Mémoire présenté devant le jury de l'EURIA en vue de l'obtention du Diplôme d'Actuaire EURIA et de l'admission à l'Institut des Actuaires

Septembre 2022

Par : Beji Fida
Titre : Lancement d'une assurance voyage au Royaume-Uni : Tarification et prise en compte de la spécificité de ce nouveau marché.

Confidentialité : Oui - (Durée: 2 ans)

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

Membre présent du jury de l'Institut des Actuaires :
YANN MILOE
DAVY SENGDY
Signature :

Entreprise :
Europ Assistance
Signature :

Membres présents du jury de l'EURIA :
DANIEL BOIVIN

Directeur de mémoire en entreprise :
QUILFEN MATTHIEU
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 :

EURIA
EURo Institut d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

T +33 (0)2 98 01 66 55
euria@univ-brest.fr

Résumé

L'assurance voyage est un contrat ayant pour objectif de protéger les voyageurs contre les imprévus qui pourraient survenir avant et/ou au cours de leurs déplacements.

Une collaboration avec une compagnie aérienne britannique a permis à Europ Assistance de faire son entrée sur le marché de l'assurance voyage au Royaume-Uni.

Le principal défi pour Europ Assistance concerne alors le respect de la pratique de ce nouveau marché, au travers de l'alignement des processus de souscription et des méthodes de tarification.

Dans ce contexte, ce mémoire apporte une méthode de tarification d'un contrat d'assurance annulation de voyage conforme à la pratique du marché britannique. Cette étude repose sur deux types de données : des ressources internes à Europ Assistance et des données disponibles en *Open Data*. Cette combinaison permet de proposer un produit compétitif sur le marché britannique, en termes d'options offertes comme en termes de placement tarifaire.

Mots clefs: Assurance voyage, nouveau marché, tarification, série temporelle, ARIMA, ETS, GLM, *Open Data*.

Abstract

Travel insurance is a contract designed to protect travellers against unforeseen events that may occur before and/or during their travels.

A collaboration with a British airline has enabled Europ Assistance to enter the travel insurance market in the United Kingdom.

The main challenge for Europ Assistance was to comply with the practices of this new market, by aligning underwriting processes and pricing methods.

In this context, this thesis provides a pricing method for a travel cancellation insurance contract in line with UK market practice. This study is based on two types of data : internal resources at Europ Assistance and Open Data. This combination makes it possible to offer a competitive product on the British market, both in terms of the options offered and in terms of price placement.

Keywords: Travel insurance, new market, pricing, time series, ARIMA, ETS, GLM, Open Data.

Note de synthèse

L'assurance voyage est un contrat ayant pour objectif de protéger les voyageurs contre les imprévus qui pourraient survenir avant et/ou au cours de leur déplacement.
En général, elle offre une sécurité financière à l'assuré en cas d'annulation, de retard de vol, de perte de bagages ou bien d'autres dépenses, le but étant de minimiser les risques pendant le voyage.

Europ Assistance, un des leaders de l'assurance voyage, a récemment mis en place un partenariat avec une compagnie aérienne britannique. À travers cette collaboration, elle souhaite commercialiser une assurance voyage au Royaume-Uni, un marché où l'entreprise n'est que très peu présente.

Avant d'entrer sur un nouveau marché, l'enjeu est de faire une étude de ce dernier et d'analyser ces spécificités éventuelles. Cette analyse a permis de constater que le processus habituel de souscription et de tarification au Royaume-Uni est différent de celui auquel Europ Assistance fait généralement appel.

En effet, en Europe continentale et dans le cadre d'une annulation de voyage vendue par une compagnie aérienne, l'assureur s'engage généralement à indemniser uniquement le prix de déplacement ou du vol de l'assuré. Cette condition permet aux assureurs de limiter le coût des sinistres et de ne pas s'exposer à des risques plus difficiles à évaluer (les frais d'hôtels réservés, nombre d'excursions planifié...).

Toutefois, sur le marché britannique, la compagnie d'assurance voyage garantit à ses clients une indemnisation de tous les frais de voyage et d'hébergement pour lesquels le voyageur s'est engagé. Ainsi, l'assurance couvre le coût total du voyage (frais d'hôtel et parfois les réservations d'excursions...) et non seulement le montant du vol, sans pour autant demander le montant total du voyage lors de la souscription.

Dans un tel contexte, les assureurs au Royaume-Uni sont exposés à un risque considérable par rapport aux assureurs sur le marché européen. Afin de limiter le montant de remboursement, les compagnies d'assurance voyage ont mis en œuvre un plafond de couverture en proposant différentes formules de police : du simple au haut de gamme pour répondre aux besoins des différents profils de voyageurs. L'assuré choisira sa gamme selon le montant qu'il s'est engagé à payer pendant son séjour. En offrant différents types de couverture, l'assureur sera en mesure d'évaluer approximativement le montant qu'il assure.

La proposition de différentes formules caractérise le marché britannique. D'une manière générale, en Europe, la compagnie d'assurance voyage propose une seule formule avec un plafond de couverture identique pour tous les voyageurs (environ 3 000 €). De plus, les assureurs en Europe connaissent le montant exact du prix du vol assuré puisqu'il est demandé au moment de la souscription, ce qui n'est pas fait au Royaume-Uni.

L'absence d'informations sur le montant des voyages au marché du Britannique oblige Europ Assistance à adapter sa structure tarifaire et à opter pour la proposition de différents plafonds de remboursement, comme ses concurrents au Royaume-Uni.

Afin de mener une étude complète et précise, un périmètre a été définie pour ce mémoire. En effet, ce mémoire s'intéresse à la tarification d'une garantie annulation de voyage uniquement, pour les voyageurs britanniques qui se rendent en Europe et pour une durée maximale de 28 jours tout en garantissant le remboursement des frais d'hébergement et des billets d'avion.

Pour établir une tarification d'une garantie annulation de voyage adéquat, il faut disposer de tous les éléments nécessaires. Pour cela, nous avons décidé d'utiliser tant des données internes que de recueillir des données en *Open Data*.

Travail sur la base de données :

Pour la tarification de ce nouveau produit, deux bases de données seront utilisées :

- Base « **Frais d'hôtel** » construite à partir de l'*Open Data* ;
- Base **«Interne»** pour la modélisation du coût du vol et pour la fréquence

Base «Frais d'hôtel» : cette base de données provenant du site l'Union des Métiers et des Industries de l'Hôtellerie (UMIH), contient les prix moyens par nuit, par mois et par gamme d'hôtels en France entre 2012 et début 2020. Les années COVID ont été exclues parce que c'est une période atypique pendant laquelle les hôtels étaient fermés la plupart des temps et aucune donnée n'a été collectée par l'UMIH.

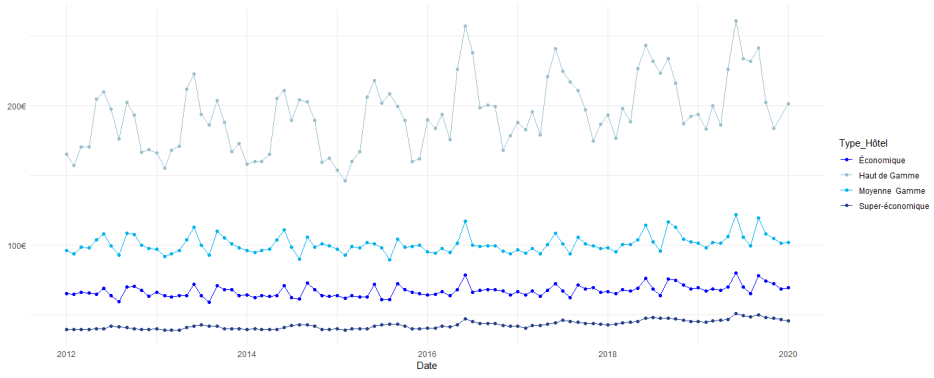


FIGURE 1 – Évolution des tarifs hôteliers en France

La figure 1 qui présente la base met en évidence une nette différence de prix entre les quatre gammes d'hôtels. Chaque gamme d'hôtels correspond à une couverture : par exemple, si un voyageur choisit une couverture économique, nous supposerons par défaut qu'il va prendre un hôtel économique.

Base « Interne » : depuis 2018, Europ Assistance collabore avec une compagnie aérienne, que l'on nommera S-Airlines, pour vendre en B to B to C des produits d'assurances voyages. La base de S-Airlines servira à modéliser le prix des billets (aller-retour) d'avion ainsi que la fréquence d'annulation, pour la compagnie aérienne B-Airlines ; futur partenaire d'Europ Assistance.

Dans un premier temps, les variables tarifaires ont fait l'objet d'une analyse univariée ainsi que bivariée. Cela a permis d'analyser la variation du coût des billets d'avion ainsi que la fréquence en fonction de ces différentes variables. Ensuite, étant donné qu'en ce moment l'inflation n'en finit pas d'augmenter, il a été décidé de rajouter un coefficient d'inflation. Les coefficients d'inflation pour les frais d'hôtel et pour le prix des billets d'avion sont respectivement obtenus à partir de l'indice de référence des loyers et l'indice de prix des services de transport aérien au Royaume-Uni.

Mise en place de la tarification :

Après avoir préparé et créé les bases de données, nous passerons à l'étape de tarification. L'approche de tarification correspond à l'approche collectif fréquence-coût.

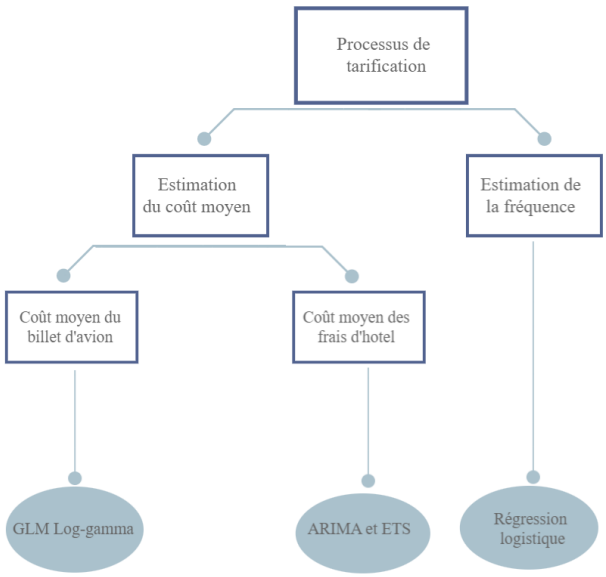


FIGURE 2 – Processus de tarification suivie.

Nous cherchons à estimer le coût moyen d'un voyage pour une police ainsi que la fréquence/probabilité pour une annulation. Chacun des coûts sera modélisé au moyen d'une méthode adaptée à la base de données associée. Ainsi, le coût de l'hôtel sera modélisé à l'aide des modèles de séries temporelles, quant au prix de l'avion, il sera modélisé par un modèle linéaire généralisé.

Concernant la fréquence, il faut noter qu'une police ne peut comporter plus d'un sinistre, puisque dès qu'un sinistre est déclenché, le contrat prend fin. Pour cette raison, nous souhaitons modéliser la probabilité qu'une police annule un voyage, puis utiliser cette probabilité comme une fréquence pour le modèle de la tarification. D'où le choix de la régression logistique.

Modélisations de frais des hôtels :

Cette étape consiste à estimer le coût moyen d'une nuit à l'hôtel par le biais d'un modèle ARIMA et d'un modèle ETS.

Afin d'obtenir le modèle ARIMA le plus adéquat pour chaque type d'hôtel, la fonction *auto.arima* a été utilisée.

	Super-économique	Économique	Moyenne gamme	Haut de gamme
RMSE base de test	0,763	1,274	3,095	7,622

Table 1 – Résultat du modèle ARIMA

D'après les résultats RMSE du tableau 1, nous constatons que le modèle ARIMA fait l'objet d'une robustesse. En effet, les RMSE obtenus varient entre 0,7 et 8 selon le gamme de l'hôtel, ce qui peut être considéré comme étant une performance acceptable vue que les frais des hôtels varient entre 39 € et 261 €. Malgré ces résultats, il est tout de même intéressant de comparer les modèles ARIMA avec les modèles ETS.

Pour les modèles ETS, un algorithme automatisé a également été utilisé. Les RMSE suivants sont obtenus :

	Super-économique	Economique	Moyenne gamme	Haut de gamme
RMSE base de test	1,391	1,372	2,769	7,673

Table 2 – Résultats du modèle ETS

Les RMSE obtenus montrent que le modèle ETS peut également être utilisé comme modèle pour prédire les frais d'hôtels. En effet, tout comme pour les modèles ARIMA, les RMSE varient entre 0,7 et 8 selon la gamme d'hôtel.

Afin de sélectionner les meilleurs modèles, Nous avons comparé les RMSE, MAE et ME de deux modèles. Une comparaison entre les frais des hôtels observés et les frais des hôtels prédites par le modèle ARIMA et le modèle ETS, à base de graphique a été également menée.

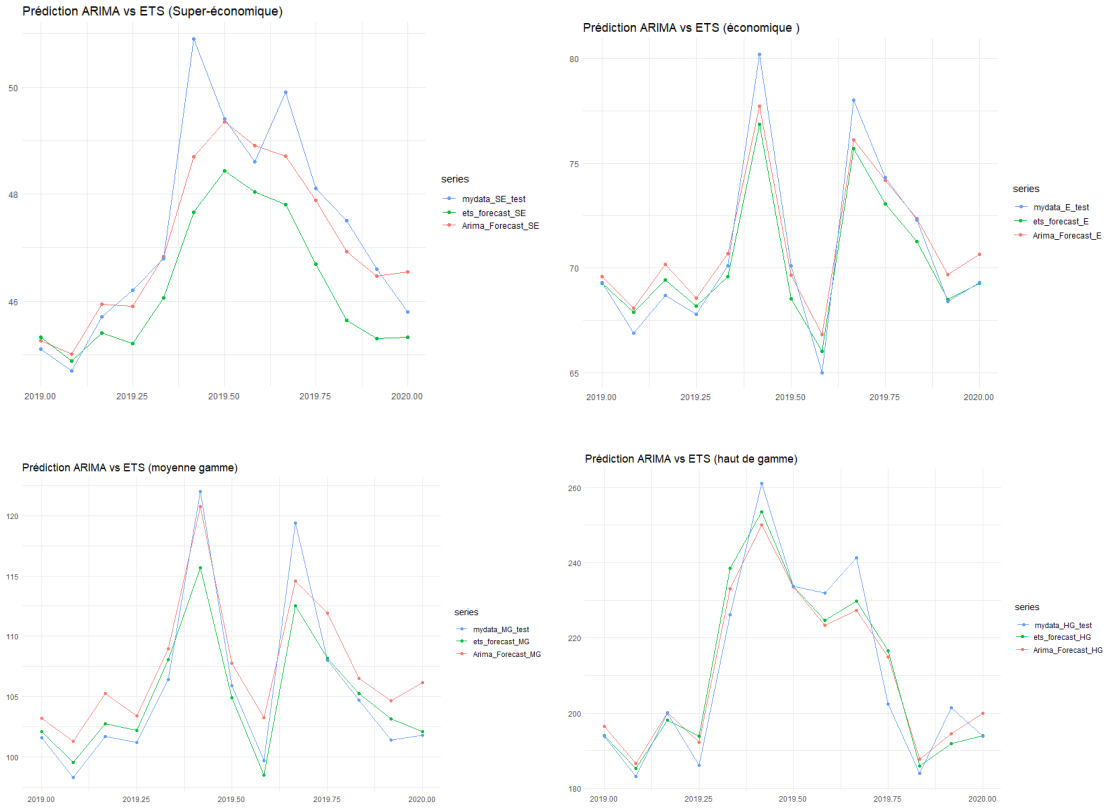


FIGURE 3 – Prédiction ARIMA vs ETS pour 2019 et début 2020

- Concernant les hôtels **super-économiques, économiques, haut de gamme**, les RMSE, MAE et ME obtenus à partir du modèle ARIMA sont bien inférieurs aux RMSE, MAE et ME obtenus par le modèle ETS. Le modèle ARIMA a ainsi réussi à bien mieux apprendre que les modèles ETS. Par ailleurs l'observation de la figure 3 montre que les estimations de l'ARIMA sont plus proches des prix observés que les estimations ETS.
- Contrairement aux résultats obtenus pour les autres catégories d'hôtels, le modèle ETS convient davantage à l'estimation des prix des hôtels **moyenne gamme** que le modèle ARIMA.

En conclusion, bien que les deux modèles soient robustes, les modèles ARIMA donnent un meilleur rendement dans la plupart des cas que le modèle ETS. Le modèle ARIMA sera donc par la suite utilisé pour estimer les frais d'hôtels.

Modélisations de coût des billets d'avion :

Cette étape, consiste à estimer le coût d'un billet aller-retour pour une seule personne. Pour réaliser la modélisation, nous avons comparé un GLM Log-gamma et un GLM Log-

normal. Le modèle qui a donné les meilleurs résultats et qui était ensuite retenu pour la prédiction est le modèle Log-gamma. Les tableaux 3 et 4 résument les résultats obtenus pour ce modèle :

	Base d'apprentissage			Base de validation		
	observé	prédiction	Erreur (%)	observé	prédiction	Erreur (%)
Coût moyen	279,413	279,5588	+0,0521 %	278,768	279,655	+3,0181 %

TABLE 3 – Résultats obtenus avec le GLM

	Base d'apprentissage	Base de validation
RMSE	162,176	161,263

TABLE 4 – RMSE

Globalement, les résultats des tests effectués sur la base de l'apprentissage ainsi que sur la base de validation montrent que le GLM obtenu peut être utilisé pour estimer les prix des billets d'avion.

Pour examiner un peu plus en détail la modélisation, nous avons comparé les prix observés ainsi que les prix prédits pour chacune des variables tarifaires, ci-dessous quelques exemples :

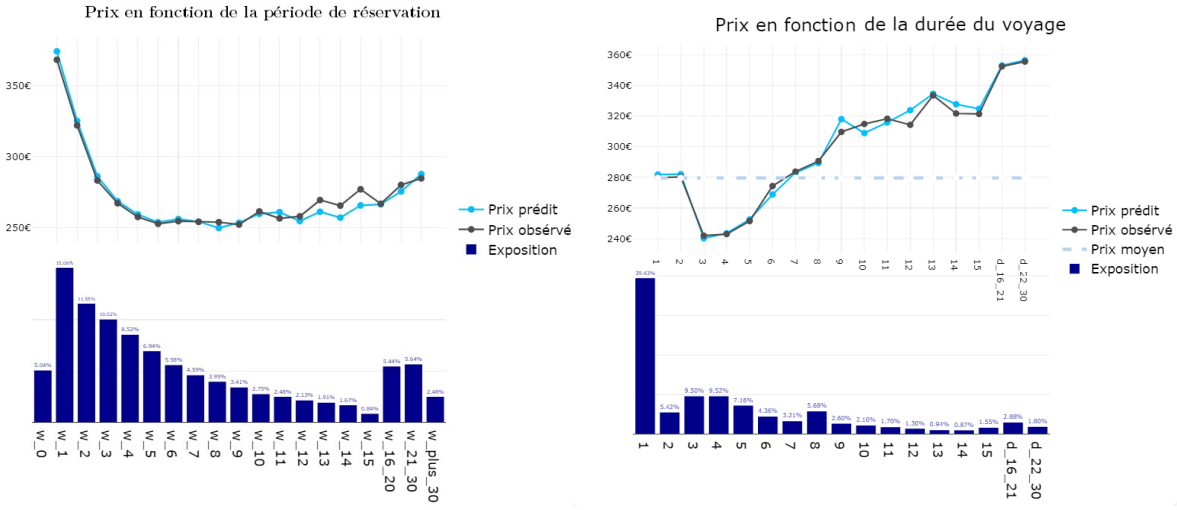


FIGURE 4 – Comparaison entre les valeurs observées et les valeurs prédites de la variable période de réservation et la variable durée du voyage.

La figure 4 montre que les coûts moyens prédit par le GLM Log-gamma suivent la tendance des coûts moyens observée sur la base de validation. Par ailleurs, les valeurs de la prédiction sont proches des valeurs observées sans pour autant remarquer un phénomène de surapprentissage.

En conclusion, pour calculer le coût moyen d'un voyage, deux modèles sont retenus : ARIMA pour le coût moyen des frais d'hôtel et GLM Log-gamma pour le coût du billet d'avion. La prochaine étape de tarification consiste à modéliser la fréquence d'annulation.

Modélisation de la fréquence d'annulation :

Cette dernière étape consiste à prédire la fréquence d'annulation à travers une régression logistique. Afin d'analyser la robustesse du modèle, il a été décidé de mesurer sa précision par le test AUC. L'AUC obtenu pour les deux bases est supérieure à 0,69 ce qui donne un modèle relativement acceptable.

	Base d'apprentissage	Base de validation
AUC	0,720	0,694

TABLE 5 – Test de l'AUC

Dans le même contexte d'étude de performance, une comparaison entre la fréquence prédite et la fréquence observée pour les diverses variables tarifaires, a été faite. La figure 5 montre quelques exemples :

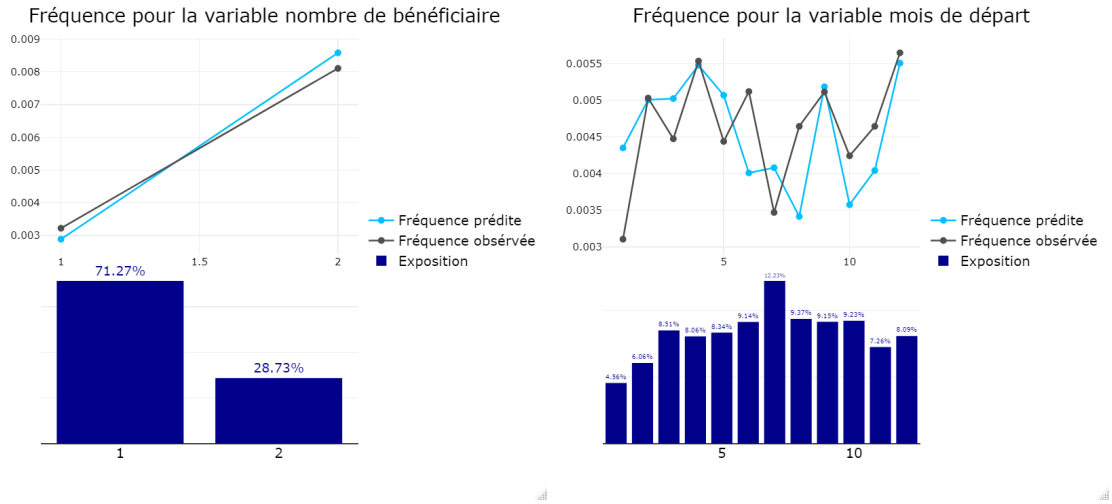


FIGURE 5 – Comparaison entre la fréquence observée et la fréquence prédite pour la variable nombre de bénéficiare et mois de départ.

Les fréquences prédites et les fréquences observées pour la variable nombre de bénéficiaires ne sont pas éloignées. Cependant, quelques différences sont remarquées pour le mois de départ. Dans l'ensemble, cependant, le modèle est un outil robuste qui peut être utilisé pour l'établissement des prix.

Application du modèle de tarification à certains profils de voyageurs :

Le processus de tarification s'achève par le calcul de la prime commerciale d'une

assurance voyage pour certains profils de voyageurs. Ensuite, cette prime sera comparée aux prix proposés par certains concurrents sur le marché de l'assurance voyage au Royaume-Uni.

Les produits d'assurance voyage proposés par la plupart des assureurs au Royaume-Uni combinent l'assistance et l'assurance annulation. Nous allons donc calculer la prime pure d'annulation obtenue par le biais des modèles déjà présentés, puis nous ajouterons la prime pure d'assistance calculée par l'équipe d'Europ Assistance pour le marché du Royaume-Uni. Enfin, la prime commerciale s'obtient en prenant en compte les commissions et les chargements tels que les frais d'acquisitions, les taxes des frais de gestion de sinistres, ... Les primes obtenues sont :

	Prime pure pour la garantie annulation			
	Plafond £1000	Plafond £1500	Plafond £2500	Hotel Plafond £5000
Profil_1 : 9 Septembre 2022 pour 14 nuits	£3,3	£4,5	£6,4	£11,8
Profil_2 : 9 Septembre 2022 pour 2 nuits	£1,1	£1,3	£1,5	£2,3
Profil_3 : 11 Février 2023 pour 14 nuits	£9,8	£13,2	£19,1	£35,1
Profil_4 : 11 Février 2023 pour 2 nuits	£2,5	£2,9	£3,6	£5,3

TABLE 6 – Prime pure annulation obtenue

Après une comparaison entre les primes d'assurance voyage au Royaume-Uni avec celles d'Europ Assistance (voir tableau 4.16), il a été constaté que les tarifs d'Europ Assitance disposent d'un bon positionnement. En effet, les primes d'Europ Assistance ne sont pas très élevées par rapport au marché britannique. Cependant, les prix incluent des taux de chargements, frais et de commissions qui varient d'un assureur et à un autre. En raison de ces variables, il est difficile d'évaluer la prime d'annulation pure obtenue à partir de nos modèles.

Conclusion :

La combinaison des données internes et de l'*Open Data*, nous a permis d'effectuer une tarification pour une garantie annulation de voyage pour un nouveau marché. Toutefois, il est important de noter que tout au long de cette tarification, nous avons pris des hypothèses et fait des choix qui peuvent être une source d'erreurs. Ceci s'intègre dans les difficultés qu'un actuaire peut rencontrer lors de la mise en place d'un nouveau tarif pour un nouveau marché.

Executive summary

Travel insurance is a contract designed to protect travellers against unforeseen events that may occur before and/or during their trip. In general, it offers financial security to the insured in case of cancellation, flight delay, loss of luggage or other expenses, with the aim of minimising risks during the trip.

Europ Assistance, one of the leaders in travel insurance, recently set up a partnership with a British airline. Through this collaboration, it will sell a travel insurance in the United Kingdom, a market in which the company has very little presence.

Before entering a new market, the challenge is to study it and analyse its possible specificities. This analysis has shown that the usual underwriting and pricing process in the UK is different from that which Europ Assistance generally uses.

Indeed, in continental Europe and in the context of a trip cancellation sold by an airline, the insurer generally undertakes to compensate only the price of the insured's travel or flight. This condition allows insurers to limit the cost of claims and to avoid exposing themselves to risks that are more difficult to assess (the cost of hotels booked, the number of excursions planned, etc.).

However, in the UK market, the travel insurance company guarantees its customers compensation for all travel and accommodation costs for which the traveller has booked. Thus, the insurance covers the total cost of the trip (hotel costs and sometimes excursion bookings, etc.) and not just the amount of the flight, without asking for the total amount of the trip when subscribing.

In such a context, insurers in the UK are exposed to a considerable risk compared to insurers in the European market. In order to limit the amount of reimbursement, travel insurance companies have implemented a ceiling of coverage by offering different policy formulas : from the simple to the high end to meet the needs of different travellers' profiles. The policyholder will choose his range according to the amount he has committed to pay during his stay. By offering different types of cover, the insurer will be able to estimate the amount it is insuring.

The offer of different packages is characteristic of the UK market. Generally speaking, in Europe, the travel insurance company offers a single formula with an identical cover limit for all travellers (approximately 3,000 €). In addition, insurers in Europe know the

exact price of the insured flight as it is requested at the time of subscription, which is not done in the UK.

The lack of information on the cost of travel in the UK market means that Europ Assistance has to adapt its pricing structure and opt to propose different reimbursement ceilings, like its competitors in the UK.

In order to carry out a complete and precise study, a perimeter has been defined for this report. This report focuses on the pricing of a trip cancellation policy only, for UK travellers to Europe, for a maximum of 28 days while guaranteeing reimbursement for accommodation and airfares.

In order to price an adequate travel cancellation insurance policy, you need to have all the necessary information. To do this, we decided to use both internal data and to collect data from Open Data.

Working on the database :

For the pricing of this new product, two databases will be used :

- "**Hotel fees**" database : built from the *Open Data* ;
- "**Internal**" database : for modelling the cost of the flight and for the frequency.

"Hotel fees" database : This database, which comes from the Union des Métiers et des Industries de l'Hôtellerie (UMIH) website, contains the average prices per night, per month and per hotel range in France between 2012 and the beginning of 2020. The COVID years have been excluded because this is an atypical period during which hotels were closed most of the time and no data was collected by the UMIH.

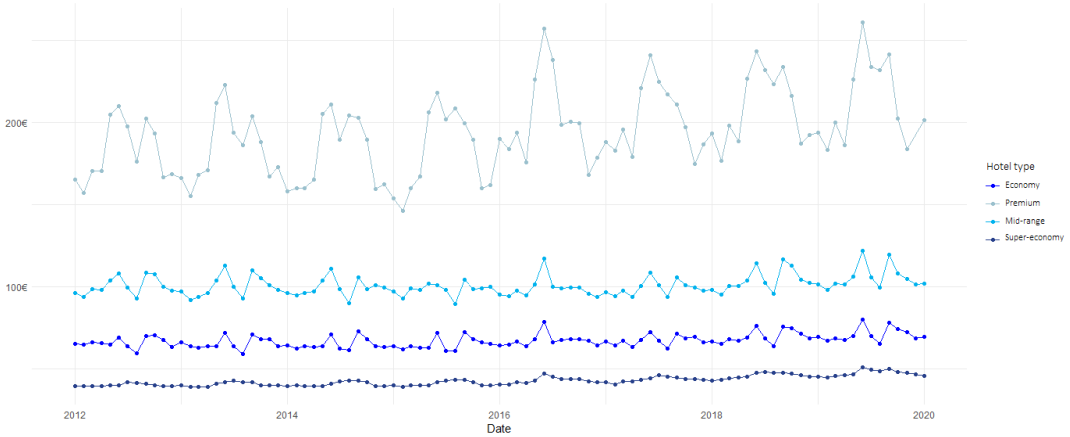


FIGURE 6 – Evolution of hotel prices in France

The figure 6 shows a clear difference in price between the four hotel ranges. Each hotel range corresponds to a coverage : for example, if a traveller chooses an economy coverage, we will assume by default that he will take an economy hotel.

"Internal" database Since 2018, Europ Assistance has been working with an airline company, which we will call S-Airlines, to sell travel insurance products on a B to B to C basis. The S-Airlines database will be used to model the price of (return) air tickets and the frequency of cancellations for the airline company B-Airlines, Europ Assistance's future partner.

Initially, the pricing variables were subjected to univariate and bivariate analysis. This made it possible to analyse the variation in the cost of air tickets as well as the frequency as a function of these different variables. Secondly, given that inflation is currently on the rise, it was decided to add an inflation coefficient. The inflation coefficients for hotel costs and airfares are derived from the Rent Reference Index and the UK Air Transport Services Price Index respectively.

Pricing process :

After preparing and creating the databases, we will move on to the pricing stage. The pricing approach corresponds to the collective frequency-cost approach.

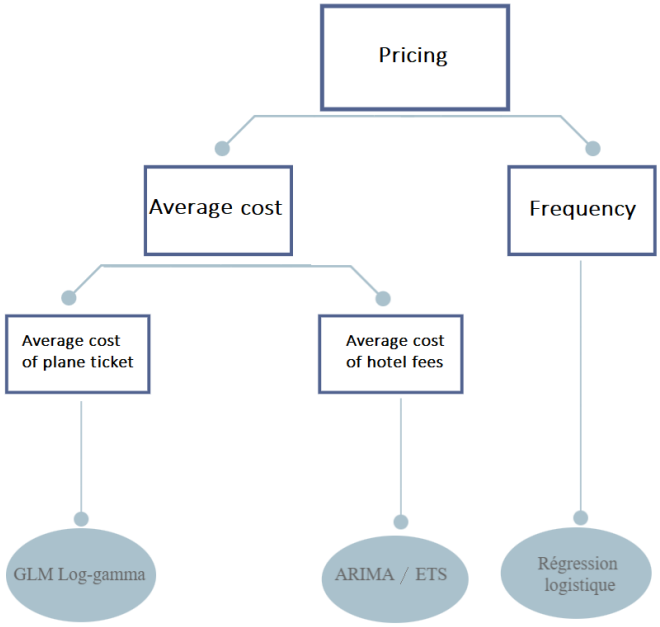


FIGURE 7 – Pricing process.

We seek to estimate the average cost of a trip for a policy and the frequency/probability of a cancellation. Each of the costs will be modelled using a method appropriate to the

associated database. Thus, the cost of the hotel will be modelled using time series models, while the price of the plane will be modelled using a generalized linear model.

Concerning the frequency, it should be noted that a policy cannot have more than one claim, since as soon as a claim is triggered, the contract ends. For this reason, we wish to model the probability of a policy cancelling a trip and then use this probability as a frequency for the pricing model. Hence the choice of logistic regression.

Hotel fees modelling :

This step consists of estimating the average cost of a night in a hotel through an ARIMA model and an ETS model.

In order to obtain the most appropriate ARIMA model for each type of hotel, the function *auto.arima* was used.

	Super-economy	Economy	Mid-range	Premium
RMSE of test database	0,763	1,274	3,095	7,622

TABLE 7 – ARIMA's results

According to the RMSE results of the table 7, we note that the ARIMA model is robust. Indeed, the RMSEs obtained vary between 0.7 and 8 depending on the range of the hotel, which can be considered as an acceptable performance given that the costs of the hotels vary between 39 € and 261 €. Despite these results, it is still interesting to compare the ARIMA models with the ETS models.

For the ETS models, an automated algorithm was also used. The following RMSEs are obtained :

	Super-economy	Economy	Mid-range	Premium
RMSE of test database	1,391	1,372	2,769	7,673

TABLE 8 – ETS's results

The RMSEs obtained show that the ETS model can also be used as a model for predicting hotel costs. Indeed, as for the ARIMA models, the RMSEs vary between 0,7 and 8 depending on the hotel range.

In order to select the best models, we compared the RMSE, MAE and ME of two models. A comparison between the observed hotel charges and the hotel charges predicted by the ARIMA model and the ETS model, based on graphs, was also conducted.

- For hotels, the RMSE, MAE and ME obtained from the ARIMA model are much lower than the RMSE, MAE and ME obtained from the ETS model. The ARIMA model has thus managed to learn much better than the ETS models. Moreover, the observation in figure 8 shows that the ARIMA estimates are closer to the observed prices than the ETS estimates.

FIGURE 8 – ARIMA's vs ETS's prediction

➢ Contrary to the results obtained for the other hotel categories, the ETS model is more suitable for estimating prices for hotels than the ARIMA model.

In conclusion, although both models are robust, the ARIMA models perform better in most cases than the ETS model. The ARIMA model will therefore be used in the following to estimate hotel costs.

Air ticket cost modelling :

This step consists of estimating the cost of a round trip ticket for one person. To perform the modelling, we compared a GLM Log-gamma and a GLM Log-normal. The model that gave the best results and was then retained for prediction was the Log-gamma model. The tables 9 and 10 summarise the results obtained for this model :

	Learning database			Validation database		
	Reality	Prediction	Error (%)	Reality	Prediction	Error (%)
Coût moyen	279,413	279,5588	+0,0521 %	278,768	279,655	+3,0181 %

TABLE 9 – GLM's results

	Learning database	Validation database
RMSE	162,176	161,263

TABLE 10 – RMSE

Overall, the results of the tests carried out on the learning basis as well as on the validation basis show that the GLM obtained can be used to estimate airfares.

To examine the modelling in a little more detail, we compared the observed prices with the predicted prices for each of the fare variables :

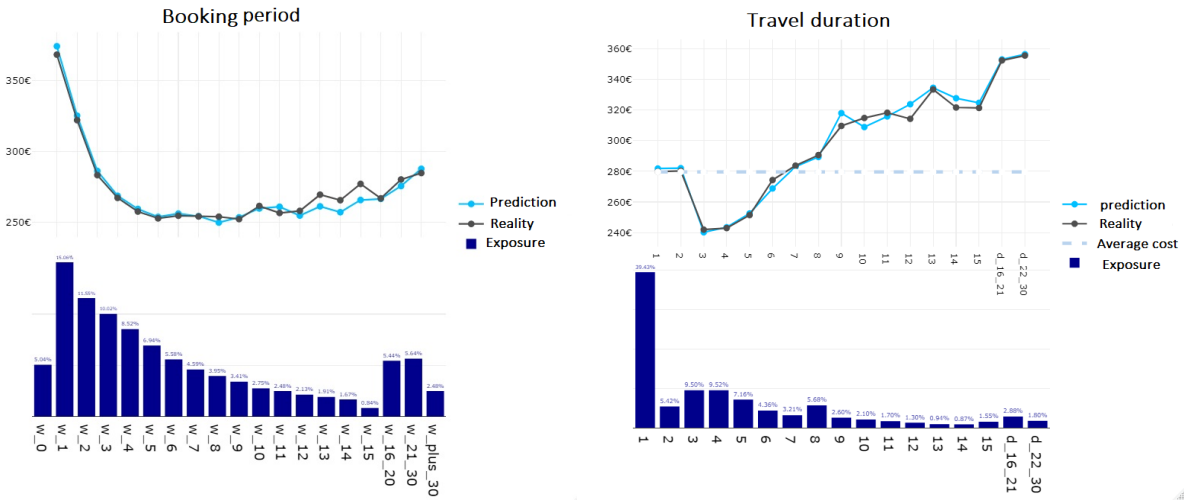


FIGURE 9 – A comparison between the observed and predicted values of the booking period and travel time variables.

Figure 9 shows that the average costs predicted by the GLM Log-gamma follow the trend of the average costs observed on the validation base. Moreover, the values of the prediction are close to the observed values without noticing an overfitting.

Cancellation frequency modelling :

This last step consists in predicting the frequency of cancellation through a logistic regression. In order to analyse the robustness of the model, it was decided to measure its accuracy by the AUC test. The AUC obtained for the two bases is higher than 0.69, which gives a relatively acceptable model.

	Learning database	Learning validation
AUC	0,720	0,694

TABLE 11 – AUC

In the same context of performance study, a comparison between the predicted and observed frequency for the various tariff variables was made. Figure 10 shows some examples :

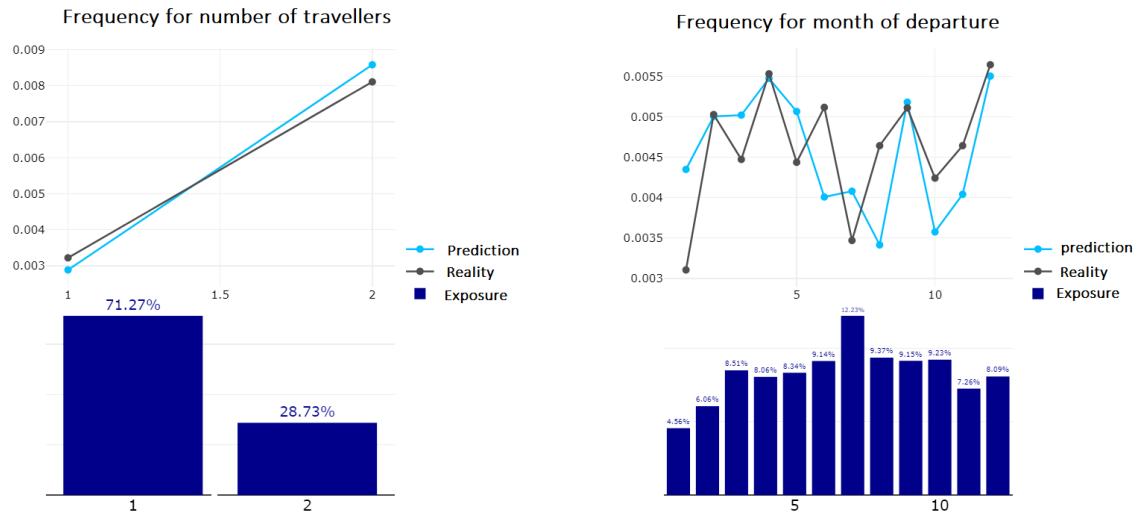


FIGURE 10 – Comparison between observed and predicted frequency for the variable number of beneficiaries and month of departure.

The predicted and observed frequencies for the variable number of travellers are not far apart. However, some differences are noted for the starting month. Overall, the model is a robust tool that can be used for pricing.

Application of the pricing model to certain traveller profiles :

The pricing process is completed by calculating the commercial travel insurance premium for certain traveller profiles. This premium will then be compared to the prices offered by some of the competitors in the UK travel insurance market.

The travel insurance products offered by most insurers in the UK combine assistance and cancellation insurance. We will therefore calculate the pure cancellation premium obtained through the models already presented and then add the pure assistance premium calculated by the Europ Assistance team for the UK market. Finally, the commercial premium is obtained by taking into account commissions and charges such as acquisition costs, taxes, claims management costs, etc. The premiums obtained are :

	Pure premium for cancellation cover			
	Limit £1000	Limit £1500	Limit £2500	Hotel Plafond £5000
Profile_1 : 9 September 2022 for 14 nights	£3,3	£4,5	£6,4	£11,8
Profile_2 : 9 September 2022 for 2 nights	£1,1	£1,3	£1,5	£2,3
Profile_3 : 11 February 2023 for 14 nights	£9,8	£13,2	£19,1	£35,1
Profile_4 : 11 February 2023 for 2 nights	£2,5	£2,9	£3,6	£5,3

TABLE 12 – Pure cancellation premium

After a comparison of travel insurance premiums in the United Kingdom with those of Europ Assistance (see table Europ Assistance (see table below), it was found that Europ Assistance's rates are well positioned. Indeed, Europ Assistance's premiums are not very high compared to the British market. However, the prices include rates for charges, fees and commissions which vary from one insurer to another. Because of these variables, it is difficult to assess the pure cancellation premium obtained from our models.

Conclusion :

The combination of in-house data and open data has allowed us to perform pricing for a travel cancellation benefit for a new market. However, it is important to note that throughout this pricing exercise, we made assumptions and choices that can be a source of error. This is part of the difficulties that an actuary may face when setting up a new tariff for a new market.

Remerciements

Tout d'abord, je remercie Etienne BONNET, directeur assurance d'Europ Assistance Groupe, ainsi qu'Olivier DANNEAUX, directeur souscription et tarification, de m'avoir donné l'opportunité d'intégrer Europ Assistance.

Je souhaite remercier l'ensemble de l'équipe Assurance d'Europ Assistance Groupe pour leur accueil chaleureux, leur soutien pendant les périodes les plus délicates et leur bonne humeur. Grâce à vous, j'ai pu passer une année d'alternance dans les meilleurs des conditions.

Je tiens à remercier plus particulièrement mon tuteur d'entreprise Matthieu QUILFEN. Sa disponibilité, ses conseils et son écoute ont été précieux pour mener à bien ce mémoire.

Je voudrais également témoigner ma reconnaissance à toute l'équipe pédagogique de l'EURIA et tout particulièrement à Alexis MERX, mon tuteur académique, pour son suivi, ses expertises et ses recommandations à l'origine de l'approfondissement de ce travail.

Je tiens à remercier très chaleureusement Sophie NAVARRO, Enzo SARTI et Flavien VALERI actuaires externes à Europ Assistance pour leur relecture attentive et leurs remarques pertinentes. J'aimerais aussi les remercier pour leur soutien et encouragement tout au long de ce mémoire. Toute ma gratitude également à Amin FARHAT qui a beaucoup contribué à ce travail grâce à ses nombreux encouragements.

Finalement, je remercie tous ceux qui ont pu contribuer d'une façon ou d'une autre à la réalisation de ce mémoire.

Table des matières

Introduction

Europ Assistance, un des leaders de l'assurance voyage, a récemment mis en place un partenariat avec une compagnie aérienne britannique. À travers cette collaboration, elle souhaite commercialiser une assurance voyage au Royaume-Uni, un marché au sein duquel l'entreprise n'est que très peu présente.

Avant d'entrer sur un nouveau marché, l'enjeu est de faire une étude sur ce dernier et d'analyser les différences éventuelles. Cette analyse a permis de constater que le processus habituel de souscription et de tarification au Royaume-Uni est différent de celui auquel Europ Assistance fait généralement appel.

En effet, en Europe, dans le cadre d'une assurance annulation voyage, l'assureur couvre uniquement le coût du vol, tandis qu'au Royaume-Uni, c'est le coût total du voyage qui est couvert, c'est à dire les prix des billets d'avions, mais également les frais d'hébergement, de location de voiture ou d'activités prévues sur place.
Par ailleurs, la structure tarifaire actuelle d'Europ Assistance n'est pas alignée avec le marché britannique, elle est plutôt adaptée à un partenariat européen. En effet, si en Europe, il est convenu qu'au moment de la souscription l'assuré indique le prix de son vol, ce n'est pas le cas au Royaume-Uni (que ce soit pour le prix du vol ou tout autre coût associé aux voyages). La tarification actuelle d'Europe Assistance, qui dépend aujourd'hui de la variable prix du billet d'avion, n'est donc pas transposable à la pratique du marché britannique.
Tous ces éléments conduisent Europ Assistance à changer sa structure tarifaire pour s'aligner à ce nouveau marché tout en proposant des tarifs avantageux afin d'attirer une nouvelle clientèle.
L'objectif de ce mémoire est d'estimer le montant des engagements qu'Europ Assistance prend envers ses assurés et de proposer ensuite des tarifs compétitifs sur le marché.

Les données Open Data permettront d'estimer le coût du voyage (frais d'hôtel principalement) en utilisant des modèles de séries temporelles qui sont les modèles ARIMA et ETS. Les données internes serviront à prédire la fréquence d'annulation du voyage du portefeuille via un modèle linéaire généralisé classique.

Une fois que le coût du voyage et la fréquence seront connus, la prime pure sera calculée par une méthode traditionnelle de tarification non-vie : fréquence x coût moyen. Nous allons enfin appliquer nos modèles à 4 profils de voyageurs différents et comparer les primes obtenues aux primes proposées par certains assureurs au Royaume-Uni.

Chapitre 1

Contexte

Ce chapitre commence par une brève introduction à l'assurance voyage, notamment ses garanties et ses canaux de distribution. La deuxième partie portera sur le marché de l'assurance voyage au Royaume-Uni : un état des lieux général accompagné d'une étude comparative de marché. Enfin, une section sur l'Open Data sera présentée pour comprendre l'idée sous-jacente à l'utilisation de ces données.

1.1 Introduction à l'assurance voyage

L'assurance voyage est un contrat ayant pour objectif de protéger les voyageurs contre les imprévus qui pourraient survenir avant et/ou au cours de leurs déplacements. En général, elle offre une sécurité financière à l'assuré en cas d'annulation, de retards de vol, de perte de bagages ou bien d'autres dépenses, le but étant de minimiser les risques pendant le voyage.

Une police d'assurance voyage peut proposer une seule garantie et par conséquent ne couvrir qu'un seul risque (par exemple, l'annulation), mais le plus souvent elle propose une large couverture en offrant différentes garanties allant de l'assistance médicale au risque traditionnel.

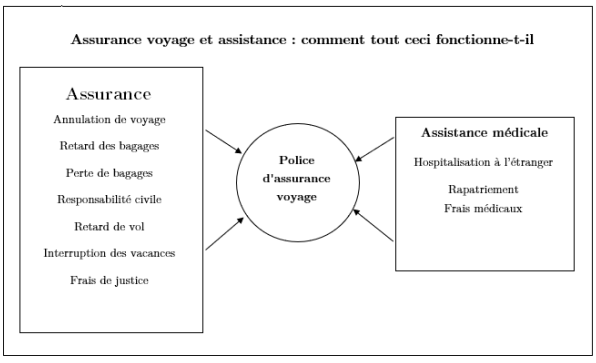


FIGURE 1.1 – Schéma des garanties qu'une assurance voyage peut proposer

1.1.1 Garanties et Contrats

Les contrats d'assurance voyage diffèrent d'un assureur à un autre et ne sont pas tous identiques. Toutefois, il existe certaines garanties clés présentes dans la plupart des contrats. Généralement, un contrat d'assurance voyage est composé de deux catégories : l'assurance (annulation, pertes de bagages, etc.) et l'assistance pour différencier les remboursements des interventions.

En ce qui concerne les garanties assurances, l'assurance annulation, qui fait l'objet de ce mémoire, est la garantie clé. L'objet de la garantie est de rembourser à l'assuré les frais qu'il a engagé directement dus à l'annulation du voyage couvert.
Plusieurs événements inclus dans la police d'assurance rendent l'assuré admissible à la protection contre l'annulation, y compris :

- Blessure, maladie ou bien décès de l'assuré ou d'un proche à lui
- Dommage matériel à la maison de l'assuré

L'assurance contre le retard ou la perte de bagages et la responsabilité civile à l'étranger font également partie des garanties assurances.

Pour ce qui est de l'assistance médicale, les frais médicaux et l'hospitalisation à l'étranger sont les garanties essentielles. En Assistance médicale, le voyageur est assuré à compter de la date de son départ en voyage jusqu'à la date de fin.

En règle générale, une police d'assurance voyage ne couvre qu'un seul voyage. Toutefois, pour répondre aux besoins des différents voyageurs, les compagnies d'assurance proposent également des contrats annuels et longs séjours.

En ce qui concerne les contrats annuels, ils s'adressent aux voyageurs fréquents qui voyagent plus d'une fois par année : ils couvrent tous les voyages au cours de cette année, jusqu'à 30 jours par voyage.

Quant aux contrats longs séjour, ils protègent les voyageurs lors d'un voyage de longue durée (plus d'un mois). Ce type d'assurance s'adresse principalement aux étudiants qui partent étudier à l'étranger ou aux personnes qui souhaitent effectuer un long séjour à l'étranger.

Dans ce mémoire, nous nous focalisons sur l'assurance annulation pour un seul contrat de voyage. En effet, il s'agit du produit qu'Europ Assitance souhaite proposer au marché du Royaume-Uni dans le cadre de sa collaboration avec B-Airlines.

1.1.2 Canaux de distribution

- ***Business to Business to Client*** **(B to B to C)**
  La plupart des produits d'assurance voyage sont vendus à travers le BtoBtoC. La compagnie d'assurance voyage conclut un contrat avec un partenaire (compagnies aériennes, agences de voyages, etc.) et ensuite ce dernier propose l'assurance à ses clients en contrepartie d'une commission sur les polices achetées.

- ***Business to Business*** **(B to B)**
  En B to B, l'assureur vend l'assurance voyage au partenaire dans le but que ce dernier offre gratuitement la couverture à tous ses clients.

- ***Business to Client*** **(B to C)**
  Les compagnies d'assurance proposent directement le produit au client, sans intermédiaire.

Le contrat d'assurance annulation qu'Europ Assistance souhaite lancer sur le marché britannique s'inscrit dans le cadre de l'assurance B to B to C. Le partenaire B-Airlines proposera une assurance voyage Europ Assistance à ceux qui voyagent avec leur compagnie aérienne et veulent acheter une assurance annulation.

Dynamique du marché de l'assurance voyage

D'après une étude menée par *Next Move Strategy Consulting*, le chiffre d'affaire de l'assurance voyage dans le monde devrait être environquatre fois supérieur en 2030 à ce qu'il était en 2019. S'élevant à près de $15,5$ milliards d'euros en 2020, le marché devrait atteindre environ 57 milliards d'euros en 2030.
En ce qui concerne le marché européen, il est prévu qu'il atteigne 10 milliards d'euros en 2027. Le marché britannique reste de loin le plus grand marché en Europe. C'est pourquoi Europ Assitance pense que la collaboration avec une compagnie britannique est une excellente opportunité pour y prendre part.
Le tableau 1.2 présente les pays ayant la plus grosse part de marché en assurance voyage en Europe en 2018.

Rang	Pays	Part du marché
1	Royaume-Uni	23,1 %
2	Allemagne	12,9 %
3	France	9,3 %
4	Norvège	9,1 %
5	Pays-Bas	7,2 %

TABLE 1.1 – Top 5 des marchés de l'assurance voyage en Europe.

1.2 État des lieux sur le marché de l'assurance voyage au Royaume-Uni

Le marché de l'assurance voyage au Royaume-Uni s'est développé entre 2014 et 2018. Cette croissance se manifeste par une hausse annuelle d'en moyenne $0,8\%$ du nombre de polices d'assurance vendues. En effet, les compagnies d'assurance voyage ont vendu $5,9$ millions assurances en 2018 contre $5,7$ millions en 2014.

Cette légère hausse résulte essentiellement de la croissance du tourisme des Britanniques. Le nombre de voyages est passé de $60,1$ millions en 2014 à $71,1$ millions en 2018, ce qui correspond à un taux de croissance annuel de $4,3\%$.
Par ailleurs, le marché britannique de l'assurance voyage est considéré comme le plus développé d'Europe. La figure 1.2 ci-dessous présente l'évolution des primes souscrites en assurance voyage en Allemagne et au Royaume-Uni entre 2014 et 2018. L'Allemagne constitue le deuxième marché le plus important.

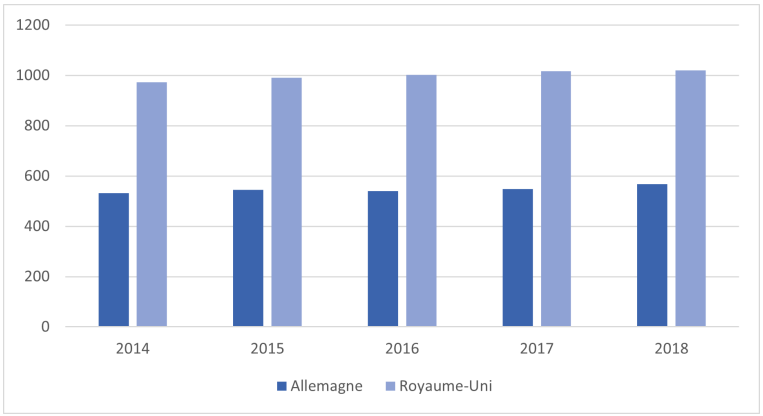


FIGURE 1.2 – Comparaison de l'évolution de la prime brute totale souscrite (en million d'euros) entre le Royaume-Uni et l'Allemagne

Les primes totales souscrites en Allemagne sont nettement inférieures à celle du Royaume-Uni et c'est en raison de divers facteurs :

- Au Royaume-Uni, la population a une fréquence relativement élevée de voyage à l'étranger par rapport aux autres populations européennes.
- Un taux relativement élevé des résidents du Royaume-Uni opte pour une assurance voyage lors de leurs déplacements.
- Le prix moyen des assurances voyages au Royaume-Uni est plus important que les autres marchés.

Afin de mieux connaître le marché de l'assurance voyage britannique, regardons ses acteurs ainsi que ses différences avec le marché européen.

1.2.1 Acteurs du marché britannique

Le graphique 1.3 montre l'évaluation de la part probable du marché de l'assurance voyage détenue par six acteurs, étude menée par *Next Move Strategy Consulting*. Ces estimations ont été basées sur une analyse combinée de l'importance de chaque concurrent au sein de chaque principal canal de distribution.

Au Royaume-Uni, trois compagnies détiennent plus de 10% du marché. AXA détient la part la plus importante des primes qui varie entre $20,5\%$ et $23,5\%$ en 2018. Elle a bénéficié de ses différents partenaires sur le marché pour vendre ses polices, en plus de ses ventes en B to C sur son site internet. Ensuite, Aviva détient entre 15% et 18% du marché, principalement grâce à son partenariat avec *Tifgroup*, deuxième plus grand courtier d'assurance voyage au Royaume-Uni.

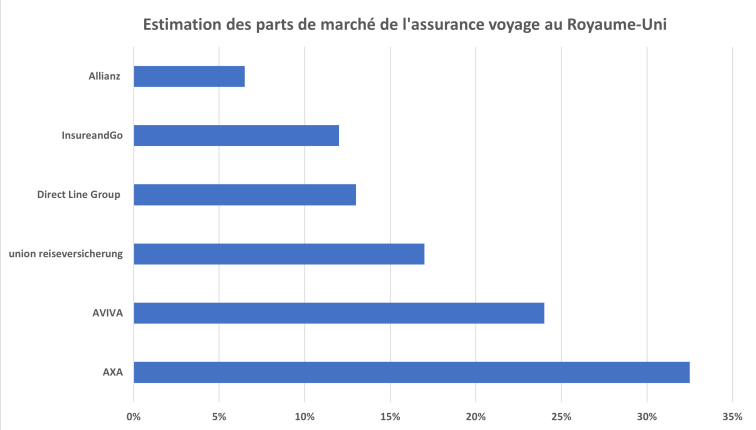


FIGURE 1.3 – Acteurs du marché de l'assurance voyage au Royaume-Uni

1.2.2 Sinistres le plus courant en assurance voyage au Royaume-Uni

Selon l'Association des assureurs britanniques (ABI) en 2017, environ 510 000 sinistres ont été enregistrés pour un coût d'environ 457 millions d'euros, dont :

- 174 000 sinistres d'annulation, qui ont coûté 172 millions d'euros, soit environ 1 000 € le sinistre.
- 159 000 sinistres d'Assistance médicale avec un coût total de 238.6 millions d'euros, soit environ 1 500 € le sinistre.

Les sinistres médicaux sont les plus coûteux, cela s'explique par le fait que les sinistres d'assistance médicale comprennent l'hospitalisation, le remboursement des frais médicaux engagés ou bien l'éventuel rapatriement sanitaire en cas de maladie ou d'accident à l'étranger.

Bien que la fréquence des sinistres médicaux soit inférieure aux taux d'annulation, 60 % des indemnités versées par l'assurance voyage au Royaume-Uni en 2019 couvrent des sinistres médicaux. La figure 1.4 montre l'évolution du montant des sinistres payés en assurance voyage au marché britannique, pour les sinistres les plus fréquents et les plus coûteux qui sont l'annulation et l'assistance médicale.

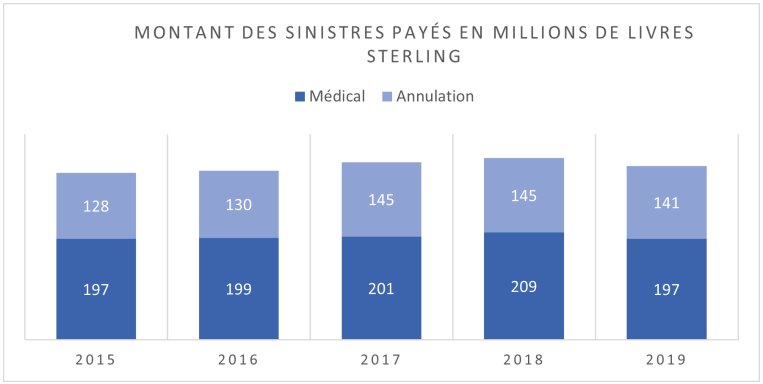


FIGURE 1.4 – Évolution de montant de sinistres payé en assurance voyage

1.2.3 Différence entre le marché européen et le marché britannique

Le marché britannique de l'assurance voyage fonctionne différemment du marché européen, que ce soit au niveau de la couverture ou bien de la méthode de tarification.

En Europe et dans le cadre d'une annulation de voyage, l'assureur s'engage à indemniser uniquement le prix de déplacement ou du vol de l'assuré. Cette condition permet aux assureurs de limiter le coût des sinistres et de ne pas s'exposer à des risques plus difficiles à évaluer (les frais d'hôtels réservés, nombres d'excursions planifiées...).

Sur le marché britannique, la compagnie d'assurance voyage garantit à ses clients une indemnisation de tous les frais de voyage et d'hébergement pour lesquels le voyageur s'est engagé. Ainsi, l'assurance couvre le coût total de voyage et non seulement le montant du vol.

Dans un tel contexte, les assureurs au Royaume-Uni sont exposés à un risque considérable par rapport aux assureurs sur le marché européen. La proposition de ce type de police peut générer un effet d'anti-sélection qui incite les personnes avec une tendance d'annulation aiguë et dont les voyages ont un coût élevé à souscrire à une assurance.

Dans ce sens et pour éviter ce phénomène, les compagnies d'assurance voyage ont mis en œuvre un plafond de couverture pour limiter le montant remboursé tout en proposant différentes formules de police : du basique au haut de gamme pour répondre aux besoins des différents profils de voyageurs.

Ainsi, l'assuré choisira sa gamme selon le montant qu'il s'est engagé à payer pendant son séjour.

Gamme	Plafond, selon l'assureur
Super-économique	Entre £500 et £1 000
Économique	Entre £1 500 et £2 000
Moyenne gamme	Entre £2 500 £ et £3 000
Haut de gamme	Au-dessus de £5 000

TABLE 1.2 – Les plafonds de remboursement selon la gamme de police.

Donc, en offrant différents types de couverture, l'assureur sera en mesure d'évaluer approximativement le montant qu'il assure.

La proposition de différentes formules caractérise le marché britannique. D'une manière générale, en Europe, la compagnie d'assurance voyage propose une seule formule avec un plafond de couverture identique pour tous les voyageurs (environ 3 000 €). De plus, les assureurs connaissent le montant exact du prix du vol assuré puisqu'il est demandé au moment de la souscription.

En effet, la tarification de l'assurance voyage au marché européen dépend principalement du coût du voyage étant donné qu'elle est en pourcentage du montant du voyage.

L'absence d'informations sur le montant des voyages au Royaume-Uni oblige Europ Assistance à adapter sa structure tarifaire et à opter pour la proposition de différents plafonds de remboursement, comme les concurrents au Royaume-Uni. Pour comprendre davantage ce marché, une étude a été réalisée.

1.2.4 Étude de marché

Pour réaliser une étude détaillée du marché, nous avons commencé par regarder les tarifs proposés par quelques assureurs sur le marché britannique.

Le profil type choisi pour réaliser ce benchmark est :

- Age : 25 ans
- Destination : France
- Mois de départ : Août 2022
- Durée de voyage : entre 1 et 30 nuits
- Plafond de couverture : £5 000

Nous avons ainsi obtenu, à partir de ce profil, les primes proposées par quatre assureurs sur ce marché qui sont : Esure, Allianz, OASIS assurance et Aviva. Les critères pour sélectionner ces assureurs sont les suivants :

- Pour Aviva et Allianz : ces assureurs font partie des 8 principaux assureurs détenant la plus grande part de marché au Royaume-Uni ;

- Pour Esure : cet assureur fait partie des assureurs offrant les primes les plus concurrentielles sur le marché ;
- Concernant OASIS assurance : il s'agit d'un nouvel assureur sur le marché. OASIS assurance est détenu par UK Oasis Group Ltd, une société active constituée le 30 octobre 2020 et basée à Londres.

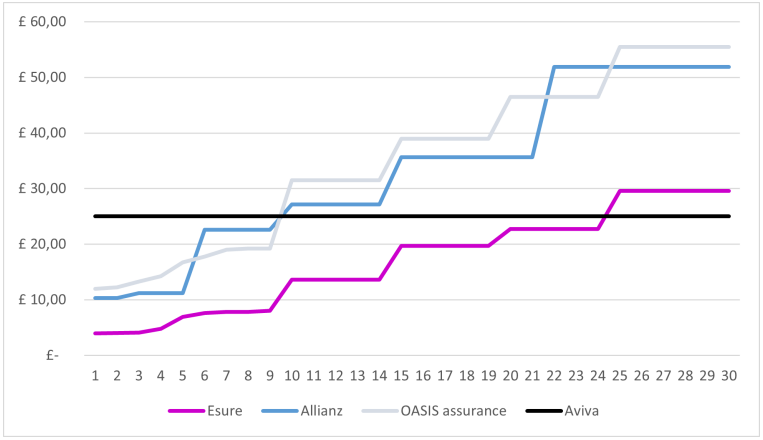


FIGURE 1.5 – Évolution de la prime selon la durée du voyage pendant le mois d'août

Nous remarquons sur la figure 1.5 qu'Aviva a une stratégie différente de celle des autres assureurs en proposant une prime constante quelle que soit la durée du voyage. Pour un voyage entre 1 et 9 nuits, cette assurance offre une prime supérieure à celle de ses concurrents. Toutefois, à partir d'une durée de 10 nuits, les prix proposés par Allianz et OASIS assurance dépassent ceux d'Aviva.

Concernant Esure, Allianz et OASIS assurance, nous remarquons qu'ils ont une stratégie similaire. En effet, leurs tarifs évoluent par palier ; la prime s'accroît par groupe et par durée du voyage. L'assureur Esure offre une prime commerciale peu élevée par rapport à l'assurance Allianz ou OASIS, et ces deux derniers assureurs (Allianz et Oasis) proposent des primes commerciales similaires avec une stratégie comparable.

Afin de pousser encore plus l'analyse de marché de l'assurance voyage au Royaume-Uni, nous avons décidé d'analyser certaines variables tarifaires utilisées par les assureurs. Pour cette analyse, nous avons choisi Allianz en tant que profil assureur à observer.

Analyses des variables tarifaires

Les analyses des variables tarifaires permettent de comprendre le lien entre les variables demandées à la souscription et le tarif. Pour souscrire à une assurance annulation au Royaume-Uni, ce sont toujours les mêmes questions posées au voyageur par tous les assureurs :

- Types de contrats : (unique, annuel, long séjour)
- Destination de voyage ;
- Date de départ/retour de voyage ;
- Nombres de voyageurs
- Date de naissance du voyageur.

Par la suite, c'est la structure tarifaire de l'assureur Allianz qui sera analysée. La police qui va faire l'objet de l'étude est une police d'assurance voyage individuelle. Ce contrat rembourse le voyageur dans le cas où il doit annuler son voyage et lui propose également une Assistance médicale pendant son séjour.
Les variables tarifaires qui seront analysées sont les suivantes :

- Mois de voyage
- Durée de voyage
- Période de réservation : La durée entre le moment de réservation et le moment de départ en vacances.

Nous avons décidé de ne pas analyser la variable de destination et la variable âge des voyageurs dans la mesure où ces variables sont actuellement absentes des bases utilisées par ce mémoire. Enfin, pour les simulations, nous avons choisi la France comme destination : elle est l'une des plus privilégiées par les voyageurs britanniques.

Analyse de la variable date de départ/ retour de voyage

Pour cette variable nous pouvons analyser la durée du voyage, le mois de départ en voyage et le mois de son retour. Concernant la durée du voyage, son analyse a été effectuée à la section 1.2.4. Pour ce qui suit, seule l'étude de la variable mois de départ sera présentée.

L'observation de l'évolution de la prime en fonction du mois de départ montre que Allianz fait une tarification par saison : Été-Automne et Hiver-Printemps.

- La saison Été-Automne s'étale de juin à novembre ;
- La saison Hiver-Printemps comprend les mois de décembre à mai.

Certes, le prix varie en fonction de la période, mais cette variation diminue avec la durée du trajet. Le chapitre 3 de la base de données fournira plus d'explications à ce sujet.

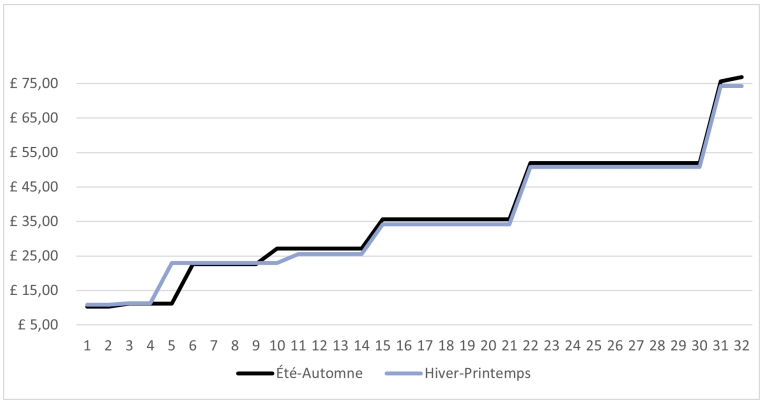


FIGURE 1.6 – Évolution de la prime selon la durée et période de départ

Analyse de la variable Période de réservation

Pour étudier l'impact de cette variable, il a été décidé de fixer toutes les variables tarifaires comme l'âge du voyageur, la durée de voyage, le mois de départ et de faire varier l'année de départ en voyage.

Nous avons constaté que les prix d'Allianz ne dépendaient pas de la durée entre la date de réservation et la date de départ en voyage (période de réservation). A titre d'exemple, la prime commerciale pour une période de réservation de 1 mois et pour un voyage en Août 2022 est égale à la prime pour une période de réservation de 1 an et pour un voyage en Août 2023.

Toutefois, une analyse des primes pour la compagnie d'assurance voyage d'Esure a montré que les prix variaient considérablement en fonction de la période de réservation. Toute chose égale par ailleurs, la prime commerciale pour une période de réservation de 1 mois et pour un voyage en août 2022 coûte 7.81£ alors qu'une assurance voyage pour une période de réservation de 1 an et un voyage en Août 2023 coûte 67.81£

Conclusion

Ces analyses préliminaires étaient importantes pour comprendre le nouveau marché qu'Europ Assistance souhaite intégrer. Premièrement, il a été constaté que les assureurs du marché britannique offraient différents plafonds pour la couverture des annulations. Par ailleurs, bien qu'il y ait peu de variables tarifaires, chaque assureur a sa propre méthode de tarification.

Par conséquent, dans ce mémoire, toutes les variables tarifaires disponibles seront étudiées et utilisées pour établir une prime adéquate au marché britannique.

Pour se faire, il faut disposer de tous les éléments nécessaires. Voilà pourquoi nous avons décidé de recueillir aussi des données en *Open Data*. La partie suivante aborde l'*Open Data* de manière un peu plus détaillée.

1.3 Open Data

Le principe de l'*Open Data* date des années 1966 lors de la sortie de la loi *Freedom of Information Act* (en français, loi d'accès à l'information). Cette loi repose sur le droit à l'information en obligeant les agences fédérales à transmettre leurs documents quelle que soit la personne et quelle que soit sa nationalité.

L'*Open Data* (en français : Les données ouvertes) sont l'ensemble des données numériques qui sont publiques, libres de droits et pour lesquelles l'accès est gratuit. Elles sont accessibles, utilisables et peuvent être partagées par toute personne tierce.

Les sources de ces données sont diverses, elles peuvent provenir de services publics ainsi que privés. Les services publics sont les services gouvernementaux, les communes ou les collectivités locales. Le secteur privé est constitué des entreprises, des organisations non-gouvernementales (ONG) ou des fondations...

En Europe l'*Open Data* s'est beaucoup développé. En 2021, le score moyen de maturité en Open Data pour les pays de l'Union européenne a atteint 81%, soit une augmentation de trois points par rapport à 2020.

En 2021, la France arrive en tête de classement pour la première fois avec un score de $97,5\%$, suivie de l'Irlande, l'Espagne, la Pologne et l'Estonie, comme le montre la figure 1.7.

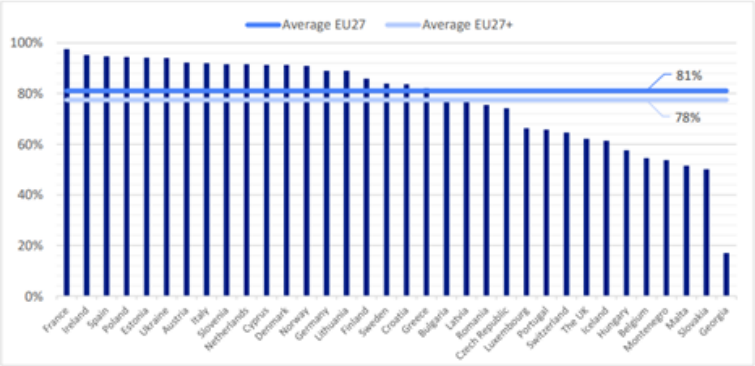


FIGURE 1.7 – Scores globaux de maturité de l'Open Data en 2021, (source)

L'un des principaux atouts de l'utilisation de l'*Open Data* est sa gratuité. Les données gouvernementales sont considérées comme étant au bénéfice du public.

Ainsi l'*Open Data* pourrait favoriser la vie démocratique en donnant aux individus de nouveaux moyens pour comprendre l'action publique.

Par ailleurs, l'*Open Data* offre également une opportunité pour la création de valeurs pour les entreprises dans différents secteurs et cette étude se concentrera sur les possibilités offertes par l'*Open Data* dans le contexte de l'assurance non-vie.

Chapitre 2

Théorie

Ce chapitre est consacré à la définition et à la présentation des différentes approches et modèles utilisés pour tarifier les produit d'assurance annulation voyage. Nous introduirons d'abord le principe de la tarification dans l'assurance non-vie. Nous exposerons ensuite les principes des modèles linéaires généralisés. Enfin, nous expliquerons la théorie derrière les modèles de série temporelle ARIMA et ETS.

2.1 Modèle de Fréquence-Sévérité

En assurance non-vie, la modélisation des primes pures peut se faire selon deux approches :

- une approche individuelle
- une approche collective

L'approche de tarification utilisée dans notre étude est l'approche collective.

L'approche collective :

Ce modèle tient compte du fait que les polices d'un portefeuille sont homogènes et qu'une police peut avoir plusieurs sinistres. Les sinistres du portefeuille sont indépendants et identiquement distribués.

Nous supposons, par la suite, que notre portefeuille satisfait aux hypothèses pour choisir un modèle collectif, malgré une éventuelle hétérogénéité.

Dans ce cas, soit X_i le coût du i-ème sinistre et N le nombre total de sinistres pour toutes les polices du portefeuille, alors la sinistralité totale peut être exprimée en tant que la somme de tous les montants de sinistres encourus par les divers individus.

En supposant que les montants des sinistres sont indépendants et identiquement distribués (i.i.d). La variable aléatoire S correspondant au montant total des sinistres est

obtenue par la formule qui suit :

$$S = \sum_{i=1}^{N} X_i$$

Ainsi, la prime pure qui correspond à la charge totale moyenne des sinistres, s'écrit selon la formule suivante :

$$\begin{aligned} \mathbb{E}(S) &= \mathbb{E}(\sum_{i=1}^{N} X_i) \\ &= \mathbb{E}(N)\mathbb{E}(X) \end{aligned}$$

Tout en tenant compte du fait que N est une variable aléatoire L^2 indépendante de $(X_i)_{i\in\mathbb{N}^*}$. $\mathbb{E}(N)$ correspond à la fréquence moyenne de sinistre et $\mathbb{E}(X)$ au coût moyen encore appelé sévérité : d'où le nom du modèle Fréquence- Sévérité.
Deux étapes sont nécessaires au calcul de la prime pure :

- Estimation de la fréquence moyenne des sinistres
- Estimation du coût moyen d'un sinistre

Ces 2 estimations sont très souvent basées sur des algorithmes de *machine learning* tels que : Les modèles linéaires généralisés (*GLMs*).

2.2 Théorie des Modèles Linéaires Généralisés *(GLM)*

À l'heure actuelle, les modèles linéaires généralisés font partie des modèles du *machine learning* les plus utilisés en tarification d'assurance non-vie.

Lorsque les assureurs travaillent en B to B to C, ils doivent souvent communiquer leur barème tarifaire et expliquer le modèle de tarifications à leurs clients. C'est pourquoi ils préfèrent recourir aux modèles GLM parce qu'ils font partie des modèles les plus faciles à interpréter tout en étant performants.

Étant donné que le GLM est une extension de la régression linéaire ordinaire, nous allons d'abord introduire ce modèle.

2.2.1 Régression linéaire ordinaire

La régression linéaire est un algorithme d'apprentissage supervisé fondé sur des principes statistiques. Elle est principalement utilisée pour modéliser la relation entre une **variable à expliquer**, généralement Y, et une ou plusieurs **variables explicatives** indépendantes désignées par X. Lorsqu'une seule variable indépendante est utilisée pour

prédire Y, il s'agit de la régression linéaire simple ou de régression linéaire, alors que lorsqu'il y a plusieurs variables indépendantes, il s'agit de la régression linéaire multiple.

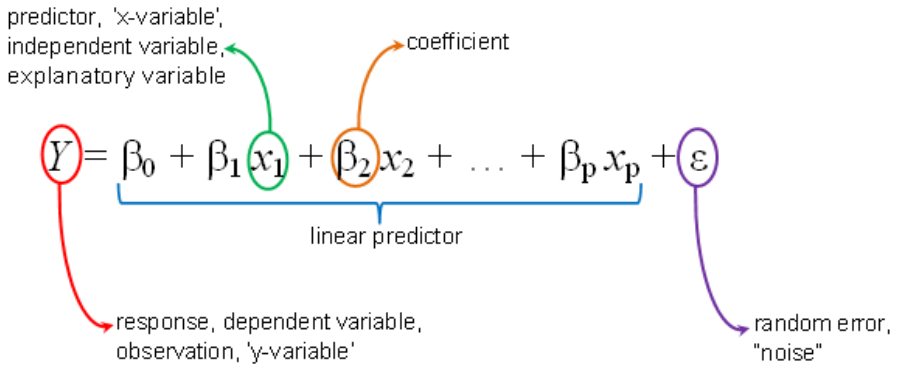


FIGURE 2.1 – Équation d'une régression linéaire ordianaire (source)

- $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$ connus sous le nom de coefficients de régression, font référence à la relation entre la variable explicative X et la variable à expliquer Y.

- Le terme d'erreur $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{pmatrix}$ est utilisé pour désigner les résidus ou la différence entre les valeurs réelles et les valeurs prédites de Y. Il indique dans quelle mesure les prédictions $\hat{Y}$ sont éloignées de la valeur réelle Y.
  Formule de calcul des résidus :
  $$\varepsilon = Y - \hat{Y}$$

Le modèle de la régression linéaire simple se repose sur les hypothèses suivantes :

- $\mathbb{E}(\varepsilon_t) = 0$
- $\mathbb{E}(\varepsilon_t^2) = \sigma^2$
- $cov(\varepsilon_i, \varepsilon_j) = 0$, pour $i \neq j$
- $\mathbb{E}(\varepsilon_t) = 0$

D'après ces hypothèses, nous déduisons que les ε_i sont des réalisations indépendantes d'une variable ε d'espérance nulle et de la même variance σ^2 pour tout x_i et qu'il ne devrait y avoir aucune variable explicative qui représente une combinaison linéaire des autres.

Cependant, la plus grande limite du modèle de régression linéaire dans l'analyse des données est l'hypothèse de la normalité. Cette hypothèse suppose que $Y \sim \mathcal{N}\left(\mathcal{X}\beta, \sigma^2\mathcal{I}_n\right)$. C'est à dire que la distribution des échantillons est normalement répartie. Si cette hypothèse n'est pas respectée, les résultats ne sont pas fiables, ce qui est une source d'erreur importante ce qui nous pousse donc à la mise en place d'un modèle linéaire généralisé.

2.2.2 Modèles Linéaires Généralisés

Le modèle linéaire généralisé est une extension du modèle linéaire simple pour les situations où la variable à expliquer Y ne suit pas une distribution normale.
En effet, Y peut, au contraire, provenir d'une grande famille de distributions connue sous le nom de famille exponentielle, qui comprend : **les distributions normale, de Poisson et binomiale**.

Le modèle linéaire généralisé se généralise en permettant au modèle linéaire simple d'être relié à la variable de réponse par une fonction de liaison. Ce modèle se repose sur trois éléments :

- Distribution de probabilités, $Y \sim f_Y(y)$, tel que la distribution de Y appartient à la famille exponentielle [1].
- Prédicteur linéaire, $\eta = \mathbf{X}\beta$
- Fonction de liaison, $g(.)$ tel que $g(.)$ est une fonction monotone, différentiable et inversible et vérifie l'équation suivante :

$$g(\mu_i) = X_i^T\beta$$

avec μ_i considéré comme étant la moyenne conditionnelle et $\mu_i = E[Y_i|X_i]$

Selon la distribution de Y [2.2.2] divers modèles sont obtenus. Dans ce mémoire, nous allons utiliser le modèle **régression logistique** et le modèle **Log-Gamma** pour prédire respectivement la fréquence d'annulation de voyage et le coût moyen d'un billet d'avion.

1. La densité de la variable Y est sous la forme suivante :

$$f_{Y_i}\left(y_i, \theta_i, \phi\right) = \exp\left[\frac{y_i\theta_i - b\left(\theta_i\right)}{a_i(\phi)} + c\left(y_i, \phi\right)\right]$$

Modèle régression logistique

La régression logistique est une transformation du modèle de régression linéaire et elle est également connue comme un modèle linéaire généralisé qui utilise un lien logit. Elle permet de modéliser de manière probabiliste des variables binaires. En effet, La régression logistique modélise la probabilité qu'une observation prenne l'une de ces deux valeurs. Étant donné que le modèle ne prédit qu'une probabilité, pas une classe, le choix du seuil de décision dépend du cas d'utilisation.

Les paramètres d'une régression logistique sont les suivants :

- **Loi :** Bernouilli $\mathcal{B}(\pi)$
- **Support :** $\{0;1\}$ **Fonction de lien :** Logit $g(x) = \ln\left(\frac{x}{1-x}\right)$
- **Expression du modèle :** $\mathbb{E}[Y \mid X] = (1 + \exp(-(X\beta)))^{-1}$

En assurance voyage, la fréquence et l'exposition concernent le portefeuille en général et non pas une police en particulier. En effet, une police ne peut comporter plus d'un sinistre, puisque dès qu'un sinistre est déclenché, le contrat prend fin. En outre, la période de couverture de l'assuré ne correspond qu'à la période entre l'inscription à l'assurance et le départ pour le voyage. Il est, donc, difficile de parler d'exposition et de fréquence d'une police en assurance voyage.
Pour cette raison, nous souhaitons modéliser la probabilité qu'une police annule un voyage, puis utiliser cette probabilité comme une fréquence pour le modèle de la tarification. D'où le choix de la régression logistique.

Modèle Log-Gamma

Le modèle Log-Gamma modélise une distribution Gamma avec un lien logarithmique. Ce modèle convient bien aux données uniquement positives comme les coût des sinistres. La distribution Gamma est flexible et peut imiter une forme log-normale.

Les paramètres d'un modèle Log-Gamma sont les suivants :

- **Loi :** Gamma $\mathcal{G}(\alpha, \beta)$
- **Support :** $\mathbb{R}^+$
- **Fonction de lien :** Ln $g(x) = \ln(x)$
- **Expression du modèle :** $\mathbb{E}[Y \mid X] = \exp(X\beta)$

Sélection de variables

Le choix des variables est une étape essentielle lorsque beaucoup de variables explicatives sont présentes. Cette sélection aide à améliorer la performance d'un modèle en

identifiant et en supprimant les variables inutiles ne contribuant pas à la précision du modèle.

À cette fin, il existe plusieurs algorithmes pour sélectionner automatiquement les variables. Les méthodes couramment employées pour sélectionner les variables sont des méthodes pas à pas. Ces méthodes font appel à des algorithmes qui reproduisent la régression en ajoutant ou en retirant certaines variables à chaque étape. Le modèle choisi est celui dont la sélection des variables permet de les réduire au minimum les critères AIC/BIC.
Généralement trois méthodes sont citées pour sélectionner des variables : :

- **Méthode *backward* :** La méthode *backward* encore connue sous le nom de méthode descendante est une méthode d'élimination de variables. Le principe de cette méthode est de partir du modèle complet avec toutes les variables explicatives présélectionnées et de rechercher, à chaque pas de l'algorithme, la variable la plus pertinente à supprimer en fonction du critère retenu. L'algorithme s'arrête quand il n'est plus possible d'améliorer le modèle en supprimant les variables.

- **Méthode *forward*** La méthode *forward* encore connue sous le nom de méthode ascendante, consiste à ajouter une variable au modèle depuis le modèle constant. En effet, l'algorithme part d'un modèle vide qui n'inclut qu'une constante et ajoute les variables les plus significatives une par une.

- **Méthode *stepwise*** La méthode stepwise est une combinaison de la méthode *backward* et de la méthode *forward*. Cet algorithme commence avec un modèle constant sans variables explicative, puis il ajoute séquentiellement les variables les plus significatives (comme pour la méthode *forward*). Après avoir ajouté chaque nouvelle variable, le modèle supprime toutes les variables en cherchant à minimiser le critère considéré (comme une sélection *backward*).

Une fois les variables sélectionnées et les données modélisées, une étape importante doit être établie : la vérification de la performance à l'aide des indicateurs de performance.

Indicateurs de performance des modèles

Les modèles de *Machine learning* peuvent être évalués grâce à plusieurs indicateurs. En voici quelques-uns :

- **L'erreur quadratique moyenne RMSE :** L'erreur quadratique moyenne (RMSE) est une méthode standard pour mesurer l'écart entre la valeur prédite et la valeur observée. La RMSE est définie comme suit

$$RMSE(\hat{y}) = \sqrt{\sum_{i=1}^{N} \frac{(\hat{y}_i - y_i)^2}{N}} \text{ (Root Mean Square Error)}$$

avec : $\hat{Y} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{pmatrix}$ sont les valeurs prédites par le modèle et $Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$ sont les valeurs de la variable à expliquer et N est le nombre d'observations. Ainsi, plus le RMSE est élevé, plus la différence entre les valeurs prédites et observées est importante, ce qui signifie que le modèle de régression s'adapte moins bien aux données. Inversement, plus le RMSE est petit, plus le modèle est performant.
Il peut être particulièrement utile de comparer le RMSE de deux modèles différents pour voir lequel est le plus adapté.

- **l'erreur absolue moyenne MAE :** L'erreur absolue moyenne (MAE) est une méthode utilisée couramment avec le RMSE pour tester la performance d'un modèle de prédiction. Cette méthode est définie par la formule suivante :

$$MAE(\hat{y}) = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \qquad \text{(Mean Absolute Error)}$$

- **AUC :** Dans le cadre d'une régression logistique, la performance d'un modèle peut être évaluée à l'aide de l'AUC (l'aire de la zone sous la courbe ROC). L'AUC est un indicateur qui mesure la performance du modèle de classification binaire sans qu'il soit nécessaire de spécifier un seuil.
Afin de comprendre cette mesure il faut tout d'abord définir la courbe **ROC**.

  - La courbe *ROC* indique la performance d'un modèle de classification pour toutes les valeurs seuils éventuelles à l'aide d'un graphique. Le ROC a deux paramètres :

    - La sensibilité (TPR) qui est défini comme suit :

$$\text{sensibilité} = \frac{\text{Vrais positifs}}{\text{Vrais positifs} + \text{Faux négatifs}}$$

    - 1 - La spécificité (FPR) qui est défini comme suit :

$$\text{Taux de faux positifs} = \frac{\text{Faux positifs}}{\text{Faux positifs} + \text{Vrais négatifs}}$$

    Une courbe ROC représente le TPR par rapport au FPR à différents seuils de classification. En abaissant le seuil de classification, le modèle classe plus d'éléments comme positifs, ce qui augmente à la fois les faux positifs et les vrais positifs.
    La figure 2.2 montre une courbe ROC typique.

    La mesure AUC est donc l'aire sous la courbe bleue. Plus l'aire est importante, meilleur est le modèle.

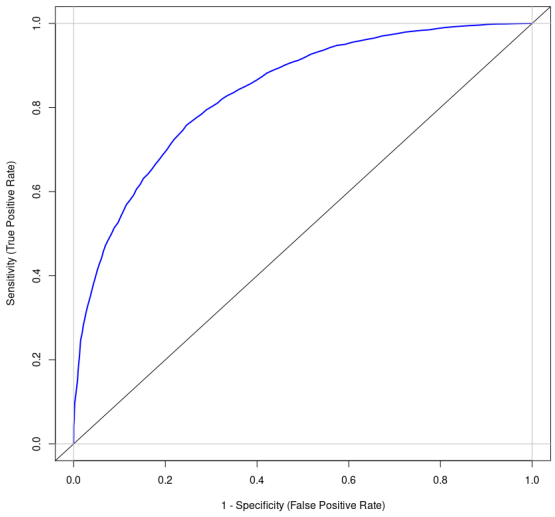


FIGURE 2.2 – Courbe ROC avec une Prédiction de probabilité parfaite (source)

D'après *Hosmer et Lemeshow* dans leur ouvrage *Applied Logistic Regression*, l'interprétation de la performance du modèle peut être jugée de la manière suivante :

- Un modèle ayant un AUC de 0,5 est équivalent à une classification aléatoire
- Un score AUC de 0,6 à 0,7 indique que la classification du modèle est relativement acceptable
- Si l'AUC est comprise entre 0,7 et 0,8, le rendement du modèle est acceptable. Au-delà de ce seuil, la classification est considérée comme étant excellente, voire exceptionnelle.

2.3 Théorie des séries temporelles

Cette partie, inspirée d'un mémoire d'actuariat rédigé par *M.BOUTTIER* et du cours des séries temporelles dispensé à l'EURIA par *M.AILLIOT* et *M.DERRIEN*, a pour but d'introduire les séries temporelles ainsi que les modèles utilisés dans ce mémoire.

2.3.1 Introduction aux séries temporelles

Une série temporelle (ou une série chronologique) est un ensemble d'observation (X_t) d'une même quantité indexée par le temps $t \in T$, où $T \subset \mathbb{Z}$ est un ensemble discret de dates avec un intervalle fixé.

Une série temporelle peut être considérée comme la réalisation d'un processus stochastique. En effet, soit $X = (X_t)_{t \in T}$ un processus stochastique en temps discret ($T = \mathbb{Z}$

ou bien $T = \mathbb{N}^*$), à savoir un ensemble dénombrable de variables aléatoires définies sur le même espace de probabilité (Ω, A, P) dans $\mathbb{R}^d$. Une réalisation $X(\omega) = (x_t)$ est appelée une trajectoire et les séries temporelles observées $x_1, ..., x_t..., x_n$ (souvent notée $X_1, ..., X_n$) font partie d'une trajectoire.

Généralement une série temporelle se décompose en trois éléments :

- Une tendance, T_t
- Une composante saisonnière, S_t
- Une composante résiduelle, ε_t

et elle peut s'écrire en fonction de ses composantes sous la forme d'un modèle additif comme suit :

$$X_t = S_t + T_t + R_t$$

Dans certains cas, la série temporelle peut être sous la forme d'un modèle multiplicatif qui se présente comme suit :

$$X_t = S_t \times T_t \times R_t$$

La tendance T_t : est une représentation de l'évolution à long terme de la série, par exemple le changement climatique sur le long terme.

La composante saisonnière S_t : est une représentation d'un phénomène qui se reproduit chaque année, mensuellement ou toutes les semaines. À titre d'exemple, en météorologie les températures sont plus basses en hiver qu'en été.
Soit S_t une saisonnalité de période p alors :

$$\forall t \in \mathbb{N},\ S_{t+p} = S_t$$

La composante résiduelle R_t : elle semble avoir un comportement aléatoire et elle est visualisée lorsque la tendance et la saisonnalité sont enlevées.

Série temporelle stationnaire

Lorsque une série temporelle ne présente pas de tendance ou de saisonnalité, elle est considérée comme étant stationnaire.
Il existe différents niveaux de stationnarité dont :

- **La stationnarité stricte :** La série (X_t) tq $X_t \in \mathbb{R}^d$, est strictement stationnaire si $(X_1, X_2, ..., X_k)$ a la même distribution que $(X_{1+h}, X_{2+h}, ..., X_{k+h})$, $\forall h$ et $k \geq 1$.
- **Stationnarité faible :** (X_t) est faiblement stationnaire ou stationnaire du second ordre si :
  - $\mu_X(t)$ est indépendante de t

- $\gamma_X(t, t+h)$ est indépendante de t, $\forall h$

μ_X et γ_X sont respectivement la fonction de la moyenne et la fonction d'autocorrélation. Elles sont définies comme suit :

- La **fonction de la moyenne** $\mu_X(.)$ est :

$$\mu_X = \mathbb{E}(X_t), \text{ pour tout } t \in \mathbb{Z}$$

- Pour présenter la fonction d'autocorrélation, il est nécessaire d'introduire la **fonction autocovariance** qui est définie comme suit :

$$\gamma_X(t, h) = Cov(X_t, X_h), \text{ pour tout } t,\ h \in \mathbb{Z}$$

- La **fonction d'autocorrélation** $\rho(.)$, est une normalisation de la fonction d'autocovariance et elle est définie par :

$$\forall t \in \mathbb{N}, \rho(t) = \frac{\gamma(t)}{\gamma(0)}$$
$$\rho(0) = 1$$

- A partir de la fonction d'autocorrélation, il est possible de présenter la **fonction d'autocorrélation partielle** $\tau(.)$ qui est définie par l'équation suivante :

$$\tau_X(t) = \text{corr}_{X_2,\ldots,X_t}(X_1, X_{t+1})$$
$$\text{ainsi : } \tau_X(1) = \rho(1) = \text{corr}(X_1, X_2)$$

Test de stationnarité

Pour vérifier la stationnarité d'une série temporelle, il existe divers tests dans la littérature, en voici quelques-uns :

- **Test ADF :** Pour bien comprendre ce test, il faut tout d'abord introduire le test DF.

  - **Test DF :** Ce test vérifie la présence de racines unitaires d'un polynôme autorégressif (AR 2.3.2). Soit $(X_t)_{t\in T}$ une série temporelle, tel que :

$$X_t = \alpha + \rho X_{t-1} + \phi \Delta X_{t-1} + \varepsilon_t$$

  avec $\varepsilon \sim b.b(0, \sigma^2), \alpha \in \mathbb{R}$
  Les hypothèses du test sont les suivantes :

$$H_0 : |\rho| = 1 \tag{2.1}$$
$$H_1 : |\rho| < 1 \tag{2.2}$$

  Donc, si le *p-value* du test est supérieur à 5 % alors H_0 est rejetée et la série est stationnaire.

- **Test ADF :**
  Ce test est une version plus avancée du test DF. En effet, l'ADF étend l'équation du test DF pour inclure d'autres ordres du processus régressif. L'équation du test est la suivante :

$$X_t = \alpha + \rho X_{t-1} + \sum_{i=1}^{k} \beta_i \Delta X_{t-i} + \varepsilon_t$$

  L'hypothèse nulle reste inchangée par rapport au test DF : :

$$H_0 : |\rho| = 1 \tag{2.3}$$
$$H_1 : |\rho| < 1 \tag{2.4}$$

  De même, si le *p-value* est supérieure à 5 %, alors il n'existe pas de racine unitaire dans la série et elle est stationnaire.

➢ **Test KPSS :**

Ce test a été motivé par le fait que les tests ADF et DF supposent que toutes les séries temporelles possèdent des racines unitaires. C'est pourquoi le test KPSS prend en compte la stationnarité tendancielle comme hypothèse nulle et la présence d'une racine unitaire comme hypothèse alternative.

Le test KPSS écrit la série temporelle sous le modèle suivant :

$$X_t = r_t + \varepsilon_t,$$

$$r_t = r_{t-1} + u_t$$

$\forall t > 0$, r_0 constant, u_t iid $\mathcal{N}(0, \sigma_u^2)$ et ε_t iid $\mathcal{N}(0, \sigma_\varepsilon^2)$.
et les hypothèses sont :

$$H_0 : \sigma_u^2 = 0$$
$$H_0 : \sigma_u^2 \neq 0$$

Si le test donne un *p-value* inférieure à 0.5 %, alors la série est stationnaire.

Bruit blanc

Le **bruit blanc** est un cas particulier d'un processus stationnaire. En effet, un bruit blanc est l'un des processus stationnaires les plus courant. Nous pouvons retrouver deux types de bruit blanc : un bruit blanc faible et un bruit blanc fort.

➢ un **bruit blanc faible** est une séquence (ϵ_t) de variables non corrélées de moyenne nulle et de variance constante.

$$E(\epsilon_t) = 0$$
$$Var(\epsilon_t) = \sigma^2$$
$$Cov(\epsilon_t, \epsilon_s) = 0, \text{ si } t \neq s$$

- un **bruit blanc fort** est une suite (ϵ_t) de variables indépendantes et identiquement distribuées (iid), tel que :

$$E(\epsilon_t) = 0,$$
$$Var(\epsilon_t) = \sigma^2$$

  Si les v.a. (ϵ_t) sont gaussiennes, alors il s'agit dans ce cas d'un **bruit blanc gaussien**. Les processus gaussiens stationnaires sont définis de manière unique par leur fonction d'autocovariance.

Après avoir introduit les séries temporelles, la partie suivante présentera les modèles de prédiction ARIMA et ETS utilisés dans ce mémoire.
En effet, les modèles *ETS* et les modèles *ARIMA* sont les deux approches les plus utilisées en matière de prédiction de séries temporelles et fournissent des approches complémentaires au problème.

Pour ce qui suit, le cadre de la modélisation est le suivant : la série temporelle (X_t) a été observée de $t = 1$ à $t = T$ et sur la base de ces observations le modèle prédira X_{T+1} ou plus généralement, X_{T+h} pour $h \geq 1$ (h est appelé l'horizon de prévision).

2.3.2 Modèle ARIMA

Le modèle ARIMA est développé à l'origine par *Box et Jenkins* en 1976. Le nom ARIMA est une abréviation de *AutoRegressive Integrated Moving Average* (moyenne mobile intégrée autorégressive). L'idée principale du modèle est que les données peuvent avoir des rapports dynamiques dans le temps, où les nouvelles valeurs dépendent des valeurs antérieures.

Préliminaires mathématiques

Opérateur retard : Cet opérateur, noté **L**, est un opérateur qui à toute observation d'une série temporelle, associe l'observation précédente :

$$LX_t = X_{t-1} \text{ pour tout } t > 1$$

D'une manière plus générale,

$$L^j X_t = X_{t-j}$$

Polynôme retard : Les propriétés de l'opérateur retard permettent d'introduire le polynôme retard, noté $\Phi(L)$.

Ce type de polynôme est employé pour simplifier l'écriture des modèles de classe ARIMA. D'une manière générale un polynôme retard est défini tel que :

$$\Phi_k(L) = 1 - \phi_1 L - \phi_2 L^2 - ... - \phi_k L^k$$

Ainsi, dans le cas d'un processus (X_t) :

$$\Phi_k(L)X_t = 1 - \phi_1 X_{t-1} - \phi_2 X_{t-2} - ... - \phi_k X_{t-k}$$

avec $\phi_1, \phi_2 ... \phi_k$ des coefficients $\in \mathbb{R}$

Différenciation

une série temporelle qui présente une tendance et/ou une saisonnalité n'est pas stationnaire. Ainsi une différentiation peut aider à enlever ces deux composantes.
Une différenciation d'ordre 1 s'écrit :

$$\begin{aligned} \forall t \in [\![2, T]\!] \quad X_t^{(1)} &= X_t - X_{t-1} \\ &= X_t - LX_t \\ &= (1 - L)X_t \end{aligned}$$

Lorsque une seule différenciation n'est pas suffisante, c'est à dire $X_t^{(1)}$ n'est pas stationnaire, il faut continuer à différencier la série.

$$\begin{aligned} \forall t \in [\![3, T]\!] \quad X_t^{(2)} &= X_t^{(1)} - X_{t-1}^{(1)} \\ &= (1 - L)X_t - (1 - L)X_{t-1} \\ &= (1 - L)X_t - (1 - L)LX_t \\ &= (1 - L)^2 X_t \\ &= X_t - 2X_{t-1} + X_{t-2} \end{aligned}$$

Plus généralement, la différenciation d'ordre k peut s'écrire comme suit :

$$\forall t \in [\![k + 1, T]\!] \quad X_t^{(k)} = (1 - L)^k X_t$$

Processus auto-régressifs : AR

Un modèle est dit auto-régressif lorsque la valeur d'un processus est régressée sur ses valeurs précédentes.
Un exemple d'un modèle AR de premier ordre :

$$X_t = \phi_1 X_{t-1} + \varepsilon_t$$

Plus généralement, une auto-régression d'ordre p $(AR(p))$ est une régression linéaire multiple dans laquelle la valeur X_t du processus est exprimée en fonction des valeurs du

processus aux temps : $t-1, t-2, ..., t-p$.
Ainsi, un modèle AR d'ordre p peut être écrit comme suit :

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t \tag{2.5}$$

Il est possible de réécrire l'équation 2.5 précédente en considérant l'opérateur retard L comme suit :

$$\Phi_p(L) X_t = (1 - \phi_1 L - \phi_2 L^2 - ... - \phi_k L^p) X_t = \varepsilon_t$$

Processus moyennes mobiles : MA

Plutôt que d'utiliser les valeurs passées dans une régression comme le modèle AR, un modèle de moyenne mobile (MA) utilise les erreurs passées multipliées par un coefficient. Un processus moyenne mobile d'ordre 1 est :

$$X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

Plus généralement, un processus moyenne mobile d'ordre q, noté $MA(q)$ vérifie la relation suivante :

$$X_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + \theta_q \epsilon_{t-q}$$
$$\text{avec : } \theta_1, ..., \theta_q \in \mathbb{R}$$

Soit Θ un polynôme retard qui vérifie l'équation suivante :

$$\Theta(L) = 1 - \theta_1 L - ... - \theta_q L^q$$

Alors le processus **MA(q)** peut s'écrire ainsi :

$$X_t = \Theta(L) \epsilon_t$$

Modèle ARIMA non saisonnier

Le modèle ARIMA non saisonnier est un modèle qui combine une différenciation, un modèle autorégressif et un modèle de moyenne mobile.
Dans un modèle ARIMA(p, d, q) X_t est modélisé par :

$$X_t^{(d)} = \phi_1 X_{t-1}^{(d)} + \cdots + \phi_p y_{t-p}^{(d)} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

$$(1-L)^d X_t = \sum_{i=1}^{p} \left(\phi_i (1-L)^d X_{t-i} \right) + \sum_{j=1}^{q} \left(\theta_j \varepsilon_{t-j} \right) + \varepsilon_t$$

$$(1-L)^d X_t = \sum_{i=1}^{p} \left(\phi_i (1-L)^d X_{t-i} \right) + \sum_{j=1}^{q} \left(\theta_j \varepsilon_{t-j} \right) + \varepsilon_t$$

$$(1-L)^d X_t = \sum_{i=1}^{p} \left(\phi_i (1-L)^d L^i X_t \right) + \sum_{j=1}^{q} \left(\theta_j L^j \varepsilon_t \right) + \varepsilon_t$$

Avec :

- $d \geq 1$ est le nombre de différenciation de la série
- $\theta_1, ..., \theta_p, \phi_1, .., \phi_p$ sont des coefficients $\in \mathbb{R}$
- $(\varepsilon_t, ..., \varepsilon_{t-q}) \sim b.b(0, \sigma^2)$

En général, l'équation du modèle ARIMA (p,q,d) peut s'écrire avec la notation d'opérateur retard et équation retard comme suit :

$$\underset{\substack{\uparrow \\ \text{AR}(p)}}{(1 - \phi_1 L - \cdots - \phi_p L^p)} \quad \underset{\substack{\uparrow \\ \text{différenciation d'ordre d}}}{(1-L)^d X_t} \quad = \underset{\substack{\uparrow \\ \text{MA}(q)}}{(1 + \theta_1 L + \cdots + \theta_q L^q)\, \varepsilon_t}$$

Ci dessous quelques cas particuliers de modèles ARIMA(p,d,q)

Bruit blanc	ARIMA(0,0,0)
Marche aléatoire	ARIMA (0,1,0)
Auto-régression (AR)	ARIMA (p,0,0)
Moyenne mobile (MA)	ARIMA (0,0,q)

TABLE 2.1 – Cas particuliers de modèles ARIMA

ARIMA avec saisonnalité

Lorsqu'il s'agit de données réelles, il existe généralement un lien entre les observations consécutives et les observations qui surviennent avec des écarts saisonniers fixes.
Dans le cadre d'un ARIMA ces relations sont prises en charge en rajoutant des termes saisonniers supplémentaires dans les modèles ARIMA non saisonniers.

Un ARIMA avec saisonnalité se présente comme suit :

$$ARIMA \quad \underbrace{(p,d,q)}_{\uparrow} \qquad \underbrace{(P,D,Q)_s}_{\uparrow}$$

Partie du modèle sans saisonnalité Partie du modèle avec saisonnalité

et son équation est la suivante :

$$(1-(\sum_{i=1}^{p}\phi_i L^i))(1-(\sum_{i=1}^{p}\Phi_i L^{si}))(1-B)^d(1-L^s)^D X_t = (1+\sum_{i=1}^{q}\theta_i L^i)(1+\sum_{i=1}^{Q}\Theta_i L^{si})\epsilon_t$$

Avec :

- $\theta_1, ..., \theta_q, \Theta_1, ..., \Theta_Q \in \mathbb{R}$
- $\phi_1, ..., \phi_p, \Phi_1, ..., \Phi_P \in \mathbb{R}$
- $\epsilon_t, ..., \epsilon_{t-q} \sim b.b(0, \sigma^2)$

2.3.3 Modèle ETS (Exponential smoothing)

Les modèles ETS (**E**xponen**t**ial **S**moothing) sont définis comme étant une série de modèles de prédiction. Ils peuvent être considérés comme des pairs et une alternative à la classe populaire des méthodes ARIMA.

A la différence des modèles ARIMA, les ETS utilisent les moyennes pondérées des observations précédentes pour prédire de nouvelles valeurs tout en combinant les composantes résiduelles, saisonnières et la tendance dans la modélisation.

Plusieurs méthodes d'ETS sont disponibles, en fonction de la présence d'une tendance et/ou d'une saisonnalité.
On en distingue trois, qui sont :

- Simple ***ETS***
- ***ETS*** pour la tendance
- ***ETS*** pour la saisonnalité

Simple ETS

Ce modèle correspond à un modèle avec des erreurs additives, sans saisonnalité et sans tendance.
L'équation de prédiction et l'équation de lissage se présentent sous les formules suivantes :

$$\text{Équation de prédiction :} \quad \hat{X}_{t+h|t} = \ell_t \tag{2.6}$$

$$\text{Équation de lissage :} \quad \ell_t = \alpha X_{t-1} + (1-\alpha)\ell_{t-1} \tag{2.7}$$

Où :

- ℓ_t est la valeur lissée de la série au temps t
- $\hat{X}_{t+h|t}$ la prévision de X_{t+h} à partir de l'information à t
- α appelé facteur de lissage ou coefficient de lissage, contrôle le taux auquel l'influence des observations des pas temporels antérieurs diminue exponentiellement. α est souvent comprise entre 0 et 1. Une valeur proche de 1 signifie que le modèle est axé sur les observations antérieures les plus récentes et un apprentissage rapide. Tandis qu'une valeur qui tend vers 0 signifie que la prédiction tient compte de l'historique des données temporelles et indique un apprentissage lent.

ETS pour la tendance

Pour ce modèle, deux méthodes d'estimation sont distinguées : **brute** ou bien **amortie**.

Méthode d'estimation brute

Le modèle ETS double est une extension de l'ETS simple en additionnant la prise en charge des tendances dans les séries temporelles.
Ses équations sont comme suit :

$$\text{Équation de prédiction :} \quad \hat{X}_{t+h|t} = \ell_t + hb_t \tag{2.8}$$

$$\text{Équation de lissage :} \quad \ell_t = \alpha X_t + (1-\alpha)\ell_{t-1} + b_{t-1} \tag{2.9}$$

$$\text{Équation de tendance :} \quad b_t = \beta^*(\ell_t - \ell_{t-1}) + (1+\beta^*)b_{t-1} \tag{2.10}$$

- hb_t permet d'ajouter une tendance linéaire pour la prévisions et b_t s'agit de la pente de cette tendance.
- ℓ_t présente la moyenne pondérée de X_t et $\hat{X}_{t|t-1}$.
- b_t correspond à la moyenne pondérée de la pente estimée au moment $t(\ell_t - \ell_{t-1})$ avec la prédiction
- β_t est un coefficient de lissage supplémentaire ajouté pour maîtriser la décroissance de l'influence de la variation de la tendance.

Méthode d'estimation amortie

Les prévisions générées par la méthode brute donnent une tendance constante (croissante ou décroissante) indéfiniment dans le futur. Ainsi cette méthode aura tendance à surestimer les prévisions, en particulier pour les horizons de prévision plus longs.

Motivés par cette observation, Gardner et McKenzie (1985) ont introduit un paramètre qui "amortit" la tendance pour la ramener à une ligne plate à un moment donné dans le futur. Ce paramètre est le facteur d'amortissement.
Ainsi l'équation de prédiction dévient :

$$\text{Équation de prédiction :} \qquad \hat{X}_{t+h|t} = \ell_t + \left(\sum_{k=1}^{h} \phi^k\right) b_t \tag{2.11}$$

$$\text{Équation de de lissage :} \qquad \ell_t = \alpha X_t + (1-\alpha)\left(\ell_{t-1} + \phi b_{t-1}\right) \tag{2.12}$$

$$\text{Équation de tendance :} \qquad b_t = \beta^* \left(\ell_t - \ell_{t-1}\right) + (1-\beta^*)\,\phi b_{t-1} \tag{2.13}$$

Si $\phi = 1$ la méthode revient à une méthode brute.
Pour $\phi \in]0,1[$, ϕ amortit la tendance de sorte qu'elle se rapproche d'une constante à un moment donné dans le futur.
Plus précisément :

$$\begin{aligned} \lim_{h\to+\infty} \hat{X}_{t+h|t} &= \ell_t + \phi \lim_{h\to+\infty} \left(\sum_{k=0}^{h-1} \phi^k\right) b_t \\ &= \ell_t + \frac{\phi}{1-\phi} b_t \end{aligned} \tag{2.14}$$

ETS pour la saisonnalité

Cette méthode propose une alternative au modèle ETS double, en permettant de prendre en compte le caractère saisonnier ainsi que la tendance.
L'ETS pour la saisonnalité utilise les équations suivantes :

- L'équation de prédiction ;
- Trois équations de lissage : une pour le niveau, une pour la tendance et une autre pour la composante saisonnière.

Il existe deux méthodes pour prédire une série temporelle avec une tendance et une saisonnalité :

- Une **méthode additive** qui est préférée lorsque les composantes saisonnières sont à peu près constantes dans le temps ;
- Une **méthode multiplicative** qui est adaptée lorsque les variations saisonnières changent proportionnellement au niveau de la série.

Méthode additive

Les équations pour la méthode additive sont comme suit :

$$\text{Équation de prédiction :} \quad \hat{X}_{t+h|t} = l_t + hb_t + s_{t+h-m(k+1)} \tag{2.15}$$

$$\text{Équation de de lissage :} \quad l_t = \alpha(X_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{2.16}$$

$$\text{Équation de tendance :} \quad b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1} \tag{2.17}$$

$$\text{Équation de la saisonnalité :} \quad s_t = \gamma(X_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m} \tag{2.18}$$

$$\tag{2.19}$$

- m est la période de la saisonnalité de la série temporelle
- k correspond à la partie entière de $\frac{h-1}{m}$
- $0 \leq \gamma \leq 1 - \alpha$
- $(\alpha, \beta^*) \in [0,1]^2$

Méthode multiplicative

Les équations pour la méthode multiplicatives sont comme suit :

$$\text{Équation de prédiction :} \quad \hat{X}_{t+h|t} = (\ell_t + hb_t)s_{t+h-m(k+1)} \tag{2.20}$$

$$\text{Équation de lissage :} \quad \ell_t = \alpha\frac{X_t}{s_{t-m}} + (1-\alpha)(\ell_{t-1} + b_{t-1}) \tag{2.21}$$

$$\text{Équation de tendance :} \quad b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1} \tag{2.22}$$

$$\text{Équation de la saisonnalité :} \quad s_t = \gamma\frac{X_t}{(\ell_{t-1} + b_{t-1})} + (1-\gamma)s_{t-m} \tag{2.23}$$

Chapitre 3

Travail sur la base de données

3.1 Périmètre du mémoire

Le montant du voyage varie selon la destination, le type d'hôtel réservé, la durée du voyage,... Le cadre de ce mémoire se restreint à un périmètre bien défini : les voyages en Europe. En effet, à la lecture d'une étude faite par l'ONS, « Bureau de la statistique nationale britannique » : en 2021 plus de 88% des voyageurs britanniques ont opté pour des destinations européennes, ces voyages sont présentés sur la figure 3.1.

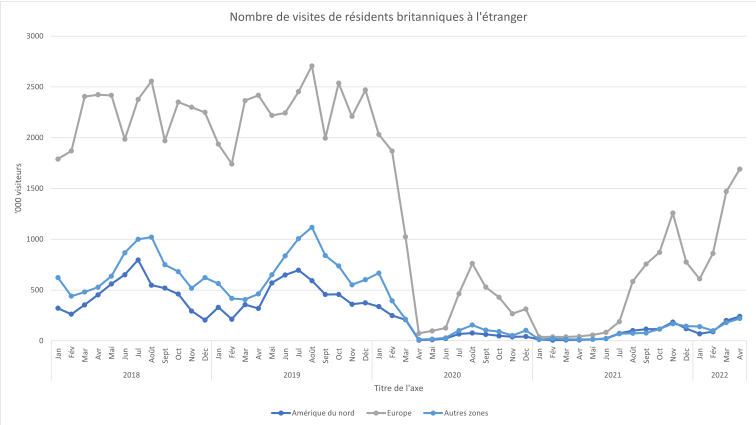


FIGURE 3.1 – Visites de résidents britanniques

Les départs à l'étranger pour les Britanniques sont très similaires avec les départs en vacances pour les Européens en général. En effet, les vacances scolaires en Grande-Bretagne sont semblables à celles en Europe. Les principales périodes de vacances au Royaume-Uni sont Noël, Pâques et les vacances d'été. Les vacances à l'université au Royaume-Uni varient en durée, mais sont généralement entre 3 et 5 semaines.

De plus, l'Espagne, la France, le Portugal et la Grèce figurent en tête de liste des pays

les plus visités [1].

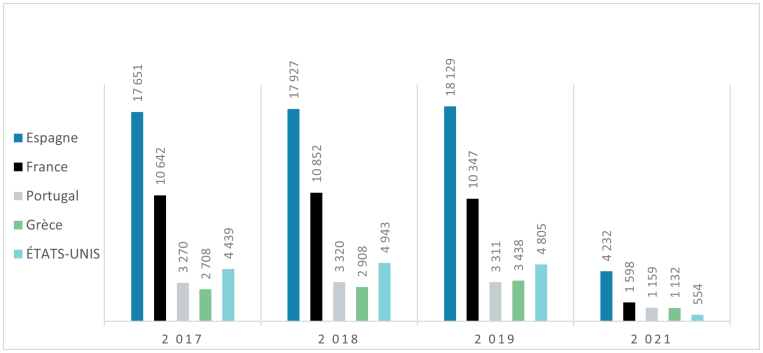


FIGURE 3.2 – Nombre de visite par pays et par année (en milliers)

La figure 3.2 montre le nombre de visites des cinq pays les plus visités par les Britanniques. En raison de la COVID-19, aucune donnée pour l'année 2020 n'a été collectée. En ce qui concerne l'année, 2021, le nombre de visites est clairement exceptionnels, également à cause de la pandémie qui a beaucoup touché le secteur du tourisme. Ainsi, il faut être prudent à l'égard des données durant cette période.

Les données qui ont été enregistrées entre 2020 et 2021 peuvent être traitées de deux façons :

- Ne pas les inclure dans les bases de données et ainsi la dernière année de référence sera l'année 2019.
- Les inclure tout en vérifiant l'homogénéité et la tendance du portefeuille durant ces deux années.

Pour ce mémoire, la base de données « **prix d'hôtel** » ne tiendra pas compte des années COVID, parce que durant cette période, aucune donnée n'a été collecté par l'UMIH [2].

Cependant, pour la base de données « **prix d'avion** » (Source : interne) il est difficile de se limiter à la période antérieure à la COVID-19 parce que le portefeuille s'étend de 2018 à 2021. Ainsi, la suppression de deux années 2020 − 2021 se traduira par une perte de l'information. Par conséquent, nous garderons les données de billet d'avion pour l'année 2020 et 2021.

Les principales dépenses en voyages sont les billets d'avion, les prix d'hébergements, la location de voiture et les activités. Le nouveau produit annulation d'Europ Assistance à destination du Royaume-Uni garantit le remboursement des frais liés aux billets d'avion et à l'hébergement uniquement.

Dans ce mémoire, un modèle de séries temporelles sera réalisé afin de prévoir les coûts futurs des hôtels en France. De même, un autre modèle linéaire généralisé servira à estimer le coût moyen d'un billet d'avion en Europe selon les données internes disponibles.

1. Travel trends estimates: overseas residents in the UK - Office for National Statistics (ons.gov.uk)
2. UMIH est l'union des métiers et des industries de l'hôtellerie

Par ailleurs, d'après l'ONS, plus de 82% des Britanniques passent entre 2 à 28 jours de vacances et la plupart des vacanciers voyagent pour une durée entre 5 et 14 jours en moyenne. Ainsi, la période de couverture pour la police d'annulation ira jusqu'à un maximum de 28 jours.

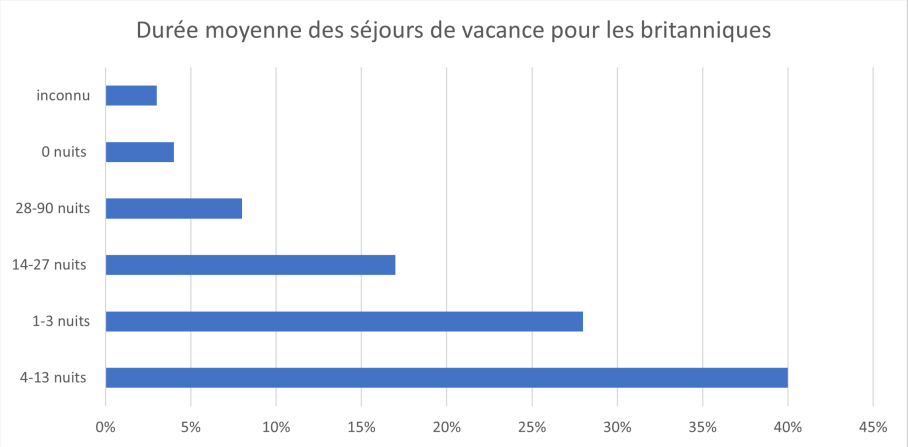


FIGURE 3.3 – Durée moyenne des séjours de vacance pour les britanniques

Pour résumer, ce mémoire s'intéresse à la tarification d'une garantie annulation de voyage pour les voyageurs britanniques qui se rendent en Europe et pour une durée maximale de 28 jours tout en garantissant le remboursement des frais d'hébergement et des billets d'avion.

Pour établir la tarification de ce nouveau produit, deux bases de données sont utilisées : La base frais d'hôtels et la base interne. La base de données sur les tarifs hôteliers provient de l'*Open Data* et servira à estimer les coûts d'hébergement. La deuxième base de données « interne » permettra de prédire la fréquence des annulations ainsi que d'estimer le coût moyen d'un billet d'avion. Ces deux bases seront présentées et analysées.

3.2 Base de données « frais d'hôtels »

Le coût du voyage est composée du montant du vol ainsi que des réservations d'hôtel. En premier lieu, une grande partie du temps est consacrée à la recherche et à la mise en place de la base de données « frais d'hôtels » à partir de l'Open Data.
Idéalement, chaque pays a sa propre base, car les prix des hôtels diffèrent d'un pays à un autre ; cependant, par manque d'information, la base « frais d'hôtels » concerne uniquement la France et par la suite les résultats seront généralisés pour les autres pays.

La France est le premier pays d'européen en matière d'*Open Data*. Selon l'indicateur adopté par l'organisation mondiale du tourisme, la France est la première destination touristique dans le monde. Sa première clientèle de touristes étrangers, comme le montre

la figure 3.4, est la Grande-Bretagne. Cette nationalité est à l'origine de 14, 1% des arrivées. Ainsi, la base « frais d'hôtel » peut refléter les dépenses d'hébergement des voyageurs britanniques.

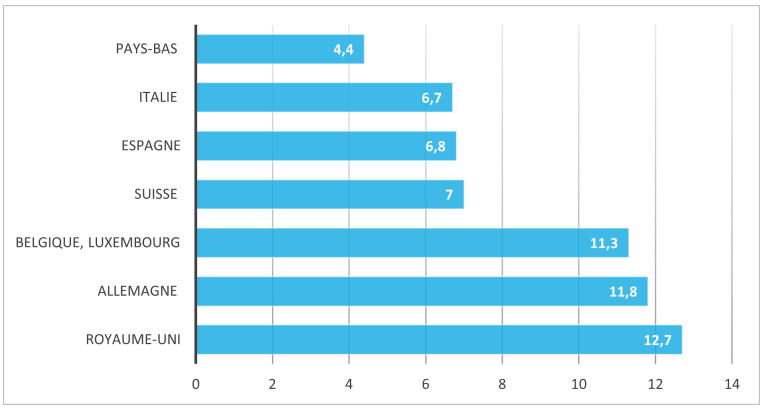


FIGURE 3.4 – Nombre (en millions) de touristes étrangers en France en 2017, par zone de résidence

Depuis 2012, l'Union des Métiers et des Industries de l'Hôtellerie (UMIH) enregistre un rendement hôtelier mensuel en France segmenté par gamme hôtelière. Elle effectue une analyse statistique quotidienne et mensuelle des parts de marché de plus de 3500 hôtels. l'UMIH réalise une analyse personnalisée des taux d'occupation ainsi que les prix moyens par nuit. Les données se présentent comme ci-dessous :

	PERFORMANCES CUMULÉES PAR SEGMENT FRANCE - DE JANVIER À OCTOBRE 2017						PERFORMANCES MENSUELLES PAR SEGMENT, FRANCE - OCTOBRE 2017					
CATÉGORIES	TO		PM HT		REVPAR HT		TO		PM HT		REVPAR HT	
	%	PTS	€	%	€	%	%	PTS	€	%	€	%
SUPER-ÉCO	68,0	2,7	43,6	-0,5	29,7	3,7	70,2	2,8	43,7	2,2	30,7	6,4
ÉCO	67,7	2,7	67,4	-1,2	45,6	2,9	70,4	2,8	69,1	0,9	48,6	5,0
MOYEN GAMME	68,7	4,0	98,4	-1,3	67,6	4,8	71,4	3,2	100,3	-0,4	71,7	4,2
HAUT DE GAMME/LUXE	72,3	4,0	195,9	-1,8	141,6	3,9	75,9	3,1	194,5	0,6	147,6	5,0
GLOBAL	68,7	3,2	89,6	-1,1	61,6	3,8	71,5	2,9	91,4	0,6	65,4	4,9

FIGURE 3.5 – Source : UMIH : Performances hôtelières mensuelles en France- Novembre 2017

Avec :

- **Taux d'occupation** (TO) : $\frac{\text{Nombre de chambres louées}}{\text{nombre de chambres disponibles pour la location}}$
- **Evolution** : soit supérieure, soit inférieure à la période de l'exercice précédent.
- **Prix moyen TTC** (PM) : $\frac{\text{Revenu d'hébergement}}{\text{Nombre de pièces louée}}$
- **Revenu moyen par chambre disponible (RevPAR TTC)** : $\frac{\text{Chiffre d'affaires hébergement}}{\text{Nombre de chambres disponibles}}$

Super-économique	1*
Économique	2*
Milieu de gamme	3*
Haut de gamme	4* et 5*

TABLE 3.1 – Les catégories d'hôtels en France

Les catégories d'hôtels sont classées ainsi :

A partir de cette base de données, nous souhaitons estimer les coûts moyens d'une nuitée dans un hôtel en France. Pour ce faire, seules les variables d'intérêt pour la construction du modèle de série temporelle seront conservées ; à savoir : le prix moyen, la date et le type d'hôtel.
Finalement, cette base compte 388 lignes dont 97 pour chaque type d'hôtel au cours de la période de 2012 – 2019.

3.2.1 Visualisation de la base

La figure 3.6 qui représente la base de données met en évidence une nette différence de prix entre les quatre gammes d'hôtels. Cette différence servira à segmenter le coût du voyage en fonction du plafond de remboursement, choisi par l'assuré.

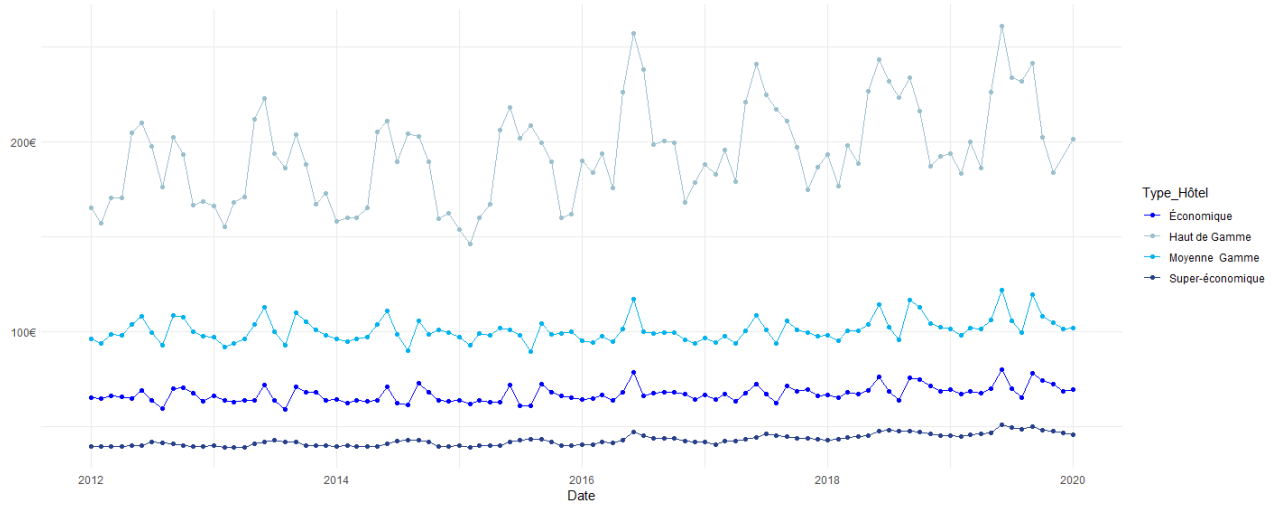


FIGURE 3.6 – Évolution des tarifs hôteliers en France

Pour plus de détails concernant les composantes et les tendances des prix, le lecteur est invité à passer à la partie 4.1.1 du chapitre 4.

3.2.2 Prise en compte de l'inflation

Les données recueillies pour les prix des hôtels vont jusqu'en janvier 2020, alors que nous souhaitons estimer les prix pour les prochains mois de 2022 et de 2023. Cet écart de quatre ans doit être pris en considération dans la modélisation. Par ailleurs, depuis juillet 2021 l'inflation a connu une forte augmentation, causée par la hausse des prix de l'énergie et des aliments, qui s'est ensuite répercutée sur les prix des services d'hébergement ainsi que les loyers.

En effet, selon une étude menée par l'INSEE et à l'observation de la figure 3.7 ci-après, l'indice de référence des loyers a augmenté de $3,6\%$ entre juillet 2021 et juillet 2022. Cette hausse de loyer entraînera sans doute une augmentation des prix des hôtels. Par conséquent, nous avons pris la décision d'ajouter un coefficient d'inflation aux prix modélisés. Ce coefficient permettra d'adapter les prix modelés à la réalité.

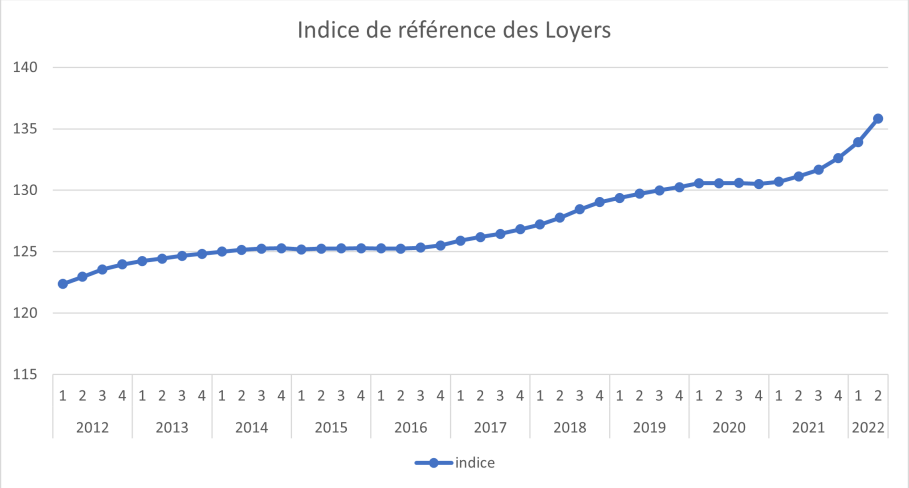


FIGURE 3.7 – Indice de référence des Loyers

$$\text{Coefficient d'inflation} = \frac{\text{Le dernier indice de référence}}{\text{L'indice de référence de 2019}}$$

3.3 Base de données « Interne » : pour le coût du vol et pour la fréquence

3.3.1 Présentation de la base

Depuis 2018, Europ Assistance collabore avec une compagnie aérienne, que l'on nommera S-Airlines, pour vendre en B to B to C des produits d'assurances voyages. Les clients qui voyagent à bord de S-Airlines et souhaitant se couvrir contre les aléas du

voyage peuvent souscrire à une assurance telle que l'assurance annulation de voyage et/ou Assistance médicale, etc.

Ici, la base de S-Airlines servira à modéliser le prix des billets d'avion ainsi que la fréquence d'annulation, pour la compagnie aérienne B-Airlines ; futur partenaire d'Europ Assistance.

Les bases de données internes d'Europ Assistance ne contiennent pas la destination du voyageur. Étant donné que la compagnie S-Airlines ne propose des voyages qu'en Europe, alors le choix de la base S-Airlines permet de s'assurer que les voyages ne concernent que le périmètre européen.

FIGURE 3.8 – Les destinations de S-Airlines

La base de modélisation S-Airlines comporte 312 437 lignes et composé de trois catégories de variables :

- Catégorie **information sur la police**, contenant des variables comme :
  - Identifiant gestionnaire de réclamations
  - Identifiant voyageur
- Catégorie **informations tarifaires**
- Catégorie **informations de modélisation**

Le tableau ci-dessous répertorie l'ensemble des informations que nous allons utiliser pour la modélisation :

Après une présentation générale de la base à utiliser, S-Airlines fera l'objet d'une analyse détaillée dans la suite de ce chapitre.

	Variable	Modalité
Informations tarifaires	Nombre de bénéficiare *(pol_ numBenef)*	1|2
	Mois de départ *(m_ dep)*	1 | 2 | 3 | ...
	Mois de réservation *(m_ subs)*	1 | 2 | 3 | ...
	Mois de retour *(m_ end)*	1 | 2 | 3 | ...
	Jours de depart *(Travel_ day_ start)*	En semaine | le week-end
	Jours de retour *(Travel_ day_ end)*	En semaine | le week-end
	Période de réservation *(booking_ week)*	0 semaine | 1 semaine | 2 semaines | ...
	Durée de voyage *(duration)*	1 nuit | 2 nuits | 3 nuits | ...
	Année de souscription (y_ subs)	2018 | 2019 | 2020 | 2021
Informations de modélisation	Montant de billets d'avion *(nv_ prix)*	en €
	Fréquence d'annulation *(Cla_ nb_ TCAN)*	0 | 1

TABLE 3.2 – Information sur les variables de modélisation

Analyse préliminaire

Les deux variables d'intérêt pour la suite de cette partie sont : montant des billets d'avion et survenance d'une annulation. Le montant des billets d'avion est le tarif que le voyageur paie sur le site de S-Airlines pour un vol aller-retour. Quant à la fréquence, il faut noter que contrairement aux produits d'assurance non-vie traditionnels (automobile, MRH, etc...), une police s'arrête une fois qu'un sinistre se déclenche. Ainsi, la fréquence individuelle ne peut être comprise qu'entre 0 et 1.
En conséquence, la fréquence des sinistres en assurance voyage est fondé sur l'observation ou pas d'un sinistre et non pas sur la quantité des sinistres.
La fréquence totale du portefeuille est définie par la formule suivante :

$$\text{Fréquence} = \frac{\text{Nombre de sinistres}}{\text{Exposition}}$$

Avec Exposition = nombre de contrat souscrit.

Informations manquantes et valeurs aberrantes

La qualité des données étant essentielle lors de la mise en œuvre d'un processus de tarification, il faut veiller à ce qu'il n'y ait pas de cas particuliers par exemple des observations loin de la réalité ou des valeurs manquantes.

Dans l'ensemble, la base de données n'a pas de valeurs manquantes. Toutefois, la figure 3.9 montre que la variable cible (montant du vol) présente $0,71\%$ d'informations manquantes. Il a été décidé de remplacer ces valeurs, selon les règles d'imputation par la moyenne.

Concernant les valeurs aberrantes, après quelques analyses de cohérence, aucun cas particulier n'a été observé, sauf pour le montant du vol. Cette variable indique qu'il existe des billets d'avion à valeur nulle. Ces cas seront traités ci-dessous.

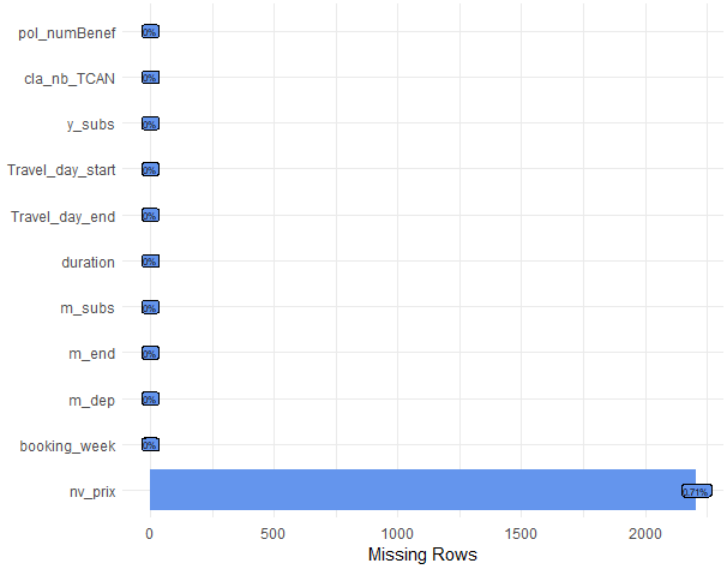


FIGURE 3.9 – Taux des valeurs manquantes par variable

Tarif très élevé/ bas pour les billets d'avion

Certaines polices présentent des montants de billets d'avion très élevés par rapport au reste des données, ceci est observable sur la figure 3.10.

Ces tarifs correspondent à ceux des sièges de première classe. Les places en première classe sont généralement limitées en nombre d'où la faible fréquence.
Ainsi, nous avons décidé de procéder à l'écrêtement des coûts extrêmes conformément à la méthode des quantiles. Cette méthode permet de déterminer un seuil de $1060,58$ € . Ainsi, les billets d'avion supérieurs à ce seuil (représentant environ 1% de la base) sont écrêtés.

L'observation de la répartition des prix des billets d'avion, montre des prix de billets très bas, voire nuls. Ces valeurs sont loin de la réalité et peuvent fausser le modèle. Nous avons décidé d'enlever les valeurs qui sont en dessous du quantile de 1 %, donc inférieures à $51,3$ €. Finalement, la base S-Airlines est composée de 306 316 lignes.

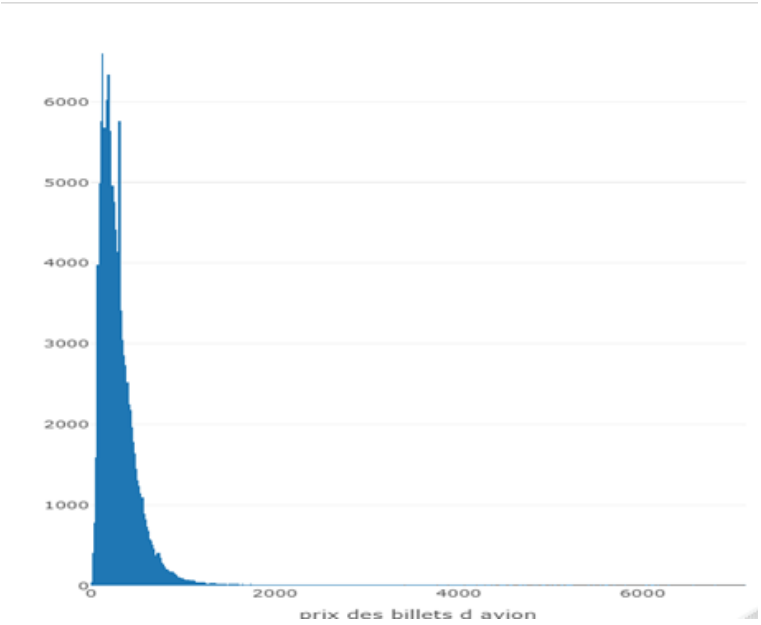


FIGURE 3.10 – Répartition des prix des billets d'avion

Faible fréquence

Les algorithmes de classification binaire, fonctionnent mieux lorsque le nombre de cas de chaque classe est à peu près égal. Lorsque le nombre d'occurrences d'une catégorie (0 : ne pas avoir d'annulation ou 1 : avoir une annulation) excède de loin celui de l'autre, des problèmes apparaissent.

La figure 3.11 montre la répartition des sinistres dans la base de données S-Airlines. Seulement $0,46\%$ des contrats ont eu un sinistre au cours de leur période de couverture tandis que $99,54\%$ n'en ont aucun. Cela montre le déséquilibre qui existe dans l'ensemble des données entre les contrats avec sinistres et ceux sans sinistres.

Toutefois vu que nous ne cherchons pas à faire une classification, mais plutôt à avoir une probabilité qu'une police ait une annulation (probabilité de défaut), cette basse fréquence n'est pas vraiment problématique pour notre modèle.

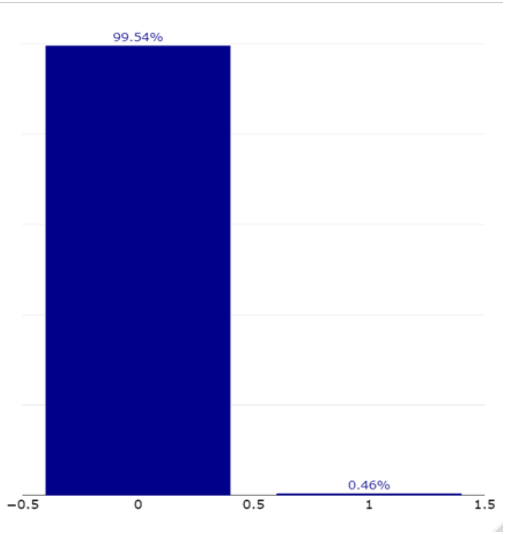


FIGURE 3.11 – Fréquence de sinistre annulation

Prise en compte de l'inflation

La prise en compte de l'inflation est importante dans la modélisation du coût des billets d'avion. En effet, l'inflation impacte directement le coût du carburant en entraînant sa hausse et cela se traduit par une augmentation du prix des billets d'avion. Il ne faut pas se baser uniquement sur les données historiques pour modéliser les prix, mais il faut aussi prendre en compte le taux d'inflation actuel.

Ainsi, pour prendre en compte l'inflation :

- L'année de réservation du voyage sera prise en compte dans le modèle de tarification
- Un coefficient d'inflation à partir de l'indice de prix des services de transport aérien communiqué par l'ONS britannique sera ajouté.

Comme la figure 3.12 le montre, l'indice de prix des services de transport aérien au Royaume-Uni a considérablement augmenté dans les deux premiers trimestre de 2022.

Le coefficient d'inflation : $\frac{\text{Indice T2 de 2022}}{\text{Indice T4 2021}}$

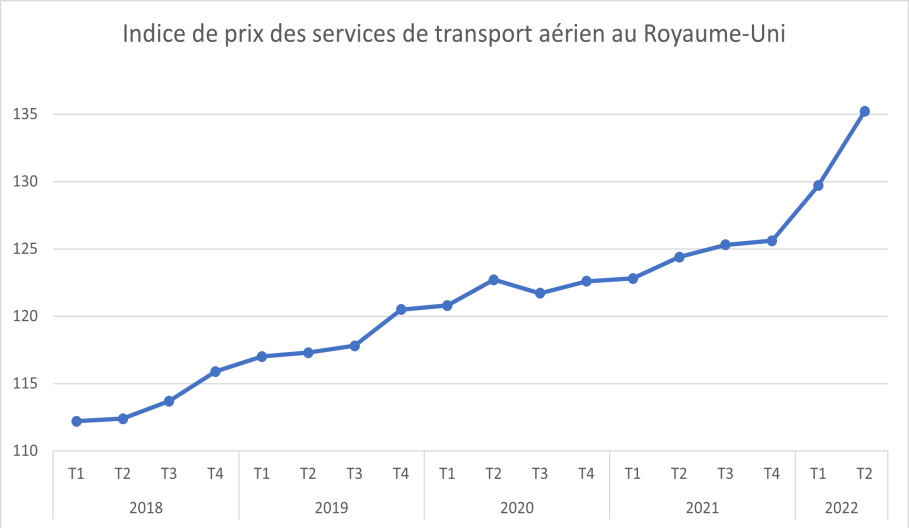


FIGURE 3.12 – Indice de prix des services de transport aérien au Royaume-Uni.

Stabilité du portefeuille

Contrairement à la base de données « frais d'hôtels », les données du portefeuille incluent la période de la COVID-19 (le portefeuille s'étale sur 2018-2021) et la période entre 2020 et 2021 est considérée comme une période atypique principalement pour l'industrie du voyage. Ainsi, il est essentiel de garantir la stabilité du portefeuille dans le temps. Pour ce faire, des observations à base de figures et de tableaux ont été réalisées.

Le tableau 3.3 ainsi que la figure 3.13 confirment que 2020 et 2021 sont deux années atypiques. En effet, nous observons une diminution de l'exposition du nombre de voyages en 2020 ainsi qu'en 2021 à cause de la fermeture des frontières. Malgré la COVID-19, le coût moyen d'un billet reste autour de la moyenne. Ceci, permet de supposer l'homogénéité du portefeuille. Quant à la diminution de la fréquence, c'est parce que les voyages touristiques ont fortement diminué pendant la période de la COVID-19.

Année de réservation	Fréquence de sinistre
2018	0,7 %
2019	0,5 %
2020	0,2 %
2021	0,2 %

TABLE 3.3 – Fréquence des sinistres par année de souscription

Les deux figures 3.14 montrent la stabilité dans le temps de la variable durée de voyage et le mois de départ. Globalement, les observations montrent que le portefeuille de S-Airlines est stable.

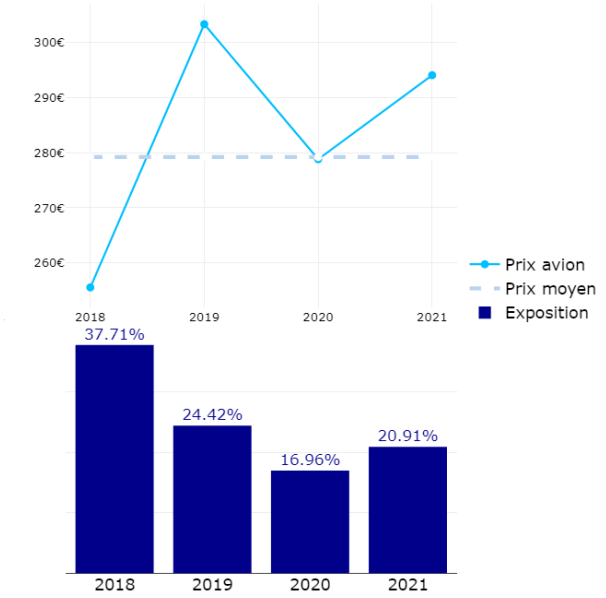


FIGURE 3.13 – Évolution du prix des billets d'avion dans le temps

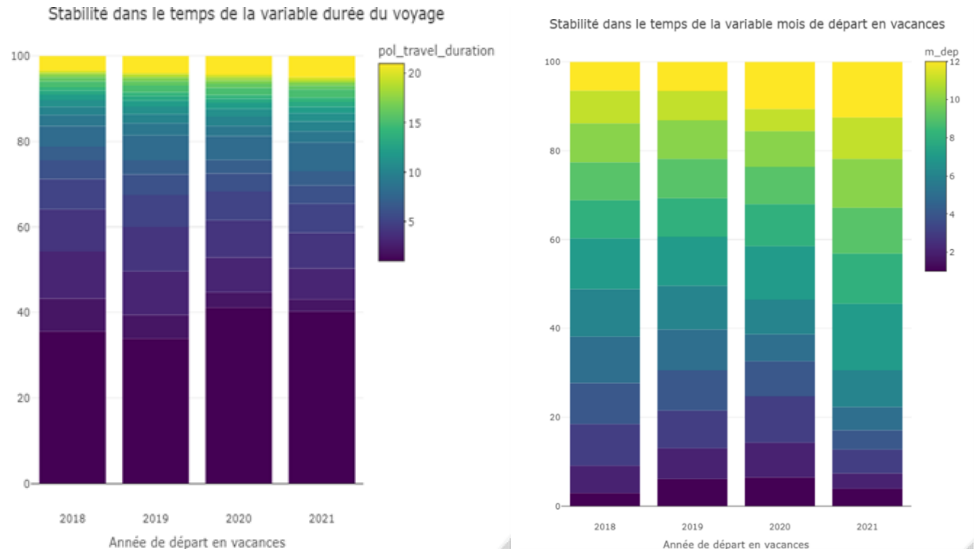


FIGURE 3.14 – Stabilité du portefeuille

En dépit des quelques effets de la pandémie sur le portefeuille, nous supposerons que généralement le portefeuille est homogène. Il est donc possible de procéder à des analyses univariées/bivariées de la base de données.

3.3.2 Analyses Univariées et bivariées

L'analyse de la base de données est l'une des étapes les plus importantes dans le processus de tarification d'un produit. En effet, elle permet de comprendre le lien entre la variable cible et les variables tarifaires ainsi que de mieux comprendre le portefeuille et les comportements des voyageurs. Par ailleurs, cette compréhension est un moyen d'expliquer aux commerciaux ainsi qu'aux partenaires la logique derrière la tarification du produit.

Analyse descriptive univariée

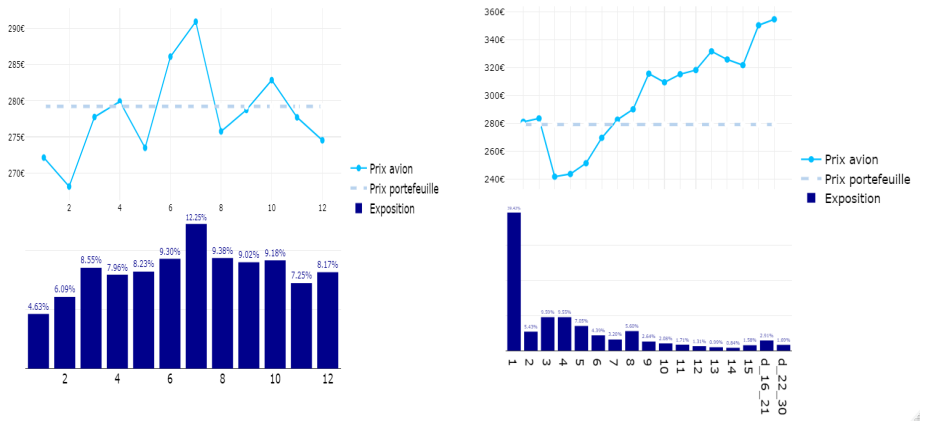
Cette partie mettra l'accent sur la variation de tarif du vol ainsi que la fréquence des sinistres en fonction des variables de la base de données. D'une manière générale, le prix d'un billet d'avion et la fréquence d'annulation dépendent principalement de la destination et de la période du voyage cependant en raison de l'absence de la variable destination de voyage, une analyse bien détaillée sera faite sur les variables qui concernent la période du voyage. Nous nous intéressons aux variables suivantes :

- mois de départ (retour) en (du) voyage ;
- jour de la semaine de départ (retour) en (du) voyage ;
- durée du voyage ;
- période de réservation : la période entre la date de la souscription à ; l'assurance et la date de départ en voyage. ;

Analyse univariée du prix moyen d'un billet d'avion

La figure 3.15a montre que le variable mois de départ, présente une variation avec deux pics significatifs en juillet et en avril, ce qui est cohérent avec les tendances du voyage. En effet, il s'agit des principaux mois de départ en vacances pour les Européens (vacances de Pâques et vacances d'été).

Quant à la durée variable du voyage, la figure 3.15b montre que plus la durée de voyage est longue, plus le prix des billets d'avion est élevé. Ceci peut s'expliquer par le fait que plus les vacances sont longues, plus la destination visitée est éloignée du pays de résidence. Ce qui peut faire vite augmenter les billets d'avion. Il est tout de même à noter qu'il y a peu d'exposition pour les voyages de plus de neuf jours.



(a) Coût moyen de billet par mois de départ

(b) Coût moyen de billet par durée de séjour

FIGURE 3.15 – Analyse univariée

La variable période de réservation influe considérablement sur les prix des billets d'avion. En effet, plus la date de départ approche, plus le coût du billet d'avion est élevé. Il existe cependant une tendance à la hausse à partir de la semaine 16, mais il s'agit d'une légère augmentation d'environ 20 €.

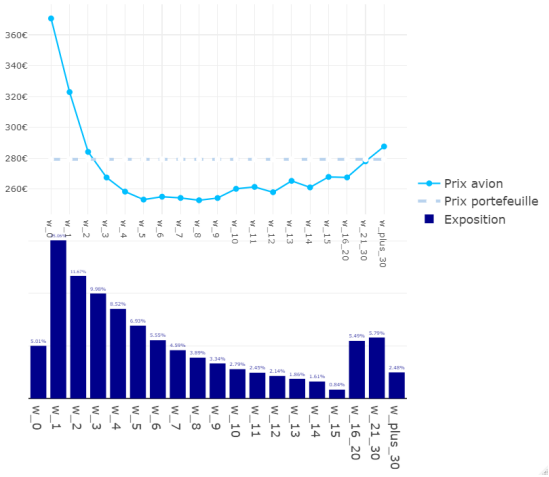
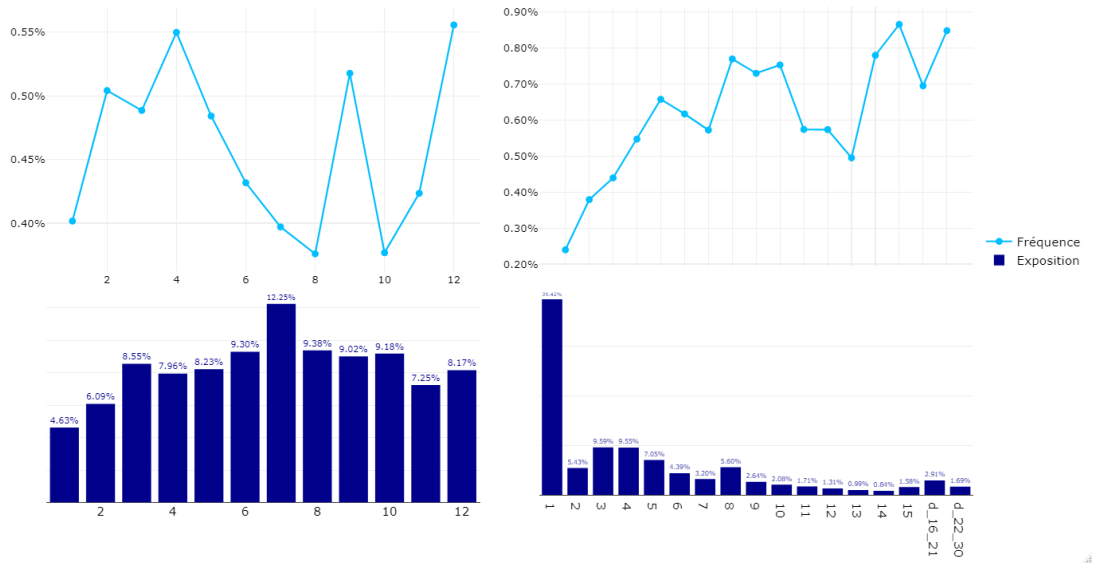


FIGURE 3.16 – Coût moyen de billet par période de réservation

Analyse univariée de la fréquence de sinistre

La même logique d'analyse s'applique à la fréquence d'annulation.

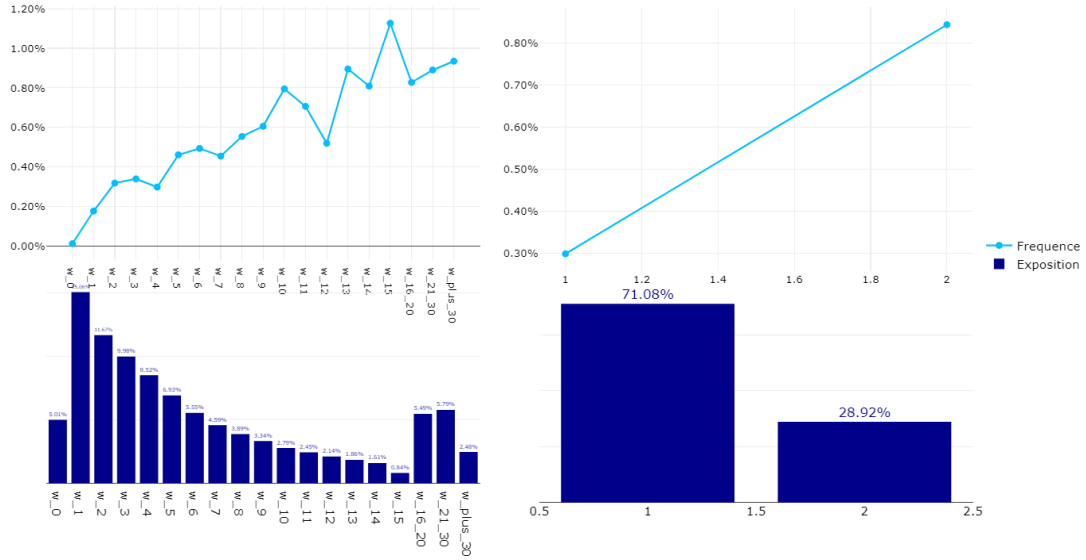


(a) Fréquence par mois de départ (b) Fréquence par durée de séjour

FIGURE 3.17 – Analyse univariée pour la fréquence

Le mois de départ ainsi que la durée du séjour exercent une influence sur la fréquence. La courbe de la fréquence en fonction du mois de départ, sur la figure 3.17b, montre que l'annulation diminue en été. Les voyageurs tombent moins malades en été et les annulations liées à la maladie diminuent. Concernant les pics d'annulation en décembre ou en avril, ceux-ci sont probablement à cause de l'augmentation du nombre des personnes malades.

La fréquence augmente en fonction de la durée de séjour. Cependant, il est à noter la faible exposition des séjours supérieurs à 15 nuits.



(a) Fréquence par période de réservation

(b) Fréquence par nombre de bénéficiaire

FIGURE 3.18 – Analyse univariée pour la fréquence

Plus une personne prend ses vacances en avances plus elle est exposée à des aléas de la vie qui peuvent causer l'annulation du voyage. Cette idée se traduit sur la figure 3.18b qui montre que la fréquence a une tendance à l'augmentation en fonction de la période de réservation.

Quant au nombre de bénéficiaires, nous observons que la fréquence d'annulation pour les polices qui couvrent deux personnes est supérieure à celles qui couvrent uniquement une personne.

Analyse descriptive bivariée

Cette analyse permet d'évaluer l'impact simultané de deux variables explicatives sur la variable à modéliser. En effet, il s'agit d'examiner l'existence d'un effet cumulatif sur la variable cible pour le prendre en compte ensuite dans la modélisation.

Différentes interactions ont été réalisé, nous allons présenter les plus intéressantes dans la suite.

Analyse bivariée du prix moyen d'un billet d'avion

Le graphe 3.19 montre la variation des prix du vol en fonction du mois de départ et le jour de départ. L'identification d'un impact bivariée se manifeste en ayant des courbes non-parallèles et qui se croisent entre elles.

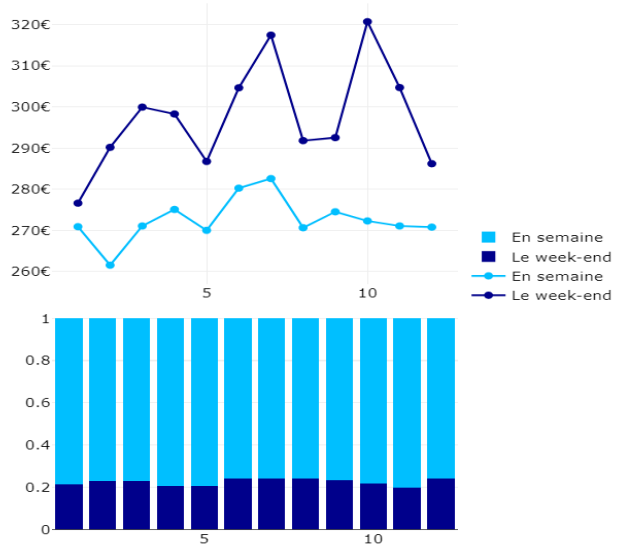


FIGURE 3.19 – Prix moyen du billet par mois de départ et jours de départ

Une première interprétation de la figure 3.19 est l'absence d'impact bivariée sur le prix du vol, car le courbe prix moyen du billet par mois de départ et jours de départ présente deux courbes assez parallèles sans interaction.

Analyse bivariée de la fréquence de sinistre

L'analyse de la figure 3.20 suggère qu'il n'y a pas d'interaction bivariée entre les variables mois de départ et durée de séjour. Majoritairement, les deux courbes sont parallèles avec quelques interactions.

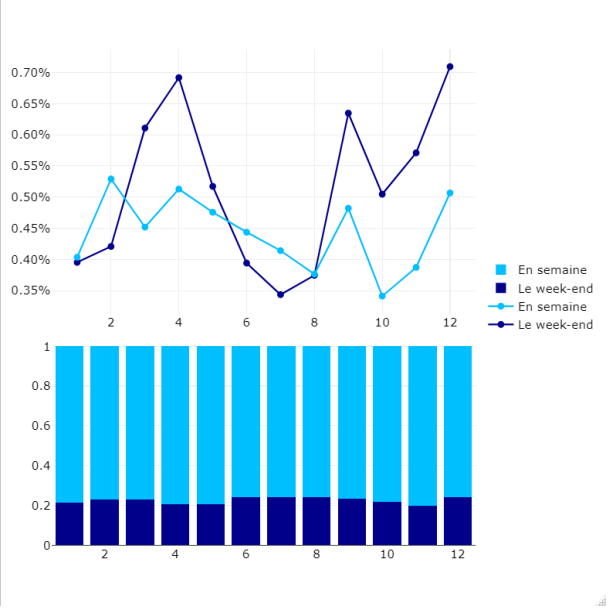


FIGURE 3.20 – Fréquence par mois de départ et jour de départ

Corrélations entre les variables

Ce chapitre s'achève par un calcul de corrélation entre les variables. Deux variables fortement corrélée peuvent impacter la performance des modèles de prédictions.
Ainsi, il est nécessaire de détecter et de résoudre tout problème de colinéarité.
Il existe différentes mesures qui consistent à détecter les corrélations entre les variables ; tels que le test de corrélation de Pearson pour les variables continues ou bien le test du χ^2 ou le V de Cramer pour les variables nominales.

Pour ce mémoire, le test du V de Cramer semble le plus adaptée pour mesurer la dépendance, étant donné qu'il prend en compte les variables catégorielles (qui composent majoritairement la base de données S-Airlines) ; de plus, ce test est connu dans la littérature par sa performance.

Théorie en bref

Le V de Cramer est obtenu à partir du test du χ^2. Ainsi, pour comprendre la théorie derrière le V de Cramer il faut avant tout présenter le test de χ^2.
Le test d'indépendance du $khi-deux$ (χ^2) détermine s'il existe une association entre des variables catégorielles. Ce test, fondé sur un tableau de contingence des deux variables étudiées, est déterminé selon la formule mathématique suivante :

$$T = \sum_{i,j} \frac{t_{i,j} - t^*_{i,j}}{t^*_{i,j}}$$

où : $t_{i,j}$: les effectifs présentant la i^{me} valeur pour la première variable et la j^{me} valeur pour la seconde variable.
$t^*_{i,j} = \frac{t_i.t_j}{n}$ avec n est l'ensemble des effectifs, t_i est l'effectif cumulé de la première variable et t_j est l'effectif cumulé de la deuxième variable. Soient Y_1 et Y_2 deux variables catégorielles. Ce test se repose sur les hypothèses suivantes :

- H_0 : Il n'existe aucun lien de dépendance entre Y_1 et Y_2
- H_1 : Il existe un lien de dépendance entre Y_1 et Y_2.

Toutefois, ce test de χ^2 indique uniquement l'indépendance ou non de deux variables sans mesurer l'intensité de l'association entre elles. Sur ce dernier point, le test de Cramer est plus pertinent parce qu'il mesure l'intensité de la relation entre deux variables tout en utilisant le test de χ^2, selon la formule suivante :

$$V = \sqrt{\frac{\mathcal{X}^2}{\mathcal{X}^2_{max}}} = \sqrt{\frac{\mathcal{X}^2}{n.min(c-1, r-1}}$$

Avec :

- $\mathcal{X}^2_{max}$ est le khi-deux maximal théorique ;
- n : l'ensemble de l'effectif ;
- c : représente le nombre de colonne du table de contingence ;
- r : correspond au nombre de lignes du table de contingence.

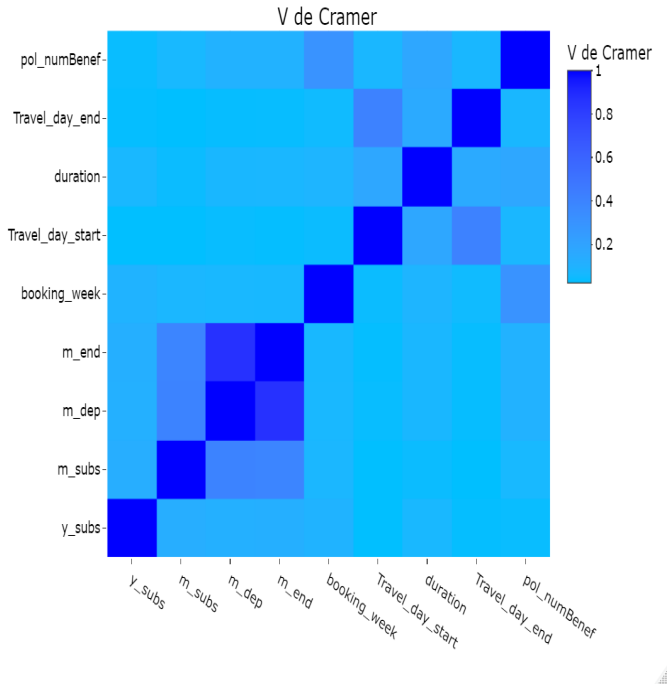


FIGURE 3.21 – V de Cramer pour la corrélation

Il ressort de la matrice de corrélation présentée sue la figure 3.21, une forte corrélation entre le mois de départ en vacances et le mois de retour de vacances. En effet compte tenu de la durée maximum du voyage dans la base de données S-Airlines qui est de 27 nuits, il est cohérent que le mois de départ et fin de voyage soient corrélés.

Ainsi, pour la suite, il a été décidé d'éliminer la variable suivante : mois de fin de voyage.

Chapitre 4

Mise en place de la tarification

Après avoir préparé et créé les bases de données, nous passerons à l'étape de tarification. L'approche de tarification utilisée est l'approche collectif fréquence-coût, déjà présenté au chapitre théorique 2.1.

La décomposition de la prime telle que $\mathbb{E}(P) = \mathbb{E}(Frq)\mathbb{E}(CM)$ vient de l'hypothèse de l'indépendance entre la fréquence et le coût moyen. Malgré quelques éventuelles corrélations, nous supposerons par la suite que la fréquence et le coût moyen sont indépendants. Ainsi :

$$\mathbb{E}(P_{annulation}) = \mathbb{E}(Frq)\mathbb{E}(prix\ avion + prix\ hotel)$$

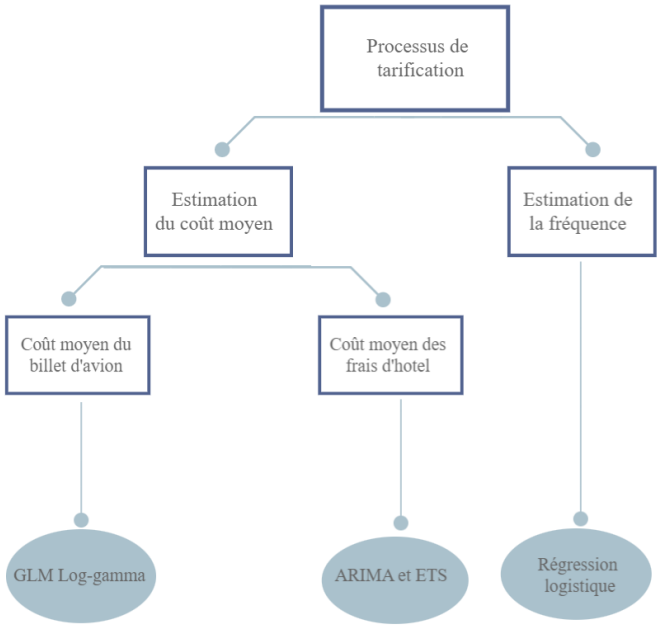


FIGURE 4.1 – Processus de tarification suivie.

Dans le présent chapitre, nous cherchons à estimer le coût moyen d'un voyage pour une police ainsi que la fréquence/probabilité d'avoir une annulation. Le coût moyen d'un voyage se compose d'un coût moyen d'hôtel plus un coût moyen de billets d'avion (aller et retour). Chacun des coûts sera modélisé au moyen d'une méthode adaptée à la base de données associée. Ainsi, le coût de l'hôtel sera modélisé à l'aide des modèles de séries temporelles, quant au prix de l'avion, il sera modélisé par un modèle linéaire généralisé. Pour ce qui est de la fréquence d'un sinistre, elle sera prédite au moyen d'une régression logistique.

4.1 Frais d'hôtel

Dans cette partie deux modèles sont mis en oeuvre : ETS et ARIMA. La comparaison entre ces deux modèles sera également abordée.

4.1.1 Modèle ARIMA

Concernant la modélisation au moyen d'une ARIMA, l'approche suivante sera utilisée :

1. Observations de la base de données et détermination de tout caractère saisonnier ou de toute tendance ;
2. Différenciation des données en cas d'absence de stationnarité ;
3. Identification des possibles modèles pour la base et choisir le meilleure parmi eux ;
4. Vérification si les résidus du modèle final sont bien des bruits blanc.

Ce processus est résumé par la figure 4.2 ci dessous.

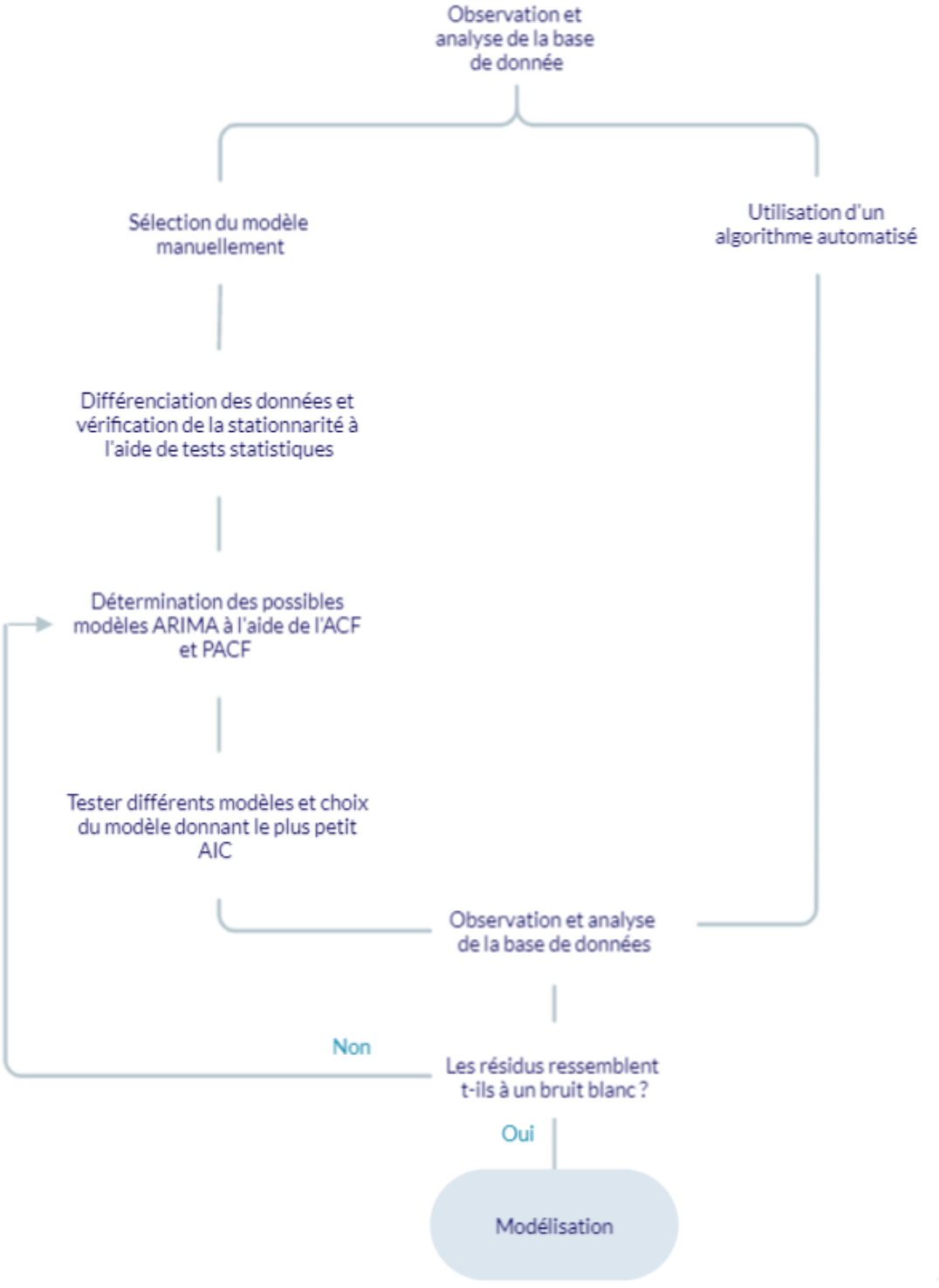


FIGURE 4.2 – Processus général de modélisation à l'aide d'un modèle ARIMA.

Première étape : Observation de la base

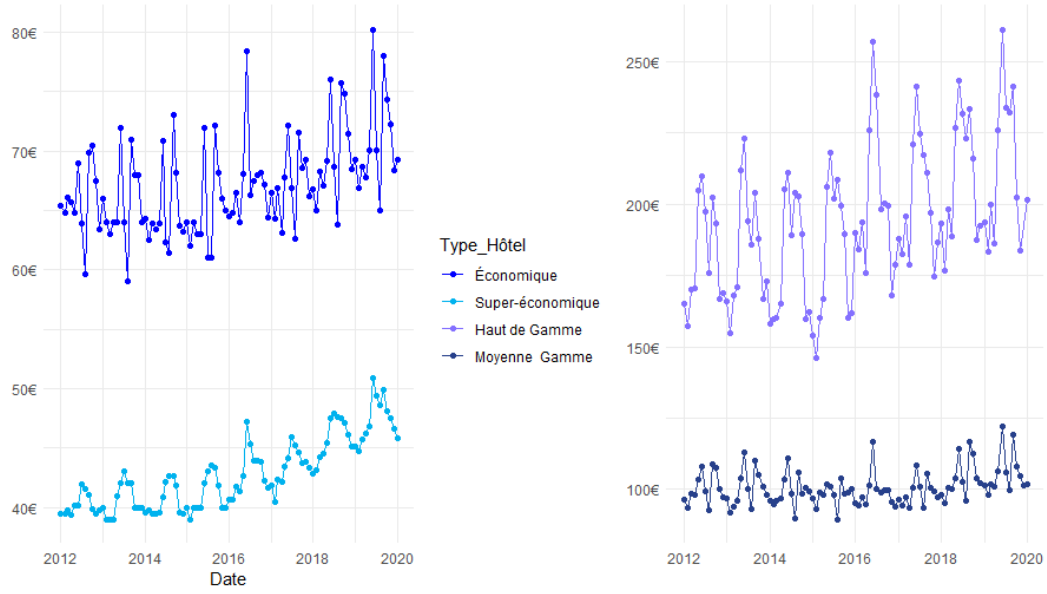


FIGURE 4.3 – Observation de la série temporelle

La figure 4.3 montre que dans l'ensemble, les prix pour une nuitée à l'hôtel ont augmenté entre 2012 et 2019. Cependant, nous observons une baisse ou une stagnation des prix en fonction du type d'hôtel entre 2014 et 2015. Cette baisse est vraisemblablement due aux attentats de janvier et du 13 novembre 2015, qui ont affecté le secteur du tourisme en France. Selon une étude réalisée par l'INSEE, les prix des hôtels ont diminué en moyenne de 1,5 % durant cette période.

Cette diminution ne durera pas parce que les prix augmentent à partir de 2016. En 2016 et particulièrement durant la période estivale, les prix des hôtels ont augmenté pour atteindre un pic considérable en juin. Cette hausse a été soutenue notamment par l'événement sportif du Championnat d'Europe de football organisé en France entre 10 juin et 10 juillet de la même année. Par la suite, les frais d'hôtel continuent d'augmenter normalement sans avoir noté une baisse ou une augmentation exceptionnelle.

Ainsi, nous soupçonnons, à partir de la figure 4.3 et l'analyse précédente, la présence d'une tendance et peut-être d'une saisonnalité, donc la non-stationnarité de la série.
Pour plus de visibilité sur la variation des séries temporelles, le lecteur est invité à consulter l'annexe A.

Test de saisonnalité : La représentation de la fonction d'autocorrélation en utilisant l'ACF permet de déterminer si une série temporelle a une composante saisonnière ou pas. Le graphique recherché est un graphique qui ressemble à des ondes sinusoïdales. Un tel

graphique indique qu'une valeur est fortement corrélée à un autre point de données dans le futur, d'où l'existence d'une saisonnalité dans l'ensemble de données.

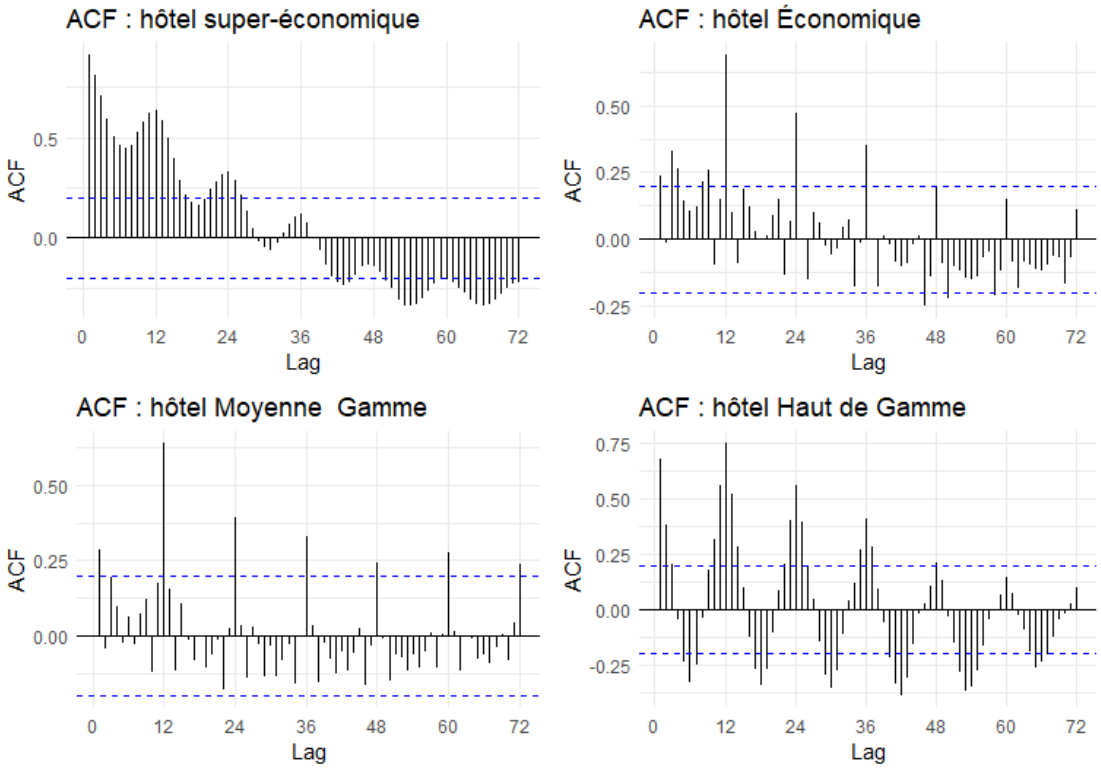


FIGURE 4.4 – Observation de la fonction d'aurocorrélation

La fonction d'autocorrélation 4.4 diffère d'un hôtel à un autre.

- Concernant les hôtels **super-économiques** : d'une manière générale, il existe une forte autocorrélation. Les coefficients de l'ACF ont une tendance sinusoïdale positive. Par ailleurs, le graphe présente des modèles alternés de décalages (Lag) positifs et négatifs, ce qui peut confirmer la présence d'une saisonnalité.
- En ce qui concerne les hôtels **haut de gamme** : la courbe ACF donne un graphique qui ressemble à des ondes sinusoïdales. Donc ces prix présentent également une saisonnalité.
- Pour les hôtels **économiques** et les hôtels **moyenne Gamme** : La tendance sinusoïdale n'est pas visualisée, mais nous pouvons noter une alternance entre des valeurs positives et négatives des coefficients de l'ACF.

Informations plus précises par le biais de la décomposition : Les figures de l'ACF 4.4 ont permis de se faire une idée sur la composante saisonnière, mais il est

difficile de tirer une conclusion concernant les hôtels économiques et moyenne gamme. Par conséquent, nous allons décomposer la base de données à l'aide de la méthode STL.

STL pour "**S**easonal and **T**rend decomposition using **L**oess" est une méthode polyvalente et robuste de décomposition des séries temporelles. Cette méthode présente de nombreux avantages, parmi lesquels elle permet de traiter tout type de saisonnalité, et non seulement des données mensuelles et trimestrielles.

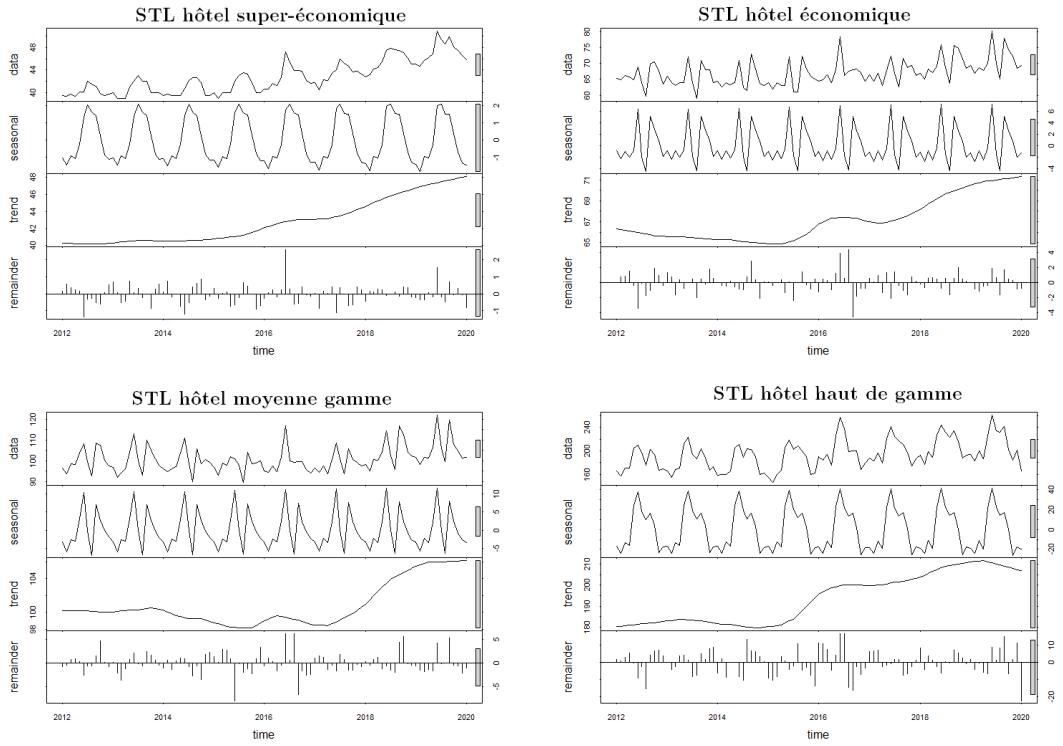


FIGURE 4.5 – Décomposition des données

À partir de la figure 4.5 il est possible de distinguer la composante saisonnière ainsi que la tendance des prix.

- Concernant les hôtels **super-économiques :** ils présentent une tendance linéaire positive avec une légère hausse en 2016 due aux événements sportifs expliqués. La composante saisonnière montre une variation sinusoïdale avec des pics pendant la période estivale.
- Concernant les hôtels **économiques :** la tendance décroît légèrement entre 2012 et 2015 et ensuite s'accentue avec un pic en 2016. En ce qui concerne la saisonnalité, elle diffère des hôtels super-économiques. Les pics de prix sont en juin et en septembre avec une baisse significatif en août. Une piste qui peut expliquer les prix du mois d'août est : les personnes qui prennent des hôtels économiques réservent leurs vacances longtemps en l'avance.

- Concernant les hôtels **moyenne gamme :** la tendance pour ce type d'hôtel est constante en 2012 et 2014 et elle diminue entre 2014 et 2017. À partir de 2018, elle part à la hausse pour stagner fin 2019, début 2020. En ce qui concerne la saisonnalité, la même chose que pour les hôtels économiques (deux pics en juin et en septembre avec une baisse pendant le mois d'août) est remarquée.
- Concernant les hôtels **haut de gamme :** Globalement la tendance est à la hausse. Toutefois, comme pour les hôtels moyenne gamme et les hôtels économiques la tendance ne varie pas considérablement entre 2012 et 2014. La saisonnalité, quant à elle, est similaire à la saisonnalité des hôtels super-économiques avec une augmentation de prix pendant la période estivale.

Deuxième étape : Différenciation des données et vérification de la stationnarité

Désaisonnalisation et différentiation : Étant donné que les données présentent un caractère saisonnier, il est essentiel de supprimer cette composante afin d'obtenir une série stationnaire. En effet, l'obtention d'une série stationnaire permettra d'estimer les paramètres du modèle ARIMA.

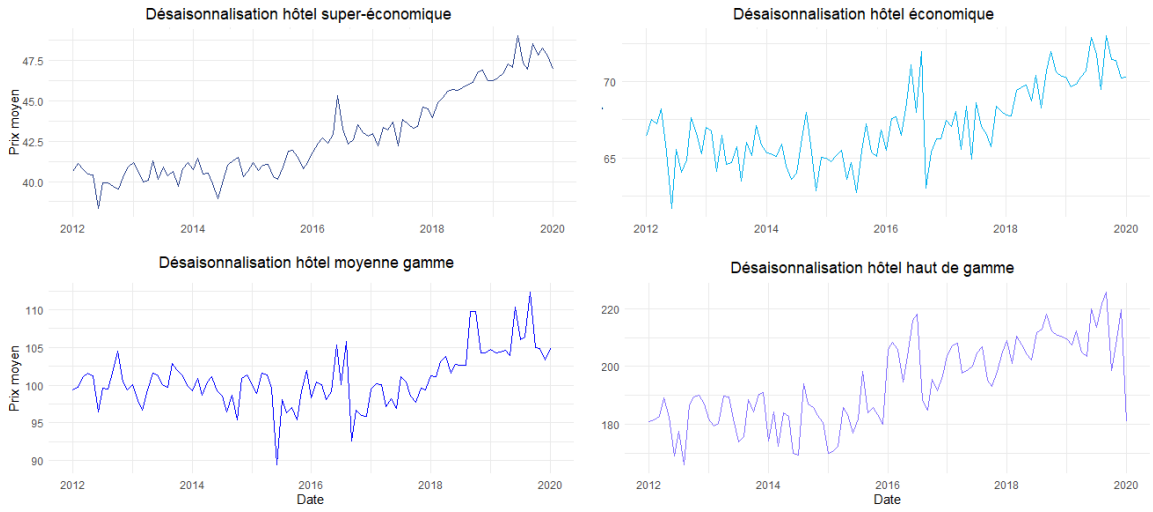


FIGURE 4.6 – Séries désaisonnalisées

Malgré la désaisonnalisation, la présentation 4.6 des séries corrigées des variations saisonnières montre l'existence d'une tendance. Il est donc essentiel de passer à la différenciation.

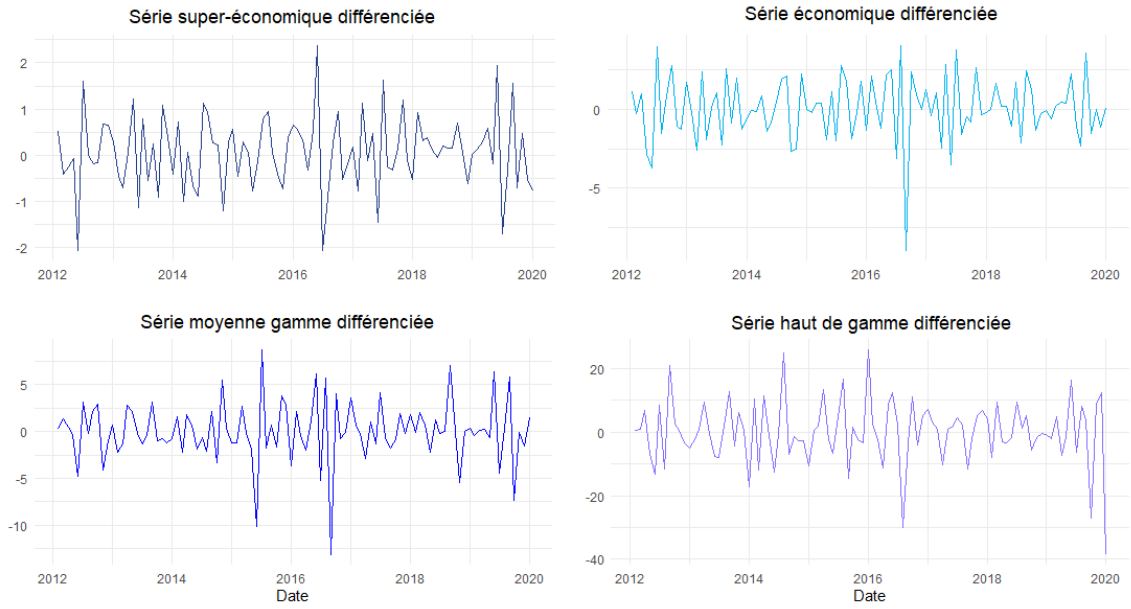


FIGURE 4.7 – Séries différenciées

Après une première différenciation et d'après la figure 4.7, les séries ne présentent plus de composante saisonnière ni tendancielle. Afin de confirmer ces observations, deux tests de stationnarité : **ADF** et **KPSS** sont réalisés.

À l'unanimité, les résultats du tableau 4.1 de deux tests permettent de conclure que la série différenciée à l'ordre 1 est stationnaire.

	Type de test	lag order	Stats de test	P_value
Super-économique	ADF	0	-13,088	0,01 <0,05
	KPSS	2	0,408	0.1>0,05
Économique	ADF	0	-16,616	0,01 <0,05
	KPSS	2	0,062	0.1>0,05
Moyenne gamme	ADF	0	-15,202	0,01 <0,05
	KPSS	2	0,039	0,1>0,05
Haut de gamme	ADF	0	-11,341	0,01 <0,05
	KPSS	2	0,088	0,1>0,05

TABLE 4.1 – Tests de stationnarité sur les séries différenciées

Troisième étape : Détermination des modèles ARIMA

Après vérification de la stationnarité, il est temps de passer à la phase de modélisation. Pour modéliser un modèle Arima, il faut tout d'abord trouver ses paramètres. La représentation de l'ACF et du PACF de la série stationnaire permet de trouver

ces potentiels paramètres. Une fois ces paramètres trouvés, plusieurs modèles sont possibles et le meilleur est celui qui donne le critère AIC le plus faible.

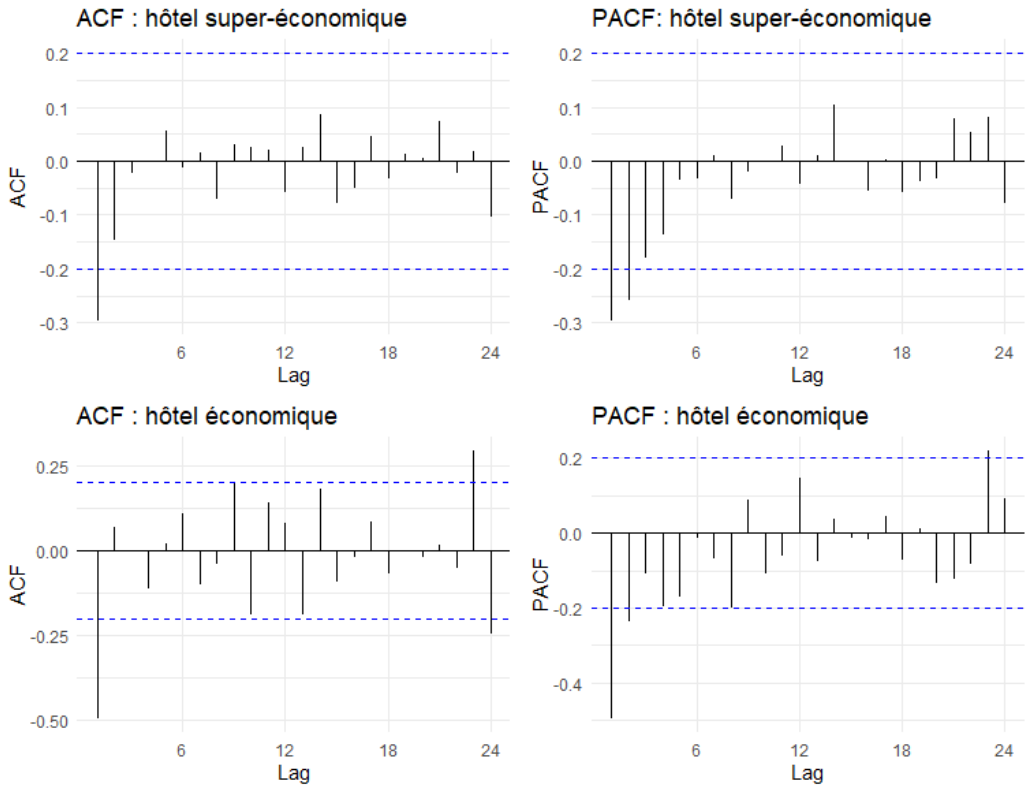


FIGURE 4.8 – Présentation de l'ACF et du PACF

L'objectif est désormais de trouver un modèle ARIMA approprié à partir de l'ACF et PACF présentés à la figure 4.8.

- pour les hôtels **super-économiques :** le pic significatif au lag 1 dans l'ACF suggère une composante MA(1) non saisonnière, ensuite aucun autre pic pour le modèle ACF n'est remarqué. Ainsi un potentiel modèle est un $ARIMA(0,1,1)(0,1,0)_{12}$. Par analogie pour le PACF, les deux pics significatifs respectivement au lag 1 et au Lag 2 suggèrent une composante AR(2) non-saisonnière, puis aucun autre pic n'est remarqué. Donc, $ARIMA(2,1,0)(0,1,0)_{12}$ est un modèle candidat.

  Les graphiques ACF et PACF des hôtels de moyenne gamme et haut de gamme sont les suivants :

- Pour les hôtels **économiques :** l'ACF avec le premier pic au lag 1 suppose un modèle MA(1) non saisonnier et le pic au lag 23 suggère une composante saisonnière de sorte que le modèle final soit $ARIMA(0,1,1)(0,1,1)_{12}$.

La représentation du PACF donne deux premiers pics proposant ainsi un modèle AR(2) non-saisonnier. De même, le pic de décalage 23 propose un AR(1) saisonnier, alors le modèle final est $ARIMA(1,1,0)(1,1,0)_{12}$.

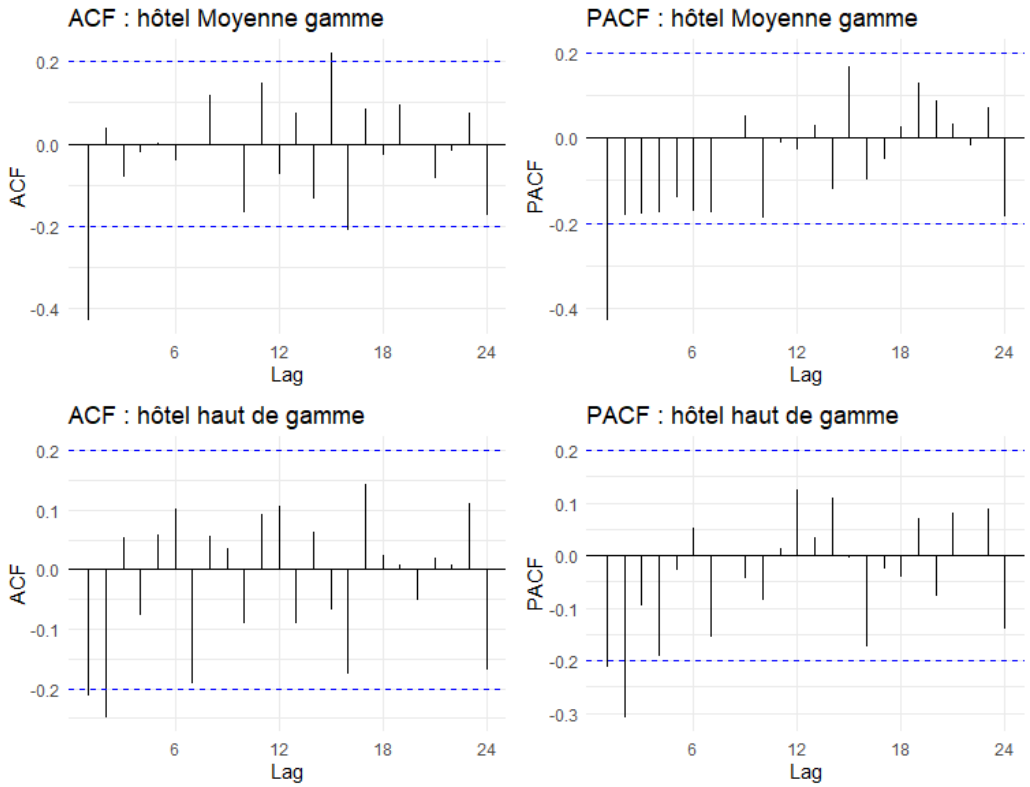


FIGURE 4.9 – Présentation de l'ACF et du PACF

Comme pour les hôtels super-économiques et économiques, une analyse du PACF et ACF est faite pour les hôtels moyenne gamme et haut de gamme.

- Hôtel **moyenne gamme :** d'après la courbe ACF, $ARIMA(0,1,1)(0,1,2)$ peut être un modèle candidat pour modéliser la série temporelle et la figure PACF propose $ARIMA(1,1,0)(0,1,0)$ comme un autre modèle potentiel pour la série.
- Hôtel **haut de gamme :** nous remarquons à partir de la figure ACF qu'un $ARIMA(0,1,2)(0,1,0)$ peut être considéré comme un modèle de la série temporelle. Tandis que le graphique PACF propose le modèle $ARIMA(2,1,0)(0,1,0)$

Comme nous avons pu le constater la sélection de l'ordre ARIMA à partir des ACF et PACF peut être jugé subjectif et difficile à appliquer. C'est pourquoi, depuis ces 25 dernières années, plusieurs études et tentatives ont été menés pour l'automatisation de la modélisation des modèles ARIMA. Finalement, plusieurs algorithmes sont disponibles à l'utilisation pour trouver le meilleur modèle ARIMA pour la série.

Modélisation en utilisant un algorithme automatisé

Pour la suite la fonction *auto.arima* sera utilisée. Cette fonction ajuste le meilleur modèle ARIMA à une série temporelle selon un critère d'information fourni (AIC, BIC ou autre). Elle effectue une recherche *stepwise* sur les ordres possibles et sélectionne les paramètres qui minimisent le critère d'information (AIC, BIC,...).

Pour la modélisation, il a été décidé de choisir comme base d'apprentissage les données entre 2012-2018 et comme base de test les données de 2019 avec janvier 2020.
Nous obtenons ainsi les résultats suivants :

	Super-économique	Economique	Moyenne gamme	Haut de gamme
Paramètre du modèle	$ARIMA(0,1,2)(0,1,1)_{12}$	$ARIMA(0,1,1)(0,1,1)_{12}$	$ARIMA(0,1,1)(2,1,0)$	$ARIMA(1,1,0)(0,1,1)_{12}$
AIC	186,4	311,04	389,48	538,21
BIC	195,46	317,83	398,53	547,31

TABLE 4.2 – Résultat de la fonction auto.arime

Quatrième étape : test de résidus

Afin de vérifier si les résidus du modèle ARIMA choisis par la fonction auto.arima sont non auto-corrélés, le test de Ljung-Box a été réalisé.
Les résultats suivants sont obtenus :

	Super-économique	Economique	Moyenne gamme	Haut de gamme
p-value de Ljung-Box test	0,976	0,441	0,442	0,329

TABLE 4.3 – p-value du Ljung-Box test

le test de Ljung-Box a donné un p-value supérieur à 5 % pour toutes les gammes d'hôtels donc l'hypothèse nulle d'absence d'autocorrélation des résidus n'est pas rejeté, ce qui confirme la stationnarité des séries.

Maintenant, que l'hypothèse de non-corrélation des résidus est vérifiée, il est possible de procéder à la modélisation de la série temporelle sur la base de 2019 et janvier 2020.

Dernière étape : prévision

Cette étape présente les RMSE obtenus pour chaque gamme d'hôtels ainsi qu'une comparaison à base de figure entre les valeurs observées et les valeurs prédites pour l'année 2019.

	Super-économique	Économique	Moyenne gamme	Haut de gamme
RMSE apprentissage	0,731	1,777	3,111	8,389
RMSE test	0,763	1,274	3,095	7,622

TABLE 4.4 – RMSE des modèles ARIMA

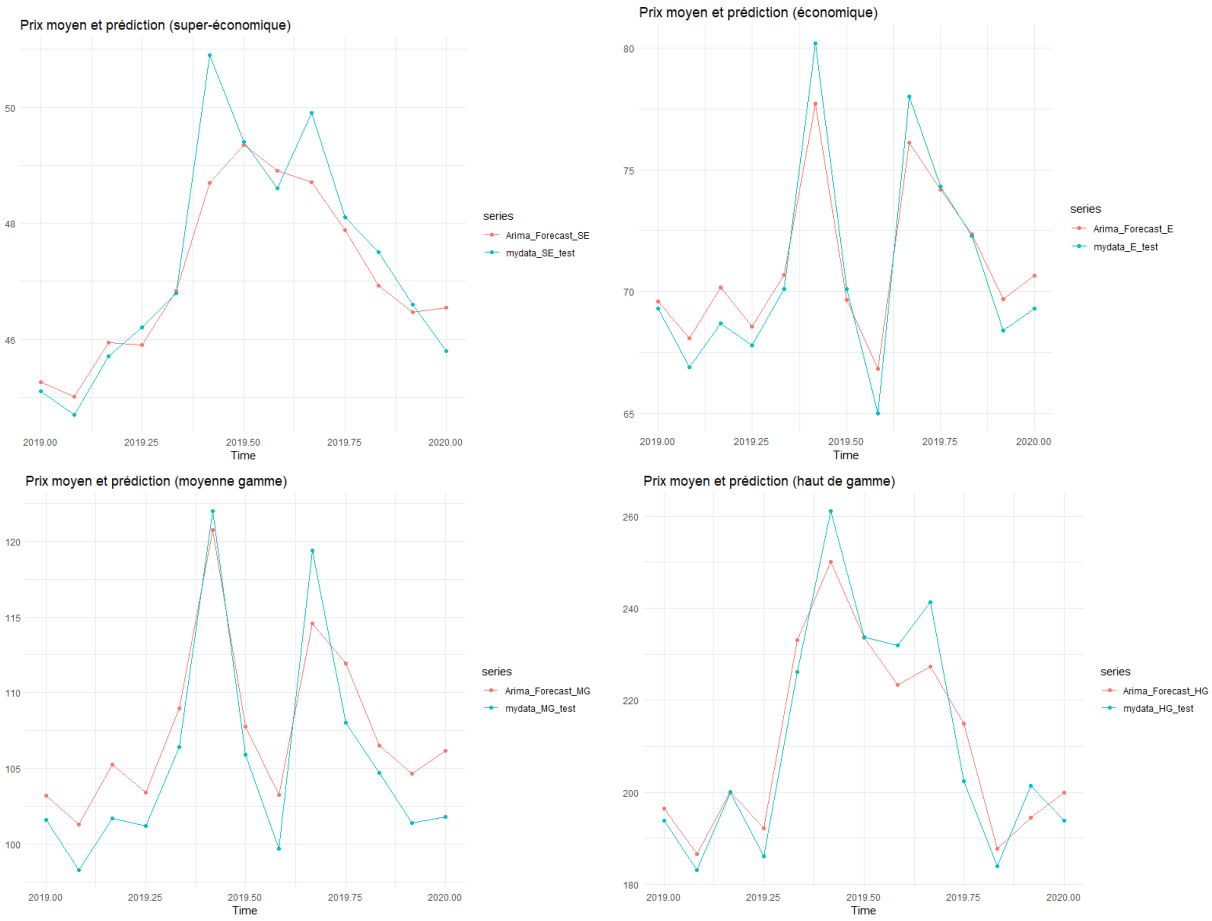


FIGURE 4.10 – Prédiction pour 2019 et début 2020

L'indice RMSE, présentée dans le tableau 4.4, augmente avec le type de l'hôtel. Le RMSE le moins élevé est celui des hôtels super-économique et le plus élèves et celui des hôtels haut de gamme. Ceci est attendu parce que les prix sont diffèrent. Les RMSE obtenus varient entre 0,7 et 8 selon le gamme de l'hôtel, ce qui peut être considéré comme étant une performance acceptable vue que les frais des hôtels varient entre 39 € et 261 €.

A partir des figures de prédictions (4.10) nous pouvons conclure :

- Pour les hôtels **super-économiques** : les prix prédits sont globalement proches des prix observés, cependant, nous constatons que la modélisation n'a pas anticipé les pics en juin et en septembre. Ce résultat est attendu, car les mois les plus chers pour les hôtels super-économiques sont généralement juillet et août.

  En outre, étant donné que janvier 2020 est un mois atypique, le modèle n'a pas prédit la baisse de prix durant ce mois. En effet, fin décembre 2019, la pandémie du coronavirus a commencé à prendre de l'ampleur dans le monde et en janvier, la Chine a confiné pour la première fois la ville, Wuhan. Ces événements ont significativement réduit le nombre de touristes en France, d'où la chute des prix.

- De la même manière, nous concluions pour les autres types d'hôtels que les valeurs prédites sont proches des valeurs observées avec une valeur prédite en janvier 2020 supérieurs à la valeur observée.

Malgré la performance du modèle ARIMA, il est tout de même intéressant de le comparer avec le modèle ETS.

4.1.2 Modèle ETS

Pour le modèle ETS, un algorithme automatisé a également été utilisé (fonction *ets* sur R). La table 4.5 et la figure 4.11 montrent respectivement les résultats de performance du modèle et la comparaison entre les valeurs prédites et les valeurs observées entre 2019 et janvier 2020.

	Super-économique	Economique	Moyenne gamme	Haut de gamme
AIC	339,880	478,455	561,280	757,656
RMSE apprentissage	0,701	1,637	2,692	8,302
RMSE test	1,391	1,372	2,769	7,673

TABLE 4.5 – Résultats du modèle ETS

Les RMSE obtenus montrent que le modèle ETS peut également être utilisé comme modèle de prédiction pour les frais d'hôtels. En effet, tout comme pour les modèles ARIMA, les RMSE varient entre 0,7 et 8 selon la gamme d'hôtel. Par ailleurs, nous constatons à partir de la figure 4.11 que les frais d'hôtels observés et les frais prédits semblent similaires. Cependant, la hausse de prix pour juin et septembre 2019 pour les hôtels économiques et moyens gamme ne semblent pas être anticipée par le modèle.

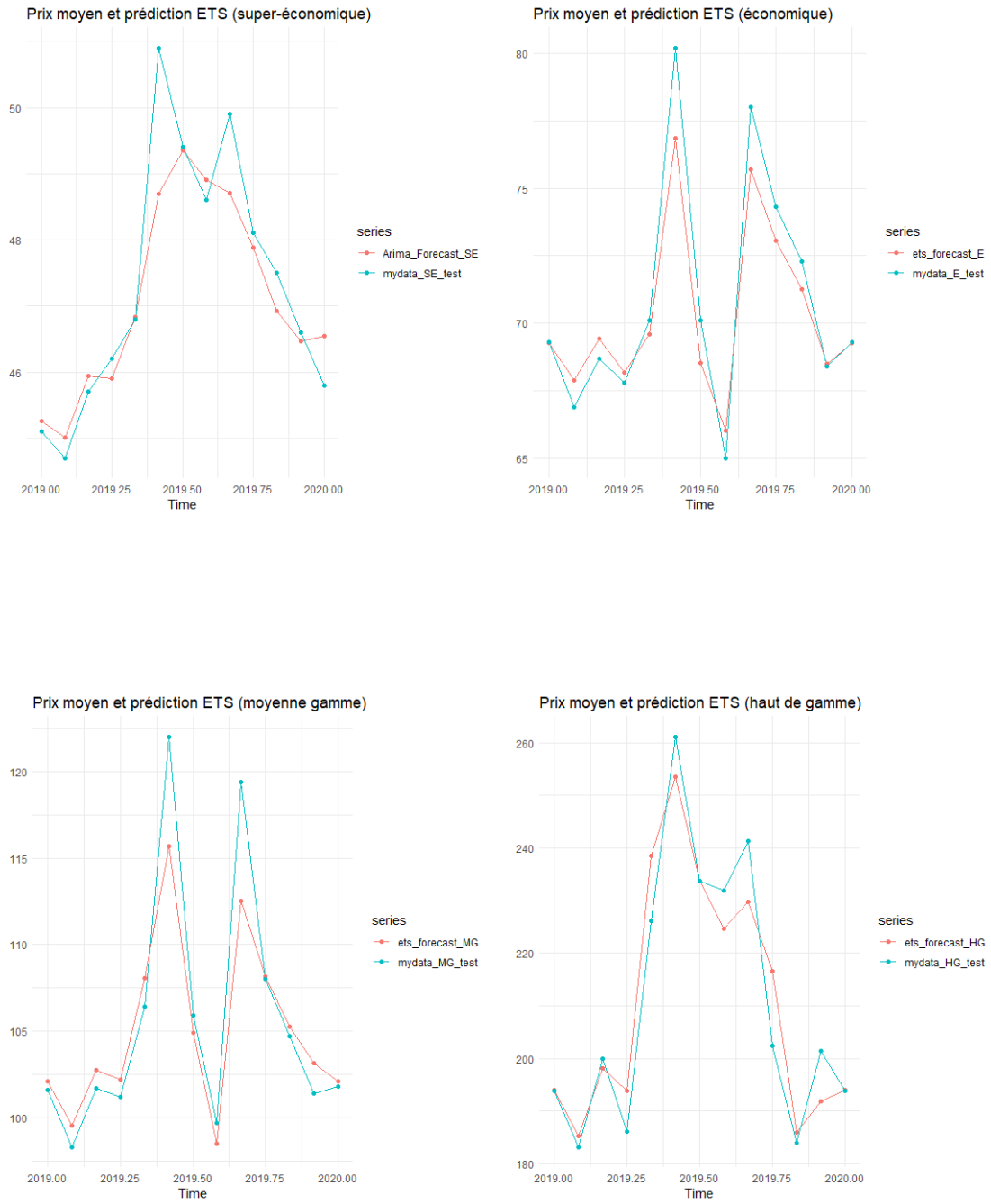


FIGURE 4.11 – Prédiction ETS pour 2019 et début 2020

L'analyse séparée de deux modèles a montré qu'ils sont tous deux performants. Dans le but de les différencier, il serait utile de les comparer.

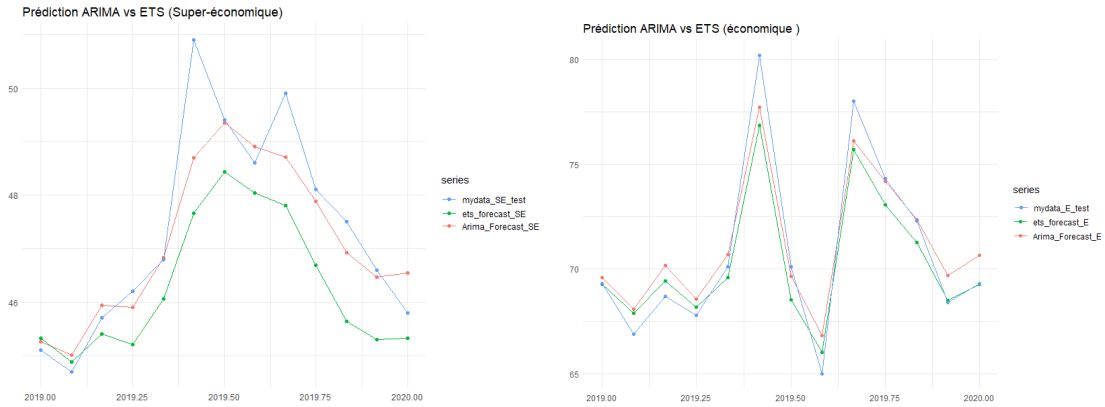
4.1.3 Comparaison entre les modèles ARIMA et ETS

Afin de comparer les deux modèles, le tableau 4.6 récapitule leurs résultats. Dans la suite, les trois indicateurs de performance suivant sont utilisés.

- $MAE(\hat{y})$
- $\text{RMSE}(\hat{y})$
- $ME(\hat{y}) = \sup_{[\![1,N]\!]} |\hat{y}_i - y_i|$ (Maximum Error)

		Super-économique	Économique	Moyenne gamme	Haut de gamme
ARIMA	MAE	0,497	1,052	2,905	6,338
	ME	0,582	1,625	9,584	58,094
	RMSE	0,763	1,274	3,095	7,622
ETS	MAE	1,102	1,021	1,812	5,909
	ME	1,935	1,895	7,670	58,878
	RMSE	1,391	1,376	2,769	7,673

TABLE 4.6 – Comparatif des résultats entre les modèles ARIMA et ETS



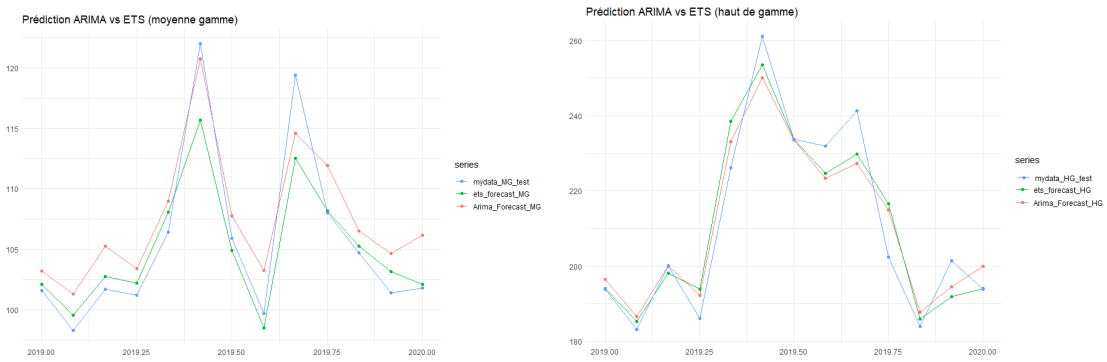


FIGURE 4.12 – Prédiction ARIMA vs ETS pour 2019 et début 2020

- Concernant les hôtels **super-économiques, économiques et haut de gamme**, les RMSE, MAE et ME obtenus à partir du modèle ARIMA sont bien inférieurs aux RMSE, MAE et ME obtenus par le modèle ETS. Le modèle ARIMA a ainsi réussi à bien mieux apprendre que les modèles ETS. Par ailleurs l'observation de la figure 4.13 montre que les estimations de l'ARIMA sont plus proches des prix observés que les estimations ETS.

- Contrairement aux résultats obtenus pour les autres catégories d'hôtels, le modèle ETS convient davantage à l'estimation des prix des hôtels **moyenne gamme** que le modèle ARIMA.

En conclusion, bien que les deux modèles soient robustes, le modèle ARIMA donne un meilleur rendement dans la plupart des cas que le modèle ETS. Ainsi, le modèle ARIMA servira par la suite pour la prédiction de frais d'hôtels.

4.2 Coût des billets d'avion

Dans cette partie, nous souhaitons estimer le coût d'un billet aller-retour pour une seule personne. La base de données contient 306 316 lignes et 7 variables explicatives catégorielles, binaires ou ordinales.

La démarche pour prédire le coût est la suivante :

1. Choix du modèle entre Log-normal et Log-gamma ;

2. Analyses des coefficients de variables ;

3. Analyse des Résultats ;

Choix du modèle

Deux modèles sont connus dans la littérature concernant la modélisation du coût moyen : le modèle Log-normal et le modèle Log-gamma. Pour sélectionner le modèle qui correspond le mieux aux données, nous avons testé l'adéquation de deux lois sur le coût moyen.

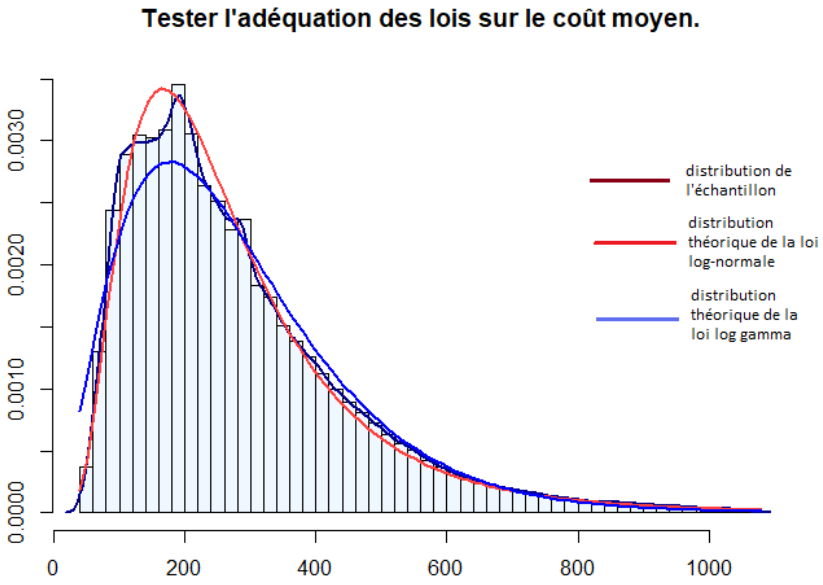


FIGURE 4.13 – Prédiction ARIMA vs ETS pour 2019 et début 2020

À première vue, la loi Log-normal est celle qui convient le mieux pour modéliser le coût. Cependant, la loi gamma semble également adaptée, donc, rien ne nous empêche de modéliser les coûts moyens avec les deux modèles afin de mieux les comparer.

Avant de passer à la modélisation, nous décidons de choisir les variables les plus significatives. Cette sélection se fait à travers la méthode *stepwise*. Nous constatons que l'algorithme a conservé toutes les variables suivantes :

- Période de réservation
- Mois de départ
- Mois de réservation
- Durée du voyage
- Jours de départ

- Jours de retour
- Année de souscription

Les résultats obtenus pour le GLM log-normal et le GLM Log-gamma sont résumés dans le tableau ci-dessous :

	GLM Log-normal	GLM Log-gamma
AIC	2 790 622	2 711 008
RMSE	161,328	161,263

TABLE 4.7 – Comparaison entre GLM log-normal et GLM Log-gamma

Bien que les résultats de deux modèles soient proches, le modèle Log-gamma parait plus performant que le modèle Log-normal. Par conséquent, nous décidons de poursuivre la tarification avec un GLM Log-gamma.

Coefficients de variables

Nous présentons dans cette partie les coefficients de certaines variables obtenus au moyen d'un GLM Log-gamma.

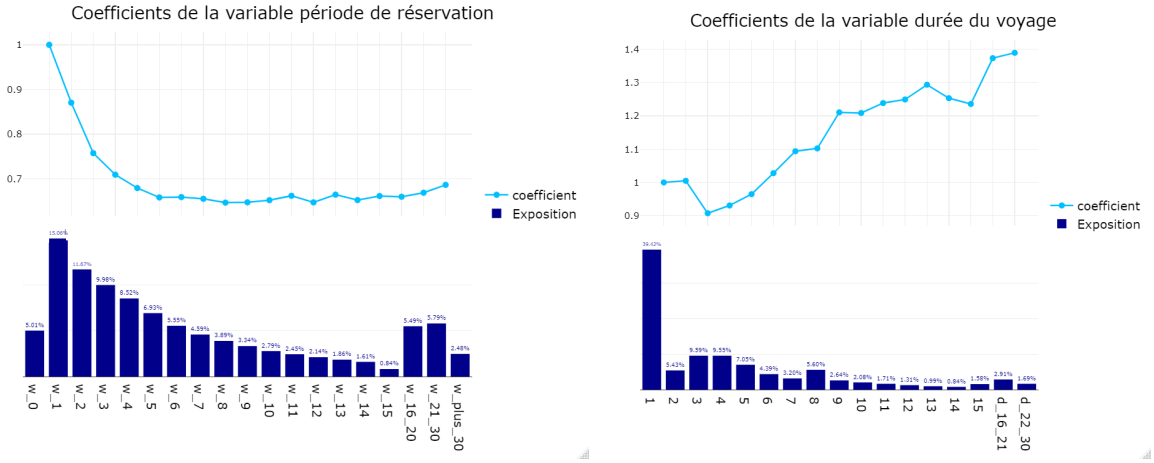


FIGURE 4.14 – Coefficients GLM pour la période de réservation et la durée du voyage

Toutes choses égales par ailleurs, plus la période entre la réservation et la date du voyage est courte, plus la prime sera élevée donc c'est cohérent avec les analyses effectuées à la partie (3.3.1).
Par ailleurs, plus le voyage est de longue durée plus la prime sera élevé.

Analyse des Résultats

Cette partie portera sur les résultats du modèle GLM obtenus. Pour réaliser la modélisation, nous avons divisé la base de données en base d'apprentissage et en base

de validation. La base d'apprentissage représente 70 % des données et l'autre 30 % correspondent à la base de validation.

	Base d'apprentissage			Base de validation		
	Observé	Prédiction	Erreur (%)	Observé	Prédiction	Erreur (%)
Coût moyen	279,413	279,5588	+0,0521 %	278,768	279,655	+3,0181 %
Coût total	59 912 046	59 943 280	+0,0521 %	25 617 412	25 698 926	+0,3181 %

TABLE 4.8 – Résultats obtenus avec le GLM

	Base d'apprentissage	Base de validation
RMSE	162,176	161,263

TABLE 4.9 – RMSE

Globalement, les résultats des tests effectués sur la base de l'apprentissage ainsi que sur la base de validation (cf.Table 4.8 et Table4.9), montrent que le modèle GLM obtenu peut être utilisé pour estimer les prix des billets d'avion.

Pour examiner un peu plus en détail la modélisation, nous avons comparé les prix observés ainsi que les prix prédits pour chacune des variables tarifaires, ci-dessous quelques exemples :

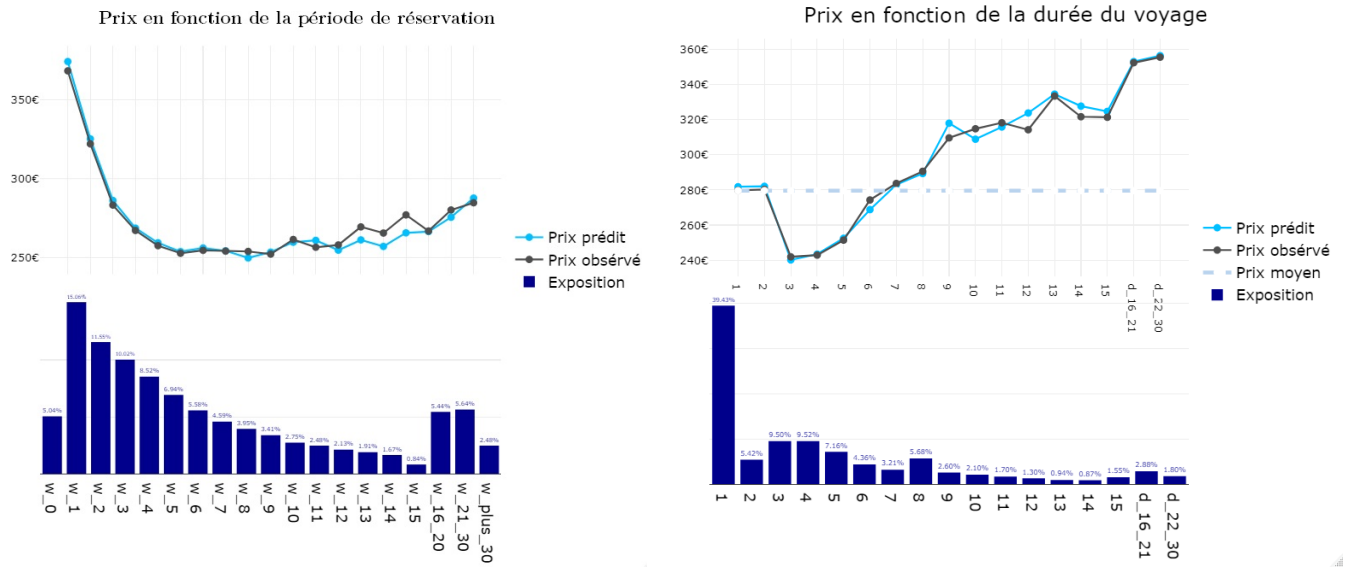


FIGURE 4.15 – Comparaison entre les valeurs observées et les valeurs prédites de la variable période de réservation et la variable durée du voyage.

La figure 4.15 montre que les coûts moyens prédits par le GLM Log-gamma suivent la tendance des coûts moyens observée sur la base de validation. Par ailleurs, les valeurs de la

prédiction sont proches des valeurs observées sans pour autant remarquer un phénomène de surapprentissage.

En conclusion, pour calculer le coût moyen d'un voyage, deux modèles sont retenus :

- ARIMA pour le coût moyen de frais d'hôtel ;
- GLM Log-gamma pour le coût moyen de l'avion

Ainsi, pour finir le processus de tarification, nous passons à la modélisation de la fréquence d'annulation.

4.3 Fréquence d'annulation

Cette partie présente les résultats obtenus à partir de la régression logistique. Au départ, les variables tarifaires sont 7 et la méthode *stepwise* de sélection de variables en a retenu seulement quatre, qui sont :

- Période de réservation ;
- Mois de départ ;
- Durée du voyage ;
- Nombre de bénéficiaire.

La sélection des variables a réduit l'indice AIC du modèle de 11 915 à 11 898. L'intervalle de prédiction obtenu par la régression logistique donne les résultats suivants :

	Minimum	Maximum	Moyenne	Médiane
Plage de prédiction	0,0001	0,0342	0,0045	0,0033

TABLE 4.10 – Résultats de la prédiction sur la base de validation

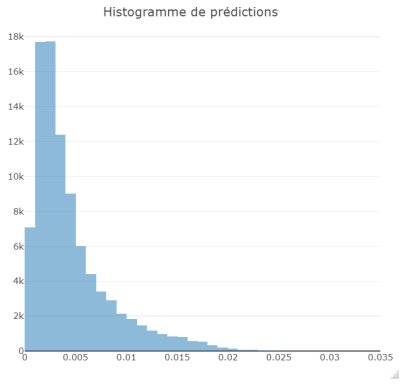


FIGURE 4.16 – Plage de prédictions

Le résultat prédit par le modèle correspond à la probabilité qu'une personne annule son voyage et c'est équivalent à la fréquence d'annulation pour une police. Les fréquences obtenues pour la base de validation sont entre 0,0001 et 0,0342. Ces fréquences sont alignées avec nos attentes. En outre, le fait que la moyenne obtenue se rapproche de la fréquence d'annulation du portefeuille confirme la bonne performance de notre modèle.

Afin de confirmer nos intuitions quant à la robustesse du modèle, il a été décidé de mesurer sa précision par le test AUC.

	Base d'apprentissage	Base de validation
AUC	0,720	0,694

TABLE 4.11 – Test de l'AUC

L'AUC (cf.Table 4.11) obtenus pour les deux bases sont supérieurs à 0,6 ce qui donne un modèle relativement acceptable. Dans le même sens que d'étudier la robustesse du modèle nous avons comparé la fréquence prédite et la fréquence observée pour les différentes variables tarifaires en voici deux exemples :

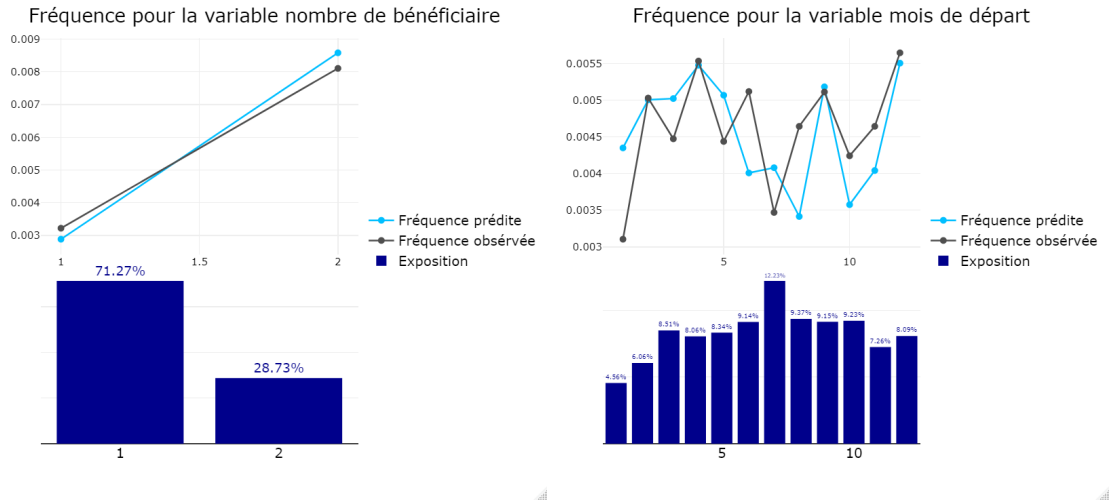


FIGURE 4.17 – Comparaison entre la fréquence observée et la fréquence prédite pour la variable nombre de bénéficiare et mois de départ.

Les fréquences estimées et les fréquences observées pour la variable nombre de bénéficiaires ne sont pas éloignées. Cependant, quelques différences sont remarquées pour le mois de départ. Dans l'ensemble, le modèle modèle reste un outil qui peut être utilisé pour prédire la fréquence.

Finalement, afin de comprendre la variation de la fréquence en fonction des variables de la police, les coefficients du modèle ont été observé, en voici quelques exemples :

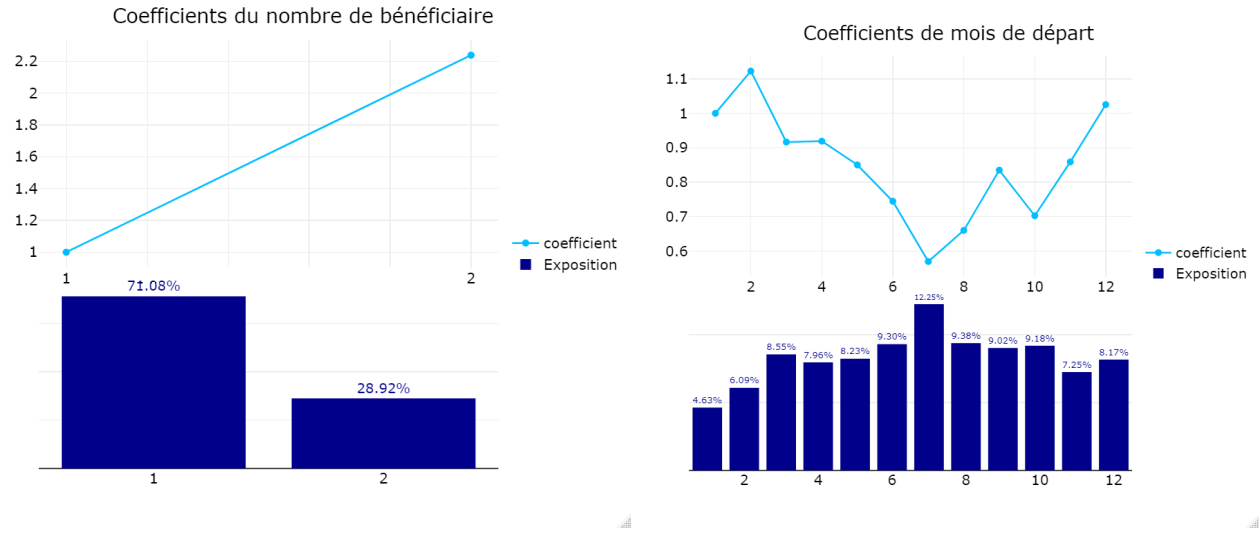


FIGURE 4.18 – Coefficient des variables nombre de bénéficiaire et mois de départ

Toutes choses égales par ailleurs, les groupes de personnes voyageant ensemble présentent un risque d'annulation 2,2 fois plus élevé que les voyageurs seuls et le risque d'annulation en été est moins important qu'en hiver, ce qui reste en ligne avec nos analyses dans la partie analyse de données.

4.4 Application du modèle de tarification à certains profils de voyageurs

Ce chapitre s'achève par le calcul de la prime commerciale de certains profils de voyageurs pour ensuite la comparer aux prix proposés sur le marché de l'assurance voyage au Royaume-Uni.

Les produits d'assurance voyage proposés par la plupart des assureurs au Royaume-Uni combinent l'assistance et l'assurance annulation. Nous allons donc calculer la prime pure d'annulation obtenue par le biais des modèles que nous avons présentés, puis nous ajouterons la prime pure d'assistance calculée par l'équipe d'Europe Assistance. Enfin, nous obtenons la prime commerciale en prenant en compte les commissions et les charges tel que les frais d'acquisitions, les taxes les frais de gestions de sinistres, ... Ces hypothèses de chargements, de commissions, de frais, etc sont des hypothèses de tarification utilisées par l'équipe souscription et tarification d'assurance voyage et par soucis de confidentialités elles ne seront pas communiquées dans ce mémoire.
La formule de prime commerciale est :

$$\text{Prime pure d'annulation} = \text{Fréquence} \times \text{Coût moyen d'un voyage}$$

Coût moyen d'un voyage = (prix d'avion×nombre de voyageur+prix d'hôtel×durée de voyage)

$$\text{Prime pure totale} = \text{Prime pure d'annulation} + \text{Prime pure d'assistance}$$

$$\text{Prime commerciale} = \frac{\text{Prime pure totale} \times (1 + \text{chargement})}{1 - \text{coût d'acquisition}}$$

4.4.1 Processus de tarification

Nous avons décidé de calculer la prime pure de divers profils de voyageurs en variant les détails du voyage comme le montre le tableau 4.12 ci-dessous et nous obtenons 4 profils différents.

Détails du voyage			
Date de départ	Durée du séjour	Nombre de bénéficiare	Plafond de couverture
- 9 Septembre 2022 - 11 Février 2023	- 2 nuits - 14 jours	- 1	- £500 - £1 000 -£2 000 -£3 000

TABLE 4.12 – Tableau présentant les différentes valeurs prises par las variables de tarification

Le processus de tarification est le suivant : selon la date de départ, la durée du séjour, nous allons estimer le coût des billets d'avion ainsi que la probabilité d'avoir un sinistre. Puis, selon le plafond de couverture choisi par le voyageur, le coût moyen d'une nuit d'hôtel multipliée par la durée du séjour sera additionné, pour finalement obtenir tous les éléments afin de calculer la prime pure de la garantie annulation.

La figure 4.19 présente le schéma de choix du type d'hôtel selon le plafond de couverture :

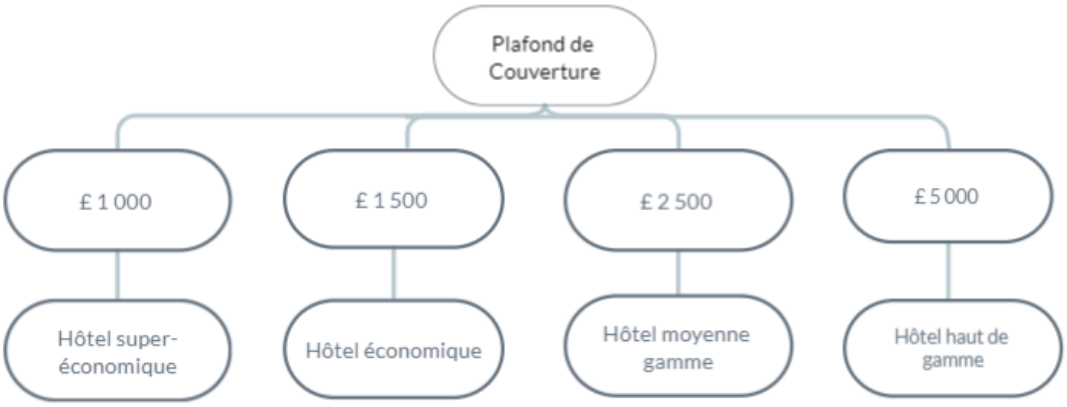


FIGURE 4.19 – Plafond de couverture et type d'hôtel correspondant

Après avoir déterminé cette prime pure, nous rajouterons la prime pure de l'assistance voyage, cette prime dépend majoritairement de la durée du voyage et par soucis de confidentialité, aucun de détail ne sera communiqué quant au calcul de cette prime. Enfin, nous obtenons la prime commerciale en prenant en compte tous les frais et les chargements.

4.4.2 Résultats et comparaison

Les tableaux 4.13, 4.14 et 4.15 ci-dessous montrent respectivement les résultats obtenus pour le tarif moyen par nuitée pour chaque type d'hôtel, le coût des billets d'avion (aller-retour) ainsi que la fréquence d'annulation par profil et finalement la prime pure pour chaque police. Les coûts et les primes seront par la suite présentée en livres sterling (£).

	Prix moyen par nuit			
	Hotel super-économique	Hotel économique	Hotel moyenne gamme	Hotel haut de gamme
Septembre 2022	£55	£82	£129	£256
Février 2023	£48	£71	£104	£179

TABLE 4.13 – Tarif moyen par nuitée pour chaque type d'hôtel.

	Période de réservation	Mois de départ	Mois de souscription	Durée	Jour de départ	Jour de retour	**Prix des billets d'avion**	**Fréquence**
Profil_1	w_4	9	8	14	semaine	semaine	**£344**	**0,3 %**
Profil_2	w_4	9	8	2	semaine	semaine	**£274**	**0,29%**
Profil_3	w_25	2	8	14	week-end	week-end	**£315**	**0,9%**
Profil_4	w_25	2	8	2	week-end	week-end	**£251**	**0,7%**

TABLE 4.14 – Coût moyen des billets d'avion et fréquence d'annulation

	Prime pure			
	Plafond £1000	Plafond £1500	Plafond £2500	Hotel Plafond £5000
Profil_1 : 9 Septembre 2022 pour 14 nuits	£3,3	£4,5	£6,4	£11,8
Profil_2 : 9 Septembre 2022 pour 2 nuits	£1,1	£1,3	£1,5	£2,3
Profil_3 : 11 Février 2023 pour 14 nuits	£9,8	£13,2	£19,1	£35,1
Profil_4 : 11 Février 2023 pour 2 nuits	£2,5	£2,9	£3,6	£5,3

TABLE 4.15 – Prime pure obtenue

Après l'obtention de la prime pure de la garantie annulation, il est possible de calculer la prime commerciale d'un contrat d'assurance voyage garantissant l'annulation ainsi que l'assistance au voyage. Puis la comparer aux différents tarifs du marché.

Il est tout de même important noter que la prime d'assistance est plus élevé que la prime d'annulation. Par exemple,une prime commerciale d'assistance est aux alentours de £16 pour une durée moyenne de 14 nuits. Ceci explique la différence de prix entre les primes pures de la garantie annulation et les primes commerciales finales (cf.Table 4.16).

Compagnie d'assurance	Profil	Plafond de couverture	Prime commerciale	Notre prime commerciale
Allianz	Profil_1	£1 000	**£18,35**	**£23,05**
		£2 500	**£21,67**	**£25,87**
		£5 000	**£27,53**	**£30,2**
	Profil_2	£1 000	**£4,93**	**£5,7**
		£2 500	**£6,94**	**£6,23**
		£5 000	**£10,84**	**£6,69**
	Profil_3	£1 000	**£17,95**	**£37,95**
		£2 500	**£21,27**	**£45,69**
		£5 000	**£27,13**	**£59,14**
	Profil_4	£1 000	**£4,91**	**£8,9**
		£2 500	**£6,92**	**£9,88**
		£5 000	**£10,82**	**£11,47**
Esure	Profil_1	£1 000	**£26,12**	**£23,05**
		£2 500	**£19,59**	**£25,87**
		£5 000	**£17,83**	**£30,2**
	Profil_2	£1 000	**£4,88**	**£5,7**
		£2 500	**£5,32**	**£6,23**
		£5 000	**£6,99**	**£6,69**
	Profil_3	£1 000	**£43,28**	**£37,95**
		£2 500	**£47,55**	**£45,69**
		£5 000	**£63,73**	**£59,14**
	Profil_4	£1 000	**£11,34**	**£8,9**
		£2 500	**£12,42**	**£9,88**
		£5 000	**£16,52**	**£11,47**
Oasis assurance	Profil_1	£1 000	**£18,75**	**£23,05**
		£2 500	**£30**	**£25,87**
		£5 000	**£41,25**	**£30,2**
	Profil_2	£1 000	**£7**	**£5,7**
		£2 500	**£11,5**	**£6,23**
		£5 000	**£15,5**	**£6,69**
	Profil_3	£1 000	**£35,5**	**£37,95**
		£2 500	**£52**	**£45,69**
		£5 000	**£73**	**£59,14**
	Profil_4	£1 000	**£12,75**	**£8,9**
		£2 500	**£19,25**	**£9,88**
		£5 000	**£26,75**	**£11,47**

TABLE 4.16 – Comparaison des primes d'assurance voyage au Royaume-Uni avec celles d'Europ Assistance.

A partir du tableau 4.16, nous constatons que nos primes commerciales se situent principalement dans la deuxième ou dans la troisième position des primes les moins coûteuses.
En effet,

- Pour le **Profil_1** :
  - Pour ce qui est de la couverture de **£1 000** : nous sommes moins cher qu'Esure mais plus coûteux que Allianz et Oasis assurance.
  - Concernant la couverture de **£2 500** et **£5 000** : nos primes sont inférieures à celles de l'Oasis et plus cher qu'Allianz et Esure.

- En ce qui concerne le **Profil_2**
  - Pour la couverture de **£1 000** notre tarif est inférieur à celui de l'assurance Oasis, mais plus élevé que les deux autres.
  - A propos de la couverture de **£2 500** la prime commerciale est inférieure à celle d'Allianz et d'Oasis assurance, mais nous restons plus chers qu'Esure.
  - Pour la couverture de **£5 000** le tarif proposé par Europ Assistance est le moins coûteux de tous les tarifs proposés par les autres compagnies d'assurance voyage.
- Concernant le **Profil_3**
  - Pour la couverture de **£1 000** la prime proposé par Europ Assitance est moins chers que la prime de la compagnie d'assurance Esure mais plus chers que Allianz et Oasis assurance.
  - Pour la couverture de **£2 500** et **£5 000** la prime commerciale obtenue est supérieure à celle d'Allianz mais reste inférieure à celle des assurances Esure et Oasis.
- Pour le **Profil_4 :** La comparaison donne les mêmes résultats pour les 3 couvertures. La prime est moins importante que l'assurance Esure et Oasis mais elle est plus chère que celle d'Allianz.

Il est donc constaté que le tarif d'Europ Assitance dispose d'un bon positionnement. Ce tarif n'est pas très élevé par rapport au marché. Cependant, ce prix contient des taux de chargements, frais et de commissions qui varient d'un assureur et à un autre. Par conséquent, en raison de ces variables, il est difficile d'évaluer la prime d'annulation pure obtenue à partir de nos modèles.

Chapitre 5

Conclusion

L'assurance voyage est une branche importante d'Europ Assitance, qui souhaite développer son activité dans différents marchés dans le monde.
Dans ce contexte, la compagnie a récemment mis en place une collaboration avec une compagnie aérienne britannique, par le biais de laquelle elle vendra des polices d'assurance voyage en B to B to C.

Cependant, après une analyse préliminaire du marché de l'assurance voyage au Royaume-Uni, il a été constaté que ce nouveau marché est différent du marché habituel d'Europ Assistance. Les différences concernent principalement l'assurance annulation de voyage, les assureurs au Royaume-Uni remboursant tout le voyage et non seulement les billets d'avion, de la même manière qu'Europ Assitance a l'habitude de rembourser dans les autres marché. De plus, au moment de la souscription Europ Assitance demande à l'assuré de renseigner son montant de voyage, chose qui n'est pas faite par les concurrents dans le marché britannique.

Deux bases de données ont servi à évaluer le coût moyen des déplacements. Une base de données interne qui contient les prix des billets d'avion d'une autre compagnie aérienne figurant déjà dans le portefeuille de l'entreprise, et une autre qui est une base de données en open Data indiquant les prix hôteliers en France ces dernières années.

Ce mémoire s'est donc intéressé à la tarification d'une garantie annulation de voyage tout en tenant compte de la spécificité du marché de l'assurance voyage britannique.
Pour estimer la prime pure de cette garantie, nous avons fait appel à une approche fréquence-coût dont les paramètres suivants ont été modélisés :

- La fréquence de sinistre par le biais d'une régression logistique
- Le montant des billets d'avion par un modèle linéaire généralisé
- Le montant d'une nuit d'hôtel à travers un modèle ARIMA des séries temporelles

Selon les tests de performance de modèle du *machine learning*, les résultats obtenus sont acceptables.

Toutefois, nos modèles présentent encore certaines limites, principalement en raison de l'absence de variables dans la base de données interne. Cette base de données ne contient pas beaucoup d'informations au sujet du voyage, par exemple la destination ou l'âge de l'assuré.
En outre, la base de prix des hôtels dont la source est l'UMIH ne contient pas de données sur les prix pour 2021 et début 2022, et prend fin en 2019. Ainsi, pour estimer les prix hôteliers en 2023, le modèle doit faire des prédictions sur 4 ans. Cela risque de fausser les résultats même si l'on tient compte de l'inflation.

Il est important de noter que tout au long de cette tarification, nous avons pris des hypothèses et fait des choix qui peuvent être une source d'erreur : par exemple, l'équivalence entre le choix du plafond de couverture et le type d'hôtel qui sera réservé par le voyageur. En effet, il se peut qu'un voyageur choisisse un plafond de £1 000 tandis qu'il a réservé un hôtel haut de gamme. En outre, afin de généraliser les prix pour le reste de pays européens, il a été décidé de rajouter un coefficient d'équivalences pour les frais d'hôtels, ce qui pourrait être une autre source d'erreurs. .

Ces limites résument parfaitement la difficulté qu'un actuaire peut rencontrer lors de la mise en place d'un nouveau tarif pour un nouveau marché ou bien pour un nouveau produit.

Étant donné que ce mémoire ne concerne que la garantie annulation, il serait intéressant de mener une autre étude complète sur la garantie assistance de voyage. Compte tenu du fait que la Grande-Bretagne a quitté l'Union européenne, les coûts médicaux peuvent être plus élevés pour les britanniques que pour les voyageurs de l'Union européenne. Cette hypothèse vient du fait que les britanniques n'aient plus le droit à la carte européenne d'assurance-maladie (CEAM), une carte qui offre la possibilité d'un remboursement des soins dans les pays membres de l'Union européenne.

Annexe A

Série temporelle

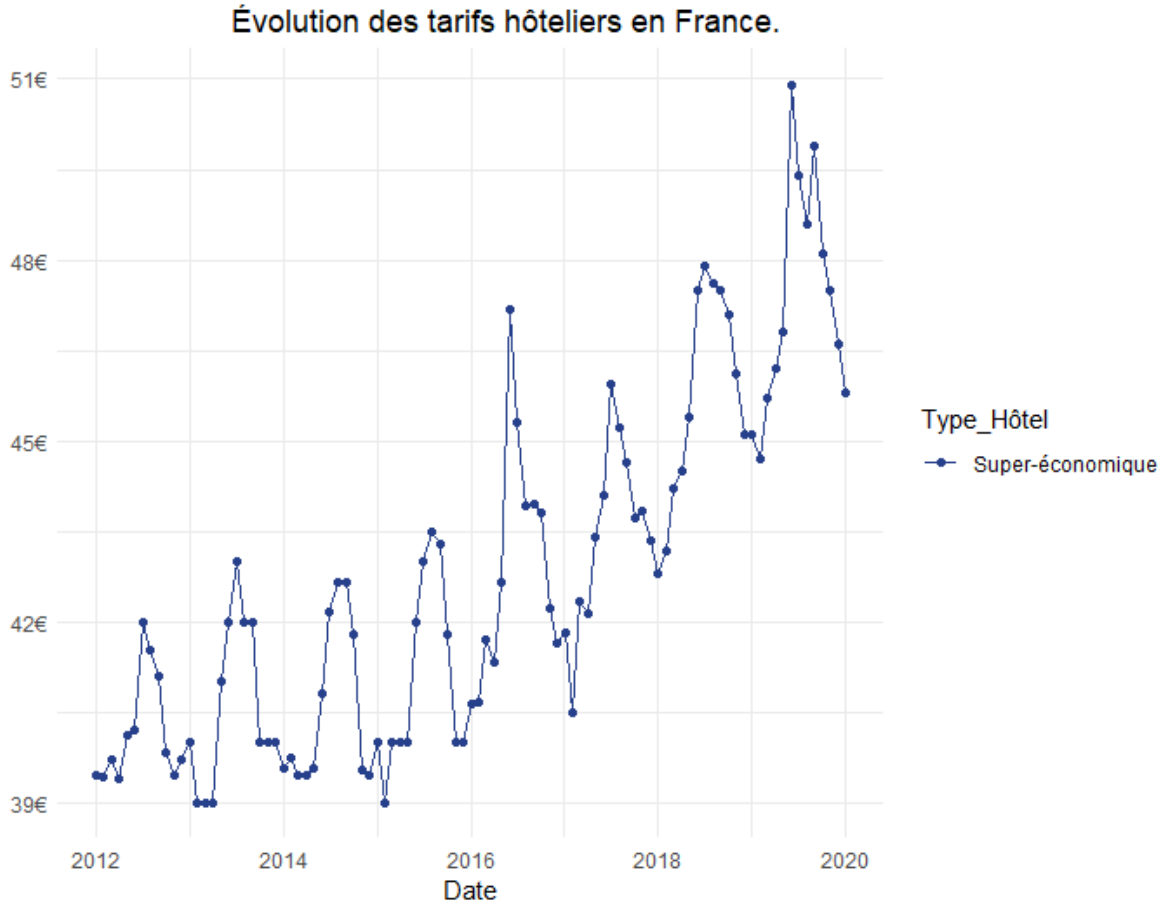


FIGURE A.1 – Évolution des prix des hôtels super-économique en France entre 2012 et janvier 2020

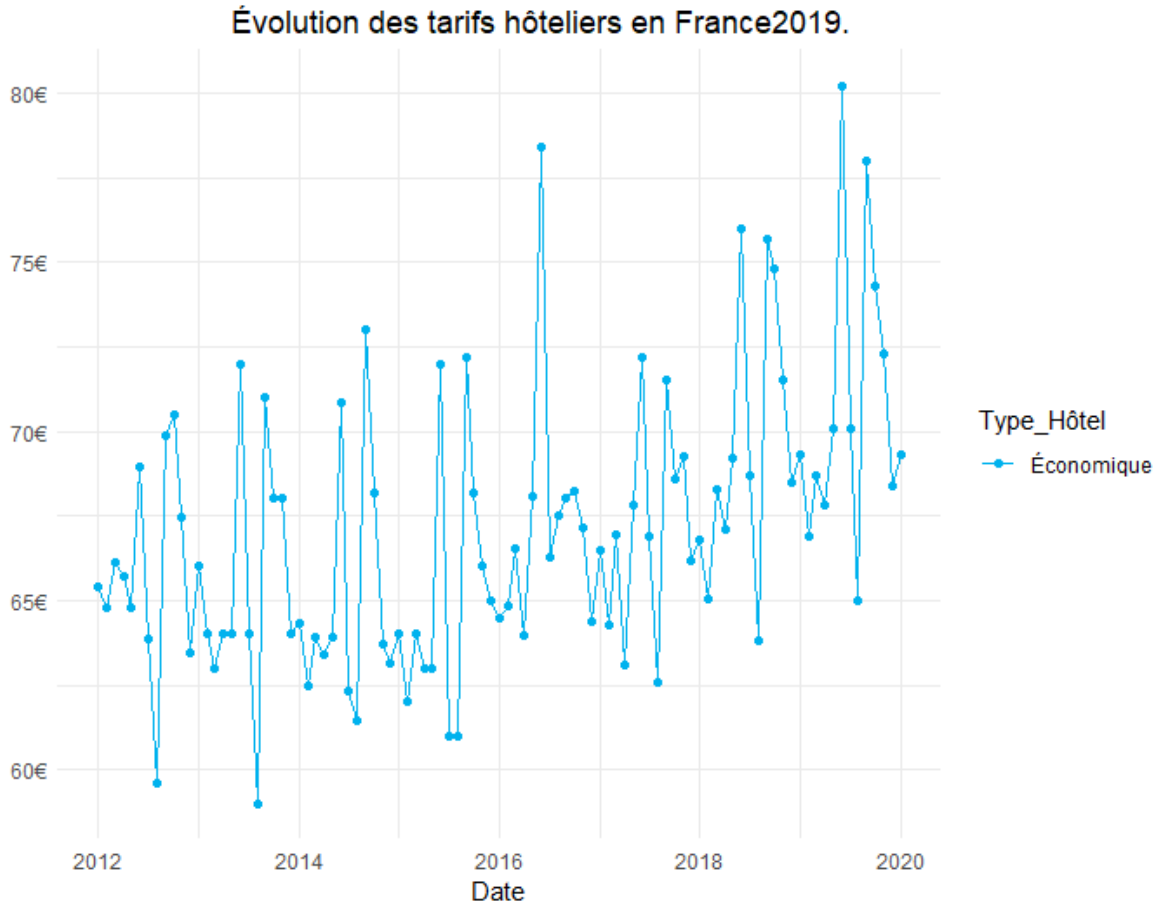


FIGURE A.2 – Évolution des prix des hôtels économique en France entre 2012 et janvier 2020

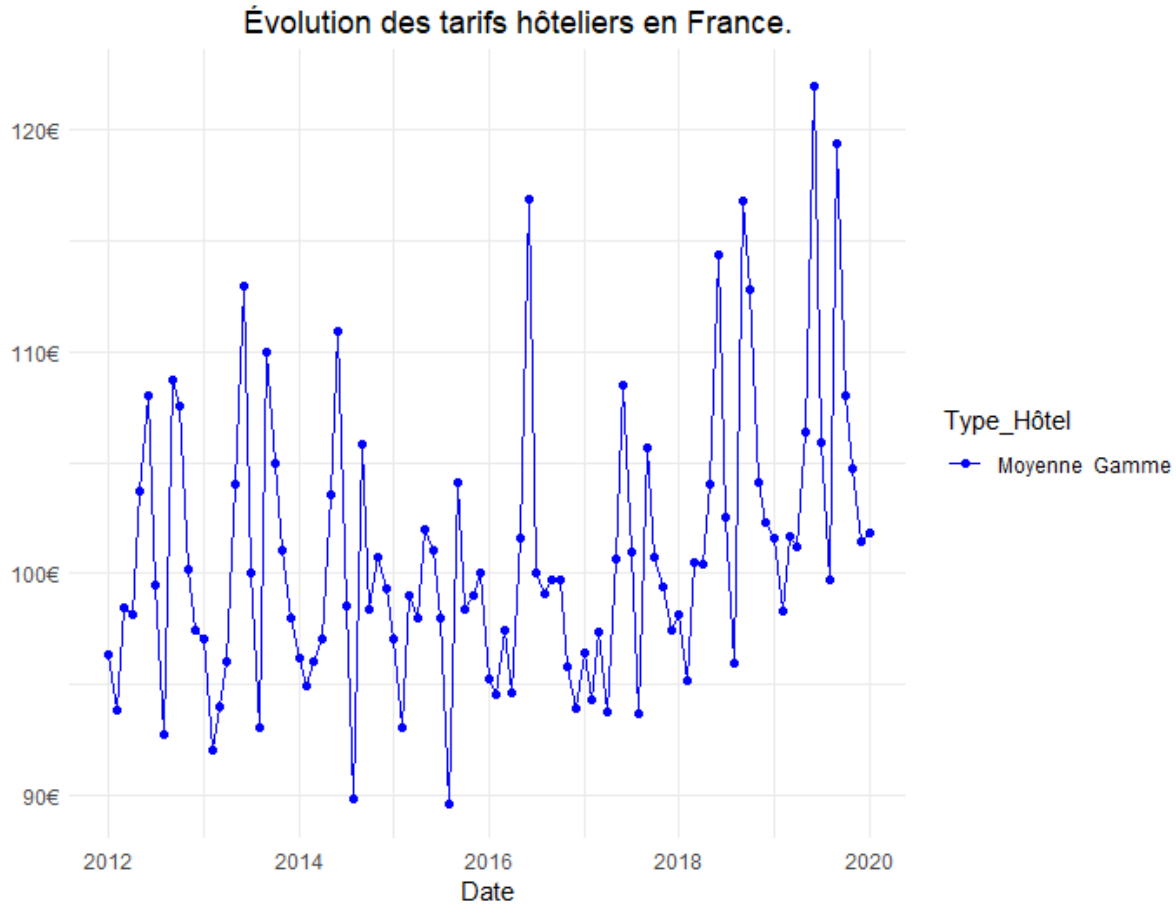


FIGURE A.3 – Évolution des prix des hôtels moyenne gamme en France entre 2012 et janvier 2020

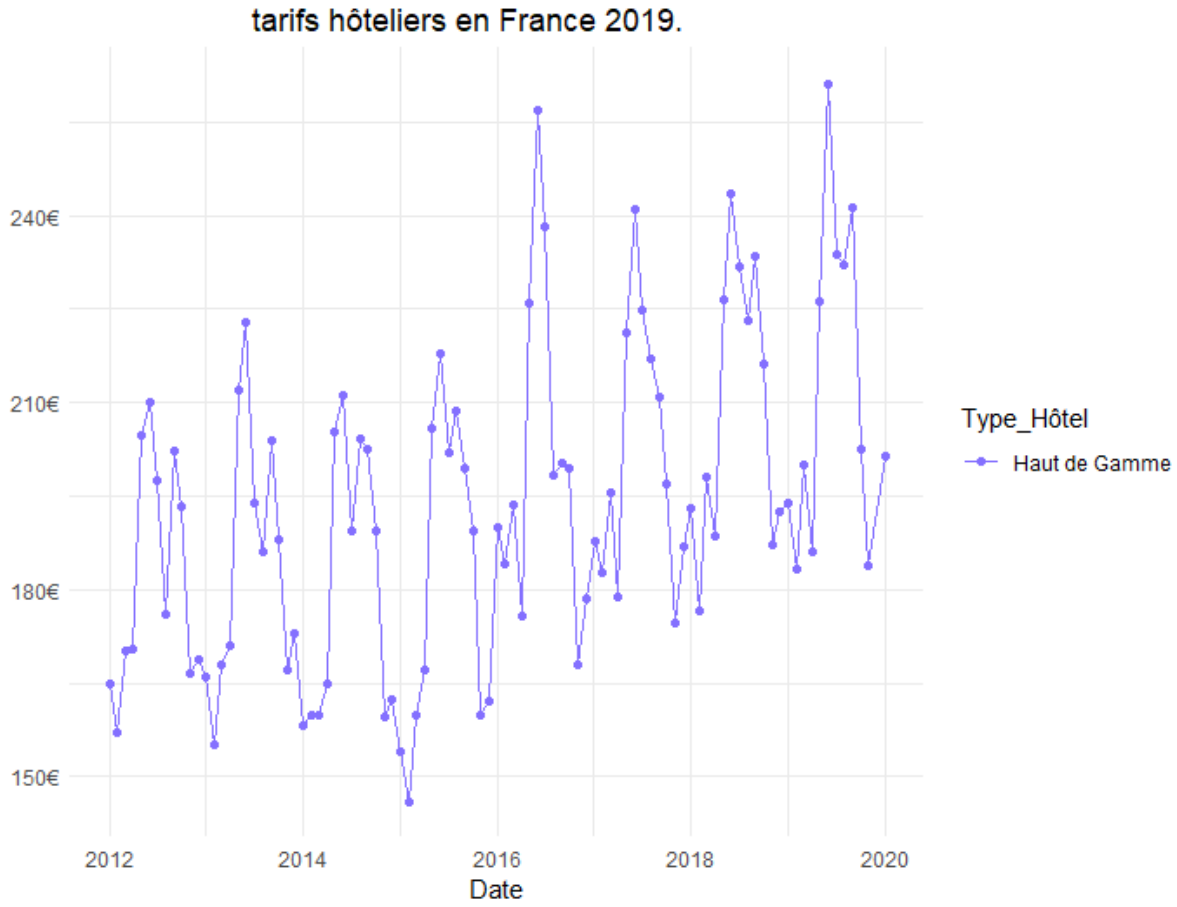


FIGURE A.4 – Évolution des prix des hôtels haut de gamme en France entre 2012 et janvier 2020

Table des figures

Bibliographie

[1] Scores globaux de maturité de l'open data. 2021.

[2] ABI. *One claim every minute - Travel insurance payouts highest since 2010*, 2018.

[3] A.CHARPENTIER. *Computational Actuarial Science with R.* 2016.

[4] Union des Métiers et des Industries de l'Hôtellerie (UMIH). Observatoire des performances hôtelières.

[5] RJ.HYNDMAR et G.ATHANASOPOULOS. *Forcasting : Principles and Practice* . 2019-2020.

[6] P.AILLIOT et J-M.DERRIEN. Cours de séries temporelles. *Support de cours M2 EURIA*, 2021-2022.

[7] F.BOUTTIER. Construction de modèles prédictifs pour déterminer l'inflation des pièces automobiles. *Mémoire Institut des Actuaires*, 2020.

[8] Office for National Statistics (ONS). *Overseas travel and tourism*, 2022.

[9] Stanley et Sturdivant Rodney X Hosmer Jr, David W et Lemeshow. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[10] M.AOUICHI. Analyse de l'assurance voyage à travers la tarification du risque. *Mémoire Institut des Actuaires*, 2021.

[11] J. MARIE-ROSE. Application sur r en tarification non-vie. *Support de cours M2 EURIA*, 2021-2022.

[12] P.AILLIOT. Modèle linéaire généralisé. *Support de cours M1 EURIA*, 2020-2021.

[13] S.NAVARRO. L'open data au service de la tarification à l'adresse. *Mémoire Institut des Actuaires*, 2021.

[14] *Finaccord consulting company. Travel Insurance and Assistance in Europe.* 2019.

[15] *Next Move Strategy Consulting. Travel Insurance Market.* 2022.