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

le 7 Septembre 2023

Par : Océane BRETON
Titre : Problématique de l'inflation en assurance auto : exploitation de rapports d'expertise pour sophistiquer le modèle de tarification

Confidentialité : Oui - (Durée: 2 ans)

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

Membres présents du jury de l'Institut des Actuaires :
David DUBOIS
Guillaume GORGE
Khouloud MANDHOUJ
Signature :

Entreprise :
Generali Vie
Signature :

Membre présent du jury de l'EURIA :
Franck Vermet

Directeur de mémoire en entreprise :
Yoann GOUYEN
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

Ce mémoire a pour objet d'adapter le processus de tarification automobile - affaires nouvelles et renouvellement du portefeuille - mis à jour tous les ans. L'objectif est de sophistiquer la prise en compte de l'inflation dans ce processus.

En effet, depuis plusieurs années, le secteur de l'assurance connaît de façon générale une augmentation du montant des sinistres à indemniser que ce soit en nombre ou en fréquence du fait du contexte climatique et de l'inflation. En assurance auto, cette inflation était de 7,9% entre 2021 et 2022 pour une réparation moyenne suite à un sinistre et plus précisément de 9,6% pour le poste des pièces, d'après le SRA (Sécurité et Réparation Automobiles) [1].

Ce constat a été le point de départ des travaux de ce mémoire. Tout d'abord, l'idée est d'être en mesure de suivre en continu l'inflation des réparations automobiles, afin de détecter des changements de politiques tarifaires de la part des constructeurs le plus en amont possible et ainsi anticiper les impacts sur le risque assuré. Dans un deuxième temps, on s'attache à identifier les facteurs discriminants de l'inflation (ex : marque du véhicule, ancienneté, etc.). Et pour finir on implémente cette inflation différenciée dans le processus de tarification des primes.

Pour réaliser cette étude, nous n'avons pas exploité des données utilisées traditionnellement en actuariat. En effet, nous nous sommes appuyés directement sur les rapports d'expertise de sinistres auto.

Mots clefs: Tarification, Inflation, Automobile, Séries temporelles, Modèles, Actuariat, Mémoire

Keywords: Pricing, Inflation, Automotive, Time series, Actuarial Issues, Master's Thesis

1. *sur les véhicules particuliers et véhicules utilitaires de moins de 3,5 tonnes, sur les réparations, collisions, stationnements hors catastrophes naturelles, vol, incendie, vandalisme et bris de glaces*

Abstract

This masterthesis aims to adapt the auto pricing process (new business and portfolio renewal), which is updated every year. The purpose is to refine the way in which inflation is incorporated into this process.

In fact, for several years now, the insurance sector in general has been experiencing an increase in the number amount and frequency to be compensated, due to the climate and inflation. In automobile insurance, this inflation was 7.9% between 2021 and 2022 for an average repair following a claim, and more precisely 9.6% for the parts item, according to the SRA - *Sécurité et Réparation Automobiles*[2].

This observation was the starting point for the work on this thesis. Firstly, the idea is to be able to continuously monitor car repair inflation, in order to detect changes in manufacturers' pricing policies as early as possible, and thus anticipate the impact on the insured risk. Secondly, we identify the discriminating factors on inflation (e.g. vehicule brand, age, etc.). Finally, we implement this differentiated inflation in the premium pricing process.

To carry out this study, we did not exploit data traditionally used in actuarial science. Instead, we relied directly on car claims adjustment reports.

Keywords: Pricing, Inflation, Automotive, Time series, Actuarial Issues, Master's Thesis

Mots clefs: Tarification, Inflation, Automobile, Séries temporelles, Modèles, Actuariat, Mémoire

2. *on private vehicles and commercial vehicles under 3.5 tonnes, on repairs, collisions, parking excluding natural disasters, theft, fire, vandalism and glass breakage*

Note de synthèse

Les compagnies d'assurances, de par leur rôle dans la société, font partie des agents économiques fortement impactés par les crises mondiales.

Au cours de ces dernières années, plusieurs événements ont perturbé l'économie. Que ce soit la pandémie de la covid-19, l'intensification de la guerre en Ukraine avec l'invasion de l'Ukraine par la Russie, des politiques accommodantes pour favoriser le pouvoir d'achat ou encore la multiplication d'événements climatiques en intensité et en fréquence ; tous ces événements ont eu un impact notable sur l'offre et la demande des consommateurs, modifiant l'équilibre de marché et entrainant une forte augmentation des prix.

Pour l'année 2022, l'indice des prix à la consommation a approché un niveau qui n'avait plus été atteint depuis 30 ans, 5,2% en moyenne en France. De plus, dans le milieu automobile, les factures moyennes de sinistres ont fortement augmenté avec une intensification de 9,6% du montant du poste des pièces sur l'année 2022 par rapport à l'année 2021, d'après l'association Sécurité et Réparations Automobiles.

Pour faire face à ces augmentations de prix, les assureurs doivent ainsi s'adapter dans leur processus de tarification IARD en estimant d'une part la fréquence de sinistre et d'autre part le montant des réparations attendus pour un sinistre. C'est le deuxième aspect qui est élaboré dans ce mémoire. Pour cela, les données historiques sont mises à profit pour donner un tarif adapté à ce qui est prévu pour l'année correspondante en formulant une hypothèse d'inflation.

Ainsi, dans l'objectif de proposer une tarification plus fine des contrats d'assurance automobile, ce mémoire se concentre à différencier cette hypothèse d'inflation en fonction des caractéristiques du véhicule assuré.

Approche suivie pour différencier l'inflation

Afin de répondre à cet objectif, l'étude s'est divisée en deux démarches principales.

La première approche se résume à la création d'un outil de détection automatique d'évolutions brutales des prix des pièces automobiles. Grâce à une alerte, les actuaires ont alors une information supplémentaire pour déterminer le taux d'inflation à injecter dans le modèle de tarification.

La seconde consiste en l'utilisation de modèles plus ou moins complexes pour déceler les facteurs discriminants de l'inflation du coût de la pièce.

Pour conclure sur ces approches, les résultats sont pris en compte et ajustés pour permettre une application opérationnelle.

Les sources de données

Deux types de sources de données ont été utilisées : un premier type constitué d'informations internes à l'entreprise et un second avec des données macro-économiques.

Comme dans les données sinistres usuellement utilisées les informations sont toutes regroupées pour donner un montant total de la facture, les informations manipulées étaient celles des rapports d'expertise automobile. Grâce à ces rapports, une étude avec une meilleure finesse a pu être réalisée puisqu'ils contenaient le détail des pièces.

Dans l'intention d'appliquer notre analyse sur des pièces remplacées, une sélection sur les rapports et les pièces a été mise en place :

— Les rapports d'expertise conservés sont ceux qui ont été réalisé sur des véhicules particuliers et utilitaires réparables techniquement et économiquement, avec une expertise en France métropolitaire.

— Pour les pièces, devant la grande variété d'information des pièces, une analyse a été réalisée pour sélectionner les pièces pertinentes à suivre. Il s'agit des "boucliers avant", "boucliers arrière", "ailes avant gauche", "portes avant" et "portes arrière". Les critères de sélection sont notamment la qualité des informations, le volume et leur significativité en termes de montant.

Pour s'assurer de la qualité des données extraites, une étude comparative avec une source de données externe nommée EUCON et une étude interne de suivi de pièces représentatives de modèles automobiles dans le temps, ont permis de confirmer que nous avions les prix constructeurs des pièces utilisées lors des réparations. Dans le choix de la granularité, le suivi peut se faire à différents niveaux, que ce soit au niveau des marques, des modèles ou de regroupements de marques. Une seconde analyse a défini le niveau marque pour cette étude.

Dans un second temps, pour expliquer les facteurs discriminants de l'inflation du coût de la pièce, des indices marchés ont été récupérés sur le site de l'INSEE. Ces informations devaient remplir certains critères pour être sélectionnées. Il s'agit de l'historique, la mise à jour régulière et la facilité de récupération pour permettre une continuité de suivi.

Ces informations macro-économiques sont :

1. l'indice de prix à la consommation
2. le cours de l'euro par rapport au dollar américain
3. l'indice des prix de certains matériaux : l'acier, l'aluminium, le magnésium et le cuivre
4. le cours du pétrole Brent
5. le prix de l'essence en retenant le sans-plomb 98
6. le prix du verre
7. le coût de la main d'oeuvre

Approche automatisée de détection de ruptures sur les évolutions des coûts des pièces détachées

Le premier objectif était de mettre en place un outil permettant de suivre en continu l'évolution des prix constructeurs des pièces automobiles.

Pour permettre cela, une procédure en 4 étapes a été réalisée :

1. Dans un premier temps, pour réduire le nombre de courbes à suivre dans le temps, une classification non supervisée a été appliquée sur les évolutions des coûts des pièces de marques. Initialement le suivi était sur 12 marques distinctes et les autres dans une treizième modalité.

2. Ensuite, sur les 6 groupes obtenus un calcul des coûts moyens de chaque pièce par groupe de marques a été formé.

3. Après cela, pour chaque groupe de marques, les courbes correspondant à leurs cinq pièces ont été agrégées en utilisant leurs répartitions trimestrielles moyennes pour garder une stabilité temporelle.

4. Pour arriver à la courbe finale, l'opération a été réitérée pour agréger les courbes des groupes de marques de manière équivalente à celle utilisée pour les pièces. C'est-à-dire qu'on utilise les parts moyennes de ces groupes.

On peut résumer les étapes de l'agrégation qui correspondent aux étapes 2, 3 et 4 dans le schéma *1*.

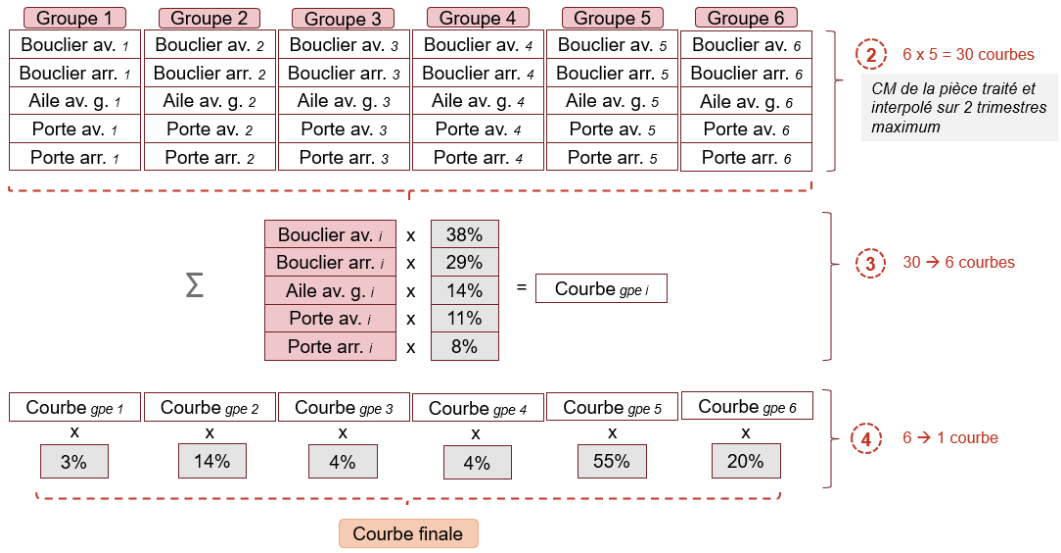


FIGURE 1 – Schéma résumé du traitement des courbes avant la détection de rupture

Sur cette courbe obtenue, deux modèles de détection de ruptures sur les séries temporelles ont été appliqués. Le premier, nommé *Prune Exact Linear Time*, est une méthode

optimale dans le sens où toutes les segmentations possibles sont testées pour ressortir les coupures idéales. La segmentation binaire est la seconde méthode, elle fait partie des approches approximatives puisqu'elle détermine une rupture à chaque itération jusqu'à ce qu'un critère d'arrêt ne soit atteint.

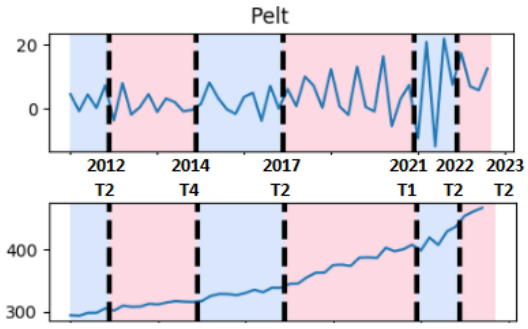


FIGURE 2 – Ruptures obtenues avec la méthode *PELT* sur la courbe différenciée et la courbe normale

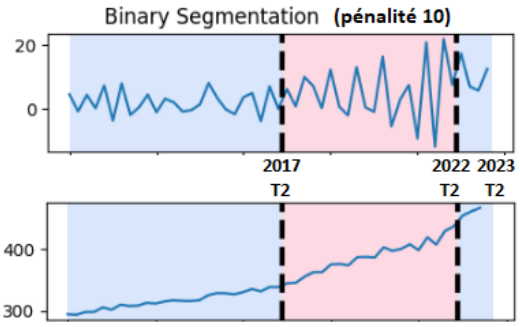


FIGURE 3 – Ruptures obtenues avec la méthode *Binary segmentation* avec une pénalité de 10, sur la courbe différenciée et la courbe normale

Dans les deux situations, le deuxième trimestre 2022 a été détecté, même avec une courbe d'entraînement qui s'arrêtait en fin 2022, ce qui signifie que la détection est rapide. Ainsi, dès cet instant, l'outil donne l'alerte de changement dans la tendance de l'évolution du prix des pièces détachées aux actuaires. Ces derniers peuvent alors prendre en compte ce nouvel élément pour définir leur hypothèse d'inflation à injecter en tarification automobile et mettre à jour leurs tarifs.

Identification des facteurs discriminants de l'inflation des pièces automobiles

Pour comprendre quels sont les caractéristiques qui impactent le plus l'évolution du prix de la pièce, plusieurs informations ont été récupérées : les informations du rapport d'expertise, du véhicule grâce aux données du SRA et les données macro-économiques.

Avec ces éléments, la base de modélisation a été formée en agrégeant les pièces en fonction des caractéristiques du véhicule marque, ancienneté du véhicule, classe de prix SRA et groupe SRA. Pour cela, comme les différentes variables d'agrégation possédaient beaucoup de modalités, des regroupements de modalités ont été formés par groupes homogènes, scindés par centiles. Pour les variables des classes de prix SRA et du groupe SRA, comme celles-ci indiquent la classe de prix à la sortie du véhicule et sa dangerosité, les coupures des modalités ont pris en compte une "normalisation" par marque, pour ne pas se retrouver avec toutes les pièces de véhicules haute gamme dans les classes de prix élevées par exemple, et inversement.

Un compromis sur le nombre de modalités pour chaque variable d'agrégation a eu lieu, en considérant un nombre suffisamment élevé pour permettre de différencier l'inflation

mais sans l'être trop pour avoir un volume suffisant de données dans la base. Ainsi, 6 modalités par variable d'agrégation ont été sélectionnées.

Par exemple quelques regroupements de modalités sont présentés dans le tableau suivant :

Variable	Marque	G1	G2	G3	G4	G5	G6
Âge		≤ 2 ans	$\in]2,3]$	$\in]3,5]$	$\in]5,7]$	$\in]7,10]$	> 10
Classe SRA	Renault	A $\rightarrow$ F	G $\rightarrow$ H	I	J	K $\rightarrow$ L	M $\rightarrow$ S
	Audi	H $\rightarrow$ L	M	N	O $\rightarrow$ P	Q $\rightarrow$ R	S $\rightarrow$ X
	BMW	I $\rightarrow$ M	N	O	P $\rightarrow$ Q	R	S $\rightarrow$ X
Groupe SRA	Renault	≤ 28	$\in]28,29]$	$\in]29,30]$	$\in]30,31]$	$\in]31,32]$	> 32
	Audi	≤ 32	$\in]32,33]$	$\in]33,34]$	$\in]34,35]$	$\in]35,37]$	> 37
	BMW	≤ 33	$\in]33,34]$	$\in]34,35]$	$\in]35,36]$	$\in]36,38]$	> 38

Une fois l'agrégation formée, sur la base obtenue, deux variables sont calculées pour les modèles :

— la variable à expliquer qui correspond à l'évolution du coût de la pièce pour la situation de risque correspondante. Pour cela, pour la pièce de la situation de risque i à l'instant t, elle est calculée :

$$Evolution_{i,t} = (\frac{CM_{i,t}}{CM_{i,t-1}} - 1) \times 100 \tag{1}$$

Par définition, cette variable cible apporte une contrainte supplémentaire sur la création des données puisqu'elle implique une condition d'observer des pièces correspondant à une même branche de risque sur deux temps consécutifs. Dans les faits, ce n'est pas toujours le cas.

— la variable exposition qui permet de renseigner une pondération sur les individus qu'on intègre dans les modèles. Comme la variable cible est calculée sur des coûts moyens à deux instants différents, l'exposition considère le nombre de pièces minimum utilisé. Pour la pièce de la situation de risque i à l'instant t, l'exposition est donnée par :

$$Exposition_{i,t} = min(Nombre\ pieces_{i,t}\ ;Nombre\ pieces_{i,t-1}) \tag{2}$$

Les informations de marché ont été ajoutées à la base ainsi obtenue, en valeurs et en variations. Comme les indices du marché peuvent prendre du temps à se propager sur le prix des pièces, plusieurs temporalités ont été prises en compte, avec des décalages temporels de zéro à trois semestres.

Pour commencer notre modélisation, nous avons sélectionné naturellement un GLM qui est facilement interprétable, pour ensuite le comparer avec des modèles habituellement plus performants mais aussi plus complexes à interpréter, l'arbre de décision, la

forêt aléatoire et l'*histogram gradient boosting*. Pour contrer cette difficulté d'interprétation de ces modèles, l'utilisation d'un algorithme d'explicabilité éprouvé SHAP sera proposé. Le résultat de ces modèles sont les suivants :

Modèle	RMSE		MAE		R^2 (%)	
	test	train	test	train	test	train
GLM pénalisé	**10.06**	9.43	7.79	7.12	**9.71**	9.31
Arbre	10.09	9.54	7.79	7.24	9.14	7.22
HGB	**10.04**	9.14	7.70	6.86	**10.13**	14.89
Forêt aléatoire	10.09	9.24	7.78	7.00	9.24	13.04

On constate sur ce tableau que :

- Tout d'abord, il n'y a pas de phénomène de surapprentissage [3], grâce à l'optimisation des paramètres par *cross validation* ;
- Ensuite, comme attendu, plus les modèles sont complexes plus les performances s'améliorent mais c'est plus visible sur le R^2 et peu sur les RMSE et MAE.
- Enfin, les modèles sont limités en performances. Une analyse a permis de montrer que cette limite est due à un problème de volumétrie de données. En effet, cette faible volumétrie ne nous permet pas de construire des variables explicatives suffisamment fines pour prédire les inflations importantes.

En revanche, tous ces modèles ont montré des convergences en termes de comportements. Tous mettent en avant comme variables explicatives les variables "marque du véhicule", "anciennetés du véhicule", "classes de prix SRA" et dans une moindre mesure "groupes SRA". Ci-dessous les résultats obtenus avec la méthode SHAP sur l'*histogram gradient boosting* :

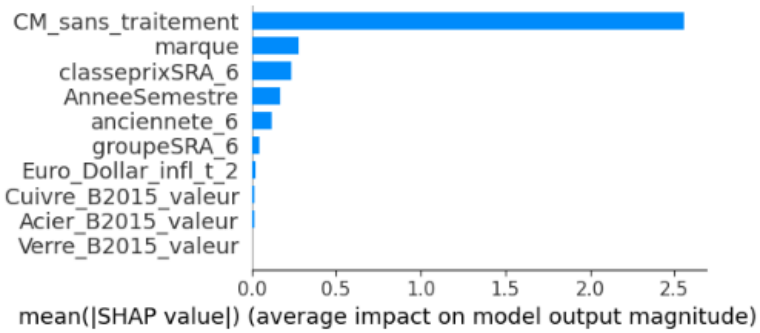


FIGURE 4 – Moyennes des valeurs absolues des mesures SHAP les plus significatives

Pour des raisons d'explicabilité et de facilité d'utilisation, nous faisons le choix de retenir le modèle GLM pénalisé pour quantifier les facteurs discriminants de l'inflation des pièces détachées.

3. Le surapprentissage, ou *overfitting*, est un comportement indésirable d'apprentissage automatique qui se produit lorsque le modèle arrive à obtenir des prédictions très précises pour les données d'entraînement mais pas pour les données de test.

Les coefficients du GLM pénalisé sont représentés dans les graphiques suivants, sur lesquels les légendes ont été retirées pour des raisons de confidentialité.

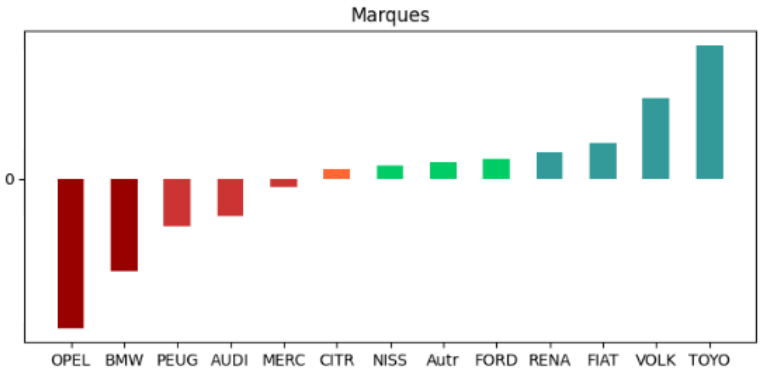


FIGURE 5 – Coefficients du GLM pénalisé pour les modalités de marques

Pour les marques comme Opel ou BMW, les coefficients sont négatifs, cela signifie qu'elles auraient plutôt tendance à avoir une évolution de leurs prix plus faible que les autres marques. A l'inverse, les coefficients de Toyota, Volkswagen ou encore Fiat sont positifs, c'est-à-dire qu'elles augmenteraient plus leurs évolutions de prix de leurs pièces automobiles.

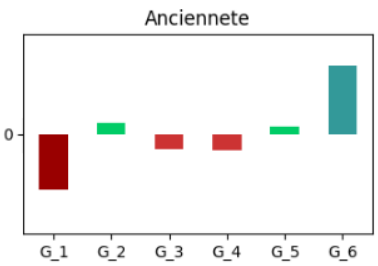


FIGURE 6 – Coefficients du GLM pénalisé pour les modalités d'ancienetés

FIGURE 7 – Coefficients du GLM pénalisé pour les modalités de classes SRA

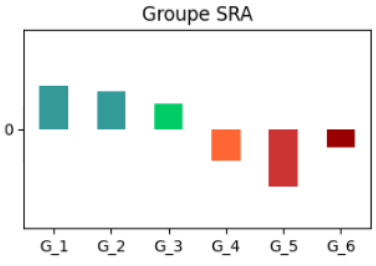


FIGURE 8 – Coefficients du GLM pénalisé pour les modalités de groupes SRA

Les coefficients correspondant aux variables "ancienneté" et "classe SRA" sont croissants en fonction des modalités des variables. Concrètement, cela signifie que les véhicules récents augmentent moins leurs pièces que les véhicules âgés et les véhicules de classes de prix SRA faibles augmentent moins leurs pièces que les véhicules de classes plus élevées. Inversement, les coefficients des modalités de groupes de puissances SRA sont décroissants. Cela signifie que, plus le véhicule est puissant et moins l'évolution de prix est importante.

Intégration opérationnelle des résultats

L'idée est d'utiliser les résultats des études précédentes pour se projeter dans une application opérationnelle pour le BET Auto.

Dans le processus de tarification, le BET Auto doit commencer par formuler une hypothèse d'inflation à injecter. Une fois l'hypothèse définie, pour distinguer cette inflation entre deux véhicules de caractéristiques différentes, les coefficients obtenus avec le GLM pénalisé pourraient être utilisés.

Cependant, les choix de politiques tarifaires vont amener certainement à ne sélectionner qu'une liste réduite de critères différenciants par rapport à celles définis par l'étude. Pour des raisons d'explicabilité et de positionnement commercial, des ajustements peuvent être appliqués. Ces derniers peuvent porter sur le nombre de variables différenciantes considérées, sur le nombre de leurs modalités ou encore sur l'amplitude des coefficients.

Conclusion

Finalement, ce mémoire a contribué à sophistiquer le modèle de tarification actuel en permettant une segmentation de l'inflation selon les profils des contrats.

Premièrement, un outil de suivi de l'évolution des coûts des pièces détachées permet d'alerter l'assureur le plus en amont possible lorsqu'un changement de politique tarifaire des constructeurs a été détecté. Grâce à cela, l'actuaire peut anticiper les questions sur l'hypothèse d'inflation qu'il injectera dans les modèles de projection.

Dans un second temps, une étude visant à identifier les facteurs discriminants de l'inflation du coût de la pièce automobile a été menée. Malgré les difficultés de modélisation rencontrées, des signaux faibles ont été détectés, et nous avons montré que la marque, l'ancienneté, la classe de prix SRA et le groupe SRA étaient des éléments importants pour différencier l'inflation des pièces détachées.

Et pour finir, une stratégie d'utilisation du GLM pénalisé obtenu pour différencier l'inflation contrat par contrat à partir de ces résultats est envisagée. Les résultats sont facilement exploitables par le Bureau d'Etudes Techniques, et après les simplifications et les ajustements dus aux contraintes commerciales, ils pourront facilement être intégrés dans les outils de tarification de Generali.

Synthesis note

Due to their role in society, insurance companies are among the most affected economic agents by global crises.

In recent years, several events have disrupted the economy. Whether it's the covid-19 pandemic, the intensification of the war in Ukraine with Russia's invasion of the country, accommodating policies to boost purchasing power or the multiplication of climatic events in terms of intensity and frequency. All these events have had a significant impact on consumer supply and demand, altering the market balance and leading to a sharp rise in prices.

For 2022, the consumer price index has approached a level not seen for 30 years, averaging 5.2% in France. What's more, in the automotive sector, average claims invoices have risen sharply, with a 9.6% increase in the cost of parts in 2022 compared with 2021, according to the Security and Reparation Automobiles association (SRA).

To cope with these price increases, insurers need to adapt their Property & Casualty (P&C) pricing process by estimating both the frequency of claims and the amount of repairs expected for a claim. It is the second aspect that is developed in this study. For this purpose, historical data is used to produce a price that is in line with the forecast for the corresponding year, with an inflation hypothesis. Thus, with the aim of proposing a more refined pricing of motor insurance contracts, this dissertation focuses on differentiating this inflation hypothesis depending to the characteristics of the insured vehicle.

Thus, with the aim of proposing a more refined pricing of motor insurance contracts, this dissertation focuses on differentiating this inflation hypothesis depending to the characteristics of the insured vehicle.

Inflation differentiation approach

To meet this challenge, the study was divided into two main approaches.

The first approach involved the creation of a tool for the automatic detection of sudden changes in auto parts prices. Thanks to an alert, actuarial teams get information to determine the inflation rate to inject into the pricing model.

The second approach involves the use of complex models to detect discriminating factors in part price inflation.

To conclude these approaches, the results are considered and adjusted to enable operational application.

Data sources

Two types of data sources were used : a first type made up of internal company information and a second with macro-economic data.

Usually, in the claims data used, the information is all grouped together to give a total invoice amount. Therefore, the information manipulated was car expert reports, thanks to these, a more detailed study could be carried out, since they contained details of the parts.

In order to apply our analysis to replaced parts, a selection of reports and parts was set up :

- The expert reports selected were those carried out on passenger cars and commercial vehicles that were technically and economically repairable, with an expertise carried out in mainland France.
- In view of the wide variety of information available on parts, an analysis was carried out to select the relevant parts to be tracked. These are "front bumper", "rear bumper", "front left fender", "front door" and "rear door". The selection criteria include the quality of the information, its volume and its significance in terms of value.

To ensure the quality of the extracted data, a comparative study with an external data source called EUCON, and an internal study of the tracking of representative car model parts over time, confirmed that we had the manufacturers' prices for the parts used in repairs. In the choice of granularity, tracking can be done at different levels, whether at brand, model or brand group level. A second analysis defined the brand level for this study.

Secondly, to explain the discriminating factors in parts cost inflation, market indices were obtained from the INSEE website. This information had to meet certain criteria in order to be selected. These criteria were historical data, regular updates, and ease of retrieval to ensure continuity of monitoring.

These macro-economic data are :

1. the consumer price index
2. the euro/US dollar exchange rate
3. the price index for several materials : steel, aluminum, magnesium and copper
4. the price of Brent crude oil
5. the price of unleaded 98 gasoline
6. the price of glass
7. labor costs

Automated approach to detect breaks in spare parts cost trends

The first objective was to set up a tool that would enable us to continuously monitor changes in manufacturers' prices for automotive parts.

To make this possible, a 4-step procedure has been developed :

1. Initially, to reduce the number of curves to be tracked over time, an unsupervised classification was applied to the evolution of branded parts costs. Initially, 12 separate brands were tracked, with the others in a thirteenth modality.
2. From the 6 groups obtained, the average cost of each part per brand group formed was calculate.
3. After this, for each brand group, the curves corresponding to their five pieces were aggregated using their average quarterly distributions to maintain temporal stability.
4. To get the final curve, the operation was repeated to aggregate the brand group curves in the same way as for the parts. In other words, we use the average shares of these groups.

We can summarize these steps in the following diagram.

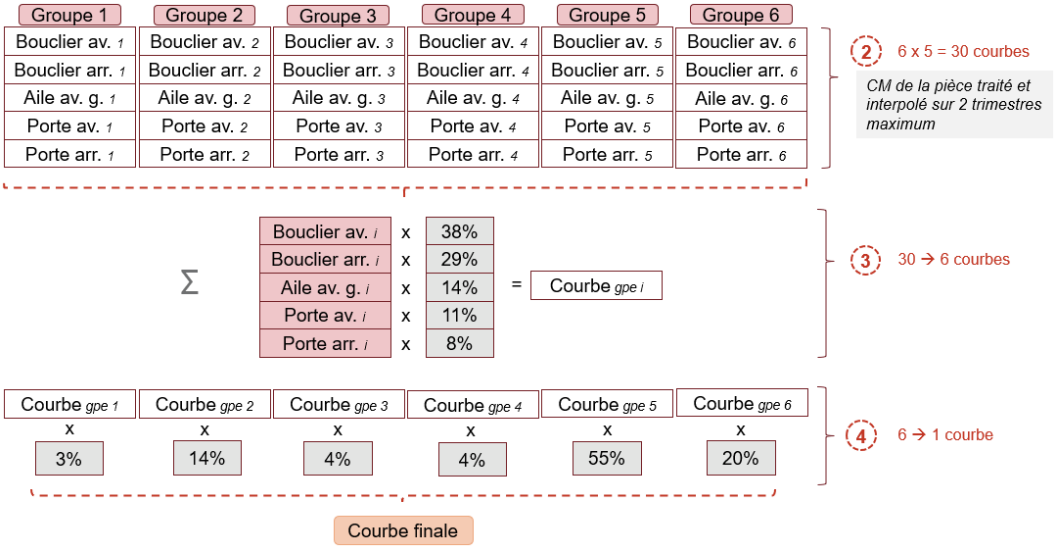


FIGURE 9 – Summary diagram of curve processing prior to break detection

On the curve obtained, two break detection models on the time series were applied. The first, known as the Prune Exact Linear Time method, being an optimal method since all possible segmentations are tested to identify the optimum cuts. Binary segmentation is the second approximate approach since it determines a break at each iteration until a stopping criterion is reached.

In both situations, the second quarter of 2022 was detected, even with a training curve that stopped at the end of 2022, which means that detection is quick. At this point, the

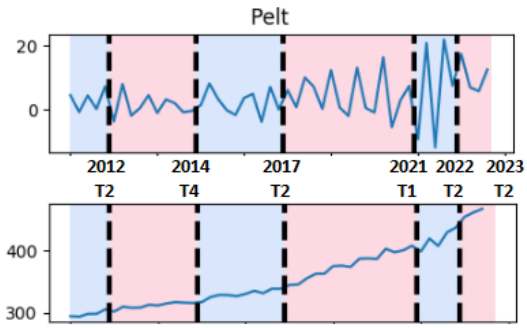


FIGURE 10 – Breaks obtained with the PELT method on the differentiated curve and the normal curve

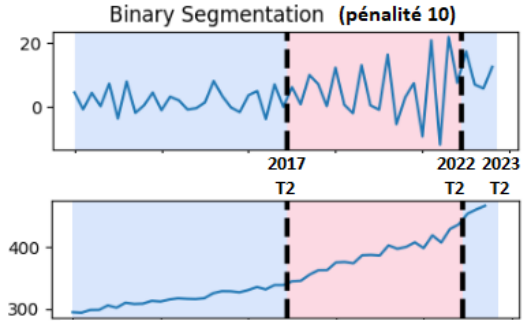


FIGURE 11 – Breaks obtained with the Binary Segmentation with a penalty of 10, on the differentiated curve and the normal curve

tool alerts the actuaries that there's a change in the trend of spare parts prices. They can then take this new element into account to define their inflation hypothesis to be injected into automobile pricing and update their rates.

Identifying discriminating factors in automotive parts inflation

To understand which characteristics, have the greatest impact on the evolution of part prices, several types of information were retrieved : information from the expert report, vehicle data from the SRA, and macro-economic data.

With these elements, the modeling base was formed by aggregating the parts according to vehicle characteristics such as brand, vehicle age, SRA price class and SRA group. For this, as the various aggregation variables had many modalities, groups of modalities were formed by homogeneous groups, split by percentiles. For the SRA price class and SRA group variables, as these indicate the vehicle's exit price class and its dangerousness, the modality segmentation took into account a "normalization" by brand, in order to not find all the parts of high-end vehicles in the high price classes, for example, and vice versa.

A compromise was reached on the number of modalities for each aggregation variable, considering a sufficiently high number to differentiate inflation but also in order to have a sufficient volume of data in the database. Thus, 6 modalities per aggregation variable were selected.

For example, some groupings of modalities are shown in the following table :

Variable	Brand	G1	G2	G3	G4	G5	G6
Age		$\leq$ 2 ans	$\in]2,3]$	$\in]3,5]$	$\in]5,7]$	$\in]7,10]$	> 10
SRA Class	Renault	A $\rightarrow$ F	G $\rightarrow$ H	I	J	K $\rightarrow$ L	M $\rightarrow$ S
	Audi	H $\rightarrow$ L	M	N	O $\rightarrow$ P	Q $\rightarrow$ R	S $\rightarrow$ X
	BMW	I $\rightarrow$ M	N	O	P $\rightarrow$ Q	R	S $\rightarrow$ X
SRA Group	Renault	≤ 28	$\in]28,29]$	$\in]29,30]$	$\in]30,31]$	$\in]31,32]$	> 32
	Audi	≤ 32	$\in]32,33]$	$\in]33,34]$	$\in]34,35]$	$\in]35,37]$	> 37
	BMW	≤ 33	$\in]33,34]$	$\in]34,35]$	$\in]35,36]$	$\in]36,38]$	> 38

Once the aggregation formed, on the database obtained, two variables are calculated for the models :

— the variable to be explained, which corresponds to the change in part cost for the corresponding risk situation. For this purpose, it is calculated this way :

$$evolution_{i,t} = (\frac{CM_{i,t}}{CM_{i,t-1}} - 1) \times 100 \tag{3}$$

for the car part in risk situation i at time t.
By definition, this target variable brings an additional constraint to data creation, since it implies a condition of observing parts corresponding to the same risk branch over two consecutive times. In practice, this is not always the case.

— the exposure variable, which is used to add a weighting for the individuals included in the models. As the target variable is calculated based on average costs at two different points in time, exposure takes into account the minimum number of rooms used.

$$Exposition_{i,t} = min(Pieces\ number_{i,t}\ ; Pieces\ number_{i,t-1}) \tag{4}$$

for the part in risk situation i at time t.

Market information was added to the resulting base, in values and variations. As market indexes can take time to propagate to part prices, several temporalities were taken into account, with time lags ranging from zero to three semesters.

To start our modeling, we naturally selected a GLM that is easy to interpret, and then compared it with models that are usually more powerful, but also more complex to interpret : the decision tree, the random forest and histogram gradient boosting. Get over the difficulty in interpreting these models, the use of a proven SHAP explicability algorithm will be proposed. The results of these models are as follows :

	RMSE		MAE		R^2 (%)	
Model	test	train	test	train	test	train
Penalized GLM	**10.06**	9.43	7.79	7.12	**9.71**	9.31
Tree CART	10.09	9.54	7.79	7.24	9.14	7.22
HGB	**10.04**	9.14	7.70	6.86	**10.13**	14.89
Random forest	10.09	9.24	7.78	7.00	9.24	13.04

The table shows that :

— First of all, there is no overfitting phenomenon, thanks to cross-validation parameter optimization.

— Therefore, as expected, the more complex are the models, the better is the performance, but this is more visible on the R^2 and less so on the RMSE and MAE.

— Finally, the models are limited in performance. An analysis has shown that this limitation is due to a problem of data volumetry. In fact, this low volume does not allow us to construct explanatory variables specific enough to predict significant inflations.

On the other hand, all these models were converging. All highlighted the variables "vehicle brand", "vehicle age", "SRA price classes" and, to a lesser extent, "SRA groups" as explanatory variables. Below are the results obtained using the SHAP method on histogram gradient boosting :

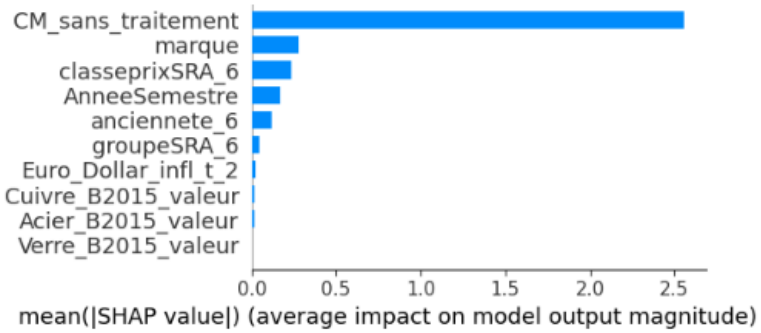


FIGURE 12 – Average absolute values of the most significant shap measurements

For explicability and ease of use reasons, we have chosen to use the penalized GLM model to quantify the discriminating factors in spare parts inflation.

The penalized GLM coefficients are shown in the following graphs, on which the legend has been removed for confidentiality reasons.

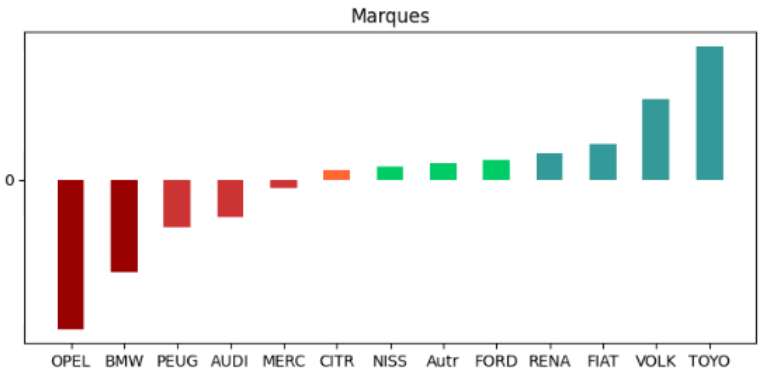


FIGURE 13 – Penalized GLM coefficients for brand modalities

For brands such as Opel and BMW, the coefficients are negative, meaning that they tend to have lower price trends than other brands. Whereas the coefficients for Toyota, Volkswagen and Fiat are positive, meaning that they would tend to increase the price of their car parts.

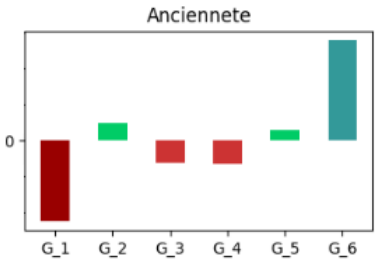


FIGURE 14 – Penalized GLM coefficients for seniority modalities

FIGURE 15 – Penalized GLM coefficients for class modalities

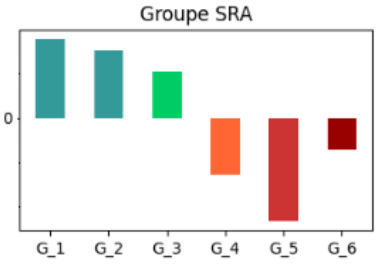


FIGURE 16 – Penalized GLM coefficients for SRA group modalities

The "age" and "SRA class" coefficients are increasing according to the modalities of the variables. In concrete terms, this means that recent vehicles increase their parts less than older vehicles, and vehicles in lower SRA price classes increase their parts less than vehicles in higher classes. Although, the coefficients of the SRA power group modalities are decreasing. This means that the more powerful is the vehicle, the lower is the price change.

Operational integration of results

The idea is to use the results of previous studies to plan an operational application for the Auto technical team.

In the pricing process, the Auto technical team start by formulating a hypothesis of inflation to be injected. Once the hypothesis has been defined, to differentiate this inflation between two vehicles with different characteristics, the coefficients obtained with the penalized GLM can be used.

However, fare policy choices will lead us to select only a reduced list of differentiating criteria compared to those defined by the study. For reasons of explicability and commercial positioning, adjustments may be made. These may concern the number of differentiating variables considered, the number of modalities, or the range of the coefficients.

Conclusion

Hence, this dissertation has contributed to the sophistication of the current pricing model by enabling inflation to be segmented according to contract profiles.

Firstly, a tool for monitoring modifications in the cost of spare parts enables the insurer to be alerted as early as possible when a change in manufacturers' pricing policies has

been detected. Thanks to this, the actuary can anticipate questions about the inflation hypothesis that he will inject into the projection models.

Secondly, a study was carried out to identify discriminating factors in the inflation of automotive parts costs. Despite the modelling difficulties encountered, weak signals were detected, and we showed that brand, age, SRA price class and SRA group were important elements in differentiating spare parts inflation.

At last, a strategy for using the penalized GLM obtained to differentiate inflation contract by contract based on these results is considered. The results can easily be used by the Technical Research Department, and after simplifications and adjustments due to commercial constraints, they can easily be integrated into Generali's pricing tools.

Remerciements

Avant de commencer ce mémoire, je tiens à adresser mes premiers remerciements à mon tuteur Yoann GOUYEN, responsable d'études d'actuariat, qui m'a admirablement accompagné tout au long de l'année. Il m'a apporté tous les outils nécessaires à la bonne réalisation de ce sujet grâce à ses précieux conseils et le temps qu'il m'a consacré. Merci pour tous ses retours constructifs qui m'ont permis de progresser dans les meilleures conditions.

Je souhaite également remercier Amandine GOMEZ, manager d'études d'actuariat pour son suivi régulier et ses recommandations sur mes missions.

Je tiens à exprimer ma reconnaissance à l'ensemble de l'équipe Wiz'You, supervisée par Etienne GUILLOU, responsable de la plateforme d'actuariat de Nantes et manager de l'équipe Wiz'You, ainsi que toute la plateforme de Nantes pour m'avoir accueillie chaleureusement et fait évoluer. Nos échanges tout au long de l'alternance ont été très enrichissants.

J'adresse également une pensée aux collaborateurs qui ont participé au sujet, que ce soit dans la construction, dans le suivi et les avis, jusqu'à la relecture. Yoann, Amandine, Marine, Alexandre, Vincent, Emeline, Freddie, Yassine, Eyrich, Jean-Sébastien, Lionel, Etienne, Rémi, merci pour votre accompagnement.

Mes remerciements s'adressent également à Franck VERMET, enseignant-chercheur et directeur de l'EURIA, Pierre AILLIOT, enseignant-chercheur, ainsi que le reste de l'équipe pédagogique pour la qualité des enseignements et l'accompagnement des étudiants dans leur projet professionnel.

Enfin, un grand merci à mes proches, ma mère, ma soeur et mes amis qui ont toujours été à mes côtés pour m'encourager dans mon parcours, leur inestimable soutien a été d'une grande aide pour ces dernières années d'études.

Table des matières

Introduction **1**

1 Contexte de l'étude **2**
1.1 Le marché et son inflation . . . 2
1.2 Le marché automobile . . . 4
1.3 L'assurance automobile . . . 12
1.3.1 L'assurance et le principe de mutualisation . . . 12
1.3.2 L'assurance automobile . . . 16
1.3.3 Le contexte de sur-sinistralité de 2022 . . . 19

2 Présentation des données **21**
2.1 Les données internes à Generali . . . 21
2.1.1 Description des données . . . 21
2.1.2 Sélection de variables pertinentes . . . 24
2.1.3 Traitement des données pour définir la base de travail . . . 24
2.1.4 Vérification de la qualité des données . . . 25
2.1.5 Définition du périmètre des pièces suivies . . . 28
2.1.6 Etude et traitement des séries temporelles . . . 30
2.1.7 Définition du maillage de l'étude . . . 31
2.2 Les données externes . . . 33

3 Détection automatisée et continue de changements de politiques tarifaires des constructeurs **39**
3.1 Présentation théorique des modèles utilisés . . . 40
3.1.1 Notions de séries temporelles . . . 40
3.1.2 Modèles de classification de séries temporelles . . . 41
3.1.3 Modèles de détection de ruptures sur les séries temporelles . . . 45
3.2 Approche historique . . . 48
3.2.1 Préparation des séries temporelles . . . 48
3.2.2 Application des modèles de détection de ruptures . . . 56
3.3 Approche opérationnelle . . . 58
3.3.1 Agrégation des courbes stables dans le temps . . . 59
3.3.2 Délai de détection de la rupture . . . 60

3.4 Conclusion sur ce détecteur d'évolution importante du prix 61

4 Identification des facteurs discriminants de l'inflation des pièces automobiles 63
4.1 Explication théorique des modèles d'explication utilisés 63
4.1.1 GLM pénalisé 64
4.1.2 Arbre de décision CART 66
4.1.3 Forêt aléatoire 68
4.1.4 Histogram Gradient Boosting 70
4.1.5 Métriques de comparaison des modèles 71
4.1.6 Méthode SHAP pour analyser les variables importantes des modèles de *Machine Learning* 73
4.2 Application des modèles 74
4.2.1 Construction de la base de modélisation 74
4.2.2 Résultats obtenus 86
4.2.3 Conclusion 91

5 Réflexion sur la mise en place opérationnelle 93
5.1 Utilisation des résultats pour sophistiquer l'intégration de l'inflation dans le processus de tarification 93
5.2 Adaptation des résultats opérationnellement 95

Conclusion 97

Annexes 105
Annexe 1 - Glossaire 105
Annexe 2 - Modèles de *datadrift* 107
Définition d'un *Model drift* et ses causes 107
Indicateurs de comparaison des données et tests statistiques 109
Modèles de Machine Learning 109
Annexe 3 - Modèles d'explication, les différents tests 111
A 3.1 - Différents traitements des valeurs aberrantes 111
A 3.2 - Tableaux des modalités des variables d'agrégation 114
A 3.3 - Encodage des variables 115

Introduction

Les assurances sont aujourd'hui des acteurs incontournables de l'économie mondiale. Les capitaux assurés représentent des sommes si importantes, que certaines assurances présentent des risques systémiques en cas de faillite. Par construction, ce monde assurantiel est soumis aux crises mondiales.

Ces dernières années, de nombreux événements ont bouleversé notre planète. Au commencement de l'année 2020, la pandémie mondiale de la covid-19 a ébranlé nos systèmes de santé, économies et habitudes. Pendant plusieurs mois, les confinements et couvre-feux ont rythmés le marché français et mondial, avec un impact notable sur l'offre et la demande. Au cours de l'année 2022, l'invasion russe d'une partie de l'Ukraine est venue déstabiliser le marché mondial entrainant une forte hausse des prix des matières premières qui s'est ensuite propagée à d'autres secteurs. Ces événements, entre autres, ont entrainé des modifications de comportement des agents économiques et par conséquent ont impacté les prix.

Dès lors, l'année 2022 a été une année particulièrement affectée par une inflation très forte, atteignant 5,2% en moyenne, un niveau non atteint depuis près de 30 ans. De plus, entre 2021 et 2022, le poste des pièces dans une facture de sinistre automobile moyenne a augmenté de 9,6%, alors que pour l'année précédente, celle-ci était de 2% seulement d'après l'association Sécurité et Réparations Automobiles.

La sur-sinistralité ainsi que le contexte inflationniste compliquent la tarification des assureurs. En effet, ces derniers doivent mettre à jour tous les ans les tarifs de leurs produits IARD en estimant la probabilité que les assurés aient un sinistre l'année suivante ainsi que le montant. On appelle cela un modèle "coût-fréquence".

Etant dans un cycle de production inversé, les données des années passées servent à la tarification future. Les données usuellement utilisées sont les montants globaux des sinistres sans distinguer les montants des différents postes (pièces, main d'oeuvre, ingrédients peintures). Pour cette étude, grâce à l'exploitation des rapports d'expertise automobiles des sinistres expertisés par Generali, les renseignements sont détaillés sur les réparations rendant la source riche d'informations.

Pour estimer le montant des coûts pour l'année que l'on tarifie, les actuaires émettent une hypothèse d'inflation. Cette dernière fait partie des ajustements sur le tarif technique qui permettent d'arriver au montant de la prime commerciale. Actuellement, elle est commune à tous les contrats d'un même produit.

De ce fait, l'objet de ce mémoire est de sophistiquer le modèle de tarification automobile en adaptant cette hypothèse d'inflation grâce à l'exploitation de ces rapports d'expertise. L'idée est de la rendre plus fine par une différenciation sur certains critères tels que la marque, l'ancienneté, etc. des véhicules.

Afin de répondre à ce sujet, ce mémoire se décompose en 5 parties.

Dans un premier chapitre, nous définissons le contexte de marché actuel avec quelques chiffres et définitions, avant de présenter le fonctionnement de l'assurance automobile et son processus de tarification.

Dans un second temps, la présentation des données internes et externes à l'entreprise, utilisées pour cette étude, sera réalisée en passant par le détail de ces données, leur analyse et la définition du périmètre utilisé. Ces données n'étant pas habituellement exploitées, une importante étude de cohérence a été réalisée.

Ensuite, les deux parties suivantes détailleront chacune les théories utilisées avant de développer leurs applications respectives.

— Grâce aux informations des rapports d'expertise, la première étape pratique de ce mémoire consiste à proposer un suivi de l'inflation des pièces automobiles sur les sinistres pour permettre au Bureau d'Etudes Techniques Auto d'anticiper l'impact de changements de politiques tarifaires des constructeurs automobiles.

— La partie qui suit s'intéresse à comprendre les facteurs discriminants de l'inflation du coût de la pièce automobile pour affiner le modèle de tarification automobile.

Et pour finir, la dernière partie est la réflexion opérationnelle de cette étude pour préparer l'intégration des résultats dans les modèles de tarification.

Chapitre 1

Contexte de l'étude

Cette première partie a pour objet d'introduire le contexte actuel qui permet d'expliquer la situation du marché automobile français et l'intérêt de l'étudier.

Pour cela, une première partie sera consacrée au contexte inflationniste de l'économie de marché dans un cadre général qui impacte beaucoup de domaines, notamment l'assurance.

Dans un second temps, une étude plus spécifique du marché de l'automobile sera réalisée. En particulier, une section sera dédiée aux parts de marché détenues par les marques afin de comprendre la répartition du marché automobile entre les différents constructeurs. Les regroupements financiers existants seront également étudiés car ils peuvent mener à des décisions financières communes. Les pièces détachées des véhicules et les variations de leur prix seront également définies.

Dans un troisième temps, des généralités plus proches de la pratique assurantielle seront abordées, permettant ainsi d'expliquer le fonctionnement de l'assurance automobile, ses principes et sa tarification.

1.1 Le marché et son inflation

L'économie de marché

Un marché est un lieu d'échange de biens ou de services entre un acheteur et un vendeur. Il en existe autant qu'il y a de variétés de biens et de services, mais on peut les regrouper en catégorie comme le marché de biens et de services qui contient celui de l'automobile. Tous ces marchés sont interconnectés et donc influencés les uns par rapport aux autres. C'est pourquoi il ne faut pas négliger les impacts potentiels de ceux-ci entre eux.

Une notion importante en théorie de l'économie ([25, 14]) est qu'en concurrence "pure et parfaite", c'est-à-dire sans monopole ou interférences à la concurrence, le prix provient de l'alignement entre l'offre et la demande. Ainsi, les prix des vendeurs rencontrent leurs acheteurs rendant possible des transactions. C'est ce mécanisme qui permet de fixer le prix d'un bien ou d'un service. Ce prix sera alors impacté par une modification de

l'offre des entreprises ou de la demande des consommateurs. C'est grâce à la présence de nombreux acteurs que la concurrence permet de maintenir une certaine stabilité et cohérence des prix avec ce système général.

L'inflation

L'inflation se décrit comme une augmentation générale et durable des prix des biens et des services qui a pour conséquence la perte de pouvoir d'achat de la population. Cette inflation est calculée à partir de l'indice des prix à la consommation (IPC) tous les ans par l'Institut National de la Statistique et des Etudes Economiques appelé INSEE.

Pour 2022, le taux d'inflation calculé pour l'ensemble des ménages en France (hors Mayotte) a atteint 5,2% en moyenne après avoir été compris entre -0,5% et 3,5% pendant plus de 30 ans [1]. Un tel taux n'avait pas été dépassé depuis les années 1980 (voir *figure 1.1*).

Une inflation excessive, tout comme son inverse, la déflation, est très mauvaise [6]. C'est pourquoi les banques centrales ont mis en place des politiques monétaires afin de mesurer, anticiper et maîtriser l'évolution des prix de leurs marchés. Il s'agit de la Banque Centrale Européenne (BCE) pour l'Europe et de la Banque de France pour la France, qui doivent mener ces politiques monétaires. La BCE notamment a pour mission fondamentale de maintenir une inflation d'environ **2% par an** dans les pays de la zone européenne.

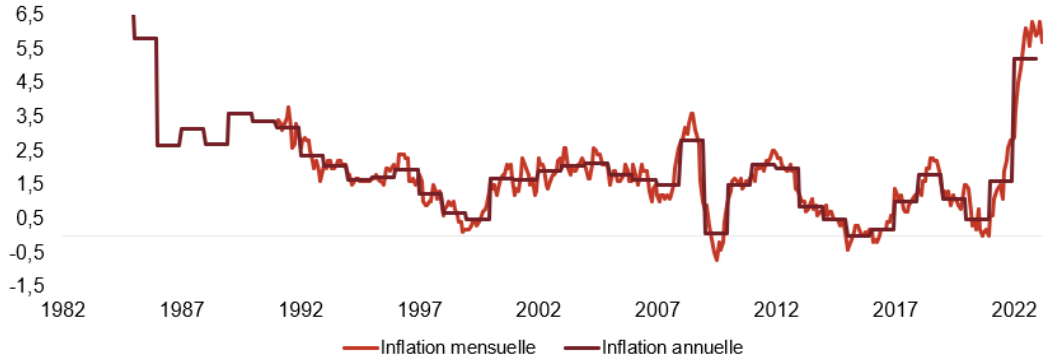


FIGURE 1.1 – Evolutions mensuelles et annuelles de l'indice d'inflation par l'INSEE (en % et en base 2015)

Mais pourquoi l'inflation a explosé en 2022 ?
Plusieurs facteurs rentrent en jeu [2] :

— la **pandémie** du Covid-19 et plus particulièrement les effets des confinements et des couvre-feux qui ont suivi. Cette épidémie a conduit de nombreux pays à

1. https://www.lemonde.fr/economie/article/2023/01/13/l-inflation-a-atteint-5-2-en-moyenne-en-2022-en-france-annonce-l-insee_6157692_3234.html
2. https ://www.vie-publique.fr/eclairage/286182-inflation-les-causes-de-la-soudaine-hausse-des-prix

fermer leurs marchés complètement ou partiellement pendant des périodes plus ou moins longues dans le but de ralentir la propagation du virus. Ces actions ont eu pour conséquences de ralentir les consommations des ménages et de les reporter à la réouverture des marchés. Ainsi, à la reprise des activités, la demande était plus forte que l'offre qui s'est trouvée dépassée dans certains domaines. Cette crise a également contribué à un changement des comportements, notamment avec l'arrivée du télétravail incitant la population à se tourner vers d'autres moyens de transports, ce qui a modifié la demande dans d'autres domaines.

— La **guerre en Ukraine** a fait augmenter les prix des matières premières qui étaient en partie exportées d'Ukraine (pétrole, gaz, huile et blé), mais aussi des pièces détachées produites en Russie ou Ukraine.

— Les politiques de **relances budgétaires massives** et les politiques trop **accommodantes** à la suite des crises des années 2010 mises en place par les Etats pour soutenir les pouvoirs d'achat ont fait augmenter la demande des populations.

— La **dépréciation de l'euro** par rapport aux monnaies étrangères a accentué l'inflation, en augmentant les prix des importations notamment. L'euro a atteint la parité avec le dollar américain en 2022 et a baissé en valeur par rapport au franc suisse avant de retrouver les taux d'avant entre 1,05 et 1,10.

Tous ces éléments favorisent une augmentation de la demande ou alors une diminution de l'offre par rapport à l'autre composante qui déséquilibre la relation du marché et fait augmenter les prix.

1.2 Le marché automobile

Comme l'ensemble des marchés, celui de l'automobile a subi les conséquences des crises mondiales depuis plus de 2 ans. Comme évoqués précédemment, il y a les conséquences de la crise sanitaire et de la guerre en Ukraine, avec par exemple la baisse des ventes de véhicules, de la production ou aussi des habitudes de transports.

Cependant, sur le marché automobile une crise supplémentaire a été observée, il s'agit de la pénurie de semi-conducteurs [3] qui a augmenté significativement les délais de livraison, allant jusqu'à plusieurs mois. Les semi-conducteurs sont des puces électroniques qui permettent aux appareils électroniques de bien fonctionner. A cause de ces pièces détachées particulières, l'offre de véhicules a fortement baissé en 2022 ce qui a accentué l'écart de l'offre et de la demande. L'association Sécurité et Réparations automobiles, autrement appelée SRA, a estimé l'inflation du poste de dépenses des pièces sur une réparation moyenne entre 2021 et 2022 de 9,6% alors que pour la facture totale, elle est estimée à 7,9% [4].

En plus des changements de comportements, grâce entre autres au télétravail et à

3. https://www.sneci.com/fr/blog/le-point-sur-la-crise-de-semi-conducteurs-et-son-impact-debut-2022/

4. Ces statistiques concernent les réparations, collisions et stationnement pour les véhicules particuliers et véhicules utilitaires de moins de 3,5 tonnes, chiffres de janvier 2023.

une volonté écologique, on peut également évoquer les réglementations qui suivent cette démarche vers une transition écologique et peuvent ainsi agir sur les comportements des utilisateurs :

— le **bonus écologique**, mis en place depuis 2007, qui correspond à une aide à l'achat d'un véhicule neuf ou d'occasion n'émettant pas de CO_2 ;
— la **prime à la reconversion**, depuis 2018, qui est aussi une prime pour l'achat d'un véhicule thermique, électrique ou hydrogène neuf ou d'occasion en échange de la mise en rebut d'un ancien véhicule ;
— un **prêt à taux zéro** sous conditions de ressources, à partir de 2023, pour l'achat d'un véhicule électrique ou hybride ;
— l'**interdiction de certains véhicules dans des zones spécifiques** appelées Zones à Faible Emission (ZFE), en fonction de leurs émissions polluantes en particules fines et oxydes d'azote, identifiées par les vignettes *Crit'Air*. Avec ces vignettes, les véhicules diesel seront interdits dans certaines zones à différents stades (voir *figure 1.3*). A Paris par exemple, à partir de 2024, les véhicules essence en 2030 et une fin des moteurs thermiques d'ici 2035.
Il y a 6 catégories de vignettes différentes :

1. La Crit'Air 1 pour les véhicules à essence immatriculés après 2011, au gaz ou hybride rechargeable ;
2. La Crit'Air 2 pour les véhicules au diesel immatriculés après 2011 ou à essence immatriculés entre 2006 et 2010 ;
3. La Crit'Air 3 pour les véhicules au diesel immatriculés entre 2006 et 2010 ou à essence immatriculés entre 1997 et 2005 ;
4. La Crit'Air 4 pour les véhicules diesel immatriculés entre 2001 et 2005 ;
5. La Crit'Air 5 pour les véhicules au diesel immatriculés entre 1997 et 2000 ;
6. Et la Crit'Air E ou 0 pour les véhicules électriques ou à hydrogène.

En 2021, les véhicules particuliers ne contenaient qu'une petite part de véhicules électriques ou hydrogènes (0,64% de véhicules en Crit'Air E en 2021, voir *figure 1.2*), ce qui va certainement augmenter dans les années à venir avec les réglementations en cours.

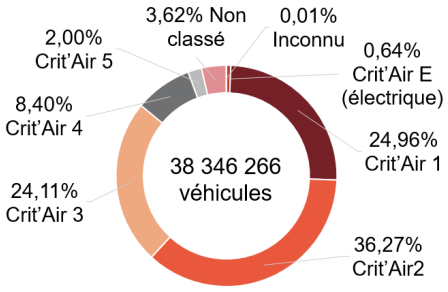


FIGURE 1.2 – Répartition des véhicules particuliers par vignettes Crit'Air en 2021 (Source)

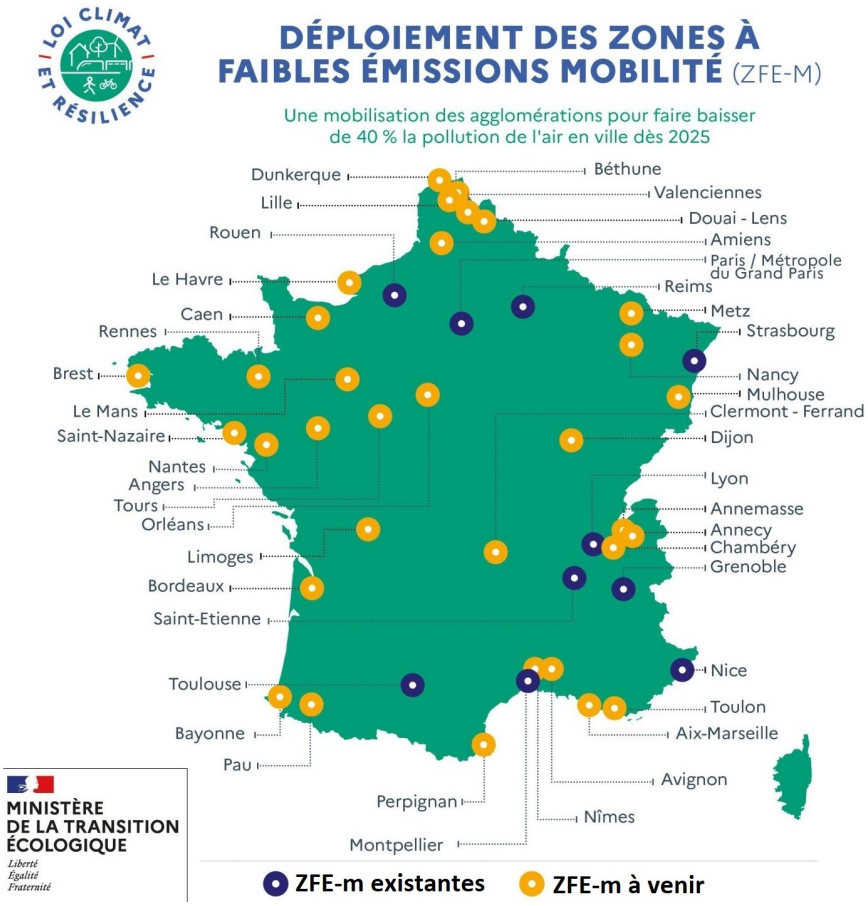


FIGURE 1.3 – Les zones ZFE actuelles et en cours en France

De façon générale, le parc automobile a connu une croissance régulière entre 2011 jusqu'en 2018, puis une tendance plus marginale par la suite, comme on peut le voir sur la *figure 1.4*. Ce parc a vieilli au cours des dernières années, passant de 9,7 ans en moyenne en 2015 à 10,5 années en 2022.

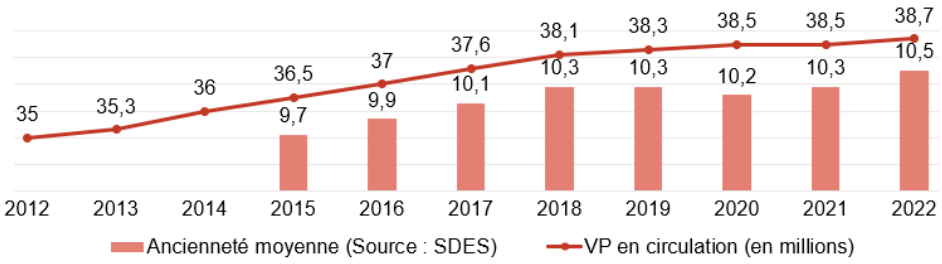


FIGURE 1.4 – Nombre et âge moyen des véhicules particuliers en circulation du parc français

Les parts de marché

Le marché français possède différentes marques de véhicules en proportions différentes. Les 18 premières marques immatriculées expliquent 92% des parts de marché français en 2022 (voir sur le *graphique 1.5*), parmi lesquelles 48% sont concentrées sur des marques françaises. Ce sont donc sur ces 18 marques qu'il faudra axer nos analyses par la suite.

On peut également voir que les marques les plus représentées sont Peugeot, Renault, Dacia et Citroën avec plus de 8% chacune. Certaines marques sont plus présentes dans d'autres pays comme Geely qu'on ne voit pas sur ces répartitions étant principalement vendue en Chine et très peu en France.

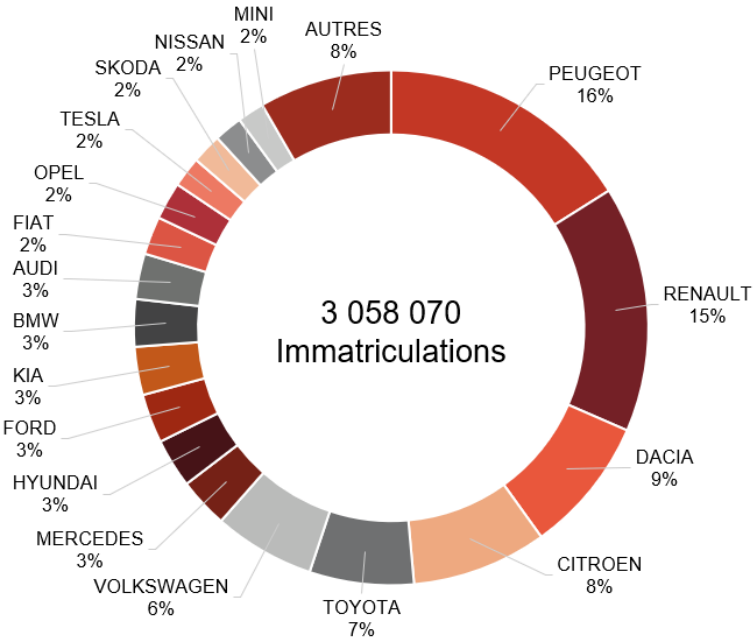


FIGURE 1.5 – Immatriculations des marques des véhicules particuliers du marché français en 2022 (Source : ccfa.fr/)

Les segmentations de marché : regroupements de marque

Dans ce mémoire, en plus de regarder les évolutions de prix par marque, il a été envisagé de considérer les évolutions de prix par groupe de marques de véhicules.

En effet, même si les politiques tarifaires peuvent sembler être prises au sein des différentes marques, on peut imaginer que des concertations au sein des regroupements de marques peuvent mener à des politiques communes.

Cependant, ces regroupements ont connu de nombreux changements au cours du temps et ainsi l'analyse des rapports d'expertise nécessite de connaître l'historique de

construction de ces regroupements de marques. C'est pourquoi, l'histoire de ces modifications est proposée par la suite.

Il en existe une douzaine actuellement qui sont composés de plus ou moins de marques dont le plus connu est **Stellantis** qui est l'alliance du 16 janvier 2021 entre le groupe Peugeot Société Automobiles (PSA) et Fiat Chrysler Automobiles (FCA). Seules les marques de véhicules de particuliers et les utilitaires ont été présentées ci-dessous.

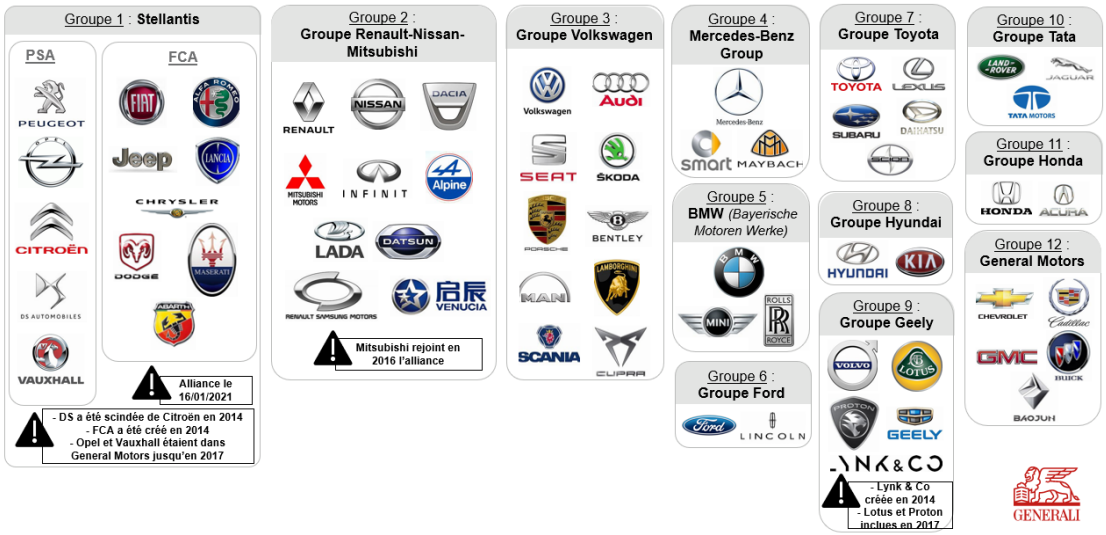


FIGURE 1.6 – Regroupements de marques

1. Le groupe Stellantis est la fusion de deux groupes comme évoqué précédemment.
   — Le groupe PSA a commencé avec **Peugeot** en 1965, puis **Citroën** a été racheté par la marque en 1976 et Chrysler Europe en 1978. Après cela, Citroën a créé **DS Automobiles** en 2014 qui était depuis 2009 une ligne haute gamme de véhicules de la marque. Enfin, en 2017, les marques **Vauxhall** et **Opel**, qui appartenaient à General Motors respectivement depuis 1925 et 1929, ont à leur tour rejoint le groupe.
   — Le groupe FCA a été créé par Fiat S.p.A en janvier 2014 avec le rachat complet du groupe Chrysler, après avoir été un actionnaire majoritaire en juin 2009. A cette date, le groupe Fiat S.p.A contenait **Abarth** (1971), **Alfa Romeo** (1986), **Fiat** et **Lancia** (1969), et Chrysler Group avait **Chrysler**, **Dodge**, **Jeep**, **Ram Trucks**. **Maserati** est une marque toute particulière du groupe FCA puisque, après avoir été reprise par Fiat en 1987, elle a été fusionnée à Ferrari en 1997, puis scindée en 2006. Et Ferrari a quitté FCA en 2015.

2. Le groupe Renault-Nissan-Mitsubishi débute en 1999 par une volonté de stratégie commune et des synergies entre **Renault** (avec **Alpine** (1965), **Dacia** (1999)) et **Nissan** (avec **Infiniti** (depuis 1989), **Datsun** (1933), **Venucia** (2010)), détenue

à parts égales. En 2008, le groupe obtient AvtoVAZ, autrement appelé **Lada** qu'il cèdera à NAMI pour 1 rouble symbolique en 2022. Enfin, **Mitsubishi** a rejoint le groupe au cours de l'année 2016.

3. Le groupe **Volkswagen** a entamé son agrandissement par l'achat d'**Audi** (1965) et NSU Motorenwerke AG (1969), puis **Skoda** (1991), **Bentley**, **Bugatti** et **Lamborghini** (1998), **Scania** (2008) et **Porsche** (2009).

4. En ce qui concerne Mercedes-Benz Group, il a initialement été la fusion entre Daimler-Benz et Chrysler de 1998, qui a ensuite perdu Chrysler en 2007 et s'est fait renommé Mercedes-Benz Group pour se concentrer sur sa marque principale Mercedes. Aujourd'hui, le groupe possède donc **Mercedes**, **Maybach** (acheté par Daimler-Benz en 1960) et **Smart** (partenariat depuis 1994).

5. Le groupe Bayerische Motoren Werke AG (BMW) vient de la marque **BMW** qui a pris le contrôle du groupe Rover en 1994 composé de Rover, Morris Garage, Mini et Land Rover, mais se sépare de Rover et Morris Garage en 2000, possédant actuellement uniquement les marques **Mini** et **Rolls Royce** en plus de BMW. En 2026, le groupe devrait également intégrer la **Alpina** d'après un accord signé en 2022.

6. Le groupe Ford est composé de **Ford** et **Lincoln** actuellement, alors qu'en 2006, il contenait huit marques : Ford, Aston Martin (cédée en 2007), Land Rover et Jaguar (vendus en 2008 à Tata Motors), Lincoln, Mercury (production arrêtée en 2010), Volvo (vendue en 2010 à Geely) et Mazda (parts d'actionnaire majoritaire vendues en 2010).

7. Le groupe **Toyota** commence lui aussi ses fusions par la marque homonyme qui a pris le contrôle de **Daihatsu** en 1967 puis créa une marque de luxe **Lexus** en 1987. **Scion** a été créée par le groupe en 2003 sur les marchés états-unien et canadien et a disparue depuis 2016. Enfin, C'est un actionnaire important de **Subaru** depuis 2005, après avoir racheté les parts à General Motors.

8. Le groupe Hyundai est un peu particulier puisqu'il s'agit d'un conglomérat sud-coréen réunissant **Hyundai Motors** et **Kia Motors** depuis le 29 mars 1999, après que Kia ait été déclarée en banqueroute.

9. Le groupe Zhejiang Geely Holding, plus couramment appelé Geely, vient de l'acquisition par **Geely Auto** de **Volvo** en 2010 (achetée à Ford), **Proton** et **Lotus** en 2017. En 2014, Geely et Volvo ont créé une marque de véhicules commercialisés uniquement en ligne nommée **Lynk & Co**, produites en Belgique et en Chine pour être livrées dans le monde entier.

10. Le groupe Tata créé par la marque indienne **Tata Motors** s'est agrandi avec l'achat de **Jaguar** et **Land Rover** au groupe Ford en 2008.

11. Le groupe Honda est uniquement composé d'**Honda** et de sa marque haut de gamme **Acura**.
12. Et pour finir le groupe General Motors commence par la **General Motors Corporation**, ou GMC, intègre la **Cadillac** en 1909 après de longues négociations, **Buick** puis **Baojun** en 2010. Aujourd'hui, après avoir cédé Saab en 2010, Opel et Vauxhall à PSA en 2017 et retiré Daewoo en 2005 et Chevrolet en 2015 du marché européen, le groupe n'a plus une présence importante actuelle en Europe.

Comme on peut le voir dans l'histoire de tous ces regroupements, ils n'ont pas été stables pendant la durée de l'historique de l'étude, c'est un aspect sur lequel il faut garder une attention toute particulière pour la suite.

Les pièces détachées

Une pièce détachée automobile est une pièce de remplacement de l'ensemble d'une voiture qui permet de remplacer une pièce défectueuse dans la voiture (voir *figure 1.7*).

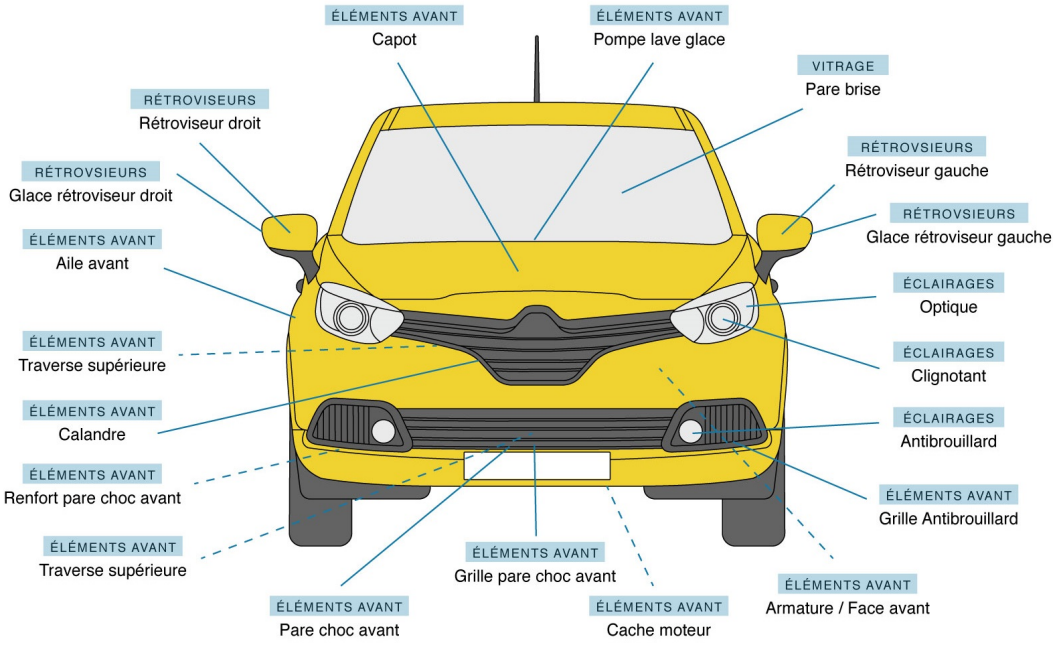


FIGURE 1.7 – Exemple de pièces détachées avant d'une Renault Captur

En 2022, ce type de pièces a subi une inflation de l'ordre de 12% d'après le groupe IDPL ([11]), chiffre réalisé sur toutes les pièces référencées de leurs 30 premiers fournisseurs. Pour limiter l'augmentation de prix de ces pièces, plusieurs alternatives sont possibles comme l'utilisation :

— de pièces de distributeurs (MDD) qui peuvent s'avérer entre 20% à 25% moins chers que les pièces dites premium. Cette possibilité s'est ouverte avec la loi Cli-

mat et Résilience de 2021 permettant aux acteurs de s'approvisionner en pièces détachées par d'autres moyens que par les constructeurs [5] ;

— de pièces issues de l'économie circulaire. Une pièce de réemploi [2], autrement appelé pièce issue de l'économie circulaire (PIEC), est une pièce automobile d'occasion qui provient d'un centre de véhicules hors d'usage (VHU) agréé par l'Etat ou est remise en état selon les spécifications du fabricant sous l'appellation échange standard. Par définition elle coûte moins cher qu'une pièce neuve, provenant du marché secondaire de l'occasion. Depuis le 1^{er} janvier 2017, les garagistes doivent proposer à leurs clients de réparer leurs voitures avec des pièces neuves et d'occasion (précisé dans le décret n°2016-703 du 30 mai 2016 [27]). Les PIEC représentaient 3,6% des pièces remplacées en 2021 d'après l'association Sécurité et Réparations Automobiles (SRA) [9].

Ainsi, les prix des pièces détachées sont variables. Cela peut venir de l'aspect neuf ou d'occasion de la pièce, de si elle provient d'un concessionnaire de la marque du véhicule ou d'un autre moyen de distribution, mais elle peut aussi dépendre de la technologie de la pièce. En effet, sur un même modèle de véhicule, une pièce détachée peut avoir des prix très différents selon la technologie utilisée sur celle-ci. Le cas le plus représentatif est celui des phares.

Ci-dessous, la liste des différentes technologies de phare [6] :

1. L'ampoule halogène qui est la technologie la plus ancienne. Cette ampoule fonctionne via son filament métallique qui chauffe pour créer la source lumineuse. Elle est encore très utilisée aujourd'hui notamment par son coût très avantageux, sa longue durée de vie et son adaptabilité sur les voitures.

2. L'ampoule Xénon qui correspond à une lumière très puissante et intense blanche. Elle est nommée ainsi puisqu'elle fonctionne par l'excitation du gaz Xénon présent dans l'ampoule avec le courant électrique qui va générer la lumière. Cette ampoule est surtout utilisée sur les véhicules haut de gamme étant chère par la rareté du gaz Xénon qui la compose.

3. Le phare LED (Light Emitting Diode ou Diode Electroluminescente) est plus récent. Il est encore très peu utilisé dans les phares aujourd'hui à cause des surchauffes de la pièce qui peut s'enflammer et du prix de la pièce qui est aussi assez coûteux. Ce type d'ampoule émet de la lumière par les diodes électroluminescentes lorsqu'elles sont traversées par le courant. Les gros avantages de cette technologie sont qu'elle est faible en consommation d'énergie, avec un éclairage puissant sans être éblouissant (voir *figure 1.8*) et est facile à adapter à toutes formes.

4. Le phare laser est lui aussi très récent. La source de lumière, c'est-à-dire le rayon laser, est créée grâce à un miroir qui concentre l'énergie dégagée des atomes. Ce

5. Source : https://www.vroomly.com/blog/-hausse-prix-pieces-detachees-comment-laffronter/

6. Sources : https://www.lesnumeriques.com/voiture/halogene-xenon-led-laser-revolution-phares-a2705.html ; https://www.avatacar.com/blog/mecanique-auto/les-differents-types-eclairage-voiture/

type de phare est également très onéreux mais permet d'éclairer sur de longues distances.

5. L'ampoule OLED (Diode Electroluminescente Organique). Elle fonctionne comme la lampe LED mais avec des diodes en matières semi-conductrices organiques composées de couches superposées. Même si l'OLED est moins lumineuse que la LED, cette source de lumière est plus économe et résistante à de fortes températures. Cette ampoule est surtout utilisée pour les feux de signalisation ou à l'arrière des voitures.

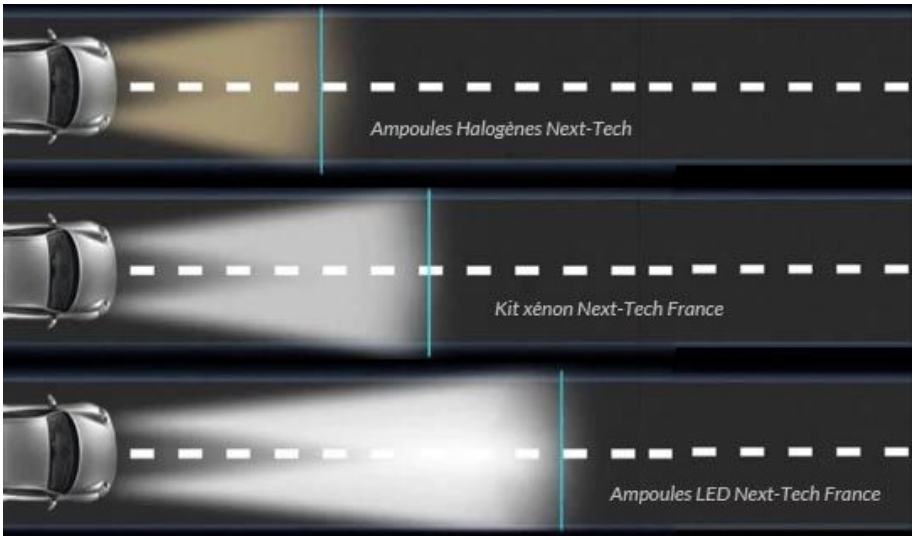


FIGURE 1.8 – Différences d'éclairages entre les phares halogène, Xénon et LED (Source : next-tech-france.com

1.3 Le fonctionnement de l'assurance automobile et son contexte de marché

Avant d'expliquer pourquoi les primes d'assurance automobiles augmentent en ce contexte de marché, expliquons comment fonctionne l'assurance puis l'assurance automobile avec sa tarification.

1.3.1 L'assurance et le principe de mutualisation

L'assurance

Une assurance est un engagement échangé entre un assureur et un client, appelé assuré, par lequel l'assureur s'engage à indemniser l'assuré de certains risques aléatoires en échange d'une prime, tout cela défini à l'avance à travers un contrat nommé la "police d'assurance".

L'inversion du cycle de production

Dans un échange traditionnel commercial, le client verse une somme en l'échange d'un bien ou d'un service dont le coût est connu par le fabricant ou le prestataire de service.

Au contraire, l'assurance a pour particularité d'encaisser la prime avant de connaître le montant des sinistres et des prestations qu'elle aura à indemniser dans l'année suivante. Ce coût de prestation qui n'est pas connu au moment de la souscription définit le mécanisme d'inversion du cycle de production.

FIGURE 1.9 – Principe d'inversion du cycle de production

L'asymétrie d'informations

Des individus rationnels qui maximisent leur utilité, c'est-à-dire leur intérêt personnel, sont donc prêts à avoir des comportements opportunistes qui risquent de compromettre le fonctionnement efficace du marché. On peut distinguer deux situations d'informations asymétriques :

1. l'antisélection ou la sélection adverse est due à un problème d'asymétrie d'informations qui se déclare au moment de la signature du contrat. C'est le cas des assurés qui ont de forts problèmes de santé et qui souscrivent à des produits qui ne demandent pas de dossier médical ;
2. l'aléa moral se produit si l'assuré change de comportement après la signature du contrat. Par exemple un client qui vient de souscrire à un contrat d'assurance habitation avec garantie vol peut inconsciemment ou consciemment baisser sa garde sur la fermeture de sa porte.

Le principe de mutualisation

Le principe fondamental de l'assurance est la mutualisation des risques.

En effet, l'assurance permet de partager le risque des assurés pour le compenser. C'est-à-dire que tous les assurés paient une prime, relativement modeste en comparaison du risque financier, qui servira à indemniser les sinistres d'une minorité d'assurés. Cela nécessite pour les assurés d'accepter de payer pour les autres sachant qu'un jour cela pourra être pour eux-mêmes et pour des sinistres qu'ils ne pourraient pas forcément régler sans aide. C'est un transfert de risque vers une mutualité de risque à travers une mécanique de solidarité.

Pour déterminer les montants de ces primes d'assurance, les assureurs considèrent le risque aléatoire assuré par le client pour définir son profil de risque. On entend par même risque aléatoire un risque [7] :

— homogène qui est de même nature avec la même probabilité de se réaliser et même montant moyen,
— dispersé dans le sens où il faut essayer de répartir au mieux les biens assurés sur les zones géographiques pour éviter le risque d'accumulation.
— divisé pour qu'un sinistre ne puisse pas menacer le principe de mutualité,
— non volontaire et futur pour que le risque reste aléatoire.

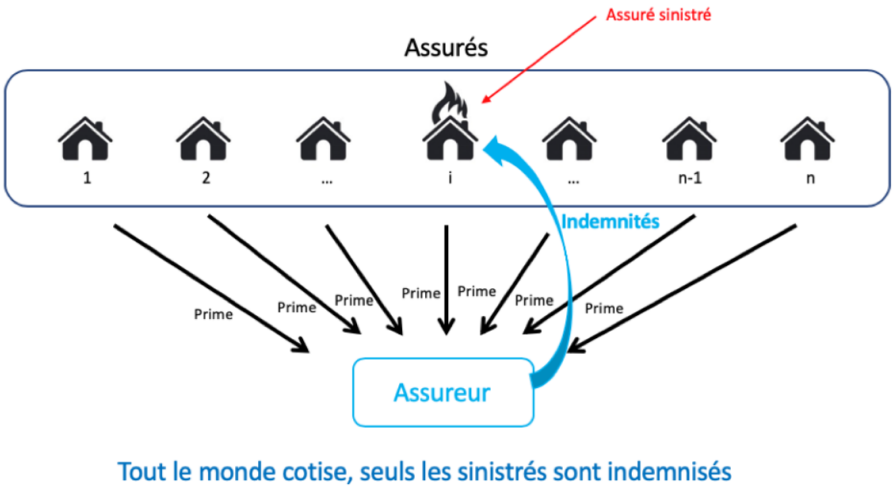


FIGURE 1.10 – Principe de mutualisation *openclassrooms.com*

Cette mutualisation doit être modérée pour éviter d'un côté la sur-mutualisation et de l'autre la personnalisation :

1. La sur-mutualisation consiste à associer une prime moyenne à tous les risques dans le portefeuille. Le risque dans ce cas est d'attirer les "mauvais risques" qui coûtent plus cher que les primes que l'assureur reçoit de leurs parts et de perdre les "bons risques" qui auraient une prime plus élevée par rapport au reste du marché.
2. La personnalisation consiste à associer une prime unique à chaque police du portefeuille. Elle empêche le principe fondamental de l'assurance qui est la mutualisation des risques. Cela signifie que les primes des assurés seraient plus hétérogènes par manque de partage des risques. De plus la personnalisation nécessite un travail plus important de tarification pour considérer chaque cas particulier ce qui est coûteux pour l'assureur.

7. Source : https://www.economie.gouv.fr/facileco/comprendre-assurance-risques

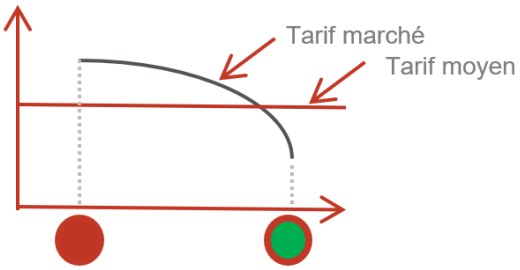


FIGURE 1.11 – Le risque de sur-mutualisation

Ainsi, la mutualisation "cible" doit ainsi être un compromis entre ces deux niveaux en créant des risques homogènes.

Cette segmentation a pour objectif de discriminer la population en sous populations de **risques homogènes**. Par exemple, en assurance automobile, on peut envisager une prime de base à 40€ qui sera ajustée en fonction du nombre de conducteurs principaux sur le véhicule, de l'ancienneté du transport et de la zone géographique dans lequel le bien est principalement utilisé et garé. Ainsi, dans l'exemple *1.12*, une voiture avec un seul conducteur principal, qui a entre 2 et 5 ans, qui se situe dans la première zone du zonier restera à 40€ de prime, alors qu'un véhicule pour 2 conducteurs principaux avec un jeune conducteur, pour un véhicule de plus de 10 ans et en zone 5 sera à 110,40€.

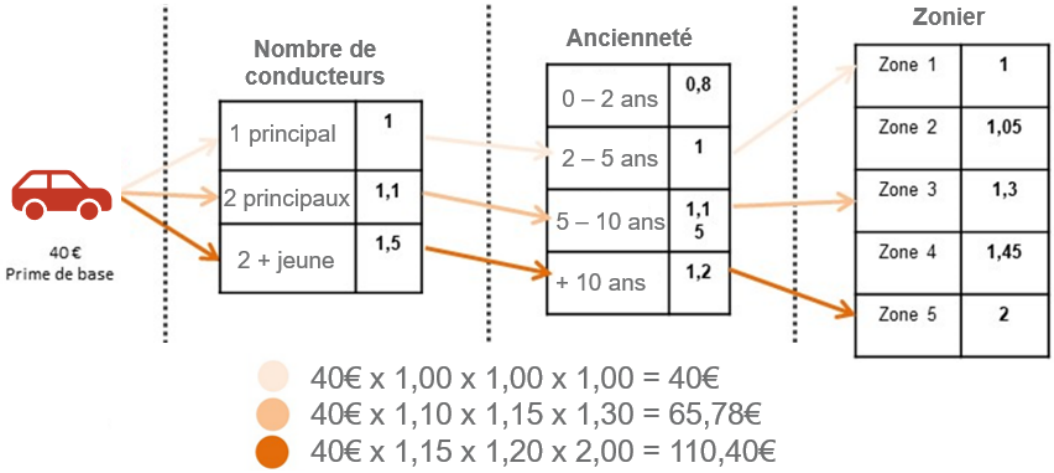


FIGURE 1.12 – Exemple simplifié de multiplication pour obtenir un risque homogène

Toutes les compositions différentes correspondants aux risques homogènes sont représentées dans des matrices avec les coefficients à appliquer à chaque étape pour simplifier la manière d'adapter la prime avec chacun de ces critères.

Le caractère obligatoire de certaines assurances

L'assuré a le choix de souscrire à certains produits d'assurance comme l'assurance-emprunteur, l'assurance-vie ... ou alors au sein du même produit de prendre telle ou telle garantie, sauf pour le cas de 3 types d'assurance en France qui sont obligatoires [8] :

1. l'assurance **responsabilité civile** qui protège l'assuré en cas de dommages matériels, immatériels ou corporels causés à une tierce personne. Parmi ces assurances, on retrouve la partie de l'**assurance automobile obligatoire** qui couvre les dommages, matériels ou corporels, de l'assuré causés auprès d'un autre automobiliste, d'un piéton ou d'un tiers à bord du véhicule (Article L.211-1 du Code de la Route). En cas de défaut de cette assurance, le propriétaire de la voiture s'expose à 3.750€ d'amende, une peine d'emprisonnement et un retrait de permis selon la gravité du sinistre.
2. le régime obligatoire de l'**assurance maladie** qu'on connait sous le nom de sécurité sociale.
3. l'**assurance habitation** pour un locataire (Loi "Alur" pour l'accès au logement et un urbanisme rénové du 26 Mars 2014).

1.3.2 L'assurance automobile

L'assurance automobile est un produit d'assurance qui permet de couvrir les dommages liés aux sinistres sur les biens automobiles et les personnes. Seule la couverture au tiers est obligatoire, le reste est optionnel, sous forme de garanties supplémentaires pour l'assuré qui choisit "à la carte" ce qu'il souhaite couvrir (par exemple formule tout risque, garantie bris de glace, ...).

L'assurance automobile fait partie du segment IARD de l'assurance, sa tarification suit le même schéma que d'autres assurances IARD.

Tarification d'une assurance automobile

La tarification auto repose sur la probabilité d'avoir un sinistre dans l'année qui suit et le montant moyen du sinistre auquel on s'attend. Ces critères sont ceux qui déterminent le niveau de la cotisation annuelle ou mensuelle d'assurance. Pour cela, ces statistiques se basent sur les caractéristiques :

1. du conducteur avec l'année du permis, son âge, ses antécédents sur la route grâce à son relevé d'informations etc.
2. du véhicule avec :
   — sa puissance, son prix d'achat, le prix de réparation de ses pièces, son année de mise en circulation,
   — son usage professionnel et/ou personnel, qu'il soit régulier ou bien ponctuel,

8. https://www.aide-sociale.fr/assurances-obligatoires/

3. la zone géographique de circulation et stationnement du contrat ou le véhicule auquel appartient le bien assuré qui correspondent à des variables latentes. Ces variables sont des informations obtenues grâce à des modèles statistiques ou algorithmiques qui différencient le risque,
4. des garanties souscrites comme la garantie bris de glace ou l'assistance 0 kilomètres.

Objectif :
L'objectif de la tarification est de segmenter le portefeuille en classes de risques homogènes afin de mutualiser le risque. Avec ces différentes combinaisons de risques obtenues, une prime pure y est associée pour couvrir la charge de sinistres uniquement. Pour arriver au tarif "final", un ajustement avec les chargements annexes, comme les frais de gestion, frais d'acquisition ou autres, est appliqué.

Processus :
Le processus de tarification classique en assurance IARD (Incendie, Accident et Risques Divers), et par conséquent en assurance automobile, se fait en 5 étapes clés :

1. **la collecte de données** avec les données contrats (exposition, adresse, âge du conducteur...), sinistres (nombre et montants de sinistres par garantie) et externes pour expliquer l'effet spatial dans le modèle.

2. le **traitement des données** pour rendre homogènes toutes ces données récoltées, en supprimant les valeurs aberrantes, en rehaussant les variables pour des données manquantes ou encore en contrôlant la cohérence des données... Pour ce dernier cas on peut par exemple regarder si le bonus-malus est cohérent avec l'âge du conducteur.

3. l'**évaluation du risque spatial** par la création de « zonier » par garantie à une maille spatiale choisie (régions, communes, ...)

4. le **calcul de la prime pure**, autrement appelé la cotisation pour l'assuré, correspondant au coût du risque.

   Décrivons mathématiquement le calcul de la prime pure.

   Soit un portefeuille d'assurance de N polices d'assurance. La prime d'assurance Π_i du i^{eme} contrat, $i \in \{1, ..., N\}$, peut s'écrire $E(S_i) + MR(S_i)$ sous réserve d'avoir un portefeuille suffisamment grand d'après la loi forte des grands nombres, avec les S_i la charge de sinistres associée au i^{eme} contrat, E(.) l'espérance et MR(.) les chargements techniques associés à la prime d'assurance. Ainsi, si la loi du nombre de sinistres I est indépendante de la loi du coût Y et les Y_i sont indépendants et identiquement distribuées, la charge totale du portefeuille peut s'écrire $S_N = \sum_{i=1}^{N} Y_i$.

   On obtient le résultat fondamental à la tarification suivant :

$$E(S) = E(I) \times E(Y) = \text{fréquence} \times \text{coût moyen} \tag{1.1}$$

C'est à dire que la charge de sinistres espérée s'exprime en fonction de la probabilité d'avoir un sinistre, soit la fréquence, et le montant moyen du sinistre attendu.

Par exemple, si en automobile il y a en moyenne un sinistre tous les cinq ans avec un coût-moyen pour le profil de risque est de 2 000€, alors la prime pure sera $PP = \frac{1}{5} \times 2000€ = 400€$.

Ce modèle qu'on appelle le modèle coût-fréquence sera en réalité modélisé par deux sous-modèles. Un modèle qui suit une loi de Poisson pour le nombre de sinistres et un modèle de loi Gamma pour les montants.

5. l'ajustement au **tarif commercial** qui intègre les frais de fonctionnement de l'assurance et des apporteurs. Il s'agit du montant payable au début de chaque période de garantie, afin que l'assurance dispose à l'avance de fonds, mais aussi que l'assuré matérialise son engagement avant que le risque survienne.

   L'assurance automobile est particulière puisqu'il s'agit d'un produit très concurrentiel. En effet, les assureurs sont amenés à proposer des prix attrayants qui sont en dessous de ce qu'on appellerait le "coût de production" pour un bien, c'est ce qu'on appelle un **produit d'appel**. Cela signifie que le but de ce produit est d'attirer des clients qui vont souscrire à ce produit qui est déficitaire mais aussi à un ou plusieurs autres qui permettent d'équilibrer les relations, voir faire une marge. Ainsi, l'engagement de l'assureur est supérieur au montant des primes perçues par son assuré, comme on peut le voir sur la *figure 1.13*. L'assureur se retrouve désavantagé par ce type de produit en lui seul.

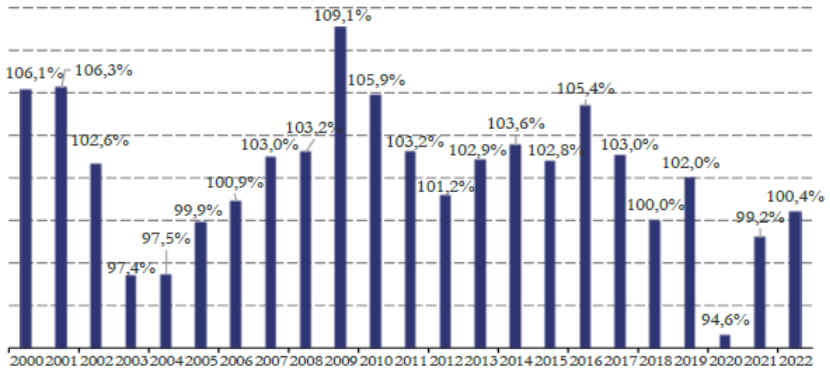


FIGURE 1.13 – Ratio combiné comptable net de réassurance en automobile (Sources : *ACPR - France Assureurs*)

Par conséquent, au-dessus de ce tarif technique (calcul de la prime pure précédent), des modifications commerciales sont réalisées pour arriver au tarif final dit

commercial. Ces choix sont propres aux entreprises. On y compte l'intégration des charges d'acquisition et de gestion, une hypothèse d'inflation, le coût du capital et une éventuelle marge.

$$Cotisation = Prime\ pure + Chargement + Marge\ beneficiaire \tag{1.2}$$

avec :

$$\begin{cases} Chargement = \text{Frais d'acquisition, frais de gestion et d'encaissement de primes et frais de règlement des sinistres} \\ Marge\ beneficiaire = \text{marge bénéficiaire du coût en capital et de la marge additionnelle.} \end{cases}$$

Si on reprend l'exemple précédent avec une prime pure de 400€ et un taux de chargements de 25%, alors la prime d'assurance technique sera de $\Pi = 400€ \times 1,25 = 500$.

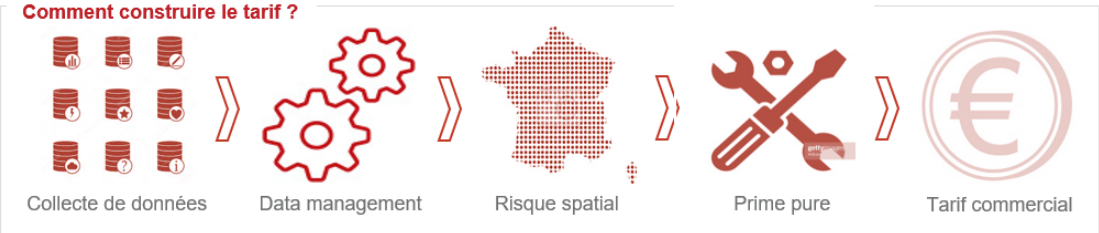


FIGURE 1.14 – Différentes étapes en tarification IARD (Incendie, Accidents et Risques Divers)

Actuellement une hypothèse d'inflation est réalisée et appliquée de manière uniforme sur tous les contrats.

L'objectif principal de ce mémoire est de transformer cette variable uniforme sur tous les contrats en variable différenciée selon des caractéristiques du véhicule identifiées grâce aux informations des rapports d'expertise auto des assurés de l'entreprise.

1.3.3 Le contexte de sur-sinistralité de 2022

En plus de l'inflation qui a fortement augmenté ces derniers mois, 2022 a été fortement touchée par un contexte de sur-sinistralité.

En effet, depuis 1999 avec les fortes tempêtes Lothar et Martin, 2022 représente la pire année des assureurs français en termes de sinistralité liée aux catastrophes naturelles, atteignant 10 milliards d'euros sur le marché français [9].

9. Article : france-les-catastrophes-naturelles-2022

Elle est l'une des conséquences du dérèglement climatique qui intensifie la gravité des événements climatiques de type grêles, tempêtes, inondations, ... mais aussi augmente leur fréquence. D'après une étude de France Assureurs [3], le coût des sinistres liés aux aléas climatiques devrait continuer de croître ces 30 prochaines années. Ils estiment que pour la période 2020 à 2050, le coût des sinistres climatiques pourrait atteindre 143 milliards d'euros en cumulé alors que pour la période 1989 à 2019 il était de 69 milliards. Cela représente une hausse de près de 93%.

Que ce soient les grêles, le péril essentiellement touché par la charge climatique en auto, les tempêtes ou encore les épisodes de sécheresses qui s'intensifient, ces aléas climatiques ont des fortes conséquences sur l'assurance automobile. Par exemple, les grêlons de plus en plus gros lors d'épisodes de grêles brisent un plus grand nombre de pare-brises et toits d'automobiles, et parfois plus intensément. Les tempêtes, quant à elles, fragilisent les arbres ou infrastructures qui peuvent céder sur les voitures assurées.

Ainsi, si le **coût** et la **fréquence** des sinistres automobiles **augmentent**, alors il est important pour tous les assureurs d'**anticiper ces charges supplémentaires** afin de pouvoir conserver un équilibre dans la relation "assureur-assuré".

Chapitre 2

Présentation des données

Pour mettre en place cette étude, on distingue deux catégories de données qui ont été utilisées :

— d'un côté les données internes à Generali qui correspondent aux données accessibles uniquement sur les systèmes d'informations de l'entreprise

— et de l'autre côté les données externes qui sont des données dites en *Open source* disponibles par tous.

Présentons-les.

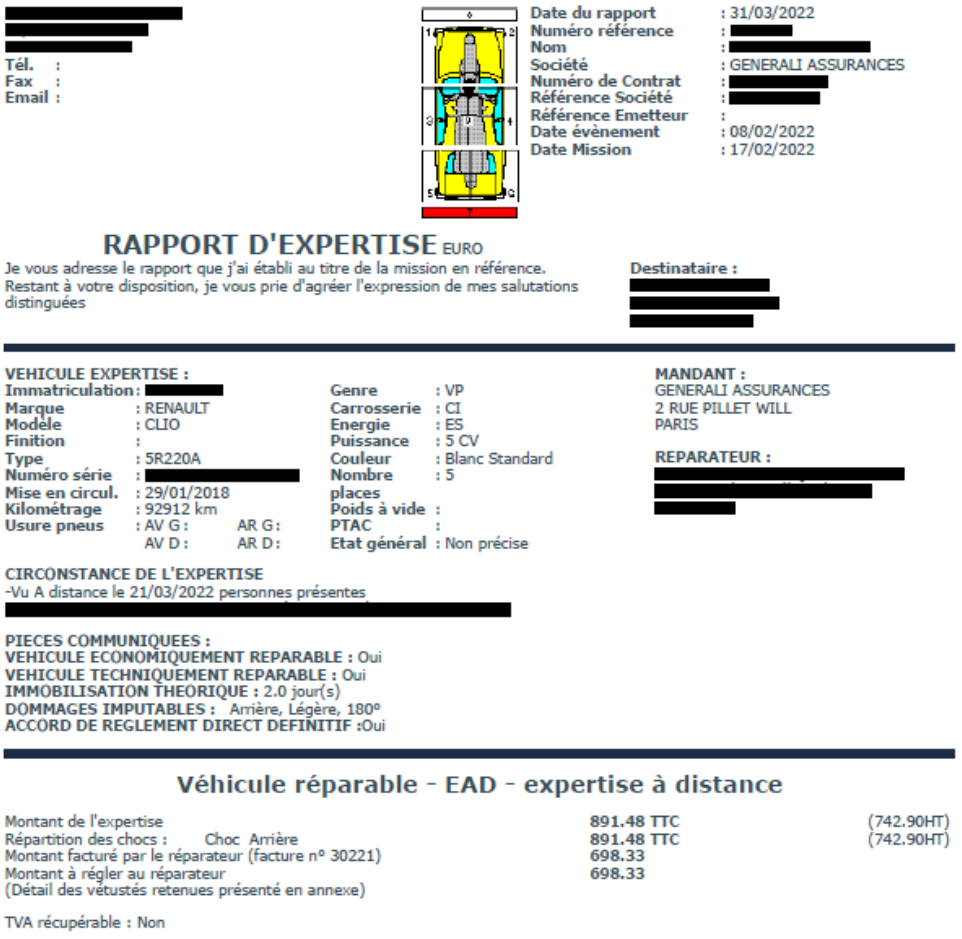
2.1 Les données internes à Generali

2.1.1 Description des données

Lorsqu'un assuré a un sinistre sur son véhicule et veut se faire indemniser par son assurance automobile, celle-ci peut demander à un expert de se déplacer pour constater l'importance des dégâts et vérifier que les réparations concernent bien les altérations du sinistre déclaré. De cette manière, elle arrive à identifier certaines fraudes tout en pouvant regarder le détail des réparations à réaliser sur le véhicule.

Ces déplacements font l'objet de **rapports d'expertise** qui contiennent toutes les informations sur le véhicule sinistré, l'assuré, le garage qui va réaliser les réparations (voir exemple *figure 2.1*) et sur le détail des réparations. Celui-ci est séparé en fonction des coûts des différents postes en auto que sont les pièces, les opérations de main d'oeuvre et les ingrédients de peinture (voir exemple *figure 2.2*), mais aussi en détaillant les montants de chaque élément de ces postes (voir exemple *figure 2.3*).

Les rapports d'expertise fournissent également des informations sur les montants de TVA appliqués et les montants déduits après vétusté. Comme la vétusté concerne l'usure des biens matériels par le temps, seules les pièces sont éligibles à cet abattement. Ainsi, en fonction du vieillissement du véhicule, si on considère que le véhicule n'est plus en bon état, alors des remises peuvent être appliquées.



[redacted]
Tél. :
Fax :
Email :

Date du rapport : 31/03/2022
Numéro référence : [redacted]
Nom : [redacted]
Société : GENERALI ASSURANCES
Numéro de Contrat : [redacted]
Référence Société : [redacted]
Référence Emetteur :
Date évènement : 08/02/2022
Date Mission : 17/02/2022

RAPPORT D'EXPERTISE EURO

Je vous adresse le rapport que j'ai établi au titre de la mission en référence.
Restant à votre disposition, je vous prie d'agréer l'expression de mes salutations distinguées

Destinataire : [redacted]

VEHICULE EXPERTISE :
Immatriculation : [redacted]
Marque : RENAULT
Modèle : CLIO
Finition :
Type : 5R220A
Numéro série : [redacted]
Mise en circul. : 29/01/2018
Kilométrage : 92912 km
Usure pneus : AV G : AR G : AV D : AR D :

Genre : VP
Carrosserie : CI
Energie : ES
Puissance : 5 CV
Couleur : Blanc Standard
Nombre places : 5
Poids à vide :
PTAC :
Etat général : Non précise

MANDANT :
GENERALI ASSURANCES
2 RUE PILLET WILL
PARIS

REPARATEUR : [redacted]

CIRCONSTANCE DE L'EXPERTISE
-Vu A distance le 21/03/2022 personnes présentes
[redacted]

PIECES COMMUNIQUEES :
VEHICULE ECONOMIQUEMENT REPARABLE : Oui
VEHICULE TECHNIQUEMENT REPARABLE : Oui
IMMOBILISATION THEORIQUE : 2.0 jour(s)
DOMMAGES IMPUTABLES : Arrière, Légère, 180°
ACCORD DE REGLEMENT DIRECT DEFINITIF : Oui

Véhicule réparable - EAD - expertise à distance

Montant de l'expertise	891.48 TTC	(742.90HT)
Répartition des chocs : Choc Arrière	891.48 TTC	(742.90HT)
Montant facturé par le réparateur (facture n° 30221)	698.33	
Montant à régler au réparateur	698.33	
(Détail des vétustés retenues présenté en annexe)		

TVA récupérable : Non

FIGURE 2.1 – Exemple de présentation de rapport anonymisé (page 1 sur 3)

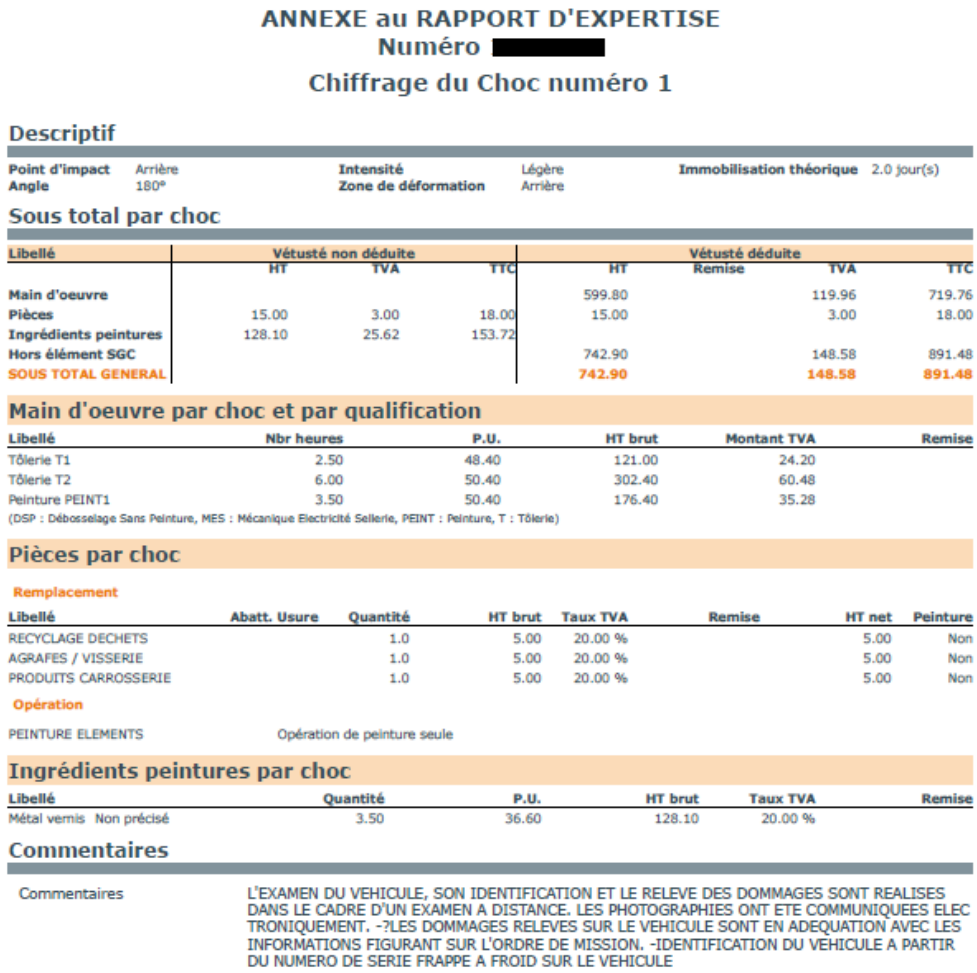
ANNEXE au RAPPORT D'EXPERTISE
Numéro [redacted]

Total Général

Libellé	Vétusté non déduite			Vétusté déduite			
	HT	TVA	TTC	HT	Remise	TVA	TTC
Main d'oeuvre				599.80		119.96	719.76
Pièces	15.00	3.00	18.00	15.00		3.00	18.00
Ingrédients peintures	128.10	25.62	153.72				
Hors élément SGC				742.90		148.58	891.48
Minoration				44.57		8.91	53.48
Total général				698.33	44.57	139.67	838.00

FIGURE 2.2 – Exemple de présentation de rapport anonymisé (page 2 sur 3)

Generali a alors à sa disposition un historique de tous les **rapports d'expertise qui ont été réalisés entre 2011 et aujourd'hui sur ses sinistres expertisés**, qui lui sont fournis par une société spécialisée en échange de données informatisées pour les métiers de l'assurance nommée **DARVA**.



ANNEXE au RAPPORT D'EXPERTISE
Numéro [redacted]

Chiffrage du Choc numéro 1

Descriptif

Point d'impact	Arrière	**Intensité**	Légère	**Immobilisation théorique**	2.0 jour(s)
Angle	180°	**Zone de déformation**	Arrière		

Sous total par choc

Libellé	Vétusté non déduite			Vétusté déduite			
	HT	TVA	TTC	HT	Remise	TVA	TTC
Main d'oeuvre				599.80		119.96	719.76
Pièces	15.00	3.00	18.00	15.00		3.00	18.00
Ingrédients peintures	128.10	25.62	153.72				
Hors élément SGC				742.90		148.58	891.48
SOUS TOTAL GENERAL				**742.90**		**148.58**	**891.48**

Main d'oeuvre par choc et par qualification

Libellé	Nbr heures	P.U.	HT brut	Montant TVA	Remise
Tôlerie T1	2.50	48.40	121.00	24.20	
Tôlerie T2	6.00	50.40	302.40	60.48	
Peinture PEINT1	3.50	50.40	176.40	35.28	

(DSP : Débosselage Sans Peinture, MES : Mécanique Electricité Sellerie, PEINT : Peinture, T : Tôlerie)

Pièces par choc

Remplacement

Libellé	Abatt. Usure	Quantité	HT brut	Taux TVA	Remise	HT net	Peinture
RECYCLAGE DECHETS		1.0	5.00	20.00 %		5.00	Non
AGRAFES / VISSERIE		1.0	5.00	20.00 %		5.00	Non
PRODUITS CARROSSERIE		1.0	5.00	20.00 %		5.00	Non

Opération

PEINTURE ELEMENTS	Opération de peinture seule

Ingrédients peintures par choc

Libellé	Quantité	P.U.	HT brut	Taux TVA	Remise
Métal vernis Non précisé	3.50	36.60	128.10	20.00 %	

Commentaires

Commentaires	L'EXAMEN DU VEHICULE, SON IDENTIFICATION ET LE RELEVE DES DOMMAGES SONT REALISES DANS LE CADRE D'UN EXAMEN A DISTANCE. LES PHOTOGRAPHIES ONT ETE COMMUNIQUEES ELEC TRONIQUEMENT. -?LES DOMMAGES RELEVES SUR LE VEHICULE SONT EN ADEQUATION AVEC LES INFORMATIONS FIGURANT SUR L'ORDRE DE MISSION. -IDENTIFICATION DU VEHICULE A PARTIR DU NUMERO DE SERIE FRAPPE A FROID SUR LE VEHICULE

FIGURE 2.3 – Exemple de présentation de rapport anonymisé (page 3 sur 3)

Tous les mois, Generali reçoit les informations de ces rapports dans un fichier au format *.json* et au format *.csv*, avec les différentes mises à jour de ces rapports. Ici, nous nous intéressons aux flux de données ***.json*** qui sont des formats faciles à exploiter, puisqu'ils sont sous forme de dictionnaires avec plusieurs clés et sous-clés qui permettent de retrouver les informations souhaitées facilement, comme si on suivait un arbre. Ce formatage de la donnée est appelé données non structurées.

Dans le format *.json* utilisé par DARVA, les clés correspondent à des noms de variables, et permettent de retrouver les informations visibles dans le rapport d'expertise au format *.pdf*. Nous disposons également d'un dictionnaire de ces noms de variables fourni par DARVA.

2.1.2 Sélection de variables pertinentes

Pour éviter le surstockage et les temps de calculs trop importants, seulement certaines variables des rapports d'expertise ont été extraites pouvant être distinguées en trois catégories :

1. les variables utiles à la définition du périmètre :
   - véhicule techniquement et économiquement réparable (Oui/Non)
   - type de véhicule pour sélectionner les véhicules particuliers (VP) et les camionnettes (qui ont un poids total autorisé en charge (PTAC) ou une masse en charge techniquement admissible inférieur à 3,5 tonnes).
   - code postal du lieu de l'expertise pour sélectionner uniquement la France métropolitaine
2. les variables qui peuvent servir de clés de jointure avec d'autres bases comme les références de sinistres, numéros de contrats, numéros de séries...
3. et les variables qui donnent le détail des pièces : le libellé, le montant, la quantité, la qualification de l'opération et le pourcentage d'abattement pour vétusté de chaque pièce.

2.1.3 Traitement des données pour définir la base de travail

Après avoir défini le périmètre d'étude sur les rapports d'expertise auto, un traitement des données a dû avoir lieu pour permettre une meilleure manipulation de celles-ci.

En effet, l'extraction des données que Generali a à sa disposition donne les informations des pièces dans des listes, par exemple "[PORTE, VIS, FORFAIT ERD]" comme nous pouvons le voir en *figure 2.4*, il a fallu alors **transposer la table pour passer d'une table avec une ligne par rapport à une table avec une ligne par pièce**.

Référence du sinistre	Libellés des pièces	Coûts des pièces
1	[PORTE, VIS, FORFAIT ERD]	[50, 1, 2]
2	[PHARE]	[40]
3	[PARE-CHOC]	[65]

Référence du sinistre	Libellés des pièces	Coûts des pièces
1	PORTE	50
1	VIS	1
1	FORFAIT ERD	2
2	PHARE	40
3	PARE-CHOC	65

FIGURE 2.4 – Exemple de pivot réalisé sur la base de données

Après cela, des traitements pratiques ont été réalisés pour **uniformiser le format des données**. Par exemple, la variable "libellé de la pièce" est une information renseignée manuellement par les experts, un travail de nettoyage du texte a été mis en place pour retirer les accents, les caractères de ponctuation etc.

Avec cette nouvelle table obtenue, nous avons constaté qu'il y avait des **éléments listés qui ne correspondaient pas à des pièces en tant que tel**, par exemple la pièce "FORFAIT ERD" qui désignait un forfait "évaluation et recyclage des déchets" ou encore "MESURE COVID" qui coïncidait au temps de nettoyage ou au matériel

supplémentaire utilisé par les réparateurs pour désinfecter les véhicules en période de pandémie. Ces forfaits ne sont pas intéressants pour l'étude donc nous les avons retirés.

Il y a également les pièces sur lesquelles uniquement une opération de main d'oeuvre ou de peinture a été réalisée. On peut repérer ces caractéristiques grâce à une variable qui donne la qualification de l'opération de la pièce. Pour être sûr de ne pas avoir ce type d'opération, les pièces de montants nuls ont aussi été retirées.

Ensuite, pour s'assurer d'avoir uniquement des **pièces qui ont une influence notable** dans le coût des réparations, nous choisissons un **seuil de 15 €** pour garder uniquement les pièces de montant "prix de base" supérieur à ce seuil. Ce filtre supplémentaire permet de retirer les pièces qui pourraient correspondre à d'autres forfaits ou autres éléments qui ne seraient pas des pièces, tout en s'assurant d'avoir des pièces de montants significatifs.

Enfin, en analysant plus en détail les données, un libellé particulier a attiré l'attention puisqu'il s'appelle **"Ensemble de pièces"**. Ce libellé, qui a pour particularité d'avoir été **utilisé uniquement avant 2015**, est un libellé général dans lequel plusieurs des pièces utilisées (ou les moins importantes) y étaient intégrées. Ainsi, pour éviter de perturber nos analyses (voir *figure 2.5*), ce libellé a été retiré.

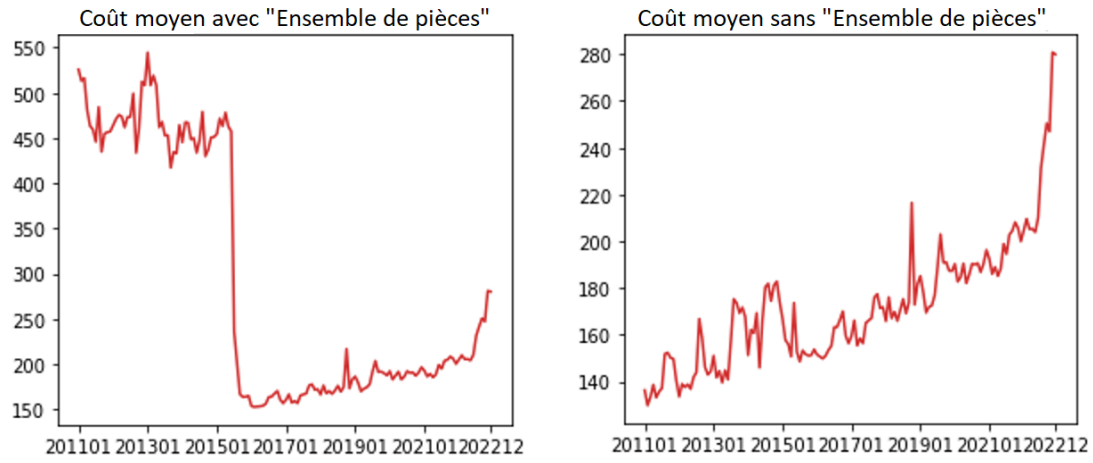


FIGURE 2.5 – Coût moyen des pièces par mois avec et sans le libellé "Ensemble de pièces" dans la table de données

2.1.4 Vérification de la qualité des données

Pour vérifier la cohérence des données de la base utilisée, nous avons utilisé d'autres sources de données pour comparer les évolutions de prix. L'objectif de cette analyse était de vérifier que les rapports d'expertise fournissaient bien le prix constructeur que nous voulions suivre et que celui-ci était cohérent avec le marché.

Ces 2 sources de données supplémentaires seront présentées avant d'expliquer la procédure de vérification.

Données du service Indemnisation : la "matrice auto"

Pour commencer, le service "Indemnisation" en collaboration avec les experts de Generali réalise tous les trimestres une **étude** concentrée sur **quelques pièces emblématiques de 12 modèles de véhicules** pour identifier les évolutions des prix de ces modèles. Chaque trimestre, les experts regardent sur un outil à leur disposition le prix constructeur de leur liste de pièces définie sur ces mêmes modèles. Par exemple pour la Renault Clio IV, il y a 8 pièces suivies concernant le choc avant gauche, étant considéré comme la partie du véhicule la plus sinistrée.

Grâce à cela, Generali a une vision rapide de l'évolution du marché avec ces modèles représentatifs de son portefeuille et du marché comme la Peugeot 206, la Citroën C3 Picasso ...

Clio 4									
Pièces	price excluding VAT								
	Aout 2021	Fevrier 2022	Var %	mai-22	Var %	août-22	Var %	nov-22	Var %
Bouclier Avant	321,27	327,69	2,0%	332,72	2%	342,69	3,0%	352,97	3,0%
Appui gauche bouclier avant	42,93	47,31	10,2%	58,05	23%	58,88	1,4%	59,99	1,9%
Absorbeur gauche bouclier avant	62,44	60,58	-3,0%	61,25	1%	61,68	0,7%	61,99	0,5%
Phare avant gauche	220,16	222,36	1,0%	276,63	24%	281,06	1,6%	287,06	2,1%
Enjoliveur inférieur phare			0,0%		0%		0,0%		0,0%
Jeu Enjoliveur inférieur phare	82,49	84,84	2,8%	123,65	46%	125,55	1,5%	128,05	2,0%
Aile avant gauche	163,06	166,32	2,0%	167,95	1%	169,13	0,7%	169,97	0,5%
Garnit avant P roue avant gauche	67,29	68,64	2,0%	69,31	1%	69,80	0,7%	70,14	0,5%
Total	**959,64**	**977,74**	**1,9%**	**1 089,56**	**11%**	**1 108,79**	**1,8%**	**1 130,17**	**1,9%**

FIGURE 2.6 – Exemple de prix constructeur récupérés dans la matrice automobile sur la Renault Clio IV

Données EUCON

Generali a fourni à certains collaborateurs du groupe une application proposée par EUCON, une entreprise spécialisée dans l'automobile, l'assurance et l'immobilier. L'interface que l'entreprise propose, informe les **prix constructeurs des pièces détachées automobiles d'une large variété de modèles**, dans 80 pays et sur les 5 dernières années.

Pour accéder au prix d'une pièce, il faut renseigner un certain nombre d'informations : la marque, le modèle, la version, la finition, la puissance, le nombre de portes et le moteur (voir *figure 2.7*).

A la suite de cela, on accède à des choix d'ensembles de pièces qui sont représentés de façon divisées pour accéder à la pièce que l'on veut consulter (voir *figure 2.8*).

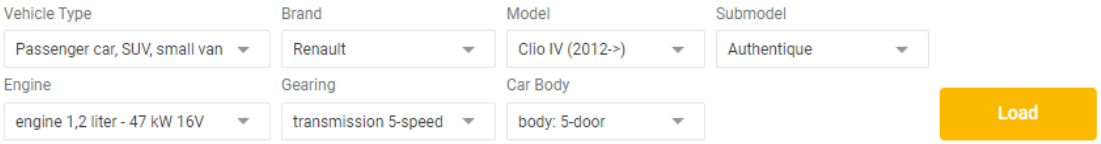


FIGURE 2.7 – Exemple de sélection d'une Renault Clio IV sur EUCON

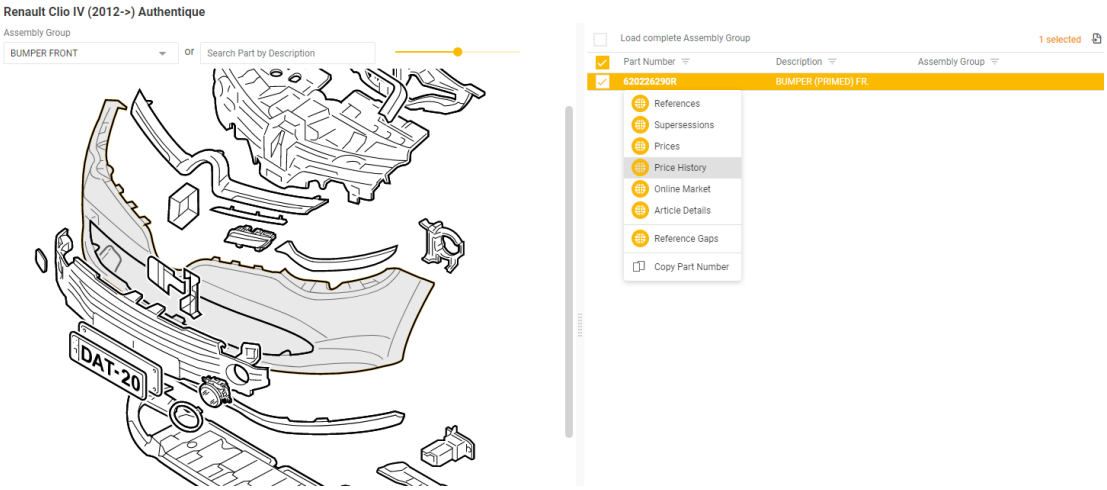


FIGURE 2.8 – Exemple de sélection d'un bouclier avant d'une Renault Clio IV sur EUCON

En sélectionnant une pièce, on obtient sa référence officielle qui permet de regarder son prix à la date de la dernière récupération dans plusieurs pays (voir *figure 2.9*), mais aussi de regarder son évolution temporelle en France sur les 5 dernières années (voir *figure 2.10*). Ce sont des informations importantes pour pouvoir se comparer aux différents marchés.

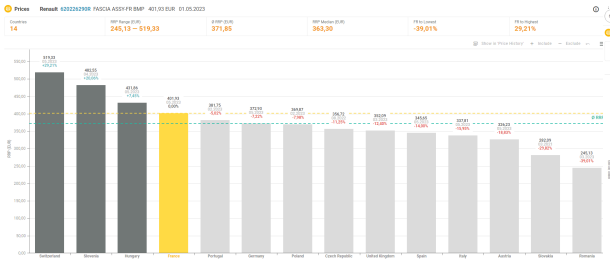


FIGURE 2.9 – Comparaison des prix du bouclier avant de la Clio IV dans différents pays

FIGURE 2.10 – Evolution du prix du bouclier avant de la Clio IV en France

Grâce à cet outil, des prix de nombreuses pièces ont été obtenus, rendant un compa-

ratif possible avec ceux que la "matrice auto" nous fournissait.

Comparaison de nos données avec celles des sources présentées

Grâce à ces éléments, une comparaison des prix des rapports d'expertise avec ceux de ces deux sources de données a été réalisée.

Par son fonctionnement, pour avoir des prix EUCON, on doit renseigner la finition du modèle que l'on veut regarder en plus des autres informations, or cette donnée a un impact, qui peut être plus ou moins important, sur le prix de la pièce. Par exemple, pour le cas du phare avant gauche d'une Peugeot 208, le prix de cette pièce pour une même période variait de 338,68€ à 1150,98€ en fonction de la version sur EUCON.

Dans DARVA on n'a pas l'information de la finition pour pouvoir l'indiquer dans l'outil EUCON, par conséquent on ne peut pas retrouver directement la pièce d'un sinistre Generali dans EUCON, en théorie. Pour pallier cela, des sondages ont été réalisés dans EUCON en considérant la marque, le modèle et la version du véhicule d'un rapport d'expertise, pour tester les différentes finitions dans EUCON. Cette analyse a permis de **retrouver les mêmes prix de pièces** entre DARVA et EUCON sur notre base de véhicules sondés.

Cette analyse a permis de conclure que nous avions bien les **prix constructeurs** des pièces utilisées lors des réparations.

2.1.5 Définition du périmètre des pièces suivies

A ce stade de l'étude, nous avons vérifié la qualité de nos données : d'une part, une analyse exploratoire nous a permis d'écarter certaines informations non pertinentes ; et d'autre part nous avons pu valider que l'information du prix des pièces dans les rapports d'expertise Generali correspondait bien au prix établis par les constructeurs automobiles.

Dans cette partie, nous allons préciser le périmètre d'étude des pièces à suivre dans le temps en s'assurant que les informations de ces pièces sont bien stables et cohérentes.

En effet, dans un premier temps, pour identifier les pièces à suivre, une méthode de recherche par mots-clés dans les libellés a été testée. Grâce à cela, nous pouvions obtenir des premiers résultats mais avec des prix qui se sont retrouvés très différents puisque par exemple en voulant avoir des "boucliers avant", d'autres pièces étaient rapportées comme "support bouclier avant" ou "appui bouclier avant"...

Pour éviter ce problème, le choix de **suivre les libellés précis** semblait plus pertinent.

Pour savoir quels libellés envisager, nous avons recherché les pièces qui étaient suivies dans la matrice auto (voir 2.1.4) et regardé les libellés qui étaient les plus présents dans

les rapports d'expertise et qui représentaient en moyenne un coût important dans la facture.

Parmi ces pièces, **certaines n'ont pas été retenues à cause d'une instabilité dans le temps**, que cela soit en coût moyen ou en quantité. En effet le libellé nommé "ensemble de pièces" qui était présent dans les rapports d'expertise avant 2015 a eu plusieurs impacts sur les données. Dans un premier temps ce libellé biaise l'analyse du coût moyen des pièces en général comme vu précédemment (voir *figure 2.5*). Il a aussi impacté le volume de certains libellés de pièces. C'est le cas par exemple de la pièce "projecteur g g" (voir *figure 2.11*) qui est une pièce représentative dans les rapports en montant, qui n'était pas présente dans les rapports d'expertise avant 2015 et qui a subi une forte évolution de volume après cette date. C'est également le cas de "pare chocs av" qui avait quelques pièces avant 2015 puis a subi une explosion en 2015.

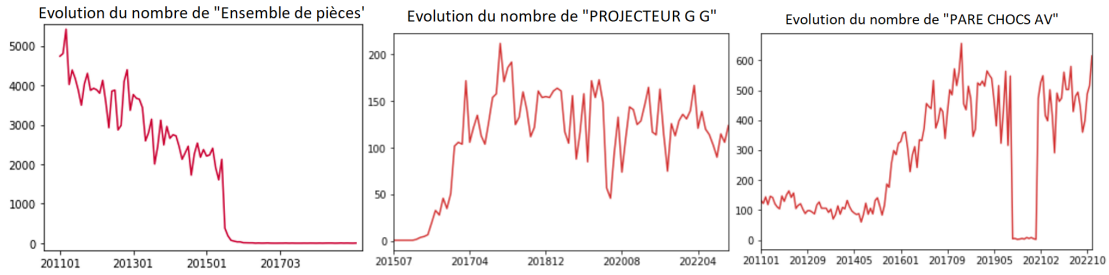


FIGURE 2.11 – Evolutions mensuelles du nombre de pièces "ensemble de pièces", "projecteur g g" et "pare-chocs av"

Après cela, nous avons également mis de côté le **"phare avant gauche"** puisqu'après une étude approfondie, on a constaté une **forte variabilité des prix** en fonction des **marques, modèles, versions et surtout finitions**. Ainsi, pour le cas particulier de la Peugeot 208 cité au-dessus, en regardant le prix de cette pièce à une même date mais avec des versions et finitions différentes, la pièce avait une variabilité de presque 1000€ (prix constaté sur EUCON, voir *2.1.4*).

Enfin, nous avons vérifié que les évolutions de prix en fonction des portes droites et gauches étaient similaires pour envisager de les regarder ensemble et ainsi gagner en volumétrie. Cette étude est présentée dans le point suivant.

A la fin de cette étude pour définir le périmètre nous arrivons à une liste de cinq pièces :

— le ***bouclier avant***,
— le ***bouclier arrière***,
— l'***aile avant gauche***,
— les ***portes avant***
— et les ***portes arrière***.

2.1.6 Etude et traitement des séries temporelles

Une fois que nous avions nos pièces à suivre, il devint alors possible de regarder les évolutions temporelles de leurs coûts moyens et de les traiter.

Traitement des valeurs aberrantes

Nous avons commencé par retirer les valeurs aberrantes en utilisant la formule empirique [1] utilisée notamment dans les "boîtes à moustache" ou "boxplots" à partir des centiles. Ainsi, une valeur est considérée comme aberrante ou non aberrante comme présenté dans la *définition 1.*

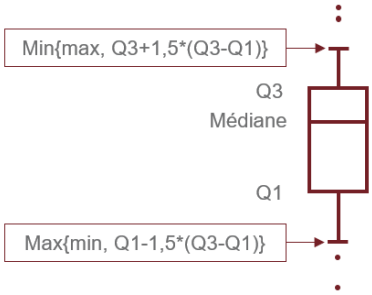


FIGURE 2.12 – Représentation d'un "boxplot"

Definition 1 (Valeur aberrante par la définition du *boxplot*) Une valeur x_i est considérée comme aberrante si :

$$x_i > min(max, Q_3 + 1.5 \times (Q_3 - Q_1)) \text{ ou } x_i < max(min, Q_1 - 1.5 \times (Q_3 - Q_1)) \quad (2.1)$$

avec Q_1 et Q_3 les centiles à 25% et 75% respectivement.

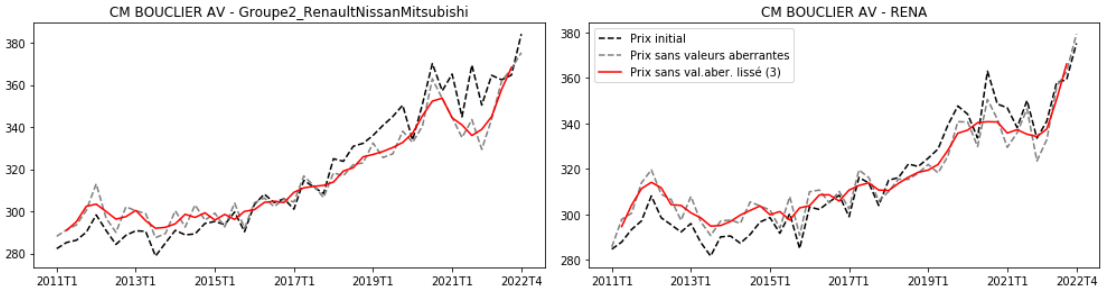


FIGURE 2.13 – Exemple de courbes avant traitement, sans valeurs aberrantes puis lissée (sur 3 trimestres) du coût du bouclier avant pour la marque Renault et pour le groupe Renault Nissan Mitsubishi

Avec ce premier traitement, nous passons de la courbe noire à la courbe grise en pointillés sur la *figure 2.13.*

1. Source : https://www.stat4decision.com/fr/le-box-plot-ou-la-fameuse-boite-a-moustache/

Lissage par moyennes mobiles

Une moyenne mobile, autrement appelée moyenne glissante, est une moyenne statistique qui permet de réduire le bruit des séries temporelles. Mathématiquement, elle se décrit comme suit :

Definition 2 (Moyennes mobiles) Soit $\{y_t\}_{t\in\{1,...,T\}}$ une série temporelle avec T éléments et $l \in \mathbb{N}$,

$$\hat{y}_t^L = \frac{1}{2L+1} \sum_{i=t-l}^{t+l} y_i, \text{ pour } t \in \{1+l, ..., T-l\}, \tag{2.2}$$

avec $L = 2l + 1$ la largeur de la fenêtre.

Plus la largeur de la fenêtre est grande et la courbe lissée, moins on aura de points d'observations. Ainsi, $\hat{y}_t^L$ n'est pas définie pour $t \leq l$ et $t > T - l$.

Avec ce deuxième traitement, nous arrivons à la courbe rouge de la *figure 2.13*. Nous constatons qu'il nous manque une valeur au début de la courbe et une à la fin à cause du lissage sur trois trimestres (un avant et un après). Grâce à cela, on constate moins de bruit et les sauts pouvant être liés à des politiques tarifaires sont plus facilement visibles.

Sauf mention contraire, la suite de l'étude utilise les données qui ont suivi ce même traitement.

2.1.7 Définition du maillage de l'étude

Une autre question soulevée est de savoir si l'étude peut se faire au niveau :

1. du groupe, comme Stellantis par exemple, en prenant les moyennes des données des pièces des marques qui composent ce groupe ;
2. de la marque, comme Renault, avec tous les modèles de la marque confondus ;
3. ou du modèle, par exemple la Clio de chez Renault.

Pour y répondre, les évolutions des "boucliers avant" ont été comparés au niveau de ces différents niveaux de finesse, en commençant du plus précis au plus général.

— Dans un premier temps, les évolutions d'une marque avec plusieurs de ses modèles ont été comparées.

  Dans l'exemple *figure 2.14*, les évolutions des pièces de la Golf, de la Polo et la Tiguan de Volkswagen ont été comparées avec les pièces de toutes les Volkswagen. Les courbes du prix des pièces des modèles suivaient relativement les mêmes évolutions que celles de la marque correspondante, il était plus difficile de faire la même conclusion sur les données qui n'étaient pas suffisamment robustes ou incomplètes sur certains trimestres comme la Tiguan ici représentée en bleu. Cela montre également l'intérêt de suivre les évolutions par marques plutôt que par modèles.

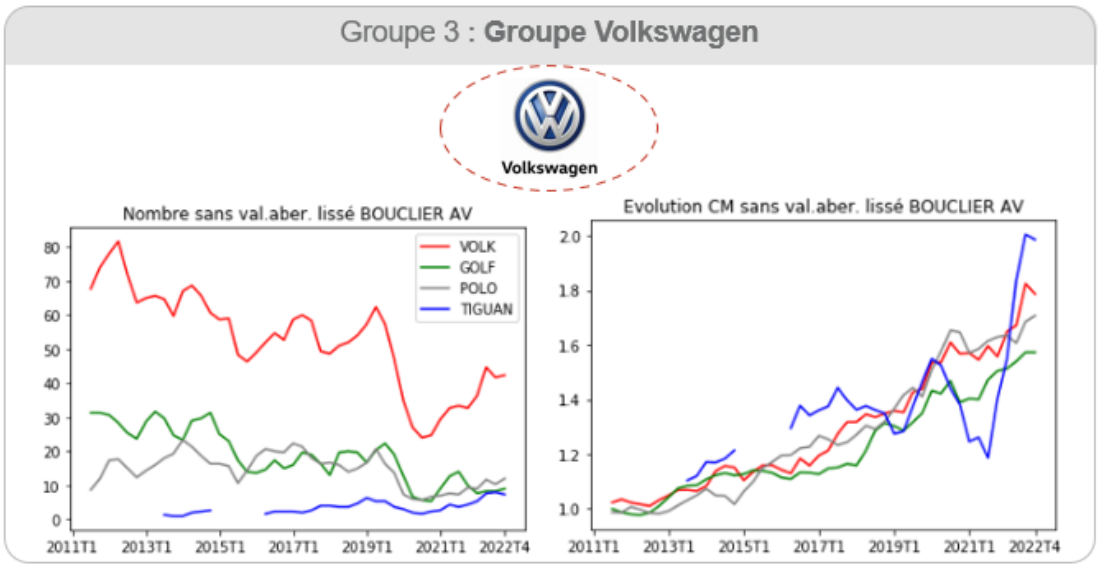


FIGURE 2.14 – Evolution du nombre de Bouclier avant et du coût moyen (base T1 2011) de Volkswagen et 3 de ses modèles Golf, Polo et Tiguan.

— Ensuite, en comparant l'évolution d'un regroupement avec plusieurs de ses marques.

Dans le groupe Volkswagen, les évolutions des nombres et des coûts moyens des boucliers avant du groupe en lui-même, ainsi que celles des marques Volkswagen, Audi et Seat, ont été représentées dans la *figure 2.15*. Dans ce cas, les tendances semblaient éloignées comme on peut le voir entre Audi et Volkswagen notamment. Dans d'autres cas, les tendances des courbes entre différentes marques d'un même groupe semblaient assez proches. C'était le cas de Peugeot, Opel et Citroën dans le sous-groupe PSA de Stellantis.

Ainsi, pour généraliser cette analyse graphique, des regroupements automatisés seront proposés dans la *partie 3.2.1* puisque le suivi par groupe n'est pas toujours favorisé, comme on l'a vu avec le groupe Volkswagen, ce qui fait que celui par marque est mis en avant.

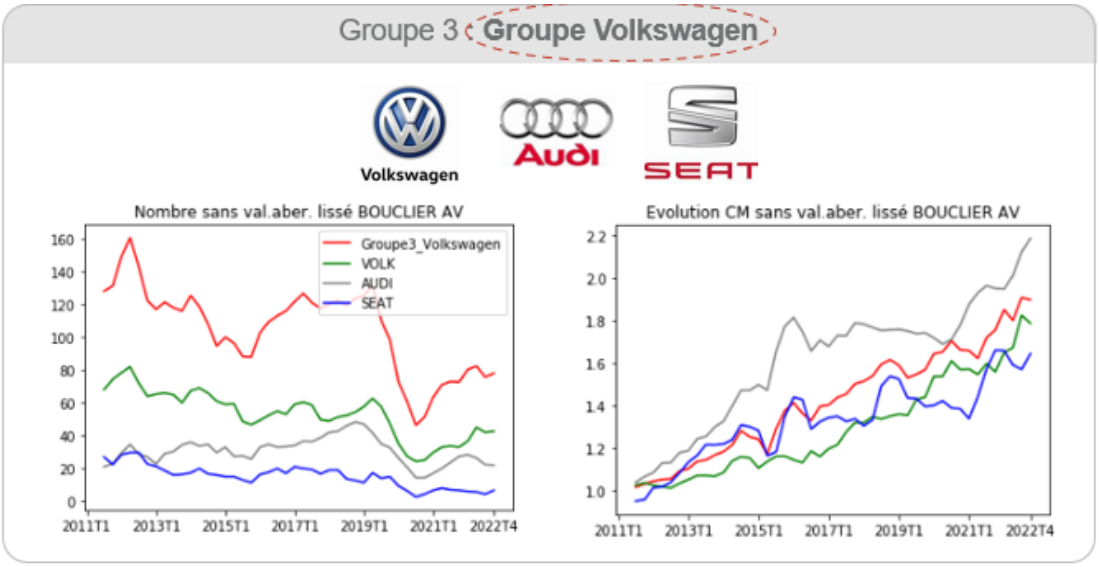


FIGURE 2.15 – Evolution du nombre de Bouclier avant et du coût moyen (base T1 2011) du groupe Volkswagen et 3 de ces marques, Volkswagen, Audi et Seat.

2.2 Les données externes

Après cette partie sur les informations fournies par les rapports d'expertise, on peut s'attendre à ce que les informations du marché influencent le prix des pièces automobiles. En effet, comme les prix sont fixés en fonction de l'offre des fournisseurs et de la demande des consommateurs, les prix du marché peuvent ainsi modifier l'équilibre du marché et impacter les prix. On peut notamment s'attendre à ce que les prix de l'essence, du pétrole ou encore des matériaux des pièces impactent le prix de la pièce directement.

Ainsi, pour enrichir les informations internes à l'entreprise, des données externes ont été recherchées et ajoutées.

Celles-ci devaient répondre à certains critères pour être intégrées à l'étude. Elles devaient :

— couvrir la même plage temporelle que celle des données Generali, c'est-à-dire de 2011 à aujourd'hui,
— être mise à jour régulièrement pour pouvoir assurer une continuité dans le suivi,
— être facilement récupérable pour permettre une mise à jour rapide à chaque lancement de l'outil

Pour répondre à ces différents points, une sélection a été faite pour que toutes les données retenues soient exploitables.

L'inflation

Commençons par l'indicateur d'inflation. Nous rappelons que l'inflation se définit comme étant une **augmentation générale et durable des prix** qui a pour conséquence de **diminuer le pouvoir d'achat** de la population.

Dans ce mémoire, on s'intéresse à l'évolution du coût des pièces automobiles qui semble intuitivement fortement corrélée à l'indicateur d'inflation du marché français en général, avec des évolutions qui peuvent cependant varier. Il parait alors intéressant de récupérer ces informations pour l'étude, c'est ce qui est fait avec les données annuelles dans un premier temps (données disponibles depuis 1982), puis les données mensuelles grâce à l'Indice de Prix à la Consommation (IPC) calculés par l'INSEE [22] (indice avec base 2015 depuis 1991), voir *figure 2.16*.

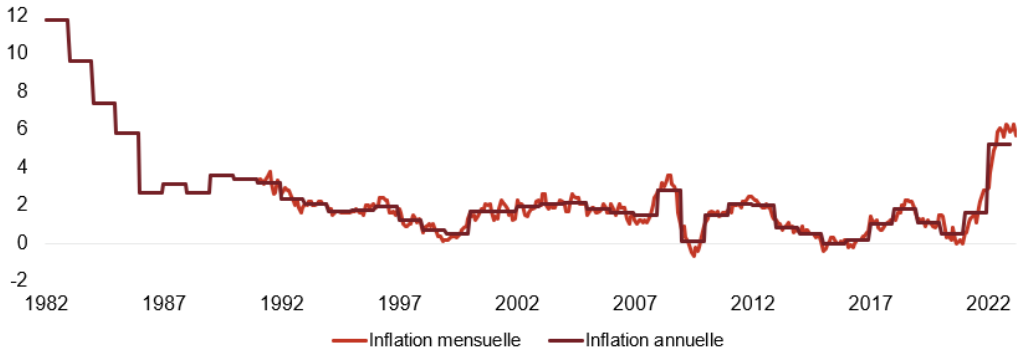


FIGURE 2.16 – Evolutions mensuelle et annuelle (base 2015) de l'indice des prix à la consommation par l'INSEE

Sur cette courbe on voit l'impact de différentes périodes qui ont eu lieu sur le marché [2]. Tout d'abord, au cours des années 1980, l'inflation a connu une forte baisse à la suite des Trente Glorieuses. Dans un second temps on voit une période assez stable, entre 0% et 2% d'inflation, jusqu'à la crise financière des *subprimes* en 2007 qui a impacté le système financier mondial ; avant de voir l'impact de la pandémie du covid et du confinement français qui a fortement baissé l'inflation avant de rattraper le retard à la réouverture du marché.

En plus de ces chocs qui ont entraîné une inflation presque nulle en 2008 et en 2020, on peut voir l'impact de la mise en vigueur de l'euro au début des années 2000, avec la monnaie scripturale en 1999 et la monnaie fiduciaire en 2002 qui a ralenti l'inflation sur ces premières années.

2. Historique de l'inflation en France avec ces crises sur france-inflation.com/inflation-depuis-1901.php et sur leconomiste.eu/decryptage-economie/186-histoire-de-l-evolution-de-l-inflation-en-zone-euro.html

Le cours de l'euro (taux de change avec le dollar américain)

D'après le mémoire [29] réalisé par l'INSEE, deux secteurs importants impactent l'inflation : le prix du pétrole (voir la *sous-partie suivante*) et le cours de l'euro par rapport aux monnaies étrangères.

On peut expliquer cela notamment par la forte connexion entre le marché français et les marchés extérieurs créant un déficit de 164 milliards d'euros en 2022 [3]entre les exportations et les importations. La **France importe beaucoup d'énergie** comme le gaz ou le pétrole à cause d'une faible production interne. Ainsi, elle est **soumise à la baisse de la valeur de sa monnaie** par rapport aux autres monnaies, augmentant ainsi la valeur des biens achetés.

Cette donnée externe a été récupérée sur le site de l'INSEE [16], *figure 2.17* avec le cours de l'euro par rapport au dollar américain de 1990 à aujourd'hui.

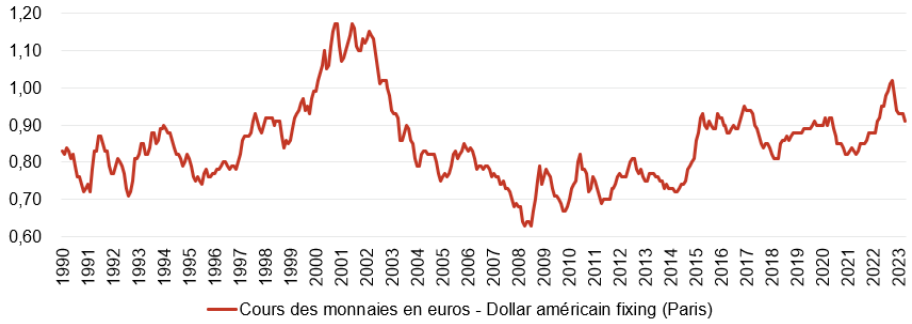


FIGURE 2.17 – Evolutions mensuelles du cours des monnaies en euros - dollars américains par l'INSEE

L'indice du prix des matériaux

De plus, le coût des matières premières utilisées dans la fabrication des pièces est apparu comme un potentiel facteur dans la détermination du prix des ventes des pièces par les constructeurs.

D'après un article de CELO Crew Technology [4], les **métaux** les plus utilisés pour les constructions de voitures sont l'**acier**, l'**aluminium**, le **magnésium** et le **cuivre** ([20], [21], [18], [19]). Ces indices commencent en 1994 pour l'aluminium et le cuivre, en 2000 pour les éléments en métaux et en 2005 pour l'acier, jusqu'à aujourd'hui.

3. Source : https://www.cerl.fr/chiffres-cles-commerce-exterieur-de-la-france-import-export/

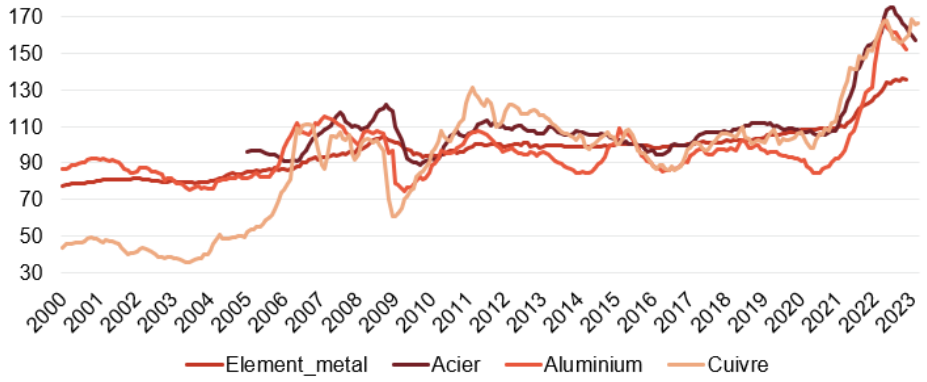


FIGURE 2.18 – Indices de prix de production d'industrie des métaux français pour le marché français obtenus par l'INSEE

On peut également citer le **pétrole** qui peut être utilisé pour les **composants plastiques** ([15]). Seuls les indicateurs qui respectent la possibilité de continuité ont été récupérés, également trouvés sur le site de l'INSEE avec les informations depuis 1990.

FIGURE 2.19 – Cours du pétrole brut Brent (Londres) importé en dollars américains par baril par l'INSEE

Le prix de l'essence

Dans un autre temps, en plus du cours du Brent, le **prix de l'essence** qui impacte le prix du **transport** a été cherché [24], avec entres autres les prix de différentes variantes d'essence, le **sans-plomb 95** depuis 2002, le **sans-plomb 95 E10** depuis 2019, le **sans-plomb 98** et le **gazole** depuis 1992. Ces données macro-économiques sont toutes corrélées les unes aux autres, on a décidé de garder le sans-plomb 98 dans la suite du mémoire, pour sa complétude même si une majeure partie des courbes couvraient l'historique des données internes.

4. Source : https://www.celofasteners.com/fr/ 2082-vis-pour-le-metal-dans-le-secteur-automobile

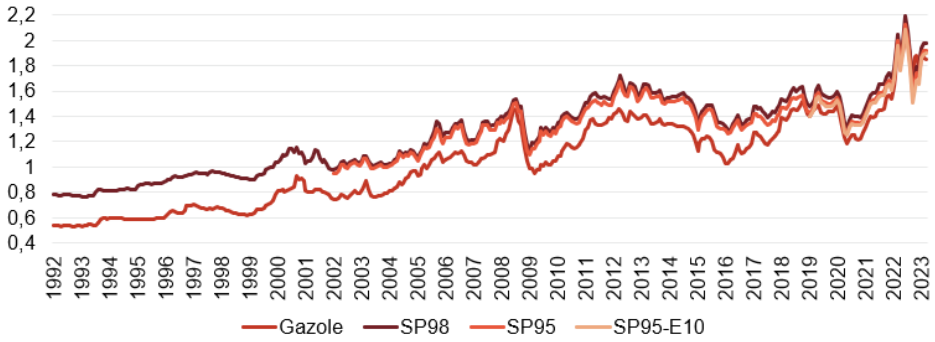


FIGURE 2.20 – Prix moyens mensuels de vente au détail en métropole des matières premières par l'INSEE

Le prix du verre

Une autre composante qui influence les prix des pièces détachées est le prix du verre, et plus particulièrement le **verre** utilisé pour la **fabrication de pare-brise**.

Plusieurs verres existent, le verre plat, creux, trempé, feuilleté, antireflet ou hydrophobe. Les pare-brises sont majoritairement fabriqués en **verre feuilleté**, composé d'une superposition de couches de verre-plastique-verre (ou plus) qui limite les éclats en cas de choc.

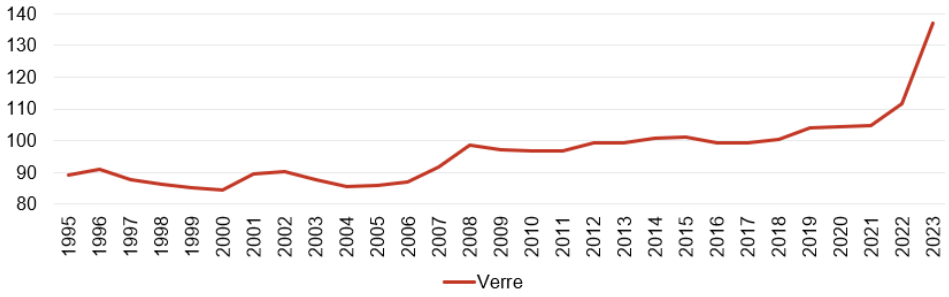


FIGURE 2.21 – Prix du verre et d'articles en verre du CPF 23.1 (base 2015) par l'INSEE

Le prix du verre feuilleté n'étant pas détaillé dans les données en source libre, l'évolution du prix du verre depuis 1995 et d'articles en verre correspondants au groupe CPF 23.1 a été récupéré auprès de l'INSEE [17]. Ce groupe inclut le verre sous toutes ses formes, quel que soit le procédé de fabrication et les articles en verre.

Le coût de la main d'oeuvre

Après avoir analysé les possibles matériaux pour la production de véhicules, un autre poste qui représente un coût significatif dans la production des pièces est celui de la main d'oeuvre. Il est alors important de suivre son évolution. Un indice de **coût du travail en France** pour l'**industrie manufacturière** (NAF rév, 2 section C) sur les salaires et charges a été trouvé sur l'INSEE, [23], de 1995 à mi-2023.

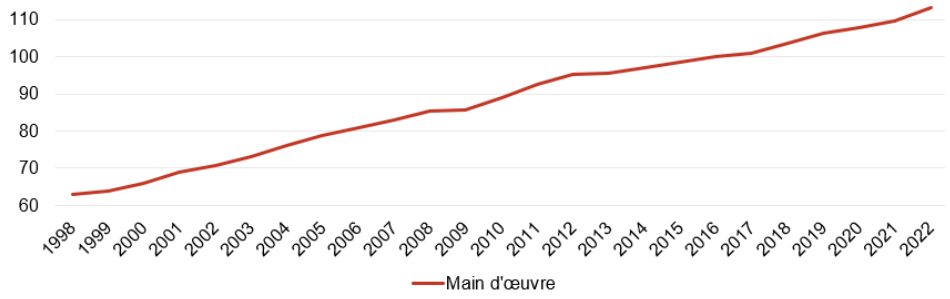


FIGURE 2.22 – Indice du coût de la main d'oeuvre de l'industrie manufacturière (base 2016) par l'INSEE

Cet indice, contrairement aux autres données externes trouvées, est fourni de manière trimestrielle et non mensuelle.

Par ailleurs, d'autres éléments de marché avaient été envisagés pour l'analyse mais la contrainte d'historique ou de continuité des données ont empêché certaines données d'être récupérées, notamment le prix de l'électricité en France.

Chapitre 3

Détection automatisée et continue de changements de politiques tarifaires des constructeurs automobiles sur les pièces détachées

Dans ce chapitre, l'idée est d'exploiter les rapports d'expertise automobiles des assurés de Generali France pour détecter les changements de pratiques tarifaires chez les constructeurs de pièces automobiles. Ainsi, en détectant le plus rapidement possible les fortes évolutions des prix des pièces automobiles qui pourraient faire pressentir une nouvelle politique tarifaire de la marque, il sera possible d'impacter cette différence sur les tarifs automobiles avant de voir la différence apparaître dans les sinistres.

Actuellement, une hypothèse d'inflation est réalisée chaque année lors de la tarification pour l'année suivante se basant sur les informations arrêtées à l'année d'avant. Si cette hypothèse prend en compte la tendance de l'évolution du prix de la pièce et qu'on détecte que celle-ci a été modifiée, alors on peut adapter son prix avec une projection de tendance plus adaptée.

FIGURE 3.1 – Schéma des années utilisées en tarification

Pour cela, une première idée était d'utiliser un modèle de dérive des données pour détecter ces changements. Certaines dérives de données concernent des changements dans les données en production par rapport aux données qui ont été utilisées en entraînement des modèles, c'est ce qu'on appelle le *datadrift*. D'autres dérives sont relatives au concept

du modèle qui peut venir d'un changement de comportements des agents économiques ou d'un changement d'objectif par rapport à celui auquel répond le modèle. Dans notre cas, comme nous n'avions pas de modèle à suivre dans le temps mais bien des données, c'était la dérive de données qui était liée à notre problème. Cependant, ces méthodes nécessitent d'avoir un large volume de données pour pouvoir comparer les distributions des variables sur différentes périodes. C'est pour cette raison que cette approche n'a pas été retenue, les recherches associées sont néanmoins disponible en *annexe 5.2*.

La méthodologie qui a été retenue est d'appliquer une méthode de détection de rupture sur les séries temporelles différenciées, pour permettre de mettre en évidence le fait que l'évolution du coût de la pièce automobile a changé.

Ce chapitre détaillera, pour commencer, la théorie des modèles qui seront utilisés par la suite, soit un modèle de classification non supervisée et des modèles de détection de ruptures. Après cela, l'application de ces modèles sera interprétée d'un point de vue théorique et opérationnel, avant de conclure sur cette partie.

3.1 Présentation théorique des modèles utilisés

Comme évoqué précédemment, plusieurs types de modèles ont été utilisés : un modèle de classification non supervisée de séries temporelles et des modèles de détection de ruptures sur les séries temporelles. Commençons par introduire les notions de séries temporelles avant de présenter ces modèles.

3.1.1 Notions de séries temporelles

Definition 3 (Série temporelle) Une série temporelle est une suite d'observations indexée par le temps.

Dans le cadre de ce mémoire, les données sont des réels qui dépendent des dates des rapports d'expertise et donc dépendent de temporalités discrètes (une donnée par jour, par mois, par trimestre...).

Nous pouvons noter les observations $(y_t)_{t\in\{1,...,T\}}$ avec t étant à valeurs entières allant de 1 à T. Dans notre cas, les temps correspondent aux moyennes des prix des pièces automobiles par trimestre de 2011 à mi-2023.

	T1-2011	T2-2011	...	T4-2022	T1-2023	T2-2023
t	1	2	...	$T-2$	$T-1$	T
	↓	↓	...	↓	↓	↓
X_t	X_1	X_2	...	X_{T-2}	X_{T-1}	X_T

Une série temporelle peut être décomposée en trois éléments, une tendance M, une composante saisonnière S et un résidu x. Généralement, un modèle additif est utilisé pour la décomposer :

$$y_t = M_t + S_t + x_t \tag{3.1}$$

pour $t \in \{1, ..., T\}$, avec $\{M_t\}_t$ la tendance sans saisonnalité, $\{S_t\}_t$ la saisonnalité ou composante saisonnière qui décrit les fluctuations périodiques répétées (de période p connue : $S_{t+p} = S_t$), et $\{x_t\}_t$ le résidu sans tendance ni composante régulière. Souvent $\{M_t\}_t$ et $\{S_t\}_t$ sont des fonctions déterministes alors que $\{x_t\}_t$ est la réalisation d'un processus aléatoire.

Definition 4 (Stationnarité d'une série temporelle) Une série temporelle est dite stationnaire (au second ordre) si :

1. l'espérance est stable dans le temps, i.e. $\forall t \in \mathbb{N}$, $E(X_t) = \mu$
2. et les covariances sont également stables dans le temps, i.e.

$$\forall (t, h) \in \mathbb{N}^2, cov(X_t, X_{t+h}) = cov(X_0, X_h) = \gamma(h) \tag{3.2}$$

Definition 5 (Fonctions d'autocovariance et d'autocorrélation) Lorsque la série temporelle est stationnaire alors sa fonction :

1. d'autocovariance vaut : $\forall (t, h) \in \mathbb{N}^2, \gamma(h) = cov(X_t, X_{t+h})$
2. et d'autocorrélation vaut : $\forall h \in \mathbb{N}, \rho(h) = \frac{\gamma(h)}{\gamma(0)}$

Maintenant que nous avons défini les notions essentielles, nous pouvons commencer à expliquer les premiers modèles.

3.1.2 Modèles de classification de séries temporelles

Dans un premier temps, l'objectif était de regrouper les marques en groupes en fonction de leurs évolutions de prix des pièces pour pouvoir réduire le nombre de marques à suivre. Pour cela, une classification a été utilisée.

Le but d'une classification est de regrouper les individus, ici les séries temporelles, dans des groupes de telle sorte que deux individus dans un même groupe se ressemblent le plus possible et deux individus dans des groupes différents soient les plus distincts. Pour cela, nous pouvons utiliser des modèles de classification non-supervisée mais, même s'il en existe un grand nombre, ils ne sont pas tous adaptés à des séries temporelles.

La méthode qui a été retenue pour sa construction et pour ses performances sur les séries temporelles est l'approche des K-moyennes avec une distance dynamique.

Méthode des centres mobiles, K-moyennes ou K-*means*

La méthode des centres mobiles suppose que le nombre de classes k final soit connu.

Pour décrire cette méthode, on introduit quelques notions comme dans le document [1]. On note $E = e_1, ..., e_n$ l'ensemble des n individus, $G = \frac{1}{n} \sum_{e \in E} e$ le centre de gravité de E et les différentes inerties utilisées dans ces modèles :

— l'inertie totale qui se décrit comme $I(E) = \frac{1}{n} \sum_{e \in E} d(e, G)^2$ avec $d(e, G)$ la distance entre e et G ;

— l'inertie en un point a comme $I_a(E) = \frac{1}{n}\sum_{e\in E} d(e,a)^2$
— et l'inertie intra-classe $I_{intra}(S) = \frac{1}{n}\sum_{j=1}^{k} n_j I(C_j)$ la moyenne des classes C_j pondérées par n_j.

Etapes de l'algorithme K-*means* (voir *figure 3.2*) :

1. *Initialisation* : on choisit k points qu'on appelle les centres des classes.
2. *Itération* : jusqu'à ce que le critère à minimiser ne décroisse plus, par exemple l'inertie intraclasse qui est faible et l'inertie interclasse qui est élevée,
   (a) on affecte les individus à la classe dont le centre de gravité est le plus proche au sens de la distance choisie
   (b) on calcule les nouveaux centres de gravité des classes créées qui deviennent les k nouveaux centres des classes.

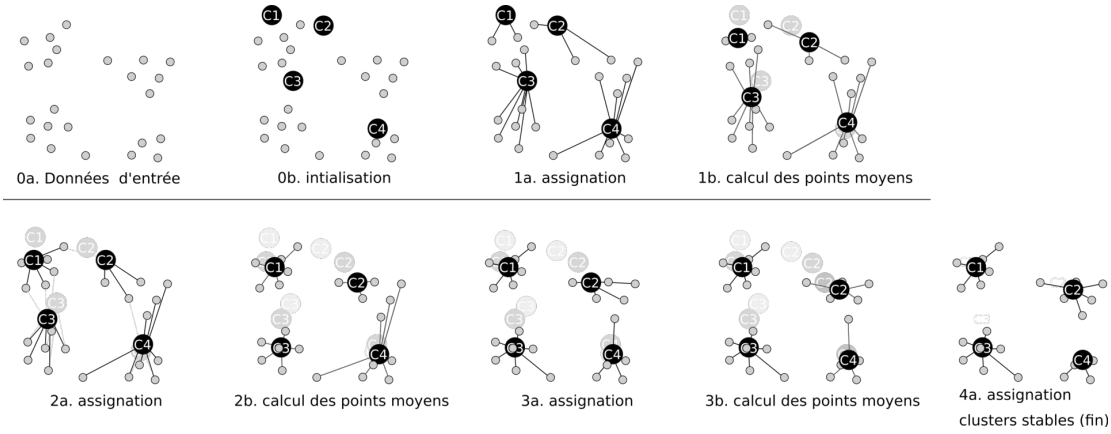


FIGURE 3.2 – Etapes de l'algorithme du K-means [1]

Ainsi, dans cet algorithme on voit deux éléments apparaître : la notion de centre de gravité et la distance choisie.

La distance utilisée par défaut est la distance euclidienne qui compare la distance des deux séries en prenant les deux points situés aux mêmes abscisses.

Definition 6 (Distance euclidienne)

$$Distance\ euclienne(x,y) = d(x,y) = ||x-y||^2 = \sqrt{\sum_{i,j}(x_i - y_j)^2} \tag{3.3}$$

Pour prendre en compte l'aspect temporel des séries, il existe une autre distance nommée la *Dynamic Time Warping*.

1. Source : www.wikiwand.com/fr/K-moyennes

La déformation temporelle dynamique ou *Dynamic Time Warping* (DTW)

Lors de l'utilisation d'un modèle des K-moyennes classique, la distance utilisée par défaut est la distance euclidienne. Certaines séries peuvent cependant avoir les mêmes courbes avec un décalage dans le temps ou une déformation qui a eu lieu avec une vitesse différente. Pour prendre en compte ces déformations temporelles et obtenir une meilleure synchronisation des séries, un algorithme pour calculer une distance dynamique appelée déformation temporelle dynamique, ou en anglais *Dynamic Time Warping* (DTW), existe (voir *figure 3.3*). Cette distance compare les chemins entre les différents points des séries temporelles pour obtenir le plus court chemin. Cette distance est celle qui a été retenue dans le modèle des K-moyennes, étant le modèle le plus adapté aux séries temporelles.

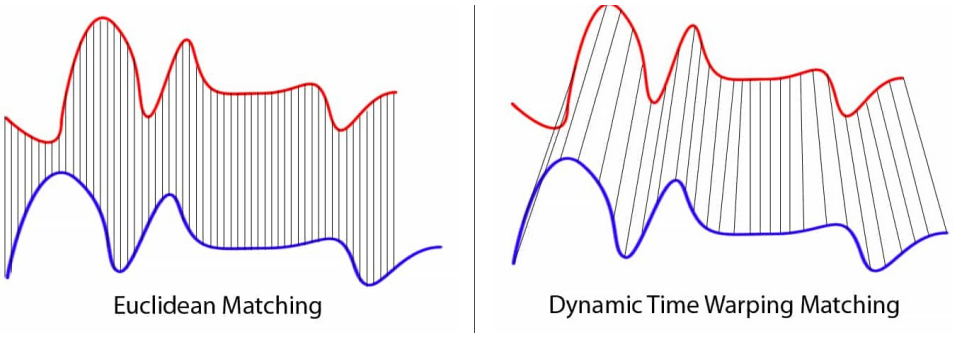


FIGURE 3.3 – Comparaison entre la distance euclidienne et la distance dynamique DTW (github.com/DBA)

Statistiquement, on introduit deux séries différentes $S = \{s_1, s_2, ..., s_m\}$, $m \in \mathbb{N}$ et $T = t_1, t_2, ..., t_n$, $n \in \mathbb{N}$.

Definition 7 (Distance DTW [2]) On peut définir de manière récursive la distance DTW entre les séries $s_1, ..., s_i$ avec $1 \leq i \leq m$ et $t_1, ..., t_j$ avec $1 \leq j \leq n$ notée $D(i,j) = D((s_1, ..., s_i), (t_1, ..., t_j))$ telle que :

$$D(i,j) = \begin{cases} |s_1 - t_1|, & \text{si i=j=1} \\ |s_i - t_j| + min\{D(i-1,j), D(i-1,j-1), D(i,j-1)\}, & \text{sinon.} \end{cases} \quad (3.4)$$

Cette distance DTW est une méthode qui recherche parmi les points de deux séries x et y ceux qui ont un appariement optimal pour permettre de suivre au mieux les tendances. Elle prend donc en compte les décalages temporels qu'il peut y avoir entre les deux séries.

2. Source : ledatascientist.com/k-means-sur-des-series-temporelles/ et en.wikipedia.org/wiki/Dynamic_time_warping

L'algorithme *Dynamic Time Warping Barycenter Averaging* (DBA)

Comme la distance utilisée n'est plus la distance euclidienne entre les séries temporelles, il faut également trouver une façon cohérente pour déterminer le centre de gravité pour les classes obtenues. C'est l'objet de l'algorithme *Dynamic Time Warping Barycenter Averaging* (DBA) [10].

En effet, cet algorithme minimise la somme des carrés de la distance DTW entre le barycentre et les séries du cluster.

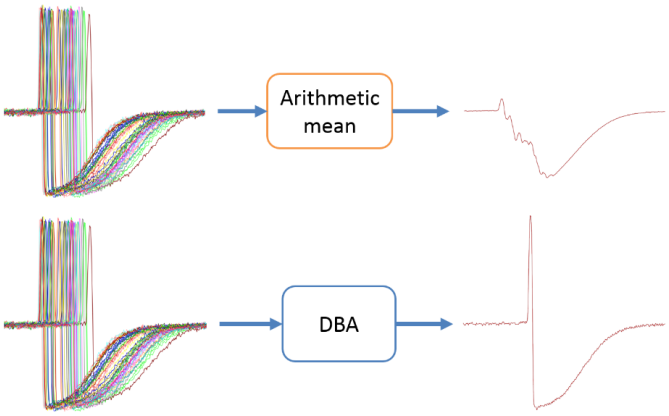


FIGURE 3.4 – Comparaison entre la moyenne arithmétique et l'algorithme DBA (*Source : github.com/fpetitjean/DBA*)

Comment choisir le nombre de clusters k ?

L'algorithme des K-moyennes, contrairement à d'autres classifications non supervisées, suppose de définir *a priori* un nombre de clusters k à construire. Il faut donc réussir à le déterminer.

Le choix du nombre de cluster k peut se faire grâce à plusieurs méthodes et celle qui a été choisie ici est la **méthode du coude**. Celle-ci revient à exécuter plusieurs k-moyennes avec plusieurs k prenant des valeurs croissantes et d'enregistrer les sommes des inerties intra-classes des modèles notées SSE pour *Sum of the Square Error*.

Ces SSE permettent notamment de tracer une courbe avec en abscisse le nombre de classes k et en ordonnée le montant du SSE. Plus on ajoute des classes à la classification et plus cette SSE sera faible puisqu'on réduit les distances de chaque point avec son centre de gravité le plus proche. La conséquence de cela est que la courbe en question est décroissante.

Sur cette courbe décroissante, on observe des ruptures dans le rythme de décroissance de cette courbe. Ces "coudes" sont alors des candidats privilégiés pour choisir un nombre de cluster.

3.1.3 Modèles de détection de ruptures sur les séries temporelles

Deuxièmement, on souhaite créer un outil de détection automatique d'un changement de politique tarifaire de marques ou regroupements de marques de véhicules dans le but de suggérer une mise à jour du processus tarifaire ou d'anticiper l'augmentation des sinistres.

Deux approches ont été envisagées, une détection de "dérive des données", autrement appelée *datadrift*, et une détection de rupture sur les moyennes des séries temporelles différenciées. Comme la volumétrie des points dont nous disposions était insuffisante pour détecter des dérives des données (*Annexe 5.2*), la seconde approche a été appliquée ici.

Les cadres similaires aux bibliographies [13] et [26] ont été repris pour cette approche.

Définition d'une rupture et du cadre

Dans le cas de la série temporelle $X = (X_1, ..., X_T)$ composée de T observations d'intervalles réguliers, une rupture est définie comme un changement brutal d'une ou plusieurs caractéristiques du signal (de la moyenne, de la variance, de la famille de distribution, du moment d'ordre r, des composantes spectrales ou encore des coefficients d'auto-régression).

Les observations X_i sont supposées indépendantes entre elles, avec $\forall i \in \{1, ..., T\}$, $X_i \sim F_i$ une distribution de probabilité de paramètre Θ_i.

Definition 8 (Rupture) X_τ est une rupture pour $\tau \in \{2, ..., T-1\}$ si $\Theta_\tau \neq \Theta_{\tau-1}$. Par convention X_1 et X_T sont des ruptures.

Dans notre cas, on recherche une rupture sur la tendance de la série temporelle et donc sur la moyenne de la série différenciée.

Pour cela, on construit une statistique de test Z permettant de choisir entre les hypothèses :

— H_0 : il n'y a pas de rupture, *i.e* $\Theta_1 = \Theta_2 = ... = \Theta_T$

— H_1 : il y a une rupture en $\tau \in \{2, ..., T-1\}$, *i.e* $\Theta_1 = \Theta_2 = ... \neq \Theta_\tau = \Theta_{\tau+1} = ... = \Theta_T$

Ainsi, une rupture est considérée lorsque la probabilité d'observer cette statistique Z est supérieure à un seuil approprié, usuellement 5%, sinon H_0 est rejeté.

Plus généralement, on peut élargir la définition des ruptures au cas à K ruptures notées $\{\tau_1, ..., \tau_K\}$ découpant la série temporelle $X = (X_1, ..., X_T)$ en $K + 1$ segments.

Le package *rupture* sur *python*

Dans ce package [3], plusieurs modèles sont disponibles pour détecter des ruptures sur une série temporelle qui n'est pas stationnaire (c'est à dire que l'espérance ou la covariance de la série peut dépendre du temps).

3. https://pypi.org/project/ruptures/

Comme détaillé dans l'écrit associé au package [5], pour identifier le bon modèle à appliquer sur des séries temporelles, il est important d'identifier les trois éléments suivants :

— La fonction de coût $c(.)$ qui est une fonction d'homogénéité pour détecter les changements.
On introduit $X = X_{0...T}$ la série complète et $X_{a...b}$ la série des temps de X allant de a+1 à b. Alors, une rupture est envisagée si $c(X_{a...b})$ est élevée, c'est à dire que le sous-signal $X_{a...b}$ est hétérogène. Sinon si $X_{a...b}$ est homogène, alors $c(X_{a...b})$ est faible et il n'y a pas de points de rupture.

La fonction de coût à minimiser par la segmentation R est :

$$V(R, X) := \sum_{k=0}^{K} c(X_{\tau_k ... \tau_{k+1}}) \tag{3.5}$$

avec c(.) la fonction de coût associée au sous-signal de X sur $[\tau_k + 1, \tau_{k+1}]$ et la segmentation R $=\{\tau_1, ..., \tau_K\}$ avec les K temps de ruptures.
Dans notre cas, la fonction de coût utilisée est une fonction paramétrique car nous avons une série de dimension finie. Le type de fonction de coût qui généralise plusieurs modèles et coûts sont ceux basés sur les estimateurs par maximum de vraisemblance. Lorsqu'on suppose que la variance est fixe, alors la fonction est définie comme :

Definition 9 (Fonction de coût c_{L_2})

$$c_{L_2}(X_{a...b}) := \sum_{t=a+1}^{b} ||X_t - \overline{X_{a...b}}||_2^2 \tag{3.6}$$

avec $\overline{X_{a...b}}$ la moyenne empirique du sous-signal $X_{a...b}$.

— La méthode de recherche correspondant à la procédure d'optimisation du problème. Si le nombre de ruptures $K \in \mathbb{N}$ est connu, alors le problème consiste à minimiser le critère V(R, X) pour obtenir $\hat{R} = \{\hat{\tau}_1, ..., \hat{\tau}_K\}$:

$$min_{|R|=K} V(R, X), \text{ avec } |R| \text{ le cardinal de } R \tag{3.7}$$

Sinon, le critère V(R, X) est à minimiser en plus d'une pénalité qui prend en compte la complexité de la segmentation notée $pen(.)$:

$$min_R V(R, X) + pen(R) \tag{3.8}$$

Dans ces deux cas, il existe des méthodes pour trouver des ruptures dites "optimales" et d'autres pour avoir des ruptures "approximatives" [5]. Les méthodes de détection optimales permettent de résoudre les problèmes (3.7) et (3.8) de

manière exacte (*Opt* et ses extensions quand le nombre de ruptures K est connu et *Prune Exact Linear Time, dite Pelt,* quand il n'est pas connu et que la pénalité est linéaire). Les méthodes de détection par approximation sont souvent appliquées lorsque les calculs des méthodes exactes sont trop complexes (*window sliding*, *binary segmentation* ou encore *bottom-up segmentation*).

Dans notre cas, nous avons comparé une méthode approximative et une méthode optimale :

— la méthode optimale *Pelt* qui est la version optimale dans le cas où le nombre de ruptures n'est pas connu, pour voir de manière précise quelles sont les ruptures obtenues. La méthode optimale consiste à énumérer toutes les segmentations possibles du signal et de retourner celle qui minimise la fonction objectif $V(R, X)$.

Definition 10 (Méthode exacte *Prune Exact Linear Time*) Pour la méthode *Pelt*, la minimisation est effectuée sur l'ensemble $\{\tau : 1 \leq |\tau| < T\}$ qui contient $\sum_{K=1}^{T-1} \binom{T-1}{K-1}$ éléments.

— et la méthode approximative *binary segmentation*. La méthode approximative entre dans la catégorie des approches de détection séquentielle, cela signifie que pour chaque itération i, l'algorithme estime un seul point de changement $\hat{\tau}_i$ compris entre 1 et T.

Definition 11 (Méthode *Binary segmentation*) La *binary segmentation* plus spécifiquement est un algorithme, optimal ou approximatif en fonction de la connaissance ou non du nombre de ruptures K. Ce modèle commence par estimer le premier point de rupture en minimisant la somme des coûts des deux sous-signaux créés par les ruptures envisagées, cela revient à résoudre l'équation suivante :

$$\hat{\tau}_i := argmin_{1 \leq t < T-1} V(R = \{t\}) = argmin_{1 \leq t < T-1} c(X_{0...t}) + c(X_{t...T}) \quad (3.9)$$

Ensuite, le signal est coupé en ce point $\hat{\tau}_i$ en créant deux sous-signaux $X_{0...\hat{\tau}_i}$ et $X_{\hat{\tau}_i...T}$. La même opération est réalisée sur chaque sous-signal comme on peut le voir sur la *figure 3.5*.
Dans le cas où le nombre de ruptures K est connu, alors l'algorithme s'arrête à la K^{eme} itération. Sinon, lorsque K est inconnu, l'algorithme séquentiel est exécuté jusqu'à la réalisation du critère d'arrêt renseigné, la pénalité.

— La contrainte. Dans le cas où le nombre de ruptures K n'est pas connu, alors il est nécessaire de renseigner une contrainte pour que les algorithmes s'arrêtent. Dans ce cas, si la pénalisation $pen(\tau)$ est trop faible par rapport à la qualité de l'ajustement, alors beaucoup de points seront détectés à cause du bruit du signal. Si elle est trop élevée, aucun ou peu de points de rupture ne seront trouvés.
La pénalité la plus utilisée est la pénalité linéaire définie par :

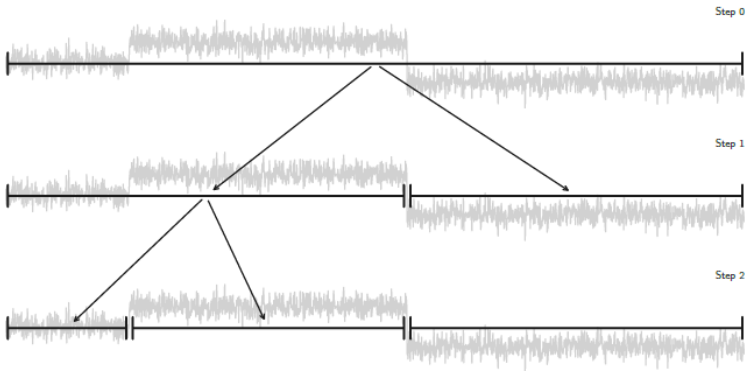


FIGURE 3.5 – Illustration de la méthode *binary segmentation*

Definition 12 (Pénalité linéaire)

$$pen_{l_0}(R) := \beta|R| \tag{3.10}$$

avec $\beta > 0$ un paramètre de lissage qui définit une appétence plus forte à la complexité avec un grand nombre de segmentations (β faible) ou à la qualité du modèle (β élevé).

Le problème à résoudre revient à trouver la meilleure segmentation $R = \{\tau_1, ..., \tau_K\}$ pour minimiser le critère égal à la somme des coûts des différents segments possibles $min_R \sum_{k=0}^{K} c(X\tau_k...\tau_{k+1}) + pen(R)$ en utilisant les méthodes détaillées au-dessus.

3.2 Approche historique

Maintenant que nous connaissons les différents modèles que nous voulons utiliser et les raisons pour lesquelles nous voulons les utiliser, nous pouvons commencer leur application en deux approches. Une première approche va tester ces modèles avec un recul sur notre historique de données pour évaluer leur pertinence et une seconde approche se tournera sur une utilisation opérationnelle.

3.2.1 Préparation des séries temporelles

A partir des rapports d'expertise, nous avons à disposition les informations des pièces automobiles, notamment de cinq pièces automobiles, que nous avons décidé de suivre pour leur stabilité dans le temps et leur importance en moyenne dans le montant de la facture.

Ces pièces retenues sont les "boucliers avant", les "boucliers arrière", les "ailes avant gauche", les "portes avant" et les "portes arrière". Ainsi, pour chaque périmètre étudié, nous avons cinq "courbes" à disposition.

Donc si nous voulons suivre 30 marques par exemple, nous aurons alors $5 \times 30 = 150$ courbes à suivre.

Pour réduire ce nombre de courbes, deux étapes ont été nécessaires.

1. Tout d'abord, nous avons utilisé une classification non supervisée pour regrouper les marques en fonction de leurs similitudes dans leur évolution du coût de leurs pièces. Grâce à cela, il a été possible de réduire le nombre de modalités de marques à suivre.
2. Dans un second temps, les informations des groupes obtenus ont été agrégées à l'aide d'une pondération pour arriver à une seule courbe finale.

Expliquons chaque étape.

Etape 1 : classification non supervisée des marques

- Approche choisie : la méthode de *clustering* qui a été utilisée pour réunir les individus en groupes similaires est l'**approche des K-moyennes avec la distance DTW**, comme présenté dans la partie 3.1.2.

Pour ce modèle, deux distances ont été envisagées : la distance euclidienne et la distance dynamique nommée DTW. Ces deux distances peuvent donner des résultats bien différents. Même si la distance euclidienne est la distance la plus couramment utilisée, celle qui a été retenue est la **distance DTW étant la plus adaptée aux séries temporelles**. Elle prend en compte les potentiels décalages de temps qu'il peut y avoir entre les évolutions des séries contrairement à la distance euclidienne.

- Traitement des données : comme évoqué précédemment, pour chaque marque étudiée, nous avons une courbe pour l'évolution du prix du bouclier avant, une pour celle du bouclier arrière, une pour l'aile avant gauche etc. Ces **évolutions sont obtenues en considérant le traitement usuel** en retirant les valeurs aberrantes et en lissant les séries temporelles sur trois trimestres, comme détaillé dans la *partie 2.1.6*. Ainsi pour chaque courbe, il peut y avoir des valeurs manquantes sur certains trimestres. Pour pallier à cela, une **interpolation sur un ou deux trimestres** a été appliquée sur ces courbes "*marque* $\times$ *pièce*".

Suite à ce traitement, pour chaque marque, dans le but de considérer les informations de ces cinq pièces en même temps, les cinq courbes des pièces ont été **agrégées en tenant compte la répartition moyenne de ces pièces dans le temps**. Cela correspond à une pondération stable dans le temps permettant ainsi de ne pas être impacté par une baisse de volume d'une des pièces à un moment donné. Les répartitions trimestrielles moyennes sont :

Pièce	Répartition (%)
Bouclier avant	38%
Bouclier arrière	29%
Aile avant gauche	14%
Porte avant	11%
Porte arrière	8%
Total	100%

Cette agrégation se définit comme suit :

Definition 13 (Agrégation des courbes des cinq pièces par marque)

$$CM_{marque_i,t} = \sum_{piece \in I} CM_{marque_i,piece,t} \times \%piece \tag{3.11}$$

avec $CM_{marque_i,piece,t}$ le coût moyen de la pièce obtenu pour la marque i en t traité des valeurs aberrantes et lissé sur trois trimestres comme détaillé *en 2.1.6*, $I = \{$bouclier avant, bouclier arrière, aile avant gauche, porte avant, porte arrière$\}$ et pour chaque *piece*, leurs répartitions moyennes données ci-dessus.

Cela revient, pour chaque marque, à calculer de manière séparée le coût moyen traité de chaque pièce et d'appliquer leur pourcentage de présence pour arriver au montant de la marque. On peut voir par exemple l'agrégation réalisée pour la Peugeot en *figure 3.6*.

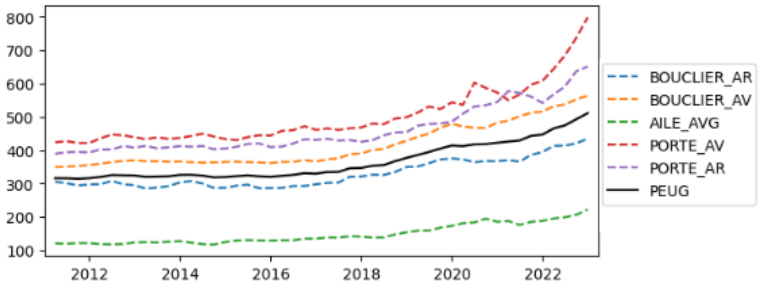


FIGURE 3.6 – Evolutions des cinq pièces suivies de la marque Peugeot (en pointillés) et de la marque obtenue à partir des cinq autres (en noir)

Comme nous n'avions pas suffisamment de données pour certaines marques formant des courbes incomplètes, nous avons choisi de suivre uniquement les **12 marques les plus représentées dans les rapports d'expertise** individuellement qui correspondent à **80% des rapports d'expertise auto**, et les autres sont réunies dans une modalité "Autres".

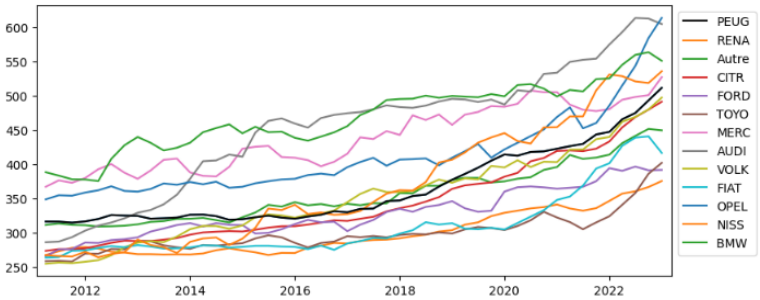


FIGURE 3.7 – Evolutions des coûts des pièces des marques agrégées par marque

Nous obtenons alors les courbes agrégées des marques visibles en *figure 3.7*, sur lesquelles nous application une normalisation avant d'appliquer le modèle de classification. La normalisation qui a été utilisée est la normalisation des valeurs entre leur minimum et leur maximum.

Definition 14 (Normalisation par échelle "min-max")

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{3.12}$$

On peut schématiser le traitement appliqué sur les courbes de cette manière :

— Etape 1 : Calcul des courbes des coûts moyens des pièces des marques en utilisant le traitement habituel - sans valeurs aberrantes et avec lissage sur trois trimestres - et en interpolant sur deux trimestres consécutifs ;
— Etape 2 : Agrégation des courbes des différentes pièces par marque, de sorte d'avoir une courbe par marque en utilisant les moyennes des répartitions des pièces ;
— Etape 3 : Normalisation des courbes des marques ;
— Etape 4 : Application des modèles de classification non supervisée.

On peut représenter ces étapes comme sur la *figure 3.8*.

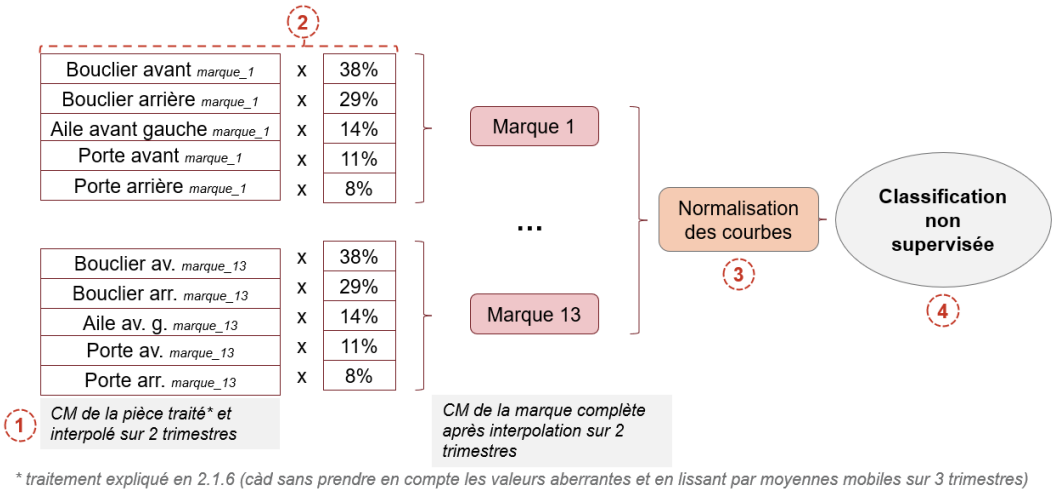


FIGURE 3.8 – Schéma du traitement des courbes avant la classification non supervisée

- Application du modèle de classification non supervisée : avant d'appliquer les modèles de classification sur toutes les marques confondues, une interrogation qui a été émise est de savoir si les marques au sein d'un même regroupement suivaient les mêmes tendances et pouvaient donc être associées. C'est la question à laquelle nous avons essayé de répondre graphiquement à la *partie 2.1.7*.

Pour répondre à cette question, le paragraphe suivant prend l'exemple particulier du groupe Stellantis venant d'une alliance récente de 2021 entre les deux sous-groupes nommés Peugeot Société Anonyme (PSA) et Fiat Chrysler Automobiles (FCA).

Premier cas : regroupement des marques au sein d'une concentration de marques

Commençons par regarder les classes obtenues avec les k-moyennes de deux et trois groupes sur les marques du groupe Stellantis, faisant partie des 12 premières marques, en *figures 3.9 et 3.10*. Sur ces graphiques, chaque couleur correspond à un cluster différent.

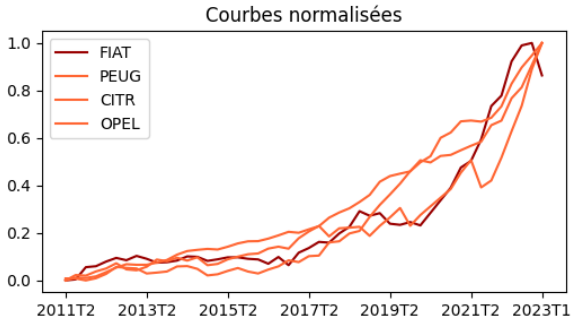


FIGURE 3.9 – K-moyennes des courbes de Stellantis avec 2 clusters

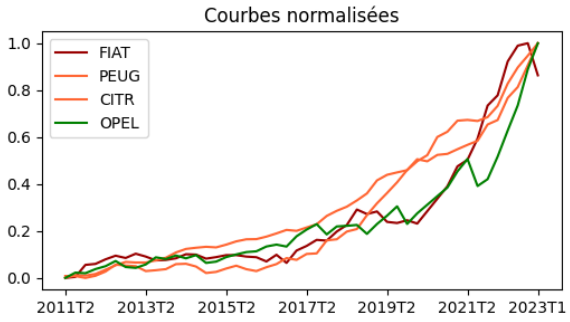


FIGURE 3.10 – K-moyennes des courbes de Stellantis avec trois clusters

Sur le premier graphique *3.9*, on constate une distinction nette entre Citroën, Peugeot et Opel qui appartiennent à PSA par rapport à Fiat qui appartient quant à lui à FCA. On remarque de plus sur le second graphique *3.10* qu'Opel, qui a rejoint le groupe en 2017, a eu des évolutions moins proches de celles de Citroën ou de Peugeot.

Ainsi on observe qu'au sein d'un même groupe les marques peuvent avoir des évolutions bien différentes. Cela peut venir de deux aspects :

1. D'abord, comme nous l'avons vu, de l'**intégration des marques aux groupes à des instants différents**. Comme Peugeot et Citroën étaient dans le même groupe depuis longtemps alors les tendances sont proches alors qu'Opel qui est arrivée plus tard se trouve plus éloignée de ces marques ;
2. Ensuite du fait que ces groupes ont des marques de **différentes gammes** et donc destinées à des publics de clients différents. On peut supposer que les regroupements de marques répartissent, selon les clients visés par chaque marque, leurs évolutions de prix pour atteindre leur espérance de marge.

Avec cet exemple, on voit la difficulté de regrouper les marques dans leurs groupes financiers.

De plus, avec le choix de suivre les 12 premières marques individuellement et les autres réunies ensemble, on constate que **les groupes ne contiennent plus beaucoup**

de marques séparées. Seuls les trois premiers regroupements ont plus d'une marque, les autres ont uniquement leur marque la plus importante comme Mercedes ou alors seulement des petites marques qui sont comprises dans le groupe "Autre". Les 12 marques suivies individuellement sont réparties ainsi dans les regroupements financiers présentés en *partie 1.2* :

	Groupe	Marques retenues
1.a	Stellantis PSA	Citroën, Opel, Peugeot
1.b	Stellantis FCA	Fiat
2	Renault-Nissan-Mitsubishi	Nissan, Renault
3	Volkswagen	Audi, Volkswagen
4	Mercedes Benz group	Mercedes
5	BMW	BMW
6	Ford	Ford
7	Toyota	Toyota
8-12	Hyundai, Geely, Tata, Honda, General Motors	

De cette façon, il semble discutable de faire des classes au sein de chaque groupe de marques.

La seconde idée est de regrouper les marques sans prendre en compte leur appartenance à un groupe. On peut s'attendre à voir des groupes avec des marques du même regroupement financier à cause de potentielles décisions tarifaires communes, même si ces décisions peuvent être différentes en fonction des stratégies commerciales et surtout des clients visés par exemple avec la gamme du véhicule.

Second cas : regroupement des marques, indépendamment du groupe de marques auxquelles elles appartiennent

Dans le cas présent, nous comparons 13 courbes d'évolution du coût des cinq pièces suivies.

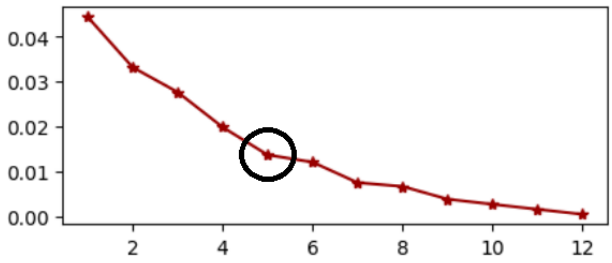


FIGURE 3.11 – SSE des k-moyennes en fonctions de k

Pour réaliser la classification finale par K-moyennes, on choisit le nombre de cinq classes à partir du graphique des SSE en *figure 3.11* sur lequel un léger coude apparaît.

Même si d'autres points de coudes apparaissent sur ce graphique, ce nombre "cinq" a été choisi permettant de suffisamment différencier les groupes de marques sans être trop restrictif.

On réalise alors la classification par méthode des k-moyennes avec la distance dynamique DTW et ce nombre de classes, que l'on représente en *figure 3.12*.

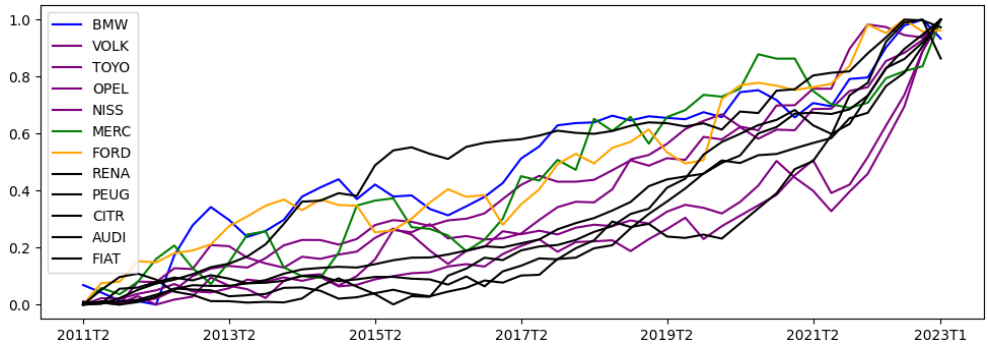


FIGURE 3.12 – Clustering K-means des 12 marques retenues avec la distance DTW et cinq classes

On obtient, à la suite de ce traitement, les 6 groupes suivants auxquels on ajoute le groupe "Autres marques" :

- Groupe 1 : BMW ;
- Groupe 2 : Volkswagen, Toyota, Opel et Nissan ;
- Groupe 3 : Mercedes ;
- Groupe 4 : Ford ;
- Groupe 5 : Renault, Peugeot, Citroën, Audi, Fiat ;
- Groupe 6 : Autres

Ces groupes seront utilisés dans la partie suivante pour diminuer le nombre de courbes à agréger.

Etape 2 : Agrégation des données pour avoir une seule courbe à suivre

Initialement, nous suivions cinq pièces sur 13 modalités de marques différentes (12 marques distinctes et les autres regroupées ensemble sous la modalité "Autre") ce qui correspondait à $5 \times 13 = 65$ courbes à étudier.

Pour arriver à une courbe finale, on a procédé en trois étapes :

1. Calculer les courbes des six groupes obtenus par classification non supervisée dans la partie précédente et non plus sur les 12 marques et les autres initialement. Grâce à cela, on réduit le nombre de courbes de $5 \times 13 = 65$ à $5 \times 6 = 30$ courbes.

Dans ce cas, une importance a été attribuée au volume d'informations que nous avions pour détecter les ruptures et alors aucun lissage n'a été appliqué [4].

2. Agréger pour chaque groupe de marques les courbes correspondant à leurs cinq pièces, en utilisant leur répartitions trimestrielles moyennes comme précédemment. L'objectif est de considérer les informations de ces cinq pièces simultanément pour chaque groupe. Cela revient à faire :

$$CM_{groupe_j,t} = \sum_{piece \in I} CM_{groupe_j,piece,t} \times \%piece \tag{3.13}$$

avec $CM_{groupe_j,piece,t}$ le coût moyen de la pièce obtenue pour chaque groupe au trimestre t traité [5], $I = \{$bouclier avant, bouclier arrière, aile avant gauche, porte avant, porte arrière$\}$ et pour $\%Piece_i$, les répartitions des pièces.

3. Réitérer l'opération qui agrège les courbes des groupes de marques de manière équivalente à celle utilisée pour les pièces. C'est-à-dire qu'on utilise les **parts moyennes** de ces groupes données par :

Groupe	Marques	Répartition
5	Renault, Peugeot, Citroën, Audi et Fiat	55%
6	Autres	20%
2	Volkswagen, Toyota, Opel, Nissan	14%
3	Mercedes	4%
4	Ford	4%
1	BMW	3%
Total		100%

Cela revient à faire :

$$CM\ final = \sum_{j \in J} CM\ Groupe\ marques_{t,\,j} \times \%Groupe\ marques_j \tag{3.14}$$

avec $CM\ Groupe\ marques_{t,\,j}$ le coût moyen en t du groupe j calculé précédemment, $J = \{Groupe\ 1, Groupe\ 2, Groupe\ 3, Groupe\ 4, Groupe\ 5, Groupe\ 6\}$ et $\%Groupe\ marques_j$ les répartitions des groupes.

On peut résumer cela dans le schéma *3.13*.

4. Pour ne pas perdre les informations du premier trimestre 2011 et surtout du deuxième trimestre 2023, les courbes des coûts moyens des pièces considérées sont celles obtenues grâce aux pièces sans valeurs aberrantes comme pour le traitement classique mais sans appliquer de lissage.

5. Avec un traitement des valeurs aberrantes comme détaillé en *partie 2.1.6* mais **sans lissage**

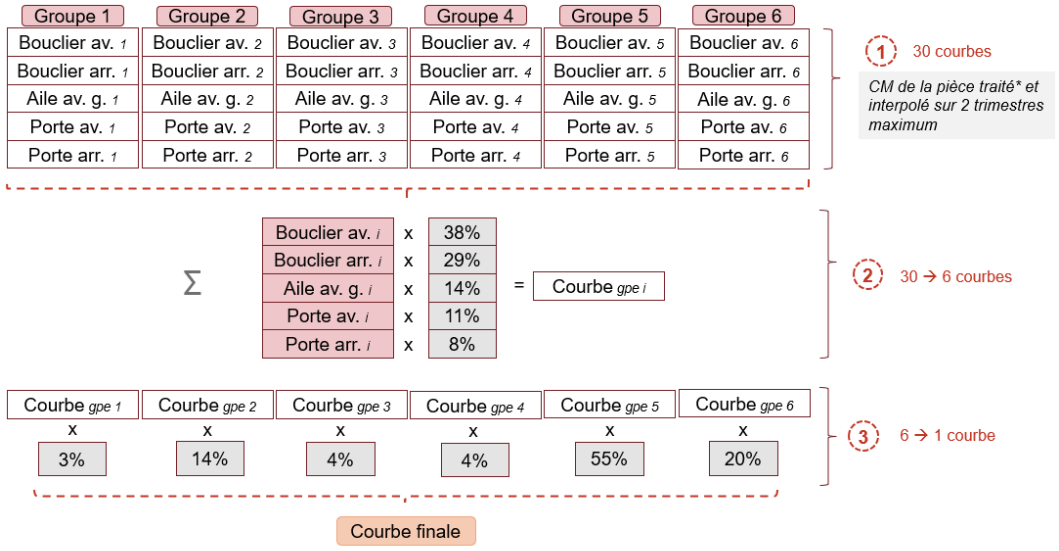


FIGURE 3.13 – Schéma résumé du traitement des courbes avant la détection de rupture

On obtient alors la courbe finale représentée en noire sur la *figure 3.14*. Il s'agit bien d'une seule courbe à suivre dans le temps sur laquelle on peut appliquer les modèles de détections de ruptures.

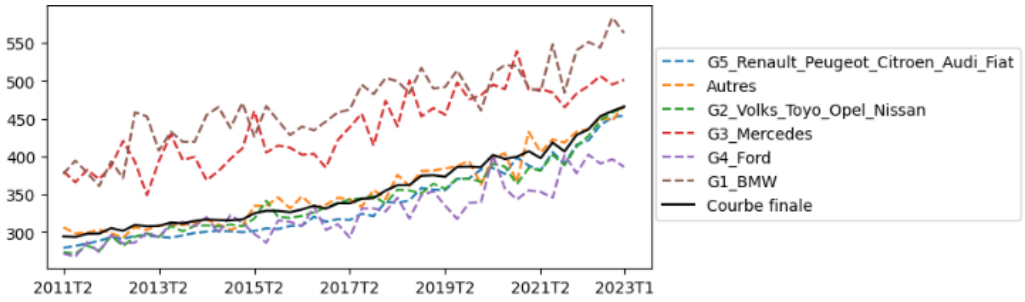


FIGURE 3.14 – Evolutions des groupes obtenus en pointillés et courbe finale en noire

3.2.2 Application des modèles de détection de ruptures

Comme présenté dans la *partie 3.1.3*, par rapport aux données que nous avons, seules deux méthodes de ruptures ont été choisies. La première est la méthode de segmentation binaire, ou *binary segmentation*, qui est une méthode approximative dans le sens où elle ne teste pas chaque segmentation mais sélectionne ses points de ruptures de manière itérative. La seconde est la méthode *Pelt* qui permet de trouver la segmentation optimale pour séparer la courbe en segments avec des tendances similaires.

La méthode approximative de segmentation binaire

Cette méthode a pour particularité de chercher **un point de rupture à chaque itération** qui sépare le ou les segments en deux **jusqu'à ce qu'un critère d'arrêt** ne soit atteint. Grâce à un paramètre nommé la "pénalité", nous pouvons ajuster la sensibilité de détection des ruptures du modèle.

Si nous choisissons une pénalité de 0 par exemple, alors beaucoup de temps de ruptures apparaissent, alors qu'avec une pénalité de 10, seules trois ruptures sont détectées. Pour comparer les deux, nous pouvons regarder les graphiques en *figure 3.15*, sur lesquels les courbes sur la première ligne représentent la série différenciée sur laquelle le modèle de détection a été appliqué, et la seconde est la courbe initiale pour constater des évolutions entre les points de ruptures.

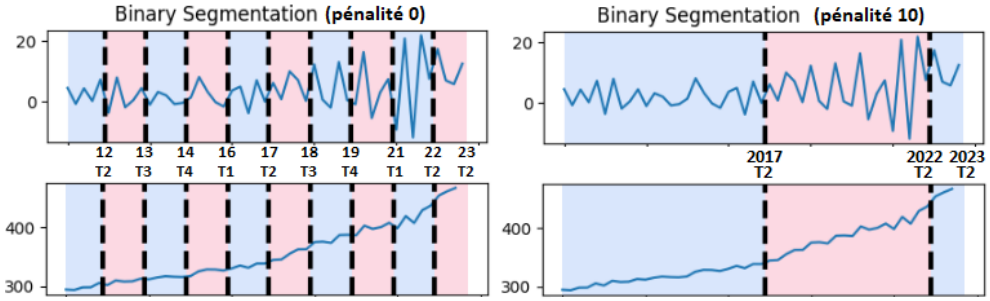


FIGURE 3.15 – Représentation des ruptures repérées grâce à la segmentation binaire sur la courbe différenciée (en haut) et la courbe initiale (en bas), avec les pénalités nulle et de 10

Afin de décider de la sensibilité désirée, le modèle a été appliqué avec plusieurs pénalités jusqu'à ce que les points de ruptures indiquées soient stables. Ainsi, nous avons eu pour différentes pénalités ces temps de ruptures :

Pénalité(s)	Ruptures obtenues avec la segmentation binaire
1	2012T2, 2013T3, 2014T4, 2017T2, 2019T4, 2021T1, 2022T2, 2023T2
2	2012T2, 2014T4, 2017T2, 2019T4, 2021T1, 2022T2, 2023T2
$3 \rightarrow 5$	2012T2, 2014T4, 2017T2, 2021T1, 2022T2, 2023T2
$6 \rightarrow 7$	2012T2, 2014T4, 2017T2, 2022T2, 2023T2
8	2012T2, 2017T2, 2022T2, 2023T2
$9 \rightarrow 97$	**2017T2, 2022T2, 2023T2**
$98 \rightarrow 198$	2022T2, 2023T2
$198 \rightarrow ...$	2023T2

Avec ce modèle, même avec une pénalité de 10 qui est assez restrictive, on obtient une rupture en milieu de l'année 2022. Cette détection est pertinente avec le contexte de forte inflation qui a été observé pour cette année. Cela nous indique que l'hypothèse de projection de l'inflation qui se base sur la tendance connue devrait être mise à jour.

La méthode optimale *Pelt*

Après les modèles approximatifs, regardons les ruptures repérées par le modèle optimal, censé être plus précis et fiable.

Pour cette méthode, il n'y a pas de critère supplémentaire à renseigner. De cette façon, ce modèle teste **toutes les segmentations envisageables** pour mettre en évidence celle qui est la meilleure selon le critère à optimiser.

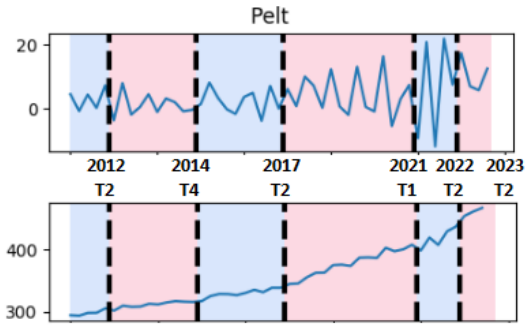


FIGURE 3.16 – Ruptures obtenues avec la méthode *PELT* sur la courbe différenciée et la courbe normale

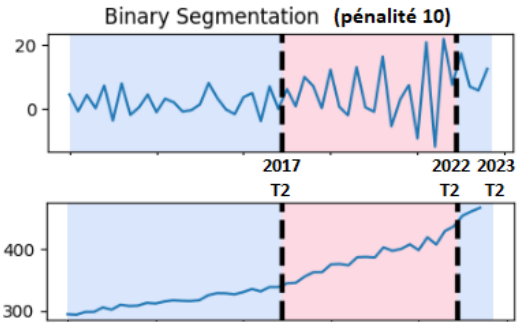


FIGURE 3.17 – Ruptures obtenues avec la méthode *Binary segmentation* avec une pénalité de 10, sur la courbe différenciée et la courbe normale

Sur les *graphiques 3.16 et 3.17*, nous avons encore représenté la courbe différenciée sur laquelle le modèle a repéré les ruptures et en dessous la courbe initiale qui permet de constater des tendances.

On remarque que les points de ruptures indiqués par ce modèle sont les mêmes que ceux fournis par la segmentation binaire avec une pénalité de 3 à 5. Ces points de ruptures ressortis sont alors 2012T2, 2014T4, 2017T2, 2021T1, 2022T2, et 2023T2. Dans ce cas, au cours de l'année 2022, une rupture est encore repérée.

Pour conclure sur cette première approche, le second trimestre 2022 a été détecté par ces nombreux modèles comme étant une rupture significative de la tendance des coûts des pièces automobiles. Ainsi, une prise en compte de ce changement pourrait être pertinente dans le processus de tarification.

3.3 Approche opérationnelle

D'un point de vue opérationnel, on ne peut pas supposer que les marques qui ont été regroupées selon leurs évolutions par classification non supervisée vont continuer à suivre des évolutions similaires. Ainsi on ne peut donc pas utiliser ces groupes dans l'étape 1 afin d'effectuer un suivi régulier.

3.3.1 Agrégation des courbes stables dans le temps

Le nouveau traitement est résumé dans la *figure 3.18*.

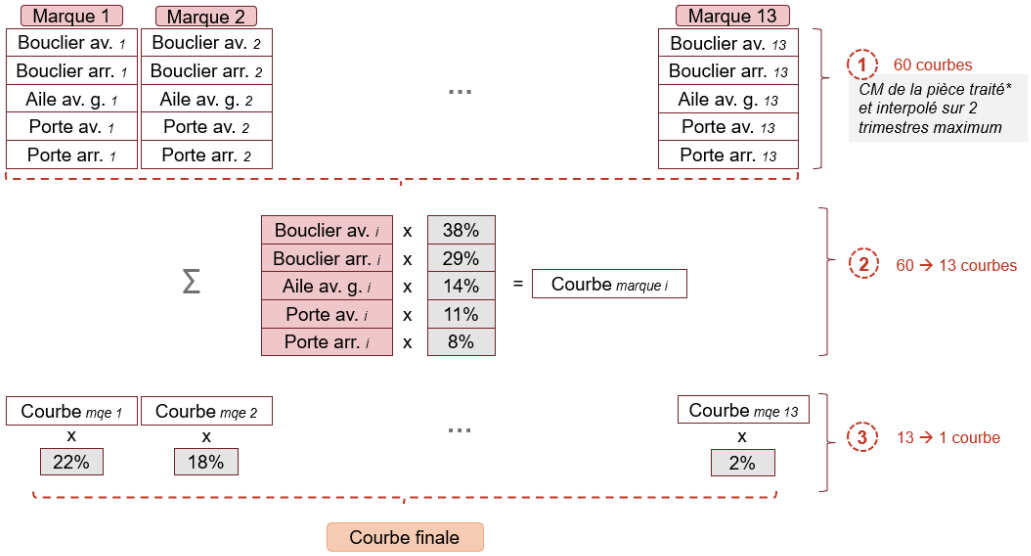


FIGURE 3.18 – Schéma résumé du traitement **opérationnel** des courbes avant la détection de rupture

Dans ce cas, on reprend les mêmes étapes que celles du traitement précédent en modifiant quelques points :

— tout d'abord, en étape 1 les pièces utilisées pour créer les courbes ne sont pas les mêmes. Au lieu de prendre en compte les pièces des groupes créés par clustering, on regarde les pièces de chacune des 12 premières marques et des autres au cours du temps [6] ;

— et en étape 3, les répartitions utilisées pour agréger les courbes des marques sont celles correspondant aux parts moyennes des marques par trimestre au cours du temps, et non plus des groupes. Ces parts sont données dans le tableau suivant :

Marque	%	Marque	%	Marque	%
Renault	18%	Mercedes	4%	Nissan	3%
Peugeot	14%	Ford	4%	Opel	2%
Citroën	11%	Audi	4%		
Volkswagen	6%	Fiat	4%		
Toyota	5%	BMW	3%	Autres	22%

On obtient alors la courbe en noir sur la *figure 3.19*.

6. sans prendre en compte les valeurs aberrantes

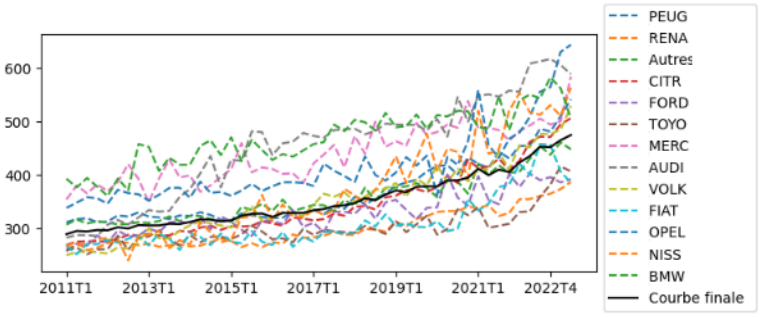


FIGURE 3.19 – Evolutions des marques obtenus en pointillés et courbe finale opérationnelle en noir

3.3.2 Délai de détection de la rupture

Avec cette courbe, se pose également la question suivante : à partir de quand arrive-t-on à détecter le point de rupture final, ici le T2 2022, qui indiquerait un changement de rupture, pour pouvoir l'utiliser opérationnellement ?

Pour cela, le modèle de segmentation binaire a été lancé sur la courbe finale réduite un temps avant, deux temps avant et trois temps avant. Les ruptures obtenues sont résumées dans le tableau suivant :

Courbe	Ruptures de la segmentation binaire avec pénalité 10
T1 2011 → T2 2023	2016T1, 2018T3, **2022T2**, 2023T2
T1 2011 → T1 2023	2016T1, 2018T3, **2022T2**, 2023T1
T1 2011 → T4 2022	2016T1, 2018T3, **2022T2**, 2022T4
T1 2011 → T3 2022	2016T1, 2018T3, 2022T3

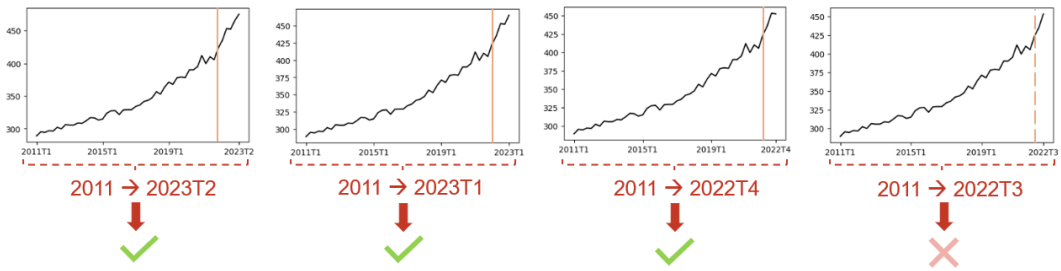


FIGURE 3.20 – Etude du délai de détection de la rupture en T2 2022

Ainsi, avec la courbe réduite à un temps avant, qui s'arrête au T1 2023, et avec la courbe réduite à deux temps avant, qui s'arrête au T4 2022, le modèle détecte bien la rupture au T2 2022.

Avant cela, avec trois temps avant, soit jusqu'au T3 2022, le modèle ne détecte pas de rupture en T2 2022. On s'y attendait puisqu'après le T2 2022 il n'y a plus qu'un seul point qui ne permet pas de définir une tendance.

3.4 Conclusion sur ce détecteur d'évolution importante du prix

Pour conclure sur cet outil de détection des ruptures sur la tendance des évolutions du coût moyen des pièces automobiles, on voit dans ces différents modèles qu'**au cours de l'année 2022, la courbe a changé de tendance**. On aurait pensé voir également l'année 2020 apparaître suite à une évolution des prix des pièces automobiles légèrement plus faible de la part des constructeurs liée au manque de ventes de véhicules, mais aucun point de rupture n'a été repéré pour cette période.

En effet, que ce soit avec la segmentation binaire ou avec le modèle *Pelt*, l'année 2022 a été repérée comme séparant les courbes de manière significative, c'est-à-dire que les segments ont des tendances suffisamment différentes. Ainsi grâce à ces modèles, on montre un **intérêt à mettre à jour les modèles de tarification** lorsqu'une rupture est repérée. Cela permettrait d'anticiper au mieux l'augmentation du coût de la pièce automobile dans l'année qui vient.

En cas d'alerte remontée par cet outil, cela permettrait d'anticiper le plus en amont possible les questions autour de l'hypothèse d'inflation qui sera injectée dans les modèles de tarification auto.

De plus, dans la partie opérationnelle, on a vu que la **rupture est détectée suffisamment rapidement** par les modèles. Dans le cas du deuxième trimestre 2022, ce point a été détecté comme rupture en utilisant les informations connues jusqu'au dernier trimestre 2022. Ainsi, si on avait lancé le modèle en début 2023, l'outil nous aurait déjà indiqué de faire attention puisque la tendance des coûts des pièces avait changé.

Chapitre 4

Identification des facteurs discriminants de l'inflation des pièces automobiles

Dans ce nouveau chapitre, l'objectif est de segmenter l'inflation à injecter dans les modèles de tarification pour permettre une meilleure précision de cette hypothèse. En effet, dans les modèles de tarification actuels, une hypothèse d'inflation uniforme à tous les contrats est réalisée. Ainsi, en distinguant cette inflation par marque par exemple, il sera possible d'ajuster cette hypothèse d'inflation en fonction des marques qui augmentent le plus le coût de leurs pièces détachées.

Pour ce faire, une modélisation de l'inflation est proposée dans cette partie, ainsi qu'une analyse du comportement des modèles pour identifier les facteurs discriminants de l'inflation.

Pour cela, les modèles qui seront appliqués par la suite vont être expliqués théoriquement. Ensuite, la seconde section se consacrera à la création de la base utile aux modèles. Enfin, le détail de l'application des modèles sera réalisé avec leur comparaison avant de conclure sur les variables discriminantes soulevées.

4.1 Explication théorique des modèles d'explication utilisés

Nous souhaitons utiliser des modèles statistiques pour identifier les facteurs discriminants de l'évolution du coût de la pièce parmi les caractéristiques du véhicule et les informations du marché.

L'objectif général des modèles est d'expliquer la variable réponse $Y = \{y_1, ..., y_n\}$ sur n individus en fonction de p variables explicatives $X_1, ..., X_p$ [1]. On peut représenter les données dans un tableau avec la première colonne correspondant à la variable à estimer et les colonnes suivantes étant les variables explicatives qui vont aider à l'estimation. On obtient alors les données suivantes :

individu 1 →	y_1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,p}$
individu 2 →	y_2	$x_{2,1}$	$x_{2,2}$	...	$x_{2,p}$
...	...	...	...	...	...
individu n →	y_n	$x_{n,1}$	$x_{n,2}$	...	$x_{n,p}$

On distingue la régression de la classification par la nature de la variable à expliquer Y. En effet, si la variable cible du modèle est numérique, c'est-à-dire à valeurs continues ou discrètes, alors le modèle sera une régression. Si en revanche la variable à expliquer est qualitative, alors la modélisation sera une classification supervisée.

Dans cette étude, nous avons commencé par sélectionner naturellement un GLM qui est facilement interprétable, pour ensuite le comparer avec des modèles habituellement plus performants mais aussi plus complexes à interpréter : l'arbre de décision, la forêt aléatoire et l'*histogram gradient boosting*. Pour contrer cette difficulté d'interprétation de ces modèles, l'utilisation d'un algorithme d'explicabilité éprouvé SHAP sera proposé.

4.1.1 GLM pénalisé

Le modèle de régression linéaire multiple

Ce modèle est défini, avec les $(W_1, ..., W_n)$ des variables aléatoires i.i.d telles que $E[W_i] = 0$ et $var(W_i) = \sigma^2 < \infty$, comme :

$$Y_i = \beta_0 + \sum_{k=1}^{p} \beta_i x_{i,k} + W_i \tag{4.1}$$

Dans ce cas, les variables explicatives sont déterministes et la variable à expliquer y_i est la réalisation de la variable Y_i.

GLM

Un modèle linéaire généralisé ou GLM est un modèle de régression qui généralise la régression linéaire [7, 8]. Dans ce cas, la variable à prédire est à valeurs continues dans $\mathbb{R}$.

Ce modèle est défini par 3 hypothèses essentielles :

1. tout d'abord les variables aléatoires $Y_1, ..., Y_n$ doivent être indépendantes pour les n individus ;
2. ensuite la loi de Y_i doit avoir une distribution connue dont les paramètres dépendent des variables explicatives ;
3. et on modélise la relation entre les variables grâce à une fonction de lien.

Dans notre cas, la variable cible est le montant de l'évolution calculée sur le coût des pièces automobiles entre deux semestres consécutifs, ainsi la variable réponse a des valeurs dans $\mathbb{R}$ avec de plus une distribution de loi normale. Cela revient alors à faire un modèle linéaire gaussien.

Le modèle se définit alors par :

$$g(E[Y_i]) = \beta_0 + \beta_1 x_{i,1} + ... + \beta_p x_{i,p} \tag{4.2}$$

avec Y_i suivant la loi $N(\mu(x_{i,1}, ..., x_{i,p}), \sigma^2)$ où $\mu(x_{i,1}, ..., x_{i,p})$ est une fonction linéaire des variables explicatives, $g(x) = x$ la fonction de lien associée à une distribution gaussienne.

GLM pénalisé

La méthode pénalisée ou sous contrainte a pour objectif de ressortir les variables les plus pertinentes à l'explication de la variable réponse pour pouvoir réduire le nombre de variables explicatives. Pour cela, une pénalité est ajoutée à la fonction à minimiser en ajoutant une condition sur les β du modèle linéaire.

Cette méthode est utilisée lorsque le modèle présente beaucoup de variables explicatives ou alors si l'on souhaite réduire le nombre de variables explicatives. Ainsi, cette réduction peut se faire en choisissant m variables parmi les p initiales d'après les résultats qui sont ressortis du modèle, ou alors en essayant de trouver la meilleure fonction f telle que $f(x) = y$ ait une meilleure approximation de la variable Y_i avec le moins de variables explicatives x.

Le *glm pénalisé*, à la différence d'un *glm* normal, vise à minimiser la fonction

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2 + \lambda \sum_{l=1}^{p} ||\beta_l|| \tag{4.3}$$

avec :

- la somme des écarts quadratiques entre les valeurs prédites et les valeurs observées $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2$,
- le paramètre de réglage λ
- la pénalité qui se définit comme étant :
  - la norme 2 pour la méthode **Ridge**, c'est-à-dire $\sum_{l=1}^{p} \beta_l^2$
  - la norme 1 pour la méthode **Lasso**, c'est-à-dire $\sum_{l=1}^{p} |\beta_l|$

  Dans ce modèle, les coefficients β_k sont testés pour voir quelles variables seront conservées. Ainsi, la pénalité appliquée pour les modèles Lasso est plus restrictive que pour les modèles Ridge du fait de la valeur absolue par rapport au carré. En effet, les coefficients auront tendance à plus facilement tendre vers zéro avec la méthode Ridge alors qu'avec Lasso les coefficients proches de zéro prendront directement cette valeur.

Pourquoi utiliser un *glm pénalisé* à la place d'un *glm* classique ?

Suite à l'apprentissage des modèles de *Machine Learning*, un surapprentissage ou un sous-apprentissage peuvent apparaître. Les erreurs d'un modèle se décomposent respectivement en une erreur de biais et une erreur de variance.

Dans les modèles GLM classiques, seule l'erreur de biais peut être réduite alors que les modèles pénalisés réduisent les deux erreurs à la fois (voir *figure 4.1*[1]). En plus de cela, la méthode pénalisée permet d'arbitrer plus rapidement sur les variables discriminantes de la variable cible.

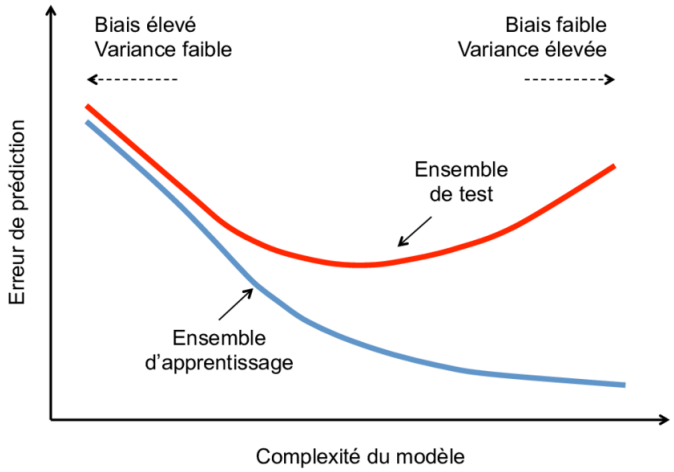


FIGURE 4.1 – Illustration du dilemme biais-variance dans les modèles de Machine Learning *(Source : Céline ROUARD)*

4.1.2 Arbre de décision CART

Un arbre de décision, autrement connu sous l'acronyme de CART (*classification and regression tree*), est un autre type de modèle permettant d'estimer la variable cible en partitionnant l'espace des valeurs des variables explicatives en rectangles de telle sorte que la variable cible soit constante [33, 28]. Dans ce cas nous n'avons plus la linéarité du modèle mais l'interprétabilité est simplifiée ce qui lui donne un aspect très apprécié pour la prise de décision. C'est un modèle qui fonctionne tant en régression qu'en classification.

Description d'un arbre

Un arbre est composé de plusieurs parties, de noeuds et de branches représentés de façon hiérarchique. Un noeud correspond à un endroit où l'arbre a subi une séparation en deux parties appelées des branches. Ainsi, un arbre commence par un noeud dit "racine" qui le sépare en deux branches, finit par des noeuds dits "finaux" qui concluent l'arbre, et entre les deux il y a des noeuds internes. Les noeuds terminaux sont alors appelés les feuilles de l'arbre.

1. Thèse : "Inférence de réseaux d'interaction protéine-protéine par apprentissage statistique", 14 février 2013

Fonctionnement de l'algorithme CART

Mathématiquement, posons le vecteur aléatoire (X, Y) à valeurs dans $\mathbb{R}^p \times \mathbb{R}$ s'il s'agit d'une régression ou $\mathbb{R}^p \times \{1, ..., K\}$ si c'est une classification. On suppose connu l'ensemble d'apprentissage $\{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$ ou $\mathbb{R}^p \times \{1, ..., K\}, i \in 1, ..., n\}$. Le noeud racine est appelé n_1 et contient toutes les observations de l'ensemble d'apprentissage.

La première étape du modèle CART sépare cet ensemble en deux sous-ensembles $\{X^j \leq s\}$ et $\{X^j > s\}$ où $j \in \{1, ..., p\}$, $X = \{X^1, ..., X^p\}$ et $s \in \mathbb{R}$ en recherchant la coupure optimale de s sur la variable j. Cela signifie que toutes les observations avec une valeur de la variable j plus petite que s sont dans le premier ensemble, sinon dans le second. Pour choisir la variable j à utiliser pour la coupure, chaque variable est étudiée pour séparer la variable cible la plus pertinente possible au sens de la prédiction.

En régression, ce critère d'erreur peut être de minimiser la MSE sur chacun des deux nouveaux ensembles créés.

Definition 15 (Mean Square Error)

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (\hat{y_i} - y_i)^2 \tag{4.4}$$

avec $\hat{y_i}$ les prédictions obtenues pour les y_i.

En classification cela peut être l'indice de Gini qui est un indice d'impureté. C'est-à-dire qu'il indique si les échantillons des sous-ensembles envisagés sont similaires (pures) ou éloignés (impures). Il est proche de 0 lorsque les échantillons sont composés d'individus similaires et proche de 1 si cela n'est pas le cas.

Definition 16 (Indice de Gini)

$$G_i = 1 - \sum_{k=1}^{n} p_{i,k}^2 \tag{4.5}$$

avec $p_{i,k}^2$ le ratio du nombre d'individus de la classe k dans le sous-échantillon du noeud i

Par exemple, en prenant ces arbres fictifs, si l'objectif est d'estimer l'inflation du coût de la pièce automobile (*figure 4.2*), alors dans le cas du problème de régression, la variable cible prend des valeurs numériques et ainsi l'arbre mènera à des valeurs numériques. S'il s'agit d'une classification, alors les réponses possibles du modèle sont les différentes modalités de la variable à expliquer comme ici les classes 1, 2 ou 3. Les feuilles correspondent alors aux classes majoritaires de la variable à estimer des observations présentes dans le noeud terminal.

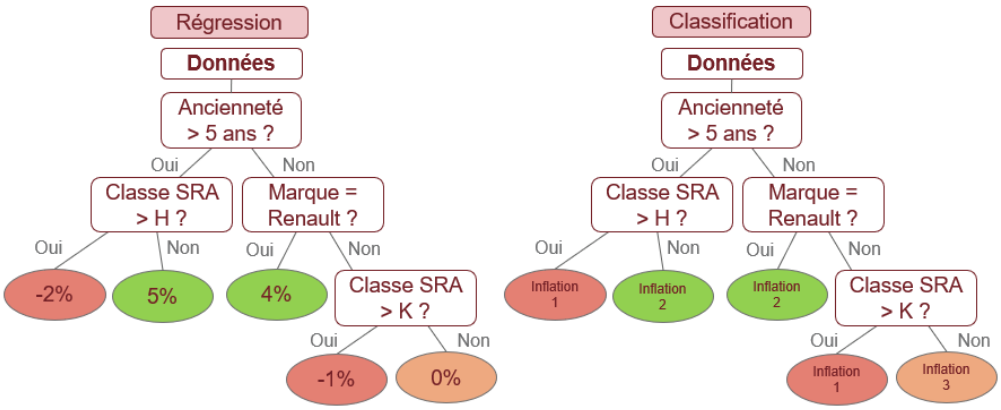


FIGURE 4.2 – Exemple fictif d'arbres de régression et de classification

Avantages et inconvénients de cette méthode

Comme vu précédemment, cette méthode est très facile à interpréter grâce au graphique de l'arbre qui permet de visualiser aisément ce qui se passe dans le modèle. Cependant ce modèle n'est pas très robuste puisqu'il est très dépendant des données de la base d'apprentissage. On peut aussi évoquer que les variables qui servent aux séparations peuvent apparaître dans plusieurs chemins de l'arbre.

Ces arbres de décisions sont plus souvent utilisés comme des étapes pour des méthodes plus sophistiquées comme la *forêt aléatoire* ou encore l'*histogramme gradient boosting*.

4.1.3 Forêt aléatoire

La *forêt aléatoire* est un modèle d'apprentissage qui consiste à reproduire un grand nombre de fois un arbre de décision CART afin de créer un modèle final grâce à la méthode du *bagging* [4]. L'objectif est de rendre moins corrélées les modèles de base en ajoutant de l'aléa dans la sélection des variables des différents modèles et ainsi rendre le modèle plus performant tout en diminuant le risque de variance.

La méthode du *bagging*

Comme le *boosting* et le *stacking*, le *bagging* est une méthode d'*ensemble learning* qui combine plusieurs méthodes de *machine learning* pour améliorer les performances de prédiction.

Le *bagging*, qui vient de la combinaison entre le *boosting* et l'*aggregating*, consiste à sélectionner aléatoirement m variables parmi les p variables explicatives à chaque étape de l'arbre et de déterminer la coupure correspondante en prenant en compte uniquement ces variables. Pour que chaque modèle s'entraîne sur un jeu de données différent, les données utilisées sont alors toujours revues par tirage aléatoire avec remise. Grâce à cela,

le modèle est moins dépendant des données d'apprentissage par rapport à l'arbre CART classique réduisant ainsi l'erreur de variance.

Contrairement au *boosting* qui construit les modèles de manière itérative afin de corriger progressivement les erreurs, le *bagging* consiste à créer plusieurs modèles indépendamment pour retenir le vote majoritaire ou la valeur moyenne des estimations (voir *figure 4.4*).

L'algorithme de la forêt aléatoire

Posons $L_n = \{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}\}$ l'ensemble d'apprentissage s'il s'agit d'une régression ou $L_n = \{(x_i, y_i) \in \mathbb{R}^p \times \{1, ..., K\}\}$ si c'est une classification.

Schématiquement, on peut le représenter comme dans l'*image 4.3*.

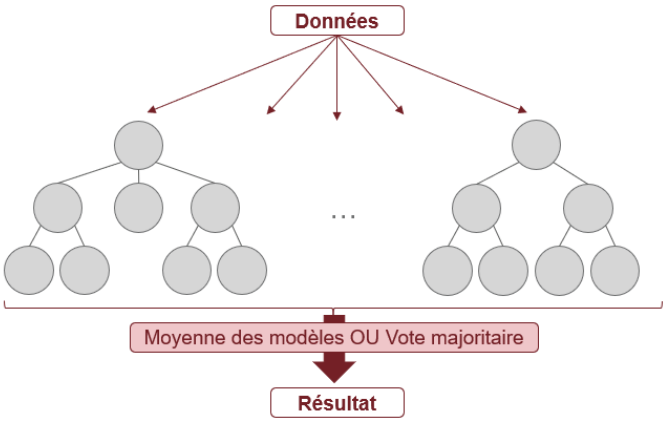


FIGURE 4.3 – Mécanisme d'une forêt aléatoire

La forêt aléatoire se construit en faisant les étapes suivantes :

1. Choisir le nombre de $B \in \mathbb{N}^*$ échantillon *bootstrap* et du nombre $m \in \mathbb{N}^*$ de variables à considérer parmi les p variables explicatives. En général, $m = \sqrt{p}$ en régression et $m = \frac{p}{3}$ en classification.
2. Pour b allant de 1 à B, on commence par tirer un sous-échantillon $L_{n,b}$ dans L_n pour construire un arbre CART sur cet échantillon. Les coupures de l'arbre sont réalisées sur les m variables explicatives tirées indépendamment et au hasard parmi les p variables. On obtient un prédicteur $\hat{\phi}_b(.)$.
3. Le prédicteur final est obtenu en faisant :
   — $\hat{\phi}(.) = \frac{1}{B} \sum_{b=1}^{B} \hat{\phi}_b(.)$ en régression, c'est-à-dire en prenant la valeur moyenne des estimations
   — $\hat{\phi}(.) = argmax_j \; card\{b : \hat{\phi}_b(.) = j\}$ en classification, cela revient à considérer le vote majoritaire sur les B prédictions.

Pour estimer les erreurs des modèles, des individus sont isolés au hasard dans la phase d'apprentissage pour pouvoir valider le modèle ensuite.

Ce type d'erreur est appelé l'erreur *out of bag* et revient à considérer l'erreur moyenne sur chaque échantillon d'apprentissage ech_i en utilisant uniquement les arbres qui n'avaient pas cet échantillon ech_i dans leur échantillon *bootstrap*. Cet algorithme est très utile lorsque le nombre de variables explicatives p est important.

4.1.4 Histogram Gradient Boosting

L'*histogram gradient boosting* consiste lui aussi à reproduire un grand nombre de fois un arbre de décision CART mais en utilisant la méthode du *boosting* [12, 31].

La méthode du *boosting*

Comme vu précédemment, le *boosting* est une méthode d'apprentissage ensembliste qui construit des arbres de façon itérative pour apprendre de ses erreurs.

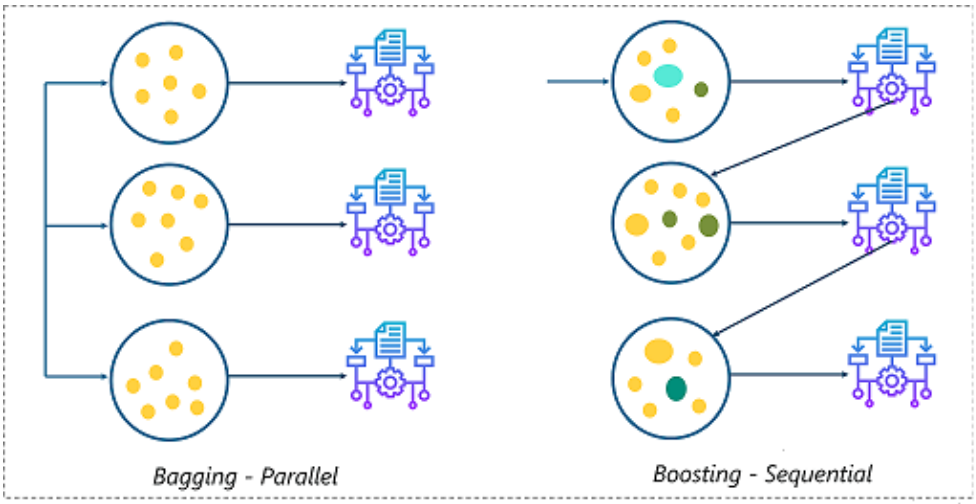


FIGURE 4.4 – Différence entre le *boosting* et le *bagging* (Source : *Edureka*)

Dans ce cas, on définit un apprenant faible comme étant un algorithme qui donne des classifieurs faibles capables de différencier deux classes au moins aussi bien que le hasard et un apprenant fort un algorithme qui fournit des classifieurs performants dans la distinction des classes. Comme l'idée initiale est d'améliorer le modèle, alors à chaque étape les poids des classifieurs sont augmentés si la performance est meilleure, inchangé dans l'autre cas.

Ainsi, grâce à ces pondérations, la méthode de *boosting* identifie les algorithmes à *"booster"* au tour suivant pour améliorer les prévisions du premier arbre (voir *figure 4.4*).

Le gradient *boosting*

Les modèles de *boosting*, comme l'*Adaboost*, sont des algorithmes basés sur une fonction de perte, construits de manière séquentielle afin d'améliorer les performances à chaque pas. Ainsi, les résultats cibles de chacun des cas sont basés sur l'erreur du gradient par rapport aux prévisions, c'est le *gradient boosting*. Cela revient à chercher le meilleur pas de la descente du gradient permettant d'obtenir les meilleures performances pour l'algorithme.

L'histogramme gradient boosting

Comme son nom l'indique, l'*histogram gradient boosting* [31] a le même principe que le *gradient boosting* classique mais avec pour particularité d'être plus rapide grâce à la discrétisation des variables sous forme d'histogrammes. Les différentes valeurs des variables sont regroupées en compartiments comme pour l'exemple en *figure 4.5*.

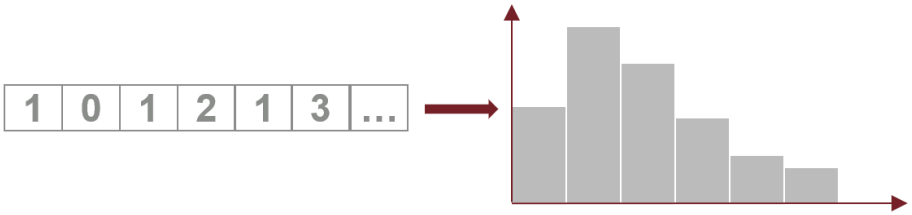


FIGURE 4.5 – Exemple fictif de discrétisation d'une variable en histogramme

Ainsi, au lieu de trouver les points de division sur les valeurs de caractéristiques triées, cet algorithme utilise les groupes de caractéristiques de l'histogramme.

Ce modèle va alors minimiser l'erreur de prédiction du modèle lorsqu'il est additionné avec les modèles précédents. C'est pour cela qu'on parle de renforcement du gradient puisque chaque nouveau modèle réduit l'erreur de prédiction totale.

En plus de réduire le temps de calcul, ce modèle a pour second avantage de ne pas impacter les performances des modèles.

4.1.5 Métriques de comparaison des modèles

En considérant y_i et $\hat{y}_i$ pour $i \in \{1, ...n\}$ les valeurs observées et les valeurs prédites de la variable cible respectivement, $\overline{y}$ la moyenne des valeurs observées de y, n le nombre de valeurs de l'échantillon, TP le nombre de vrais positifs, TN le nombre de vrais négatifs, FP le nombre de faux positifs et FN les faux négatifs, nous obtenons un ensemble de métriques qui sont définies ainsi :

Régression	Formule	Classification	Formule
RMSE *root mean square error*	$\sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$	Accuracy (*bonnes prédictions*)	$\frac{TP+TN}{TP+FN+TN+FP}$
MAE *mean absolute error*	$\sum_{i=0}^{n-1} \frac{\lvert y_i - \hat{y}_i \rvert}{n}$	Précision	$\frac{TP}{TP+FP}$
Variance expliquée	$1 - \frac{Var(y-\hat{y})}{Var(y)}$	Rappel *recall*	$\frac{TP}{TP+FN}$
R^2	$1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$	F1 *(déséquilibres)*	$2 \times \frac{precision \times recall}{precision + recall} = \frac{TP}{TP + 0{,}5 \times (FN+FP)}$

En général en régression, la métrique utilisée pour constater de la performance du modèle est le $\underline{R^2}$, autrement appelé coefficient de détermination. Cet indicateur permet de comparer les erreurs quadratiques moyennes obtenues avec le modèle et celles obtenues avec un modèle basique en considérant la moyenne des réponses. Ainsi, lorsque le R^2 est proche de 1, cela signifie que les prédictions du modèle sont exactes et que le modèle est parfait. Inversement lorsqu'il est proche de 0 le modèle n'est pas très bon puisque les prédictions sont éloignées des valeurs réelles, on peut aussi dire qu'il est "aussi bon que l'aléatoire". Dans certains cas, le R^2 peut être négatif et cela signifie que les prédictions du modèle sont moins bonnes que si l'on prédisait toujours la valeur moyenne.

Cette métrique est souvent regardée avec la $\underline{RMSE}$ ou la $\underline{MAE}$ qui permettent d'indiquer l'erreur moyenne du modèle.

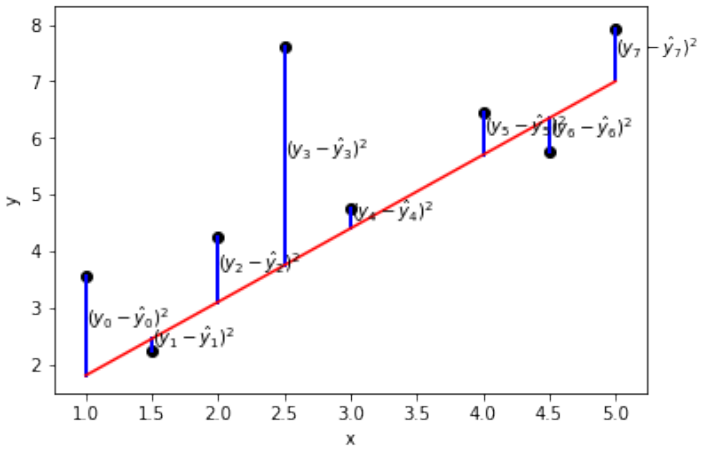


FIGURE 4.6 – Schéma des erreurs de prédictions (écarts entre $y_i - \hat{y}_i$)

En classification, les quatre métriques précédentes se regardent en parallèle fournissant chacune une information intéressante.

1. l'*accuracy* indique le taux de prédictions correctes. Il faut cependant faire attention car c'est une métrique qui peut être mal interprétée, notamment dans le cas d'une classification très déséquilibrée. C'est pourquoi on regarde plutôt les 3 métriques suivantes.
2. la précision donne un pourcentage sur le nombre de prédictions positives correctes. C'est à dire que quand la précision est haute (proche de 1), alors la majorité des prédictions positives du modèle sont des positifs bien prédits.
3. le rappel se demande si le modèle a réussi à capter tous les individus positifs, il communique une bonne information sur les individus appartenant aux classes positives mais ne donne aucune information sur les négatifs.
4. en synthèse de ces deux dernières métriques, le F1 score est une moyenne harmonique entre la précision et le rappel, il est favorisé dans le cas de classes fortement déséquilibrées.

4.1.6 Méthode SHAP pour analyser les variables importantes des modèles de *Machine Learning*

Pour les modèles de *machine learning* qui sont difficilement interprétables comme l'*histogramme gradient boosting* ou la *forêt aléatoire*, le package SHAP développé sur *python* permet de rendre plus transparent les résultats de tout modèle d'apprentissage automatique.

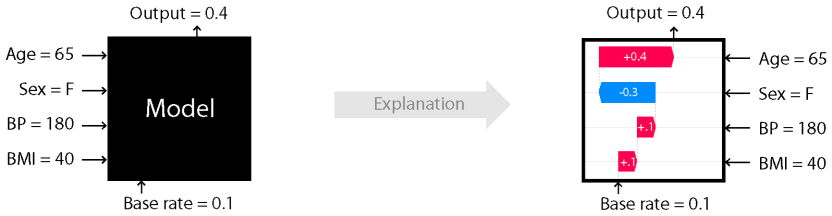


FIGURE 4.7 – Représentation du principe de la méthode SHAP [2]

En effet, comme ce type de modèle est difficile à décrypter, le package SHAP, qui vient de *SHapley Additive exPlanations*, permet de donner une valeur de Shapley à chaque caractéristique d'entrée pour visualiser son impact sur la variable cible [32]. Il permet de quantifier l'impact de chaque variable ou modalité de variable en fonction de l'encodage utilisé sur la variable à expliquer. Concrètement, ces *Shapley values* sont calculées pour chaque combinaison de variables. Pour que ces combinaisons soient ensuite résumées grâce à une moyenne.

2. Source : https://github.com/slundberg/shap)

Cette approche provient de la théorie des jeux qui consiste à répartir équitablement les gains d'un jeu aux différents participants. Le principe est de comprendre les interactions entre les individus pour optimiser le gain.

4.2 Application des modèles

4.2.1 Construction de la base de modélisation

Dans les rapports d'expertise nous avons de nombreuses informations sur la voiture, le sinistre et la réparation, qui nous ont déjà servi à :

- constituer nos périmètres d'études sur les rapports étudiés comme la nature du sinistre, le type de véhicule, etc., mais aussi sur les pièces étudiées avec l'opération réalisée sur la pièce, le montant de la pièce, le libellé de la pièce etc. ;
- récolter des informations sur les pièces comme le montant unitaire et sur les véhicules sinistrés avec la marque, le modèle, le code SRA [3], etc.

Afin de comprendre quels paramètres influencent le plus le prix de la pièce et en particulier l'augmentent, nous avons voulu enrichir les données avec des informations externes. Ainsi, une récolte d'informations supplémentaires a eu lieu pour récupérer des informations plus poussées sur le véhicule grâce au code SRA et des informations sur le marché.

Pour mener à la base de modélisation, 7 étapes ont été réalisées :

1. ajout d'informations externes aux données d'apprentissage ;
2. problématique du nombre de modalités utilisées pour la construction des variables d'agrégation des données ;
3. définition de la variable d'exposition ;
4. définition de la variable cible ;
5. traitement des valeurs extrêmes ;
6. stratégie mise en place pour tester les modèles, avec :
   - une construction des données qui dépend directement du nombre de modalités choisi pour les variables d'agrégation ;
   - et des tests de différents types sur les modèles ;
7. conclusion sur les données de modélisation finalement retenues ;

Après la définition de cette table, une étape d'analyse des performances et des comportements des modèles permettra de conclure ce chapitre.

3. Ce code est une chaîne de 7 caractères qui indique la marque du véhicule avec les deux premières lettres, le modèle avec les deux chiffres suivants puis la version, finition, puissance du véhicule avec les 3 derniers. Par exemple le code FI49037 désigne une Fiat (FI) 500L (49) 1.6 MJT 120 Trekking (037).

Etape 1 : Récupérations d'informations externes supplémentaires

- Informations véhicules par le SRA : dans un premier temps, une base de données véhicules est fournie par l'association SRA [4]. On y trouve, associés au **code SRA** du véhicule, la marque, le modèle, la version, la puissance, l'alimentation, la classe de prix, le nombre de portes du véhicule, etc.

Grâce à ce code SRA qui est renseigné dans les rapports d'expertise ou dans les sinistres de Generali, il est possible de faire un raccordement pour obtenir des informations sur le véhicule sinistré et donc assuré.

Dans les rapports d'expertise et les bases sinistres, prêts de **72,07% des rapports ont un code SRA renseigné cohérent** avec la marque du véhicule sinistré. Dans les autres cas, soit le code était mal renseigné - "999999", "*", " " - soit la marque ne correspondait pas aux deux premières lettres du code SRA. Pour la suite du mémoire, on se retrouve avec une base d'étude de 622 267 sinistres. Il y a alors 112 823 pièces étudiées qui ont des informations SRA.

- Informations du marché : dans un second temps, les variables macro-économiques, comme le prix de l'essence à la date du rapport, l'indice du prix de l'acier, etc., ont été récupérées sur le site de l'INSEE (voir la *partie 2.2* qui présente les données externes).

Ces informations vont être associées au rapport d'expertise en fonction de la date du rapport et de la granularité de la segmentation choisie. En effet, si la granularité est le semestre, alors une moyenne des mois constituants ce semestre permet de donner une valeur de marché. C'est la même chose pour les autres temporalités.

Les **valeurs "brutes"** ont été considérées dans les variables explicatives mais aussi les **évolutions** de ces indices de marché calculées avec la même formule de calcul d'une évolution. Pour l'indice i à l'instant t, l'évolution de l'indice est :

$$evolution_{i,t} = (\frac{CM_t}{CM_{t-1}} - 1) \times 100 \text{ pour l'indice du marché à l'instant t} \tag{4.6}$$

De plus, pour prendre en compte le potentiel **temps de propagation** des inflations des prix du marché sur les prix des pièces automobiles, les valeurs brutes et évolutions des prix ont été ajoutées avec des **décalages de un, deux, ou trois unités** de la segmentation temporelle.

C'est-à-dire que pour une pièce, on compare l'évolution de la pièce au temps t avec l'évolution du coût des éléments de marché au temps t, t-1, t-2 et t-3 pour prendre en compte un temps de diffusion potentiel de l'inflation sur le coût des pièces. Si l'unité de temps est semestrielle alors t-1 correspond au semestre d'avant, t-2 au semestre de l'année d'avant et t-3 au semestre 1 an et demi avant.

4. Informations : https://www.index-assurance.fr/pratique/devis-souscription/code-sra-dune-voiture-avec-groupe-classe-cnit-ou-type-mines

Dans le tableau suivant, nous avons les premières lignes des valeurs d'évolutions associées au prix horaire de la main d'oeuvre manufacturière (base 2016) en fonction du décalage considéré :

Année Semestre	Décalage Coût de la MO	t 0	t-1 1	t-2 2	t-3 3
2011 S1	93.00	0%	0%	0%	0%
2011 S2	94.05	1,17%	0%	0%	0%
2012 S1	95.35	1.45%	1,17%	0%	0%
2012 S2	96.65	1.31%	1.45%	1,17%	0%
2013 S1	96.00	-0.73%	1.31%	1.45%	1,17%
2013 S2	97.40	1.53%	-0.73%	1.31%	1.45%
2014 S1	97.20	-0.22%	1.53%	-0.73%	1.31%
2014 S2	98.20	1.07%	-0.22%	1.53%	- 0.73%

Etape 2 : Modalités des variables d'agrégation

On souhaite calculer l'inflation du coût de la pièce en fonction de certaines situations de risque qui dépendent de caractéristiques du véhicule. Cependant ces variables que sont la pièce, la marque, la classe SRA et le groupe SRA peuvent avoir de nombreuses modalités. Pour permettre de trouver des pièces dans le plus de cas possible et sur deux temporalités consécutives, des regroupements de modalités ont été proposés avec le principe **des regroupements par centiles** pour avoir des groupes homogènes en volume.

Definition 17 (Branche de risque) On définit une branche ou situation de risque comme étant une combinaison des modalités des variables d'agrégation. Si ces variables d'agrégation sont la marque, l'ancienneté, la classe SRA, le groupe SRA et la temporalité, alors elles sont de la forme "marque × ancienneté × classe SRA × groupe SRA × Année Semestre" avec chaque variable qui prend une modalité différente. Prenons 3 exemples de branches de risques aléatoires pour ces variables d'agrégation :

	Marques	Ancienneté	Classe SRA	Groupe SRA	Temporalité
Exemple 1	Audi	Modalité 1	Modalité 1	Modalité 1	2011 S1
Exemple 2	Renault	Modalité 6	Modalité 4	Modalité 2	2021 S2
Exemple 3	Autres	Modalité 5	Modalité 3	Modalité 3	2016 S2
Nombre total	13	6	6	6	23

- Construction des groupes de modalités des variables d'agrégation : plusieurs regroupements en 3, 4, 5 ou plus de modalités par variable ont été testés pour éviter d'avoir des branches de risques trop précises qui correspondraient à des volumes de pièces détachées trop faibles mais également pour ne pas perdre en capacité de différencier

l'inflation. Pour la coupure en 5 groupes, les centiles à 20%, 40%, 60% et 80% ont été considérés pour réaliser les séparations, pour 4 groupes, les centiles à 25%, 50% et 75%, et pour les 3 groupes ceux à 33% et 66% par exemple.

Pour chaque variable, la segmentation était alors la suivante :

1. Pour la variable des pièces, seules les pièces sélectionnées en *partie 2.1.5* ont été retenues, c'est-à-dire le "bouclier avant", le "bouclier arrière", l'"aile avant gauche", la "porte avant" et la "porte arrière".

2. Pour considérer l'ancienneté du véhicule, l'âge à la date d'émission du rapport d'expertise a été favorisé pour mieux répondre à la problématique de l'étude qui consiste à suivre le coût de la pièce constructeur dans le temps, soit à la date du rapport. Cette ancienneté est très proche de celle à la date du sinistre, souvent moins d'un an, sauf dans quelques cas où le rapport a été mis à jour et remplacé.

   On a alors obtenu les différentes segmentations suivantes pour les anciennetés avec 3, 4 et 5 modalités par exemple :

Nombre de modalités	Modalité n°1	Modalité n°2	Modalité n°3	Modalité n°4	Modalité n°5
3	≤ 3 ans	$\in]3, 7]$	> 7		
4	≤ 2 ans	$\in]2, 5]$	$\in]5, 8]$	> 8	
5	≤ 2 ans	$\in]2, 4]$	$\in]4, 6]$	$\in]6, 9]$	> 9

3. Pour la classe SRA, deux codifications étaient possibles, une première qui donne la classe de prix du véhicule à sa sortie et une deuxième pour la classe de prix des réparations des pièces du véhicule à sa sortie [5]. Même si la classe de prix des réparations des pièces semblait la plus adaptée à notre étude, la classe qui a été retenue est celle qui donne la classe de prix en général, parce qu'elle possède moins de véhicules en catégorie "hors classe" (lignes retirées des segmentations).

   Ces classes vont de A à ZA de telle sorte que les lettres proches de A ont une faible valeur de prix et celles proches de ZA ont une forte valeur de prix.

   Comme les différentes marques font partie de gammes de véhicules différentes, une **normalisation par marque** a été appliquée pour éviter par exemple que toutes les pièces de véhicules BMW soient dans les fortes valeurs de prix. Par exemple pour Renault, BMW et Audi nous obtenons ces segmentations pour ces mêmes 3 exemples :

5. Ces deux codifications sont proposées par le SRA. La classe de prix est la valeur à neuf TTC du véhicule, hors option et remise.

Marque	Nombre de modalités	Modalité n°1	Modalité n°2	Modalité n°3	Modalité n°4	Modalité n°5
Renault	3	A → H	I → J	K → S		
	4	A → G	H → I	J → K	L → S	
	5	A → F	G → H	I → J	K → L	M → S
BMW	3	I → N	O → Q	R → X		
	4	I → M	N → O	P → Q	R → X	
	5	I → M	N → O	P	Q → R	S → X
Audi	3	H → M	N → P	Q → X		
	4	H → M	N	O → Q	R → X	
	5	H → L	M → N	O	P → Q	R → X

4. Pour le groupe SRA qui donne la puissance du véhicule à l'origine, de même que pour la classe, une **normalisation par marque** a été faite pour prendre en compte les gammes des marques. Dans ce cas, on a :

Marque	Nombre de modalités	Modalité n°1	Modalité n°2	Modalité n°3	Modalité n°4	Modalité n°5
Renault	3	≤ 29	$\in]29, 31]$	> 31		
	4	≤ 29	$\in]29, 30]$	$\in]30, 31]$	> 31	
	5	≤ 28	$\in]28, 29]$	$\in]29, 31]$	$\in]31, 32]$	> 32
BMW	3	≤ 34	$\in]34, 36]$	> 36		
	4	≤ 34	$\in]34, 35]$	$\in]35, 36]$	> 36	
	5	≤ 33	$\in]33, 34]$	$\in]34, 36]$	$\in]36, 37]$	> 37
Audi	3	≤ 33	$\in]33, 35]$	> 35		
	4	≤ 33	$\in]33, 34]$	$\in]34, 36]$	> 36	
	5	≤ 32	$\in]32, 33]$	$\in]33, 35]$	$\in]35, 37]$	> 37

5. Enfin, pour la temporalité, les tests de volumes ont été réalisés sur des moyennes du prix de la pièce associées de façon trimestrielle, semestrielle ou annuelle. En effet, pour le reste du mémoire la maille trimestrielle est favorisée par rapport à la volonté d'avoir un suivi raffiné de l'inflation. Dans ce cas, on souhaite comprendre les facteurs discriminants de l'inflation du coût de la pièce. Pour s'assurer d'avoir des montants d'évolutions du coût de la pièce, il est important d'avoir deux valeurs de prix consécutifs.

   De cette façon, comme pour chaque semestre le volume de pièces est plus important que pour chaque trimestre ou mois, en agrégeant les données sur des temporalités plus larges, nous pouvons obtenir plus de volumes.

FIGURE 4.8 – Représentation de deux branches de risques grâce auxquelles on peut calculer une évolution du coût de la pièce

- Problématique du nombre de modalités des variables d'agrégation avec ces variables, il est alors possible d'agréger les pièces par situation de risque. Les variables d'agrégation choisies sont la marque, l'ancienneté du véhicule à la date du rapport, la classe SRA, le groupe SRA et la temporalité.

Cependant, d'autres décisions sur la structure de ces données auront des conséquences sur l'apprentissage des modèles :

1. Tout d'abord, en fonction de l'agrégation, le nombre de lignes obtenues va être modifié. Ainsi, ce premier critère va dépendre directement du nombre de modalités par variable et du pas temporel retenu ;
2. Ensuite, c'est la diversité des branches de risque observées par rapport à la combinatoire théorique. En effet, si le nombre de modalités par variable est élevé, alors les branches de risques sont plus fines pour permettre de mieux modéliser l'inflation, mais cette finesse a pour conséquence de ne pas avoir des pièces pour chaque combinatoire ;
3. Pour finir, les combinatoires doivent être retrouvées pour deux temporalités successives pour pouvoir calculer une inflation, ainsi, avec une grande finesse dans les branches de risque fait diminuer le nombre de pièces par branche de risque consécutives.

Tout cela revient à faire un compromis entre la finesse et la précision. En effet, plus le nombre de modalités est élevé mais en étant fin pour décrire l'inflation, moins il y a de pièce par branche de risque pour calculer une inflation. A l'inverse, plus le nombre de modalités est faible, plus on a un volume important de pièces par agrégation et plus on aura de difficultés à expliquer l'inflation.

Toutes ces remarques nécessitent de réaliser des simulations pour appréhender les impacts en termes de données.

Le *tableau 4.9* fait guise de récapitulatif sur les volumes obtenus avec quelques-unes des combinatoires. Il prend en compte les modalités des variables citées précédemment [6].

6. sans prendre en compte les modalités "hors classe" pour les variables classe et groupe SRA

combinatoire : **sans les modalités HC**

	pièce	marque	ancienneté	classe sra	groupe sra	*Nombre de pas temporels - 1*	Max lignes théoriques	Nombre lignes obtenues	% obt / théo	Inflations successives et même branche	%/théo	%/obt
trimestre	5	13	5	5	5	47	381 875	62 672	16%	**32 160**	8%	**51%**
semestre						23	186 875	46 882	25%	**28 744**	15%	**61%**
annee						11	89 375	32 416	36%	**20 771**	23%	**64%**
trimestre	5	13	4	4	4	47	195 520	50 204	26%	**30 203**	15%	**60%**
semestre						23	95 680	35 137	37%	**24 298**	25%	**69%**
annee						11	45 760	22 793	50%	**16 235**	35%	**71%**
trimestre	5	13	4	4	3	47	146 640	45 802	31%	**29 130**	20%	**64%**
semestre						23	71 760	31 187	43%	**22 647**	32%	**73%**
annee						11	34 320	19 623	57%	**14 545**	42%	**74%**
trimestre	5	13	4	3	3	47	109 980	40 663	37%	**27 439**	25%	**67%**
semestre						23	53 820	26 901	**50%**	**20 302**	38%	**75%**
annee						11	25 740	16 517	64%	**12 656**	49%	**77%**

FIGURE 4.9 – Volumes du nombre d'inflations obtenues pour quelques combinatoires

Interprétation du tableau 4.9

Sur ce tableau, une ligne correspond à une combinatoire du nombre de modalités par variables utilisées. Par exemple, sur la première ligne du deuxième bloc, avec 5 pièces détachées, 13 marques, 4 groupes d'ancienneté, 4 groupes de classes SRA et 4 groupes de groupes SRA sur les 48 trimestres on peut avoir au maximum $5 \times 13 \times 4 \times 4 \times 4 \times (48-1) = 195.520$ combinaisons correspondant aux branches de risques. On perd un trimestre en calculant l'évolution par définition du calcul (*équation 4.7*).

Dans ce cas, $\frac{50204}{195520} = 26\%$ des combinaisons théoriques ont au moins une pièce. Sur ces 26% de pièces, seules 60% sont des pièces sur lesquelles l'évolution a été calculée avec la même branche de risque et pour des temps consécutifs. Finalement, sur toutes les pièces que nous avions par combinaisons initialement, nous nous retrouvons avec seulement 60% des lignes qui sont exploitables, soit 15% des branches théoriques maximales.

Par conséquent, pour essayer d'expliquer au maximum l'inflation et avoir suffisamment de volumes de données, une agrégation semestrielle avec 4 à 8 modalités pour les autres variables d'agrégation a semblé correspondre à un meilleur compromis.

Etape 3 : Construction de la variable cible

Dans les différents modèles utilisés, comme l'arbre de décision par exemple, il y a une possibilité de renseigner une exposition pour définir les individus auxquels on attribue plus d'importance lors de l'apprentissage.

La variable cible des modèles est l'**évolution du coût de la pièce automobile en fonction de la branche de risque considérée**. Cette évolution se calcule, pour une

pièce de situation de risque i à l'instant t :

$$evolution_{i,t} = (\frac{CM_{i,t}}{CM_{i,t-1}} - 1) \times 100 \tag{4.7}$$

Il est alors important d'avoir **deux coûts consécutifs** pour obtenir une valeur.

Ainsi, une fois les agrégations par situation de risque finalisées, on a vérifié que cette branche de risque était la même sur deux temps qui se suivaient, avant de calculer les évolutions du prix des pièces.

Dans ce cas, la variable cible pour la régression est la variable numérique obtenue. Dans un second temps, des classifications ont été testées avec comme variable cible l'évolution du coût de la pièce calculée avec ces valeurs classées dans 4, 6 ou 10 groupes, de telle sorte qu'elle soit qualitative.

Etape 4 : Construction de la variable exposition

Pour calculer l'évolution du coût de la pièce automobile, nous avons besoin d'avoir deux prix consécutifs concernant le même périmètre (voir la *formule 4.7*). Ainsi, pour prendre en compte le nombre de pièces utilisées pour faire le calcul de l'évolution du coût de la pièce, une variable exposition a été créée telle que, pour la pièce de situation de risque i à l'intant t :

$$Exposition_{i,t} = min(Nombre\ pieces_{i,t}\ ; Nombre\ pieces_{i,t-1}) \tag{4.8}$$

Cela signifie que, par rapport à la construction de la variable cible qui prend en compte le coût moyen de la pièce en i et en $i-1$, on considère pour l'exposition le **minimum entre le nombre de pièces utilisées en i et celui des pièces en $i-1$**. Cela signifie que plus le nombre de pièces est élevé sur une branche de risque, plus le modèle accorde de l'importance celle-ci lors de l'apprentissage.

Etape 5 : Agrégation des pièces et traitement des données de la base pour les modèles d'explication

Après avoir défini ces situations de risque en agrégeant les pièces, différents traitements plus ou moins restrictifs ont été testés sur ces données, pour permettre d'avoir une inflation plus cohérente avec le marché (voir *annexe 5.2*).

En effet, lors de l'étude des prix des pièces des rapports d'expertise, nous avons constaté qu'on retrouvait le prix constructeur des pièces dans une majorité des cas. Ainsi, lorsque que notre besoin était d'étudier des séries temporelles du coût moyen des pièces, la mise en place d'un traitement des valeurs aberrantes était relativement simple ; cependant ici les coûts moyens se calculent sur peu de pièces correspondant aux branches de risques spécifiques et il est alors plus difficile de distinguer les valeurs incohérentes avec celles observées sur le marché.

Pour éviter cela, des restrictions en amont et en aval de l'agrégation des pièces ont été appliquées.

- Ecrêtement des lignes en amont de l'agrégation : pour éviter d'avoir de fortes variations dans les séries des coûts moyens des pièces pour chacune des 5 pièces détachées suivies, les lignes de pièces supprimées sont celles dont les **prix étaient en dessous du centile 15% ou au-dessus du centile 85% des prix des pièces de la même marque et de la même année**. Ce traitement sur les centiles s'est fait pour chaque pas temporel, comme illustré sur le graphique 4.10.

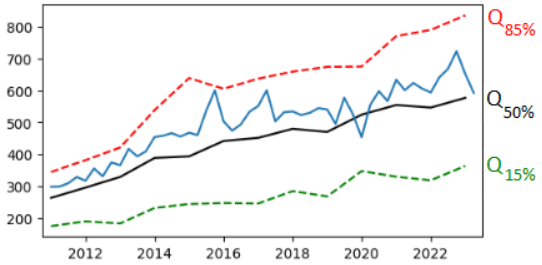


FIGURE 4.10 – Evolution du coût moyen du bouclier avant d'Audi sans traitement avec les limites calculées sur chaque année

Avec ce traitement, on retient 70% des pièces de la base par rapport à l'étape précédente.

Pièce	Nombre avant traitement	Nombre après traitement	% conservé
Bouclier avant	42 237	35 479	84%
Bouclier arrière	32 188	26 560	83%
Aile avant gauche	16 419	4 262	26%
Porte avant	13 050	7 328	56%
Porte arrière	8 820	5 262	60%
Total	112 714	78 891	70%

- Ecrêtement des lignes en aval de l'agrégation : ensuite, même si nous avons fait un retraitement en amont pour éviter d'avoir des inflations calculées extrêmes après les agrégations, certains montants de la variable calculée obtenus restaient trop élevés par rapport aux autres. Ainsi, en aval nous avons décidé de réduire d'avantage les lignes de la base en retirant les pièces qui avaient des inflations calculées en dessous de leur centile 10% et au-dessus de leur centile 90%.

Pour la variable cible, on reconnaît une distribution de **loi normale** en regardant la forme de l'histogramme de la variable ainsi que le diagramme "quantile-quantile", autrement appelé *qqplot*, qui compare les quantiles d'une loi normale avec ceux de la variable créée (voir *figure 4.11*). Cette intuition de normalité pour la variable a été validée par le test statistique de D'Agostino et Pearson réalisé à l'aide de la fonction *python normaltest* [30]. Ce test combine l'asymétrie et l'aplatissement pour produire un test collectif de normalité.

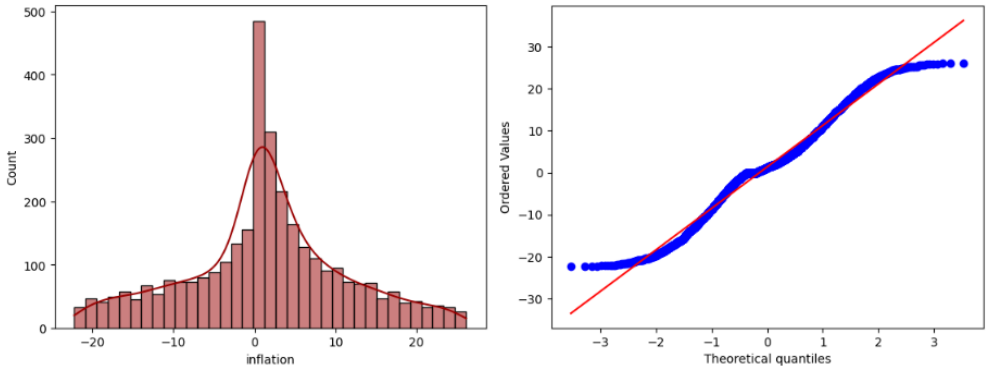


FIGURE 4.11 – Histogramme et qqplot avec une loi normale de la variable cible calculée

En connaissant la fonction de distribution de la variable cible on peut alors choisir la fonction de lien adaptée pour le GLM. Dans ce cas, la fonction de lien est l'identité.

Etape 6 : Tests sur les différents paramètres

Dans l'optique de trouver le modèle optimum permettant d'expliquer l'inflation des pièces détachées, de nombreux tests ont été lancés. Tout d'abord, des tests sur le périmètre de la base d'apprentissage puis sur les modèles en eux-mêmes avant de finir sur la modification de la variable cible.

- Sur la base initiale : des modèles ont été créés avec différentes bases en entrée pour prendre en compte :

— les combinatoires différentes du nombre de modalités pour les variables d'agrégation en branche de risque. Elles ont la forme "x pièces $\times$ x marques $\times$ x anciennetés $\times$ x classes SRA $\times$ x groupes SRA $\times$ Année Semestre" avec "x" les nombres de modalités par variable comme dans le tableau suivant :

Pièces	Marques	Ancienneté	Classe SRA	Groupe SRA	Temporalité
5	13	4	3	3	Année Semestre
5	13	6	6	6	Année Semestre
5	13	8	8	8	Année Semestre

— les années considérées, c'est-à-dire en retirant les années post-covid, en considérant une seule année, tout l'historique, etc. ;
— les pièces considérées ;
— le nombre de variables explicatives pour limiter le bruit.

Ces différents tests ont été réalisés via une régression avec la variable cible en numérique mais aussi à l'aide d'une classification avec les valeurs de la variable classée dans 4 groupes, 6 groupes ou 10 groupes de modalités.

- Sur les modèles et leurs paramètres : une fois que ces tests ont mené à un périmètre pour la base initiale et une intuition des modèles à retenir, les paramètres des modèles ont été optimisés par validation croisée avec la fonction *GridSearchCV*. Cette fonction teste toutes les combinaisons des paramètres demandés pour calculer les valeurs de la métrique souhaitée et donner ainsi la meilleure combinaison de paramètres pour cette métrique.

Dans notre cas, pour la régression, la métrique que nous avons optimisée est le RMSE. Le R^2 semblait moins pertinent dans le sens où il consiste à comparer les prédictions avec la valeur moyenne, or une inflation en 2011 est différente d'une inflation en 2022 qui est une période de forte inflation. Ainsi cette métrique a été regardée à la fin mais pas utilisée pour l'optimisation. En classification il s'agissait de l'*accuracy*.

→ Difficulté à expliquer les inflations extrêmes : dans ces différents cas, on a constaté une difficulté des modèles à prédire des inflations fortes. En effet, comme dans les cas plus extrêmes nous avons moins d'observations, il est compliqué de trouver des critères qui permettent de les expliquer.

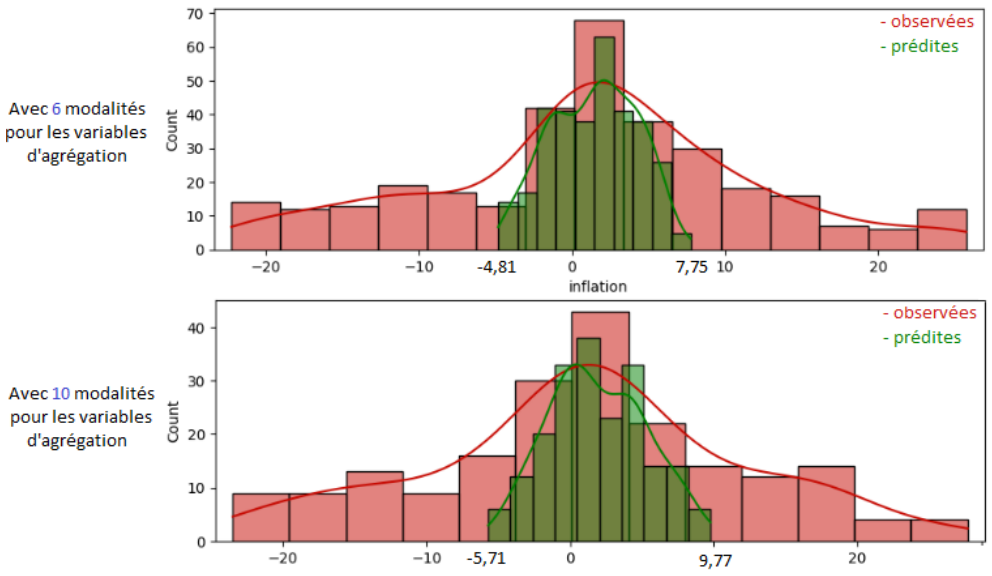


FIGURE 4.12 – Histogramme des valeurs observées et prédites d'un *histogram gradient boosting* pour 6 modalités (en haut) et 10 modalités (en bas) par variable d'agrégation

Comme nous pouvons le voir sur les *figures en 4.12*, en ajoutant des modalités dans les variables d'agrégation des pièces, nous arrivons à élargir légèrement l'intervalle des valeurs prédites en passant de [-4,71 ;7,75] à [-5,71 ;9,77].

Ainsi, on en conclut deux choses avec ce constat :

— D'abord que plus les données d'apprentissage construites sont fines, c'est-à-dire avec un grand nombre de modalités pour les variables d'agrégation, plus les mo-

dèles améliorent les prédictions "extrêmes". Cependant, le gain reste mesuré, ce qui nous amène à conclure que nos données ont une limite "physique" qui ne permet pas d'espérer une optimisation forte des performances des modèles.

— Et puis comme le R^2 est une métrique qui est proche de la variance expliquée, on comprend aisément pourquoi cette métrique n'est pas très significative dans les modèles.

Etape 7 : Base d'apprentissage obtenue

Avec ces bases, on peut alors créer une base d'apprentissage avec 90% des lignes sélectionnées aléatoirement et créer une base de test avec les 10% restants, cela pour avoir suffisamment de lignes pour l'apprentissage.

Pour les modèles que l'on va présenter, nous avons finalement conclu sur ces choix :

1. Tout d'abord, pour faire un compromis entre le nombre de lignes de la base et la finesse de l'étude, les variables d'agrégation avaient **6 modalités** pour l'ancienneté du véhicule à la date du rapport, les classes de prix SRA et les groupes de puissance SRA, et une temporalité **semestrielle**. Cela revient à choisir la combinatoire "$6 \times 6 \times 6$" d'après ma notation.
   Ces différents segments sont présentés en *annexe 5.2.*

2. De plus, dans les différents tests, nous avions constaté que les modèles étaient plus performants :
   — soit en prenant qu'une seule pièce et en gardant plusieurs années d'observations,
   — soit en considérant les 5 pièces suivies d'une plus courte période.
   Pour pouvoir conserver les informations historiques, seuls les "**boucliers avant**" ont été gardés.

3. Par ailleurs, comme l'année 2020 a été très particulière avec la pandémie du covid et son confinement, tout l'historique de la table a été conservé sauf cette année 2020. Ainsi les années **de 2011 à mi 2023 en retirant l'année 2020** ont été gardées pour s'assurer d'avoir suffisamment de volume pour l'apprentissage et la validation des modèles.

4. Enfin, les **variables explicatives ont été réduites** pour éviter le bruit dans les résultats en retirant
   — une des variables parmi celles fortement corrélées entre elles comme le Gazole avec le SP-98 ou l'acier et l'aluminium qui ont une corrélation explicable par lien direct,
   — et celles qui avaient des coefficients très proches de zéro dans le GLM pénalisé ont aussi été retirées au fur et à mesure, en vérifiant que les performances n'étaient pas détériorées.

On obtient une base d'apprentissage de 3 242 lignes pour 22 variables explicatives sur lesquels nous pouvons lancer les modèles et optimiser les paramètres.

4.2.2 Résultats obtenus

Avec cette base d'entraînement, les métriques obtenues pour la régression sont les suivantes :

	RMSE		MAE		R^2 (%)	
Modèle	test	train	test	train	test	train
Régression linéaire	10.09	9.54	7.73	7.21	0.09	0.07
GLM pénalisé	**10.06**	9.43	7.79	7.12	**9.71**	9.31
Arbre	10.09	9.54	7.79	7.24	9.14	7.22
HGB	**10.04**	9.14	7.70	6.86	**10.13**	14.89
Forêt aléatoire	10.09	9.24	7.78	7.00	9.24	13.04

On constate sur ce tableau que :

— Tout d'abord, il n'y a pas de phénomène de surapprentissage [7], grâce à l'optimisation des paramètres par *cross validation* ;

— Ensuite, plus les modèles sont complexes plus les performances s'améliorent, cela est notamment visible sur le R^2 mais assez peu sur les autres critères.

Et pour la classification avec 4 modalités pour la variable cible et des performances de hasard de 25% :

	Accurary		F1		Précision		Rappel	
Modèle	test	train	test	train	test	train	test	train
GLM pénalisé	0.36	0.37	0.35	0.37	0.35	0.37	0.36	0.37
Arbre	0.35	0.33	0.28	0.28	0.25	0.26	0.35	0.33
HGB	0.34	0.50	0.32	0.50	0.32	0.51	0.34	0.50
Forêt aléatoire	0.30	0.41	0.28	0.40	0.28	0.41	0.30	0.41

On constate que nous n'avons pas de problème de déséquilibre des classes avec ces métriques dont les résultats sont similaires. Cependant on voit également que, malgré l'optimisation des paramètres, un léger surapprentissage apparaît pour l'*histogram gradient boosting* et la forêt aléatoire.

Nous avons montré précédemment que les modèles étaient limités en performances très probablement dû à une limite physique de volume de données, mais ils donnent tous des résultats très proches sur lesquels nous allons nous concentrer pour la suite.

Analysons deux des modèles de régression, le GLM qui est facilement interprétable et l'*histogram gradient boosting* qui est plus complexe.

7. Le surapprentissage, ou *overfitting*, est un comportement indésirable d'apprentissage automatique qui se produit lorsque le modèle arrive à obtenir des prédictions très précises pour les données d'entraînement mais pas pour les données de test.

- Les modèles de régression

Le GLM pénalisé : commençons par regarder les résultats du GLM pénalisé qui a pour avantage d'être facilement interprétable. Celui-ci donne des coefficients proches de zéro lorsque la modalité de la variable qualitative ou la variable numérique n'est pas considérée comme significative. Ce modèle revient à faire pour chaque individu i :

$$inflation_i = intercept + \sum_{i \in I} \beta_i variables\ vehicule_i + \sum_{j \in J} \beta_j variables\ marche_j + \beta_t temps \tag{4.9}$$

avec I={marque, ancienneté, classe SRA, groupe SRA} et J comprenant les variables des indices marché qui ont été conservées suite à la réduction de variables.

Ainsi, lorsqu'on regarde les variables qui ont des coefficients de plus de 0,1 et de moins de -0,1, on voit principalement les modalités de marques, d'anciennetés du véhicule, de classes de prix et de groupes de puissances SRA apparaître.

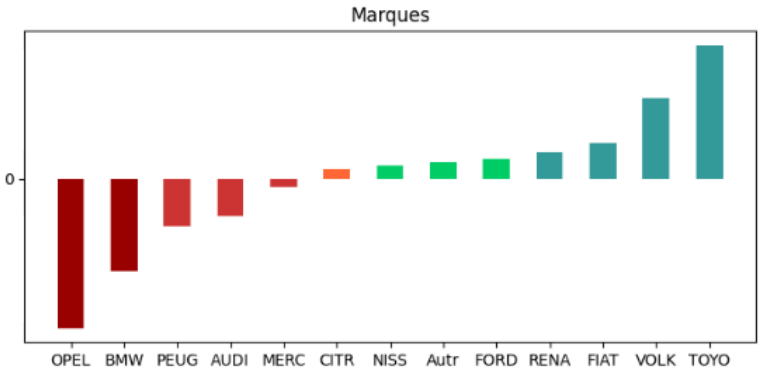


FIGURE 4.13 – Coefficients du GLM pénalisé pour les modalités de marques [8]

Pour les marques comme Opel ou BMW, les coefficients β sont négatifs, cela signifie qu'elles auraient plutôt tendance à avoir une évolution de leurs prix plus faible que les autres marques. A l'inverse, les coefficients de Toyota, Volkswagen ou encore Fiat sont positifs, c'est-à-dire qu'elles augmenteraient plus leurs évolutions de prix de leurs pièces automobiles.

Les coefficients correspondant aux variables "ancienneté" et "classe SRA" sont croissants en fonction des modalités des variables. Concrètement, cela signifie que les véhicules récents augmentent moins leurs pièces que les véhicules âgés et les véhicules de classes de prix SRA faibles augmentent moins leurs pièces que les véhicules de classes plus élevées.

8. Pour des raisons de confidentialité des résultats, l'échelle de l'axe y a été retiré

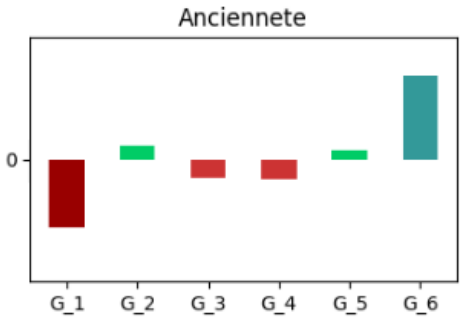


FIGURE 4.14 – Coefficients du GLM pénalisé pour les modalités d'ancienneté

FIGURE 4.15 – Coefficients du GLM pénalisé pour les modalités de classes SRA

Inversement, les coefficients des modalités de groupes de puissances SRA sont décroissants. Cela signifie que, plus le véhicule est puissant et moins l'évolution de prix est importante.

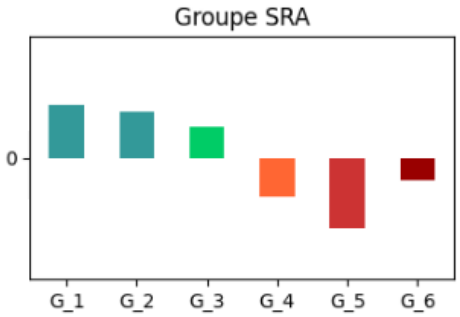


FIGURE 4.16 – Coefficients du GLM pénalisé pour les modalités de groupes SRA

L'*histogram gradient boosting* : ce modèle qui a de meilleures performances en contrepartie d'une interprétabilité moindre, a été appliqué.

Pour permettre d'interpréter ces modèles de *machine learning* plus complexes, on utilise l'algorithme SHAP, basé sur les valeurs de Shapley et ses représentations.

La première représentation permet de mettre en avant les variables les plus significatives dans les modèles. Sur cette *figure 4.17*, à la grande différence avec le GLM, la variable coût moyen sans traitement apparaît en premier, avant les variables similaires. Cette variable ne créé pas de phénomène dit de leakage [9]dans le modèle dans le sens où les modèles n'ont pas de performances trop élevées.

9. On définit le *data leakage* comme étant une fuite de données qui améliore artificiellement les modèles si un lien direct avec la variable cible existe. Dans ce cas, même si ce coût moyen de la pièce fait

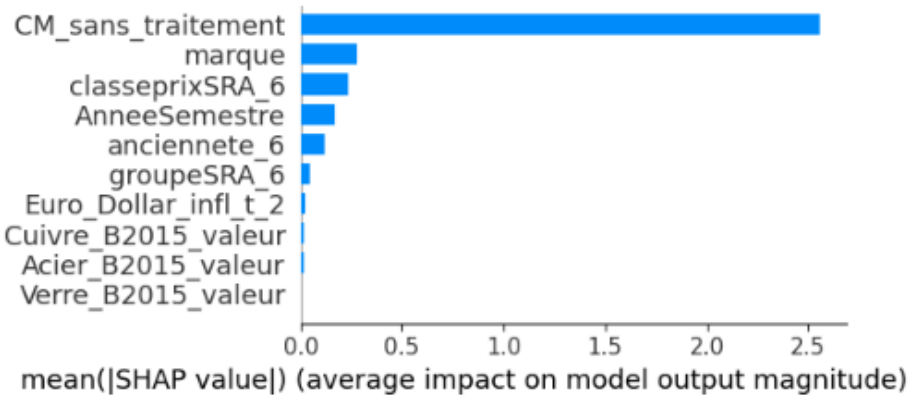


FIGURE 4.17 – Moyennes des valeurs absolues des mesures *shap* les plus significatives

Ensuite, une seconde représentation permet d'interpréter le sens dans lequel les modalités influencent la variable cible. Sur ce *graphique 4.18* les couleurs et la position des points par rapport à l'axe vertical, représentant des individus, permettent de donner une indication.

Ainsi, plus la couleur est froide alors plus l'individu correspond à une modalité faible, et inversement plus la couleur est chaude alors plus la modalité est élevée. Si le point est à gauche, la modalité correspondante impacte négativement la variable cible par rapport à la moyenne des prédictions, c'est-à-dire qu'il contribue à la baisse de l'évolution du prix de la pièce automobile. Sinon, elle contribue à sa hausse.

Dans ce cas, le package SHAP permet de ressortir des variables significatives comme pour le GLM mais les tendances sont plus difficiles à comprendre.

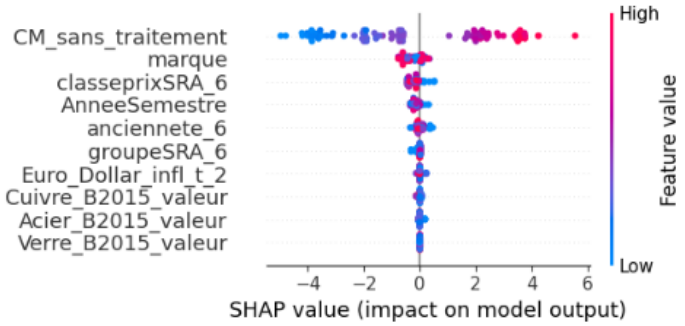


FIGURE 4.18 – Impact des modalités des variables sur la variable cible avec les valeurs *shap*

partie du calcul de la variable cible, comme toutes les composantes de ce calcul ne sont pas présentes dans les données explicatives, en théorie il n'y a pas de leakage avéré.

- Etude de corrélation entre la variable cible et les variables de marché :

Pour essayer de comprendre pourquoi aucune variable de marché ne ressort comme étant significative dans l'explication de l'inflation du coût de la pièce automobile, même avec un décalage temporel, une étude des évolutions de ces variables a été réalisée. Cette étude se compose en 2 parties :

1. Pour commencer, on a regardé les corrélations entre la variable cible qui ont été calculées à partir de l'évolution du coût moyen de la pièce pour la branche de risque considérée et les données macro-économiques. Celles-ci sont **très faibles**, que ce soit avec les variables qui donnent les caractéristiques du véhicule ou celles des informations de marché.

2. Après cela, nous avons souhaité comprendre pourquoi la variable sur l'évolution du coût de la pièce n'était liée à aucune variable de marché, même avec un décalage temporel. On a alors tracé les évolutions du coût moyen du prix d'une pièce d'une marque donnée, ici le bouclier avant d'une Citroën, avec l'évolution des indicateurs de marché comme les indices des prix de certains matériaux et le prix de l'essence.

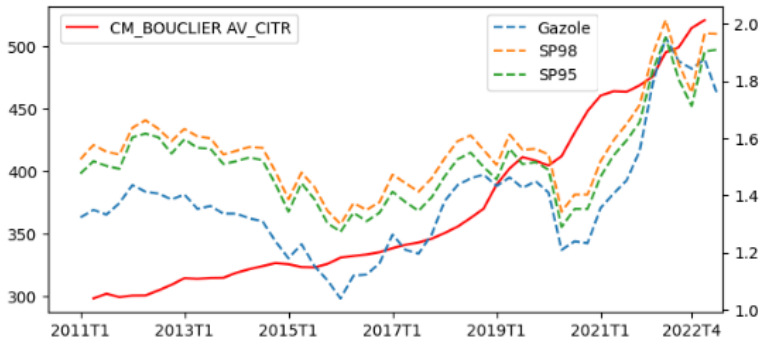


FIGURE 4.19 – Evolution du coût moyen du bouclier avant d'une Citroën et évolution des prix de l'essence

Dans tous ces cas, on constate que, contrairement au marché, le prix de la pièce augmente presque pendant toute la durée de l'historique. Même si le confinement en 2020 a eu un impact très important sur le marché, les constructeurs n'ont pas pour autant changé leur dynamique de prix des pièces détachées pour contrer la baisse dans leurs ventes de véhicules.

Que ce soit en 2015 ou en 2021, le marché a eu une diminution assez importante. De son côté, le coût moyen de la pièce est resté en constante évolution en 2015 et sa baisse en 2020 était beaucoup plus faible que le choc subit sur le marché.

Cette analyse approfondie permet de mieux comprendre pourquoi aucune variable de marché ne ressort comme étant discriminante de l'inflation du coût de la pièce.

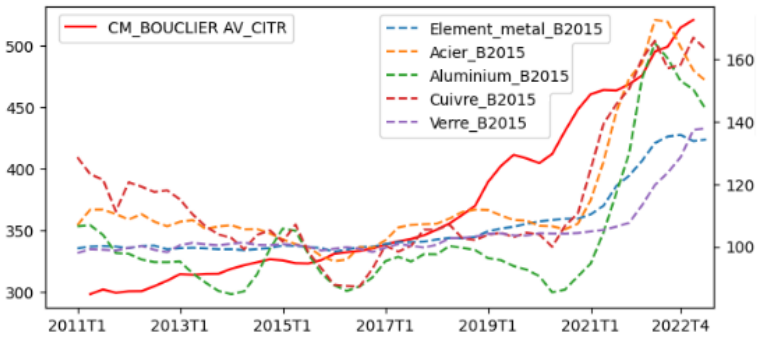


FIGURE 4.20 – Evolution du coût moyen du bouclier avant d'une Citroën et évolution des indices des prix des métaux

4.2.3 Conclusion

Au cours de cette analyse on a pu mettre en avant la difficulté de prédire l'inflation à une maille aussi fine à cause d'une limitation due aux données. En effet, pour cette étude, il fallait créer des agrégations sur de nombreuses variables ce qui a eu comme conséquence que chaque agrégation avait peu de données pour obtenir des coûts moyens robustes et sur deux temporalités consécutives.

En plus de cela, comme les prix des pièces automobiles ne suivent pas les mêmes évolutions que les indices de marché récupérés, ils ne peuvent pas aider ces modèles à comprendre cette variable créée.

Cependant, même avec cette difficulté, tous les tests réalisés ont montré une cohérence dans les signaux faibles détectés par les différents modèles. A chaque fois les variables les plus significatives étaient la **marque**, l'**ancienneté du véhicule à la date du rapport**, la **classe de prix SRA** et le **groupe SRA**.

Cela nous permet de conclure sur les variables discriminantes de l'inflation du coût de la pièce détachée.

Chapitre 5

Réflexion sur la mise en place opérationnelle

Pour donner suite aux parties précédentes, nous souhaitons pouvoir utiliser les résultats des études pour les intégrer au processus de tarification et répondre à la demande opérationnelle du Bureau d'Etudes Technique (BET) Auto.

Ainsi, avec le *chapitre 3*, un outil de détection de ruptures, dans les tendances des évolutions des coûts moyens des pièces automobiles, a été créé pour détecter au plus tôt les évolutions de politiques tarifaires sur les pièces détachées des constructeurs automobiles ; et anticiper leur impact sur les coûts moyens des réparations.

De plus, dans le *chapitre 4*, une étude visant à identifier les facteurs discriminants de l'inflation du coût de la pièce automobile a été menée. Malgré les difficultés de modélisation rencontrées, des signaux faibles ont été détectés, et nous avons montré que la marque, l'ancienneté, la classe de prix SRA et le groupe SRA étaient des éléments importants pour différencier l'inflation des pièces détachées.

Dans ce nouveau chapitre, l'idée est d'utiliser les résultats des études précédentes pour se projeter dans une application opérationnelle pour le BET Auto.

Pour cela, une première partie concerne la mise à jour de l'hypothèse d'inflation avec l'intégration des résultats dans le processus actuel et une seconde partie portera sur une rapide réflexion des ajustements possibles pour rendre l'indicateur plus opérationnel.

5.1 Utilisation des résultats pour sophistiquer l'intégration de l'inflation dans le processus de tarification

Pour rappel, la tarification est revue chaque année vers le milieu d'année pour mettre à jour les montants des primes des assurés de l'année à venir. Pour cela, l'assureur se base sur les informations dont il dispose, c'est-à-dire les données arrêtées au 31 décembre de l'année précédente, comme présenté sur la *figure 5.1*.

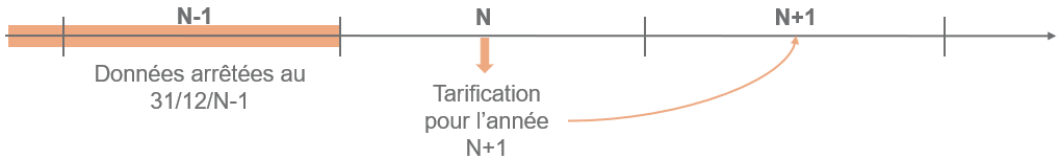


FIGURE 5.1 – Schéma des années utilisées en tarification

Ainsi, pour définir cette hypothèse de projection d'inflation, l'assureur peut s'appuyer sur différents éléments dont il dispose comme le contexte actuel, des indices macro-économiques, l'évolution historique des coûts des pièces de ses assurés sinistrés, etc.

Dans le cas où une rupture a été identifiée, l'assureur a alors un élément supplémentaire pour revoir son hypothèse.

Par exemple, on peut définir que l'inflation de l'année passée sera la même que l'année qui vient. De cette façon, si par exemple l'inflation de l'année N-1 était de 5%, alors on peut imaginer une inflation de 5% pour l'année N+1 sans détection de rupture. Si une rupture a été repérée et qu'elle est :

— à la baisse, alors cette hypothèse peut se retrouver inchangée pour rester prudent ou si la raison de la rupture n'est pas comprise ;
— à la hausse, alors l'hypothèse peut être adaptée à l'évolution identifiée.

Une fois cette hypothèse définie, pour distinguer cette inflation entre un véhicule de marque Audi et une autre marque ou entre un véhicule récent ou plus ancien par exemple, les coefficients obtenus avec le GLM en *partie 4.2.2* pourraient être utilisés.

La procédure mise en place peut se représenter comme dans la *figure 5.2* suivante :

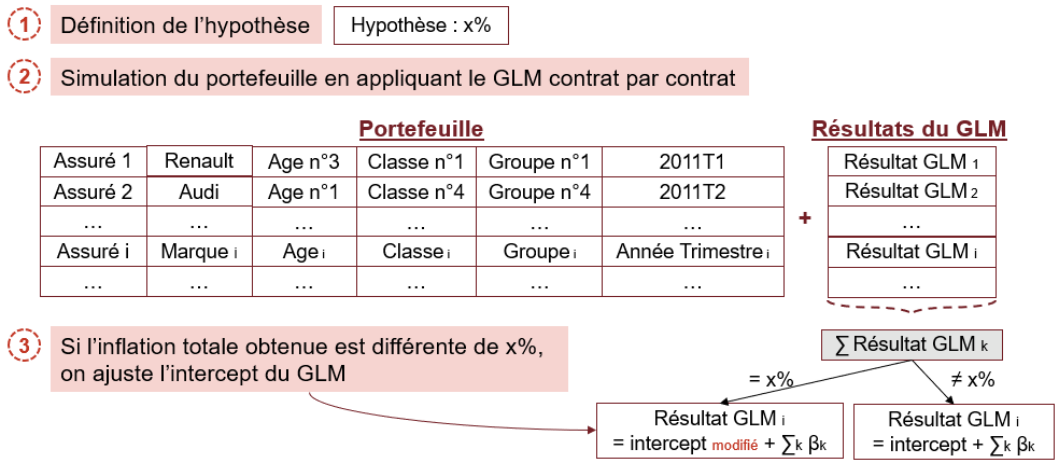


FIGURE 5.2 – Schéma expliquant l'utilisation des coefficients pour différencier l'hypothèse d'inflation selon les critères de différenciation

Ces étapes sont :

1. Le BET Auto fixe un niveau d'inflation global de $x\%$ à injecter.
2. On simule sur le portefeuille les inflations différenciées contrat par contrat en appliquant le GLM.
3. On ajuste l'*intercept* du GLM pour avoir une inflation totale injectée du niveau $x\%$, tout en conservant la différenciation de l'inflation sur chaque contrat.

Opérationnellement, les conclusions du GLM pourraient ainsi être utilisée rapidement pour différencier l'inflation, à travers l'utilisation des coefficients en adaptant l'*intercept* du modèle en fonction du niveau d'inflation global qui est souhaité par Generali.

Cependant, les choix de politiques tarifaires vont amener certainement à ne sélectionner qu'une liste réduite de critères différenciants par rapport à celles définis par l'étude.

5.2 Adaptation des résultats opérationnellement

Dans l'environnement de marché assurantiel, les tarifs que l'assureur propose doivent être explicables pour, d'une part être capable de détailler à ses assurés le montant de leur prime commerciale, et d'autre part pour respecter les recommandations de l'Autorité de Contrôle Prudentiel et de Résolution (ACPR).

Ainsi, pour faciliter cette explicabilité du tarif et son le pilotage par la direction, les résultats d'un tel modèle pourraient être adaptés pour une application opérationnelle.

1. Dans un premier temps, le nombre de variables pourra être modifié. Ainsi, même si dans l'étude, 4 variables sont ressorties comme discriminant l'inflation, l'expérience métier va favoriser l'usage de quelques-unes seulement avec un ordre de priorité d'intégration étant marque, ancienneté, classe SRA puis groupe SRA (ayant une corrélation non négligeable avec la classe SRA mais fournissant des informations supplémentaires sur la dangerosité du véhicule). En effet, que ce soit pour expliquer aux assurés les raisons de leur tarification ou alors pour prendre le temps de s'adapter, l'assureur pourra choisir seulement 2 ou 3 variables par exemple.

2. Ensuite, comme nous l'avons vu sur les graphiques des coefficients du GLM que nous représentons à nouveau ici, certaines modalités peuvent être regroupées pour éviter d'avoir beaucoup de situations de risques différentes.
   Dans ce cas, on pourrait regrouper les modalités de sorte à n'avoir plus que 2 ou 3 modalités par variable discriminante. Par exemple :

Hypothèses de segmentation opérationnelle	Ancienneté du véhicule	Classe de prix SRA	Groupe SRA
1	G_1	G_1 et G_2	$G_1 \rightarrow G_3$
2	$G_2 \rightarrow G_5$	G_3 et G_4	$G_4 \rightarrow G_6$
3	G_6	G_5 et G_6	

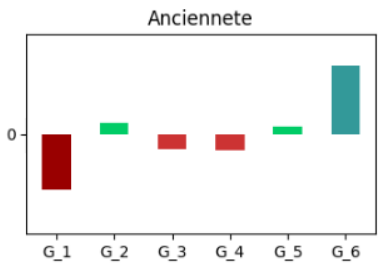


FIGURE 5.3 – Coefficients du GLM pénalisé pour les modalités d'anciennetés

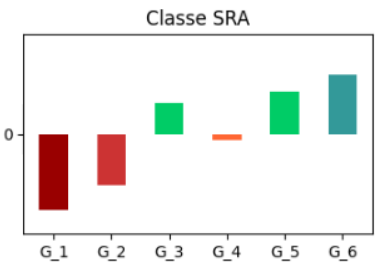


FIGURE 5.4 – Coefficients du GLM pénalisé pour les modalités de classes SRA

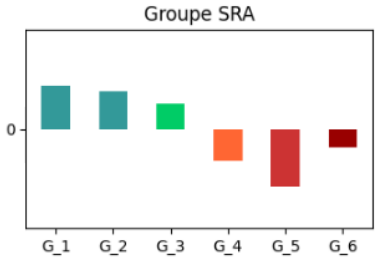


FIGURE 5.5 – Coefficients du GLM pénalisé pour les modalités de groupes SRA

3. Il peut également y avoir des ajustements possibles sur ces distinctions pour permettre d'être plus justes envers les assurés et prudents. En effet, d'après le modèle, Opel et Toyota sont les marques avec les impacts sur l'inflation les plus opposés. Ainsi, si l'on veut lisser légèrement la répartition, on peut diminuer les amplitudes de ces coefficients pour avoir des montants plus rapprochés.

Cette différenciation, qui répond aux choix de politiques tarifaires de Generali, a pour vocation d'être intégrée dans les outils métiers de l'entreprise pour prendre en charge cette inflation plus fine sur les contrats d'assurance automobile.

Conclusion

Les assurances, du fait de leur rôle dans la société, font partie des agents aisément impactés par les crises mondiales.

En raison de contextes inflationnistes et de sur-sinistralité de l'année 2022, les assureurs ont été amenés à adapter leur processus de tarification afin de réagir face aux nouvelles conditions du marché. Pour les produits d'assurance IARD, ce processus revient à estimer, d'une part, la fréquence de sinistres, et d'autre part, le montant moyen attendu d'après les informations historiques. Pour ce faire, l'assureur se base sur les données des années précédentes, par conséquent, afin d'estimer un montant de sinistre futur, une hypothèse d'inflation doit être réalisée. Avec une inflation de 9,6% en 2022 sur les pièces dans une réparation moyenne [1], cette dernière est devenue un enjeu important pour les assureurs.

Ainsi, pour estimer le montant moyen d'un sinistre auquel on s'attend pour l'année tarifée, une hypothèse d'inflation doit être réalisée.

Suite à ce constat contextuel, le sujet de ce mémoire consiste à affiner l'hypothèse d'inflation qui est actuellement faite de manière uniforme sur les contrats, afin de l'appliquer à une maille plus fine notamment en tenant compte des caractéristiques du véhicule.

Pour répondre à cette problématique, les informations des rapports d'expertise automobile [2]ont été exploitées en raison de leur grande finesse. Pour cela 3 étapes principales ont été appliquées :

Le premier objectif était de mettre en place un outil permettant de suivre en continu l'évolution des prix constructeurs des pièces détachées, et ainsi alerter l'assureur le plus en amont possible lorsqu'un changement de politique tarifaire des constructeurs est avéré. Cela permet aux actuaires d'anticiper au mieux les questions sur l'hypothèse d'inflation qu'il injectera dans les modèles de projection.

Pour cela, un outil de détection, basé sur des modèles *Prune Exact Linear Time* et de segmentation binaire, permet d'identifier les temps de ruptures de la tendance des évolutions des prix des pièces. Et si ce temps est proche, alors l'outil donne une alerte.

1. D'après l'association "Sécurité et Réparation Automobiles"
2. sur les véhicules particuliers et utilitaires de moins de 3,5 tonnes, réparables et sinistrés en France métropolitaine.

Cet outil avait notamment identifié le second trimestre 2022 comme étant un instant modifiant significativement la tendance de l'évolution du coût de ces pièces automobiles.

La seconde partie, quant à elle, avait pour objet d'identifier les facteurs discriminants de l'inflation du coût de la pièce dans le but de sophistiquer le processus de tarification.

Pour atteindre cet objectif, des modèles de GLM pénalisés qui sont facilement interprétables et d'autres plus complexes de *Machine Learning* comme l'*histogram gradient boosting* ont été utilisés.

L'analyse des performances de ces modèles a montré la complexité inhérente à la modélisation de l'inflation des pièces automobiles. Ceci s'explique par la difficulté d'obtenir une base de modélisation qui décrit l'inflation avec une granularité fine et qui possède une volumétrie conséquente de données. Ainsi, ce compromis entre ces deux éléments a eu pour effet de réduire la possibilité d'apprentissage des modèles sur les fortes variations par manque d'exemples. De plus, les informations macro-économiques suivent des évolutions différentes des prix des pièces automobiles comme on a pu le voir en 2020 avec la pandémie de la covid-19 où les indices du marché, comme le pétrole, ont baissé à cette période contrairement aux prix des pièces détachées. Cela était dû à la faiblesse de la demande en matières premières et à la baisse de la production de pièces à la suite des confinements. Ainsi, ces informations externes n'ont pas expliqué cette variable, même en considérant des décalages temporels.

Malgré cela, les différents modèles que nous avons obtenus, que cela soit en régression ou en classification, ont permis de mettre en évidence le caractère discriminant des variables "marque", "ancienneté", "classe de prix SRA" et "groupe SRA" ; même s'il s'agit de signaux relativement faibles.

Enfin, dans la troisième partie nous nous sommes projetés dans une exploitation opérationnelle de nos résultats.

Ainsi, par sa facilité d'utilisation, nous montrons une stratégie d'utilisation du GLM pénalisé obtenu pour différencier l'inflation contrat par contrat.

Cependant, pour des raisons d'explicabilité et d'intégration, les politiques tarifaires vont certainement amener à ne sélectionner qu'une liste réduite de critères différenciants par rapport à celles définis par l'étude. Afin de réduire cette finesse, un choix de quelques-unes des variables peut être adopté, une réduction dans les modalités de ces variables différenciant ; et dans le but d'être plus prudent, une diminution de l'amplitude des coefficients peut être envisagée.

Ce mémoire a contribué à sophistiquer le modèle de tarification actuel en permettant une segmentation de l'inflation selon les profils des contrats. Les résultats sont facilement exploitables par le Bureau d'Etudes Techniques, et après les simplifications et les ajustements dus aux contraintes commerciales, ils pourront facilement être intégrés dans les outils de tarification de Generali.

A la suite de cette étude, bien qu'intense et structurée, il est possible d'évoquer plusieurs pistes de réflexion afin de prolonger ce sujet. Tout d'abord, pour compenser la

limite liée au volume de données, une nouvelle analyse peut être réalisée en considérant un périmètre plus important de pièces étudiées. Il serait également intéressant de rechercher de nouvelles sources de données externes telles que celles d'EUCON afin de suivre les prix constructeurs de manière plus directe. Un croisement de données permettrait d'augmenter la véracité des prix observés. Enfin, il est possible d'envisager de compléter la base d'apprentissage avec quelques variables explicatives telles que la carrosserie, l'identification des véhicules électrique ou les complexités des technologies intégrées qui seraient ensuite étudiées par le GLM.

Table des figures

1 Schéma résumé du traitement des courbes avant la détection de rupture . iii
2 Ruptures obtenues avec la méthode *PELT* sur la courbe différenciée et la courbe normale iv
3 Ruptures obtenues avec la méthode *Binary segmentation* avec une pénalité de 10, sur la courbe différenciée et la courbe normale iv
4 Moyennes des valeurs absolues des mesures SHAP les plus significatives . . vi
5 Coefficients du GLM pénalisé pour les modalités de marques vii
6 Coefficients du GLM pénalisé pour les modalités d'anciennetés vii
7 Coefficients du GLM pénalisé pour les modalités de classes SRA vii
8 Coefficients du GLM pénalisé pour les modalités de groupes SRA vii
9 Summary diagram of curve processing prior to break detection xi
10 Breaks obtained with the PELT method on the differentiated curve and the normal curve xii
11 Breaks obtained with the Binary Segmentation with a penalty of 10, on the differentiated curve and the normal curve xii
12 Average absolute values of the most significant shap measurements xiv
13 Penalized GLM coefficients for brand modalities xiv
14 Penalized GLM coefficients for seniority modalities xv
15 Penalized GLM coefficients for class modalities xv
16 Penalized GLM coefficients for SRA group modalities xv

1.1 Evolutions mensuelles et annuelles de l'indice d'inflation par l'INSEE (en % et en base 2015) 3
1.2 Répartition des véhicules particuliers par vignettes Crit'Air en 2021 (Source) 5
1.3 Les zones ZFE actuelles et en cours en France 6
1.4 Nombre et âge moyen des véhicules particuliers en circulation du parc français 6
1.5 Immatriculations des marques des véhicules particuliers du marché français en 2022 (Source : ccfa.fr/) 7
1.6 Regroupements de marques 8
1.7 Exemple de pièces détachées avant d'une Renault Captur 10
1.8 Différences d'éclairages entre les phares halogène, Xénon et LED (Source : next-tech-france.com 12
1.9 Principe d'inversion du cycle de production 13

1.10 Principe de mutualisation *openclassrooms.com* 14
1.11 Le risque de sur-mutualisation 15
1.12 Exemple simplifié de multiplication pour obtenir un risque homogène . . . 15
1.13 Ratio combiné comptable net de réassurance en automobile (Sources : *ACPR - France Assureurs*) 18
1.14 Différentes étapes en tarification IARD (Incendie, Accidents et Risques Divers) 19

2.1 Exemple de présentation de rapport anonymisé (page 1 sur 3) 22
2.2 Exemple de présentation de rapport anonymisé (page 2 sur 3) 22
2.3 Exemple de présentation de rapport anonymisé (page 3 sur 3) 23
2.4 Exemple de pivot réalisé sur la base de données 24
2.5 Coût moyen des pièces par mois avec et sans le libellé "Ensemble de pièces" dans la table de données 25
2.6 Exemple de prix constructeur récupérés dans la matrice automobile sur la Renault Clio IV 26
2.7 Exemple de sélection d'une Renault Clio IV sur EUCON 27
2.8 Exemple de sélection d'un bouclier avant d'une Renault Clio IV sur EUCON 27
2.9 Comparaison des prix du bouclier avant de la Clio IV dans différents pays 27
2.10 Evolution du prix du bouclier avant de la Clio IV en France 27
2.11 Evolutions mensuelles du nombre de pièces "ensemble de pièces", "projecteur g g" et "pare-chocs av" 29
2.12 Représentation d'un "boxplot" 30
2.13 Exemple de courbes avant traitement, sans valeurs aberrantes puis lissée (sur 3 trimestres) du coût du bouclier avant pour la marque Renault et pour le groupe Renault Nissan Mitsubishi 30
2.14 Evolution du nombre de Bouclier avant et du coût moyen (base T1 2011) de Volkswagen et 3 de ses modèles Golf, Polo et Tiguan. 32
2.15 Evolution du nombre de Bouclier avant et du coût moyen (base T1 2011) du groupe Volkswagen et 3 de ces marques, Volkswagen, Audi et Seat. . . 33
2.16 Evolutions mensuelle et annuelle (base 2015) de l'indice des prix à la consommation par l'INSEE 34
2.17 Evolutions mensuelles du cours des monnaies en euros - dollars américains par l'INSEE 35
2.18 Indices de prix de production d'industrie des métaux français pour le marché français obtenus par l'INSEE 36
2.19 Cours du pétrole brut Brent (Londres) importé en dollars américains par baril par l'INSEE 36
2.20 Prix moyens mensuels de vente au détail en métropole des matières premières par l'INSEE 37
2.21 Prix du verre et d'articles en verre du CPF 23.1 (base 2015) par l'INSEE . 37
2.22 Indice du coût de la main d'oeuvre de l'industrie manufacturière (base 2016) par l'INSEE 38

3.1 Schéma des années utilisées en tarification 39
42figure.3.2
3.3 Comparaison entre la distance euclidienne et la distance dynamique DTW (github.com/DBA) 43
3.4 Comparaison entre la moyenne arithmétique et l'algorithme DBA (*Source : github.com/fpetitjean/DBA*) 44
3.5 Illustration de la méthode *binary segmentation* 48
3.6 Evolutions des cinq pièces suivies de la marque Peugeot (en pointillés) et de la marque obtenue à partir des cinq autres (en noir) 50
3.7 Evolutions des coûts des pièces des marques agrégées par marque 50
3.8 Schéma du traitement des courbes avant la classification non supervisée . 51
3.9 K-moyennes des courbes de Stellantis avec 2 clusters 52
3.10 K-moyennes des courbes de Stellantis avec trois clusters 52
3.11 SSE des k-moyennes en fonctions de k 53
3.12 Clustering K-means des 12 marques retenues avec la distance DTW et cinq classes 54
3.13 Schéma résumé du traitement des courbes avant la détection de rupture . 56
3.14 Evolutions des groupes obtenus en pointillés et courbe finale en noire . . . 56
3.15 Représentation des ruptures repérées grâce à la segmentation binaire sur la courbe différenciée (en haut) et la courbe initiale (en bas), avec les pénalités nulle et de 10 57
3.16 Ruptures obtenues avec la méthode *PELT* sur la courbe différenciée et la courbe normale 58
3.17 Ruptures obtenues avec la méthode *Binary segmentation* avec une pénalité de 10, sur la courbe différenciée et la courbe normale 58
3.18 Schéma résumé du traitement **opérationnel** des courbes avant la détection de rupture 59
3.19 Evolutions des marques obtenus en pointillés et courbe finale opérationnelle en noir 60
3.20 Etude du délai de détection de la rupture en T2 2022 60

4.1 Illustration du dilemme biais-variance dans les modèles de Machine Learning *(Source : Céline ROUARD)* 66
4.2 Exemple fictif d'arbres de régression et de classification 68
4.3 Mécanisme d'une forêt aléatoire 69
4.4 Différence entre le *boosting* et le *bagging* (Source : *Edureka*) 70
4.5 Exemple fictif de discrétisation d'une variable en histogramme 71
4.6 Schéma des erreurs de prédictions (écarts entre $y_i - \hat{y}_i$) 72
73figure.4.7
4.8 Représentation de deux branches de risques grâce auxquelles on peut calculer une évolution du coût de la pièce 79
4.9 Volumes du nombre d'inflations obtenues pour quelques combinatoires . . 80

4.10 Evolution du coût moyen du bouclier avant d'Audi sans traitement avec les limites calculées sur chaque année 82
4.11 Histogramme et qqplot avec une loi normale de la variable cible calculée . 83
4.12 Histogramme des valeurs observées et prédites d'un *histogram gradient boosting* pour 6 modalités (en haut) et 10 modalités (en bas) par variable d'agrégation 84
87figure.4.13
4.14 Coefficients du GLM pénalisé pour les modalités d'anciennetés 88
4.15 Coefficients du GLM pénalisé pour les modalités de classes SRA 88
4.16 Coefficients du GLM pénalisé pour les modalités de groupes SRA 88
4.17 Moyennes des valeurs absolues des mesures *shap* les plus significatives . . 89
4.18 Impact des modalités des variables sur la variable cible avec les valeurs *shap* 89
4.19 Evolution du coût moyen du bouclier avant d'une Citroën et évolution des prix de l'essence 90
4.20 Evolution du coût moyen du bouclier avant d'une Citroën et évolution des indices des prix des métaux 91

5.1 Schéma des années utilisées en tarification 94
5.2 Schéma expliquant l'utilisation des coefficients pour différencier l'hypothèse d'inflation selon les critères de différenciation 94
5.3 Coefficients du GLM pénalisé pour les modalités d'anciennetés 96
5.4 Coefficients du GLM pénalisé pour les modalités de classes SRA 96
5.5 Coefficients du GLM pénalisé pour les modalités de groupes SRA 96
5.6 Différents exemples de dérives du modèle 107
5.7 Histogramme de la variable "inflation" calculée avant et après le premier traitement des valeurs aberrantes pour une des combinatoires 111
5.8 Qqplot normal de la distribution de la variable inflation avant et après le premier traitement des valeurs aberrantes 112
5.9 Evolution du coût moyen du bouclier avant d'Audi avec les limites calculées sur toutes les années confondues 112
5.10 Evolution du coût moyen du bouclier avant d'Audi avec les limites calculées sur chaque année 112

Annexes

Annexe 1 - Glossaire

— Le **BET** (Bureau d'Etudes Techniques) est l'appellation d'une équipe qui réalise des études techniques comme le BET Auto ou le BET Indemnisation. Elles sont respectivement liées aux branches Auto et Indemnisation analysant les résultats techniques, réalisant la tarification ou des indemnisations et nous assistant sur les spécificités techniques.

— **CM** (Coût moyen)

— **DARVA** (Développement d'Applications sur Réseaux à Valeur Ajoutée) est une société spécialisée en échange de données informatisé et de solutions Internet pour les métiers de l'assurance.

— **ERD** (Evaluation et Recyclage des Déchets) correspond aux forfaits attribués au traitement des déchets.

— **Fréquence des sinistres** : Nombre de sinistres divisé par l'exposition (correspondant à la proportion de jours présents sur l'année).

— L'**IARD** (Incendie, Accidents et Risques Divers) est un sigle utilisé dans le domaine des assurances pour désigner les clauses d'un contrat dont l'objet porte sur les dommages et les biens.

— Le format **JSON** (JavaScript Object Notation) est un format de données textuelles léger et structuré. Il est facile à lire ou à écrire pour des humains et également des machines.

— Le **SRA** (Sécurité et Réparations Automobiles) est une association française regroupant les sociétés et mutuelles d'assurance automobile et chargée d'études techniques et de statistiques, qui donne, via son site, des statistiques sur le marché automobile français (*SRA pieces de reemploi.fr*).

— Les éléments **SGC** ou **SRGC** (Sous réserves de Garantie Contractuelle) sont généralement des accessoires ou des équipements qui sont pris en charge par l'assureur selon le type de garantie mais qui ne peuvent pas être intégrés dans le corps du rapport d'expertise, de par leur statut.

— La **VRADE** (Valeur de Remplacement A Dire d'Experts) désigne le prix estimé d'un véhicule par un expert mandaté par l'assureur lorsque celui-ci est non réparable ou issu d'un vol non retrouvé.

Annexe 2 - Modèles de *datadrift*

Dans un premier temps, un modèle de dérive des données a été envisagé pour assurer le maintien de cette détection dans le temps, mais aussi pour détecter les changements en eux-mêmes, en vérifiant la continuité des caractéristiques des données utilisées.

Définition d'un *Model drift* et ses causes

On distingue deux types de *Model drift*[3] :

— la dérive des données, autrement appelée *datadrift*, qui correspond à un changement des données utilisées en entraînement par rapport aux données en production, qui peut mener à une baisse de la performance du modèle.

— et la dérive conceptuelle du modèle, autrement le *concept drift* (voir *figure 5.6*) qui peut se traduire par un changement naturel des **comportements des agents économiques** qui peuvent évoluer avec le temps, ou alors par l'**objectif** du modèle qui n'est plus adapté.

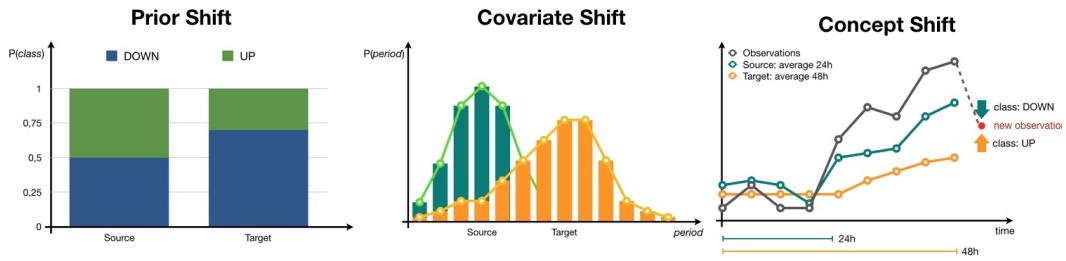


FIGURE 5.6 – Différents exemples de dérives du modèle

Ces phénomènes peuvent avoir des impacts très importants puisque la réponse du modèle créé peut s'avérer erronée à cause de données qui ont changé silencieusement. Il est alors très important de s'assurer de la cohérence du modèle et des données utilisées dans celui-ci pour éviter la prise de décision basée sur un résultat qui n'est plus adapté.

La dérive peut être distinguée de manières différentes[4] :

1. Dans le cas du *data drift* : on rappelle que cette dérive vient d'un changement de données entre celles utilisées en entrée du modèle d'entraînement et celles en entrée du modèle en production.

   On peut le traduire de manière mathématiques comme $P[X|Y] \neq P[X|Y']$ avec $P[X|Y]$ la distribution de probabilité pour des données en entrée Y et de sortie

3. Sources : https://www.free-work.com/fr/tech-it/blog/actualites-informatiques/model-drift-en-machine-learning-explications-et-enjeux, https://www.avisia.fr/news/tribune-expert/gestion-drift-modele-ml/, datascientest.com/definition-data-drift

4. Source : https://www.dataversity.net/data-drift-vs-concept-drift-what-is-the-difference/

X et $P[X|Y']$ la distribution de probabilité pour des données d'entrée différentes Y'. Ainsi il s'agit du modèle entraîné sur deux bases en entrée différentes Y et Y'.

Ainsi les dérives peuvent venir :

- tout d'abord en amont, sur la **distribution de la variable cible** qui n'est plus le même entre les bases d'apprentissage et de test. On appelle ce phénomène le *prior shift*. C'est alors la caractéristique de la variable qui est impactée.
- ensuite, il y a le *covariate drift* qui correspond à des **changements dans les distributions** des variables de la base d'apprentissage et celles de l'échantillon d'application. On peut l'interpréter comme une "mauvaise création d'échantillons d'apprentissage et de test" qui sont très éloignés des données actuelles réelles.
  - Cela peut venir d'un problème dans la création de la base d'apprentissage qui possède un périmètre qui ne correspond pas au public réel sur lequel on veut appliquer le modèle.
    - Par exemple si la base possède beaucoup plus d'hommes que de femmes par rapport à la réalité, alors le résultat sera biaisé si leurs comportements sont différents.
    - Autre exemple, si le modèle a été entraîné sur des données de températures hivernales alors que la saison a un impact sur la variable à expliquer, alors il est difficile d'estimer les températures estivales. Ainsi, les saisonnalités des variables qui peuvent changer en fonction de la période de l'année ou au cours du temps, comme la température ou la durée d'ensoleillement dans la journée peuvent l'être doivent être prises en compte.
    - Autre exemple avec les marques de véhicules qui ont des politiques tarifaires et stratégies marketing différentes en fonction des publics visés. Ainsi, les prix de l'automobile en France ne sont pas les mêmes que dans les autres pays. La dimension géographique doit aussi être regardée.
  - Le *drift* peut également survenir avec **un événement** particulier, comme la crise pandémique du covid-19, qui peut modifier les comportements des agents économiques. Ainsi les données collectées avant cet événement peuvent être très différentes d'après celui-ci.
  - Il peut également venir avec la création ou suppression de modalités pour certaines variables.

2. dans le cas du *concept drift* par un changement dans la relation fonctionnelle entre les données d'entrée et de sortie d'un modèle. Ce modèle fonctionne malgré la dérive de concept mais ne donne plus de résultats exacts.

   Mathématiquement, on a $P_{t_1}[Y|X] \neq P_{t_2}[Y|X]$ avec les prédictions du modèle à deux temps différents t_1 et t_2.

Elle peut aussi être de différents rythmes : soudaine, graduelle, ou encore récurrente [5].

Indicateurs de comparaison des données et tests statistiques

— La divergence de Kullback - Leibler ou "entropie relative" qui calcule la différence entre une distribution de référence qui provient de l'échantillon de test et une autre distribution de probabilités.

— Le test de Kolmogorov - Smirnov qui compare les distributions des échantillons avec celle d'une loi normale.

— La distance de Wasserstein qui, sur un espace mathématique donné, mesure la distance entre deux distributions de probabilités.

— L'indice de stabilité de la population (ISP) qui permet de mesurer, dans un intervalle de temps, l'évolution d'une population entre deux échantillons.

Toutes ces méthodes consistent alors à comparer la distribution des variables à différentes périodes ou à regarder l'importance des variables du modèle de prédiction et celle des variables du modèle de drift. Elles ont l'avantage d'être précises mais l'inconvénient de ne pas être automatisables puisque la comparaison des variables doit se faire unes à unes.

Pour cela, d'autres types de méthodes ont été créées, c'est l'objet du sous-point suivant.

Modèles de Machine Learning

Détail de plusieurs méthodes :

— L'**adversarial validation** est une approche qui teste un individu au hasard sur pour savoir s'il appartient à l'échantillon de test ou d'apprentissage, et si le résultat est proche du hasard, soit 50% l'un ou l'autre, alors il n'y a pas de drift suggéré.

— Data Drift Model, **DDM** est basé sur le taux d'erreur de prédiction comme pour une variable binomiale, c'est à dire que lorsque la prédiction n'est pas adaptée avec la valeur obtenue sur les autres données, alors la prédiction donne 1, 0 sinon. Cette méthode permet de détecter les dérives graduelles, si elles ne sont pas trop lentes à arriver, et les dérives soudaines.

— Early Data Drift Model, **EDDM** est une méthode qui est plus adaptée aux dérives lentes et graduelles. Elle est basée sur le principe PAC selon lequel le taux d'erreur de l'apprenant diminue à mesure que le nombre d'échantillons analysés augmente, tant que la distribution des données est stable.

5. Source : https://www.sciencedirect.com/science/article/pii/S1319157821003062

— Le Fast Hoeffding Drift Detection Model **FHDDM** fonctionne aussi avec une fenêtre glissante qui calcule la probabilité de prédiction correcte, comme pour le DDM et l'EDDM, et, si cette probabilité dépasse un seuil défini par le théorème d'inégalité de Hoeffding, alors il considère qu'il y a une dérive. Cette méthode détecte les dérives soudaines.

— La technique **ADWIN** (Adaptive Windowing) consiste à commencer à faire croître dynamiquement la fenêtre d'étude tant qu'il n'y a pas de changement apparent dans la distribution des données, et à la réduire quand un changement est détecté, tout cela en comparant la moyenne de deux sous-fenêtres utilisées.

Les méthodes DDM, EDDM et FHDDM comparent les réponses du modèle appris précédemment et le nouveau modèle appris avec d'autres données. L'ADWIN compare les distributions des données comme dans les tests statistiques précédemment listés.

Annexe 3 - Modèles d'explication, les différents tests

A 3.1 - Différents traitements des valeurs aberrantes

En regardant l'histogramme de la variable à expliquer (voir *figure 5.7*), une distribution de loi normale semblait apparaître avec une queue lourde à droite à cause de valeurs extrêmes créées par des prix renseignés par les experts qui sont démesurés par rapport à ceux du marché.

Ainsi, des traitements plus ou moins restrictifs ont été appliqués pour permettre d'avoir une inflation plus cohérente avec le marché.

Premier cas de traitement des valeurs aberrantes

a) GLM pénalisé gaussien Pour ce cas, le premier traitement est plus simple que celui de la *partie 2.1.6*, puisqu'il consiste à retirer les inflations qui ne sont pas dans l'intervalle $[mediane - x, mediane + x]$ de telle sorte que la distribution de la loi puisse être normale avec une symétrie à la médiane.

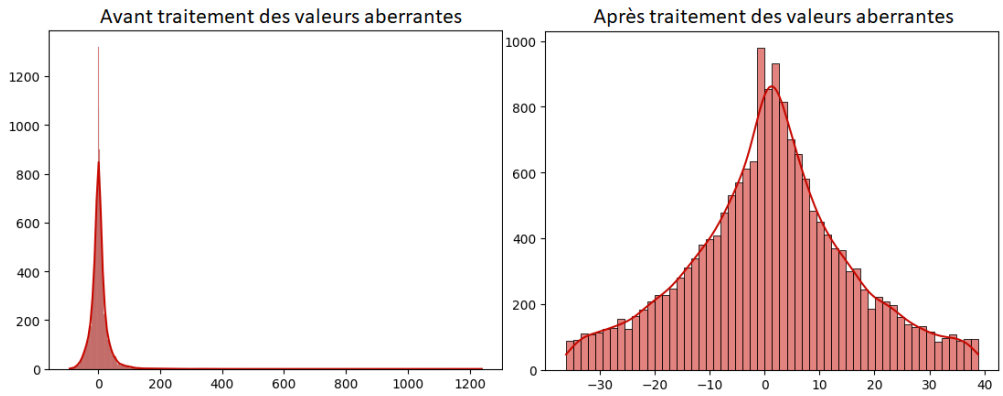


FIGURE 5.7 – Histogramme de la variable "inflation" calculée avant et après le premier traitement des valeurs aberrantes pour une des combinatoires

b) Arbre de décision, Histogram Gradient Boosting (HGB) et Forêt aléatoire Pour appliquer ces modèles, il n'y a pas de conditions sur la loi de distribution de la variable cible, un traitement moins restrictif a été appliqué et ainsi les bases utilisées ont plus de lignes que pour le *glm* précédent. Pour cela, les lignes avec des inflations calculées en dessous du centile 10% et au-dessus du centile 90% ont été retirées pour avoir des valeurs plus acceptables d'inflation.

Second cas du traitement des valeurs aberrantes (plus restrictif)

Dans ce second cas, le traitement, en plus d'être fait sur les centiles, s'est fait en amont des agrégations de risque et sont similaires pour tous les modèles.

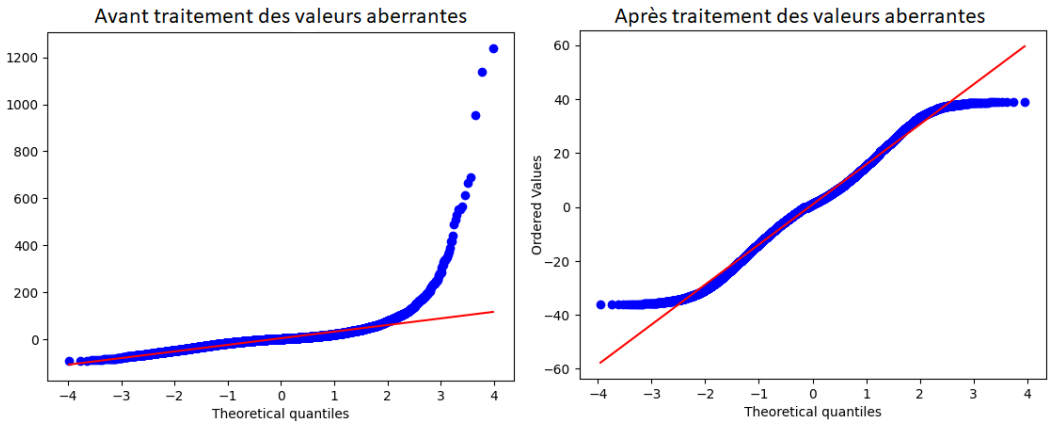


FIGURE 5.8 – Qqplot normal de la distribution de la variable inflation avant et après le premier traitement des valeurs aberrantes

Pour éviter les sauts importants d'un prix à un autre sur les coûts moyens des pièces, seule la pièce la plus importante sur les cinq pièces suivies a été gardée, il s'agit du *bouclier avant*.

De plus, les pièces qui avaient des prix en dessous du centile 10% des pièces de la même marque et de la même année ou au-dessus du centile 90% ont été retirées. Une attention particulière a été faite sur la temporalité puisque sans prendre en compte cette condition les pièces de prix forts n'étaient pas considérés.

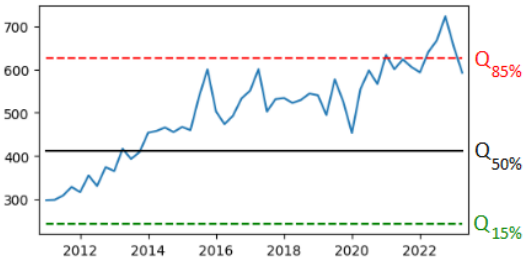


FIGURE 5.9 – Evolution du coût moyen du bouclier avant d'Audi avec les limites calculées sur toutes les années confondues

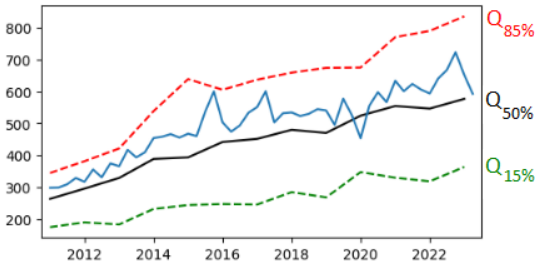


FIGURE 5.10 – Evolution du coût moyen du bouclier avant d'Audi avec les limites calculées sur chaque année

Dans le tableau suivant, les coûts médians des boucliers avant des différentes marque pour les périodes annuelles ou historique sont représentées dans le tableau suivant :

Marque	2011	2012	...	2016	2017	...	2022	2023	2011 → 2023
Autre	263,2	295,5	...	441,5	451,4	...	546,1	576,6	317,09
Audi	281,6	283,0	...	300,0	310,4	...	376,4	404,6	412,66
BMW	375.6	381.0	...	437.1	450.1	...	520.9	519.5	443,95
Citroën	285,0	286,4	...	312,6	317,5	...	436,3	436,4	306,275
Fiat	238,6	248,9	...	247,4	257,1	...	401,3	401,3	266,74
Ford	243,7	276,9	...	301,5	315,8	...	368,4	373,1	301,52
Mercedes	411,0	407,2	...	434,4	478,1	...	492,7	522,1	465,82
Nissan	259,4	228,7	...	335,2	361,0	...	520,9	506,3	347,15
Opel	342,5	353,0	...	384,0	386,0	...	513,6	574,0	384,5
Peugeot	301,5	309,7	...	320,2	334,2	...	475,1	515,8	335,12
Renault	271.,	282,1	...	245,7	248,1	...	323,0	328,6	273,13
Toyota	225,8	248,1	...	255,9	261,3	...	304,8	336,2	265,38
Volkswagen	260,4	266,6	...	296,5	311,6	...	443,8	455,1	314,16

Ensuite après les agrégations comme pour l'autre base, les valeurs avant le centile 10% et après le centile 90% ont été retirées. Dans ce cas on obtient également une distribution normale.

Ainsi, les valeurs les plus fortes de l'inflation était alors plus ou moins 15-20% dans ce second cas alors que dans le premier c'était plus ou moins 30-40% en fonction des agrégations réalisées.

A 3.2 - Tableaux des modalités des variables d'agrégation

Pour les variables d'agrégation avec **6 modalités** pour l'ancienneté du véhicule à la date du rapport, les classes de prix SRA et les groupes de puissance SRA, et une temporalité **semestrielle**.

Les modalités regroupées, en prenant en compte la "normalisation" des segmentations par marques pour les classes et groupes SRA, sont les suivantes :

Variable	G1	G2	G3	G4	G5	G6
Âge	$\leq$ 2 ans	$\in]2,3]$	$\in]3,5]$	$\in]5,7]$	$\in]7,10]$	> 10

Variable	Marque	G1	G2	G3	G4	G5	G6
	Renault	A $\rightarrow$ F	G $\rightarrow$ H	I	J	K $\rightarrow$ L	M $\rightarrow$ S
	Peugeot	A $\rightarrow$ G	H	I	J	K $\rightarrow$ L	M $\rightarrow$ S
	Citroën	A $\rightarrow$ G	H	I	J	K $\rightarrow$ L	M $\rightarrow$ S
	Volkswagen	A $\rightarrow$ G	H $\rightarrow$ I	J	K $\rightarrow$ L	M	N $\rightarrow$ V
	Mercedes	G $\rightarrow$ K	L $\rightarrow$ M	N	O	P $\rightarrow$ Q	R $\rightarrow$ W
Classe	Audi	H $\rightarrow$ L	M	N	O $\rightarrow$ P	Q $\rightarrow$ R	S $\rightarrow$ X
SRA	Ford	B $\rightarrow$ F	G	H $\rightarrow$ I	J $\rightarrow$ K	L	M $\rightarrow$ T
	BMW	I $\rightarrow$ M	N	O	P $\rightarrow$ Q	R	S $\rightarrow$ X
	Fiat	A $\rightarrow$ E	F	G	H	I $\rightarrow$ J	K $\rightarrow$ R
	Toyota	C $\rightarrow$ F	G $\rightarrow$ H	I $\rightarrow$ K	L	M	N $\rightarrow$ V
	Opel	B $\rightarrow$ F	G	H $\rightarrow$ I	J	K $\rightarrow$ L	M $\rightarrow$ R
	Nissan	B $\rightarrow$ G	H $\rightarrow$ I	J $\rightarrow$ K	L	M	N $\rightarrow$ V
	Autres	A $\rightarrow$ E	F $\rightarrow$ G	H $\rightarrow$ J	K $\rightarrow$ L	M $\rightarrow$ N	O $\rightarrow$ ZA

Variable	Marque	G1	G2	G3	G4	G5	G6
	Renault	≤ 28	$\in]28,29]$	$\in]29,30]$	$\in]30,31]$	$\in]31,32]$	> 32
	Peugeot	≤ 29	$\in]29,30]$	$\in]30,30]$	$\in]30,31]$	$\in]31,32]$	> 32
	Citroën	≤ 28	$\in]28,29]$	$\in]29,30]$	$\in]30,31]$	$\in]31,32]$	> 32
	Volkswagen	≤ 29	$\in]29,30]$	$\in]30,31]$	$\in]31,32]$	$\in]32,33]$	> 33
	Mercedes	≤ 31	$\in]31,32]$	$\in]32,33]$	$\in]33,35]$	$\in]35,36]$	> 36
Groupe	Audi	≤ 32	$\in]32,33]$	$\in]33,34]$	$\in]34,35]$	$\in]35,37]$	> 37
SRA	Ford	≤ 28	$\in]28,29]$	$\in]29,30]$	$\in]30,31]$	$\in]31,32]$	> 32
	BMW	≤ 33	$\in]33,34]$	$\in]34,35]$	$\in]35,36]$	$\in]36,38]$	> 38
	Fiat	≤ 28	$\in]28,28]$	$\in]28,29]$	$\in]29,30]$	$\in]30,31]$	> 31
	Toyota	≤ 28	$\in]28,29]$	$\in]29,30]$	$\in]30,32]$	$\in]32,32]$	> 32
	Opel	≤ 29	$\in]29,29]$	$\in]29,30]$	$\in]30,31]$	$\in]31,32]$	> 32
	Nissan	≤ 29	$\in]29,30]$	$\in]30,30]$	$\in]30,31]$	$\in]31,32]$	> 32
	Autres	≤ 29	$\in]29,30]$	$\in]30,31]$	$\in]31,32]$	$\in]32,34]$	> 34

A 3.3 - Encodage des variables

Il existe plusieurs façons d'encoder les variables dans les processeurs des modèles, celles-ci dépendent des types de variables.

Il existe 3 types de variables :

1. les variables **numériques** qui sont des variables qui prennent leurs valeurs discrètes ou continues dans $\mathbb{R}$ ou $\mathbb{N}$,
2. les variables qualitatives qui peuvent contenir un ordre particulier, par exemple la variable ancienneté du véhicule avec 6 modalités allant des âges les plus faibles aux âges les plus forts. Ainsi les modalités peuvent être ordonnés. Ces variables sont appelées **ordinales**,
3. et les variables qualitatives sans relation d'ordre comme la marque du véhicule. On ne peut pas dire que Audi va avant Renault par exemple, cela n'a pas de sens. Ces variables sont appelées **nominales**.

En fonction de ces catégories, les variables ne sont pas encodées de la même manière. Ici, deux fonctions d'encodage ont été utilisées pour les variables qualitatives :

— l'*ordinal encoder* pour coder les variables ordinales en y associant des numéros ordonnés par l'ordre indiqué. Par exemple, pour la variable ancienneté, les âges peuvent être classés du plus faible au plus fort comme ceci :

Variable	Modalité	Encodage ordinal
Age	Faible	1
	Moyen	2
	Fort	3

— le *one hot encoder* pour coder les variables nominales qui ne peuvent pas être ordonnées. Pour cela, pour chaque modalité de la variable, une variable "indicatrice" de cette modalité est créée. Par exemple pour ces 3 marques, on peut le résumer comme cela :

		Encodage One Hot		
Variable	Modalité	Marque_Renault	Marque_Audi	Marque_BMW
Marque	Renault	1	0	0
	Audi	0	1	0
	BMW	0	0	1

Bibliographie

[1] Pierre AILLIOT. Cours de modèles linéaires et de séries temporelles, 2020. [Années académiques 2020 à 2023].

[2] Vroomly Ariane. Pièce de réemploi : définition et législation. https://www.vroomly.com/blog/piece-de-reemploi-definition-et-legislation/, 2022. [En ligne ; Mis à jour le 01 juin 2022].

[3] France Assureurs. Changement climatique : quel impact sur l'assurance à l'horizon 2050. https://www.franceassureurs.fr/lassurance-protege-finance-et-emploie/lassurance-protege/actualites-protege/changement-climatique-quel-impact-sur-lassurance-a-lhorizon-2050/, 2021. [Publié le 28/10/2021].

[4] Léo BREIMAN. Random forests, 2001. [Machine learning 45, 5-32].

[5] Nicolas VAYATIS Charles TRUONG, Laurent OURDRE. Selective review of offline change point detection methods. signal processing, 2020. [Signal Processing, 167 :107299].

[6] Banque de France. Inflation et déflation. https://abc-economie.banque-france.fr/inflation-et-deflation, 2023. [Mis à jour le 17/04/2023].

[7] Arthur E. HOERL et Robert W. KENNARD. Ridge regression : Biased estimation for nonorthogonal problems, 1970. [Technometrics, 12 :1, 55-67, DOI].

[8] John A. Nelder et Robert W. M. WEDDERBURN. Generalized linear models, journal of the royal statistical society, 1972. [Series a (general), 135(3), 370-384].

[9] SRA (Sécurité et Réparations Automobiles). Etude sur les pièces de réemploi. https://www.sra.asso.fr/statistiques/Communication, 2022.

[10] Pierre GANCARSKI François PETITJEAN, Alain KETTERLIN. A global averaging method for dynamic time warping with applications to clustering, 2011. [Pattern Recognition 44 (2011) 678-693].

[11] Sylvain GAUTHIER. L'inflation touche aussi les pièces détachées, quelles solutions pour y remédier ? https://www.auto-moto.com/en-bref/linflation-touche-aussi-les-pieces-detachees-quelles-solutions-pour-y-remedier-16979, 2023. [Publiée le 20/06/2023].

[12] Thomas FINLEY Taifeng WANG Wei CHEN Weidong MA Qiwei YE et Tie-Yan LIU Guolin KE, Qi MENG. Lightgbm : A highly efficient gradient boosting decision tree, 2017.

[13] Flore HARLE. Détection de ruptures multiples dans des séries temporelles multivariées : application à l'inférence de réseaux de dépendance. https://theses.hal.science/tel-01382051/document, 2016.

[14] Julien HAY. Cours de macroéconomie, 2020. [Année académique 2020-2021].

[15] INSEE. Cours des matières premières importées : Pétrole brut brent (londres), prix en dollars us par baril. https://www.insee.fr/fr/statistiques/serie/010002077, 2023. [Identifiant 010002077, mise à jour le 23/06/2023].

[16] INSEE. Cours des monnaies en euros - dollar américain fixing (paris). https://www.insee.fr/fr/statistiques/serie/010002053, 2023. [Identifiant 010002053, mise à jour le 23/06/2023].

[17] INSEE. Indice de prix de production de l'industrie française pour le marché français (cpf 23.1) du verre et articles en verre. https://www.insee.fr/fr/statistiques/serie/010534630, 2023. [Identifiant 010534630, prix de marché en base 2015 avec les données mensuelles brutes, mise à jour le 30/06/2023].

[18] INSEE. Indice de prix de production de l'industrie française pour le marché français (cpf 24.42) de l'aluminium. https://www.insee.fr/fr/statistiques/serie/010534272, 2023. [Identifiant 010534272, Prix de base en base 2015 avec des données mensuelles brutes, mise à jour le 30/06/2023].

[19] INSEE. Indice de prix de production de l'industrie française pour le marché français (cpf 24.44) du cuivre. https://www.insee.fr/fr/statistiques/serie/010534659, 2023. [Identifiant 010534659, Prix de marché en base 2015 avec des données mensuelles brutes, mise à jour le 30/06/2023].

[20] INSEE. Indice de prix de production de l'industrie française pour le marché français (cpf 25.1) des éléments en métal pour la construction. https://www.insee.fr/fr/statistiques/serie/010534285, 2023. [Identifiant 010534285, Prix de base en base 2015 avec les données mensuelles brutes, mise à jour le 30/06/2023].

[21] INSEE. Indice de prix de production de l'industrie française pour l'ensemble des marchés (cpf 24.3) des autres produits de première transformation de l'acier. https://www.insee.fr/fr/statistiques/serie/010535488, 2023. [Identifiant 010535488, Prix de base en base 2015 avec des données mensuelles brutes, mise à jour le 30/06/2023].

[22] INSEE. Indice des prix à la consommation - base 2015 - glissement annuel - ensemble des ménages - france métropolitaine - ensemble. https://www.insee.fr/fr/statistiques/serie/001768594, 2023. [Identifiant 001768594, mise à jour le 15/06/2023].

[23] INSEE. Indice du coût du travail (salaires et charges) de l'industrie manufacturière (naf rév. 2 section c). https://www.insee.fr/fr/statistiques/serie/010599838, 2023. [Identifiant 010599838, indice en base 2016].

[24] INSEE. Prix moyens mensuels de vente au détail en métropole du supercarburant sans plomb, indice d'octane 98 (1 litre). https://www.insee.fr/fr/statistiques/serie/000442589, 2023. [Identifiant 000442589, mise à jour le 15/06/2023].

[25] Chloé JIMENEZ. Cours de microéconomie, 2020. [Année académique 2020-2021].

[26] Anna KEREBEL. Analyse de séries temporelles paléo-écologiques, 2021. [Rapport de stage avec l'Université de Bretagne Occidentale].

[27] LégiFrance. Décret n° 2016-703 du 30 mai 2016 relatif à l'utilisation de pièces de rechange automobiles issues de l'économie circulaire. https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000032610837, 2016. [En ligne ; version du 30 mai 2016].

[28] Richard OLSHEN et Charles J. STONE Léo BREIMAN, Jerome H. FRIEDMAN. Cart : Classification and regression trees, 1984. [Wadsworth International].

[29] Kévin MILIN. Modélisation de l'inflation en france par une approche macrosectorielle. https://www.insee.fr/fr/statistiques/3202498, 2017. [Publié le 21/11/2017 par l'INSEE].

[30] Egon S. PEARSON Ralph B. D'AGOSTINO. An omnibus test of normality for moderate and large sample size, 1971. [Biometrika, 58, 341-348].

[31] scikit-learn developers (BSD LICENSE). Histogram-based gradient boosting regression tree. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingRegressor.html, 2021.

[32] Hugh CHEN et al. Scott .M. LUNDBERG, Gabriel ERION. From local explanations to global understanding with explainable ai for trees. https://www.insee.fr/fr/statistiques/serie/010599838, 2020. [Nature Machine Intelligence 2, 56-67].

[33] Franck VERMET. Cours d'arbres de décisions et de régression linéaire, 2020. [Années académiques 2020 à 2023].