pas afin d'approcher un minimum local.

Cependant, la méthode de Newton présente le désavantage d'imposer le calcul de l'inverse de la Hessienne, ce qui peut vite s'avérer couteux en temps et en capacité de calcul. Les méthodes Quasi-Newton vont essayer de contourner ce problème en approximant l'inverse de la Hessienne à partir du gradient. Cependant, cette approche ne permet plus de connaître la taille du pas nécessaire afin d'approcher le minimum local. L'algorithme BFGS résout ce problème en utilisant une recherche linéaire dans la direction choisie afin de déterminer la distance à parcourir dans cette direction.

Comme la taille de la Hessienne est directement proportionnelle au nombre de variables du modèle, celle-ci peut rapidement devenir très importante. L'algorithme L-BFGS tient compte de ce facteur et ne stocke plus l'ensemble de l'approximation de la matrice inverse et considère plutôt une simplification de la matrice inverse de la précédente itération.

[32] La dérivée partielle d'une fonction correspond à la dérivée d'une variable sur cette même fonction en considérant que toutes les autres variables sont constantes.

F. Annexe 6 : prédictions du modèle Prophet par mois de soin

1. Soins de ville courants

Mois de soin	Observée	Prédiction SARIMA	Prédiction Prophet	Erreur SARIMA	Erreur Prophet	% erreur SARIMA	% erreur Prophet
Janvier	3 142	3 223	3 060	**- 81**	**82**	**- 2,50 %**	**2,68 %**
Février	2 857	2 503	2 884	**354**	**- 27**	**14,14 %**	**- 0,94 %**
Mars	3 089	3 161	3 111	**- 72**	**- 22**	**- 2,29 %**	**- 0,71 %**
Avril	2 610	2 768	2 841	**- 158**	**- 231**	**- 5,72 %**	**- 8,13 %**
Mai	2 634	2 912	2 923	**- 278**	**- 289**	**- 9,53 %**	**- 9,89 %**
Juin	2 886	2 874	2 972	**12**	**- 86**	**0,43 %**	**- 2,89 %**
Juillet	3 007	3 040	2 834	**- 33**	**173**	**- 1,08 %**	**6,10 %**
Août	2 687	2 605	2 727	**82**	**- 40**	**3,19 %**	**- 1,47 %**
Septembre	3 002	3 029	2 914	**- 27**	**88**	**- 0,90 %**	**3,02 %**
Octobre	3 388	3 309	2 996	**80**	**392**	**2,41 %**	**13,08 %**
Novembre	3101	2 689	3 000	**411**	**101**	**15,30 %**	**3,37 %**
Décembre	3 354	2 838	2 980	**515**	**374**	**18,16 %**	**12,55 %**
Total	**35 756**	**34 950**	**35 242**	**806**	**515**	**2,31 %**	**1,46 %**

Figure IX-57 – Résultat de la prédiction du modèle Prophet et SARIMA pour le poste de consommation Pharmacie apprécié par mois de soin (en millions d'euros)

2. Pharmacie

Mois de soin	Observée	Prédiction SARIMA	Prédiction Prophet	Erreur SARIMA	Erreur Prophet	% erreur SARIMA	% erreur Prophet
Janvier	4 837	4 867	4 648	**- 30**	**189**	**- 0,63 %**	**4,07 %**
Février	4 358	3 568	4 447	**790**	**- 89**	**22,14 %**	**- 2,00 %**
Mars	3 378	4 472	4 751	**- 1 094**	**- 1 373**	**- 24,47 %**	**- 28,90 %**
Avril	2 276	3 929	3 965	**- 1 655**	**- 1 691**	**- 42,13 %**	**- 42,65 %**
Mai	3 559	4 140	4 287	**- 581**	**- 728**	**- 14,04 %**	**- 16,98 %**
Juin	4 560	3 872	4 445	**727**	**155**	**18,78 %**	**3,49 %**
Juillet	4 412	4 067	3 834	**345**	**578**	**8,48 %**	**15,08 %**
Août	3 630	3 119	3 514	**512**	**116**	**16,40 %**	**3,3 %**
Septembre	5 019	5 332	4 435	**- 313**	**584**	**- 5,88 %**	**13,17 %**
Octobre	5 080	4 371	4 533	**709**	**547**	**16,22 %**	**12,07 %**
Novembre	5 009	3 781	4 556	**1 228**	**453**	**32,48 %**	**9,94 %**
Décembre	5 133	3 825	4 079	**1 308**	**1 054**	**34,19 %**	**25,84 %**
Total	**51 289**	**49 345**	**51 494**	**1 944**	**- 205**	**3,94 %**	**- 0,4 %**

Figure IX-58 – Résultat de la prédiction du modèle Prophet et SARIMA pour le poste de consommation Soins de ville courants apprécié par mois de soin (en millions d'euros)

3. Dentaire

Mois de soin	Observée	Prédiction SARIMA	Prédiction Prophet	Erreur SARIMA	Erreur Prophet	% erreur SARIMA	% erreur Prophet
Janvier	917	920	895	**- 2**	**22**	**- 0,23 %**	**2,46 %**
Février	956	960	953	**- 4**	**3**	**- 0,43 %**	**0,31 %**
Mars	611	1 189	1 109	**- 578**	**- 498**	**- 48,62 %**	**- 44,91 %**
Avril	88	694	1 053	**- 606**	**- 965**	**- 87,38 %**	**- 91,64 %**
Mai	675	1 030	1 124	**- 355**	**- 449**	**- 34,44 %**	**- 39,95 %**
Juin	1 203	968	1 149	**235**	**54**	**24,29 %**	**4,7 %**
Juillet	1 153	1 023	957	**130**	**196**	**12,67 %**	**20,48 %**
Août	553	348	526	**205**	**27**	**59,09 %**	**5,13 %**
Septembre	1 059	1 446	999	**- 388**	**60**	**- 26,89 %**	**6,01 %**
Octobre	1 095	1 088	1 049	**7**	**46**	**0,69 %**	**4,39 %**
Novembre	1 180	825	1 082	**355**	**98**	**43,01 %**	**9,06 %**
Décembre	1 251	933	1 007	**318**	**244**	**34,12 %**	**24,23 %**
Total	**10 741**	**11 423**	**11 903**	**- 682**	**- 1 162**	**- 5,97 %**	**- 9,76 %**

Figure IX-59 – Résultat de la prédiction du modèle Prophet et SARIMA pour le poste de consommation Dentaire apprécié par mois de soin (en millions d'euros)

4. Optique

Mois de soin	Observée	Prédiction SARIMA	Prédiction Prophet	Erreur SARIMA	Erreur Prophet	% erreur SARIMA	% erreur Prophet
Janvier	365	550	540	**- 185**	**- 175**	**- 33,64 %**	**- 32,41 %**
Février	583	772	594	**- 189**	**- 11**	**- 24,48 %**	**- 1,85 %**
Mars	335	776	641	**- 442**	**- 306**	**- 56,88 %**	**- 47,74 %**
Avril	38	663	552	**- 625**	**- 514**	**- 94,27 %**	**- 93,12 %**
Mai	319	690	520	**- 371**	**- 201**	**- 53,79 %**	**- 38,65 %**
Juin	659	724	603	**- 65**	**56**	**- 8,94 %**	**9,29 %**
Juillet	676	736	576	**- 60**	**100**	**- 8,16 %**	**17,36 %**
Août	504	583	474	**- 79**	**30**	**- 13,59 %**	**6,33 %**
Septembre	620	784	577	**- 164**	**43**	**- 20,89 %**	**7,45 %**
Octobre	701	771	588	**- 70**	**113**	**- 9,13 %**	**19,22 %**
Novembre	567	685	583	**- 119**	**- 16**	**- 17,30 %**	**- 2,74**
Décembre	796	871	709	**- 75**	**87**	**- 8,66 %**	**12,27 %**
Total	**6 162**	**8 606**	**6 957**	**- 2 444**	**-794**	**- 28,40 %**	**- 11,41 %**

Figure IX-60 – Résultat de la prédiction du modèle Prophet et SARIMA pour le poste de consommation Optique apprécié par mois de soin (en millions d'euros)

5. Prothèses auditives

Mois de soin	Observée	Prédiction SARIMA	Prédiction Prophet	Erreur SARIMA	Erreur Prophet	% erreur SARIMA	% erreur Prophet
Janvier	205	185	201	**19**	**4**	**10,26 %**	**1,99 %**
Février	198	202	194	**- 4**	**4**	**- 1,99 %**	**2,06 %**
Mars	135	246	221	**- 112**	**-86**	**- 45,29 %**	**- 38,91 %**
Avril	36	204	181	**- 169**	**-145**	**- 82,48 %**	**- 80,11 %**
Mai	133	210	187	**- 77**	**-54**	**- 36,75 %**	**- 28,88 %**
Juin	214	205	207	**9**	**7**	**4,51 %**	**3,38 %**
Juillet	229	211	170	**18**	**59**	**8,56 %**	**34,71 %**
Août	140	139	126	**1**	**14**	**0,77 %**	**11,11 %**
Septembre	211	272	186	**- 60**	**25**	**- 22,12 %**	**13,44 %**
Octobre	233	253	206	**- 20**	**27**	**- 8,08 %**	**13,11 %**
Novembre	230	206	217	**24**	**13**	**11,69 %**	**5,99 %**
Décembre	335	219	211	**117**	**124**	**53,32 %**	**58,77 %**
Total	**2 298**	**2 552**	**2 307**	**- 254**	**- 8**	**- 9,94 %**	**- 0,35 %**

Figure IX-61– Résultat de la prédiction du modèle Prophet et SARIMA pour le poste de consommation Prothèses auditives apprécié par mois de soin (en millions d'euros)