Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 09/11/2023

Par : **Daniel NKAMENI**

Titre : **Etude de la stabilité d'un modèle de segmentation en assurance-crédit**

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

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

Membres présents du jury de la filière

Nom : Olivier LOPEZ

Membres présents du jury de l'Institut des Actuaires

Entreprise : SIAPARTNERS

Nom : Ronan DAVIT
Signature :

Directeurs de mémoire en entreprise :

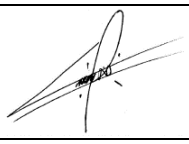
Nom : Julien LEBLOA
Signature :

Nom : Slim SAANOUNI
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



Secrétariat :

Bibliothèque :

Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
5, avenue Henry Le Chatelier - 91120 PALAISEAU, FRANCE

Résumé

Comme tout type d'assurance, la viabilité de l'activité d'assurance-crédit repose sur la modélisation de la sinistralité, indispensable pour anticiper les pertes futures. Cette démarche aboutit au calcul du capital règlementaire, le SCR, assurant la solvabilité de l'entreprise d'assurance à 99,5 % sur un an.

Au cœur de cette modélisation, la segmentation joue un rôle clé en créant des groupes homogènes d'observations. Son utilité est incontestée, mais l'ACPR s'interroge sur la stabilité des modèles de segmentation face aux petites fluctuations des données d'entraînement. Malheureusement, la littérature n'offre pas de méthodologie complète pour évaluer cette stabilité en assurance-crédit.

Ce mémoire vise à combler ce vide en proposant une méthodologie claire et opérationnelle pour analyser la stabilité des modèles de segmentation. Il se veut un guide méthodologique applicable à d'autres branches de l'assurance et de la banque.

Notre démarche débute par une définition de la stabilité, suivie de la présentation de notre méthodologie. Nous étudions ensuite trois modèles de segmentation - CART, k-prototypes et CAH - à travers des analyses graphiques, des variations par rapport aux scénarios de référence, des études de pentes et des tests statistiques. Il en ressort que le modèle CART est le plus stable, et que la stabilité est influencée par la taille de l'échantillon, les interactions entre les variables et la nature de ces dernières. Nous terminons nos travaux en mettant en évidence que l'étude et l'amélioration de la stabilité des modèles de segmentation permettent d'obtenir des SCR de souscription plus robustes en assurance-crédit.

Ce mémoire se situe à la frontière entre la recherche et l'opérationnel, en fournissant une méthodologie détaillée et reproductible pour évaluer la stabilité des modèles de segmentation en assurance-crédit et au-delà.

Mots-clés : Stabilité, Segmentation, Assurance-crédit, CART, k-prototypes, CAH.

Abstract

Like any type of insurance, the viability of credit insurance activity relies on modeling loss events, which is essential for anticipating future losses. This process leads to the calculation of the regulatory capital, known as SCR (Solvency Capital Requirement), ensuring the solvency of the insurance company at a 99.5% confidence level over one year.

At the core of this modeling process, segmentation plays a key role by creating homogeneous groups of observations. Its utility is undisputed, but the ACPR (Prudential Supervision and Resolution Authority) questions the stability of segmentation models in the face of small fluctuations in training data. Unfortunately, the literature does not provide a comprehensive methodology for assessing this stability in credit insurance.

This thesis aims to fill this gap by offering a clear and operational methodology for analyzing the stability of segmentation models. It aims to serve as a methodological guide applicable to other branches of insurance and banking.

Our approach begins with a definition of stability, followed by the presentation of our methodology. We then study three segmentation models - CART, k-prototypes, and HAC - through graphical analyses, variations compared to reference scenarios, slope studies, and statistical tests. It emerges that the CART model is the most stable, and stability is influenced by sample size, interactions between variables, and the nature of these variables. We conclude our work by highlighting that studying and improving the stability of segmentation models lead to more robust underwriting SCR calculations in credit-insurance.

This thesis lies at the intersection of research and operational aspects, providing a detailed and replicable methodology for assessing the stability of segmentation models in credit insurance and beyond.

Keywords: Stability, Segmentation, Credit Insurance, CART, k-prototypes, CAH.

Note de synthèse

1. Contexte et introduction

L'assurance-crédit couvre les risques liés aux clients (aussi appelés acheteurs) des assurés, offrant une indemnisation en cas de non-remboursement des créances commerciales détenues sur ces acheteurs (Ndoye, 2019). La directive Solvabilité 2 définit un capital de solvabilité requis (SCR) pour les entreprises d'assurance, calculé soit selon une formule standard de l'Autorité de contrôle prudentiel et de résolution (ACPR), soit via un modèle interne adapté au portefeuille de l'assureur. La Compagnie française d'assurance pour le commerce extérieur (Coface) utilise un modèle interne partiel (combinaison de formule standard et de modèle interne) pour le calcul de son SCR de souscription non-vie en raison de la nature spécifique de l'assurance-crédit.

Le calcul de ce SCR repose sur la modélisation de la sinistralité, estimant les pertes nettes de recouvrement sur un an. Cette modélisation inclut des simulations des probabilités de défaut (PD), des proportions d'utilisation des garanties (UGD), des spécificités contractuelles (CS), et des taux de perte (LGD) des acheteurs, aboutissant à la perte ultime en cas de défaut.

$$Perte\ ultime = Expo \times UGD \times CS \times LGD$$

Où $Expo$ est l'exposition sur un acheteur. Elle correspond au montant maximal des factures que l'assureur-crédit prendra en charge pour un assuré vis-à-vis d'un acheteur (avant quotité garantie, franchise ou réassurance).

Les simulations susmentionnées sont effectuées grâce à des lois de probabilité calibrées sur les portefeuilles d'acheteurs. Elles servent par la suite à construire une distribution du résultat technique dont le quantile à 0,5% correspond au SCR de souscription.

$$SCR_{Souscription\ Assurance-crédit} = -VaR_{0,5\%}(Résultat\ Technique)$$

Pour garantir une meilleure qualité des résultats et une adéquation avec la réalité, les calibrages de ces lois de probabilité doivent être effectués sur des groupes de données homogènes. La segmentation, une technique couramment utilisée en analyse de données statistiques, permet de construire ces groupes homogènes.

2. Problématique et objectifs de l'étude

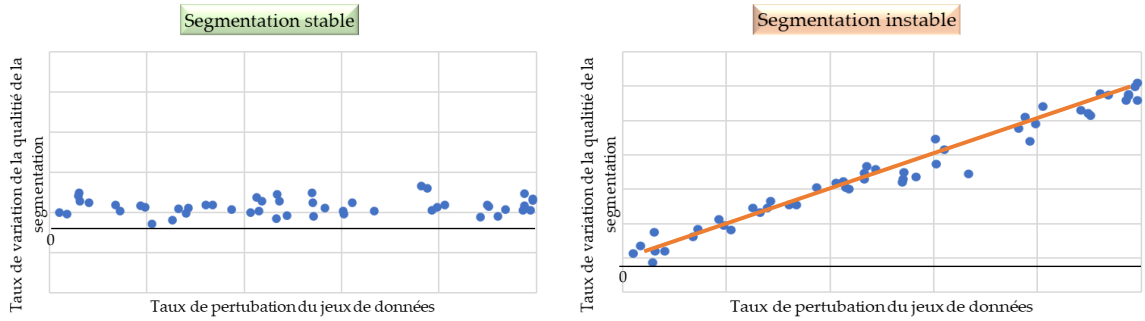
L'ACPR, ainsi que certaines études statistiques telles que (Kassambara, 2018) et (Amit, 2017), se posent des questions concernant la stabilité des modèles CART (Classification and Regression Trees) qui sont actuellement utilisés par la Coface pour effectuer les segmentations dans son modèle interne partiel. En effet, le régulateur se demande si

les structures de segmentation issues des modèles CART restent robustes face à de légères perturbations dans les données utilisées pour les construire. Malheureusement, bien que la segmentation soit largement utilisée, la littérature scientifique ne propose pas de méthodologie complète pour l'étude de la stabilité d'un modèle de segmentation en assurance-crédit.

L'objectif de ce mémoire est de proposer une méthodologie claire et détaillée pour l'étude et la validation de la stabilité d'un modèle de segmentation dans le cadre de l'assurance-crédit.

3. Définition de la stabilité

D'après Liu, et al. (2022), un algorithme de segmentation est considéré comme stable s'il fournit des segmentations très similaires, voire identiques, lorsqu'il est appliqué à de petites perturbations du jeu de données initial. Les figures ci-dessous illustrent le concept de stabilité. Notons que les taux utilisés dans cette analyse sont en valeurs absolues.



Soit m la pente de la courbe linéaire représentée sur la deuxième figure et c son ordonnée à l'origine, de telle sorte que l'équation de cette courbe soit $y = mx + c$. Pour quantifier cette définition de la stabilité dans le cadre de nos analyses, nous proposons la formule suivante pour un coefficient d'instabilité :

$$Coef\_Instab = \frac{|m| + (100 \times |c|)}{2}$$

Un modèle stable afficherait des valeurs presque identiques de la mesure de qualité de la segmentation, quel que soit le niveau de perturbation des données. Cela se traduit par une courbe linéaire horizontale ($m = 0$) et proche de l'axe des abscisses ($c = 0$), ce qui donne un $Coef\_Instab = 0$ pour un modèle parfaitement stable. Plus le modèle est instable, plus la valeur du $Coef\_Instab$ est élevée. Il convient de noter que ce coefficient est simple à calculer et facile à interpréter, ce qui en facilite l'adoption et l'utilisation dans les travaux actuariels.

D'après Homa, et al. (2020), ne pas étudier la stabilité d'un modèle de segmentation revient à prendre le risque que la segmentation obtenue soit contextuelle

plutôt que structurelle. En assurance-crédit, utiliser un modèle de segmentation instable lors du calcul du besoin en capital expose l'entreprise d'assurance au risque de sous-estimer ou de surestimer ce besoin.

4. Données de l'étude

Dans la suite, nous concentrerons nos analyses sur la probabilité de défaut (PD) car, la méthodologie utilisée pour évaluer la stabilité des modèles de segmentation liés à la PD peut être facilement appliquée aux trois autres phénomènes mentionnés précédemment, sans requérir d'adaptations ou d'ajustements spécifiques.

La Coface nous a fourni les données de son activité d'assurance-crédit couvrant la période de 2007 à 2022. La période de calibrage s'étend de 2007 à 2021 et inclut un total de 24 388 428 acheteurs. Le calcul du SCR de souscription sera effectué à partir des données du quatrième trimestre de l'année 2022. Chaque acheteur est caractérisé par plusieurs informations, notamment son année de rattachement à la Coface, sa probabilité de défaut historique, la présence d'un défaut au cours de l'année de rattachement, sa tranche d'exposition, la région de l'entité Coface à laquelle il est rattaché, la cible économique de son activité, son secteur d'activité et sa tranche de notation.

5. Méthodologie et résultats de l'étude

La méthodologie d'analyse de la stabilité d'un modèle de segmentation, développée dans le cadre de cette étude conformément aux recommandations de la littérature, peut être résumée par les étapes suivantes. Pour chaque étape, les résultats obtenus après son implémentation sont également présentés.

5.1. Choix des modèles de segmentation candidats

À partir des travaux empiriques réalisés dans le cadre de la segmentation, nous avons choisi trois modèles candidats : le modèle CART, le modèle de k-prototypes et le modèle de classification ascendante hiérarchique (CAH).

5.2. Sélection d'une métrique pour évaluer la qualité d'une segmentation

Nous avons choisi d'utiliser un indice de Gini modifié pour tenir compte des spécificités de l'assurance-crédit. Cet indice varie de 0 à 1, et plus il est élevé, meilleure est la segmentation.

5.3. Sélection d'une métrique reflétant la structure d'un sous-échantillon

Nous utilisons la proportion de défaut d'un sous-échantillon pour représenter sa structure. Le taux de perturbation dans un sous-échantillon est alors mesuré par la variation de sa proportion de défaut par rapport à un sous-échantillon de référence.

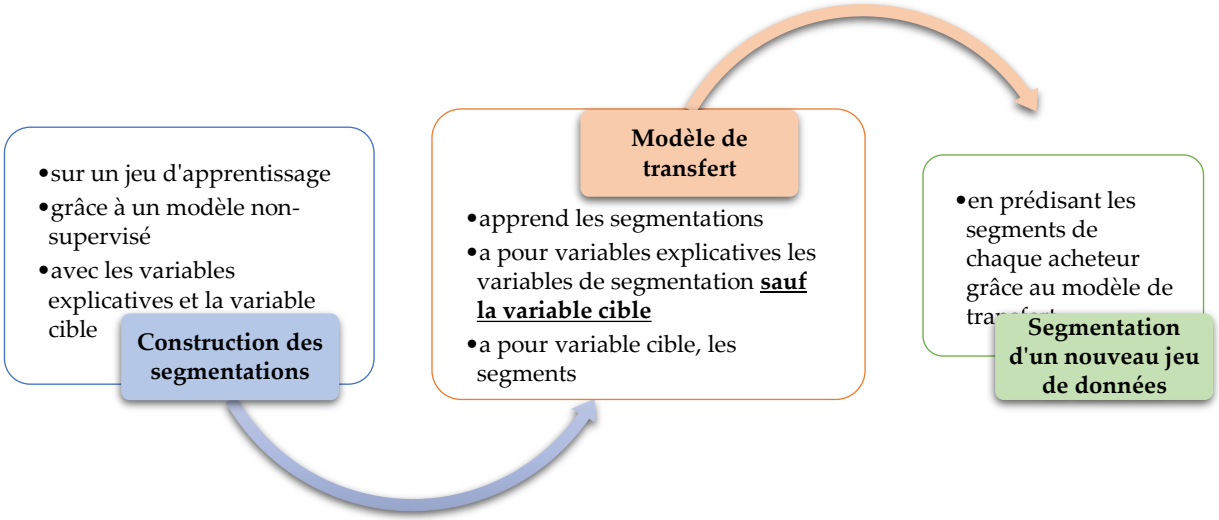
5.4. Division des données en jeu d'apprentissage et en jeu de validation

Nous construisons la base d'apprentissage en effectuant un tirage aléatoire stratifié sans remise de 80 % de la base de données initiale. Les variables de stratification sont l'année de rattachement d'un acheteur et sa zone géographique.

5.5. Calibrage des modèles de segmentation candidats sur les données

Le calibrage d'un modèle candidat s'effectue sur l'ensemble de données d'apprentissage. Il vise à déterminer le nombre optimal de segments ainsi que les meilleurs hyperparamètres du modèle. Le nombre de segments est sélectionné à l'aide de la méthode du coude, tandis que les hyperparamètres sont choisis grâce à la validation croisée. La variable cible utilisée est l'occurrence des défauts.

Contrairement au modèle CART, les modèles k-prototypes et de classification ascendante hiérarchique ne peuvent pas généraliser directement les segmentations apprises d'un jeu de données à un autre. Pour surmonter cette limitation, nous utilisons des modèles de transfert, qui fonctionnent comme suit :



A l'issue de cette phase de calibrage, nous obtenons les modèles optimaux suivants :

Modèles de segmentation	Nombres de segments retenus	Modèles de transfert retenu	Performances des modèles de transfert en validation croisée	
			Taux de bon classement	Sensibilités
CART	18	Transfert assuré par des tables de correspondances		
k-prototypes	6	XGBoost	98,79%	94,88%
CAH	18	XGBoost	97,54%	98,02%

5.6. Création des sous-échantillons d'entraînement à partir du jeu d'apprentissage

Nous avons créé 30 sous-échantillons par tirage aléatoire stratifié sans remise de 60% du jeu de données d'apprentissage initial, introduisant ainsi de légères perturbations

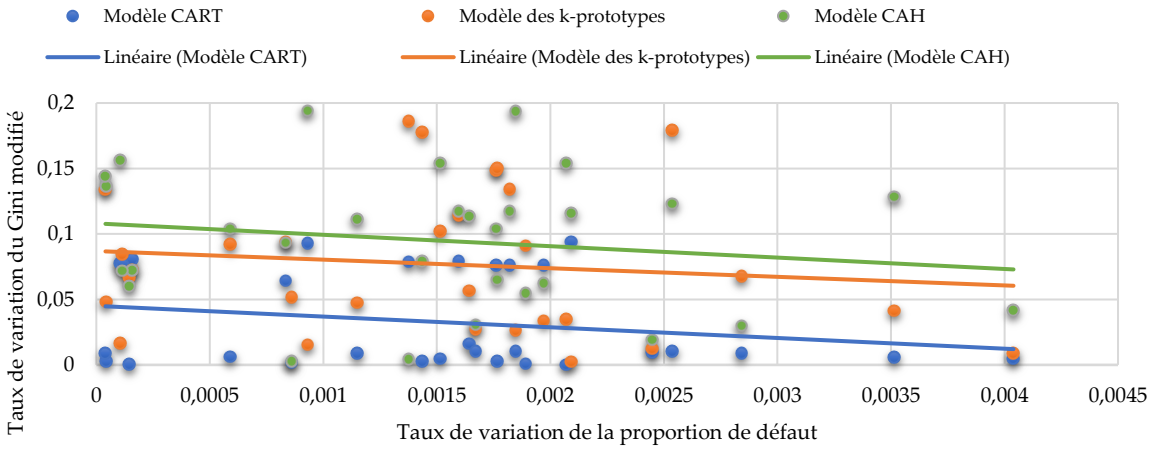
dans les sous-échantillons. Cette méthode est recommandée par Dudoit et al. (2002), Levine et al. (2002).

5.7. Entraînement des modèles candidats sur chaque sous-échantillon et calcul des valeurs de la métrique de qualité sur le jeu de validation

À l'issue de cette étape, nous disposons de 30 valeurs de l'indice de Gini modifié pour chaque modèle candidat.

5.8. Utilisation de visualisations graphiques et d'analyses de pentes pour sélectionner le modèle le plus stable

La figure suivante montre les taux de variation absolue du Gini modifié en fonction des taux de variation absolue de la proportion de défaut pour nos trois modèles. Ces taux de variation sont calculés par rapport à des valeurs de référence pour chaque sous-échantillon.



Nous constatons une grande variabilité dans les taux de variation de l'indice de Gini modifié pour les modèles de k-prototypes et de CAH, tandis que le modèle CART présente des variations de structure de segmentation proches de 0% dans la plupart des cas, indiquant ainsi une plus grande stabilité. Cette observation est corroborée par le fait que le coefficient d'instabilité du modèle CART est le plus bas parmi les trois modèles (6,3411 contre 7,6277 pour le modèle de k-prototypes et 9,7516 pour le modèle CAH).

5.9. Validation de cette sélection statistiquement à l'aide des tests d'adéquation à des lois uniformes

Pour chaque modèle, nous effectuons un test d'adéquation des 30 valeurs de l'indice de Gini modifié à une loi uniforme. Au seuil de 1%, le modèle CART est le seul à rejeter l'hypothèse nulle selon laquelle les valeurs de l'indice de Gini modifié suivent une distribution uniforme. Cela est dû au fait que, contrairement aux deux autres modèles,

les valeurs de l'indice de Gini modifié pour le modèle CART sur le jeu de données de validation sont très similaires. Par conséquent, nous conservons le modèle CART pour la suite en raison de sa meilleure stabilité.

5.10. Identification des facteurs susceptibles d'influencer la stabilité du modèle retenu à l'étape précédente

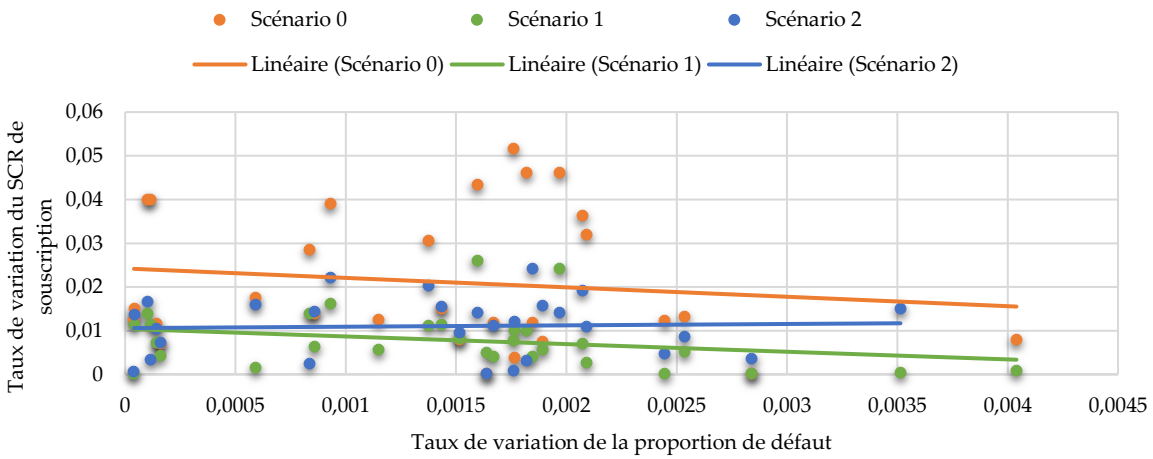
Les potentielles sources d'instabilité identifiées, en s'inspirant de la littérature et de l'analyse descriptive des données, sont les suivantes : des échantillons d'entraînement de petite taille, des variables explicatives ayant un faible pouvoir de prédiction, des variables explicatives fortement corrélées entre elles, ainsi que les prétraitements des données, notamment la discrétisation de la variable cible.

5.11. Prise en compte des sources d'instabilité en calculant le coefficient d'instabilité

Nous avons analysé les impacts que la prise en compte des sources d'instabilité identifiées à l'étape précédente peut avoir sur le SCR de souscription. Pour des raisons de confidentialité, le SCR analysé sera le SCR avant application des mesures correctrices. Considérons les scénarios suivants :

Scénario 0	Scénario 1	Scénario 2
Modèle CART avec toutes les variables explicatives et avec l'occurrence de défaut comme variable cible.	Modèle CART sans les variables «***Cible économique***» et «***Région de l'entité Coface***» et avec l'occurrence de défaut comme variable cible.	Modèle CART sans les variables «***Cible économique***» et «***Région de l'entité Coface***» et avec la probabilité de défaut comme variable cible.

La figure suivante présente les taux de variation du SCR de souscription en fonction des taux de perturbation des données d'entraînement pour chaque scénario :



Dans le scénario 0, d'importantes variations du SCR de souscription sont observées, atteignant parfois plus de 5%, avec une part significative excédant 3%.

Cependant, l'élimination les sources d'instabilité identifiées, notamment en retirant la variable fortement corrélée « *Cible économique* » et la moins prédictive « *Région de l'entité Coface* » dans le scénario 1, le coefficient d'instabilité passe de 6,3413 à 1,5059 et on observe une réduction significative des variations du SCR. Dans ce cas, la plupart des taux de variation du SCR tombent en dessous de 1,6%, et la majorité se situe en dessous de 1%. Le scénario 2, qui utilise la probabilité de défaut comme variable cible affiche un coefficient d'instabilité de 1,0984 et montre également une baisse générale des taux de variation du SCR de souscription.

Par ailleurs, il est important de noter que les variations qui subsistent sont potentiellement dues à d'autres modèles intervenant dans les calculs du modèle interne partiel. En pratique, la Coface effectue des backtests et applique des marges correctrices pour neutraliser ces bruits latents.

En résumé, la robustesse du SCR de souscription en assurance-crédit dépend étroitement de la stabilité des modèles de segmentation utilisés pour son calcul. Grâce à l'analyse de stabilité et à l'élimination des sources d'instabilité, un SCR plus stable et plus robuste a été obtenu.

6. Limites de l'étude

Le cadre de la présente étude est restrictif car elle comporte certaines limites. Parmi celles-ci, on peut citer le nombre restreint de modèles de segmentation et de sources d'instabilité examinés, ainsi que l'utilisation de métriques génériques comme la proportion de défaut pour représenter la structure d'un sous-échantillon. Conduire des analyses autour de ces limites pourrait ouvrir de nouvelles perspectives pour notre étude.

7. Conclusion

Notre étude met en évidence l'importance cruciale de la stabilité des modèles de segmentation, en particulier dans le contexte de l'assurance-crédit. Nous proposons des métriques et des méthodes appropriées pour évaluer cette stabilité, en suivant une méthodologie détaillée basée sur la littérature académique.

En fin de compte, ce mémoire se présente comme une boîte à outils située à la frontière entre la recherche scientifique et l'opérationnel. Il se positionne comme un point de départ pour une meilleure compréhension et une gestion plus précise de la stabilité des modèles de segmentation en assurance-crédit et dans d'autres secteurs où la segmentation est utilisée. Les recommandations découlant de nos résultats sont destinées à guider les professionnels de l'assurance, de la banque et d'autres secteurs vers la construction des modèles de segmentation plus robustes, améliorant ainsi la qualité de leurs analyses.

Executive summary

1. Context and Introduction

Credit insurance covers risks related to insured parties' clients, also referred to as buyers, by providing compensation in case of non-repayment of commercial claims held against these buyers (Ndoye, 2019). The Solvency 2 directive defines a required solvency capital (SCR) for insurance companies, which can be calculated either using a standard formula provided by the Prudential Supervision and Resolution Authority (ACPR) or through an internal model tailored to the insurer's portfolio. The French Export Credit Insurance Company (Coface) uses a partial internal model (a combination of the standard formula and an internal model) to calculate its non-life underwriting SCR due to the specific nature of credit insurance.

The calculation of this SCR relies on modeling credit losses, estimating net recovery losses over one year. This modeling includes simulations of the probabilities of default (PD), usage given defaults (UGD), contractual specificities (CS), and loss given defaults (LGD) of buyers, resulting in estimations ultimate losses in case of default.

$$Ultimate\ loss = Expo \times UGD \times CS \times LGD$$

Where $Expo$ represents the exposure to a buyer, corresponding to the maximum amount of invoices that the credit insurer will cover for an insured party concerning a buyer (before deductible, franchise, or reinsurance).

The aforementioned simulations are performed using probability distributions calibrated based on buyer portfolios. They are subsequently used to construct a distribution of the technical result, with the 0.5th percentile corresponding to the underwriting SCR.

$$SCR_{Credit\ Insurance\ underwriting} = -VaR_{0,5\%}(Technical\ Result)$$

To ensure higher quality results and alignment with reality, the calibration of these probability distributions must be conducted on homogeneous data groups. Segmentation, a commonly used technique in statistical data analysis, allows the construction of these homogeneous groups.

2. Problem Statement and Study Objectives

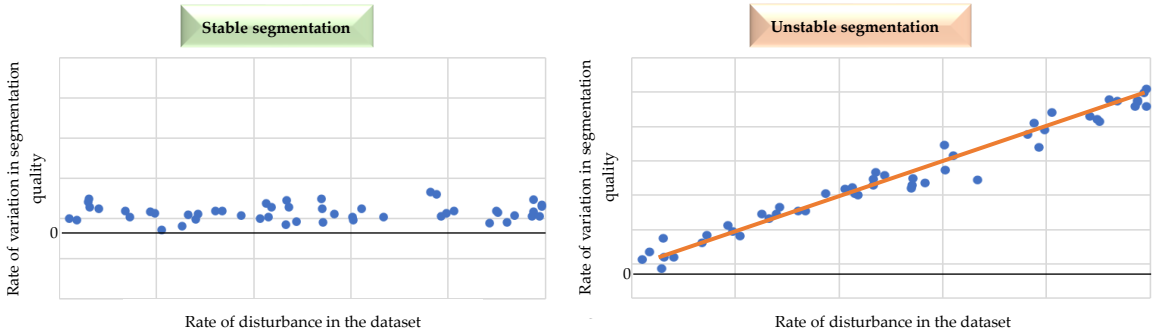
ACPR, along with certain statistical studies such as (Kassambara, 2018) and (Amit, 2017), has raised questions regarding the stability of CART (Classification and Regression Trees) models currently used by Coface for segmentations within its partial

internal model. Indeed, the regulator is concerned about whether the segmentation structures derived from CART models remain robust when faced with minor perturbations in the data used to construct them. Unfortunately, despite the widespread use of segmentation, the scientific literature does not provide a comprehensive methodology for studying the stability of a segmentation model in credit insurance.

The objective of this thesis is to propose a clear and detailed methodology for the study and validation of the stability of a segmentation model in the context of credit insurance.

3. Definition of Stability

According to Liu et al. (2022), a segmentation algorithm is considered stable if it provides highly similar or even identical segmentations when applied to small perturbations of the original dataset. The figure below illustrates the concept of stability. Please note that the rates used in this analysis are in absolute values.



Let m be the slope of the linear curve represented in the second figure, and c its y-intercept, such that the equation of this curve is $y = mx + c$. To quantify this definition of stability in the context of our analyses, we propose the following formula for an instability coefficient:

$$Instability\_Coef = \frac{|m| + (100 \times |c|)}{2}$$

A stable model would exhibit nearly identical values for the segmentation quality measure, regardless of the level of data perturbation. This results in a horizontal linear curve ($m = 0$) close to the x-axis ($c = 0$), resulting in an $Instability\_Coef = 0$ for a perfectly stable model. It is worth noting that this coefficient is easy to calculate and straightforward to interpret, making it suitable for adoption and use in actuarial work.

According to Homa et al. (2020), failing to study the stability of a segmentation model poses the risk that the obtained segmentation is contextual rather than structural. In credit insurance, using an unstable segmentation model when

calculating capital requirements exposes the insurance company to the risk of underestimating or overestimating these requirements.

4. Study Data

In the following sections, we will focus our analysis on the Probability of Default (PD), since the methodology used to assess the stability of segmentation models related to PD can be easily applied to the three other phenomena mentioned earlier without requiring specific adaptations or adjustments.

Coface has provided us with data covering its credit insurance activities from 2007 to 2022. The calibration period spans from 2007 to 2021 and includes a total of 24,388,428 buyers. The calculation of the Underwriting Solvency Capital Requirement (SCR) will be performed using data from the fourth quarter of 2022. Each buyer is characterized by various information, including their year of affiliation with Coface, historical probability of default, occurrence of a default during the affiliation year, exposure category, the region of Coface entity to which they are affiliated, the economic target of their activity, industry sector, and rating category.

5. Methodology and Study Results

The methodology for analyzing the stability of a segmentation model, developed in this study in accordance with the literature's recommendations, can be summarized by the following steps. For each step, the results we obtained after its implementation are also presented.

5.1. Selection of Candidate Segmentation Models

Based on empirical work conducted in the context of segmentation, we have chosen three candidate models: the CART model, the k-prototypes model, and the hierarchical agglomerative clustering (CAH) model.

5.2. Selection of a Metric to Assess Segmentation Quality

We choose to use a modified Gini index to account for the specificities of credit insurance. This index ranges from 0 to 1, and the higher it is, the better the segmentation.

5.3. Selection of a Metric Reflecting Subsample Structure

We use the default proportion within a subsample to represent its structure. The level of perturbation in a subsample is then measured by the variation in its default proportion compared to a reference subsample.

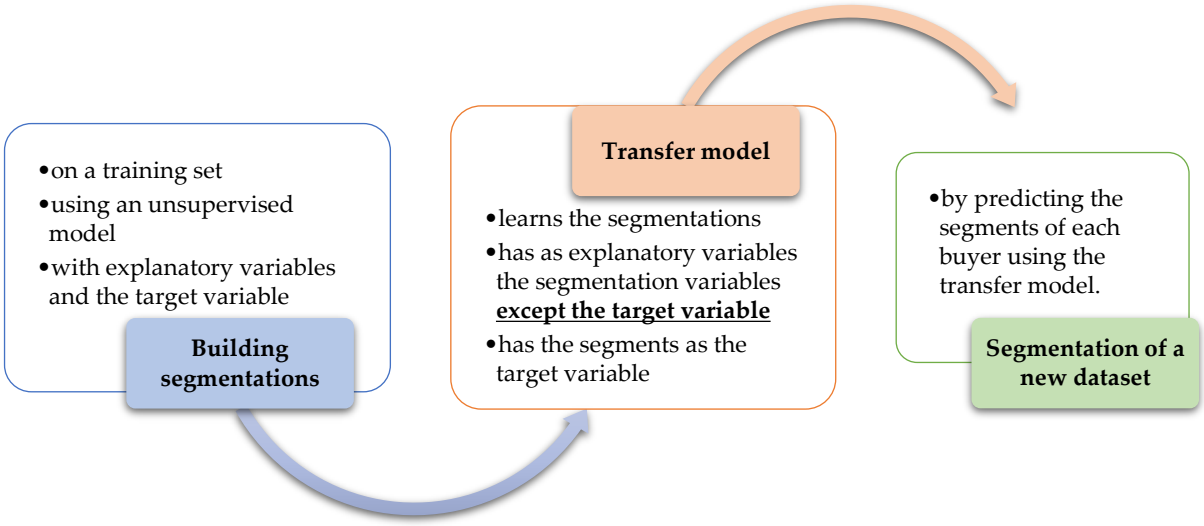
5.4. Splitting Data into Training and Validation Sets

We construct the training dataset by performing a stratified random sampling without replacement of 80% of the initial database. The stratification variables are the buyer's affiliation year and geographical zone.

5.5. Calibration of Candidate Segmentation Models on the Data

Calibrating a candidate model is done on the entire training dataset. It aims to determine the optimal number of segments as well as the best hyperparameters for the model. The number of segments is selected using the elbow method, while hyperparameters are chosen through cross-validation. The target variable used is the occurrence of defaults.

Unlike the CART model, the k-prototypes and hierarchical agglomerative clustering models cannot directly generalize segmentations learned from one dataset to another. To overcome this limitation, we use transfer models, which operate as follows:



At the end of this calibration phase, we obtain the following optimal models:

Segmentation models	Number of retained segments	Retained transfer models	Performance of transfer models in cross-validation.	
			Accuracy	Recall
CART	18	Transfer ensured by correspondence tables.		
k-prototypes	6	XGBoost	98,79%	94,88%
CAH	18	XGBoost	97,54%	98,02%

5.6. Creation of Training Subsamples from the Training Data Set

We generated 30 subsamples through stratified random sampling without replacement, each consisting of 60% of the original training data set, thereby

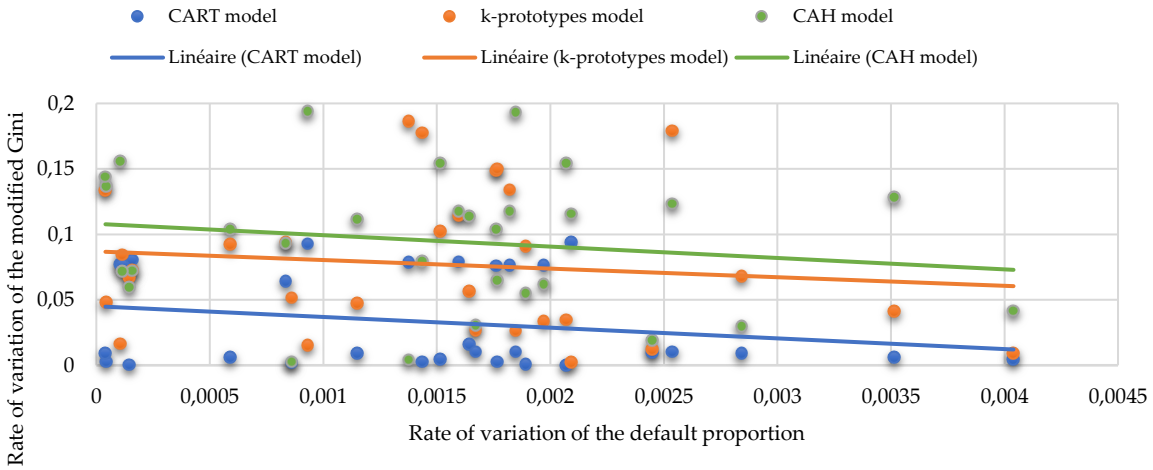
introducing slight perturbations into these subsamples. This method is recommended by Dudoit et al. (2002), Levine et al. (2002).

5.7. Training Candidate Models on Each Subsample and Calculating Quality Metric Values on the Validation Set

At the end of this step, we have 30 values of the modified Gini index for each candidate model.

5.8. Using Graphical Visualizations and Slope Analyses to Select the Most Stable Model

The following figure displays the absolute variation rates of the modified Gini index against the absolute variation rates of the default proportion for our three models. These variation rates are calculated for each subsample relative to reference values.



We observe a significant variability in the rates of variation of the modified Gini index for the k-prototypes and CAH models, while the CART model exhibits segmentation structure variations close to 0% in most cases, indicating greater stability. This observation is supported by the fact that the instability coefficient of the CART model is the lowest among the three models (6.3411 compared to 7.6277 for the k-prototypes model and 9.7516 for the CAH model).

5.9. Statistically Validating This Selection Using Uniformity Tests

For each model, we conduct a goodness-of-fit test for the 30 values of the modified Gini index against a uniform distribution. At the 1% significance level, the CART model is the only one to reject the null hypothesis that the values of the modified Gini index follow a uniform distribution. This is because, unlike the other two models, the values of the modified Gini index for the CART model on the validation data set are very similar. Therefore, we retain the CART model for further analysis due to its better stability.

5.10. Identifying Factors That Could Influence the Stability of the Model Selected in the Previous Step

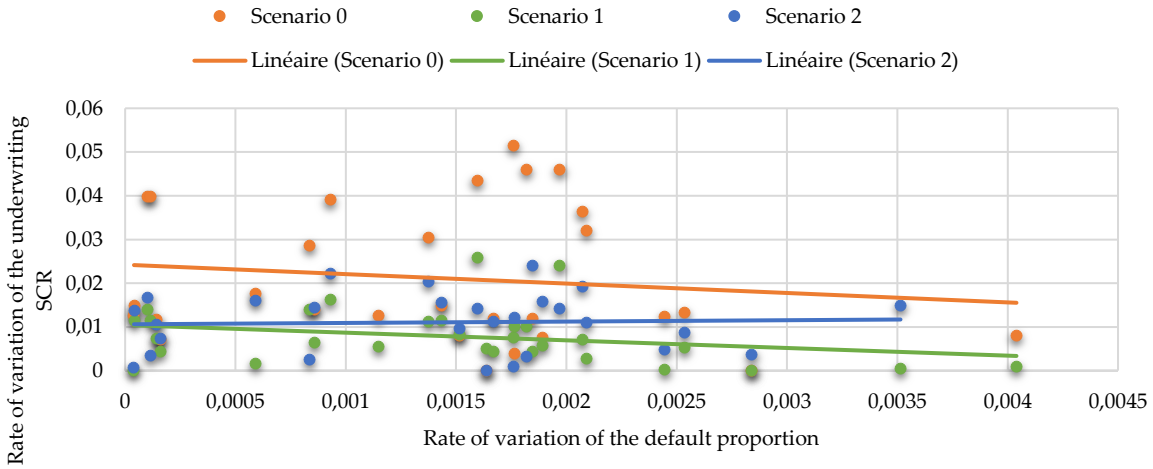
The potential sources of instability identified, drawing from the literature and descriptive data analysis, include the following: small training samples, explanatory variables with low predictive power, highly correlated explanatory variables, and data preprocessing, such as discretization of the target variable.

5.11. Accounting for Sources of Instability by calculating the Instability Coefficient

We analyzed the impacts that accounting for the identified sources of instability in the previous step may have on the underwriting SCR. For confidentiality reasons, the analyzed SCR will be presented without corrective measures. Consider the following scenarios:

Scenario 0	Scenario 1	Scenario 2
CART model with all explanatory variables and with default occurrence as the target variable.	CART model without the "***Economic target***" and "***Coface entity region***" variables and with default occurrence as the target variable.	CART model without the "***Economic target***" and "***Coface entity region***" variables and with probability of default as the target variable.

The following figure presents the rates of variation of the underwriting SCR as a function of rates of perturbation of the training data for each scenario:



In scenario 0, significant variations in the underwriting SCR are observed, sometimes exceeding 5%, with a significant portion exceeding 3%. However, after eliminating identified sources of instability, such as removing the highly correlated variable "*Economic Target*" and the less predictive variable "*Coface Entity Region*" in scenario 1, the instability coefficient decreases from 6.3413 to 1.5059, and a significant reduction in SCR variations is observed. In this case, most SCR variation rates fall below 1.6%, with the majority below 1%. Scenario 2, which uses the default probability

as the target variable, has an instability coefficient of 1.0984 and also shows a general decrease in underwriting SCR variation rates.

Furthermore, it is important to note that the remaining variations may potentially be due to other models involved in the calculations of the partial internal model. In practice, Coface conducts backtests and applies corrective margins to neutralize these latent noises.

In summary, the robustness of the underwriting SCR in credit insurance is closely dependent on the stability of the segmentation models used in its calculation. Through stability analysis and the elimination of instability sources, a more stable and robust underwriting SCR was achieved.

6. Limitations of the Study

The framework of this study is restrictive as it has certain limitations. Among these, we can mention the limited number of segmentation models and sources of instability examined, as well as the use of generic metrics like the default proportion to represent the structure of a subsample. Conducting analyses around these limitations could open up new perspectives for our study.

7. Conclusion

Our study highlights the crucial importance of segmentation model stability, particularly in the context of credit insurance. We propose appropriate metrics and methods to assess this stability, following a detailed methodology based on the academic literature.

Ultimately, this thesis serves as a toolbox situated at the intersection of scientific research and practical application. It serves as a starting point for a better understanding and more precise management of segmentation model stability in credit insurance and other sectors where segmentation is utilized. The recommendations stemming from our results are intended to guide professionals in insurance, banking, and other sectors toward the construction of more robust segmentation models, thereby improving the quality of their analyses.

Remerciements

Je tiens à exprimer ma profonde gratitude envers toutes les personnes qui ont contribué de manière significative à la réalisation de ce mémoire. Leur soutien, leurs conseils et leur encouragement ont été essentiels tout au long de ce parcours académique et professionnel.

Tout d'abord, je souhaite remercier chaleureusement Ronan DAVIT, qui m'a proposé ce sujet de mémoire. Sa confiance en moi et son orientation précieuse ont été les pierres angulaires de ce projet.

Un grand merci à mes tuteurs, Julien LEBLOA et Slim SAANOUNI, pour leur encadrement attentif, leurs lectures critiques et leurs conseils éclairés. Leur expertise a été d'une importance cruciale pour donner forme à ce mémoire.

Je tiens également à exprimer ma reconnaissance envers l'ENSAE de Paris pour la qualité exceptionnelle de son enseignement et pour avoir créé un environnement propice à l'apprentissage et à la recherche en Actuariat.

Mes tuteurs académiques, Olivier LOPEZ et Caroline HILLAIRET, méritent une mention spéciale pour leurs précieuses contributions, leurs orientations avisées et leurs relectures attentives.

Je n'oublie pas de remercier chaleureusement l'équipe AQS de SIA Partners et l'équipe Actuariat de la Coface pour leurs accueils chaleureux, leur collaboration et leurs précieuses informations qui ont enrichi mon travail.

Enfin, un remerciement tout particulier va à ma famille et à mes proches pour leur soutien inébranlable, leur patience et leur encouragement constant.

Ce mémoire est le fruit d'un effort collectif, et je suis profondément reconnaissant envers chacune de ces personnes pour leur contribution à sa réussite.

Sigles et abréviations

ACPR	Autorité de Contrôle Prudentiel et de Résolution
ALM	Asset and Liability Management
APE	Activité Principale Exercée
AQS	Actuarial and Quantitative Services
B2B	Business to Business
BE	Best Estimate
BSCR	Base Solvency Capital Requirement
CAH	Classification Ascendante Hiérarchique
CART	Classification and Regression Trees
COFACE	Compagnie Française d'Assurance pour le Commerce Extérieur
CS	Contractual Specifications
EAD	Exposure at Default
ED	Eta-Deux
EIOPA	European Insurance and Occupational Pensions Authority
ENSAE	Ecole Nationale de la Statistique et de l'Administration Economique
FR	First Reserve
LGD	Loss Given Default
MCR	Minimum Capital Requirement
MIP	Modèle Interne Partiel
MSE	Mean Square Error
NAV	Net Asset Value
NOA	Notification of Overdue Amount
ORSA	Own Risk and Solvency Assessment
PD	Probabilité de Défaut
PV	P-Valeur
QRT	Quantitative Reporting Template
SCR	Solvency Capital Requirement
SE	Sous-Echantillon
SQL	Structured Query Language
ST	Statistique du Test
TBC	Taux de Bon Classement
UGD	Usage Given Default
VaR	Value at Risk
VC	V de Cramer
XGBoost	eXtreme Gradient Boosting

Table des matières

Introduction

L'assurance-crédit est un type d'assurance non-vie qui intervient principalement dans le cadre des transactions Business to Business (B2B) et offre à un assuré une protection contre les défauts de paiement de ses clients. En France, les entreprises d'assurance-crédit, comme toutes les autres entreprises d'assurance, sont soumises à la directive Solvabilité 2. Dans le but de protéger les assurés et le système financier, cette directive définit un capital de solvabilité requis (SCR) que ces entreprises doivent détenir pour continuer à exercer leur activité. Ce capital garantit avec une probabilité de 99,5% qu'une entreprise d'assurance pourra faire face à ses engagements envers ses assurés à un horizon d'un an. Le calcul de ce capital se fait soit grâce à une formule standard proposée par l'Autorité de contrôle prudentiel et de résolution (ACPR), soit grâce à un modèle interne validé par l'ACPR, offrant entre autres, l'avantage d'une meilleure adaptation à la structure du portefeuille de l'assureur.

La Compagnie française d'assurance pour le commerce extérieur (Coface), en raison des spécificités liées à son activité d'assurance-crédit et à la structure de son portefeuille, a opté pour l'utilisation d'un modèle interne partiel[1] pour le calcul de son SCR de souscription non-vie. Une étape de ce modèle consiste à calibrer des lois de probabilité sur des données afin d'effectuer des simulations. Ces dernières servent à construire une distribution du résultat technique dont le quantile à 0,5% correspond au SCR de souscription. Pour garantir une meilleure qualité des résultats et une adéquation avec la réalité, les calibrages de ces lois de probabilité doivent être effectués sur des groupes de données homogènes. La segmentation, une technique couramment utilisée en analyse de données statistiques, permet de construire ces groupes homogènes. Elle s'avère particulièrement utile lorsque la quantité de données à traiter devient importante (Shalev, et al., 2014).

Malgré le caractère indispensable et les avantages de la segmentation dans ce modèle interne partiel, l'ACPR s'interroge sur la stabilité des modèles de segmentation utilisés. Il convient en effet de se demander si les structures des segmentations obtenues grâce à ces modèles varient avec de petites perturbations dans les données d'entraînement, soulevant ainsi des questions sur la fiabilité des résultats. Malheureusement, bien que la segmentation soit largement utilisée, la littérature scientifique ne propose pas de méthodologie complète pour l'étude de la stabilité d'un modèle de segmentation.

L'objectif de ce mémoire est de proposer une méthodologie claire et détaillée pour l'étude et la validation de la stabilité d'un modèle de segmentation dans le cadre de l'assurance-crédit.

[1] *Combinaison de la formule standard et d'un modèle interne*

Nous nous attacherons spécifiquement à :

- Présenter l'importance de la segmentation en analyse et en modélisation statistique de manière générale, et plus spécifiquement, le rôle de la segmentation dans la modélisation de la sinistralité en assurance-crédit.
- Définir la stabilité d'un modèle de segmentation, élaborer des métriques adaptées pour son évaluation, et explorer les facteurs susceptibles d'influencer la stabilité d'un tel modèle.
- Proposer une méthodologie complète pour l'évaluation de la stabilité d'un modèle de segmentation.
- Appliquer cette méthodologie aux données et au modèle de segmentation de la Coface afin d'évaluer la stabilité de ce dernier.
- Démontrer la contribution de la stabilité des modèles de segmentation à la robustesse du SCR de souscription en assurance-crédit.

Ce mémoire se situe à la frontière entre la recherche scientifique et l'opérationnel. Il se présente comme une boîte à outils destinée à aider les acteurs des secteurs de l'assurance et de la banque à intégrer une phase d'étude de la stabilité lors de la conception de leurs modèles de segmentation. Il est divisé en trois grandes parties, chacune comprenant deux chapitres. La première partie présente le contexte de l'étude, expliquant ce qu'est l'assurance-crédit dans le cadre de la directive Solvabilité 2 et présentant le modèle interne partiel de la Coface, objet de nos travaux. La deuxième partie présente les bases théoriques de l'étude, comprenant une revue de la littérature visant à définir la stabilité d'un modèle de segmentation, ainsi que la méthodologie qui sera utilisée pour son étude dans la suite du mémoire. Enfin, la troisième partie présente les résultats de nos travaux, en commençant par la présentation des données de l'étude, suivie de l'application de la méthodologie aux données, la formulation de recommandations basées sur nos résultats, et la présentation des limites de notre étude.

Partie 1 : Contexte de l'étude

Chapitre 1 : Assurance-crédit et Solvabilité 2

Sommaire

Le but de ce chapitre est de présenter globalement l'assurance-crédit et la directive Solvabilité 2. Le calcul des besoins en fonds propres est une nécessité et une obligation pour toute activité d'assurance, et l'assurance-crédit ne fait pas exception à cette règle, bien qu'elle présente quelques particularités. Dans ce chapitre, nous aborderons également ces particularités ainsi que les branches du pilier 1 de la directive Solvabilité 2 qui permettent de calculer le besoin en fonds propres de l'activité d'assurance-crédit.

1.1. Présentation de l'assurance-crédit

L'assurance-crédit est un type d'assurance non-vie qui se distingue par le fait qu'elle couvre des risques liés aux clients des assurés plutôt qu'aux assurés eux-mêmes. En cas de non-remboursement des créances commerciales détenues sur des entreprises clientes, l'assurance-crédit garantit à un assuré une indemnisation (Ndoye, 2019). Les raisons de ce non-remboursement peuvent être regroupées en deux catégories :

- Les raisons internes aux clients de l'assuré, allant du retard de paiement à la faillite ;
- Les raisons externes, notamment à l'exportation, telles que des incidents politiques ou des catastrophes naturelles.

Ainsi, un sinistre ou un évènement de défaut en assurance-crédit survient en cas de défaillance des clients de l'assuré, appelés « acheteurs ». Cette notion est essentielle car elle guide la vision adoptée dans toute modélisation visant à quantifier les risques en assurance-crédit. La plupart des analyses en assurance-crédit se concentrent sur les acheteurs. L'assurance-crédit est principalement utilisée dans le cadre des activités B2B (Business to Business), où les assurés et les acheteurs sont des entreprises (Allianz, 2023). Le terme « assurance-crédit commerciale » est parfois employé pour rappeler que l'assuré est une entreprise commerciale. En France, l'assurance-crédit correspond à la branche 14 du Code des assurances.

Dans le domaine de l'assurance-crédit, l'assureur couvre un montant limite de créances en cas de défaut d'un acheteur. Ce montant limite est appelé garantie et il est défini lors de la souscription de la police. Toutefois, il peut être révisé à l'initiative de

l'assureur ou de l'assuré. Une police d'assurance-crédit contient plusieurs agréments ou contrats, chacun associé à un acheteur de l'assuré considéré. Le schéma suivant illustre de manière très simplifiée l'activité d'assurance-crédit :

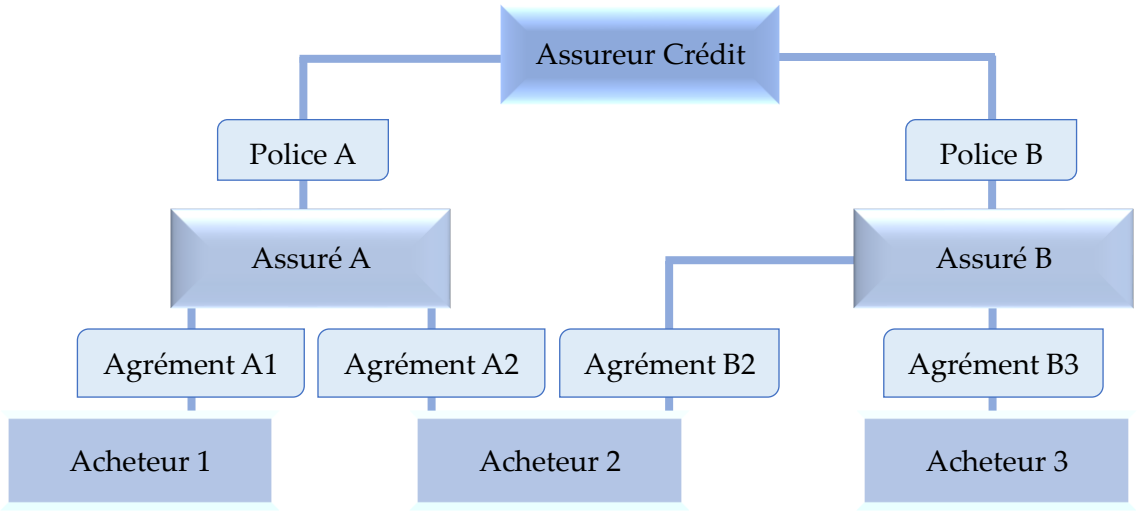


Figure 1 : Illustration de l'assurance-crédit

Comme illustré sur cette figure, il est possible et fréquent que des agréments soient délivrés à différents assurés pour le même acheteur. Par exemple, les agréments A2 et B2 portent sur le même acheteur, mais l'agrément A2 couvre la créance de l'assuré A, tandis que l'agrément B2 couvre la créance de l'assuré B. Nous verrons dans le chapitre suivant que les méthodes de modélisation et de quantification des risques en assurance-crédit sont également influencées par cette spécificité.

1.1.1. Fonctionnement d'une police d'assurance-crédit

Avant de conclure une transaction avec un acheteur, l'assuré doit informer l'assureur-crédit. Ce dernier procédera à une étude de la santé financière de l'acheteur et décidera ensuite d'autoriser ou de refuser la transaction en fonction des informations recueillies. Si l'assureur autorise la transaction, il attribuera un montant limite pour cette transaction entre l'assuré et cet acheteur. Comme mentionné précédemment, ce montant limite pourra être ajusté à la demande de l'assuré ou être réduit en cas de dégradation de la solvabilité constatée par l'assureur.

En cas de défaut de paiement, l'assureur pourra indemniser l'assuré en payant la facture à la place de l'acheteur. Par la suite, l'assureur se substituera à l'assuré dans la gestion de la créance et pourra notamment engager une procédure judiciaire pour récupérer les montants dus et ainsi procéder au recouvrement.

Le schéma suivant illustre ce fonctionnement :

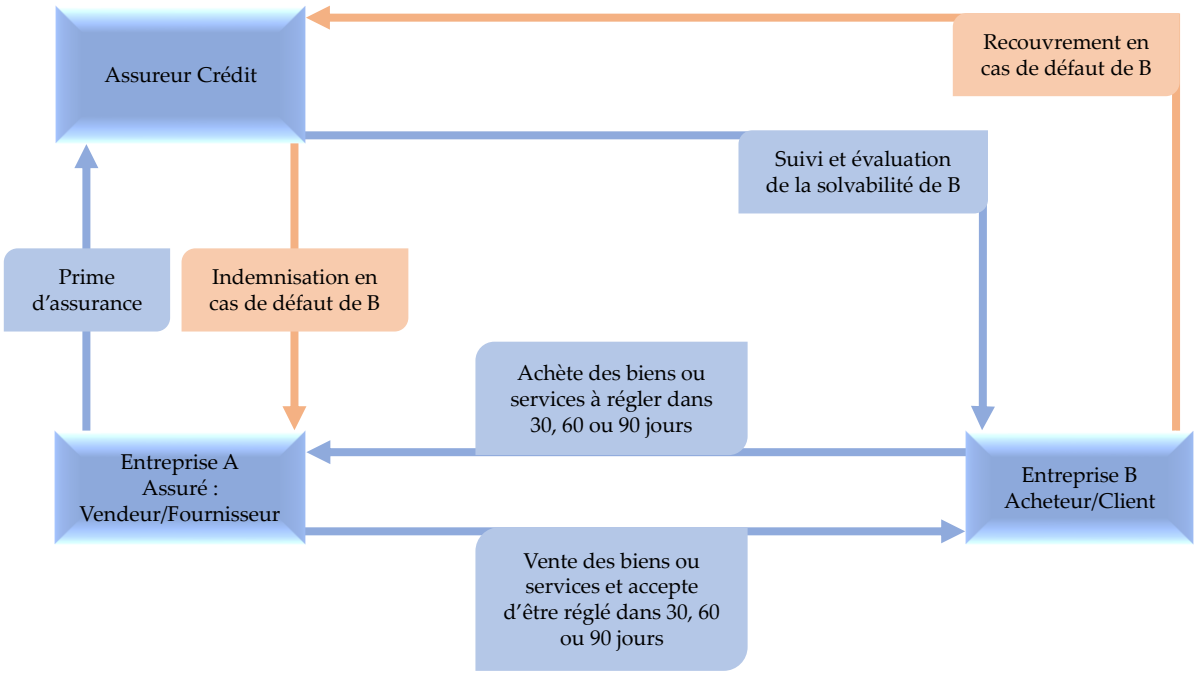


Figure 2 : Fonctionnement de l'assurance-crédit

Une police d'assurance-crédit peut également contenir des clauses telles que des franchises et des quotités garanties. La quotité garantie fixe la part des sinistres que l'assureur prend en charge. Elle varie généralement entre 60% et 90% (Ooreka, 2023). La prime d'un contrat d'assurance-crédit prend en compte tous ces éléments et sera généralement forfaitaire pour les petites entreprises, ou fixée en fonction d'un pourcentage du chiffre d'affaires ou de la part du chiffre d'affaires faisant l'objet du contrat d'assurance. Ce pourcentage varie généralement entre 0,1% et 0,5% (Altassura, 2023).

1.1.1.1. Evaluation du risque acheteur en assurance-crédit

Nous avons vu précédemment qu'une étude de la santé financière des acheteurs est effectuée en continu par l'assureur afin de réduire les impayés potentiels. En effet, les assureurs crédit disposent généralement d'une grille de notation pour un grand nombre de sociétés dans les régions du monde où ils exercent leur activité. L'avantage des assureurs crédit par rapport aux grandes agences de notation est d'être plus proches des sociétés, y compris les plus petites. Cette proximité leur permet de disposer d'une base de données fiable, constamment mise à jour et analysée pour produire des scores reflétant la santé financière des acheteurs avec un niveau de précision acceptable. Cette base de données est alimentée par des informations financières publiques ainsi que par la connaissance des particularités du secteur (géographique et d'activité) couvert.

<table>
<tr><th colspan="2">Sources publiques</th><th colspan="2">Sources spécifiques</th></tr>
<tr><th>Sources</th><th>Données</th><th>Sources</th><th>Données</th></tr>
<tr><td>Instituts Nationaux de Statistiques</td><td rowspan="3">Données d'identification des entreprises : SIREN/SIRET, code APE, statuts, effectifs, comptes sociaux, date de création, etc.</td><td>Entreprises concernées (via courrier, téléphone ou visite)</td><td>Actionnariat, métier, rapports des commissaires aux comptes, principaux clients, compléments des rapports annuels, documents comptables, carnet de commandes, business plan, etc.</td></tr>
<tr><td>Registres du Commerce et des Sociétés</td><td rowspan="2">Historiques des entreprises concernées</td><td rowspan="2">Performances financières des entreprises, dynamique de paiement des créances, croissance des entreprises, etc.</td></tr>
<tr><td>Journaux officiels</td></tr>
<tr><td>Sites Internet</td><td>Rapports annuels, actualités des entreprises, documentation commerciale, etc.</td><td rowspan="2">Analyses du secteur d'activité et de la zone géographique des entreprises concernées</td><td rowspan="2">Conjoncture de la branche, situation économique et politique de la zone géographique, évolution du cours des matières premières, etc.</td></tr>
<tr><td>Greffes des tribunaux du commerce</td><td>Procédures collectives, privilèges, etc.</td></tr>
</table>

Tableau 1 : Sources de données utilisées pour la natation

1.1.1.2. Le recouvrement

Avant ou après l'indemnisation d'un assuré, l'assureur crédit peut entamer le recouvrement de la créance auprès de l'acheteur en défaut. Le recouvrement des créances constitue l'un des services essentiels d'une police d'assurance-crédit. En effet, il s'agit d'une activité à part entière qui contribue à préserver la rentabilité de l'assureur crédit. Grâce au recouvrement des créances, les primes en assurance-crédit peuvent être maintenues à un niveau acceptable pour la plupart des assurés.

Après la déclaration d'un sinistre, l'assureur dispose d'un délai contractuel pour se rapprocher de l'acheteur en défaut afin de comprendre les raisons de l'impayé. Ensuite, l'assureur détermine la méthode de recouvrement la plus adaptée. Selon la méthode retenue, l'assureur décide s'il doit indemniser le sinistre ou non.

Dans la majorité des cas, l'impayé est dû à des retards souvent liés à des problèmes de trésorerie. Dans de tels cas, l'assureur peut proposer de mettre en place un échéancier de paiement afin de résoudre le problème à l'amiable. Le sinistre peut donc ne pas être indemnisé par l'assureur, et on appelle les montants récupérés : « économies d'indemnités ».

Dans des cas plus complexes, comme les litiges (lorsque l'assuré et son acheteur ne sont pas d'accord sur une facture émise) ou même les procédures collectives (liquidation d'une société par exemple), c'est généralement une décision de justice qui intervient. Les sinistres sont alors indemnisés, et les montants éventuellement récupérés sont conservés par l'assureur. On parle ici de « recours après indemnisation ».

Le schéma ci-dessous présente le processus de recouvrement en assurance-crédit :

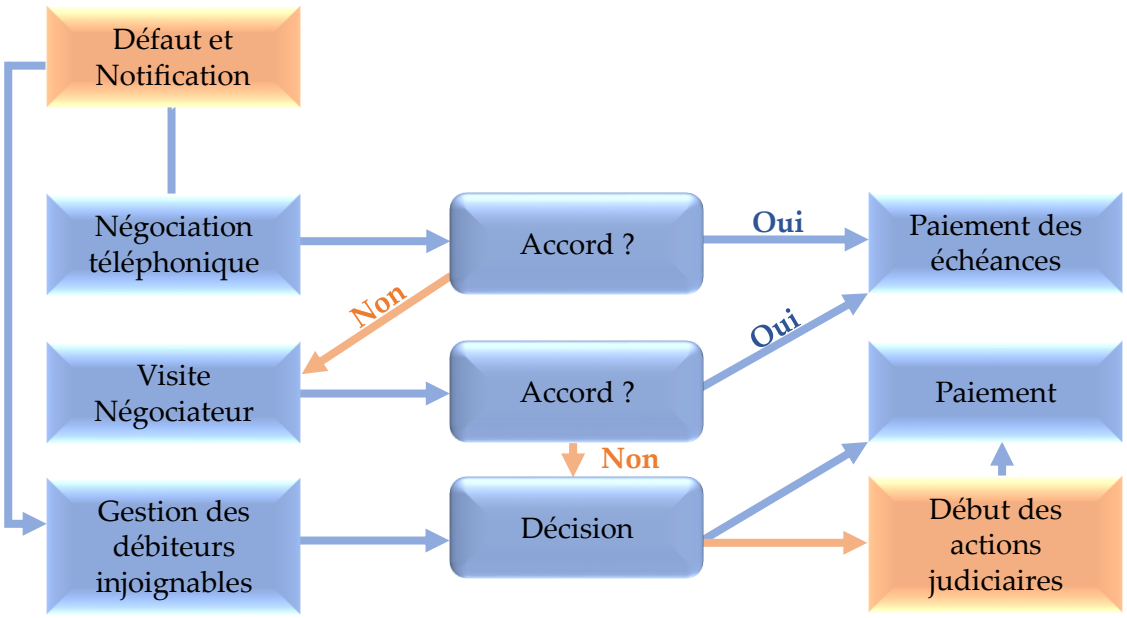


Figure 3 : Le recouvrement en assurance-crédit

1.1.2. Notion d'exposition en assurance-crédit

La plupart des études en assurance-crédit ont pour variable centrale l'exposition. Les travaux de ce mémoire ne font pas exception. Nous verrons dans le *Chapitre 2* que l'exposition est la variable clé à partir de laquelle la perte ultime et le résultat technique d'une entreprise d'assurance-crédit sont calculés. L'exposition correspond au montant maximal des factures que l'assureur-crédit prendra en charge pour un assuré vis-à-vis d'un acheteur (avant quotité garantie, franchise ou réassurance). Ce montant est déterminé pour chaque nouvel acheteur et peut varier durant la vie du contrat d'assurance (l'agrément). Il existe quelques notions complémentaires liées à l'exposition :

- **Exposition acquise :** Moyenne de l'exposition sur une période T (en général 12 mois), pondérée par le temps de présence dans le portefeuille. L'exposition acquise de l'année t sur l'agrément α est donnée par :

$$Expo_{t,\alpha}^{aq} = \frac{\sum_{m=1}^{12} I_{t,\alpha}(m) \times Expo_{t,\alpha}(m)}{\sum_{m=1}^{12} I_{t,\alpha}(m)}$$

Où :

- $Expo_{t,\alpha}(m)$ est l'exposition du mois m de l'année t sur l'agrément α
- $I_{t,\alpha}(m)$ prend la valeur 1 si l'agrément α était dans le portefeuille au mois m de l'année t et 0 sinon.

L'exposition acquise permet de prendre en compte le profil de risque en considérant uniquement les garanties qui sont réellement exposées à l'occurrence des sinistres sur une période donnée (Friedland, 2013), (Brown, et al., 2007).

- **Exposition déformée :** Exposition simulée sur un portefeuille pour l'exercice suivant ;
- **Exposition en nombre :** Nombre d'acheteurs dans le portefeuille ;
- **Exposition sous risque :** Exposition × Quotité garantie ;
- **Exposition en défaut (EAD[2]) :** Exposition d'un agrément à la date de rattachement du sinistre. Plusieurs facteurs peuvent expliquer la différence entre l'exposition et l'exposition en défaut pour un acheteur donné :
  - Expiration de certains agréments ;
  - Variation des montants garantis ;
  - Défauts sur certains agréments uniquement.

A titre d'illustration, pour l'agrément suivant, si le sinistre survient en décembre, l'EAD sera de 100.

Mois	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Exposition	0	0	80	100	80	110	110	80	80	100	100	100

Tableau 2 : Exemple d'agrément en assurance-crédit

1.1.3. Avantages de l'assurance-crédit

Un retard ou un défaut de paiement peut affecter la liquidité d'une entreprise et, dans certains cas, entraîner la faillite. En effet, des études montrent qu'environ 80% des entreprises ont déjà fait face à des impayés et que près de 25% des dépôts de bilan sont dus à des impayés (COFACE, 2023). L'assurance-crédit permet d'éviter cela en sécurisant les revenus des entreprises.

L'assurance-crédit permet également de faciliter les transactions et de les rendre plus liquides. En l'absence d'assurance-crédit, les entreprises sont obligées d'exiger des avances ou des paiements avant service lors de leurs transactions. Ceci peut réduire les volumes des ventes et ralentir l'exploration des nouveaux marchés.

[2] *Exposure at default*

Les assureurs-crédit s'engagent en général à suivre de près la solvabilité des acheteurs de leurs assurés. Ceci permet aux assurés de réduire leurs frais d'accès aux informations de solvabilité de leurs acheteurs en amont de toute transaction.

En cas d'impayés, l'assureur-crédit se charge en général du recours auprès de l'acheteur en situation d'impayé. Ceci permet aux assurés d'économiser sur les frais de recours et de justice.

En période de crise, l'assurance-crédit peut s'avérer particulièrement utile. En effet, lors de la pandémie de la COVID en 2020, l'économie mondiale a connu un ralentissement général et plusieurs entreprises se sont trouvées en situation d'impayés. Les entreprises ayant souscrit à une assurance-crédit ont pu bénéficier d'indemnisations suite aux impayés de leurs acheteurs. De même, lors de la crise de 2008-2009, les assureurs-crédit ont indemnisé à hauteur de 9 milliards d'euros leurs assurés, ce qui a permis à beaucoup d'acteurs d'éviter la faillite (Lefebvre, 2017).

Les travaux réalisés dans ce mémoire portent sur un modèle interne partiel (combinaison de la formule standard et d'un modèle interne). Les modèles internes et la formule standard sont les outils permettant d'implémenter les exigences du pilier 1 de la directive Solvabilité 2. Nous allons donc présenter la directive Solvabilité 2 dans la section suivante et expliquer comment cette dernière s'adapte à l'activité d'assurance-crédit.

1.2. Aperçu de Solvabilité 2

Les travaux sur la directive Solvabilité 2 ont débuté en 2001. À cette époque, la Commission Européenne souhaitait combler les lacunes de la directive Solvabilité 1 en construisant un dispositif permettant de mieux cerner et quantifier les risques auxquels les assureurs européens font face, dans le but de mieux protéger les assurés contre ces risques (Dreyfuss, 2015). Pour atteindre cet objectif, la directive Solvabilité 2, qui était alors un projet, s'est vu attribuer des objectifs bien précis.

1.2.1. Principaux objectifs de la directive Solvabilité 2

La directive Solvabilité 2 apporte des règles de solvabilité modernes auxquelles toutes les entreprises d'assurance européennes doivent se soumettre. Ceci dans le but :

- d'assurer une protection maximale des assurés ;
- d'avoir des règles de solvabilité homogènes et harmonisées en Europe ;
- de mieux cerner et mesurer les risques des entreprises d'assurance ;

- de fournir un terrain commun sur lequel les entreprises d'assurance peuvent exercer et être contrôlées.

Pour atteindre ces objectifs, plusieurs étapes caractérisent la nouvelle directive. Il s'agit, dans un premier temps, de l'identification des risques, puis de la mesure, la quantification et le suivi de ces risques, et enfin de la communication des calculs et analyses aux autorités compétentes afin de leur permettre d'évaluer la santé financière de l'entreprise d'assurance.

En Europe, la directive Solvabilité 2 est au secteur des assurances ce que les accords de Bâle sont au secteur bancaire (Charbonneau, 2003). Structurée en trois piliers, la directive Solvabilité 2 fournit des consignes détaillées permettant aux entreprises d'assurance d'atteindre les objectifs présentés plus haut.

1.2.2. Les 3 piliers de la directive Solvabilité 2

Sous Solvabilité 2, le risque est abordé selon une approche économique. La directive adopte une vision prospective du bilan pour refléter le véritable profil de risque de l'assureur. Cette vision globale du risque est rendue possible grâce à une structure en 3 piliers.

Le premier pilier est de nature quantitative et fournit les règles pour évaluer le passif et l'actif en fonction des valeurs de marché. Le deuxième pilier, plus qualitatif que quantitatif, concerne la gestion interne des risques et le suivi par les autorités de contrôle. Il implique la mise en place d'un dispositif de suivi et de gestion de tous les risques (financiers, techniques et opérationnels) auxquels l'assureur est exposé. Cela permet aux compagnies d'assurance d'assurer une cohérence entre leurs stratégies et leurs capacités financières pour les soutenir.

Le troisième pilier concerne la diffusion et la communication des résultats et analyses issus des deux premiers piliers. Tous les assureurs doivent communiquer les mêmes types d'informations à l'attention des assurés, des investisseurs et des autorités de contrôle. Par exemple, les assureurs doivent fournir les rapports suivants aux autorités de contrôle :

- Des rapports narratifs présentant de manière descriptive la politique prudentielle mise en place ;
- Des rapports quantitatifs sous forme de tableaux de bord couvrant l'ensemble des activités de l'assureur : provisions techniques, programme de réassurance, gestion d'actifs, etc.

Le tableau suivant résume les exigences des 3 piliers de la directive Solvabilité 2 :

Pilier 1 **Exigences Quantitatives**	**Pilier 2** **Exigences Qualitatives**	**Pilier 3** **Exigences d'Information**
• Provisions techniques • MCR et SCR • Classification des fonds propres	• Contrôle interne • Gestion des risques • Harmonisation des procédure	• Transparence et discipline de marché • Publication et communication
➢ Inventaires (trimestriels et annuels) ➢ Nouvelles normes pour le calcul des provisions ➢ Calcul des nouveaux capitaux de solvabilité ➢ Pilotage et projection à court terme	➢ ORSA[3] et gestion des risques ALM[4] ➢ Fonctions clés et organisations	➢ Rapport pour l'autorité de contrôle ➢ Rapport pour le public QRT[5]

Tableau 3 : Les trois piliers de la Solvabilité 2

1.2.3. Le pilier 1 de Solvabilité 2 et focus sur l'assurance-crédit

La directive Solvabilité 2 impose aux compagnies d'assurance de respecter un niveau de solvabilité suffisant pour être en mesure de répondre à leurs engagements envers leurs assurés. Le pilier 1 définit, entre autres, les règles de calcul de ce niveau de solvabilité. Afin de mieux comprendre ces règles, nous présentons ci-dessous un bilan d'une compagnie d'assurance sous Solvabilité 2 :

<table>
<tr><td rowspan="5">ACTIFS</td><td colspan="2">Surplus de Capital</td><td rowspan="3">Fonds Propres</td><td rowspan="5">PASSIFS</td></tr>
<tr><td>SCR – MCR</td><td rowspan="2">SCR</td></tr>
<tr><td>MCR</td></tr>
<tr><td colspan="2">Marge de Risque</td><td rowspan="2">Provisions Techniques</td></tr>
<tr><td colspan="2">Best Estimate</td></tr>
<tr><td colspan="3">Autres Passifs</td></tr>
</table>

Tableau 4 : Bilan d'une compagnie d'assurance sous Solvabilité 2

Deux éléments importants qui figurent au passif d'une compagnie d'assurance sont les provisions techniques et les fonds propres. Ce mémoire se focalisera essentiellement sur les fonds propres. Le premier pilier fournit un ensemble de règles visant à définir des seuils quantitatifs pour les fonds propres des entreprises d'assurance. Ces seuils exigés par la réglementation en vigueur dépendent des

[3] *Own Risk and Solvency Assessment*
[4] *Asset and Liability Management*
[5] *Quantitative Reporting Template*

données issues des activités des différentes compagnies d'assurance. Deux niveaux de fonds propres sont définis : le Capital Minimum Requis (MCR[6]) et le Capital de Solvabilité Requis (SCR[7]).

Le MCR représente le niveau minimum absolu de fonds propres tolérés par le régulateur. En dessous de cette limite, une surveillance automatique est déclenchée et l'assureur concerné peut perdre l'autorisation d'exercer.

1.2.4. Le Capital de Solvabilité Requis (SCR) en assurance-crédit

Le Capital de Solvabilité Requis est le montant de fonds propres que doit disposer une compagnie d'assurance pour avoir une probabilité de 99,5% de pouvoir faire face à tous ses engagements envers ses assurés à un horizon d'un an. En d'autres termes, le SCR est le niveau de fonds propres qui garantit à une compagnie d'assurance qu'elle pourra faire face au cours de l'année à une ruine économique qui se produit une fois tous les 200 ans. La ruine économique est une situation où la valeur de marché de l'actif est inférieure au Best Estimate (BE) des passifs. Le SCR d'une entreprise est constitué de plusieurs sous-modules. La figure suivante en fait une présentation :

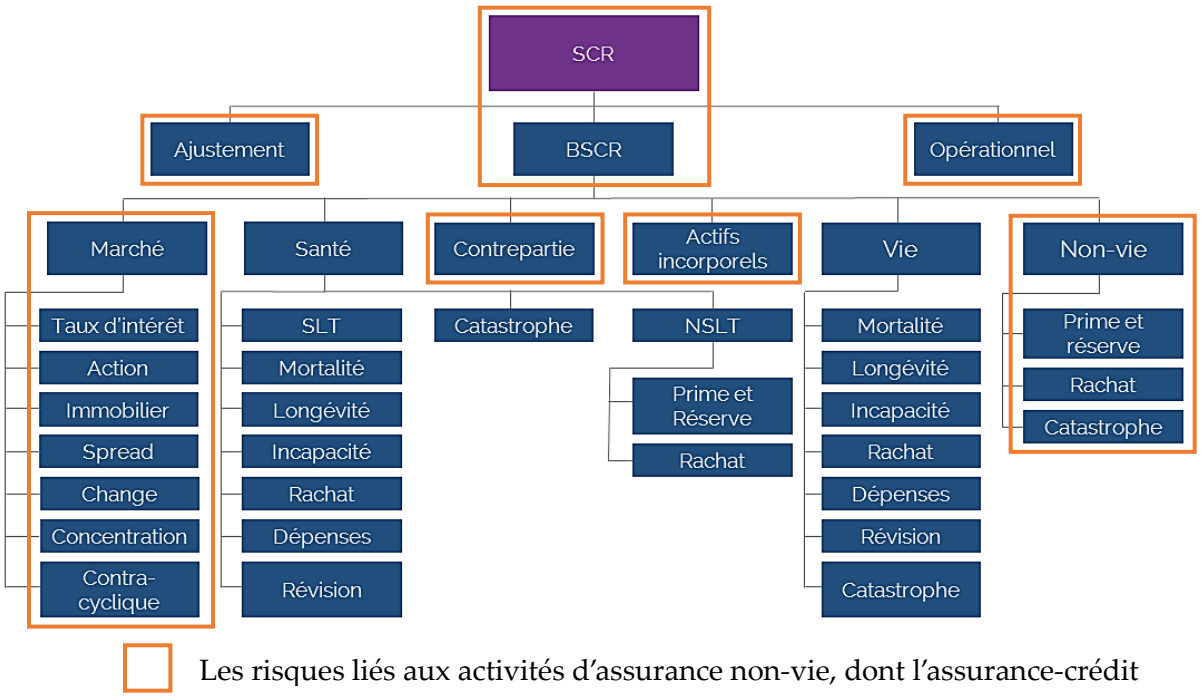


Figure 4 : Le SCR en assurance-crédit

Comme illustré sur cette figure, l'activité d'assurance-crédit, qui est une activité d'assurance non-vie, est soumise à 4 sous-modules de risques. Ces derniers sont les

[6] *Minimum Capital Requirement*
[7] *Solvency Capital Requirement*

risques de marché, les risques de contrepartie, les risques liés aux actifs incorporels et les risques de souscription non-vie.

Pour mieux illustrer le calcul des besoins en capital liés à chaque sous-module, considérons le bilan économique simplifié suivant :

Actifs à l'instant $\boldsymbol{t}, \boldsymbol{A_t}$	Fonds Propres à l'instant $\boldsymbol{t}, \boldsymbol{FP_t}$
	Best Estimate des Passifs à l'instant t, BE_t

Tableau 5 : Bilan économique simplifié

Les actifs sont évalués en valeur de marché. Le Best Estimate des passifs en date t est défini comme l'espérance actualisée des cash-flows de passifs sous la probabilité risque-neutre (Boumezoued, 2022). Le bilan étant équilibré, on a $FP_t = A_t - BE_t$. FP_t est donc l'espérance actualisée des marges futures sous la probabilité risque-neutre, que nous appellerons par la suite NAV (Net Asset Value).

Le SCR peut donc être défini comme étant le montant minimal de NAV au moment de l'évaluation ($t = 0$) tel que la probabilité d'avoir une NAV négative au cours de l'année ($t = 1$) soit plus petite que 0,5% (Mehalla, 2021).

$$SCR = inf\{x \in \mathbb{R} : \ \mathbb{P}(NAV_1 \leq 0 | NAV_0 = x) \leq 0{,}005\}$$

Rappelons que NAV_1 (représentant la valeur des fonds propres disponibles à l'année 1) est une variable aléatoire.

Le SCR peut également être calculé comme le quantile de la distribution des pertes à travers la Value-at-Risk. Cette approche est la plus utilisée en pratique. Dans ce cas, le SCR sera défini comme le montant minimal en dessous duquel la perte de l'entreprise d'assurance se trouve avec une probabilité de 0,995.

Définition : La VaR associée à une variable aléatoire Yet de niveau $\alpha \in [0,1]$ est donnée par $VaR_\alpha(Y) = \underset{m \in \mathbb{R}}{inf}\{\mathbb{P}(Y \leq -m) \leq \alpha\}$. En posant $m' = -m$, on a :

$$VaR_\alpha(Y) = -\underset{m' \in \mathbb{R}}{sup}\{\mathbb{P}(Y \leq m') \leq \alpha\} = -q_\alpha(Y)$$

Où $q_\alpha(Y)$ est le quantile d'ordre α de Y.

Au moment de l'évaluation du SCR ($t = 0$) la perte L est une variable aléatoire définie par $L = NAV_0 - D(0,1) \times NAV_1$, où $D(0,1)$ est le facteur d'actualisation entre l'année 0 et l'année 1. Le SCR est donc donné par :

$$SCR = \underset{x \in \mathbb{R}}{inf}\{\mathbb{P}(L \leq x) \geq 0{,}995\}$$

En remplaçant L, on a :

$$SCR = \underset{x \in \mathbb{R}}{inf}\{\mathbb{P}(x + D(0,1) \times NAV_1 \leq NAV_0) \leq 0{,}005\}$$

En posant $x = y + NAV_0$ on a :

$$SCR = NAV_0 + \inf_{y \in \mathbb{R}} \{\mathbb{P}(y + D(0,1) \times NAV_1 \leq 0) \leq 0{,}005\}$$

En appliquant la définition de la VaR donnée plus haut, on a :

$$SCR = NAV_0 + VaR_{0,5\%}(D(0,1) \times NAV_1)$$

D'où

$$SCR = NAV_0 - q_{0,5\%}(D(0,1) \times NAV_1)$$

Le calcul de $q_{0,5\%}(D(0,1) \times NAV_1)$ n'est pas toujours évident. Les compagnies d'assurance qui disposent des ressources et de suffisamment de données peuvent réaliser ce calcul grâce à un modèle interne. Mais elles ont également la possibilité d'utiliser une méthode plus simple et moins complexe, la formule standard.

1.2.4.1. La formule standard du SCR et les modèles internes

La formule standard est une méthode réglementaire fournie par le régulateur. Le modèle interne, quant à lui, est plus complexe à mettre en œuvre mais a le mérite de tenir compte de la structure du portefeuille de l'assureur. Il est donc plus précis et permet une meilleure optimisation des fonds propres de l'assureur. Toutefois, un modèle interne doit être justifié très soigneusement et soumis au régulateur pour validation avant son utilisation. Il est également possible pour une compagnie d'assurance d'utiliser la formule standard sur certains sous-modules et un modèle interne sur d'autres sous-modules. Cette combinaison donne lieu à un modèle interne partiel (Dreyfuss, 2015). La figure suivante présente les avantages et les inconvénients de la formule standard et du modèle interne (Le Vallois, 2021) :

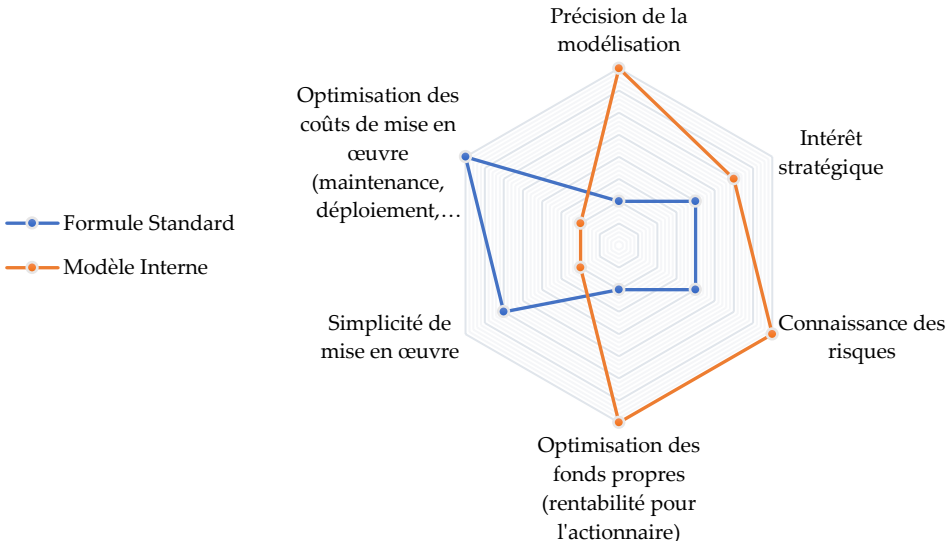


Figure 5 : Formule standard vs modèle interne

1.2.4.1.1. La formule standard

Comme évoqué précédemment, le calcul du SCR se fait suivant une approche modulaire. L'application de la formule standard implique le calcul des besoins en capital pour chaque risque élémentaire (action, immobilier, spread, etc.). Le SCR d'un risque élémentaire est calculé en remplaçant $q_{0,5\%}(D(0,1) \times NAV_1)$ par $NAV^{choqué}$ obtenu en appliquant un choc (en pourcentage) au facteur de risque étudié[8]. Ce choc fourni par l'EIOPA[9] permet de simuler la survenance d'un risque durant l'année 1. Les chocs utilisés varient en fonction du facteur de risque étudié et conformément à la directive Solvabilité 2, ils doivent correspondre au scénario d'un risque qui se produit une fois tous les 200 ans (probabilité d'occurrence de 0,5%). On a donc pour un risque élémentaire i :

$$SCR_i = NAV_0 - NAV_i^{choqué}$$

Les SCR des risques élémentaires sont ensuite agrégés grâce à des coefficients de corrélation fournis par l'EIOPA pour obtenir le besoin en capital d'un sous-module (marché, souscription non-vie, etc.). Ces coefficients de corrélation permettent de prendre en compte l'effet de diversification entre les différents sous-modules (Germain, 2022). Pour un sous-module y, on a :

$$SCR_y = \sqrt{\sum_{(i,j)\in R_y} \rho_{i,j} SCR_i SCR_j}$$

Où R_y est l'ensemble des risques élémentaires constituant le sous-module y et $\rho_{i,j}$ est le coefficient de corrélation entre les risques élémentaires i et j.

Le capital de solvabilité requis de base (BSCR[10]) est ensuite obtenu en agrégeant de la même manière les SCR des sous-modules concernés. Dans le cas de l'assurance-crédit, on a $BSCR = \sqrt{\sum_{(i,j)\in R} \rho_{i,j} SCR_i SCR_j}$ avec R = {Marché, Contrepartie, Souscription non-vie, Actifs Incorporels}.

On obtient enfin le SCR de l'entreprise d'assurance comme la somme du BSCR, du chargement de capital au titre du risque opérationnel et des ajustements au titre de la capacité d'absorption des pertes. On a :

$$SCR = Ajustement + BSCR + SCR_{Op}$$

[8] *Un facteur de risque est une composante ou une branche d'un risque élémentaire susceptible de contenir un risque. Exemple : LTEI pour le risque action*

[9] *European Insurance and Occupational Pensions Authority*

[10] *Base Solvency Capital Requirement*

Les risques opérationnels sont les risques liés aux erreurs humaines, aux pannes informatiques, etc.

1.2.4.1.2. Le modèle interne

Un assureur qui souhaite utiliser un modèle interne doit au préalable le faire valider et approuver par le régulateur. L'utilisation d'un modèle interne présente de nombreux avantages, parmi lesquels la possibilité pour l'assureur de construire un SCR adapté à son portefeuille d'assurés et à son activité. Toutefois, il est important de noter que la conception, la production et la mise en service d'un modèle interne sont en général des processus très longs et coûteux.

Nous avons précédemment défini le SCR en fonction de la perte comme suit :

$$SCR = \underset{x\in\mathbb{R}}{inf}\{\mathbb{P}(L \leq x) \geq 0{,}995\}$$

Cette expression peut également s'écrire :

$$SCR = \underset{x\in\mathbb{R}}{inf}\{\mathbb{P}(-L \leq -x) \leq 0{,}005\}$$

D'après la définition de la VaR donnée plus haut, on a :

$$SCR = VaR_{0,5\%}(-L)$$

Cette formule nous permet d'avoir une intuition sur la méthode de calcul du SCR grâce à un modèle interne. Nous parlons ici d'intuition car la définition de L (et donc de $-L$) peut varier en fonction du risque considéré et de la structure du modèle. Les modèles internes utilisent en général des méthodes de simulations ou des méthodes stochastiques pour générer un grand nombre de valeurs de $-L$. En disposant d'un grand nombre de réalisations de $-L$, il est ensuite possible de calculer le SCR comme la VaR empirique de ces réalisations.

Puisque la simulation des réalisations de $-L$ prend en compte l'activité de l'assureur ainsi que les données issues de cette activité, les modèles internes sont a priori plus précis et permettent de mieux capter le risque de l'assureur (cf. *Figure 5*).

Dans le chapitre suivant, nous présenterons le modèle interne partiel de la Coface sur lequel porteront les travaux de ce mémoire. Nous définirons également le périmètre de notre étude ainsi que les parties de ce modèle interne sur lesquelles nous nous focaliserons.

Chapitre 2 : Modèle Interne Partiel de la Coface et rôle de la segmentation

Sommaire

Nous avons défini ce que sont des modèles internes et des modèles internes partiels dans le chapitre précédent. Dans ce chapitre, nous présenterons le modèle interne partiel de la Coface et le rôle que joue la segmentation dans ce dernier. Ce chapitre permettra de définir le périmètre de notre étude ainsi que les objectifs visés. Avant de nous lancer dans le vif du sujet, il est important de faire une brève présentation de la Coface.

2.1. Présentation de la Coface

La Coface est une entreprise d'assurance non-vie créée en 1946. Elle est apparue au lendemain de la Seconde Guerre mondiale dans un contexte d'expansion du commerce extérieur. À l'origine, la Coface était spécialisée dans l'assurance-crédit à l'exportation française. Son activité a évolué par la suite pour lui permettre de commercialiser les produits suivants :

- **L'assurance-crédit** : Protège les entreprises contre le risque d'impayés en leur permettant d'être couvertes et indemnisées en cas de non-paiement de leurs créances commerciales, que ce soit sur leur marché domestique ou à l'export ;
- **Le single risk** : Couvre des risques commerciaux et politiques dans le cadre d'opérations ponctuelles, complexes, d'un montant élevé (généralement supérieur à 5 M€) et dont la durée de crédit est comprise entre 12 mois et 7 ans ;
- **L'information et la notation d'entreprise** : Consistent à collecter, traiter et analyser toute information permettant de quantifier la solvabilité d'une entreprise grâce à une note ou un score ;

- **L'affacturage** : Une technique financière par laquelle une société d'affacturage (le factor) finance et, le cas échéant, gère le poste clients d'une entreprise en acquérant ses créances clients ;
- **Le cautionnement** : Consiste en un engagement de payer le bénéficiaire de la caution en cas de défaillance éventuelle ou de manquement par le cautionné à ses obligations contractuelles.

Aujourd'hui, la Coface est implantée dans plus de 67 pays et continue toutefois à assurer la gestion des garanties publiques à l'exportation pour le compte de l'État Français. Les chiffres d'affaires de la Coface pour le premier semestre de 2022 sont donnés dans le tableau suivant (COFACE, 2022) :

Activité	**Chiffre d'affaires en milliers d'euros**	
	30 Juin 2022	30 Juin 2021
Assurance-crédit	790 666	674 425
Single Risk	13 338	10 915
Information et autres services	26 209	23 173
Affacturage	35 038	31 548
Cautionnement	29 640	27 977

Tableau 6 : Chiffre d'affaires de la Coface par activité aux premiers trimestres 2022 et 2021

L'activité d'assurance-crédit représente donc la majeure partie du chiffre d'affaires total de la Coface (plus de 89% au premier semestre de 2022). De plus, contrairement aux autres modules de risque et aux autres activités, la souscription en assurance-crédit comporte de d'importantes singularités. Pour ces deux raisons, le modèle interne analysé dans ce mémoire a été développé uniquement pour l'activité d'assurance-crédit. L'objectif étant d'adopter une vision « Risques » dans le pilotage de l'entreprise, comme recommandé par la directive Solvabilité 2. Dans la section suivante, nous verrons plus en détail où interviennent la formule standard et le modèle interne dans les calculs du besoin en capital de la Coface.

2.2. Périmètre du modèle interne partiel de la Coface

Nous avons vu au chapitre précédent que le SCR de l'activité d'assurance-crédit était destiné à couvrir les risques de souscription, les risques de contrepartie, les risques de marché et les risques opérationnels. La structure du SCR de la Coface ne s'écarte pas de ce schéma.

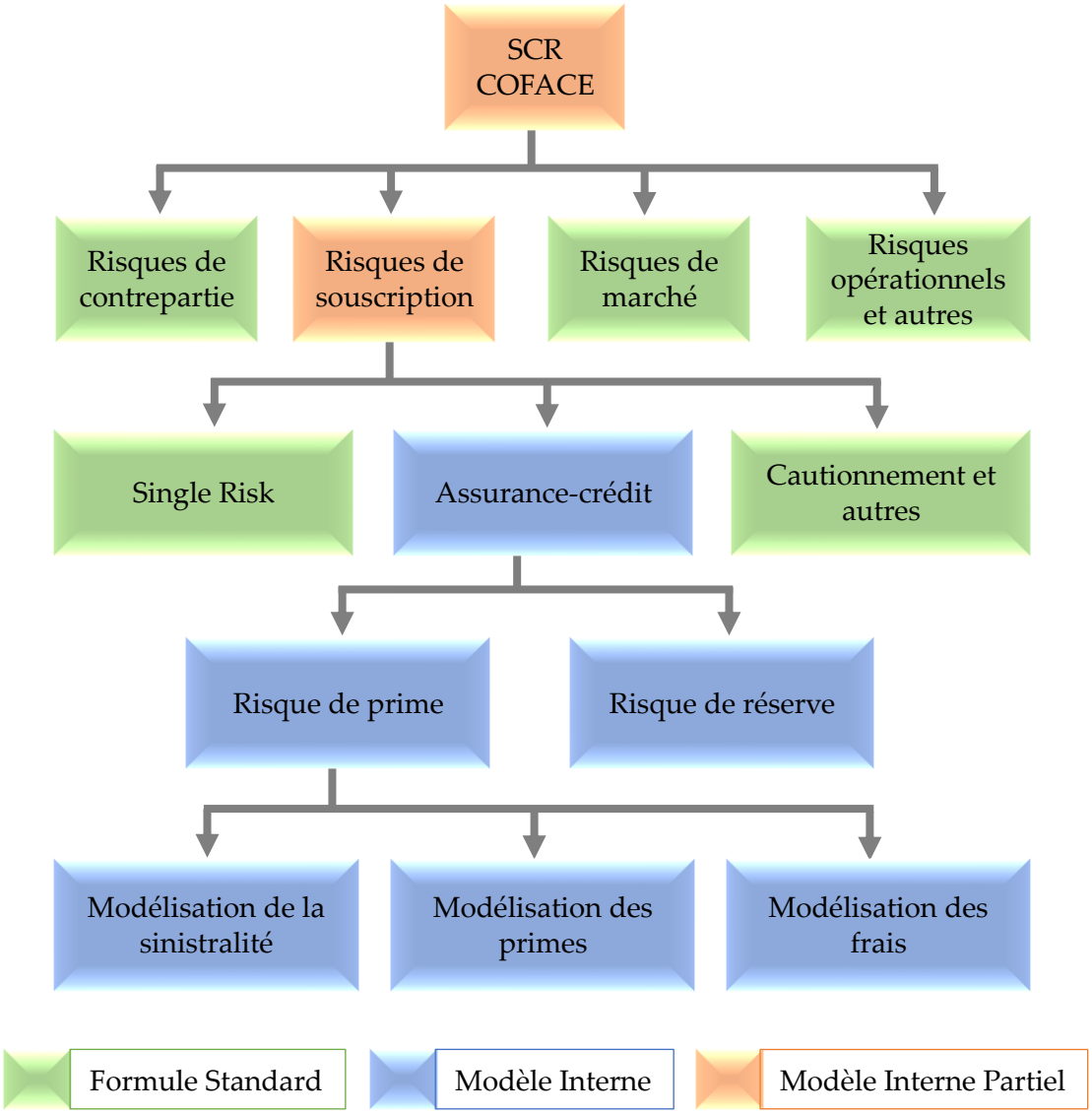


Figure 6 : Périmètre du MIP de la Coface

Comme nous le remarquons, seul le SCR de souscription non-vie est calculé grâce à un modèle interne partiel. Les risques de contrepartie, les risques de marché et les risques opérationnels sont calculés grâce à la formule standard. Du point de vue des activités, le SCR de souscription non-vie est obtenu à partir des SCR des différentes activités de la Coface. Le SCR de souscription de l'activité d'assurance-crédit est calculé grâce à un modèle interne, tandis que les SCR de souscription des autres activités (Single Risk, Cautionnement et autres) sont calculés grâce à la formule standard. Le SCR de souscription globale est ensuite obtenu par agrégation de tous les SCR de souscription à l'aide des coefficients de corrélation calibrés sur des données historiques.

Pour quantifier le besoin en capital nécessaire pour se couvrir contre le risque de souscription non-vie en assurance-crédit, la Coface modélise le risque de prime et le

risque de réserve. La modélisation du risque catastrophe est incluse dans celle des risques de prime et de réserve.

Afin de modéliser et évaluer le risque de prime, trois modélisations doivent être effectuées :

- La modélisation des primes
- La modélisation des frais
- La modélisation de sinistralité (cf. section *2.4*)

Dans la section suivante, nous allons présenter de manière générique le fonctionnement du modèle interne servant à évaluer le risque de souscription de l'activité d'assurance-crédit de la Coface.

2.3. Fonctionnement global du modèle interne partiel étudié

Le modèle interne partiel de la Coface calcule le SCR de souscription de son activité d'assurance-crédit comme une Value-at-Risk du résultat technique en vision économique à horizon 1 an. D'après les résultats obtenus à la section *1.2.4*, on a :

$$SCR_{Souscription\ Assurance-crédit} = -VaR_{0,5\%}(Résultat\ Technique)$$

Pour obtenir le résultat technique, le modèle interne partiel décompose le compte de résultat pour appréhender les risques de l'activité d'assurance-crédit :

- **Le risque de prime et de catastrophe :** Il s'agit des pertes liées à une insuffisance des primes pour couvrir la sinistralité de l'année à venir. Ces pertes potentielles sont évaluées grâce à l'analyse des principaux postes du compte de résultat pour l'année à venir, à savoir :
  - Les primes nettes de participation aux bénéfices ;
  - Les sinistres de l'exercice courant (règlements, frais de gestion des sinistres et provisions) ;
  - Les frais administratifs et commerciaux ;
  - Le solde de réassurance.
- **Le risque de réserve :** Il s'agit des pertes liées à une insuffisance de provisions pour couvrir les sinistres survenus sur les exercices antérieurs, mais non encore réglés.

Le modèle interne partiel génère un grand nombre de simulations des situations économiques probables à 1 an et construit un compte de résultat pour chacune de ces simulations. Sur la base de ces comptes de résultats, une distribution empirique du résultat technique est obtenue. La VaR empirique à 0,5% de cette distribution représente le SCR de souscription de l'activité d'assurance-crédit.

Le tableau ci-dessous présente un exemple de compte de résultat :

				Poste du compte de résultat	Risques Couverts
+	=			**Primes**	
	+			Primes émises	Risque de Primes
	-			Participation aux bénéfices	
-		=		**Charge de sinistralité**	
		+		Sinistres payés durant l'année	Risque de Primes + Risque de Réserve
		+		Variation du Best Estimate de sinistres	
		+		Variation du Best Estimate de primes	Risque de Primes
-			=	**Frais**	
			+	Frais sur primes	Risque de Primes
			+	Frais sur sinistres payés durant l'année	Risque de Primes + Risque de Réserve
			+	Variations du Best Estimate de frais sur sinistres	Risque de Primes + Risque de Réserve
=				**Résultat technique brut de réassurance**	

Tableau 7 : Compte de résultat d'une entreprise d'assurance-crédit

La simulation des situations économiques par le modèle interne partiel nécessite de modéliser chacun des postes du compte de résultat ci-dessus. Les méthodes permettant de modéliser ces postes sont basées sur des hypothèses qui varient selon le poste considéré.

2.3.1. Modélisation des primes

La modélisation des primes est relativement simple. Au fil des années, les primes collectées en assurance-crédit connaissent une évolution assez régulière. La modélisation des primes consiste donc à projeter cette évolution sur une période d'un an. De plus, le modèle interne partiel intègre une variabilité des primes collectées à un an. Cette variabilité est calculée de manière à refléter l'incertitude inhérente aux estimations de primes basées sur les erreurs de prédiction historiques observées entre les budgets et les primes réellement réalisées.

2.3.2. Modélisation de la participation aux bénéfices

La participation aux bénéfices correspond aux ristournes de primes accordées par la Coface à ses assurés en cas de faible sinistralité sur leur police. Elle est modélisée via un pourcentage déterministe des primes acquises à un an.

2.3.3. Modélisation des frais

La modélisation des frais distingue deux types de frais :

- **Les frais considérés comme sensible aux primes :** Il s'agit des frais d'acquisition des polices et des frais de gestion des polices. Ils sont modélisés via un pourcentage déterministe des primes acquises à un an ;
- **Les frais sensibles aux sinistres réglés :** Il s'agit de l'ensemble des frais engagés pour le règlement d'un sinistre ou pour le recouvrement. Ils sont également modélisés via un pourcentage déterministe des sinistres à un an.

2.3.4. Modélisation des Best Estimates

Les Best Estimates, dans ce cas, renvoient aux meilleures estimations des provisions. Il peut s'agir des provisions pour primes (Best Estimates de primes) ou pour sinistres à payer (Best Estimates de sinistres). Ils sont modélisés dans le modèle interne partiel au moyen d'une procédure Bootstrap. Cette méthode consiste à rééchantillonner aléatoirement des observations un grand nombre de fois afin de générer une longue série de réalisations probables de ces observations. Le Bootstrap permet ainsi de réaliser un grand nombre de simulations des règlements de sinistres ou des acquisitions des primes par année de rattachement. Les Best Estimates sont ensuite calculés à partir de ces données simulées.

2.3.5. Application de la réassurance

Après le calcul des résultats techniques bruts de réassurance, l'étape suivante est l'application des traités de réassurance. Le résultat de réassurance est obtenu comme suit :

		Poste du compte de résultat	Risque couvert
+	=	**Charge de sinistres cédée**	
	+	Sinistres cédés	Risque de Primes + Risque de Réserve
	+	Variation de BE de sinistres cédés	
	+	Variation de BE de primes cédés	Risque de Primes
+		**Frais sur sinistres payés cédés**	Risque de Primes + Risque de Réserve
-		**Primes cédées**	Risque de Primes
+		**Commissions rétrocédées**	Risque de Primes + Risque de Réserve
=	**Résultat de Réassurance**		

Tableau 8 : Compte de résultat de réassurance d'une entreprise d'assurance-crédit

La modélisation de la réassurance consiste à appliquer les traités et les conditions contractuelles qui la régissent. Un module dédié du modèle interne partiel récupère la sinistralité et les éléments comptables nécessaires à l'application des traités, puis applique la réassurance telle que définie dans les contrats. Le résultat technique net de réassurance est ensuite obtenu comme suit :

	Poste du compte de résultat	**Risque Couvert**
+	**Résultat technique brut de Réassurance**	Risque de Primes + Risque de Réserve
+	**Résultat de réassurance**	Risque de Primes + Risque de Réserve
=	**Résultat technique net de réassurance**	

Tableau 9 : Résultat net de réassurance d'une entreprise d'assurance-crédit

2.4. Modélisation de sinistralité et présentation des phénomènes segmentés

La modélisation de la sinistralité vise à produire des estimations fiables des pertes liées aux sinistres nettes de recouvrement générées par un portefeuille à 1 an. La modélisation des pertes financières est une activité fréquente dans le domaine bancaire et dans d'autres secteurs de l'assurance. Toutefois, l'assurance-crédit contient un certain nombre de spécificités qui doivent être prises en compte afin d'obtenir des estimations adaptées à cette activité. A titre de rappel, l'une de ces spécificités est la possibilité pour plusieurs assurés d'avoir des contrats d'assurance sur un même acheteur. En cas de défaut, la perte ultime pour un acheteur est donnée par :

$$Perte\ ultime = Expo \times UGD \times CS \times LGD$$

Où :

- $Expo$ est l'exposition connue de l'acheteur ;
- UGD[11] est le ratio représentant la part de l'exposition réellement en défaut pour cet acheteur ;
- CS[12] est la proportion du défaut garantie par la Coface (Exemples : Exclusions via quotités et/ou clauses contractuelles) ;
- LGD[13] est la proportion du défaut non récupérée à l'ultime par la Coface (1 – Taux de recouvrement et récupération).

La modélisation de la sinistralité passe donc par la modélisation des défauts à travers la probabilité de défaut (PD), la modélisation de la proportion d'utilisation de

[11] *Usage Given Default*
[12] *Contractual Specifications*
[13] *Loss Given Default*

la garantie (UGD), la modélisation des spécificités contractuelles (CS) et la modélisation du taux de perte (LGD). Ces phénomènes sont modélisés à travers des simulations pour chaque acheteur couvert par la Coface dans un grand nombre de situations économiques, elles aussi simulées dans le cadre du modèle interne partiel.

La maille de simulation retenue par la Coface est la maille acheteur. Elle permet de capturer les caractéristiques des expositions qui portent le risque sous-jacent à l'activité. Lorsqu'un acheteur est défaillant, l'ensemble de ses polices est considéré en défaut. Cette hypothèse prudente est motivée par le fait que lorsqu'une entreprise n'arrive pas à payer ses dettes auprès d'un assuré, il y a de plus fortes chances qu'elle n'arrive pas à payer ses dettes auprès du reste des assurés avec lesquels elle entretient des relations commerciales. Des taux d'UGD, de CS et de LDG sont ensuite simulés puis appliqués aux expositions de cet acheteur. Une tâche très importante à réaliser avant la simulation de ces 4 phénomènes (PD, UGD, CS et LGD) est la segmentation. Les segmentations sont spécifiques à chaque phénomène.

2.4.1. Modélisation de la probabilité de défaut (PD)

La Coface utilise l'approche de Merton[14] pour générer des défauts dans le cadre de son modèle interne partiel. Cette approche mesure la capacité d'une entreprise à rembourser sa dette à partir de la valeur de ses actifs. D'après Vasicek (1987)[15], une entreprise fait défaut si le niveau de ses actifs franchit un seuil de défaut calibré sur la base des probabilités de défaut historiques.

De plus, différentes études macro-économiques démontrent un risque de contagion significatif du défaut (Duffie, 2011). En effet, des entreprises ont tendance à faire défaut simultanément du fait de leur dépendance à des facteurs de risque communs (Das, et al., 2002). Ces facteurs peuvent être économiques, sectoriels ou géographiques. L'approche de Merton a l'avantage de tenir compte de ce phénomène. Elle considère que la capacité d'un acheteur à rembourser sa dette dépend d'un facteur propre à cet acheteur (facteur idiosyncratique) et d'un facteur commun à tous les acheteurs d'un portefeuille ou d'un segment (facteur systémique). Les corrélations des performances permettent de mesurer le lien entre les facteurs systémiques et les performances d'un acheteur. Ces corrélations sont très importantes pour quantifier la vraisemblance de la défaillance jointe de deux acheteurs appartenant au même portefeuille ou au même segment. Un portefeuille avec des corrélations de performance élevées produira plus de défaut qu'un portefeuille avec des corrélations de performances plus faibles (Somnath, 2015).

[14] *(Merton, 1974)*
[15] *(Vasicek, 1987)*

L'aspect systémique dans le modèle de Merton peut être représenté par un seul facteur (modèle mono-facteur) ou par plusieurs facteurs (modèle multi-facteurs). D'après la littérature, les modèles mono-facteur ne permettent pas de refléter correctement les effets de la segmentation ou de la diversification du portefeuille (Herve, 2002). De plus, il est possible que deux facteurs systémiques très liés aux performances des acheteurs connaissent des cycles économiques différents. Le modèle interne partiel de la Coface retient donc un modèle de Merton multi-facteurs pour la simulation des probabilités de défaut. Les facteurs de risque retenus sont les suivants :

- Un facteur de risque « Monde » commun à tous les acheteurs ;
- Des facteurs de risques spécifiques :
    - au pays de chaque acheteur ;
    - au secteur d'activité de chaque acheteur.

Le modèle de Merton utilisé par la Coface tient également compte des corrélations entre ces différents facteurs. Pour illustrer ces explications, considérons la formule suivante :

$$Y_i = \sum_{j=1}^{K} \rho_{i,j} Z_j + \sqrt{1 - \sum_{j=1}^{K} \rho_{i,j}^2}\, \varepsilon_i$$

Y_i représente les performances (montant des actifs) de l'acheteur i, Z_j est le facteur j commun à tous les acheteurs (systémique), K est le nombre de facteurs de risque systémiques utilisés, ε_i est un facteur de risque propre à l'acheteur i (idiosyncratique) et $\rho_{i,j}$ est la corrélation entre le facteur Z_j et les performances de l'acheteur i. Elle représente la réaction de l'acheteur i à une variation du facteur de risque systémique Z_j.

On dira qu'un acheteur est en défaut si ses actifs sont en dessous d'un certain seuil D_i ($Y_i \leq D_i$). Ce seuil est calibré sur la base des probabilités de défaut historiques PD_i comme suit :

$$\mathbb{P}\left(\sum_{j=1}^{K} \rho_{i,j} Z_j + \sqrt{1 - \sum_{j=1}^{K} \rho_{i,j}^2}\, \varepsilon_i \leq D_i \right) = PD_i$$

En supposant que Z et ε_i suivent des lois normales $\mathcal{N}(0,1)$ on a :

$$D_i = \mathcal{N}^{-1}(PD_i)$$

La Coface calcul la probabilité de défaut historique d'un acheteur i comme suit :

$$PD_i = \frac{\sum_t \sum_{\alpha \in i} EAD_{t,\alpha} \mathbb{1}_{\{def_{t,\alpha}\}}}{\sum_t \sum_{\alpha \in i} Expo_{t,\alpha}^{aq}}$$

Où :

- α est un agrément et $\alpha \in i$ signifie que l'agrément α appartient à l'acheteur i ;
- $Expo_{t,\alpha}^{aq}$ est l'exposition acquise de l'année t sur l'agrément α (cf. section *1.1.2*) ;
- $EAD_{t,\alpha}$ est l'exposition en défaut de l'année t sur l'agrément α (cf. section *1.1.2*) ;
- $\mathbb{1}_{\{def_{t,\alpha}\}}$ prend la valeur 1 si l'agrément α est en défaut au cours de l'année t et 0 sinon.

Après le calibrage des seuil D_i, des corrélations $\rho_{i,j}$ et des autres éléments intervenant dans la formule ci-dessus, l'étape suivante est la simulation des défauts à partir des lois de probabilité des Z_j et des ε_i. Toutefois, le calibrage et l'application du modèle à facteurs nécessitent de disposer de groupes homogènes en probabilité de défaut ainsi que des facteurs de risques qui leur sont associés. Le calibrage des paramètres s'effectue à l'intérieur de chaque segment.

Pour construire ces groupes homogènes, des algorithmes de segmentation sont utilisés. Ces algorithmes utilisent des méthodes statistiques calibrées sur des variables soigneusement choisies dans le but de produire les segments les plus discriminants possibles.

2.4.2. Modélisation des autres phénomènes (UGD, LGD et CS)

Dans la section précédente, nous avons présenté la méthodologie permettant de modéliser et de générer des défauts. Cette section s'applique uniquement aux acheteurs en défaut obtenus à la section précédente. Le but est de présenter le passage du défaut d'un acheteur au montant ultime des sinistres à payer.

Dans le modèle interne partiel de la Coface, une approche très granulaire est adoptée. En effet, l'UGD, la LGD et les CS sont calibrés à la maille agrément. De plus, ces trois phénomènes sont modélisés séparément et de manière interdépendante.

2.4.2.1. Le taux d'utilisation de la garantie (UGD)

L'UGD est défini comme le niveau d'utilisation de la garantie au moment du défaut. L'UGD d'un agrément α est donné par :

$$UGD_\alpha = \frac{NOA_\alpha}{EAD_\alpha}$$

Où NOA_α est le montant de sinistre déclaré par l'assuré pour l'agrément α (Notification of Overdue Amount) et EAD_α est l'exposition en défaut sur cet agrément.

Après le calcul de l'UGD, les agréments sont regroupés en classes homogènes d'UGD grâce à des algorithmes de segmentation. Ensuite, une loi de probabilité est

ajustée par segment sur les distributions empiriques de l'UGD. Cette loi est choisie parmi les options suivantes :

- Lognormale ;
- Weibull ;
- Gamma ;
- Exponentielle.

Comme pour la PD, la loi de probabilité retenue permet d'effectuer des simulations d'un grand nombre d'UGD à l'intérieur de chaque segment. Les segments retenus sont issus d'un algorithme de segmentation reposant sur des variables explicatives caractérisant l'UGD.

2.4.2.2. Les spécificités contractuelles (CS)

Le taux de CS correspond au taux de conversion de l'exposition en défaut à la perte maximale possible relative à la déclaration d'un sinistre. Le taux de CS pour un agrément α est donné par :

$$CS_\alpha = \frac{FR_\alpha}{min(NOA_\alpha, EAD_\alpha)}$$

Où

- FR_α correspond à la première réserve pour l'agrément α. Elle représente le montant de la perte possible pour la Coface hors récupération et après application des spécificités contractuelles ;
- NOA_α est le montant de sinistre déclaré par l'assuré pour l'agrément α ;
- EAD_α est l'exposition en défaut sur l'agrément α.

Le taux de CS quantifie les clauses suivantes :

- Les caractéristiques contractuelles : franchises, quotités garanties, etc. ;
- L'inéligibilité d'un sinistre (absence de couverture effective) ;
- Le paiement d'une proportion des factures entre la date de notification de l'assureur (NOA) et le calcul de la première réserve.

La modélisation des CS est séparée en 2 composantes. La première composante modélise la survenance d'un sinistre sans suite (première réserve nulle) par une loi de Bernoulli. Comme pour l'UGD, cette loi est calibrée par segments de CS homogènes. La deuxième composante applique des taux déterministes calculés par segment de CS homogènes sur les agréments ayant des premières réserves non nulles. Les sinistres sans suite peuvent être dus aux raisons suivantes :

- Retards de déclaration sur certaines factures ;

- Déclaration de factures sur une activité non couverte ;
- Facturation à un acheteur non prévu dans la police.

Les segments sont issus d'un processus identique à celui implémenté pour l'UGD.

2.4.2.3. Le taux de perte (LGD)

La LGD tient compte des récupérations avant et après indemnisation par la Coface. Elle permet de passer de la première réserve à la perte ultime après récupérations et recouvrements. La LGD pour un agrément α se calcule via la formule :

$$LGD_\alpha = 1 - \frac{FR_\alpha - Indemn_\alpha + Recov_\alpha}{FR_\alpha}$$

Où :

- FR_α correspond à la première réserve pour l'agrément α
- $Indemn$ est le montant des indemnités versées ;
- $Recov$ est le montant de recouvrement post-indemnité

Une loi Beta est ajustée sur la distribution empirique des LGD par segment. Ces segments sont obtenus à partir d'un processus identique à celui utilisé pour l'UGD. Tout comme l'UGD et la CS, l'appartenance à un segment de LGD homogènes est déterminée par un certain nombre de variables explicatives.

La figure suivante résume le passage de l'exposition à la perte ultime grâce aux 4 phénomènes (PD, UGD, CS et LGD) présentés ci-dessus :

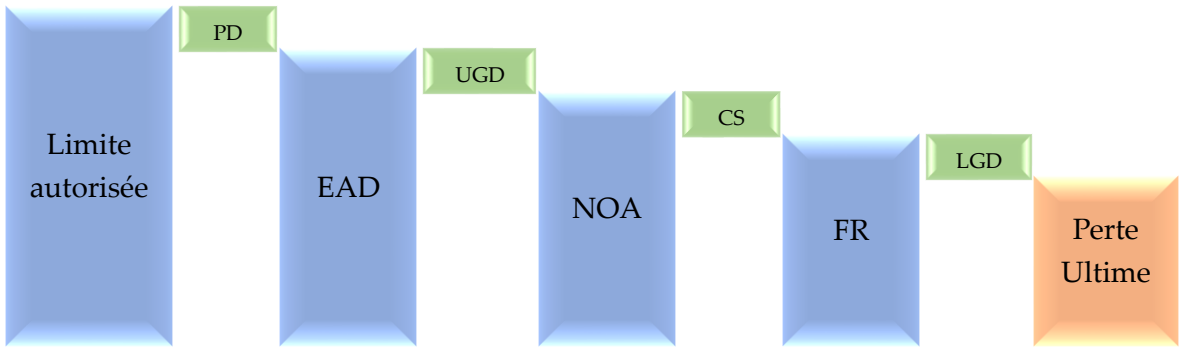


Figure 7 : Modélisation de la sinistralité en assurance-crédit

Nous remarquons que la modélisation de chaque phénomène qui intervient dans le calcul de la perte ultime de la Coface nécessite une segmentation. Cette segmentation de chaque phénomène permet de construire des classes de risque homogènes. Le calibrage des lois de probabilité à l'intérieur de ces segments permet de produire des simulations plus robustes et plus fiables. Dans le *Chapitre 3*, nous utiliserons la littérature académique et les travaux empiriques pour justifier davantage l'utilité de la segmentation.

2.5. Méthodologie de segmentation actuelle

Dans cette section, nous présenterons la méthode de segmentation utilisée par la Coface ainsi que les motivations de l'étude réalisée dans le cadre de ce mémoire. Les quatre phénomènes liés à la sinistralité, à savoir la probabilité de défaut (PD), le taux d'utilisation de la garantie (UGD), les spécificités contractuelles (CS) et le taux de perte (LGD), sont segmentés à l'aide des arbres de classification CART[16] (que nous appellerons « *modèle CART* » dans la suite de nos travaux). Les variables cibles dans ces modèles CART sont des discrétisations des phénomènes obtenues grâce à la méthode des quartiles et aux avis d'experts.

L'objectif des modèles CART dans ce cas n'est pas de prédire ces variables cibles, mais de regrouper en segments des individus (acheteurs ou agréments) ayant des valeurs proches pour les phénomènes considérés. Les feuilles des arbres CART correspondent aux regroupements de ces individus. Les arbres sont construits grâce à des variables explicatives soigneusement choisies. Le choix de ces variables explicatives est effectué sur la base de leur liaison avec les phénomènes considérés. Plus de détails sur les arbres de classification CART, la sélection des variables explicatives, la construction des variables cibles et l'évaluation de la qualité des segmentations sont présentés dans les *Chapitre 4* et *Chapitre 5* de ce mémoire.

2.6. Intérêt de notre étude

L'ACPR, ainsi que certaines études statistiques telles que (Kassambara, 2018) et (Amit, 2017), se posent des questions concernant la stabilité des modèles CART. En effet, le régulateur se demande si les structures de segmentation issues des modèles CART restent robustes face à de légères perturbations dans les données utilisées pour les construire.

La Coface souhaite donc répondre à cette interrogation grâce à une étude approfondie de la stabilité de ses modèles de segmentation. C'est dans ce contexte que s'inscrivent les travaux de ce mémoire. Dans la partie suivante, nous définirons la stabilité, ainsi que les métriques et les méthodes permettant de l'évaluer. Nous présenterons également une méthodologie pour étudier la stabilité d'un modèle de segmentation, ainsi que les facteurs susceptibles d'influencer cette stabilité. Dans la *Partie 3* : Résultats de l'étude, nous appliquerons cette méthodologie à l'un des phénomènes présentés ci-dessus afin de lui apporter une validation pratique.

[16] *Classification and Regression Trees (Genuer, et al., 2017)*

Partie 2 : Bases théoriques de l'étude

Chapitre 3 : Littérature sur la segmentation et sur la stabilité des modèles de segmentation

Sommaire

Dans le chapitre précédent, nous avons vu que la modélisation des 4 phénomènes intervenant lors de la sinistralité en assurance-crédit (PD, UGD, CS et LGD) nécessitait une étape de segmentation. Dans ce chapitre, nous allons parcourir la littérature scientifique afin de comprendre l'utilité de la segmentation. Nous aborderons également, grâce à la littérature, la problématique de la stabilité des modèles de segmentation évoquée précédemment.

3.1. La segmentation

Dans cette section, nous allons définir la segmentation, présenter les différents types, et analyser son utilité d'après la littérature.

3.1.1. Définition

La segmentation est un processus de partitionnement des données en sous-ensembles de telle sorte que les données dans chaque sous-ensemble soient les plus similaires possibles. Cette similarité est calculée grâce à une métrique de distance prédéfinie (Madhulatha, 2012). Les données utilisées pour construire les segments ne contiennent généralement pas de variable cible. L'objectif est d'obtenir des groupes de données qui sont très similaires à l'intérieur de chaque groupe, mais très différents d'un groupe à un autre (Cormack, 1971).

Il est important de noter que la segmentation est appliquée dans plusieurs domaines tels que les analyses statistiques, le marketing, la pharmacie, le traitement d'image, et bien d'autres (Milligan, et al., 1987). La définition que nous avons donnée ci-dessus correspond le mieux à la problématique que nous allons résoudre dans le cadre de ce mémoire.

Pour un problème relativement simple, tel que la division de 25 observations en 5 segments distincts, il existe environ $2{,}4 \times 10^5$ solutions possibles (Anderberg, 1973). Les méthodes de segmentation doivent donc trouver la répartition optimale en utilisant des outils qui évitent de tester toutes les possibilités. Il existe des centaines d'algorithmes de segmentation, mais aucun n'est systématiquement meilleur que les autres. C'est pourquoi, pour un jeu de données que l'on souhaite segmenter, il est important d'identifier et d'implémenter des métriques et des méthodes de validation adaptées. La meilleure segmentation sera celle qui se démarque du lot à cette étape de validation (Milligan, et al., 1987).

La construction d'une segmentation sur un jeu de données intervient après deux étapes préliminaires :

- La confirmation de la présence d'une structure segmentée dans les données.
- Le choix du nombre optimal de segments à construire.

En pratique, la première étape est souvent ignorée, car les besoins conduisent à produire des segments sur les données disponibles, en mettant éventuellement à jour la segmentation au fur et à mesure que le volume de données augmente et que des frontières plus claires apparaissent.

En revanche, la deuxième étape est très importante et est quasiment toujours prise en compte. En effet, un nombre de segments trop bas produit une segmentation avec beaucoup d'hétérogénéité à l'intérieur de chaque segment, tandis qu'un nombre de segments trop élevé produit beaucoup d'homogénéité entre les segments. Ces deux situations sont justement ce que l'on cherche à éviter lors de la construction d'une bonne segmentation.

3.1.2. Les étapes d'une segmentation

D'après les travaux d'Anderberg (1973), Cormack (1971), Everitt (1980) et Lorr (1983)[17], un problème de segmentation peut être regroupé en 7 étapes. Bien que l'application de ces 7 étapes puisse varier d'un contexte à un autre, la construction de toute segmentation devrait passer par ces étapes :

1. **Sélection des individus à segmenter :** La segmentation doit se faire dans un contexte bien connu et sur des observations ou individus sélectionnés à l'avance. Ceci facilitera la validation de la segmentation et l'interprétation des résultats à la fin du processus ;
2. **Sélection des variables qui serviront à effectuer la segmentation :** Les variables doivent caractériser les individus et le phénomène d'intérêt s'il y en a.

[17] *(Anderberg, 1973), (Cormack, 1971), (Everitt, 1980), (Lorr, 1983)*

Ces variables doivent également contenir suffisamment d'information pour rendre la segmentation possible ;

3. **Traitement des variables :** Il s'agit de la préparation des données et de la prise en compte de tous les facteurs pouvant influencer la construction des segments. De tels facteurs peuvent être les valeurs aberrantes ou les valeurs manquantes ;
4. **Choix d'une mesure de similarité ou de dissimilarité :** Cette mesure reflète le niveau d'homogénéité ou d'hétérogénéité entre les observations. Le choix de cette mesure doit être fait en tenant compte de la nature des variables disponibles (quantitatives, qualitatives ou mixtes) ;
5. **Choix de la méthode ou du modèle de segmentation :** Cette étape est intimement liée à la précédente, car le choix de la méthode de segmentation dépend du choix de la mesure de similarité/dissimilarité et également de la nature et la structure des données disponibles ;
6. **Détermination du nombre optimal de segments :** Comme mentionné précédemment, le nombre de segments ne doit être ni trop faible ni trop élevé. C'est un point auquel la littérature sur la segmentation a accordé une attention particulière au cours des dernières années ;
7. **Validation de la segmentation :** Cette étape consiste à tester, valider et interpréter la segmentation obtenue. L'interprétation de la segmentation doit se faire dans le contexte fixé initialement dans le but de s'assurer que les résultats obtenus font sens. C'est également à cette étape que les choix effectués précédemment sont acceptés ou rejetés.

3.1.3. Les mesures de distance

Il existe plusieurs mesures de distance en segmentation, mais deux sont plus fréquemment utilisées que les autres : la distance de Manhattan et la distance euclidienne (Madhulatha, 2012).

- **La distance de Manhattan :** Elle correspond à la somme des valeurs absolues des écarts entre chaque variable pour deux observations. Soient $X = (x_1, x_2, \dots, x_p)$ et $Y = (y_1, y_2, \dots, y_p)$, la distance de Manhattan entre ces deux observations est donnée par :

$$d = \sum_{i=1}^{p} |x_i - y_i|$$

- **La distance euclidienne :** Elle correspond à une distance à vol d'oiseau entre deux observations. Pour deux observations $X = (x_1, x_2, \dots, x_p)$ et $Y = (y_1, y_2, \dots, y_p)$, la distance euclidienne entre elles est donnée par :

$$d = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

Ces deux distances sont illustrées pour $p = 2$ sur la figure suivante :

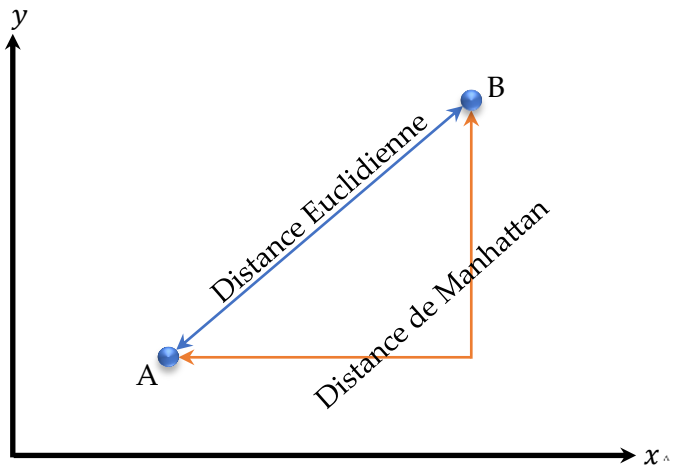


Figure 8 : Deux exemples de mesure de distance

Il est important de rappeler qu'il s'agit là des deux mesures les plus utilisées dans les algorithmes de segmentation de façon générale. En effet, certains algorithmes de segmentation peuvent avoir leurs propres mesures de distance. Les mesures présentées ci-dessus sont fournies à titre illustratif et ont pour but de faciliter la compréhension des notions abordées dans ce chapitre.

3.1.4. Le choix du nombre de segments

Nous avons vu dans les sections précédentes que le choix du bon nombre de segments est crucial pour produire une segmentation optimale. Certains algorithmes de segmentation exigent de préciser le nombre de segments avant leur exécution, tandis que d'autres algorithmes sélectionnent automatiquement le meilleur nombre de segments. Dans les deux cas, ces nombres de segments doivent être validés. La méthode la plus utilisée dans la littérature pour choisir et/ou valider un nombre de segments est la méthode du coude (Diday, Edwin, et al., 1976).

La méthode du coude consiste à choisir le nombre de segments tel que l'ajout d'un segment supplémentaire améliore très peu la qualité de la segmentation. Cette qualité est captée grâce à une mesure convenablement choisie. Les mesures de qualité peuvent porter sur des indicateurs statistiques tels que l'indice de Gini, la variance, le MSE (Mean Squared Error), etc., ou des indicateurs liés à l'activité qui a généré les données segmentées.

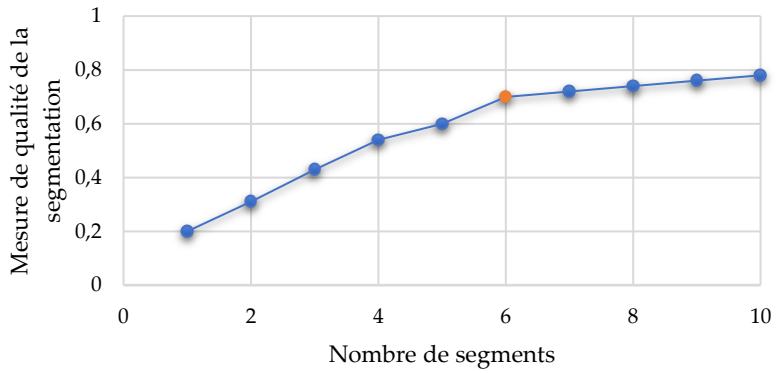


Figure 9 : La méthode du coude

Dans le cas illustré sur la figure ci-dessus, nous choisirons 6 segments, car au-delà de 6, on n'observe pas d'amélioration considérable de la mesure de qualité de la segmentation.

3.2. Les algorithmes de segmentation

D'après Bair (2013)[18], il existe deux grandes familles d'algorithmes de segmentation. D'une part, nous avons les algorithmes de segmentation supervisée, et d'autre part, les algorithmes de segmentation non-supervisée.

3.2.1. Les algorithmes de segmentation supervisée

La majeure partie des études sur la segmentation s'est focalisée sur la segmentation non-supervisée (Awasthi, et al., 2010). Il n'existe donc pas encore de consensus sur une théorie formelle de la segmentation supervisée. Comme son nom l'indique, la segmentation supervisée est un type de segmentation qui incorpore une variable réponse dans le processus de segmentation (Dettling, et al., 2002). La segmentation supervisée utilise la variable réponse pour orienter la construction des segments afin de garantir que ces derniers soient des classes homogènes de ladite variable.

La littérature présente très peu de modèles de segmentation supervisée. Toutefois, un modèle qui revient très souvent est le modèle d'arbre de classification et de régression (CART). Initialement construits pour la classification et la régression, plusieurs études montrent que les CART sont très performants et très pratiques pour effectuer des segmentations supervisées (Hancock, et al., 2003), (Baert, et al., 2007). La méthodologie d'implémentation des CART pour la segmentation reste identique à celle utilisée dans le cas de la régression et de la classification. En revanche, dans le cas

[18] *(Bair, 2013)*

de la segmentation, l'attention est portée sur les feuilles de l'arbre qui constituent les segments, et non sur les prédictions du modèle.

3.2.1.1. Quelques avantages des CART (Hancock, et al., 2003)

- Les CART sont des modèles non-paramétriques et ne nécessitent donc pas d'hypothèses sur la distribution de la variable réponse.
- Les CART ont la capacité de gérer de grandes bases de données.
- Les CART sont intuitifs et produisent des résultats faciles à interpréter.
- L'implémentation des CART est facile et compréhensible par des non-statisticiens.

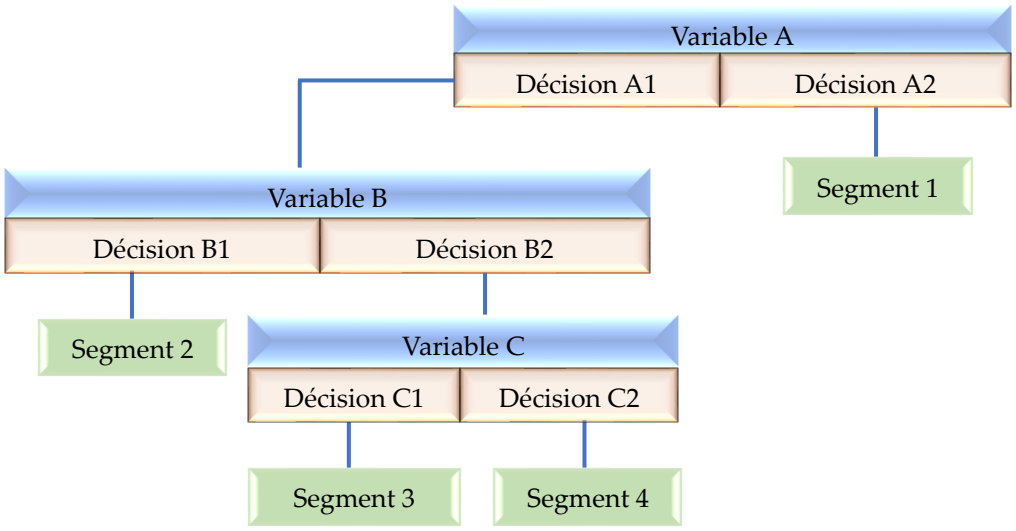


Figure 10 : Illustration du modèle CART

3.2.2. Les algorithmes de segmentation non-supervisée

L'implémentation de ces algorithmes ne nécessite pas de variable réponse. D'après Madhulatha (2012)[19], les différents algorithmes de segmentation non-supervisée peuvent être regroupés en deux types : les algorithmes hiérarchiques et les algorithmes de partitionnement.

3.2.2.1. Les algorithmes hiérarchiques

Ces algorithmes déterminent les segments successifs en utilisant les segments construits précédemment. Les algorithmes hiérarchiques peuvent être agglomératifs ou divisifs. Les algorithmes agglomératifs débutent avec chaque élément comme étant un segment et les fusionnent progressivement (grâce à une mesure de distance) pour former des segments plus grands. Les algorithmes divisifs quant à eux commencent avec toutes les observations dans un segment et les divisent progressivement en

[19] *(Madhulatha, 2012)*

segments plus petits. Après exécution, ces deux algorithmes produisent tous les deux n segments, où n est le nombre d'observations à segmenter. L'étape finale consiste donc à « couper » la hiérarchie construite à un niveau qui permet d'avoir des segments avec une forte homogénéité intra-segments et une forte hétérogénéité inter-segments. Le meilleur point de coupure (et donc le nombre de segments optimal) est choisi sur la base d'une méthode de validation telle que la méthode du coude, présentée précédemment.

D'après la littérature, l'algorithme de segmentation hiérarchique le plus utilisé est la classification ascendante hiérarchique (Chevalier, et al., 2013) que nous appellerons « *modèle CAH* » dans la suite de ce mémoire.

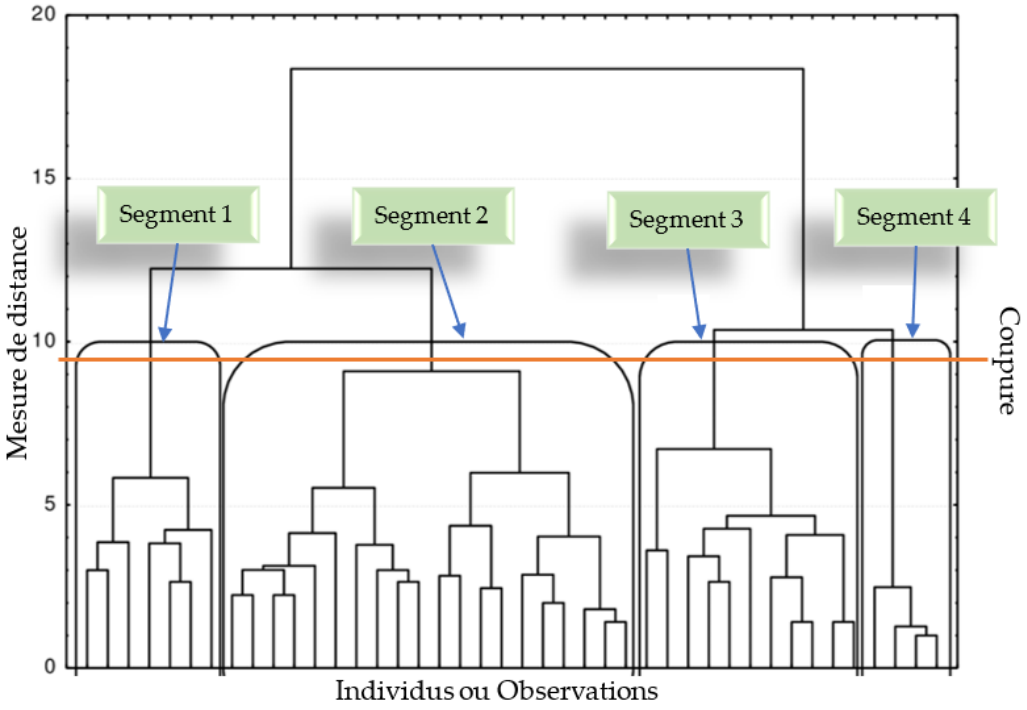


Figure 11 : Illustration du modèle CAH

3.2.2.2. Les algorithmes par partitionnement

Contrairement aux algorithmes hiérarchiques, les algorithmes de partitionnement effectuent les segmentations en une seule phase. La mesure de distance est utilisée pour calculer les distances entre les observations, et la segmentation est construite de manière statique ou itérative (Ball, et al., 1965). D'après Blashfield (1977)[20], les algorithmes de segmentation par partitionnement se distinguent grâce à cinq caractéristiques :

- **Le point de départ de la segmentation :** Certains algorithmes, tels que les k-means, choisissent le point de départ aléatoirement, tandis que d'autres

[20] *(Blashfield, 1977)*

algorithmes permettent à l'utilisateur de choisir le point de départ (Jancey, 1966) ;

- **Le nombre de passage dans la base de données :** Certains algorithmes, comme les k-means, effectuent un seul passage dans la base de données et affectent les observations au segment dont le centroïde est le plus proche, tandis que d'autres algorithmes font plusieurs passages en mettant à jour les centroïdes à chaque passage ;
- **Le critère d'affectation d'une observation à un segment :** Certains algorithmes utilisent un critère unique d'affectation, tel que la distance, tandis que d'autres utilisent plusieurs critères présentés sous forme de matrice ;
- **Le nombre de segments :** La majeure partie de ces algorithmes nécessite une spécification du nombre de segments avant exécution ;
- **La gestion des valeurs aberrantes :** La majeure partie de ces algorithmes affecte les valeurs aberrantes à l'un des segments, et très peu isolent ces points. Il pourrait donc être important de traiter les valeurs aberrantes avant l'exécution de ces algorithmes.

D'après la littérature, l'algorithme de segmentation par partitionnement le plus utilisé est l'algorithme des k-means(Ahmed, et al., 2020).

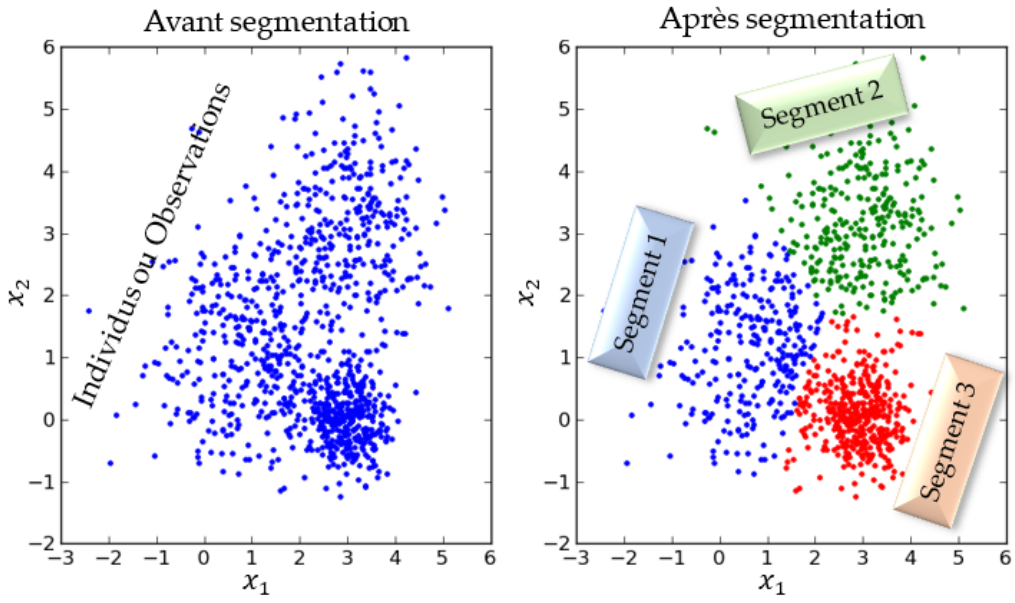


Figure 12 : Illustration du modèle de k-plus proches voisins

3.3. Utilité de la segmentation

D'après Blashfield et Aldenderfer (1978)[21], le nombre d'articles scientifiques utilisant la segmentation est passé de 25 en 1964 à 501 en 1976. Milligan, et al., (1987)[22] remarquent qu'entre 1958 et 1973, plus de 1 600 articles scientifiques sur la segmentation ont été publiés. Ils ajoutent qu'en 1985 uniquement, on avait 1 658 références sur le sujet. Il ne fait donc aucun doute que la segmentation est une méthode extrêmement utilisée en analyse statistique et dans plusieurs autres domaines. Nous ne pouvons donc pas prétendre être capables de présenter l'utilité de la segmentation dans sa globalité. Nous nous limiterons à son utilité dans le cadre de la modélisation de la sinistralité en assurance-crédit.

Nous avons vu dans la section précédente que la segmentation avait la capacité de regrouper des observations d'une base de données en groupes homogènes. C'est cette homogénéité à l'intérieur des groupes dont la modélisation de la sinistralité en assurance-crédit a besoin. En effet, comme vu au *Chapitre 2* de ce mémoire, la modélisation des phénomènes intervenant lors de la sinistralité en assurance-crédit (PD, UGD, CS et LGD) nécessite de calibrer des lois de probabilité sur ces phénomènes. Les lois de probabilité ainsi calibrées servent à faire des simulations dans le but d'obtenir une distribution du résultat technique et de calculer le SCR de souscription (cf. *Chapitre 1*). Le calibrage de ces lois de probabilité nécessite une distribution empirique homogène. À titre d'illustration, considérons la figure suivante :

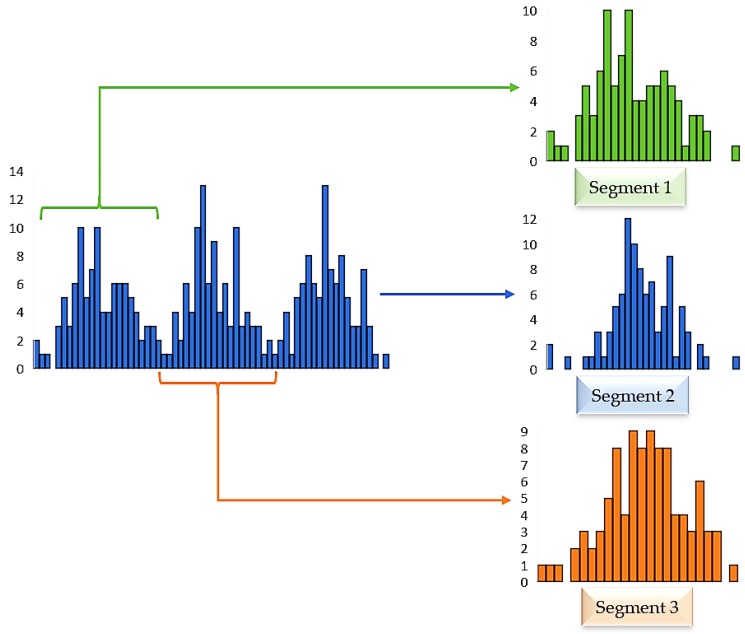


Figure 13 : Illustration de l'utilité d'une segmentation

[21] *(Blashfield, et al., 1978)*
[22] *(Milligan, et al., 1987)*

Le phénomène que l'on tente de simuler ici possède une distribution empirique multimodale. Il est donc difficile, voire impossible, de calibrer une loi de probabilité sur cette dernière. Toute tentative produira des simulations irréalistes et difficilement interprétables.

Après la segmentation, nous obtenons des distributions empiriques intra-segment unimodales. Il est plus réaliste et a priori plus facile de calibrer des lois de probabilité sur ces distributions unimodales. Les simulations qui en découleront auront donc de plus grandes probabilités d'être conformes à la réalité.

Maintenant que la segmentation a été présentée en détail et son utilité dans le cadre de notre étude démontrée, nous allons nous intéresser à un sujet lié aux modèles de segmentation : la stabilité. Rappelons que les travaux de ce mémoire visent à étudier la stabilité des modèles de segmentation utilisés dans la modélisation de la sinistralité en assurance-crédit.

3.4. La stabilité et intérêt de son étude

3.4.1. Définition

D'après Von Luxburg (2010)[23], un algorithme de segmentation est dit stable s'il produit des segmentations similaires lorsque appliqué à plusieurs jeux de données générés par le même processus générateur de données sous-jacent. De même, selon Liu, et al. (2022)[24], un algorithme de segmentation est considéré comme stable s'il fournit des segmentations très similaires, voire identiques, lorsqu'il est appliqué à de petites perturbations du jeu de données initial.

Nous remarquons que deux termes reviennent dans ces définitions ainsi que dans plusieurs autres définitions fournies par la littérature (Müller, et al., 2014), (Dolnicar, et al., 2009) : « perturbations » et « similaires ». Ces deux termes constituent la base de toute méthodologie d'analyse de la stabilité d'un modèle de segmentation. Dans la littérature, ces termes sont employés dans plusieurs problématiques. Notre analyse sera orientée uniquement vers la résolution de la problématique de la présente étude.

- **Similarité entre deux segmentations :** Le moyen le plus évident et le plus compréhensible de quantifier la similarité entre deux segmentations est d'utiliser des quantités qui caractérisent ces segmentations. Ces quantités dépendent en général du type de phénomène étudié et du type de données

[23] *(Von Luxburg, 2010)*
[24] *(Liu, et al., 2022)*

disponibles (Choi, et al., 2010). Rand (1971)[25] recommande d'utiliser les valeurs de la mesure de qualité des segmentations pour évaluer leur similarité. Ainsi, on dira que deux segmentations sont similaires si les valeurs prises par la mesure de qualité sont proches pour ces deux segmentations.

- **Perturbations du jeu de données initial :** La littérature présente plusieurs méthodes pour générer des jeux de données « perturbés » à partir d'un jeu de données initial. Les plus répandues sont les méthodes de Bootstrap (Jain, et al., 1987), (Kerr, et al., 2001), (Dudoit, et al., 2003) et les méthodes de sous-échantillonnage (Dudoit, et al., 2002), (Levine, et al., 2001), (Ben-Hur, et al., 2002). Les méthodes de Bootstrap consistent à faire des tirages aléatoires avec remise dans le jeu de données initial. Les jeux de données « perturbés » obtenus dans ce cas ont la même taille que le jeu de données initial. Les méthodes de sous-échantillonnage consistent à faire des tirages sans remise dans le jeu de données initial. Cette méthode génère donc des jeux de données « perturbés » de tailles inférieures au jeu de données initial. Liu, et al. (2022) recommandent d'utiliser les méthodes de sous-échantillonnage pour générer des jeux de données « perturbés ». En effet, les auteurs estiment que les méthodes de Bootstrap introduisent un biais supplémentaire dû au fait que certaines observations peuvent apparaître dans les jeux de données plus d'une fois.

Les figures ci-dessous illustrent le concept de stabilité grâce aux deux termes que nous venons de définir. Notons que les taux utilisés dans cette analyse sont en valeurs absolues :

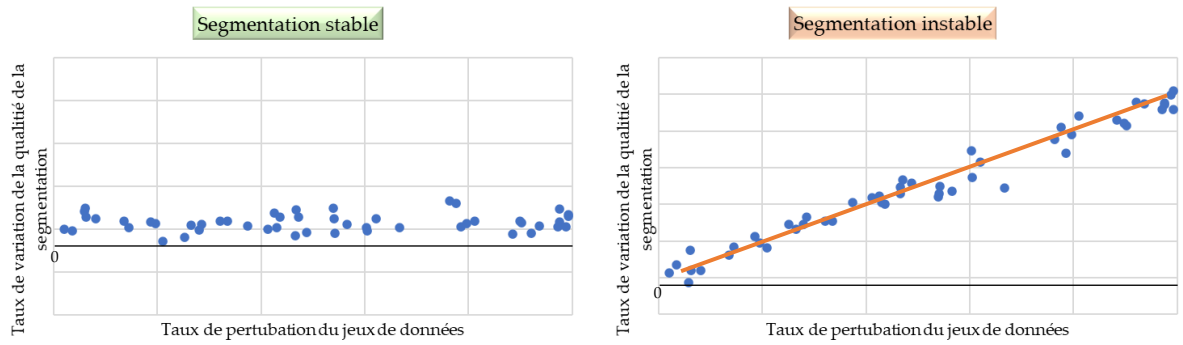


Figure 14 : Analyse de la stabilité d'un modèle de segmentation

Comme indiqué dans la définition de la stabilité, cette illustration concerne des petites perturbations du jeu de données. Le seuil en dessous duquel des perturbations sont considérées comme petites est le niveau maximal de modification du jeu de données anticipé sur la période d'utilisation de la segmentation.

[25] *(Rand, 1971)*

Pour quantifier cette définition de la stabilité, nous proposons un coefficient d'instabilité en nous inspirant de la définition de la stabilité au sens de Lyapunov[26]. Soit m la pente de la courbe linéaire représentée sur la deuxième figure et c son ordonnée à l'origine, de telle sorte que l'équation de cette courbe soit $y = mx + c$. Le coefficient d'instabilité, que nous utiliserons pour comparer nos modèles dans le cadre de ce mémoire sera donné par :

$$Coef\_Instab = \frac{|m| + (100 \times |c|)}{2}$$

Pour un modèle stable, la mesure de qualité de la segmentation prendra des valeurs presque identiques, quel que soit le niveau de perturbation des données. La courbe linéaire sera donc horizontale ($m = 0$) et proche de l'axe des abscisses ($c = 0$). Ainsi, on obtiendra un $Coef\_Instab = 0$ pour un modèle parfaitement stable. Plus un modèle sera instable, plus la valeur du $Coef\_Instab$ sera élevée. Sur la base de nos résultats, nous donnerons un intervalle de valeurs acceptables pour ce coefficient. Notons que ce coefficient a l'avantage d'être simple à calculer et facilement interprétable, ce qui faciliterait son adoption et son utilisation dans des travaux actuariels. Cependant, nous ne nous limiterons pas uniquement à ce coefficient pour analyser la stabilité des modèles de segmentation. Nous proposerons d'autres outils et méthodes dans le *Chapitre 4* qui serviront à la fois d'alternatives et de base d'évaluation de ce coefficient d'instabilité.

3.4.2. Intérêt de l'étude de la stabilité des modèles de segmentation

D'après Homa, et al. (2020)[27], ne pas étudier la stabilité d'un modèle de segmentation revient à prendre le risque que la segmentation obtenue soit contextuelle plutôt que structurelle. En effet, un modèle instable présente un risque élevé de construire une segmentation qui ne reflète pas la réalité. Un tel modèle deviendra obsolète dès que la population segmentée connaîtra le moindre changement. En assurance-crédit, utiliser un modèle de segmentation instable lors du calcul du besoin en capital expose l'entreprise d'assurance au risque de sous-estimer ou de surestimer ce besoin. En cas de sous-estimation, l'entreprise d'assurance est exposée au risque de ne pas pouvoir faire face à ses sinistres au cours de l'année à venir.

L'étude de la stabilité peut également servir d'outil de validation des modèles de segmentation. En plus d'avoir une valeur optimale pour la mesure de qualité de la segmentation choisie, un modèle de segmentation doit être stable. Lorsque deux modèles ont des valeurs proches pour cette mesure de qualité, Liu, et al. (2022) recommandent de choisir le modèle le plus stable.

[26] *(Gaid, et al., 2015)*
[27] *(Homa, et al., 2020)*

3.4.3. Les sources d'instabilité des modèles de segmentation

D'après Shai, et al. (2006)[28], il existe deux sources d'instabilité en segmentation des données. Une source est liée à la structure intrinsèque des données, tandis que l'autre est liée au processus générateur des données segmentées. La première source d'instabilité concerne la présence d'une symétrie dans le jeu de données. En l'absence de symétrie naturelle, les modèles de segmentation construits sur ces données peuvent être instables. Il est donc important, dans ce cas, de choisir le modèle de segmentation qui fournit la segmentation la plus stable possible. La deuxième source d'instabilité concerne la variance et les bruits introduits dans le jeu de données par le processus ou le phénomène qui les a générés. Elle peut également être liée (dans une moindre mesure) aux prétraitements effectués sur les données avant la segmentation. Plusieurs études se sont focalisées sur cette deuxième source d'instabilité et sont arrivées à la conclusion que l'instabilité qu'elle génère diminue avec une augmentation de la taille du jeu de données (Ben-Hur, et al., 2002), (Caponnetto, et al., 2006).

Il ressort de tout ceci qu'il est très difficile en pratique de contrôler la stabilité d'un modèle de segmentation en partant des sources d'instabilité. Les méthodologies utilisées en pratique consistent à analyser la stabilité de plusieurs modèles de segmentation candidats et de choisir celui qui fournit la segmentation la plus stable (Von Luxburg, 2010). La méthodologie utilisée dans ce mémoire pour étudier la stabilité des modèles de segmentation (présentée au chapitre suivant) sera élaborée dans ce même ordre d'idées.

[28] *(Shai, et al., 2006)*

Chapitre 4 : Méthodologie de l'étude

Sommaire

À ce stade, le contexte ainsi que les objectifs de l'étude sont définis. De plus, nous avons justifié les raisons sous-jacentes à cette étude par le biais d'une revue minutieuse de la littérature. Dans ce chapitre, nous allons synthétiser toutes ces informations afin de proposer une méthodologie qui sera employée pour analyser la stabilité des modèles de segmentation dans le domaine de l'assurance-crédit et au-delà. Nous exposerons également les bases mathématiques, statistiques et actuarielles des modèles et des métriques que nous utiliserons. La figure suivante offre une vue sommaire de l'approche méthodologique adoptée pour notre étude :

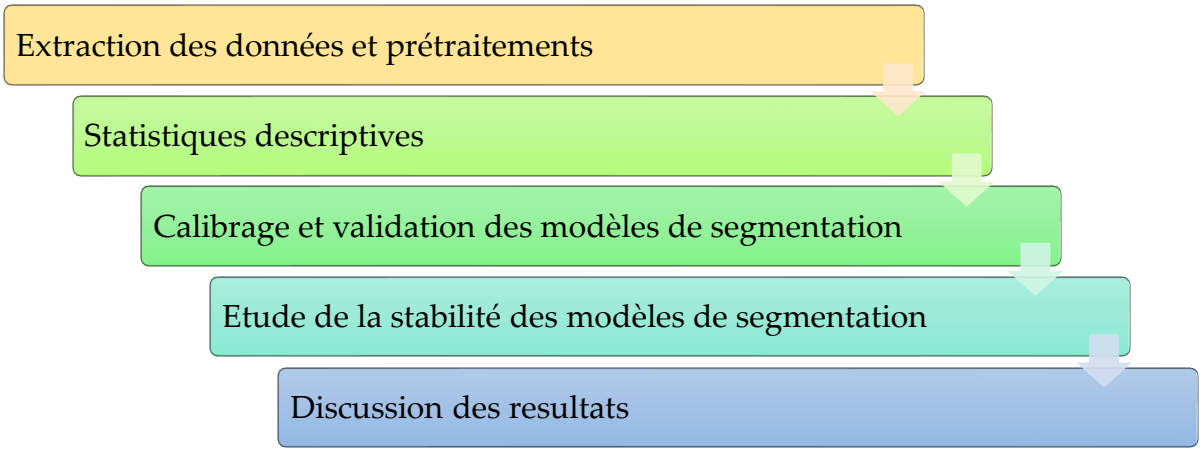


Figure 15 : Méthodologie de l'étude

4.1. Extraction des données, prétraitements et statistiques descriptives

Ces étapes consistent à présenter les bases de données de travail ainsi que leurs sources. Le *Chapitre 5* leur est entièrement dédié. Il s'agira d'effectuer des analyses de base et des prétraitements sur les données afin d'obtenir des bases de données adaptées aux modélisations ultérieures. Les analyses descriptives obtenues à cette étape permettent de saisir la nature des données, ce qui facilite la préparation des étapes suivantes.

4.2. Méthodologie retenue pour l'étude de la stabilité d'un modèle de segmentation

Selon la littérature exposée dans le chapitre précédent, un modèle de segmentation est considéré comme stable lorsque de petites perturbations appliquées au jeu de données utilisé pour sa construction n'engendrent pas de changements significatifs dans la qualité et la structure de la segmentation (Liu, et al., 2022). La méthodologie que nous adoptons pour cette étude est basée sur cette définition.

Au début des analyses, nous disposons de deux bases de données : la base A et la base B. La base A est utilisée pour entraîner et valider les modèles de segmentation. Elle est constituée des informations sur les acheteurs issues des différentes entités Coface. La base B contient des données de clôture qui sont utilisées dans le calcul du SCR de souscription. Plus de détails sur ces bases sont donnés dans les sections *5.1* et *6.7*. La première étape consiste à diviser la base A en deux grandes parties : l'une pour la construction des modèles (échantillon d'apprentissage) et l'autre pour l'évaluation de leurs performances (échantillon de validation). Actuellement, Coface effectue cette division en utilisant des échantillonnages aléatoires stratifiés sans remise, allouant 80% aux données d'apprentissage et 20% aux données de validation. Pour tenir compte des réalités métier et de la dimension temporelle (Salazar, et al., 2022), la Coface construit les strates pour cette division en se basant sur l'année et les zones géographiques des acheteurs.

4.3. Méthodologie de calibrage des modèles de segmentation candidats

En s'appuyant sur les travaux empiriques exposés dans le chapitre précédent, nous avons sélectionné les modèles de segmentation suivants pour notre étude :

- Le modèle d'arbres de classification et de régression (CART).

- Le modèle de k-prototypes : une variante des modèles k-means conçue pour prendre en compte les variables catégorielles.
- Le modèle de classification ascendante hiérarchique.

Comme indiqué précédemment, les modèles des k-prototypes et de classification ascendante hiérarchique sont non-supervisés et nécessitent des modèles supervisés pour appliquer les segmentations développées sur les échantillons d'entraînement à de nouvelles données. Dans chaque cas, l'un des modèles suivants sera retenu pour assurer ce rôle et sera désigné comme le modèle de transfert :

- Le modèle de forêts aléatoires.
- Le modèle d'eXtreme Gradient Boosting (XGBoost) que nous appellerons « *modèle XGBoost* » dans la suite de nos travaux.

Ces modèles sont exposés en détail dans la section *4.5.1*.

Avant d'examiner la stabilité de ces modèles de segmentation, il est essentiel de les calibrer en fonction de nos données. Le processus de calibrage impliquera de déterminer les nombres optimaux de segments pour les modèles de segmentation et de sélectionner les hyperparamètres optimaux pour les modèles de transfert. Les algorithmes de calibrage sont présentés ci-dessous.

4.3.1. Algorithme de calibrage du modèle CART

i. Construire un arbre de classification et de régression (CART) sur l'échantillon d'apprentissage ;
ii. Segmenter de l'échantillon d'apprentissage à l'aide de ce modèle ;
iii. Calculer la métrique de qualité (Gini modifié) pour cette segmentation ;
iv. Répéter les étapes i à iii pour divers nombres de segments (feuilles de l'arbre) ;
v. Sélectionner le nombre de segments qui engendre la plus grande amélioration de la métrique de qualité (méthode du coude).

À ce stade, nous disposons du nombre optimal de segments (ou d'une gamme de nombres de segments fournissant des résultats acceptables) pour le modèle CART. Pour reproduire une segmentation obtenue en calibrant un modèle CART sur un jeu d'entraînement, il est nécessaire de construire des tables de correspondance (*mapping en anglais*). Ces tables permettent de segmenter un nouveau jeu de données en se basant uniquement sur les variables explicatives. Une table de correspondance présente la structure suivante :

Variable 1	**Variable 2**	**…**	**Variable p**	**Segment**
Modalités : $A1_1, A1_2, \dots, A1_{n_{A1}}$	Modalités : $A2_1, A2_2, \dots, A2_{n_{A2}}$	…	Modalités : $Ap_1, Ap_2, \dots, Ap_{n_{Ap}}$	A
Modalités : $B1_1, B1_2, \dots, B1_{n_{B1}}$	Modalités : $B2_1, B2_2, \dots, B2_{n_{B2}}$	…	Modalités : $Bp_1, Bp_2, \dots, Bp_{n_{Bp}}$	B
⋮	⋮	⋱	⋮	⋮
Modalités : $Z1_1, Z1_2, \dots, Z1_{n_{Z1}}$	Modalités : $Z2_1, Z2_2, \dots, Z2_{n_{Z2}}$	…	Modalités : $Zp_1, Zp_2, \dots, Zp_{n_{Zp}}$	Z
Pour une variable i et un segment S, $Si_j \in M_i$ pour $j \in \{1, 2, \dots, n_{Si}\}$. Où M_i est l'ensemble des modalités de la variable i et $n_{Si} \in \{1, 2, \dots, card(M_i)\}$.				

Tableau 10 : Table de correspondances du modèle CART

Pour une combinaison donnée de modalités de variables explicatives, ce tableau permet de déterminer dans quel segment devrait se trouver l'acheteur possédant cette combinaison. Sachant qu'un segment correspond ici à une feuille de l'arbre CART, la combinaison de modalités de ce segment est celle qui conduit à cette feuille spécifique dans l'arbre de décision.

Par la suite, nous examinerons la stabilité du modèle CART optimal en utilisant la méthodologie décrite dans la section *4.4*.

4.3.2. Algorithme de calibrage des modèles non-supervisés

i. Segmenter l'échantillon d'apprentissage en utilisant l'un des modèles non supervisés (k-prototypes et classification ascendante hiérarchique) ;

ii. Calculer la métrique de qualité (Gini modifié) pour cette segmentation ;

iii. Répéter les étapes i et ii pour différentes valeurs de segments ;

iv. Sélectionner le nombre de segments offrant la meilleure amélioration de la métrique de qualité (méthode du coude) ;

v. Effectuer l'apprentissage de la segmentation optimale choisie à l'étape précédente en utilisant les modèles supervisés suivants :

- Random Forest ;
- XGBoost ;

vi. Optimiser ces modèles de transfert par validation croisée sur l'échantillon d'apprentissage ;

vii. Répéter les étapes i à vi pour le deuxième modèle de segmentation non supervisée ;

viii. Sélectionner le meilleur modèle de transfert et les hyperparamètres optimaux.

À ce stade, nous disposons de deux combinaisons optimales de modèles de segmentation et de transfert, dont les stabilités seront étudiées de manière similaire à celle du modèle CART.

4.4. Méthodologie d'étude de la stabilité des modèles retenus

Tout au long de l'étude, les 20% de la base réservés à la validation resteront inchangés. En suivant les méthodes proposées par Dudoit, et al. (2002), Levine, et al. (2001) et Ben-Hur, et al. (2002)[29], plusieurs sous-échantillons sont extraits de la base d'apprentissage. Chaque modèle de segmentation retenu lors de l'étape de calibrage sera entraîné sur ces sous-échantillons. Les modèles ainsi entraînés seront ensuite utilisés pour segmenter l'ensemble de données de validation. Par conséquent, pour chaque sous-échantillon, l'indice de Gini modifié pourra être calculé sur l'ensemble de données de validation. Les détails mathématiques et statistiques de cet indice de qualité sont présentés dans la section *4.6*.

Pour un modèle donné, cette étape peut être répétée plusieurs fois en variant un ou plusieurs facteur(s) susceptible(s) d'influencer la stabilité dudit modèle. Ces facteurs peuvent inclure les hyperparamètres, les types de prétraitement utilisés, les choix des variables de segmentation et la taille des sous-échantillons.

Concernant la création des sous-échantillons, nous suggérons de maintenir un échantillonnage aléatoire stratifié sans remise dans l'ensemble d'apprentissage. Le nombre de sous-échantillons n est fixé à 30. Selon la littérature, il peut être difficile de déceler les effets des perturbations sur les segmentations avec moins de 30 sous-échantillonnages. Il est important de noter que les perturbations dans l'ensemble d'entraînement sont introduites par les sous-échantillonnages, qui veillent à ce que la structure de la base évolue tout en restant proche de la base de données originale. De plus, pour chaque sous-échantillon d'entraînement, des valeurs de SCR de souscription seront également calculées sur la base B. Le *Tableau 11* récapitule les étapes de cette méthodologie.

Pour chaque scénario (consistant en un modèle de segmentation et une variation d'un facteur potentiellement influent sur sa stabilité), nous obtiendrons n valeurs de la métrique de qualité, calculées à partir de la base de données de validation, ainsi que n valeurs du SCR de souscription, calculées sur la base B. Nous analyserons ensuite la stabilité dans chaque cas par le biais des méthodes suivantes :

- Visualisations graphiques ;
- Analyses des variations par rapport aux scénarios de référence ;

[29] *(Dudoit, et al., 2002), (Levine, et al., 2001) et (Ben-Hur, et al., 2002)*

- Analyses des pentes et des coefficients d'instabilité (cf. section *3.4.1*) ;
- Tests statistiques d'adéquation à des lois uniformes.

<table>
<tr><th colspan="2">Base A</th><th rowspan="2">Base B</th></tr>
<tr><th>Base d'apprentissage (80% de la base A)</th><th>Base de validation (20% de la base A)</th></tr>
<tr><td>SE 1 : Obtenu de la base d'apprentissage par tirage aléatoire stratifié sans remise</td><td rowspan="4">➢ Calcul de n valeurs de la métrique de qualité
➢ Visualisations graphiques des n valeurs de la métrique de qualité
➢ Tests d'adéquation à des lois uniformes
➢ Calcul du coefficient d'instabilité
➢ Etude de la stabilité statistique.</td><td rowspan="4">➢ Calcul de n valeurs du SCR de souscription
➢ Visualisations graphiques des n valeurs du SCR de souscription
➢ Tests d'adéquation à des lois uniformes
➢ Etude de la stabilité business.</td></tr>
<tr><td>SE 2 : Obtenu de la base d'apprentissage par tirage aléatoire stratifié sans remise</td></tr>
<tr><td>…</td></tr>
<tr><td>SE $\boldsymbol{n}$: Obtenu de la base d'apprentissage par tirage aléatoire stratifié sans remise</td></tr>
<tr><td colspan="3">SE = Sous-Echantillon d'entraînement</td></tr>
</table>

Tableau 11 : Division des bases de données et pour la suite des analyses.

4.5. Modèles de segmentation et métriques d'évaluation

Dans cette section, nous allons exposer les bases mathématiques, statistiques et actuarielles des modèles qui sont impliqués dans les divers algorithmes présentés dans les sections précédentes.

4.5.1. Les modèles supervisés

Le terme « supervisé » implique que le modèle est « guidé » par des variables explicatives qui sont associées à une variable réponse. Dans ce cas, les données sont étiquetées par la variable réponse. Un modèle supervisé est donc un type de modèle qui utilise un ensemble de données d'entraînement étiquetées pour apprendre à effectuer des prédictions sur de nouvelles données non étiquetées (Pádraig, et al., 2008). Les modèles supervisés peuvent être utilisés pour la régression (variable réponse continue) ou pour la classification (variable réponse catégorielle). Dans cette section, nous présenterons les modèles supervisés suivants :

- Le modèle d'arbres de classification et de régression (CART) ;
- Le modèle de forêts aléatoires ;
- Le modèle d'Extreme Gradient Boosting (XGBoost).

Par la suite, nous considérerons que les données sont composées de p variables explicatives X^j, qui peuvent être continues ou catégorielles, ainsi qu'une variable

réponse Y, qui peut être continue ou catégorielle avec m modalités $\{\mathcal{T}_l : l = 1, \dots, m\}$. Nous supposerons également que ces variables sont observées sur un échantillon de n individus.

4.5.1.1. Le modèle d'arbres de classification et de régression (CART)

Les modèles CART sont des modèles supervisés qui reposent sur la construction d'arbres de décision. Ces arbres divisent de manière récursive l'espace d'entrée en sous-espaces plus petits et plus homogènes en fonction des variables explicatives. Cette segmentation se fait en utilisant des règles de décision basées sur des seuils pour les différentes caractéristiques des données (Breiman, et al., 1984).

À chaque nœud de l'arbre, le modèle cherche à identifier la meilleure variable explicative ainsi que la valeur seuil optimale pour diviser les données en deux groupes. Cette division est effectuée de manière à maximiser la pureté ou la précision des sous-groupes résultants. L'indice de Gini est généralement employé pour les problèmes de classification, tandis que l'erreur quadratique moyenne (MSE[30]) est utilisée pour les problèmes de régression. En sélectionnant judicieusement les variables et les seuils, les modèles CART visent à accroître la réduction de l'impureté à chaque étape. L'impureté mesure la dispersion des étiquettes de classe (dans le cas de la classification) ou des valeurs cibles (dans le cas de la régression) au sein d'un nœud.

La division du nœud κ génère deux descendants : un à gauche, κ_G, et un à droite, κ_D. Soit D_κ, D_{κ_G} et D_{κ_D} les impuretés de ces nœuds. Pour toute division admissible au nœud κ (c'est-à-dire lorsque κ_G et κ_D ne sont pas vides), l'algorithme sélectionne la variable qui maximise la somme $D_{\kappa_G} + D_{\kappa_D}$. Le problème résolu à chaque nœud κ de l'arbre est donc le suivant :

$$\max_{\{divisions\ de\ X^j : j=1,\dots,p\}} D_\kappa - \left(D_{\kappa_G} + D_{\kappa_D}\right)$$

Si Y est continue, alors D_κ correspond à la variance et on a :

$$D_\kappa = \frac{1}{|\kappa|} \sum_{i \in \kappa} (y_i - \bar{y}_\kappa)^2$$

Où $|\kappa|$ est l'effectif du nœud κ.

Si Y est catégorielle, alors D_κ peut être :

- L'entropie :

$$D_\kappa = -2 \sum_{l=1}^{m} |\kappa| p_\kappa^l \log(p_\kappa^l)$$

[30] *Mean Squared Error*

Où p_κ^l est la proportion de la modalité $\mathcal{T}_l$ de Y dans le nœud κ.

- La concentration de Gini :

$$D_\kappa = \sum_{l=1}^{m} p_\kappa^l (1 - p_\kappa^l)$$

L'arrêt de la croissance de l'arbre peut se produire à un nœud dans deux cas :

- Lorsqu'il n'existe plus de partition admissible (le nœud est homogène) ;
- Lorsque le nombre d'observations qu'il contient est inférieur à un seuil prédéfini. Ce seuil est appelé « minibucket ».

À la fin de la construction de l'arbre (cf. *Figure 12*), chaque nœud terminal devient une feuille à laquelle est assignée la valeur moyenne des observations du nœud si la variable réponse est continue, ou la modalité la plus fréquente dans le nœud si la variable réponse est catégorielle.

Cette méthode de construction génère un arbre très détaillé, parfois avec un nombre excessif de feuilles. Cela peut conduire à un surajustement (*overfitting en anglais*) des données d'entraînement. Pour obtenir un modèle plus parcimonieux, il est nécessaire d'effectuer un élagage (*pruning en anglais*) de l'arbre. Cette opération vise à déterminer un nombre optimal de feuilles, situé entre une seule feuille et le nombre de feuilles de l'arbre avant élagage. Dans notre contexte, le nombre optimal de feuilles, qui correspond également au nombre de segments, sera celui qui produit la segmentation la plus performante pour notre ensemble de données. Autrement dit, il s'agit de choisir celui qui obtient la valeur la plus élevée pour la croissance de la métrique évaluant la qualité de la segmentation.

4.5.1.2. Le modèle de forêts aléatoires

Les forêts aléatoires (*Random Forests en anglais*) représentent une technique d'apprentissage automatique largement utilisée pour la classification et la régression. Ce modèle combine plusieurs arbres de décision afin de former un modèle global plus robuste et plus précis (Breiman, 2001). Les arbres de décision pris individuellement présentent des performances relativement modestes. Grâce à une méthode appelée « Bagging », le modèle de forêts aléatoires parvient à amalgamer les performances de ces arbres de décision pour aboutir à une performance globale améliorée (Liaw, et al., 2002). Le Bagging est une combinaison de techniques de Bootstrap et d'agrégation des résultats.

4.5.1.2.1. Le Bootstrap

L'étape de Bootstrap implique l'échantillonnage aléatoire avec remplacement des données d'entraînement, formant ainsi chaque sous-ensemble utilisé dans la construction des arbres de décision. Chaque arbre est ainsi formé sur un échantillon unique, introduisant ainsi de la diversité dans les arbres et améliorant les performances globales de la forêt aléatoire. De plus, il est possible d'effectuer un échantillonnage aléatoire avec remplacement des variables explicatives à chaque division d'un nœud. Cette deuxième randomisation, connue sous le nom de « feature bagging », contribue également à accroître la diversité des arbres et à améliorer les performances globales du modèle (Liaw, et al., 2002).

4.5.1.2.2. L'agrégation des résultats

Pour réaliser une prédiction sur une nouvelle observation, chaque arbre de la forêt génère sa propre prédiction, et la prédiction finale est obtenue en agrégeant ces prédictions individuelles. En classification, une méthode courante est le vote majoritaire, tandis qu'en régression, la moyenne des prédictions des arbres est employée. Cette approche permet de réduire la variance des prédictions (Cutler, et al., 2012).

La figure ci-dessous illustre le fonctionnement d'un modèle de forêts aléatoires :

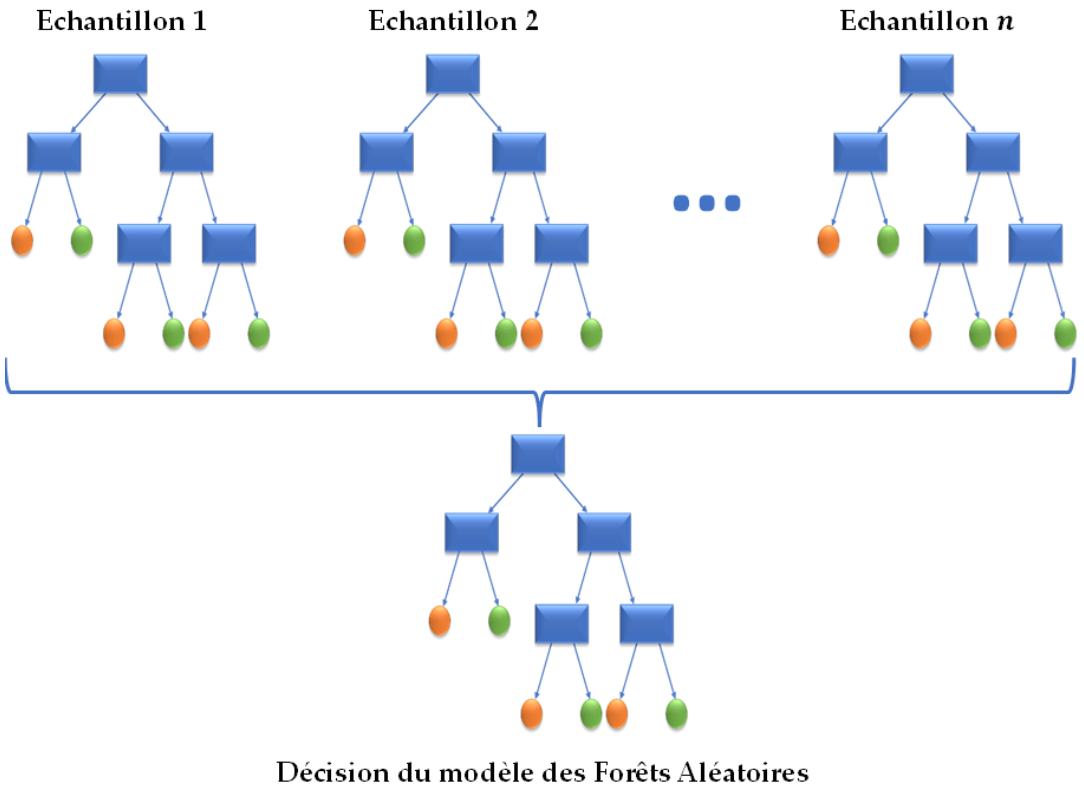


Figure 16 : Illustration du modèle de forêts aléatoires

4.5.1.3. Le modèle d'eXtreme Gradient Boosting (XGBoost)

Les méthodes de gradient boosting consistent en l'assemblage séquentiel de plusieurs modèles « faibles » pour constituer un modèle de prédiction plus puissant (Schapire, 2003). Le modèle XGBoost est une méthode de boosting basée sur les arbres de décision. Lors de sa création, le modèle XGBoost utilise des arbres de décision pour approximer et minimiser une fonction de perte. Chaque arbre de décision vise à corriger les erreurs résiduelles du précédent. Les prédictions de chaque arbre sont agrégées pour former une prédiction globale du modèle XGBoost, minimisant ainsi la fonction de perte globale.

Le XGBoost se caractérise par sa parcimonie (Chen, et al., 2016). Il a recourt à des paramètres de régularisation pour contrôler les poids des arbres, évitant ainsi tout surajustement du modèle. En outre, le XGBoost effectue un élagage des arbres. Comme présenté dans la section *4.5.1.1*, l'élagage contribue à réduire le risque de surajustement.

Un autre avantage du modèle XGBoost est sa rapidité. Ceci est attribuable à l'utilisation de techniques d'optimisation avancées pour accélérer le processus d'apprentissage et améliorer les performances du modèle. Il adopte une approche d'apprentissage par lots (*batch learning en anglais*) qui tire parti des statistiques d'agrégation pour minimiser les coûts de calcul (Nielsen, 2016). La figure suivante illustre le fonctionnement du modèle XGBoost :

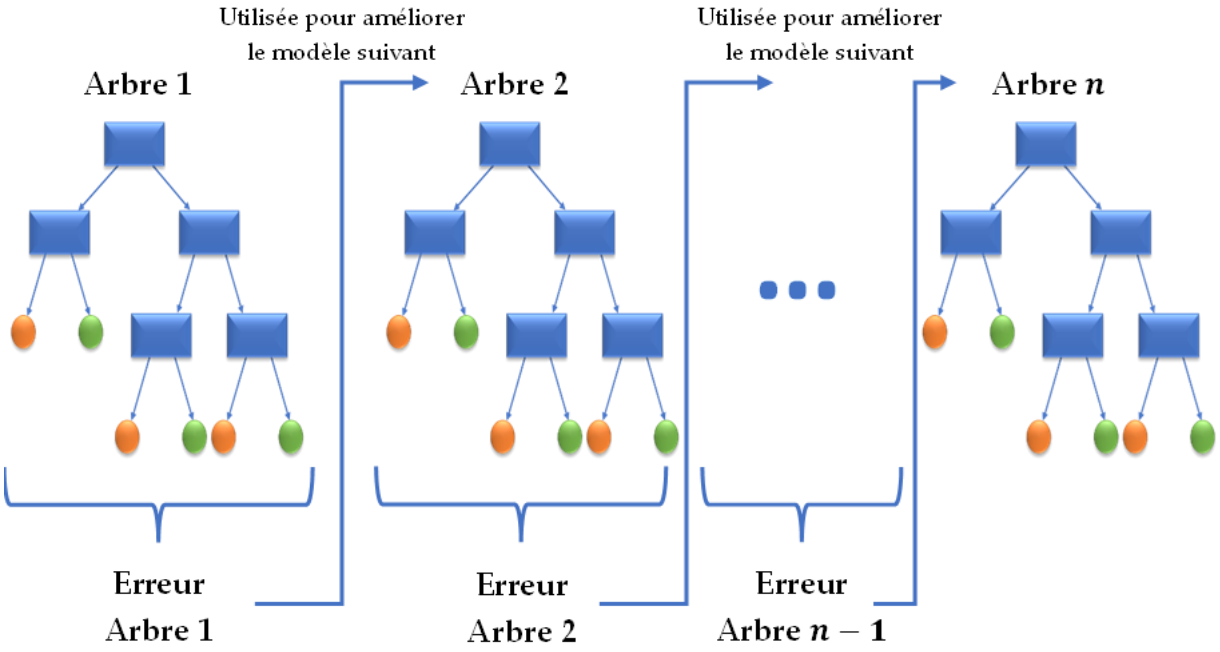


Figure 17 : Illustration du modèle XGBoost

4.5.2. Les modèles non-supervisés

Un modèle non supervisé en apprentissage statistique est un type de modèle utilisé pour explorer et analyser des données sans nécessiter d'étiquettes ou de variable réponse préalable. Contrairement aux modèles supervisés, qui sont entraînés sur des données étiquetées en vue de prédire des valeurs de sortie spécifiques, les modèles non supervisés cherchent à découvrir des structures, des motifs ou des relations intrinsèques dans les données sans préconnaissance des catégories ou des regroupements (Hastie, et al., 2009). La capacité de ces modèles à saisir des structures au sein des données les rend particulièrement adaptés à la segmentation d'un jeu de données en sous-groupes homogènes. Dans le contexte de ce mémoire, nous nous focaliserons particulièrement sur les deux modèles non supervisés suivants :

- Le modèle de k-prototypes
- Le modèle de classification ascendante hiérarchique

Le choix de ces modèles est motivé d'une part par leur simplicité en termes de compréhension et d'interprétation, et d'autre part par leur utilisation fréquente dans des problématiques similaires à la nôtre (cf. section *3.2.2*).

4.5.2.1. Le modèle de k-prototypes

Le modèle des k-means est utilisé pour segmenter des observations en fonction de variables continues uniquement, tandis que la méthode des k-modes permet de réaliser des segmentations basées exclusivement sur des variables catégorielles. En 1998, Zhexue Huang[31] propose une extension de ces deux modèles, capable de construire des segmentations en prenant en compte des variables mixtes (continues et catégorielles). C'est le modèle de k-prototypes. Le modèle de k-prototypes fusionne l'algorithme des k-means pour les variables continues avec une approche fondée sur la correspondance simple pour les variables catégorielles. Comme le modèle de k-means, il itère entre la mise à jour des centroïdes et l'affectation des données aux clusters jusqu'à convergence.

Revenons aux notations des variables mentionnées précédemment et supposons que parmi nos p variables, q sont continues (les variables 1 à q), tandis que les autres sont catégorielles. Considérons deux observations, a et b tirées de notre ensemble de données. Nous avons alors :

$$a = \begin{pmatrix} a_1 \\ \vdots \\ a_q \\ \vdots \\ a_p \end{pmatrix}; \; b = \begin{pmatrix} b_1 \\ \vdots \\ b_q \\ \vdots \\ b_p \end{pmatrix}$$

[31] *(Huang, 1998)*

La distance entre les observations a et b, telle qu'elle est utilisée dans le modèle de k-prototypes, est calculée comme suit :

$$d(a,b) = \sum_{j=1}^{q} \|a_j - b_j\|^2 + \gamma \sum_{j=q+1}^{p} \mathbb{1}(a_j \neq b_j)$$

Où $\mathbb{1}(a_j \neq b_j)$ prend la valeur 1 si a_j est différent de b_j et 0 sinon. La distance utilisée dans le modèle de k-prototypes est donc une somme de la distance euclidienne pour les variables continues et de la distance de Hamming (Dorman, et al., 2022) pour les variables catégorielles. Le paramètre γ représente un facteur de pondération, ajustant le poids relatif de chaque catégorie de variable dans la création de la segmentation.

En comparaison avec les modèles des k-means et k-modes, le modèle de k-prototypes présente un inconvénient en termes de temps de calcul, étant plus coûteux.

4.5.2.2. Le modèle de classification ascendante hiérarchique

La classification ascendante hiérarchique (CAH) fait partie de la famille des algorithmes hiérarchiques (cf. section *3.2.2.1*). Lorsqu'il s'agit de segmenter n observations en k segments, le modèle débute avec n segments, chacun ne contenant qu'une seule observation. Ces segments sont ensuite fusionnés de manière hiérarchique (Jain, et al., 1988). Pour réaliser ces fusions, une mesure de distance entre les segments est utilisée.

La première étape entreprise par le modèle CAH consiste donc à calculer les distances entre chaque observation, afin de construire une matrice de distance comme celle illustrée ci-dessous :

Observations	**1**	**2**	...	$\boldsymbol{j}$	...	$\boldsymbol{n-1}$	$\boldsymbol{n}$
1	0						
2	$d(2,1)$	0					
⋮							
$\boldsymbol{j}$	$d(j,1)$	$d(j,2)$		0			
⋮							
$\boldsymbol{n-1}$	$d(n-1,1)$	$d(n-1,2)$		$d(n-1,j)$		0	
$\boldsymbol{n}$	$d(n,1)$	$d(n,2)$		$d(n,j)$		$d(n,n-1)$	0

Tableau 12 : Table de distance du modèle CAH

Comme on peut le prévoir, le temps de calcul de cette matrice augmente de manière exponentielle avec le nombre d'observations, ce qui constitue le principal inconvénient du modèle CAH (Sonagara, et al., 2014).

Des exemples de mesures de distance entre deux observations incluent la distance euclidienne et la distance de Manhattan (cf. section *3.1.3*). Pour calculer des distances en utilisant des variables mixtes (continues et catégorielles), le modèle CAH adopte des mesures de distance plus appropriées, comme la distance de Gower[32], qui est définie de la manière suivante :

$$d(a,b) = \sqrt{1 - \frac{\sum_{k=1}^{p} d_{abk}\delta_{abk}}{\sum_{k=1}^{p} \delta_{abk}}}$$

Où δ_{abk} prend la valeur 1 si les observations a et b peuvent être comparées au niveau de la variable k et 0 sinon (en raison des valeurs manquantes par exemple). En ce qui concerne d_{abk},

- Si la variable k est continue, $d_{abk} = 1 - \frac{|a_k - b_k|}{R_k}$.

  Où R_k est l'étendue de la variable k

- Si la variable k est catégorielle, $d_{abk} = \mathbb{1}(a_k = b_k)$

Lors de la fusion des segments, le modèle CAH a plusieurs méthodes pour évaluer la distance entre deux segments. Parmi celles-ci, nous avons notamment :

- **La distance minimale (*single linkage en anglais*) :** Elle définit la distance entre deux segments $S1$ et $S2$ comme la distance entre leurs deux éléments les plus proches.

$$D(S1,S2) = \min_{a \in S1, b \in S2} d(a,b)$$

- **La distance maximale (*complete linkage en anglais*) :** Elle définit la distance entre deux segments $S1$ et $S2$ comme la distance entre leurs deux éléments les plus éloignés.

$$D(S1,S2) = \max_{a \in S1, b \in S2} d(a,b)$$

- **La distance moyenne (*average linkage en anglais*) :** Elle définit la distance entre deux segments $S1$ et $S2$ comme la moyenne de toutes les distances entre les éléments de ces segments.

$$D(S1,S2) = \frac{1}{|S1|.|S2|} \sum_{a \in S1} \sum_{b \in S2} d(a,b)$$

Où $|S|$ représente le nombre d'observations dans le segment S.

[32] *(Gower, 1971)*

- **La distance de Ward (*Ward's linkage en anglais*) :** Elle définit la distance entre deux segments S1 et S2 comme l'écart quadratique entre les centroïdes μ_{S1} et μ_{S2} de ces segments.

$$D(S1, S2) = \frac{|S1|.|S2|}{|S1 \cup S2|} \|\mu_{S1} - \mu_{S2}\|^2$$

L'une des étapes de nos analyses consistera à sélectionner les distances les plus appropriées pour notre jeu de données.

4.6. L'indice de Gini pour l'évaluation de la qualité des segmentations

Selon les définitions fournies au *Chapitre 3*, une segmentation efficace est celle qui maximise la similarité entre les membres d'un segment (similarité intra-segments) tout en maximisant la différence entre les membres de segments distincts (dissimilarité inter-segments). Pour évaluer ces similarités/dissimilarités, il est essentiel de définir une mesure appropriée. Dans nos analyses, nous avons opté pour une mesure en lien avec l'activité d'assurance-crédit sur laquelle se concentrent nos travaux. Cette mesure, que nous nommerons « indice de Gini modifié », s'inspire de l'indice de Gini largement utilisé et préconisé par la littérature pour des problématiques similaires aux nôtres (cf. section *3.1.4*).

À l'origine, l'indice de Gini est une mesure statistique permettant de représenter la répartition d'une variable (salaire, revenus, patrimoine, etc.) au sein d'une population (Gastwirth, 1972). En matière de segmentation, cet indice peut être adapté pour évaluer l'homogénéité des individus au sein d'un segment, en fonction d'une ou plusieurs variables. Considérons la figure suivante :

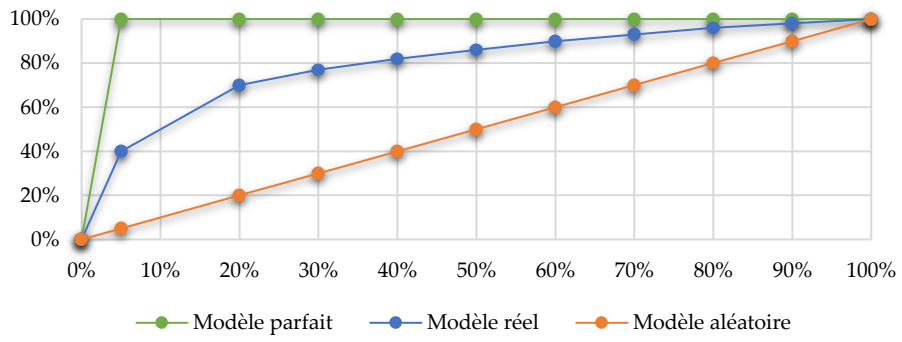


Figure 18 : Calcul de l'indice de Gini

L'indice de Gini est donné par :

$$Gini = \frac{Aire\ réelle}{Aire\ parfaite}$$

L'aire parfaite représente la région située entre la courbe de Lorentz (Gastwirth, 1971) du modèle parfait (si chaque observation constituait un segment) et la courbe de Lorentz du modèle aléatoire. L'aire réelle, quant à elle, correspond à la région entre la courbe de Lorentz du modèle de segmentation évalué et la courbe de Lorentz du modèle aléatoire.

Dans le cadre de notre étude, nous adopterons un indice de Gini modifié pour prendre en compte les particularités de l'activité d'assurance-crédit. Cet indice de Gini modifié est actuellement utilisé par Coface dans le cadre de son modèle interne partiel. Les aires impliquées dans son calcul sont déterminées à l'aide de la méthode des trapèzes (Vergnes, 1980) au moyen des algorithmes suivants :

4.6.1. Aire réelle

i. Calculer les sommes d'EAD et d'Expositions (cf. section *1.1.2*) par segments :

$$EAD_s = \sum_{i \in s} EAD_i \; ; Expo_s = \sum_{i \in s} Expo_i$$

ii. Calculer le rapport R_s entre l'EAD et l'Exposition pour chaque segment :

$$R_s = \frac{EAD_s}{Expo_s}$$

iii. Ranger les segments par ordre décroissant des valeurs de R_s

iv. Calculer les fréquences cumulées d'EAD et d'Expositions suivant l'ordre induit par R_s

$$EAD_s^{FC} = \frac{\sum_{i=1}^{s} EAD_{(i)}}{\sum_s EAD_s}; \; Expo_s^{FC} = \frac{\sum_{i=1}^{s} Expo_{(i)}}{\sum_s Expo_s}$$

Où (i) est le $i^{ème}$ segment d'après l'ordre induit par R_s, $EAD_0^{FC} = 0$ et $Expo_0^{FC} = Expo_1^{FC}$

On a finalement :

$$Aire\ réelle = -1 + \sum_s \left((EAD_s^{FC} + EAD_{s-1}^{FC}) \times (Expo_s^{FC} - Expo_{s-1}^{FC}) \right)$$

4.6.2. Aire parfaite

i. Calculer le rapport R_i entre l'EAD et l'Exposition de chaque observation (agrément ou acheteur) :

$$R_i = \frac{EAD_i}{Expo_i}$$

ii. Ranger les segments par ordre décroissant des valeurs de R_i

iii. Calculer des fréquences cumulées d'EAD et d'Expositions suivant l'ordre induit par R_i

$$EAD_i^{FC} = \frac{\sum_{k=1}^{i} EAD_{(k)}}{\sum_i EAD_i};\ Expo_i^{FC} = \frac{\sum_{k=1}^{i} Expo_{(k)}}{\sum_i Expo_i}$$

Où (k) est la $k^{ième}$ observation d'après l'ordre induit par R_i, $EAD_0^{FC} = 0$ et $Expo_0^{FC} = Expo_1^{FC}$

On a finalement :

$$Aire\ parfaite = -1 + \sum_i \left(\left(EAD_i^{FC} + EAD_{i-1}^{FC} \right) \times \left(Expo_i^{FC} - Expo_{i-1}^{FC} \right) \right)$$

L'indice de Gini modifié ainsi calculé varie entre 0 et 1. Plus cet indice se rapproche de 1, meilleure est la qualité de la segmentation.

4.7. Les métriques d'évaluation des pouvoirs prédictifs des modèles de transfert

Dans la section *4.3.2*, nous avons exposé l'utilité des modèles de transfert dans notre algorithme de segmentation non supervisée. Pour rappel, ces modèles sont destinés à apprendre les segmentations construites par les modèles de segmentation non supervisée afin de les reproduire sur de nouveaux jeux de données. Cette reproduction des segmentations apprises est rendue possible grâce à la capacité prédictive des modèles de transfert (modèles supervisés), une capacité absente chez les modèles non supervisés. Afin de sélectionner un modèle de transfert efficace, il est crucial de pouvoir quantifier et évaluer les performances prédictives des modèles candidats.

Soit y les valeurs observées de la variable réponse catégorielle de notre ensemble de données, et $\hat{y}$ les valeurs prédites de cette variable par le modèle à évaluer. Supposons que notre ensemble de données comprenne n observations et que la variable y ait m modalités. Pour l'évaluation de ce modèle, nous utiliserons les métriques suivantes en validation croisée :

4.7.1. Le taux de bon classement

Cette métrique, également appelée « accuracy », mesure la proportion de prédictions correctes d'un modèle. Elle est calculée comme suit :

$$TBC = \frac{1}{n} \sum_{i=0}^{n} \mathbb{1}(\hat{y}_i = y_i)$$

Où $\mathbb{1}(\hat{y}_i = y_i)$ prend la valeur 1 si la valeur prédite correspond à la valeur réelle pour l'observation i, et 0 sinon. Cette métrique est donc comprise entre 0 et 1.

Un modèle peu performant aura un taux de bon classement proche de 0, tandis qu'un modèle performant aura un taux de bon classement de 1.

4.7.2. La sensibilité

La sensibilité mesure la capacité d'un modèle à prédire correctement toutes les modalités de la variable réponse. La prise en compte de cette métrique est particulièrement recommandée dans les cas où le jeu de données est déséquilibré, c'est-à-dire lorsque les modalités de la variable réponse n'ont pas des fréquences similaires dans la base de données (Burduk, 2020). Dans de tels cas, il est possible d'obtenir des taux de bon classement acceptables, pourtant certaines modalités sont mal prédites. La sensibilité d'un modèle de classification est calculée par :

$$Sensibilité = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_i}\sum_{j\in i}\mathbb{1}(\hat{y}_j = y_j)$$

Où n_i est le nombre d'observations pour lesquelles la variable réponse y à la modalité i, et $j \in i$ signifie que la modalité de la variable réponse de l'observation j est i.

La sensibilité d'un modèle varie entre 0 et 1. Plus elle est élevée, meilleur est le modèle.

4.8. Les tests statistiques

Dans le cadre de nos analyses, nous réaliserons plusieurs tests statistiques. Ces tests seront effectués lors de l'analyse des liens entre les variables de l'étude et lors de l'étude de la stabilité de nos modèles de segmentation. Dans le premier cas, il s'agira du test du chi-deux et du test de Kruskal-Wallis, tandis que dans le second cas, il s'agira des tests d'adéquation à des lois uniformes.

4.8.1. Test de Khi-deux et V de Cramer

Le coefficient de corrélation usuel de Pearson n'est pas approprié pour les variables catégorielles. Cependant, dans notre étude, toutes les variables explicatives sont catégorielles. Pour analyser les liens entre ces variables, le test d'indépendance du chi-deux de Pearson est plus approprié (Tallarida, et al., 1987).

Soient deux variables catégorielles X et Y, ayant toutes les deux n observations et respectivement r et s modalités. Notons $n_{i,j}$ le nombre d'observations de ayant à la fois la modalité i de la variable X et j de la variable Y, $n_{i,.}$ Le nombre d'observations ayant la modalité i de la variable X et $n_{.,j}$ ne nombre d'observations ayant la modalité j de la variable Y. Pour tester l'hypothèse nulle H_0 : X indépendante de Y contre l'hypothèse

alternative $H_1 : X$ liée à Y, on utilise la statistique du test de Khi-deux, définie comme suit :

$$T = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{i,j} - \frac{n_{i,.}n_{.,j}}{n}\right)^2}{\frac{n_{i,.}n_{.,j}}{n}}$$

Cette statistique est distribuée suivant une loi du Khi-deux à $(r-1)(s-1)$ degrés de liberté et l'hypothèse H_0 est rejeté au seuil α si la p-valeur $= \mathbb{P}\left(\chi^2_{(r-1)(s-1)} > T\right)$ est inférieure à α.

Comme nous le remarquons, ce test permet uniquement de déterminer si deux variables catégorielles sont indépendantes ou non. Il ne fournit pas directement d'information sur l'intensité de la relation. Pour quantifier cette intensité, nous utiliserons le V de Cramer (Sun, et al., 2010), défini comme suit :

$$V = \sqrt{\frac{T}{n \times \min(s-1, r-1)}}$$

Le V de Cramer est compris entre 0 et 1. D'après la littérature, si $0 < V < 0{,}2$, alors la liaison entre les deux variables est faible. Si $0{,}2 < V < 0{,}6$, alors la liaison entre les deux variables est modérée, et si $V > 0{,}6$, alors la liaison entre les deux variables est forte.

4.8.2. Test de Kruskal-Wallis et Eta-deux

Nous serons également amenés à étudier les liens entre nos variables explicatives et les variables réponses. Ces variables réponses correspondent aux phénomènes intervenant lors de la sinistralité (PD, UGD, CS et LGD), peuvent être continues. Cependant, nous n'avons a priori aucune raison de penser qu'elles suivent une distribution normale. Nous utiliserons donc le test de Kruskal-Wallis. Ce test non-paramétrique est le plus adapté pour analyser la liaison entre une variable catégorielle et une variable continue si cette dernière ne suit pas une distribution normale (McKight, et al., 2010).

Soit X une variable catégorielle à g modalités et Y une variable continue. Soit n le nombre total d'observations, n_i le nombre d'observations de la modalité i, $r_{i,j}$ le rang (parmi toutes les observations) de l'observation j de la modalité i. Pour tester l'hypothèse nulle $H_0 : X$ indépendante de Y contre l'hypothèse alternative $H_1 : X$ liée à Y, on utilise la statistique du test de Kruskal-Wallis définie comme suit :

$$H = (n-1)\frac{\sum_{i=1}^{g} n_i \left(\frac{\sum_{j=1}^{n_i} r_{i,j}}{n_i} - \frac{n+1}{2}\right)^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(r_{i,j} - \frac{n+1}{2}\right)^2}$$

Cette statistique suit une loi du Khi-deux à $(g-1)$ degrés de liberté et l'hypothèse H_0 est rejeté au seuil α si la p-valeur $= \mathbb{P}\left(\chi^2_{(g-1)} > H\right)$ est inférieure à α.

Tout comme dans le cas du test du Khi-deux, il est nécessaire de définir une statistique permettant de quantifier l'intensité de la liaison. C'est le rôle que joue l'eta-carré (η^2) (Richardson, 2011). Il est défini comme suit :

$$\eta^2 = \frac{H - g + 1}{n - g}$$

L'eta-carré est compris entre 0 et 1. D'après la littérature, si $0{,}01 < \eta^2 < 0{,}06$, alors la liaison entre les deux variables est faible. Si $0{,}06 < \eta^2 < 0{,}14$, alors la liaison entre les deux variables est modérée, et si $\eta^2 > 0{,}14$, alors la liaison entre les deux variables est forte.

4.8.3. Test d'adéquation à une loi uniforme

L'une des étapes de l'analyse de la stabilité de nos modèles de segmentation consistera à évaluer la distribution des valeurs de la mesure de qualité des segmentations issues des différentes perturbations des jeux de données d'entraînement (cf. section *4.4*). Le procédé consistera, pour un modèle de segmentation donné, à tester l'adéquation de l'échantillon des valeurs de la mesure de qualité des segmentations à une loi uniforme. Le rejet statistique de cette adéquation pourra permettre de valider la stabilité du modèle considéré. En effet, pour un modèle instable, l'indice de Gini modifié (qui sert à mesurer la qualité d'une segmentation dans notre cas) prendra des valeurs distinctes et parfois éloignées pour chaque sous-échantillon d'entraînement. Ces valeurs seront donc réparties dans un large intervalle, donnant ainsi lieu à un distribution similaire à une loi uniforme sur cet intervalle. En revanche, pour un modèle stable, les valeurs de l'indice de Gini modifié seront très proches, voire identiques, et leur distribution sera très différente de celle d'une loi uniforme. La *Figure 19* illustre ce raisonnement.

Pour tester l'adéquation d'un échantillon à une loi uniforme continue, nous utiliserons le test de Kolmogorov-Smirnov (Berger, et al., 2014). Soit un échantillon $(x_1, \dots, x_n)$ de loi inconnue P. L'hypothèse nulle de ce test est H_0 : la loi P à pour fonction de répartition F_0 (qui est une loi uniforme continue dans notre cas). Si cette hypothèse est vraie, alors la fonction de répartition empirique $\hat{F}$ de l'échantillon doit être proche de F_0. Pour rappel, la fonction de répartition empirique définie sur $\mathbb{R}$ et à valeurs dans [0,1] vaut :

$$\hat{F}(x) = \begin{cases} 0, & pour\ x < x_{(1)} \\ \dfrac{i}{n}, & pour\ x_{(i)} \leq x \leq x_{(i+1)} \\ 1, & pour\ x \geq x_{(n)} \end{cases}$$

Où les $X_{(i)}$ sont les statistiques d'ordre de l'échantillon (valeurs de l'échantillon rangées par ordre croissant). La statistique du test, également appelée distance de Kolmogorov-Smirnov, vaut :

$$D_{KS}(F_0, \hat{F}) = \max_{i=1,\dots,n} \left\{ \left| F_0(x_{(i)}) - \frac{i}{n} \right|, \left| F_0(x_{(i)}) - \frac{i-1}{n} \right| \right\}$$

Sous H_0, la fonction de répartition de cette statistique est indépendante de F_0. Elle n'a pas d'expression explicite. Ses quantiles et la p-valeur du test sont calculés numériquement.

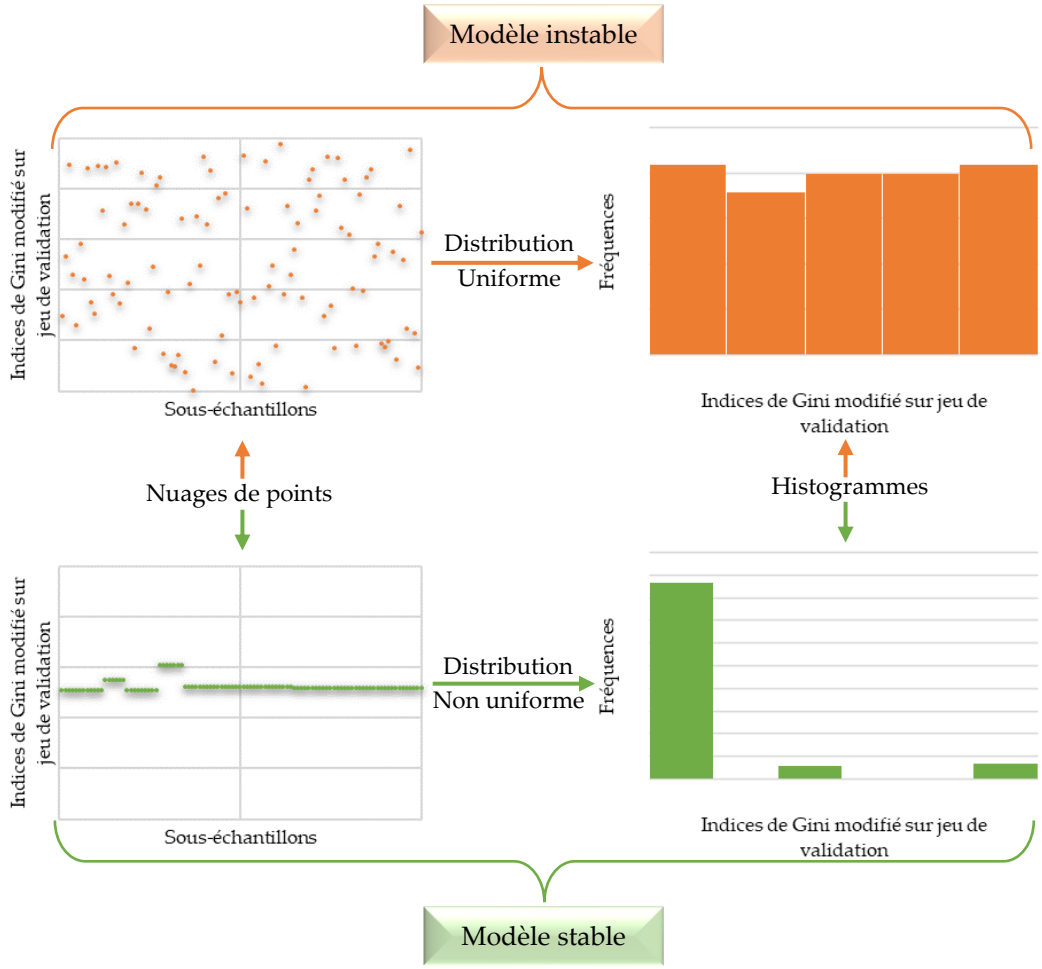


Figure 19 : Distributions de l'indice de Gini modifié suivant la stabilité d'un modèle

Précisons qu'au cours de notre travail, le test d'adéquation à une loi uniforme, ainsi présenté, pourra servir à d'autres analyses que celle de la stabilité.

Ce chapitre marque la fin de la présentation des bases théoriques de notre étude. Dans la partie suivante, nous appliquerons la méthodologie présentée ci-dessus aux données réelles de la Coface. Cette méthodologie nous permettra d'étudier la stabilité de leur modèle de segmentation dans le contexte de l'assurance-crédit tel que présenté dans la première partie de ce mémoire.

Partie 3 : Résultats de l'étude

Chapitre 5 : Données et statistiques descriptives

Sommaire

Dans ce chapitre, nous allons présenter les données sur lesquelles nous appliquerons la méthodologie adoptée dans le chapitre précédent. Tout d'abord, nous exposerons la source de ces données ainsi que les prétraitements réalisés sur celles-ci. Ensuite, nous effectuerons une analyse descriptive minutieuse en examinant chaque variable afin de mettre en évidence les structures de dépendance présentes dans l'ensemble de données. À partir de cette analyse descriptive, nous sélectionnerons les variables les plus pertinentes pour la suite de nos travaux.

Nous avons vu au *Chapitre 2* que les quatre phénomènes intervenant dans la sinistralité en assurance-crédit sont successivement la probabilité de défaut (PD), le taux d'utilisation de la garantie (UGD), les spécificités contractuelles (CS) et le taux de perte (LGD). Notre étude vise à examiner la stabilité des modèles de segmentation des données en fonction de ces quatre phénomènes. Néanmoins, les méthodologies de segmentation se ressemblent grandement, voire sont identiques, pour chacun de ces phénomènes.

Compte tenu de cette similitude et dans le but de faciliter l'interprétation et la compréhension de nos travaux et résultats, nous concentrerons nos analyses sur la probabilité de défaut (PD). Ce choix est motivé par le fait que le défaut est le premier phénomène qui intervient lors de la survenue des sinistres en assurance-crédit (cf. *Figure 7*). De plus, la méthodologie employée pour évaluer la stabilité des modèles de segmentation liés à la PD pourra facilement être reproduite pour les trois autres phénomènes, sans requérir d'adaptation ou d'ajustement spécifique.

5.1. Extraction et traitements des données

Pour notre étude, la Coface nous a fourni les données de son activité d'assurance-crédit couvrant la période de 2007 à 2022. La période historique (désignée comme base A dans la section *4.4*) s'étend de 2007 à 2021, tandis que le SCR de souscription sera calculé à partir des données du quatrième trimestre de l'année 2022 (base B).

Il convient de noter que pour la probabilité de défaut, la segmentation est effectuée en fonction des acheteurs. Autrement dit, chaque observation (ligne) dans ces bases de données correspond à un acheteur (client) d'un assuré de la Coface. La base A, utilisée pour calibrer les modèles, comporte un total de 24 388 428 observations. Les variables de base incluses dans cet ensemble de données sont les suivantes :

- **L'année :** Indique l'année au cours de laquelle un acheteur était associé à un assuré de la Coface par une transaction commerciale.
- **Pays de l'acheteur :** Indique le pays où un acheteur exerce son activité.
- **Cible économique :** Permet de déterminer si l'activité d'un acheteur est orientée vers le marché intérieur ou l'exportation.
- **Région de l'entité :** Indique la région de l'entité de la Coface à laquelle un acheteur donné est rattaché. En cas d'acheteurs couverts par plusieurs agréments, l'entité retenue est celle du contrat avec l'exposition la plus élevée.
- **Secteur d'activité :** Indique le type ou le secteur d'activité économique d'un acheteur.
- **Rating :** Fournit un score de solvabilité établi à partir d'un modèle de notation. Ce dernier est calibré en utilisant plusieurs variables spécifiques aux acheteurs, telles que la taille, le chiffre d'affaires, le capital, le taux de rotation, etc (cf. section *1.1.1.1*). **Le score de solvabilité varie de 0 à 10 et augmente avec la probabilité de solvabilité d'un acheteur.**
- **Exposition :** Donne l'exposition d'un acheteur, calculée conformément à la formule présentée dans la section *1.1.2* de ce mémoire.
- **Exposition en défaut :** Indique l'exposition d'un acheteur au moment de son défaut.
- **Probabilité de défaut :** Indique la probabilité de défaut d'un acheteur, calculée selon la formule exposée dans la section *2.4.1* de ce mémoire.

Ces variables sont extraites de l'entrepôt de données de la Coface à l'aide de multiples requêtes SQL[33]. En se basant sur les analyses statistiques et les avis d'experts de la Coface, des prétraitements ont été réalisés sur les variables de base dans le but d'obtenir l'ensemble de données que nous utiliserons pour le calibrage de nos modèles de segmentation. Le *Tableau 13* présente ces prétraitements ainsi que les raisons qui ont conduit à leur mise en œuvre.

Rappelons que la Coface calcule la probabilité de défaut d'un acheteur en tant que rapport entre son exposition en défaut et son exposition acquise (cf. section *1.1.2*). Par conséquent, il est possible d'obtenir des probabilités de défaut supérieures à 1. Cette variable tend ainsi à quantifier davantage la gravité des défauts que la fréquence

[33] *Structured Query Language*

des défauts. Pour mesurer cette dernière, nous introduisons une nouvelle variable qui prend la valeur 0 si un acheteur n'a pas connu de défaut, et 1 si cet acheteur a connu au moins 1 défaut.

N°	Prétraitements	Justifications
1	Les douze zones géographiques suivantes ont été définies en se basant sur les pays des acheteurs : ➢ Afrique et Moyen Orient ➢ Amérique du nord ➢ Amérique Latine ➢ Asie Pacifique ➢ Europe de l'Est ➢ Europe Centrale ➢ France ➢ Allemagne et Autriche ➢ Grande Chine ➢ Italie ➢ Espagne ➢ Royaume Uni	Ces regroupements de pays ont pour objectif de différencier les pays avec les expositions les plus significatives, tout en rassemblant les autres de manière économiquement justifiée. La réduction du nombre de catégories facilite également la suite des analyses et renforce la solidité des modèles.
2	L'exposition a été regroupée en 8 tranches : ➢ Tranche 1 : *[MIN ; 9K[* ➢ Tranche 2 : *[9K ; 20K[* ➢ Tranche 3 : *[20K ; 50K[* ➢ Tranche 4 : *[50K ; 500K[* ➢ Tranche 5 : *[500K ; 1M[* ➢ Tranche 6 : *[1M ; 10M[* ➢ Tranche 7 : *[10M ; 50M[* ➢ Tranche 8 : *[50M ; MAX[*	La construction de ces tranches permet d'isoler les grandes expositions et de prendre en considération la structure du portefeuille de la Coface. De plus, regrouper les expositions en tranches facilite le transfert des segmentations élaborées grâce au modèle CART. En effet, comme indiqué à la section *4.3.1*, ce transfert nécessite l'élaboration de tables de correspondance entre les variables explicatives et les segments. Cependant, ces tables sont complexes à manipuler lorsque la base de calibrage contient des variables continues.
3	Le rating, variant initialement entre 0 et 10, a été regroupé en 5 tranches : ➢ Tranche 1 : *[0 ; 3]* ➢ Tranche 2 : *4* ➢ Tranche 3 : *5* ➢ Tranche 4 : *[6 ; 7]* ➢ Tranche 5 : *[8 ; 10]*	Les acheteurs regroupés selon leurs scores de solvabilité (ratings) présentent des comportements similaires. Cette tendance a été constatée dans le passé et a été confirmée par des études statistiques réalisées par la Coface.

Tableau 13 : Prétraitements effectuées sur nos variables

Les variables retenues dans la base de données de calibrage après ces prétraitements sont les suivantes :

Variables	Natures	Modalités	Proportions de valeurs manquantes
Année	Catégorielle	*«2001» à «2021»*	0,000%
Probabilité de défaut	Continue		0,004%
Occurrence des défauts	Catégorielle	*«0» et «1»*	0,004%
Tranche d'exposition	Catégorielle	*«Tranche 1», «Tranche 2», «Tranche 3», «Tranche 4», «Tranche 5», «Tranche 6», «Tranche 7» et «Tranche 8»*	0,000%
Exposition en défaut	Continue		0,000%
Région de l'entité Coface	Catégorielle	*«Asie Pacifique», «Europe Centrale et de l'Est», «Allemagne et d'Autriche», «Amérique Latine», «Afrique et du Moyen Orient», «Amérique du Nord», «Europe du Nord» et «Europe de l'Ouest»*	0,077%
Zone géographique de l'acheteur	Catégorielle	*«Afrique et Moyen Orient», «Amérique du nord», «Amérique Latine», «Asie Pacifique», «Europe de l'Est», «Europe Centrale», «France», «Allemagne et Autriche», «Grande Chine», «Italie», «Espagne» et «Royaume Uni»*	0,000%
Cible économique	Catégorielle	*«Intérieur» et «Export»*	0,000%
Secteur d'activité	Catégorielle	*«Industrie métallurgique», «Agroalimentaire», «Industrie chimique», «Construction», «Activités Commerciales», «Electronique et IT», «Services», «Textile, Bois et Papier» et «Autres»*	1,567%
Tranche rating	Catégorielle	*«Tranche 1», «Tranche 2», «Tranche 3», «Tranche 4 et Tranche 5»*	0,000%

Tableau 14 : Variables de l'étude

Nous avons attribué la catégorie « *Autres* » aux acheteurs dont le secteur d'activité n'était pas spécifié. Par la suite, nous avons pris la décision de retirer les acheteurs présentant des valeurs manquantes pour au moins une variable. Cette décision est motivée par le fait que les proportions de valeurs manquantes pour les autres variables sont très faibles, c'est-à-dire inférieures à 1%.

La variable « *Exposition en défaut* » ne sera pas prise en compte lors du processus de calibrage des modèles de segmentation. Ceci s'explique par le fait qu'au moment du calcul du SCR de souscription de l'année $n+1$ au cours de l'année n, l'exposition en défaut de l'année $n+1$ n'est pas encore connue.

5.2. Présentation des variables de l'études

Dans cette section, nous allons effectuer une analyse descriptive univariée des différentes variables de notre étude. Nous nous intéresserons dans un premier temps à la variable d'intérêt, puis aux variables explicatives.

5.2.1. La variable d'intérêt

La probabilité de défaut d'un acheteur est calculée selon la formule présentée à la section *2.4.1*. Sur l'ensemble des acheteurs étudiés, cette variable varie de 0 à 275,692, avec une moyenne de 0,022. Comme mentionné précédemment, cette variable a tendance à représenter la sévérité des défauts plutôt que leur fréquence. Les valeurs très élevées de cette variable concernent les acheteurs dont l'exposition au moment du défaut était supérieure à l'exposition acquise au cours de l'année (cf. section 1.1.2). L'évolution de la probabilité de défaut moyenne sur la période de l'étude, présentée dans la figure ci-dessous, révèle une décroissance constante de celle-ci.

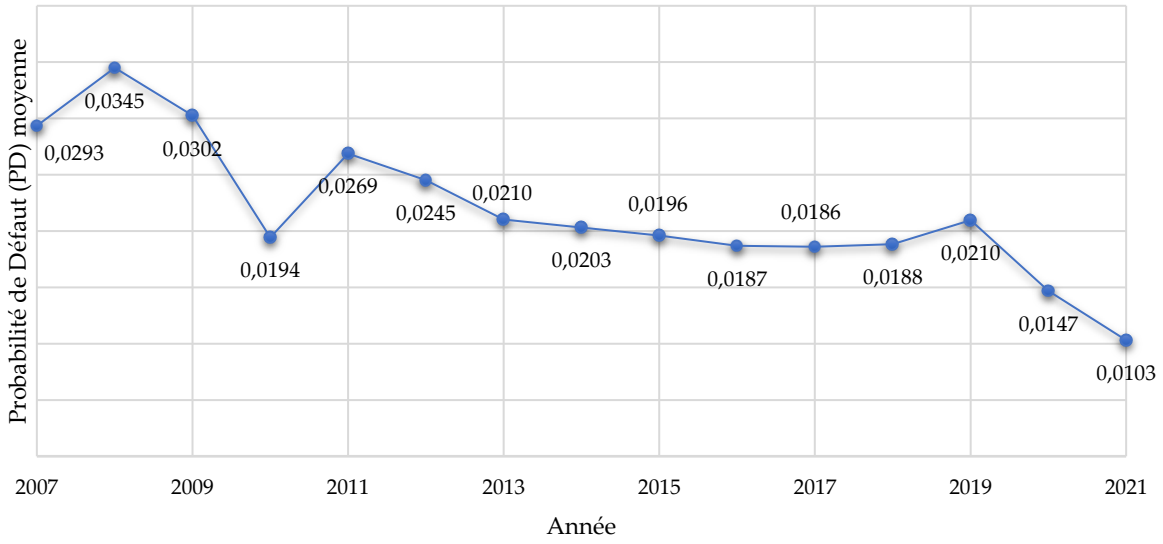


Figure 20 : Evolution de la probabilité de défaut moyenne

Cette décroissance peut s'expliquer par l'amélioration des politiques commerciales de la Coface et les progrès dans les méthodes d'identification et de suivi des acheteurs potentiellement insolvables. Cette explication est renforcée par le fait que des tendances similaires sont observées lors d'une analyse par secteur d'activité et

par zone géographique des acheteurs (cf. *Figure 49*). Il ressort de cette analyse qu'il existe un lien fort entre l'année et la probabilité de défaut. Par conséquent, il est important de conserver la répartition des acheteurs en fonction de l'année de rattachement lors des tirages aléatoires afin d'éviter d'introduire des perturbations dans les données. La variable « *Année* » fera donc partie des variables de stratification dans tous les tirages aléatoires stratifiés que nous réaliserons au cours de cette étude. En ce qui concerne la répartition spatiale des défauts, la figure ci-dessous montre que la probabilité de défaut moyenne varie considérablement d'une zone géographique à une autre.

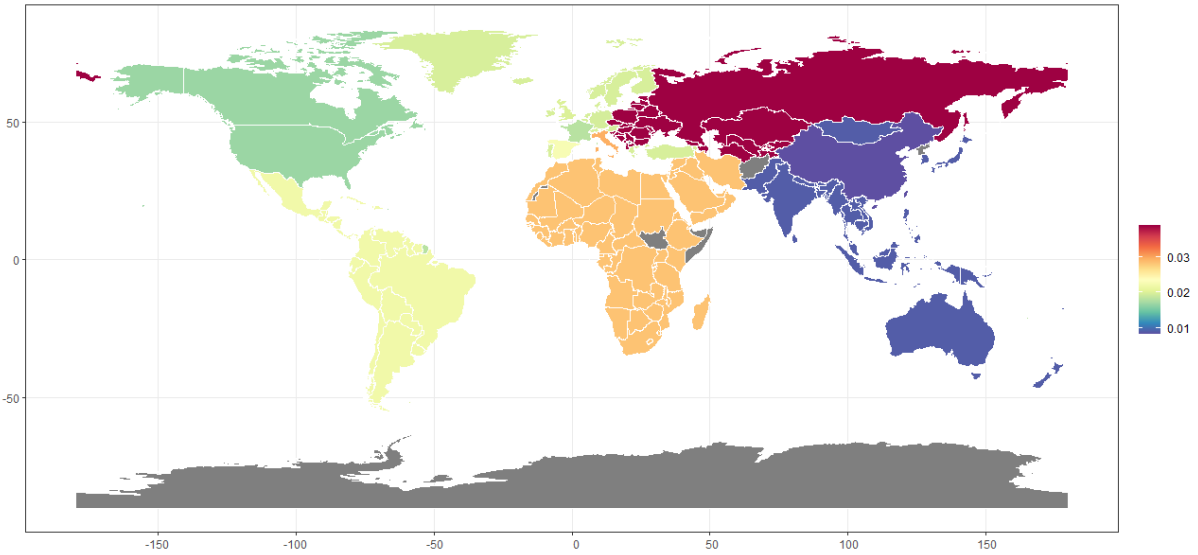


Figure 21 : Répartition spatiale de la probabilité de défaut moyenne

En examinant l'ensemble de la période de l'étude, on constate que la probabilité moyenne de défaut est la plus élevée en Europe de l'Est et la plus faible en Grande Chine. Comme le montre la *Figure 49* en annexe, ce classement des zones géographiques des acheteurs en termes de probabilités de défaut moyenne varie très peu entre 2007 et 2021. Sur cette période, la probabilité moyenne de défaut est plus élevée en Europe de l'Est chaque année, à l'exception de 2012 où l'Afrique et le Moyen-Orient prennent la tête. La Grande Chine et l'Asie-Pacifique occupent le bas du classement tout au long de la période.

En ce qui concerne la fréquence de défaut, 1,398% des acheteurs considérés dans notre étude ont connu au moins un défaut sur la période de 2007 à 2021, soit 340 950 sur un total de 24 388 428 acheteurs. La *Figure 50* en annexe révèle également une diminution de la fréquence des défauts au fil des années, dans toutes les zones géographiques des acheteurs de la Coface. Cette tendance pourrait être due aux mêmes raisons évoquées précédemment concernant la probabilité de défaut moyenne.

La figure suivante illustre la répartition spatiale des fréquences de défauts sur la période allant de 2007 à 2021.

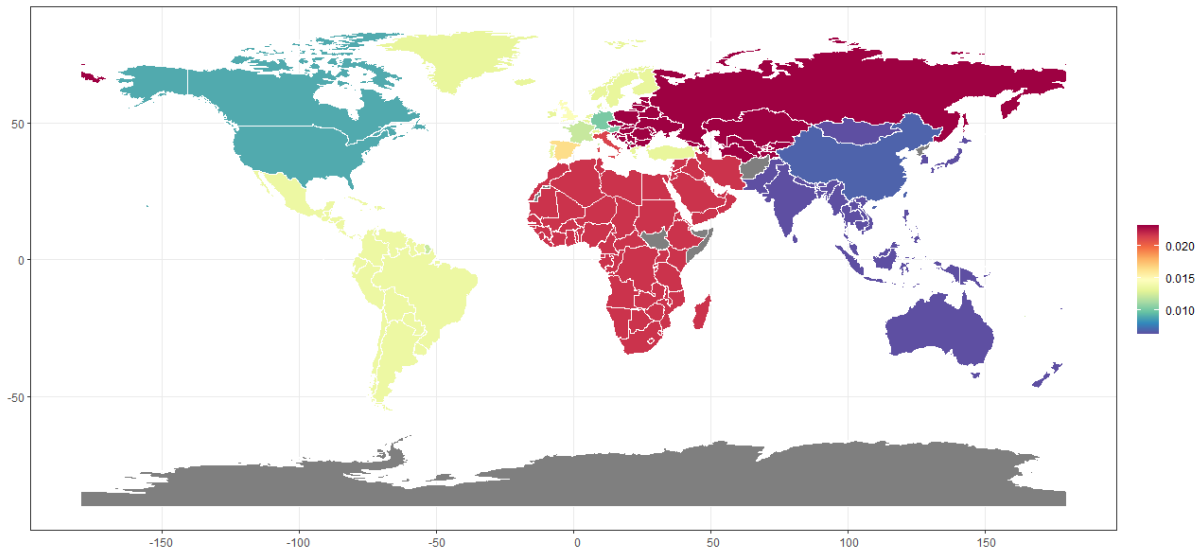


Figure 22 : Répartition spatiale de la proportion de défaut

Selon cette figure, les fréquences des défauts semblent suivre une répartition spatiale similaire à celle de la probabilité de défaut moyenne. Toutefois, l'Afrique et le Moyen-Orient passent au rouge, ce qui signifie que dans cette région, les défauts sont fréquents mais de faible sévérité.

L'exposition moyenne de la Coface au cours de la période 2007 à 2021 semble suivre une tendance inverse à celle de la probabilité de défaut moyenne. Pour mieux comprendre cela, examinons la figure ci-dessous qui présente la répartition spatiale de l'exposition moyenne sur la période allant de 2007 à 2021.

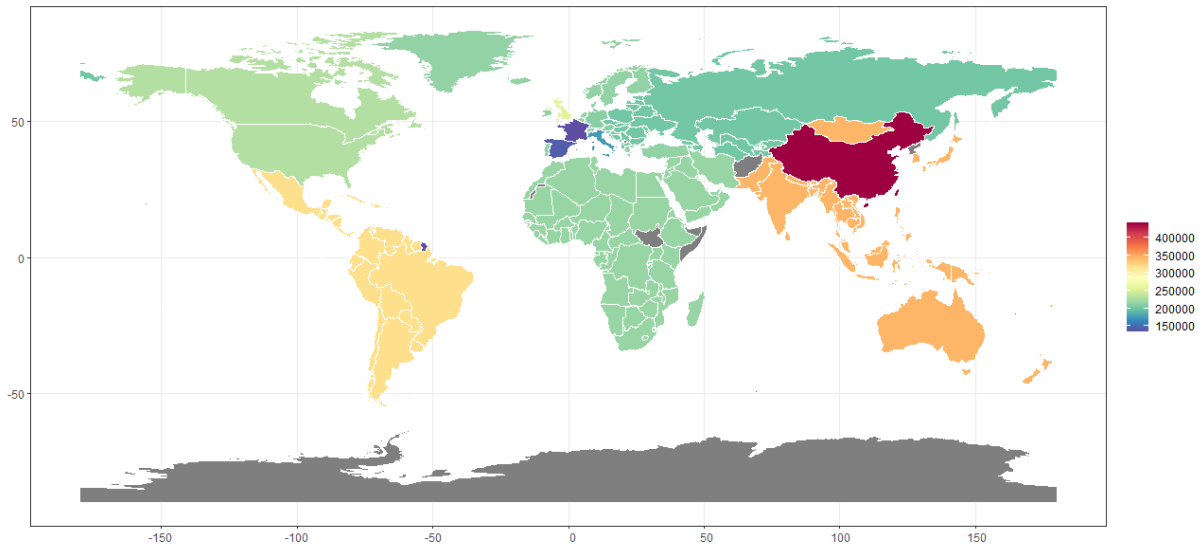


Figure 23 : Répartition spatiale de l'exposition moyenne

Nous remarquons que les zones géographiques des acheteurs où les probabilités de défaut moyennes étaient les plus élevées sont aussi les zones avec les expositions les plus faibles. Ceci pourrait être attribuable à des motivations purement commerciales. En effet, il semblerait que la Coface choisisse de limiter ses expositions dans les régions à risque et d'augmenter celles dans les régions moins risquées. Cette même tendance est observée sur l'ensemble de la période de l'étude, comme illustré sur la figure ci-dessous.

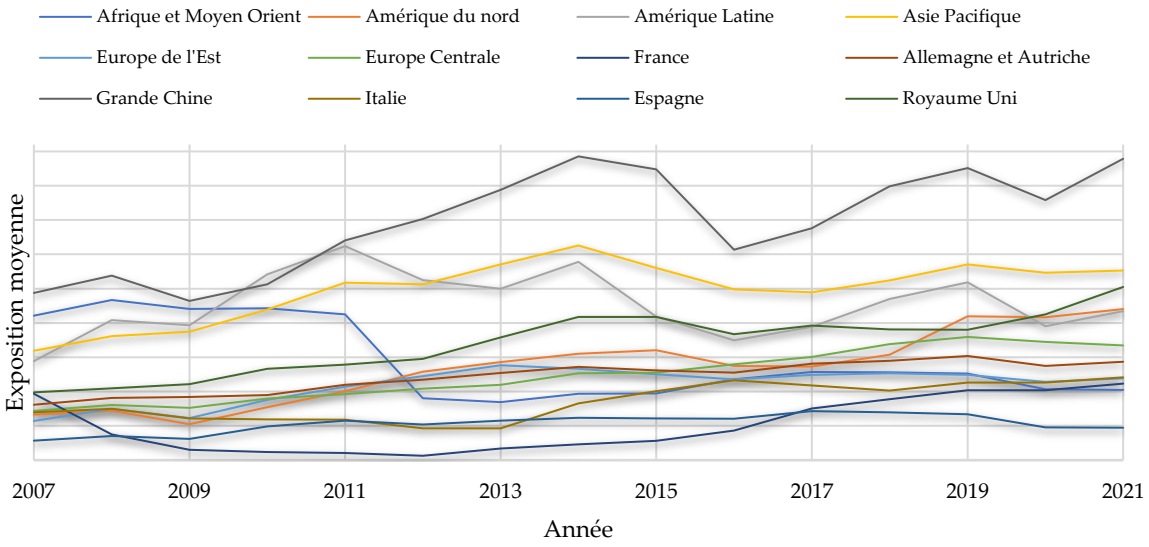


Figure 24 : Evolution de l'exposition moyenne par zone géographique des acheteurs

Cette figure révèle également une croissance presque constante de l'exposition moyenne sur la période considérée. En combinant cette information avec la décroissance de la probabilité de défaut moyenne observée précédemment, nous soupçonnons immédiatement une forte réduction de l'exposition en défaut (EAD). Ce constat est confirmé par la *Figure 51* en annexe. En effet, la baisse de l'EAD moyen présentée sur cette figure renforce l'explication que nous avons avancée précédemment pour justifier la décroissance de la probabilité de défaut moyenne. Au fil des années, l'exposition de la Coface n'a cessé de croître, tandis que l'exposition en défaut a suivi une tendance inverse. Ainsi, nous sommes confortés dans l'idée que la diminution de la probabilité de défaut (qui est le rapport entre l'EAD et l'exposition acquise) est liée à une meilleure anticipation et à un meilleur suivi de la solvabilité des acheteurs par la Coface au fil des années. En effet, l'évolution et l'harmonisation des normes comptables et des méthodes statistiques permettent aux entreprises d'assurance-crédit comme la Coface d'obtenir des informations plus fiables sur les potentiels acheteurs et de calculer avec une précision accrue leur probabilité de solvabilité (rating).

La dépendance claire et forte entre la probabilité de défaut et la zone géographique d'un acheteur justifie l'utilisation de cette dernière en tant que variable

de stratification, au même titre que la variable « *Année* ». Cette relation est confirmée par des tests statistiques présentés à la section *5.4*. Terminons cette section en présentant l'évolution de l'exposition moyenne de la Coface par secteur d'activité des acheteurs.

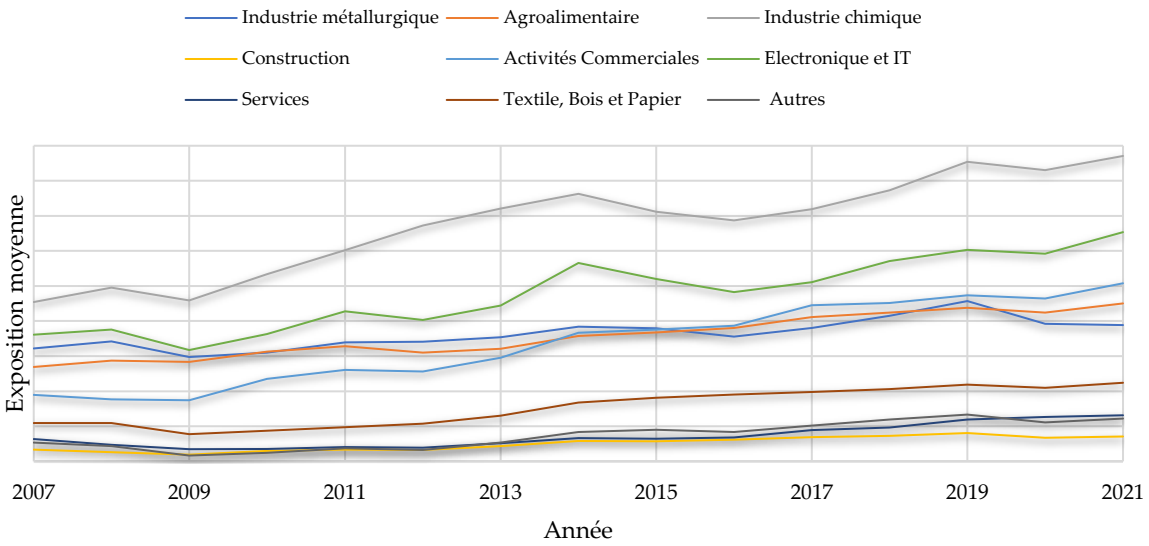


Figure 25 : Evolution de l'exposition moyenne par secteur d'activité

La figure ci-dessus présente l'exposition moyenne de la Coface par secteur d'activité et par année. Outre la croissance de l'exposition moyenne sur la période de 2007 à 2021, nous observons l'existence de frontières assez nettes entre les expositions moyennes des acheteurs opérant dans différents secteurs d'activité, et ce, tout au long de la période étudiée. Les expositions moyennes dans les industries chimiques sont les plus élevées, suivies des entreprises de l'électronique et des technologies de l'information. En revanche, les expositions moyennes dans le secteur de la construction sont les plus faibles.

5.2.2. Les variables explicatives

Il s'agit de toutes les variables présentées à la section *5.1*, à l'exception de l'EAD, de l'occurrence des défauts et de la probabilité de défaut. Dans cette section, nous présenterons les répartitions de ces variables dans l'échantillon d'acheteurs étudié. Dans certains cas, nous porterons également notre attention sur l'évolution de ces répartitions au fil des années.

L'exposition pour la majorité des acheteurs se situe entre 0 et 500 mille euros. En effet, la *Figure 26* montre que l'exposition pour environ 96% des acheteurs se trouve dans les quatre premières tranches. La tranche d'exposition la plus fréquente est la tranche 4 (50 à 500 mille euros) avec 27% des expositions, tandis que la tranche 8 (50

millions d'euros ou plus) est la moins fréquente, représentant moins de 1% des expositions. La *Figure 52* en annexe montre que cette répartition reste pratiquement stable entre 2007 et 2021. Comme évoqué à la section *5.1*, le regroupement de l'exposition en tranches permet effectivement d'isoler les expositions importantes, éliminant ainsi les biais que ces valeurs atypiques pourraient introduire dans les modèles de segmentation.

En ce qui concerne le secteur d'activité des acheteurs, la répartition est assez hétérogène. La majorité des acheteurs (22,5%) opèrent dans le secteur de la construction, suivis par 17,2% travaillant dans l'industrie métallurgique. Le secteur d'activité le moins fréquent est l'électronique, qui représente 5,6% des acheteurs. Nous remarquons qu'entre 2007 et 2021, les proportions des acheteurs travaillant dans les secteurs des services et de l'agroalimentaire n'ont cessé de croître (cf. *Figure 53*). Ceci pourrait s'expliquer par une croissance de la rentabilité et de la stabilité économique de ces secteurs au fil des années. Cette explication est d'autant plus renforcée par le fait que les expositions dans ces secteurs ont également augmenté entre 2007 et 2021 (cf. *Figure 25*).

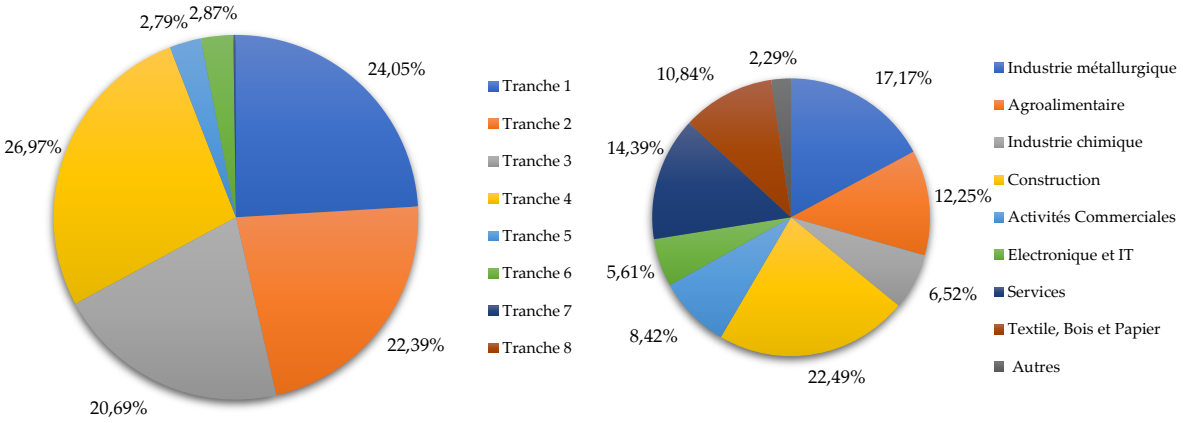


Figure 26 : Répartition des acheteurs par tranche d'exposition (à gauche) et par secteur d'activité (à droite)

La plus grande proportion (18,6%) des acheteurs de notre base exerce ses activités en Allemagne et en Autriche, suivie de la France (17,2%) et de l'Europe centrale (13,4%). La zone géographique la moins représentée dans notre base est la Grande Chine, regroupant seulement 1,1% des acheteurs. Cette répartition semble varier considérablement dans le temps, avec une augmentation des acheteurs en provenance d'Amérique du Nord et une diminution des acheteurs en France et au Royaume-Uni. Ces résultats justifient également le regroupement des pays en zones géographiques.

En effet, avec ce regroupement, aucune zone géographique ne se démarque significativement des autres, ce qui évite d'influencer les résultats des modèles de segmentation de manière prépondérante.

En ce qui concerne les entités auxquelles les acheteurs sont rattachés, la majeure partie (33,1%) est rattachée aux entités Coface présentes en Europe de l'Ouest, tandis que la plus faible proportion des acheteurs (3,0%) est rattachée aux entités Coface d'Amérique latine. Nous notons également que seulement 3,5% des acheteurs sont rattachés aux entités Coface d'Europe centrale et de l'Est, bien que plus de 20,5% des acheteurs exercent leurs activités dans ces régions. Ces variables, qui semblent fortement corrélées à première vue, pourraient donc fournir des informations différentes sur la structure de notre portefeuille d'assurés. Une analyse statistique des liens et des interactions entre ces variables sera effectuée dans la section *5.4*.

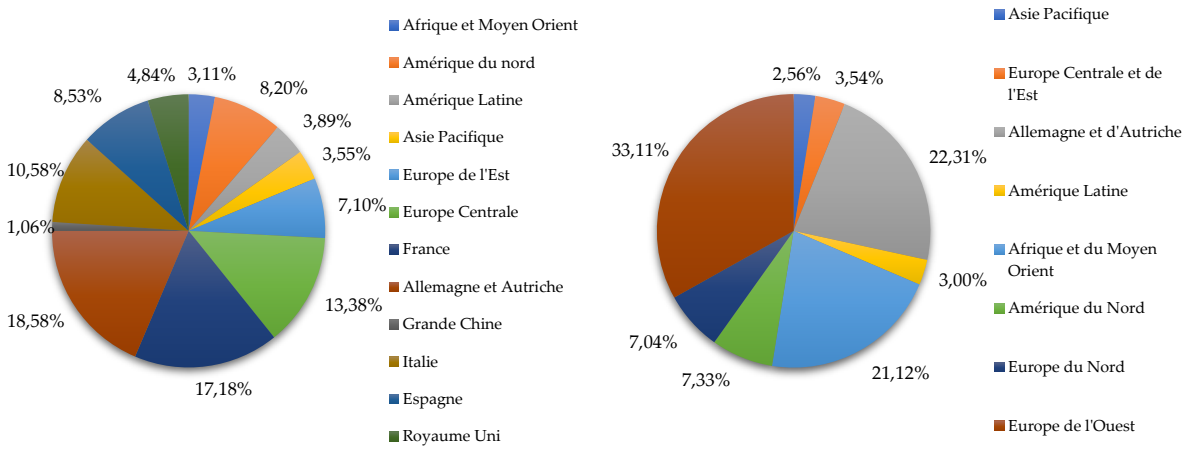


Figure 27 : Répartition des acheteurs par zone géographique (à gauche) et par entité de rattachement (à droite)

Le rating des acheteurs est une composante essentielle à prendre en compte avant la souscription d'une police d'assurance-crédit, car il renseigne sur la probabilité de solvabilité d'un acheteur. Le portefeuille de la Coface est principalement constitué d'acheteurs classés dans les tranches de rating 4 ou 5. En effet, 47,2% des acheteurs de notre base de données sont classés dans la tranche 4, et 15,0% dans la tranche 5. Seuls 7,1% des acheteurs se situent dans la tranche de rating 1. Ce résultat révèle une certaine prudence de la Coface dans l'exercice de son activité. En observant l'évolution de cette répartition au fil des années sur la *Figure 55* en annexe, on remarque également que la Coface a tiré des leçons de la crise financière de 2008. En effet, on constate une nette diminution des acheteurs de la tranche 1 entre 2008 (15,9%) et 2009 (9%), pour se stabiliser autour de 3% à partir de 2017. Notons que cette proportion était de 22,0% en

2007. On remarque également une augmentation de la proportion d'acheteurs de la tranche 4, passant de 39,7% en 2008 à 47,8% en 2009. Cette proportion est restée supérieure à 44,0% pour les années suivantes, atteignant même 50,8% en 2015.

En ce qui concerne la cible des acheteurs, la majorité d'entre eux se concentre sur le marché intérieur (72,2%). Cette répartition reste quasi-identique entre 2007 et 2021. Cela peut s'expliquer par le fait qu'un acheteur qui se concentre sur le marché intérieur a une meilleure compréhension du marché, maîtrise davantage ses risques, et possède une connaissance approfondie des détails logistiques liés à son activité ainsi que des tendances économiques du marché, ce qui pourrait réduire sa probabilité de défaut.

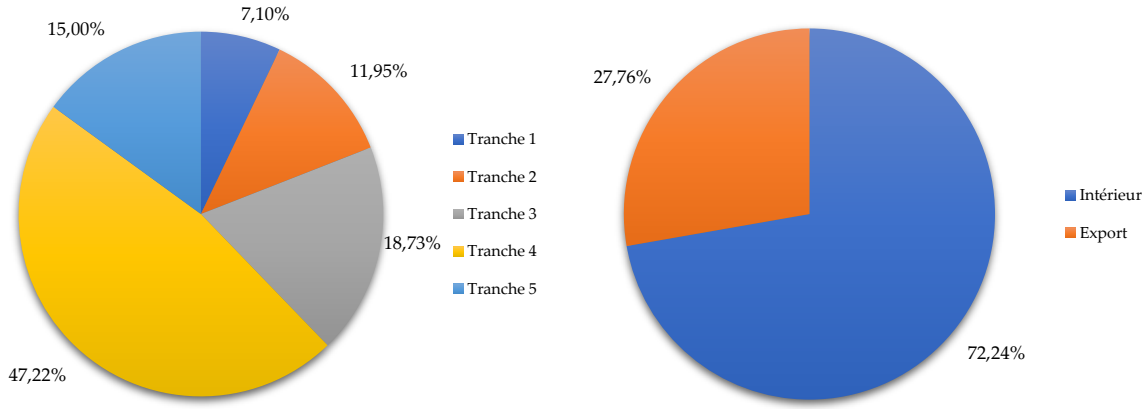


Figure 28 : Répartition des acheteurs par tranche de rating (à gauche) et par cible économique (à droite)

À ce stade, nous avons présenté l'ensemble des variables de l'étude. Les sections suivantes de ce chapitre auront pour objectif d'étudier les interactions entre ces variables afin de garantir qu'elles sont appropriées pour la construction des modèles de segmentation.

5.3. Statistiques descriptives bivariées

Dans cette section, nous effectuerons une analyse descriptive des liens et des dépendances entre les variables explicatives et les variables cibles. Ces analyses nous aideront à mieux comprendre les relations qui existent entre les 6 variables présentées précédemment et le défaut d'un acheteur. Le défaut sera représenté à travers la probabilité de défaut d'une part, et l'occurrence des défauts d'autre part.

5.3.1. La tranche d'exposition

L'exposition est une variable essentielle dans l'analyse de la sinistralité en assurance-crédit. D'après nos analyses, cette variable semble également être liée au défaut d'un acheteur. En effet, les figures ci-dessous montrent des variations de la probabilité de défaut moyenne et de la fréquence des défauts en fonction des tranches d'exposition. Nous observons une diminution de la probabilité de défaut avec l'augmentation de l'exposition. Ceci peut s'expliquer par le fait que les acheteurs sur lesquels la Coface choisit d'avoir une exposition importante sont généralement de bonne qualité. Cette stratégie est logique et elle est appliquée dans plusieurs secteurs, tels que le secteur bancaire. En effet, plus un client d'une banque est susceptible d'être solvable, plus la banque est disposée à lui accorder un prêt plus élevé.

Cependant, en examinant l'occurrence des défauts, nous remarquons une augmentation de la proportion de défauts avec l'exposition. Il semblerait donc que les acheteurs sur lesquels la Coface a des expositions élevées fassent des défauts plus fréquemment, mais que ces défauts soient de faible intensité. Cependant, en raison de leur faible gravité, il est fort probable que ces défauts soient plus liés à des retards de paiement qu'à des dépôts de bilan.

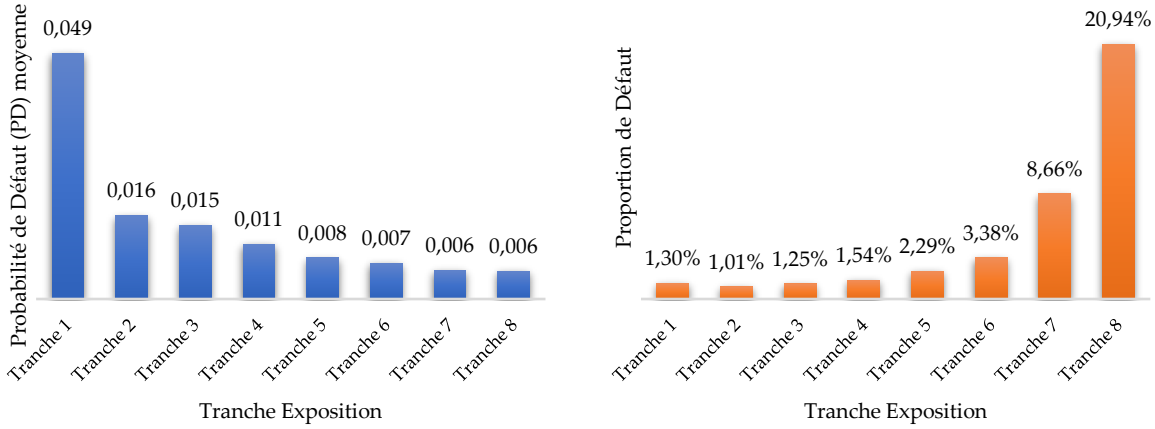


Figure 29 : Probabilité de défaut moyenne (à gauche) et proportion de défaut (à droite par tranche d'exposition

5.3.2. La région de l'entité Coface

Les répartitions des probabilités de défaut moyenne et des fréquences de défaut par région de rattachement des acheteurs semblent être très similaires. Les entités Coface qui enregistrent les probabilités de défaut les plus élevées sont celles d'Europe centrale et d'Europe de l'Est (0,05). Ce sont également les entités de cette région qui enregistrent les fréquences de défaut les plus élevées (2,7%). Les entités de la région Asie-Pacifique ont les acheteurs avec les probabilités de défaut les plus faibles (0,008) et la plus faible

proportion de défaut (0,54%). Dans l'ensemble, on observe une variation significative des défauts en passant d'une région de rattachement à une autre. Il semble donc que ces deux variables soient liées. La signification statistique de ces liens sera étudiée dans la section *5.4*.

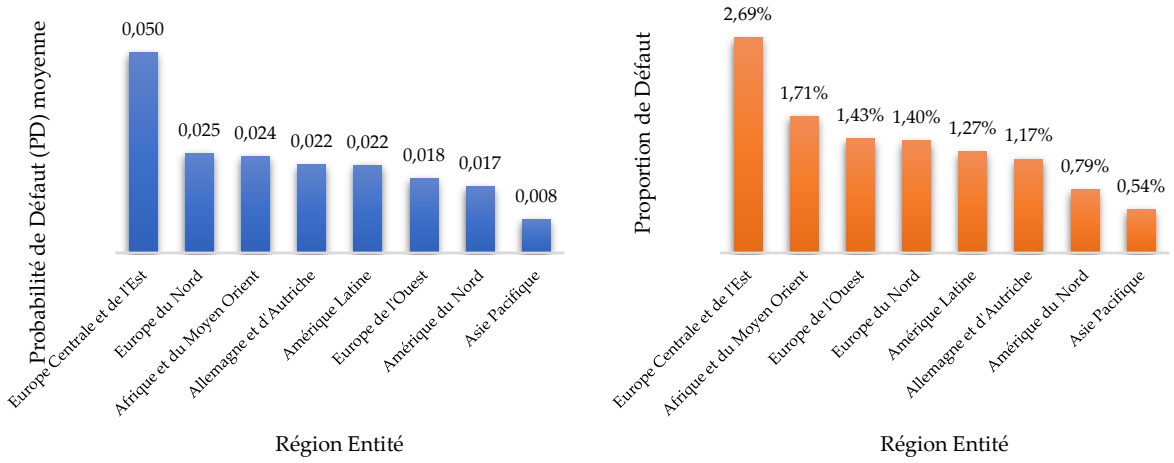


Figure 30 : Probabilité de défaut moyenne (à gauche) et proportion de défaut (à droite) par région de rattachement

5.3.3. La zone géographique de l'acheteur

Intuitivement, on s'attendrait à observer une forte corrélation entre la zone géographique d'un acheteur et ses risques de défaut. En effet, des études démontrent que les contextes économiques et sociaux dans lesquels une entreprise opère ont un impact sur son activité. Nos analyses semblent confirmer cette hypothèse. Les figures suivantes révèlent une disparité significative des défauts parmi les acheteurs exerçant dans différentes zones géographiques du monde. Les répartitions de la probabilité de défaut moyenne et de la fréquence des défauts semblent également similaires dans ce contexte.

La probabilité de défaut moyenne la plus élevée concerne les acheteurs d'Europe de l'Est (0,039), suivis par les acheteurs d'Italie (0,03). Les acheteurs d'Afrique et du Moyen-Orient viennent en troisième position, avec une probabilité de défaut moyenne de 0,028. Ces trois zones géographiques occupent également les premières places en termes de fréquence des défauts, avec respectivement 2,32%, 2,13% et 2,18% de défauts pour l'Europe de l'Est, l'Italie et l'Afrique du Moyen-Orient. Il semble que les acheteurs de ces régions soient confrontés à des difficultés économiques, administratives et logistiques qui peuvent affecter leur solvabilité, voire entraîner des retards de paiement. Cette hypothèse est renforcée par le fait que les expositions de la Coface dans ces zones sont plus faibles que dans d'autres zones (cf. *Figure 24*).

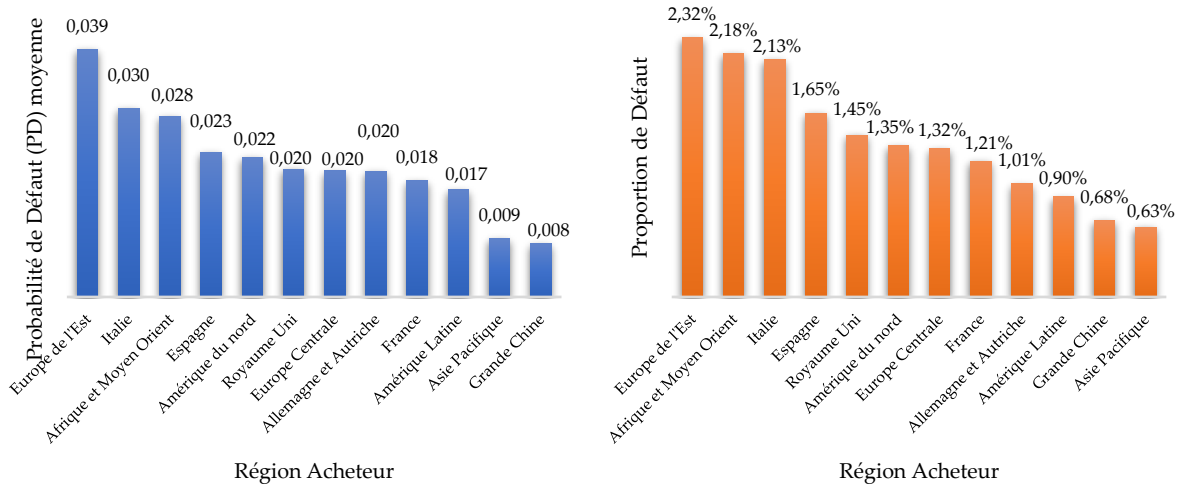


Figure 31 : Probabilité de défaut moyenne (à gauche) et proportion de défaut (à droite) par zone géographique des acheteurs

5.3.4. La cible économique

Comme nous l'avons soupçonné lors de nos analyses à la section *5.2.2*, les acheteurs qui ciblent le marché extérieur présentent une probabilité de défaut moyenne plus élevée, soit 0,023 par rapport à 0,021 pour les acheteurs dont l'activité est orientée vers le marché domestique. Cet écart se creuse davantage lorsque l'on examine l'occurrence des défauts. En effet, la fréquence de défaut chez les acheteurs orientés vers l'exportation s'élève à 1,81%, tandis qu'elle est de 1,24% pour les acheteurs axés sur le marché domestique. Cela représente une différence d'environ 46%, comparée à environ 10% dans le cas de la probabilité de défaut moyenne. Ces chiffres suggèrent que la cible économique semble avoir une corrélation plus forte avec l'occurrence des défauts qu'avec la probabilité de défaut (la gravité). Les différentes interactions des variables explicatives avec la probabilité de défaut et l'occurrence des défauts renforcent notre idée d'analyser les impacts de l'une ou l'autre de ces variables cibles sur la stabilité du modèle de segmentation optimal. L'intensité de ces relations sera analysée dans la section suivante à l'aide de tests statistiques.

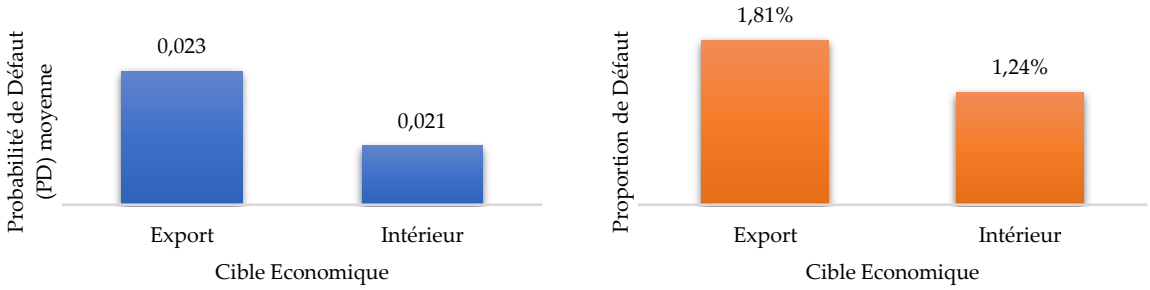


Figure 32 : Probabilité de défaut moyenne (à gauche) et proportion de défaut (à droite) par cible économique des acheteurs

5.3.5. Secteur d'activité

D'après les figures ci-dessous, la répartition de la probabilité de défaut moyenne et celle de la fréquence de défaut par secteur d'activité présentent des différences marquées. Bien que les quatre secteurs affichant les probabilités de défaut moyenne les plus élevées soient également ceux avec les fréquences de défaut les plus élevées, leur position diffère selon qu'on examine la fréquence des défauts ou la gravité des défauts. Ces quatre secteurs sont le textile, le bois et le papier (probabilité de défaut moyenne de 0,029 et fréquence de défaut de 2,08%), la construction (probabilité de défaut moyenne de 0,026 et fréquence de défaut de 1,46%), les activités commerciales (probabilité de défaut moyenne de 0,025 et fréquence de défaut de 1,48%) et l'agroalimentaire (probabilité de défaut moyenne de 0,023 et fréquence de défaut de 1,71%). La disparité observée dans les probabilités de défaut moyenne et les fréquences de défaut des acheteurs exerçant dans différents secteurs d'activité pourrait indiquer l'existence d'une corrélation entre le secteur d'activité et les risques de défaut.

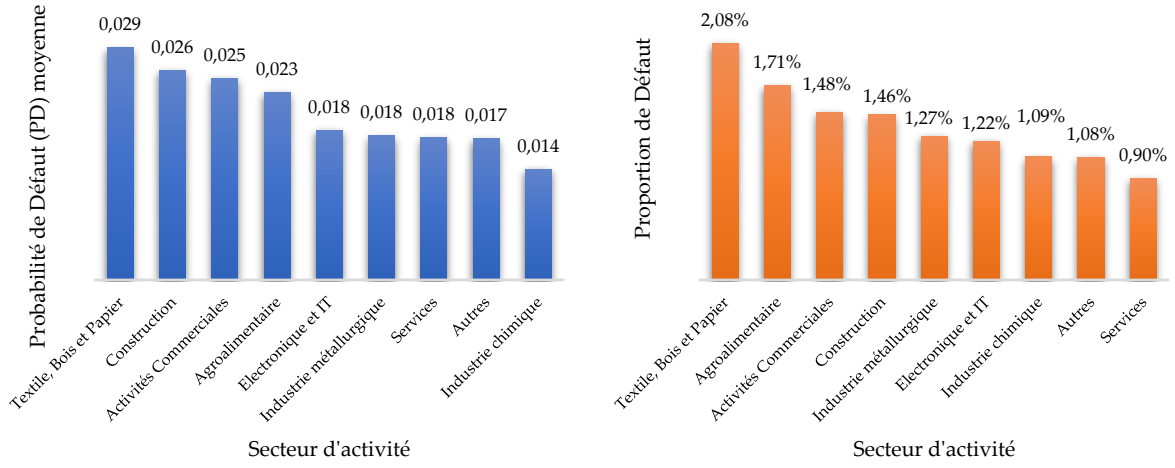


Figure 33 : Probabilité de défaut moyenne (à gauche) et proportion de défaut (à droite) par secteur d'activité des acheteurs

5.3.6. Tranche rating

Tout comme pour la zone géographique, on s'attend a priori à observer une corrélation significative entre le rating d'un acheteur et sa probabilité de défaut. Ceci s'explique par le fait que le score de rating est conçu pour refléter la solvabilité d'un acheteur, laquelle est intrinsèquement liée au risque de défaut. Nos analyses confirment cette hypothèse. En effet, les figures suivantes mettent en évidence une diminution de la probabilité de défaut moyenne à mesure que la notation augmente. Une tendance similaire est également observée en ce qui concerne la fréquence de défaut. La baisse de la probabilité de défaut moyenne en fonction de l'augmentation du rating semble

être plus marquée que celle de la fréquence de défaut. Cela pourrait indiquer une relation plus étroite entre le rating et la probabilité de défaut.

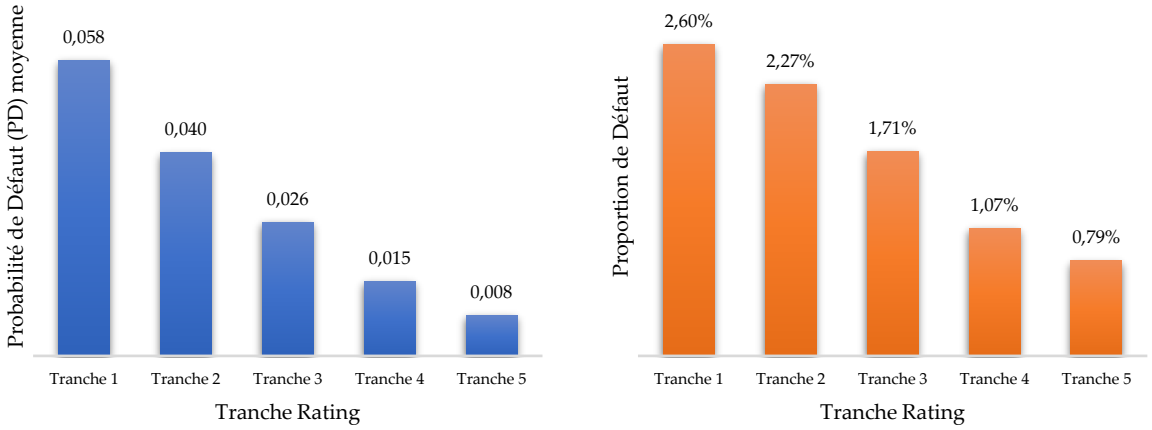


Figure 34 : Probabilité de défaut moyenne (à gauche) et proportion de défaut (à droite) par tranche rating

Dans les sections précédentes, nous avons examiné l'ensemble de nos variables pour nous familiariser avec celles-ci et comprendre les diverses interactions qui les sous-tendent. Une fois que nous aurons quantifié ces interactions et les aurons validées à l'aide de tests statistiques, nous pourrons procéder à la construction des modèles de segmentation et à l'étude de leur stabilité.

5.4. Choix des variables pour la segmentation

La revue de la littérature présentée au *Chapitre 3* montre que la segmentation des acheteurs en fonction d'un phénomène, en l'occurrence le défaut, doit être réalisée en se basant sur les variables liées à ce phénomène. Dans les sections précédentes, nous avons présenté les variables de notre étude et avons analysé de manière descriptive les liens qu'elles entretiennent avec les défauts des acheteurs. Dans cette section, nous allons utiliser les tests statistiques exposés dans les sections *4.8.1* et *4.8.2* pour valider statistiquement ces relations.

Nous testons la significativité statistique des liens entre nos variables explicatives et les deux variables représentant le défaut : la probabilité de défaut (variable continue) et l'occurrence des défauts (variable catégorielle). Comme mentionné à la section *4.8.2*, le test de Kruskal-Wallis est le plus approprié dans le premier cas, tandis que le test du chi-deux est approprié dans le second cas. Le test de normalité de Kolmogorov-Smirnov, qui produit une p-valeur de 0,000, nous permet de conclure, au seuil de 5%, que la probabilité de défaut ne suit pas une distribution normale (cf. *Figure 56* en annexe). Ceci confirme la pertinence de l'utilisation du test de Kruskal-Wallis.

Considérons le tableau suivant :

	Test de Kruskal-Wallis sur la probabilité de défaut			Test de Khi-deux sur l'occurrence des défauts		
	ST	PV	ED	ST	PV	VC
Région de l'entité Coface	29606	0,0000	0,0010	29613	0,0000	0,0316
Zone géographique de l'acheteur	47805	0,0000	0,0016	47866	0,0000	0,0401
Tranche rating	69271	0,0000	0,0023	68674	0,0000	0,0481
Cible économique	13813	0,0000	0,0005	13983	0,0000	0,0217
Tranche d'exposition	76223	0,0000	0,0026	78752	0,0000	0,0515
Secteur d'activité	24484	0,0000	0,0008	24577	0,0000	0,0287
ST = Statistiques du test, PV = p-valeurs, ED = Eta-deux et VC = V de Cramer						

Tableau 15 : Tests statistique des liaisons avec les variables cibles

Ces tests statistiques confirment les observations issues des analyses descriptives. En effet, les liens entre toutes les variables explicatives et la probabilité de défaut sont statistiquement significatifs au seuil de 1%. Il en va de même pour les liens entre les variables explicatives et l'occurrence des défauts. La probabilité de défaut et l'occurrence des défauts seront utilisées tour à tour comme variables explicatives lors de l'étude de la stabilité de nos modèles de segmentation. L'objectif sera d'analyser l'impact de l'utilisation de l'une ou l'autre de ces deux variables sur la stabilité des modèles de segmentation que nous construirons. Intéressons-nous à présent aux liens entre nos différentes variables explicatives en examinant le tableau suivant :

	Zone géographique de l'acheteur		Tranche rating		Cible économique		Tranche d'exposition		Secteur d'activité	
	PV	VC	PV	VC	PV	VC	PV	VC	PV	VC
Région de l'entité Coface	0,000	0,684	0,000	0,090	0,000	0,139	0,000	0,063	0,000	0,091
Zone géographique de l'acheteur			0,000	0,123	0,000	0,339	0,000	0,080	0,000	0,101
Tranche rating					0,000	0,052	0,000	0,108	0,000	0,040
Cible économique							0,000	0,148	0,000	0,191
Tranche d'exposition									0,000	0,070
PV = p-valeurs et VC = V de Cramer										

Tableau 16 : Tests statistique des liaisons entre variables explicatives

Ce tableau présente les p-valeurs des tests du khi-deux entre les variables explicatives, ainsi que les indices V de Cramer associés. Les résultats des tests révèlent que toutes les variables explicatives sont interconnectées, mais à des degrés variables. La variable « *Cible économique* » semble avoir des liens forts avec toutes les autres variables, bien qu'elle affiche les liens les moins marqués avec les variables expliquées. D'un point de vue statistique, cette situation soulève une problématique, car elle pourrait engendrer de la colinéarité. De même, les variables « *Région de l'entité Coface* » et « *Zone géographique de l'acheteur* » affichent une corrélation très marquée (V de Cramer = 0,684). Les méthodologies statistiques recommandent, dans de tels cas, de retirer la variable qui présente le lien le moins fort avec la variable cible, ce qui serait ici la « *Région de l'entité Coface* ».

Ces analyses mettent en évidence que les variables « *Cible économique* » et « *Région de l'entité Coface* » devraient être manipulées avec attention. Cependant, étant donné que les modèles de segmentation que nous mettrons en place sont tous non paramétriques, et que notre objectif n'est pas de prédire une variable cible, mais plutôt de créer des segments homogènes, nous préserverons ces deux variables dans nos modélisations. Toutefois, nous dédierons une section du *Chapitre 6* à l'analyse de leur impact sur la stabilité du modèle optimal.

Chapitre 6 : Résultats et discussion

Sommaire

Ce chapitre mettra en application la méthodologie présentée dans le *Chapitre 4*. Il débutera en exposant la configuration de nos modèles de segmentation ainsi que le choix des hyperparamètres optimaux. Ensuite, il se penchera sur l'étude de la stabilité proprement dite, et se conclura par une analyse des facteurs susceptibles d'influencer la stabilité d'un modèle de segmentation en assurance-crédit.

Dans ce chapitre, les variables explicatives sont celles présentées à la section *5.2.2*, tandis que la variable cible sera soit l'occurrence des défauts, soit la probabilité de défaut. La construction de nos modèles et les premières études de stabilité seront effectuées en utilisant l'occurrence des défauts. Nous justifions ce choix en raison de notre souhait de mener ces premières études sans l'influence des valeurs atypiques que peut prendre la probabilité de défaut pour certains acheteurs. Néanmoins, une analyse de l'effet de l'utilisation de la probabilité de défaut sur la performance et la stabilité du modèle de segmentation optimal sera réalisée dans la section *6.6.3* de ce mémoire.

Comme annoncé à la section *4.4*, pour toutes nos analyses, la base A (données couvrant la période de 2007 à 2021) sera divisée en deux sous-échantillons grâce à un tirage aléatoire stratifié sans remise. La stratification se fait en fonction des variables « *Année* » et « *Zone géographique de l'acheteur* ». Cette division permet d'obtenir un jeu d'apprentissage (représentant 80% de la base A) utilisé pour la construction et le calibrage des modèles, ainsi qu'un jeu de validation (représentant 20% de la base A) destiné aux tests et à l'étude de la stabilité de ces modèles.

Jeux de données	Proportions	Nombres d'acheteurs
Jeu d'apprentissage	80%	23 793 589
Jeu de validation	20%	5 948 394
Total (base A)	100%	24 388 428
Variables de stratification : « *Année* » et « *Zone géographique de l'acheteur* »		

Tableau 17 : Tailles des échantillons

6.1. Segmentation des acheteurs et choix des nombres de segments

La segmentation des acheteurs en groupes homogènes, en préambule aux simulations pour le calcul du SCR de souscription, sera effectuée à l'aide de trois modèles : le modèle CART, le modèle de k-prototypes et le modèle de classification ascendante hiérarchique. Ces modèles présentent des caractéristiques statistiques et mathématiques distinctes (cf. section *4.5*) et doivent donc être mis en œuvre selon des schémas précis. Cependant, ils ont tous en commun la particularité de permettre à l'utilisateur de spécifier le nombre de segments à créer. L'objectif principal de cette section est de sélectionner les nombres optimaux de segments pour nos modèles de segmentation en utilisant la méthode du coude, présentée à la section *3.1.4*.

6.1.1. Le modèle CART

Le modèle CART est un modèle supervisé qui permet d'effectuer des segmentations sur de nouveaux jeux de données en construisant des tables de correspondances (cf. section *4.3.1*). Chaque feuille de l'arbre correspond à un segment, et les acheteurs sont assignés à un segment en fonction des modalités des variables explicatives qui conduisent à cette feuille dans l'arbre. Le contrôle du nombre de segments revient donc à contrôler le nombre de feuilles. Le modèle d'arbre de décision que nous avons mis en place utilise la concentration de Gini comme mesure d'impureté (cf. section *4.5.1.1*). Aucune limite sur le nombre de feuilles de l'arbre n'a été fixée a priori. L'objectif est d'utiliser la technique d'élagage pour ajuster le nombre de feuilles a posteriori, tout en surveillant l'évolution de l'indice de Gini modifié.

Toujours dans le but de construire des arbres initiaux très détaillés avant l'élagage, nous avons choisi de ne pas imposer de seuils d'amélioration pour qu'un découpage soit considéré comme valable. L'indice de Gini modifié est calculé sur le jeu de données d'apprentissage selon la formule donnée à la section *4.6*.

Considérons les figures suivantes :

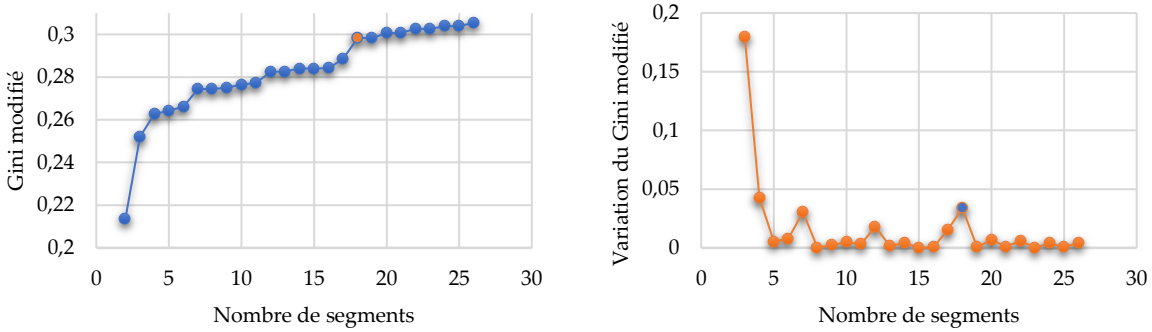


Figure 35 : Indice de Gini modifié (à gauche) et variation du Gini modifié (à droite) par nombre de segments pour le modèle CART

Ces figures illustrent l'évolution de l'indice de Gini modifié en fonction du nombre de segments (feuilles de l'arbre). Nous pouvons observer que la croissance de cet indice se stabilise après 18 segments. De plus, en examinant la variation de l'indice d'un nombre de segments à un autre, nous notons que le taux de croissance le plus élevé de l'indice de Gini modifié se produit également à 18 segments. Par conséquent, nous choisissons de retenir 18 segments pour le modèle CART dans la poursuite de nos analyses.

6.1.2. Le modèle de k-prototypes

Contrairement au modèle CART, le modèle de k-prototypes est un modèle non supervisé, ce qui signifie qu'il n'y a pas de spécification explicite d'une variable cible. Cependant, lors de la construction de ce modèle, nous incluons la variable « *Occurrence de défaut* » parmi les variables de segmentation. Ainsi, le modèle est construit en utilisant un ensemble de 7 variables, dont les 6 variables explicatives présentées à la section *5.2.2*. Bien que les fortes relations entre les variables explicatives et l'occurrence des défauts puissent suffire à garantir l'homogénéité des segments en termes de défaut, l'introduction de la variable cible « *Occurrence de défaut* » dans cette segmentation non supervisée renforce cette homogénéité. En effet, l'utilisation de cette variable cible apporte une certaine forme de supervision ou d'orientation à la construction des segments.

Les modèles des k-prototypes sont des modèles non paramétriques essentiellement basés sur le calcul des distances. Le modèle que nous mettons en œuvre utilise la mesure de distance présentée à la section *4.5.2.1*, où le poids attribué à la distance euclidienne est identique à celui attribué à la distance de Hamming ($\gamma = 1$). Le nombre maximal d'itérations est fixé à 100, bien que toutes nos exécutions

convergent bien avant ce seuil. Nous utilisons l'indice de Gini modifié, calculé sur l'échantillon d'apprentissage, pour déterminer le nombre optimal de segments.

Considérons les figures suivantes :

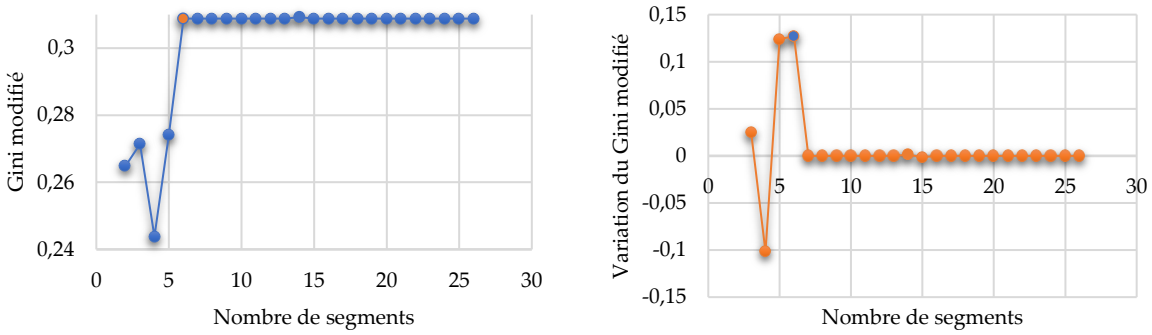


Figure 36 : Indice de Gini modifié (à gauche) et variation du Gini modifié (à droite) par nombre de segments pour le modèle de k-prototypes

Dans le cas du modèle de k-prototypes, nous remarquons une stabilisation de la croissance de l'indice de Gini modifié après 6 segments. Cette observation est confirmée par la figure de droite, qui montre que, après 6 segments, les variations de l'indice de Gini modifié sont nulles ou presque. La méthode du coude nous amène donc à retenir 6 segments pour le modèle de k-prototypes.

6.1.3. Le modèle de classification ascendante hiérarchique (CAH)

Le modèle de Classification Ascendante Hiérarchique (modèle CAH) est un modèle non supervisé, tout comme le modèle de k-prototypes. Cependant, contrairement à ce dernier, qui est un modèle qui fonctionne par partitionnement (cf. section *3.2.2.2*), le modèle CAH est construit à l'aide d'un algorithme hiérarchique (cf. section *3.2.2.1*). La construction de ce modèle commence par le calcul des distances entre les acheteurs. Pour cette tâche, nous utilisons la distance de Gower (cf. section *4.5.2.2*) en raison de la nature mixte de nos variables. Ces variables sont identiques à celles utilisées lors de la construction du modèle de k-prototypes. Ensuite, nous procédons au choix de la meilleure mesure de distance entre les segments. Nous avons le choix entre la distance minimale, la distance maximale, la distance moyenne et la distance de Ward (cf. section *4.5.2.2*). Nous optons pour la mesure de distance qui produit la répartition la plus uniforme des acheteurs dans les segments. Ce choix est motivé par deux raisons. Premièrement, une répartition uniforme des acheteurs dans les segments traduit un découpage plus fin des données. Deuxièmement, cela s'avère bénéfique pour la mise en œuvre des modèles de transfert qui suivra. En effet, une répartition uniforme des acheteurs dans les segments signifie qu'un jeu de données ayant pour variable cible le

segment sera plus équilibré[34]. Selon certaines études, construire un modèle d'apprentissage automatique sur un jeu de données déséquilibré peut introduire des biais dans les prédictions.

Pour les 18 segments, nous avons réalisé des tests d'adéquation des effectifs d'acheteurs dans ces segments par rapport à des lois uniformes (cf. section *4.8.3*) pour chaque type de distance. Le tableau suivant résume les résultats :

Distances	Statistiques du test	P-valeurs
Minimale	0,9444	0,0000
Maximale	0,7750	0,0000
Moyenne	0,5145	0,0001
Ward	0,1379	0,8384

Tableau 18 : Tests d'adéquation des effectifs d'acheteurs par segment à des lois uniformes

Au seuil de 1%, l'hypothèse nulle d'une distribution uniforme est rejetée pour toutes les distances, à l'exception de la distance de Ward. Nous choisissons donc cette dernière pour la poursuite de nos analyses. En ce qui concerne le choix du nombre optimal de segments, considérons les figures suivantes :

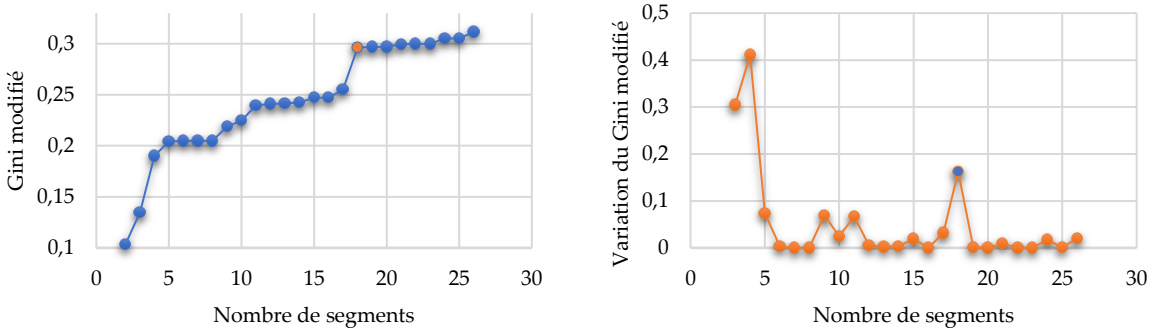


Figure 37 : Indice de Gini modifié (à gauche) et variation du Gini modifié (à droite) par nombre de segments pour le modèle CAH

Tout comme pour le modèle CART, le choix du nombre optimal de segments ici est très clair. À 18 segments, nous observons la plus grande amélioration de l'indice de Gini modifié. Par conséquent, nous retenons 18 segments pour le modèle CAH dans la poursuite de nos analyses.

Les modèles des k-prototypes et de Classification Ascendante Hiérarchique n'ont pas pu être calibrés sur l'ensemble de la base de données d'apprentissage en raison de leurs exigences en termes de temps de calcul et d'exécution. Conformément aux recommandations des travaux de Deng et al. (2021)[35], nous les avons construits sur des

[34] *Le mot « équilibré » ici signifie que les effectifs des différentes modalités de la variable cible sont proches voire égaux.*

[35] *(Deng, et al., 2021)*

sous-échantillons de taille 4 758 717, soit 20% de notre jeu d'apprentissage initial. En effet, selon Deng et al. (2021), si une segmentation est construite sur un échantillon représentatif d'une population, alors cette segmentation est très proche de celle qui aurait été construite sur la population totale. Afin d'assurer cette représentativité, nous avons utilisé un tirage aléatoire stratifié sans remise pour obtenir des sous-échantillons à partir de la base d'apprentissage. Les variables « *Année* », « *Zone géographique de l'acheteur* » et « *Occurrence de défaut* » ont été utilisées pour définir les strates de ces tirages.

6.2. Calibrage des modèles de transfert

Comme indiqué à la section *4.3.2*, les modèles non supervisés n'ont généralement pas la capacité directe de généraliser des segmentations apprises sur un jeu de données à un autre. Le modèle CAH ne fait pas exception à cette règle. Bien que des améliorations récentes sur le modèle de k-prototypes lui permettent de prédire des segmentations sur de nouveaux jeux de données, l'utilisation de cette fonctionnalité reste impossible dans notre cas. Cela s'explique par le fait que l'introduction de la variable cible dans la construction des segments, comme expliqué précédemment, implique que cette variable soit connue lors des prédictions avec le modèle de k-prototypes. Ce qui ne serait pas le cas en pratique, car au moment du calcul du SCR de souscription pour une année à venir, les défauts de cette année ne sont pas encore connus.

Toutes ces considérations justifient pleinement le recours à des modèles de transfert. Ces modèles apprennent les segmentations construites sur les jeux de données d'apprentissage afin de les reproduire sur de nouveaux jeux de données. Le schéma suivant illustre leur utilisation :

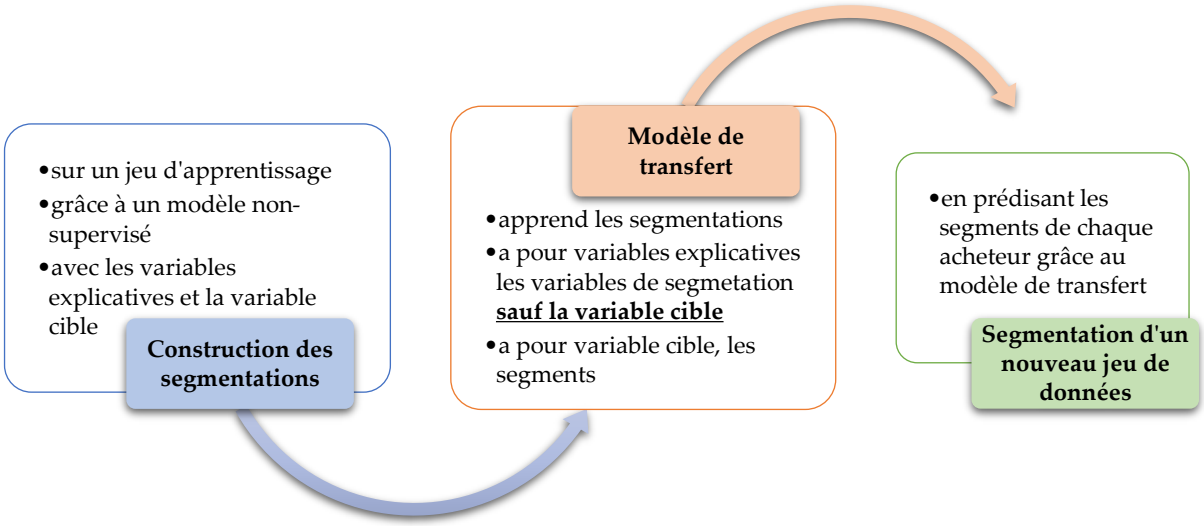


Figure 38 : Fonctionnement d'un modèle de transfert

En tant que modèle de transfert, nous aurons le choix entre le modèle de Forêts Aléatoires et le modèle d'eXtreme Gradient Boosting (XGBoost). Ces deux modèles candidats ont été retenus en raison de leur large utilisation dans la littérature, ainsi que du fait qu'ils utilisent deux approches différentes du Machine Learning : le bagging pour le modèle de Forêts Aléatoires et le boosting pour le modèle XGBoost (cf. section *4.5.1*).

Le calibrage de ces modèles consiste à sélectionner les hyperparamètres les mieux adaptés à notre jeu de données d'acheteurs. Nous effectuerons ce calibrage en évaluant le taux de bon classement et la sensibilité de ces modèles lors des validations croisées sur le jeu de données d'apprentissage. Commençons donc par expliquer ce qu'est la validation croisée.

6.2.1. La validation croisée

La validation croisée est une technique en Machine Learning qui permet d'évaluer les performances d'un modèle. Elle vise à estimer la capacité de généralisation du modèle, c'est-à-dire sa capacité à effectuer des prédictions correctes sur de nouvelles données. Contrairement à la méthode classique de division d'un jeu de données en jeux d'apprentissage et de test, la validation croisée présente l'avantage de diviser le jeu de données en plusieurs sous-échantillons afin d'effectuer plusieurs tests. Ceci permet d'obtenir des résultats plus robustes et de prévenir le surajustement. Dans cette étude, nous avons mis en œuvre la validation croisée en suivant les étapes de l'algorithme ci-dessous :

i. Sélectionner un hyperparamètre à calibrer ;
ii. Choisir une liste de valeurs pour ce paramètre à tester (en se basant sur la littérature et les études empiriques) ;
iii. Fixer une valeur de la liste pour l'hyperparamètre choisi à l'étape i ;
iv. Diviser de manière aléatoire le jeu de données d'apprentissage en 10 sous-échantillons ;
v. Pour k allant de 1 à 10 :
   - Construire le modèle en utilisant 9 des sous-échantillons et en excluant l'échantillon k ;
   - Calculer le taux de bon classement sur l'échantillon k ;
vi. Calculer la moyenne des 10 valeurs du taux de bon classement ;
vii. Répéter les étapes iii à vi pour chaque valeur d'hyperparamètre de la liste choisie à l'étape ii ;
viii. Sélectionner la valeur optimale de l'hyperparamètre choisi à l'étape i : celle qui présente la moyenne la plus élevée du taux bon classement ;
ix. Répéter les étapes i à viii pour chacun des hyperparamètres à calibrer.

À l'issue de cet algorithme, nous disposons des meilleurs hyperparamètres pour nos modèles. Ensuite, nous comparerons les performances de ces hyperparamètres aux performances des hyperparamètres par défaut afin de nous assurer que le calibrage a été bénéfique.

6.2.2. Liste et choix des hyperparamètres

Dans cette section, nous présenterons la liste des hyperparamètres de nos deux modèles de transfert, leurs utilités et leurs valeurs par défaut. Nous examinerons également les performances de ces modèles avant et après calibrage.

6.2.2.1. Le modèle de forêts aléatoires

Les travaux empiriques et la littérature recommandent le calibrage de cinq hyperparamètres pour ce modèle. Dans le logiciel R, il s'agit de :

- **num.trees :** Contrôle le nombre d'arbres de décision dans la forêt aléatoire. La valeur par défaut est de 500. Lors du calibrage, nous avons testé les valeurs de 50, 100, 200, 500 et 1000.
- **mtry :** Contrôle le nombre de variables utilisées pour le découpage à chaque nœud. La valeur par défaut est la racine carrée du nombre de variables, soit 2 dans notre cas avec 6 variables. Lors du calibrage, nous avons exploré les valeurs de 2, 3, 4 et 5.
- **splitrule :** Spécifie la fonction d'impureté utilisée pour le découpage des nœuds. La fonction par défaut est la concentration de Gini. Nous l'avons conservée lors de nos analyses.
- **min.node.size :** Contrôle le nombre minimal d'observations dans un nœud. La valeur par défaut est 1. Lors du calibrage, nous avons testé toutes les valeurs entre 1 et 10.
- **sample.fraction :** Spécifie la proportion d'observations à tirer aléatoirement pour la construction des arbres (cf. section *4.5.1.2.1*). La valeur par défaut est de 0,632. Lors du calibrage, nous avons exploré l'intervalle]0,1].

6.2.2.2. Le modèle XGBoost

Le modèle XGBoost est l'un des modèles les plus appréciés des Data Scientists en raison de ses excellentes performances. Cependant, ces performances nécessitent un effort conséquent en matière de calibrage. Le calibrage d'un modèle XGBoost sous R consiste à optimiser les hyperparamètres suivants :

- **nrounds :** Contrôle le nombre d'itérations de boosting (cf. section *4.5.1.3*). La valeur par défaut est de 100. Lors du calibrage, nous avons testé les valeurs de 50, 100, 150 et 200.

- **max_depth :** Contrôle la profondeur maximale des arbres de décision. La valeur par défaut est de 6 niveaux. Lors du calibrage, nous avons exploré les valeurs de 2, 3, 4, 5 et 6.
- **eta :** Contrôle le taux d'apprentissage. Il multiplie la contribution de chaque arbre de décision avant de l'ajouter à l'approximation en cours. La valeur par défaut est de 0,3. Plus ce paramètre est faible, plus le modèle est robuste contre le surajustement, mais il nécessite plus de temps pour s'exécuter. Lors du calibrage, nous avons testé les valeurs de 0,01, 0,015, 0,025, 0,05, 0,1 et 0,3.
- **gamma :** Spécifie la réduction minimale de l'impureté nécessaire pour effectuer la division d'un nœud. La valeur par défaut est de 0. Lors du calibrage, nous avons examiné les valeurs de 0, 0,05, 0,1, 0,5, 0,7, 0,9 et 1.
- **colsample_bytree :** Spécifie la proportion de variables à tirer aléatoirement lors de la construction de chaque arbre. La valeur par défaut est de 1. Lors du calibrage, nous avons testé les valeurs de 0,4, 0,6, 0,8 et 1.
- **subsample :** Contrôle la proportion d'observations à tirer aléatoirement lors de la construction de chaque arbre. La valeur par défaut est de 1. Lors du calibrage, nous avons exploré les valeurs de 0,5, 0,75 et 1.

Le tableau suivant récapitule les hyperparamètres retenus après le calibrage de nos modèles de transfert pour le modèle de k-prototypes et pour le modèle de classification ascendante hiérarchique :

Modèles non-supervisés	Modèles de transfert	Hyperparamètres après calibrage
k-prototypes	Forêts Aléatoires	num.trees : 500 mtry : 4 splitrule : "gini" min.node.size : 3 sample.fraction : 0,889
	XGBoost	nrounds : 200 max_depth : 4 eta : 0,3 gamma : 0 colsample_bytree : 0,6 subsample : 0,5
Classification Ascendante Hiérarchique	Forêts Aléatoires	num.trees : 500 mtry : 5 splitrule : "gini" min.node.size : 4 sample.fraction : 0,838
	XGBoost	nrounds : 200 max_depth : 6 eta : 0,3 gamma : 0 colsample_bytree : 1 subsample : 0,75

Tableau 19 : Liste des hyperparamètres retenus après calibrage

Analysons à présent les performances de ces modèles de transfert en validation croisée. Ces performances sont présentées dans le tableau suivant :

Modèles non-supervisés	Modèles de transfert	Avant calibrage	Après calibrage
k-prototypes	Forêts Aléatoires	TBC : 69,62%	TBC : 89,41%
		Sensibilité : 51,67%	Sensibilité : 83,31%
	XGBoost	TBC : 98,89%	TBC : 98,79%
		Sensibilité : 92,90%	Sensibilité : 94,88%
Classification Ascendante Hiérarchique	Forêts Aléatoires	TBC : 87,31%	TBC : 94,57%
		Sensibilité : 87,19%	Sensibilité : 95,17%
	XGBoost	TBC : 97,15%	TBC : 97,54%
		Sensibilité : 97,70%	Sensibilité : 98,02%
TBC = Taux de Bon Classement			

Tableau 20 : Performances des modèles de transfert avant et après calibrage

Il ressort de ces analyses que le calibrage améliore considérablement les performances des modèles de transfert. Nous remarquons également que les modèles de transfert affichent de très bonnes performances en validation croisée. Ceci nous conforte dans l'idée de leur utilisation et assure la fiabilité des segmentations qui seront construites grâce à ces modèles.

6.3. Récapitulatif des modèles retenus

À l'issue de tout ce qui précède, nous disposons d'une liste finale de modèles de segmentation que nous étudierons pour évaluer leur stabilité. Le modèle XGBoost affiche les meilleures performances en validation croisée et sera donc retenu comme modèle de transfert, avec les hyperparamètres correspondants. Le tableau suivant résume les résultats obtenus jusqu'ici :

Modèles de segmentation	Nombres de segments retenus	Modèles de transfert retenu	Performances des modèles de transfert en validation croisée	
			Taux de bon classement	Sensibilités
CART	18	Transfert assuré par des tables de correspondances		
k-prototypes	6	XGBoost	98,79%	94,88%
CAH	18	XGBoost	97,54%	98,02%

Tableau 21 : Récapitulatif des modèles de segmentation retenus

Dans la section suivante, nous allons étudier la stabilité de ces trois modèles selon la méthodologie décrite à la section *4.4*.

6.4. Etude de la stabilité des modèles de segmentation

L'étude de la stabilité se fera pour chacun des 3 modèles présentés à la section précédente, suivant la méthodologie détaillée à la section *4.4*. Pour chaque modèle, nous construisons, grâce à des tirages aléatoires stratifiés sans remise, 30 sous-échantillons de taille 14 276 153 chacun à partir du jeu de données d'apprentissage, soit 60% de ce dernier. Chaque modèle est construit sur chacun de ces sous-échantillons. Chaque modèle ainsi construit est utilisé pour segmenter le jeu de données de validation grâce à la procédure ou au modèle de transfert associé. Nous étudions ensuite la stabilité de nos 3 modèles de segmentation en analysant les variations de l'indice de Gini modifié sur le jeu de données de validation pour chaque sous-échantillon.

Tirer 60% du jeu de données d'apprentissage permet de garantir des perturbations non négligeables dans les sous-échantillons grâce aux tirages aléatoires. Toutefois, nous analyserons dans la section *6.6.1* les impacts que la variation de cette proportion peut avoir sur la stabilité du modèle optimal. Cette analyse nous donnera une idée des impacts que la taille du jeu de données d'entraînement peut avoir sur la stabilité d'un modèle de segmentation.

Afin de garantir la comparabilité des résultats obtenus pour chacun de nos modèles, nous effectuons les tirages selon un schéma identique pour les 3 modèles. En effet, pour construire le sous-échantillon i, nous fixons une graine[36] de $100 \times i$ lors de tous les tirages le concernant. Nous sommes ainsi certains que les variations que nous observerons lors de l'analyse des stabilités des modèles seront propres aux caractéristiques intrinsèques de ces derniers et non aux variations dans les schémas de tirage des sous-échantillons lorsque l'on passe d'un modèle à un autre.

6.4.1. Résultats de l'étude

Dans cette section, nous présenterons les résultats issus de l'implémentation de notre méthodologie d'étude de la stabilité. Nous commencerons par visualiser graphiquement nos résultats, puis nous effectuerons une analyse des pentes (cf. section *3.4*), et enfin, nous validerons ces résultats par des tests statistiques. À la section *3.4.1*, nous avons évoqué l'utilité de pouvoir quantifier les perturbations dans les structures des sous-échantillons. Pour cela, nous utiliserons le taux de variation des proportions de défaut dans chaque sous-échantillon par rapport à un sous-échantillon

[36] *En programmation, une graine (ou seed en anglais) est un nombre utilisé pour initialiser un générateur de nombres aléatoires. Dans notre cas, il détermine la séquence qui sera utilisée pour tirer aléatoirement les acheteurs. Utiliser une même graine permet de reproduire la même séquence aléatoire, ce qui est utile pour la reproductibilité et la comparaison des résultats.*

de référence. Nous faisons ce choix car nous pensons qu'une différence entre les valeurs du phénomène d'intérêt de deux sous-échantillons est un signe que ces deux échantillons ont des structures différentes du point de vue de ce phénomène.

6.4.1.1. Visualisation des résultats

Nous présentons ici les variations de l'indice de Gini modifié et des proportions de défaut dans les sous-échantillons pour chacun de nos modèles de segmentation. Nous examinerons également des histogrammes afin de comprendre les distributions des valeurs de cet indice dans les sous-échantillons. Cette compréhension, combinée aux résultats des tests statistiques de la section *6.4.2*, renforcera la robustesse de notre choix du modèle optimal et de nos conclusions.

6.4.1.1.1. Le modèle CART

Comme illustré sur la figure ci-dessous, les valeurs de l'indice de Gini modifié calculées grâce au modèle CART sur le jeu de validation semblent très proches d'un sous-échantillon à un autre. En effet, pour certains sous-échantillons, les valeurs prises par l'indice sont quasiment identiques. De plus, l'histogramme de ces valeurs (cf. *Figure 57* en annexe) montre qu'elles sont concentrées dans de très petits intervalles ($[0{,}25; 0{,}27]$ et $]0{,}28; 0{,}29]$). Cette proximité des valeurs de l'indice pourrait être un signe de stabilité de ce modèle.

Sur ce même graphique, nous observons une variation assez forte de la proportion de défaut lorsque l'on passe d'un sous-échantillon à un autre. Ceci montre que les sous-échantillons sont effectivement différents et que la construction de ces derniers par tirage aléatoire stratifié sans remise introduit bien des perturbations dans les jeux de données. Les variations facilement perceptibles des proportions de défaut entre les sous-échantillons nous confortent dans le choix du taux de variation de la proportion de défaut comme mesure de perturbation dans un jeu de données.

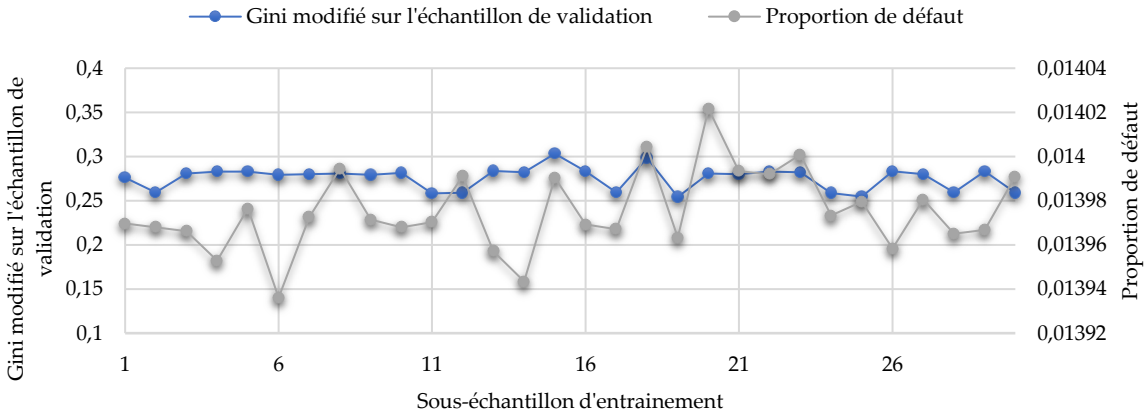


Figure 39 : Gini modifié sur l'échantillon de validation par sous-échantillon pour le modèle CART

6.4.1.1.2. Le modèle de k-prototypes + XGBoost

S'agissant du modèle de k-prototypes, les valeurs prises par l'indice de Gini modifié sur le jeu de données de validation semblent très variables et non stationnaires. En effet, nous observons sur la figure ci-dessous que chaque sous-échantillon conduit à une valeur de l'indice assez éloignée de celles produites par les autres sous-échantillons. Dans l'ensemble, les valeurs de l'indice semblent varier dans un intervalle assez large. De plus, l'histogramme de ces valeurs (cf. *Figure 58* en annexe) montre une répartition étalée sur l'intervalle [0,20; 0,29] avec des fréquences non négligeables pour chaque sous-intervalle. Ces remarques pourraient être des signes d'instabilité (cf. section *3.4.1*). L'analyse des pentes faite à la section suivante nous apportera plus de précision.

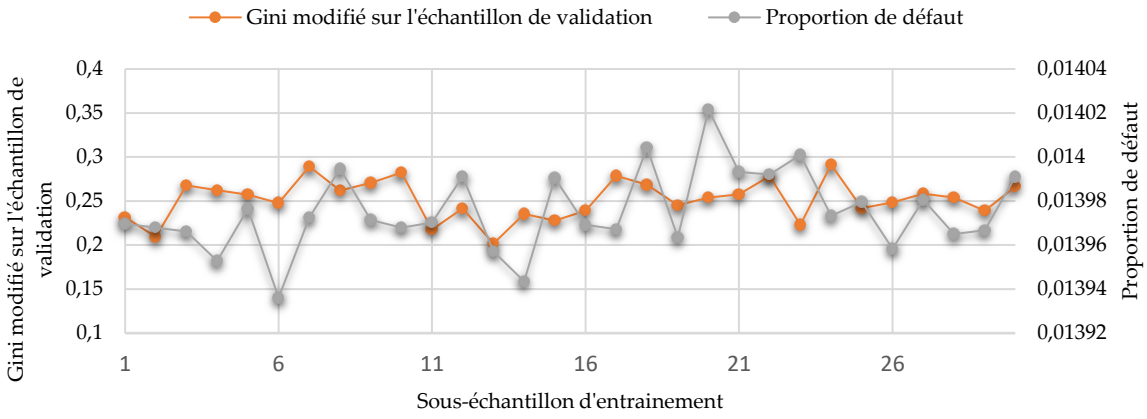


Figure 40 : Gini modifié sur l'échantillon de validation par sous-échantillon pour le modèle de k-prototypes

6.4.1.1.3. Le modèle CAH + XGBoost

D'après la *Figure 41*, la distribution des valeurs de l'indice de Gini modifié pour le modèle CAH ne semble pas s'éloigner de celle du modèle de k-prototypes. En effet, on observe des variations assez grandes des valeurs de l'indice quand on passe d'un sous-échantillon à un autre. De plus, l'histogramme de ces valeurs présenté sur la figure en annexe montre que l'indice prend des valeurs sur l'intervalle [0,20; 0,28] avec de grandes fréquences. Ceci se rapproche d'une distribution de probabilité uniforme sur cet intervalle. Dans la section *6.4.2*, nous validerons cette possible instabilité grâce à un test d'adéquation à une loi uniforme.

Le modèle CART semble être le seul modèle à présenter des signes de stabilité du point de vue statistique. Dans la section suivante, nous procéderons à une analyse des pentes afin de confirmer ou d'infirmer ce constat.

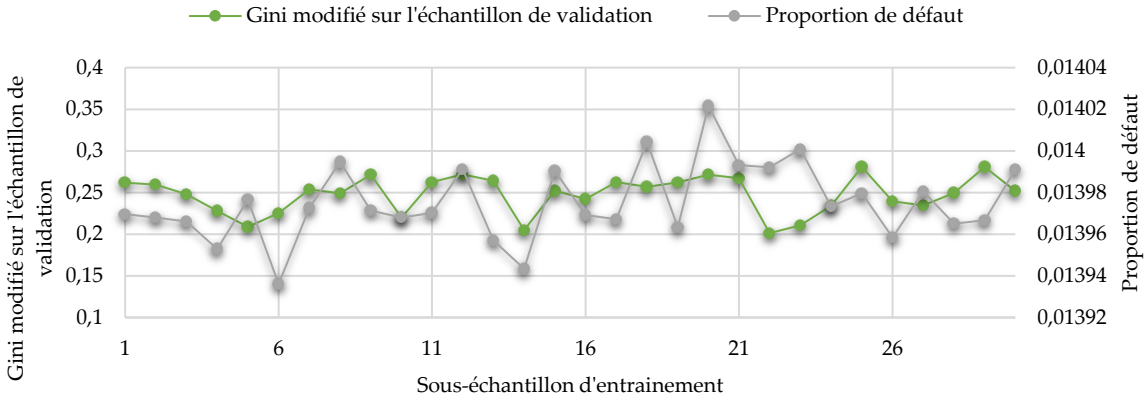


Figure 41 : Gini modifié sur l'échantillon de validation par sous-échantillon pour le modèle CAH

6.4.1.2. Analyse des pentes et des coefficients d'instabilité

Pour effectuer ces analyses, nous traçons sur le même graphique les taux de variation en valeur absolue de la proportion de défaut et les taux de variation en valeur absolue du Gini modifié correspondants pour nos trois modèles (cf. section *3.4.1*). Ces taux de variation sont calculés pour chaque sous-échantillon par rapport aux valeurs de référence suivantes :

Proportion de défaut	**Indice de Gini modifié**		
	Modèle CART	Modèle de k-prototypes	Modèle CAH
0,013992634	0,280611211	0,2456336	0,235124537

Tableau 22 : Valeurs de référence pour les calculs des taux de variation

Ces valeurs de référence sont calculées à partir d'un sous-échantillon fixe également construit par un tirage aléatoire stratifié sans remise de 60% du jeu de données d'apprentissage. La figure ci-dessous présente ces tracés :

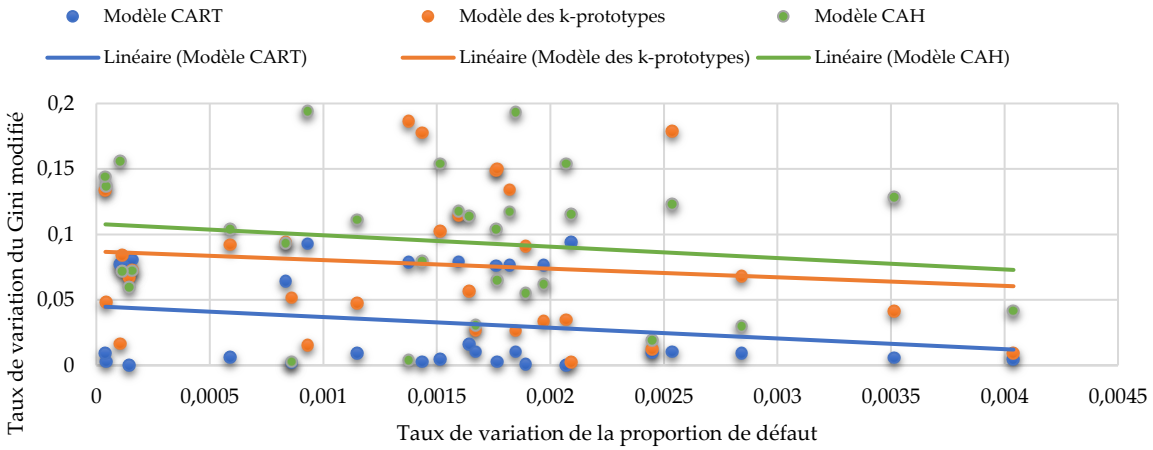


Figure 42 : Taux de modification de la structure des segmentations en fonction des taux de pertubation des échantillons d'entraînement

La première observation que nous faisons est la forte dispersion des taux de variation de l'indice de Gini modifié pour le modèle de k-prototypes et le modèle CAH. En effet, pour des niveaux de perturbation similaires des données, ces deux modèles produisent des variations plus importantes de l'indice de Gini modifié. Pour ces deux modèles, la majorité des taux de variation de l'indice sont au-dessus de 5%. À titre d'exemple, pour une perturbation de 0,14% du jeu de données, la variation de la structure de la segmentation mesurée par le taux de variation de l'indice de Gini modifié est de 7,94% pour le modèle CAH et 17,75% pour le modèle de k-prototypes. Ceci confirme les soupçons soulevés concernant leur instabilité dans les sections précédentes.

Dans le cas du modèle CART, cette même perturbation de 0,14% du jeu d'entraînement ne modifie la structure de la segmentation que de 0,25%. La *Figure 42* montre que pour toutes les perturbations du jeu d'entraînement testées, la plupart des variations de la structure de la segmentation obtenue en utilisant le modèle CART sont proches de 0%. Il semble donc que le modèle CART soit le plus stable des trois. L'analyse des tendances semble confirmer cette conclusion. En effet, nous observons que la courbe de tendance linéaire du modèle CART est la plus basse et presque horizontale (pente proche de 0). Bien que les pentes des courbes de tendance linéaire des autres modèles soient également très faibles, leurs courbes sont au-dessus de celle du modèle CART. Cela implique qu'en réponse à des perturbations dans le jeu de données d'entraînement, le modèle de k-prototypes et le modèle CAH présentent en moyenne de plus grandes modifications dans les structures des segmentations que le modèle CART. Pour quantifier ces observations, nous avons calculé des coefficients d'instabilité en utilisant la formule fournie à la section *3.4.1*. Le tableau suivant présente ces coefficients :

Modèles	**Pentes (*m*)**	**Ordonnés à l'origine (*c*)**	**Coefficients d'instabilité**
CART	-8,1751	0,0451	6,3411
k-prototypes	-6,5613	0,0869	7,6277
CAH	-8,6988	0,1080	9,7516

Tableau 23 : Coefficients d'instabilité

Le modèle CART affiche le coefficient d'instabilité le plus faible. Cela semble confirmer les observations faites à la section précédente. Les tests statistiques effectués dans la section suivante nous permettront de valider ces résultats.

6.4.2. Validation des résultats par des tests statistiques

Conformément à la méthodologie présentée à la section *4.8.3*, nous effectuons pour chaque modèle un test d'adéquation des 30 valeurs de l'indice de Gini modifié à une loi uniforme. Pour rappel, ces 30 valeurs sont calculées sur le jeu de données de

validation après que les modèles ont été calibrés sur les 30 sous-échantillons tirés du jeu de données d'apprentissage. Les résultats de ces tests sont résumés dans le tableau suivant :

Modèles	Statistiques du test	p-valeurs
CART	0,38108	0,0016
k-prototypes	0,19061	0,1845
CAH	0,20104	0,1419

Tableau 24 : Test d'adéquation des valeurs de l'indice de Gini modifié sur l'échantillon de validation à des lois uniformes

Au seuil de 1%, le modèle CART est le seul à rejeter l'hypothèse nulle selon laquelle les valeurs de l'indice de Gini modifié suivent une distribution uniforme. Ceci est dû au fait que, contrairement aux deux autres modèles, les valeurs de l'indice de Gini modifié sur le jeu de données de validation pour le modèle CART sont très similaires. Nous avions déjà remarqué cela lors de l'analyse des histogrammes à la section *6.4.1.1*. Cette analyse des histogrammes, associée au non-rejet de la distribution uniforme des valeurs de l'indice de Gini modifié pour les modèles des k-prototypes et de la CAH, indique que ces valeurs sont réparties sur un large intervalle, confirmant ainsi l'instabilité de ces modèles, selon la définition donnée à la section *3.4.1*.

6.5. Choix du modèle le plus stable

Au vu de tout ce qui précède, nous avons de bonnes raisons de penser que le modèle CART est le plus stable parmi les trois modèles étudiés jusqu'à présent. En effet, nous avons constaté que, comparativement au modèle de k-prototypes et au modèle CAH, des perturbations dans le jeu de données d'entraînement produisent moins de variations dans la structure des segmentations construites par le modèle CART. La figure suivante présente les variations des indices de Gini modifié pour nos trois modèles :

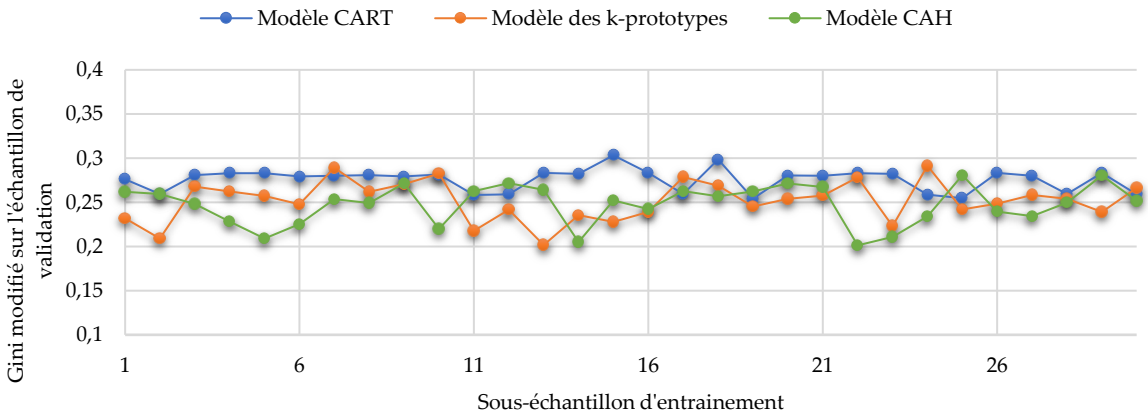


Figure 43 : Gini modifié sur l'échantillon de validation par sous-échantillon pour les 3 modèles

D'après cette figure, les oscillations du modèle CART sont les plus faibles. L'analyse des pentes, l'analyse des coefficients d'instabilité et les tests d'adéquation à des lois uniformes réalisés dans les sections précédentes confirment la meilleure stabilité du modèle CART. De plus, la *Figure 42* montre que les variations des structures des segmentations issues du modèle CART suivant des perturbations dans les données d'entraînement sont majoritairement proches de 0%. Ceci pourrait pousser à conclure sur la stabilité absolue du modèle CART. Toutefois, nous pensons qu'il est préférable d'approfondir les analyses et d'élargir la liste des modèles de référence pour tirer une telle conclusion. Ceci pourrait faire l'objet d'une prochaine étude.

Rappelons pour conclure que le modèle CART a également l'avantage d'être un modèle très simple à implémenter et facilement interprétable, ce qui en fait un atout pour son utilisation dans des travaux actuariels où l'interprétabilité des modèles est une nécessité.

Dans les sections suivantes, nous utiliserons ce modèle optimal pour analyser la stabilité vis-à-vis du SCR de souscription et pour étudier les facteurs susceptibles d'influencer la stabilité d'un modèle de segmentation.

6.6. Analyse des facteurs pouvant influencer la stabilité d'un modèle de segmentation

Nous allons conclure notre étude en analysant quelques facteurs susceptibles d'influencer la stabilité d'un modèle de segmentation. Le choix des facteurs à analyser a été inspiré par la littérature et par les conclusions tirées dans les sections précédentes. Toutefois, des recherches plus approfondies pourraient révéler d'autres sources d'instabilité. Le modèle CART, qui est le modèle le plus stable de notre liste, sera utilisé pour les analyses que nous effectuerons dans cette section.

6.6.1. Taille de l'échantillon d'entraînement

Plusieurs études sont d'accord pour affirmer qu'il existe une corrélation positive entre la taille de l'échantillon d'apprentissage et les performances d'un modèle prédictif construit sur cet échantillon (Dhiman, et al., 2023), (Ogundimu, et al., 2016), (De Jong, et al., 2019). D'après la *Figure 44*, ce lien positif semble également exister lorsque l'on s'intéresse à la stabilité d'un modèle de segmentation.

Nous remarquons qu'à mesure que la taille des échantillons d'entraînement augmente, les courbes de tendance deviennent horizontales et se rapprochent de l'axe des abscisses. La *Figure 60* en annexe confirme cette observation. En effectuant des tirages aléatoires stratifiés avec remise de 80% de la base d'apprentissage pour

construire des sous-échantillons d'entraînement, nous constatons que la plupart des taux de variation de l'indice de Gini modifié sont proches de 0%.

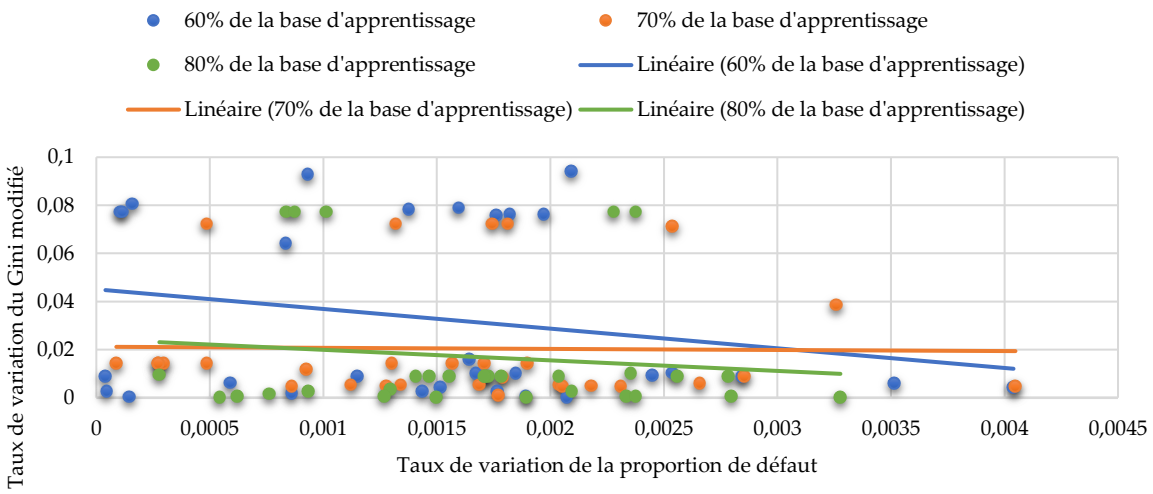


Figure 44 : Impacts des tailles d'échantillons

Dans le *Tableau 27* en annexe, on observe également des coefficients d'instabilité plus faibles pour les tirages à 70% (1,28) et à 80% (3,41). Ceci peut s'expliquer par le fait qu'augmenter la taille de l'échantillon d'entraînement augmente la probabilité de tenir compte de tous les scénarios possibles dans la structure du portefeuille. La prise en compte de ces divers scénarios réduit la possibilité de rencontrer de fortes perturbations naturelles dans les données au cours d'une année, ce qui rend les segmentations obtenues plus stables dans le temps. Nous nous sommes limités à 80% car nous estimons que 20% de la base d'apprentissage est le seuil minimal nécessaire pour provoquer des perturbations observables dans les jeux de données d'entraînement. De plus, en deçà de 80%, nous parvenons à obtenir des résultats interprétables.

6.6.2. Le choix des variables

Que ce soit dans le cadre d'une prédiction ou d'une segmentation, le choix des variables d'entraînement exerce une influence non négligeable sur les performances du modèle utilisé. Dans cette section, nous allons analyser cette influence sur la stabilité. Considérons les 4 scénarios suivants :

Scénario 0	Scénario 1	Scénario 2	Scénario 3
Toutes les variables	Sans la variable «***Région de l'entité Coface***»	Sans la variable «***Cible économique***»	Sans les variables «***Cible économique***» et «***Région de l'entité Coface***»

Tableau 25 : Choix des variables explicatives

Dans la section *5.4*, nous avons remarqué une forte corrélation entre les variables « *Région de l'entité Coface* » et « *Zone géographique de l'acheteur* ». Nous avions également

noté de fortes corrélations entre toutes les variables explicatives et la variable « *Cible économique* ». Ces deux observations nous ont incités à approfondir l'analyse de l'impact de ces variables sur la stabilité du modèle optimal afin de tirer des conclusions généralisables. La figure suivante présente les pentes des trois scénarios définis ci-dessus.

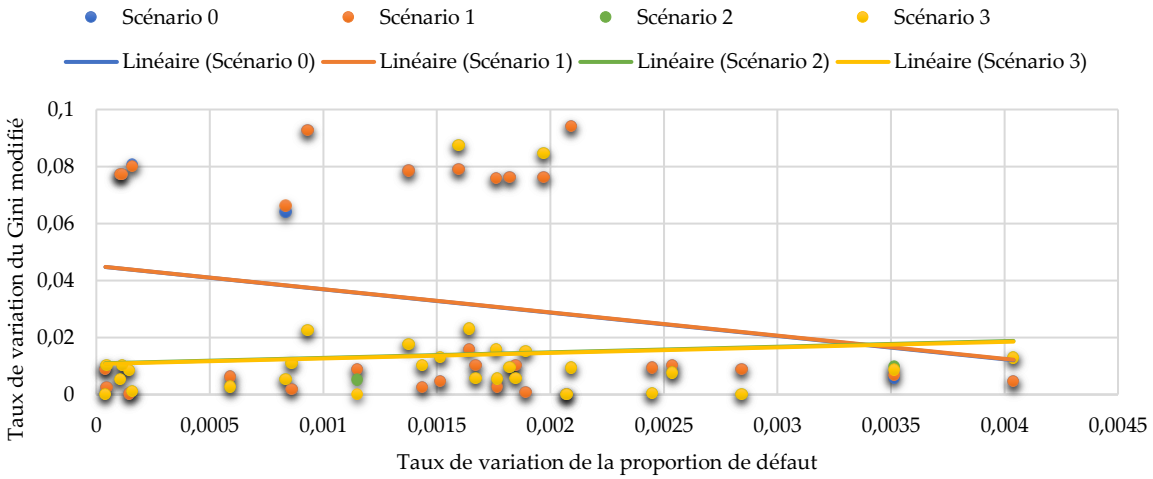


Figure 45 : Impacts des choix des variables explicatives

Nous remarquons que le scénario 1 et le scénario 0 se confondent. Cela signifie qu'exclure la variable « *Région de l'entité Coface* » de la construction du modèle n'a aucun ou très peu d'impact, que ce soit sur les performances ou sur la stabilité du modèle de segmentation optimal. Ceci s'observe également au niveau des coefficient d'instabilité (cf. *Tableau 28* en annexe), qui sont presque identiques pour ces deux scénarios (6,34 pour le scénario 0 et 6,33 pour le scénario 1). Cette observation intrigante nous a incités à examiner la contribution des différentes variables à la prédiction de l'occurrence des défauts. Pour ce faire, nous avons calculé l'importance de chaque variable dans ladite prédiction à l'aide d'un modèle de forêts aléatoires. Le modèle de forêts aléatoires calcule cette importance en se basant sur l'impureté de Gini (cf. section *4.5.1.1*).

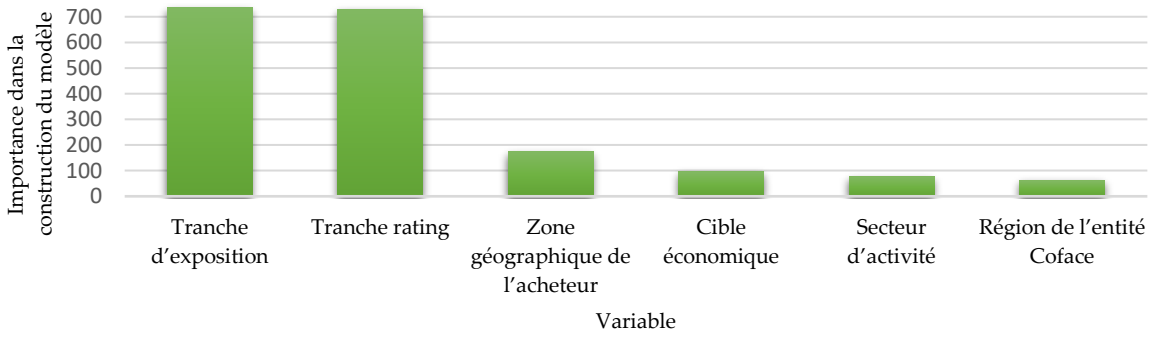


Figure 46 : Importance des variables dans la prédiction de l'occurrence d'un défaut

D'après la *Figure 46* ci-dessus, qui présente ces importances, la variable « *Région de l'entité Coface* » est la moins importante dans la prédiction de l'occurrence d'un défaut. Ceci pourrait expliquer le fait qu'elle n'a pas d'impact sur la qualité de notre segmentation.

En revanche, le scénario 2 a un effet très significatif sur la stabilité du modèle de segmentation étudié (le modèle CART). En effet, construire ce modèle sans la variable « *Cible économique* » améliore significativement la stabilité du modèle de segmentation, mais pas nécessairement la qualité de la segmentation qui reste du même ordre de grandeur que celle du scénario 1 (cf. *Figure 61* en annexe). De plus, le coefficient d'instabilité de ce scénario est de 1,52, contre 6,34 pour le scénario 0. Il semblerait donc que les fortes liaisons entre cette variable et les autres variables explicatives aient des impacts négatifs sur la stabilité du modèle de segmentation. Et ce, malgré le fait que l'importance de cette variable dans la prédiction de la variable cible soit non négligeable et supérieure à celle de la variable « *Région de l'entité Coface* ». Le scénario 3, qui se confond avec le scénario 2, vient confirmer l'observation faite au scénario 1 : la variable « *Région de l'entité Coface* » n'a aucun effet ni sur la qualité ni sur la stabilité du modèle de segmentation étudié.

6.6.3. La nature des variables

Nous nous intéresserons particulièrement à la nature de la variable cible. Dans la section *5.1*, lors de la présentation des variables, nous avions mentionné que la cible pouvait être soit la probabilité de défaut, soit l'occurrence de défaut. Dans cette section, nous analysons l'impact que l'un ou l'autre de ces choix peut avoir sur la stabilité du modèle CART. La figure ci-dessous présente les pentes pour les deux variables cibles.

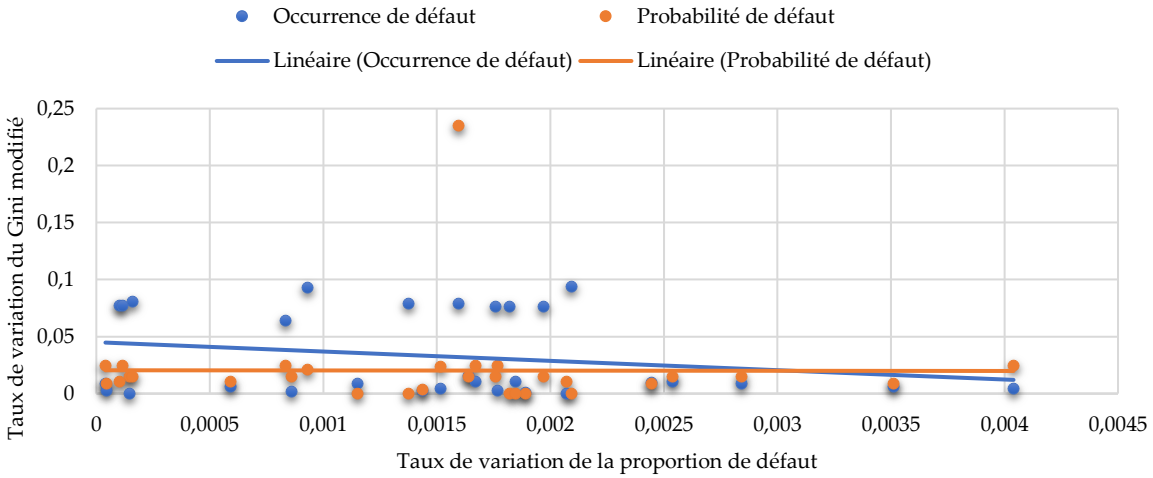


Figure 47 : Impact de la nature de la variable cible

Il ressort de cette figure que l'utilisation de la probabilité de défaut (variable continue) permet d'obtenir un modèle plus stable. Le coefficient de stabilité dans ce

cas est de 1,10, alors qu'il est de 6,34 dans le cas de l'utilisation de l'occurrence des défauts (cf. *Tableau 29* en annexe). Ceci s'explique par le fait qu'en discrétisant la variable « *Probabilité de défaut* » pour obtenir la variable « *Occurrence de défaut* », nous perdons une partie de l'information. Bien que cette perte d'information soit compensée par l'élimination des effets des valeurs atypiques que peut prendre la probabilité de défaut sur la segmentation, les conséquences sur la stabilité semblent être considérables. Toutefois, il est important de nuancer cette conclusion en analysant la *Figure 62* en annexe. D'après cette figure, le modèle de segmentation construit sur la probabilité de défaut est certes plus stable, mais il est également moins performant. En effet, l'indice de Gini modifié de ce modèle reste bien en dessous de celui du modèle construit sur l'occurrence de défaut.

Ce facteur rejoint d'une certaine manière le facteur analysé précédemment car il concerne le choix des variables. Toutefois, il met en lumière à la fois un avantage et un inconvénient de la discrétisation des variables, rejoignant ainsi la littérature qui est très mitigée sur le sujet. En effet, selon des travaux empiriques, il est important de faire un arbitrage entre les avantages et les inconvénients de la discrétisation des variables en fonction du modèle et des objectifs visés, ce qui n'est pas toujours évident. Approfondir les analyses concernant cet arbitrage pour l'ensemble de nos variables explicatives pourrait constituer une étape ultérieure pour notre étude.

6.7. Analyse de la stabilité vis-à-vis du SCR de souscription

Dans cette section, nous allons analyser les impacts que de petites perturbations dans les jeux de données d'entraînement peuvent avoir sur le SCR de souscription. Pour des raisons de confidentialité, le SCR analysé sera le SCR avant application des mesures correctrices. Cette analyse sera effectuée à l'aide du modèle CART, qui a été identifié comme le modèle le plus stable dans les sections précédentes. Les calculs des SCR de souscription se feront selon la méthodologie décrite dans la section *2.3* et en utilisant la base B (cf. section *4.4*), correspondant aux données du quatrième trimestre de 2022.

Nous allons également analyser les impacts que la prise en compte des sources d'instabilité présentées à la section *6.6* peut avoir sur le SCR de souscription, qui est le résultat final du modèle interne partiel de la Coface. Pour ce faire, considérons les scénarios suivants :

Scénario 0	Scénario 1	Scénario 2
Modèle CART avec toutes les variables explicatives et avec l'occurrence de défaut comme variable cible.	Modèle CART sans les variables «***Cible économique***» et «***Région de l'entité Coface***» et avec l'occurrence de défaut comme variable cible.	Modèle CART sans les variables «***Cible économique***» et «***Région de l'entité Coface***» et avec la probabilité de défaut comme variable cible.

Tableau 26 : Prise en compte des sources d'instabilité

Ces scénarios sont testés sur des bases de données construites par tirage aléatoire stratifié avec remise de 60% de la base de données d'apprentissage (cf. section *4.4*). Le schéma de construction de ces bases de données, et donc le schéma de perturbation des données, est identique pour les trois scénarios. La figure suivante présente les taux de variation du SCR de souscription en fonction des taux de perturbation dans les données :

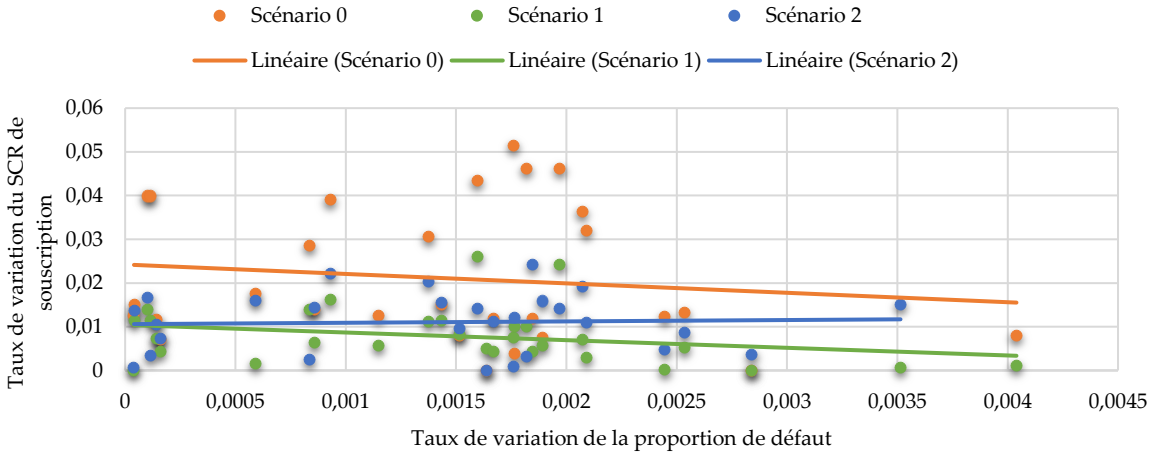


Figure 48 : Taux de variation du SCR de souscription en fonction des taux de perturbation des échantillons d'entraînement

Dans le scénario 0, nous observons de fortes variations du SCR de souscription pouvant atteindre plus de 5%. Dans ce scénario, une part considérable des variations du SCR dépasse 3%. Cependant, après l'élimination des sources d'instabilité identifiées à la section *6.6*, nous constatons une réduction significative des variations du SCR pour des perturbations équivalentes des jeux de données d'entraînement. En effet, dans le scénario 1, en supprimant la variable « *Cible économique* », fortement corrélée aux autres variables explicatives, ainsi que la variable « *Région de l'entité Coface* », qui contribue le moins à la prédiction de l'occurrence des défauts, la plupart des taux de variation du SCR tombent en dessous de 1,6%. De plus, la majorité des taux de variation du SCR de souscription se situent en dessous de 1%. Le scénario 2, qui utilise la probabilité de défaut comme variable cible, montre également une baisse générale des taux de variation du SCR de souscription. Bien que les taux de variation du SCR issus de ce scénario soient en moyenne plus élevés que ceux du scénario 1, la plupart restent en dessous de 2%, ce qui est bien meilleur que les 5% initiaux. Toutefois, dans ce cas précis, il serait préférable d'utiliser l'occurrence de défaut comme variable cible pour parvenir à une stabilité globale accrue du modèle interne partiel.

Il convient de noter que plusieurs autres modèles et algorithmes, en plus des modèles de segmentation, interviennent dans le calcul du résultat final du modèle interne partiel (cf. section *2.3*). Une partie significative des variations du SCR de

souscription observées après l'élimination des sources d'instabilité pourrait découler de ces autres modèles ou algorithmes. Par exemple, d'autres travaux réalisés sur le modèle interne partiel de la Coface indiquent qu'une partie des variations du SCR observées dans les scénarios 1 et 2 pourrait être due au modèle de calibrage des corrélations utilisé dans le modèle de Merton (cf. section *2.4.1*). En pratique, pour prendre en compte ces sources de bruit supplémentaires et les neutraliser, la Coface effectue des backtests et calcule des marges correctrices qu'elle ajoute à ses différents phénomènes après modélisation.

Il ressort de cette section que la robustesse du SCR de souscription en assurance-crédit dépend de la stabilité des modèles de segmentation utilisés dans son calcul. En analysant la stabilité du modèle de segmentation retenu et en éliminant les sources d'instabilité identifiées grâce à cette analyse, nous avons réussi à obtenir un SCR plus stable et plus robuste. Les résultats présentés dans la *Figure 48* ci-dessus et analysés dans cette section soulignent une fois de plus l'importance de l'étude de la stabilité des modèles de segmentation dans la modélisation de la sinistralité en assurance-crédit, ainsi que dans d'autres secteurs d'activité éventuels.

6.8. Discussions finales et recommandations

Au *Chapitre 3* de ce mémoire, nous avons défini la stabilité et expliqué les raisons pour lesquelles un modèle de segmentation devrait être stable. Cependant, nous avons constaté que plusieurs modèles de segmentation fréquemment utilisés ne sont pas stables. De plus, la segmentation est largement utilisée dans plusieurs secteurs d'activité sans qu'aucune analyse de stabilité ne soit préalablement effectuée. En assurance-crédit, la segmentation est nécessaire voire incontournable dans le calcul du SCR de souscription. Or, un modèle instable peut avoir des impacts considérables sur les valeurs du SCR de souscription, comme nous l'avons vu dans la section *6.7*. La méthodologie d'étude de la stabilité d'un modèle de segmentation proposée au *Chapitre 4* et utilisée tout au long de ce mémoire peut être résumée par les étapes suivantes :

i. Choisir des modèles de segmentation candidats.
ii. Sélectionner une métrique pour évaluer la qualité d'une segmentation (cf. section *4.6*).
iii. Sélectionner une métrique reflétant la structure d'un sous-échantillon (cf. section *6.4.1*).
iv. Diviser les données en jeux d'apprentissage et en jeux de validation.
v. Calibrer les modèles de segmentation candidats sur les données (cf. section *6.1*).
vi. Créer plusieurs sous-échantillons d'entraînement (minimum 30) à partir du jeu d'apprentissage (cf. section *6.4*).

vii. Entraîner les modèles candidats sur chaque sous-échantillon et calculer les valeurs de la métrique de qualité sur le jeu de validation (cf. section *6.4.1.1*).
viii. Utiliser des visualisations graphiques et l'analyse des pentes pour sélectionner le modèle le plus stable (cf. section *6.4.1.2*).
ix. Valider cette sélection statistiquement à l'aide des tests d'adéquation à des lois uniformes (cf. section *6.4.2*).
x. Identifier les facteurs susceptibles d'influencer la stabilité du modèle retenu à l'étape précédente.
xi. Faire varier ces facteurs en calculant un coefficient d'instabilité à chaque fois (cf. section *6.6*).
xii. Identifier la configuration la plus stable des facteurs recensés à l'étape x. Cette configuration est celle qui minimise le coefficient d'instabilité.

Sur la base de nos travaux et de nos résultats, nous formulons les recommandations suivantes à l'attention de tous les acteurs des secteurs de l'assurance, de la banque, de l'académie et d'autres secteurs employant la segmentation dans leurs travaux :

- Mettre en place une méthodologie claire permettant d'évaluer la stabilité des modèles de segmentation utilisés. Le *Chapitre 4* de ce mémoire présente une méthodologie claire et détaillée permettant d'étudier la stabilité d'un modèle de segmentation. Ce chapitre ainsi que le résumé présenté ci-dessus peuvent servir de guides à adapter à des contextes spécifiques si nécessaire.
- Veillez à ce que les coefficients d'instabilité des modèles de segmentation tels que définis à la section *3.4* et utilisés dans des contextes similaires au nôtre, se situent dans l'intervalle [0; 1,10]. Dans la section *3.4.1* de ce mémoire, nous avons défini un coefficient d'instabilité que nous avons testé dans le cadre de notre étude. Nous avons réussi à obtenir des valeurs aussi basses que 1,10 pour ce coefficient en tenant compte des facteurs susceptibles d'influencer la stabilité d'un modèle de segmentation. Cependant, il convient de noter que des études plus approfondies pourraient permettre de réduire davantage cet intervalle et de généraliser ce coefficient.
- Veiller à construire des modèles de segmentation sur des échantillons de taille aussi élevée que possible. Nous avons montré dans la section *6.6.1* que la taille des échantillons peut avoir un impact sur la stabilité des modèles de segmentation.
- Effectuer une sélection minutieuse des variables de segmentation. Que ce soit dans une segmentation supervisée ou non supervisée, il est nécessaire d'utiliser des variables qui apportent une information supplémentaire au modèle. Dans le cas d'une segmentation supervisée, nous avons montré que l'importance des

variables dans la prédiction de la variable réponse devrait être prise en compte dans le choix des variables de segmentation (cf. section *6.6.2*).

- Construire des segmentations à partir des variables qui sont peu ou pas du tout corrélées entre elles. D'après les analyses effectuées dans la section *6.6.2*, il semblerait également que la présence d'une ou plusieurs variables fortement corrélées avec les autres puisse avoir des impacts négatifs sur la stabilité d'un modèle de segmentation.
- Tenir compte de la nature des variables lors de la construction des segmentations. En fonction des données, il est possible que certains prétraitements ou certaines transformations, tels que la discrétisation (cf. section *6.6.3*), aient des impacts sur la stabilité d'un modèle de segmentation. Il est donc utile d'effectuer systématiquement une analyse de ces impacts afin de décider quels prétraitements conserver dans la modélisation.

6.9. Limites de l'étude

Notre étude se présente comme un guide méthodologique pour l'étude de la stabilité des modèles de segmentation en assurance-crédit et dans d'autres secteurs. Cependant, son cadre est restrictif car elle comporte certaines limites. Ces limites comprennent les suivantes :

- **Ressources informatiques :** Comme mentionné à la section *6.1*, l'entraînement des modèles de k-prototypes et des modèles CAH s'est fait sur des sous-échantillons d'observations en raison de la voracité de ces modèles en termes de ressources informatiques et de temps de calcul. Bien que cette pratique soit justifiée dans la littérature, il aurait été préférable d'utiliser l'ensemble des données disponibles pour la construction de ces modèles.
- **Nombre de modèles de segmentation analysés :** Dans notre étude, nous avons examiné trois modèles de segmentation parmi les centaines que propose la littérature. Bien que ces trois modèles soient les plus couramment utilisés et recommandés pour des problématiques similaires à la nôtre, il aurait été possible d'obtenir des informations complémentaires en examinant d'autres modèles disponibles.
- **Sources d'instabilité étudiées :** Nous avons analysé trois facteurs susceptibles d'influencer la stabilité d'un modèle de segmentation. Le choix de ces trois facteurs s'inspirait de la littérature. Cependant, une recherche plus approfondie pourrait révéler d'autres sources d'instabilité. Une exploration approfondie des sources d'instabilité pourrait constituer une perspective intéressante pour ce mémoire, offrant ainsi une base plus solide pour une calibration améliorée de notre coefficient d'instabilité.

- **Métriques utilisées :** Pour effectuer l'analyse des pentes à la section *6.4.1.2*, nous avons choisi les variations de la proportion de défaut pour quantifier les perturbations dans les jeux d'entraînement. Cette métrique semble adaptée à notre étude et nous permet d'obtenir des résultats interprétables. Toutefois, une proportion, en raison de son caractère général, pourrait masquer certaines informations concernant la structure des jeux de données, laissant penser à une similarité entre deux jeux de données totalement différents, par exemple. Il n'est donc pas impossible qu'une analyse plus approfondie conduise à la proposition d'une métrique plus précise.

Conclusion

La problématique de cette étude consistait à déterminer comment étudier et valider la stabilité d'un modèle de segmentation dans le cadre de l'assurance-crédit. Ce mémoire a souligné l'importance cruciale de la stabilité des modèles de segmentation en général, avec une emphase particulière sur son rôle dans le calcul du SCR de souscription en assurance-crédit. De plus, nous avons proposé des métriques et des méthodes appropriées pour évaluer cette stabilité. En nous basant sur les recommandations de la littérature, nous avons développé une méthodologie rigoureuse pour étudier la stabilité d'un modèle de segmentation en assurance-crédit. Cette méthodologie comprend des visualisations graphiques, des analyses des variations par rapport aux scénarios de référence, des analyses des pentes et des tests statistiques.

À partir des analyses des pentes, nous avons conçu une formule permettant de calculer un coefficient d'instabilité. D'après nos résultats, nous recommandons, dans le contexte de l'assurance-crédit et de situations similaires à celles traitées dans ce mémoire, l'utilisation de modèles de segmentation dont le coefficient d'instabilité se situe entre 0 et 1,10. Notre méthodologie propose dans un premier temps de valider la stabilité d'un modèle de segmentation au moyen d'un test d'adéquation à une loi uniforme. En effet, nos résultats indiquent que la stabilité d'un modèle peut être confirmée si les valeurs de l'indice de Gini modifié (une métrique mesurant la qualité d'une segmentation) issues des perturbations de l'échantillon d'entraînement ne suivent pas une distribution uniforme. Une fois la stabilité d'un modèle validée, nous suggérons d'utiliser un coefficient d'instabilité pour améliorer davantage sa stabilité en jouant sur les facteurs susceptibles d'influencer celle-ci.

Nos travaux ont été appliqués à trois modèles de segmentation, à savoir le modèle CAH, le modèle de k-prototypes, et le modèle CART utilisé par la Compagnie française d'assurance pour le commerce extérieur (Coface). Cette application a apporté une validation pratique à notre méthodologie. Le modèle CART se révèle être le plus stable. De plus, ce modèle a l'avantage d'être très interprétable, ce qui est un atout du point de vue métier. Il ressort également de nos analyses que des échantillons de petite taille, des choix de variables inappropriés, et des prétraitements inadaptés peuvent être des sources d'instabilité. Nous avons terminé notre étude en démontrant la contribution de la stabilité des modèles de segmentation à la robustesse du SCR de souscription en assurance-crédit. En effet, nous montrons que plus un modèle est stable, moins le SCR de souscription obtenu est sensible à de petites variations dans le jeu de données de calibrage. Il convient de noter que cette méthodologie peut être adaptée et appliquée à d'autres secteurs que l'assurance-crédit, élargissant ainsi son

champ d'application. Les recommandations formulées à la suite de nos résultats visent à guider les acteurs de l'assurance, de la banque, et d'autres secteurs dans la construction de leurs modèles de segmentation, contribuant ainsi à renforcer la robustesse et la qualité de leurs analyses.

Néanmoins, il est important de souligner que notre étude n'épuise pas toutes les perspectives de recherche dans ce domaine. Des pistes supplémentaires peuvent être explorées, notamment en approfondissant l'analyse de la stabilité pour d'autres modèles de segmentation ou en investiguant d'autres sources potentielles d'instabilité. De plus, l'application de notre méthodologie à un éventail plus large de données pourrait fournir des informations complémentaires et contribuer à son amélioration continue.

En fin de compte, ce mémoire se positionne comme un point de départ pour une meilleure compréhension et une gestion plus précise de la stabilité des modèles de segmentation en assurance-crédit et dans d'autres secteurs où la segmentation est utilisée. En encourageant la communauté actuarielle, bancaire et académique à intégrer cette dimension dans leurs pratiques, nous aspirons à optimiser les résultats fournis par les modèles de segmentation et à renforcer la confiance dans les outils qui contiennent ou utilisent de tels modèles.

Bibliographie

Ahmed, Mohiuddin, Raihan, Seraj et Syed, Mohammed. 2020. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics.* 2020, Vol. 9.8, 1295.

Allianz. 2023. Assurance-crédit entreprise contre les impayés clients. [En ligne] 2023. [Citation : 19 06 2023.] https://www.allianz-trade.fr/assurance-credit.html.

Altassura. 2023. Prime d'assurance crédit. [En ligne] 2023. [Citation : 19 06 2023.] https://www.assurance-credit-entreprise.fr/glossary/prime-assurance-credit/.

Amit, Kumar Ojha. 2017. Use a Classification and Regression Tree (CART) for Quick Data Insights. [En ligne] ISIXSIGMA, 15 05 2017. [Citation : 30 06 2023.] https://www.isixsigma.com/lean-methodology/use-a-classification-and-regression-tree-cart-for-quick-data-insights/.

Anderberg, Michael R. 1973. The broad view of cluster analysis. *Cluster analysis for applications.* 1973, Vol. 1.1, 1-9.

Awasthi, Pranjal et Reza, Zadeh. 2010. Supervised clustering. *Advances in neural information processing systems.* 2010, 23.

Baert, Bram, et al. 2007. Transdermal penetration behaviour of drugs: CART-clustering, QSPR and selection of model compounds. *Bioorganic & medicinal chemistry.* 2007, Vol. 15.22, 6943-6955.

Bair, Eric. 2013. Semi-supervised clustering methods. *Wiley Interdisciplinary Reviews: Computational Statistics.* 2013, Vol. 5.5, 349-361.

Ball, Geoffrey H. et David, J. Hall. 1965. ISODATA, a novel method of data analysis and pattern classification. *Stanford research institute.* Menlo Park, CA, 1965, Vol. 4.

Ben-Hur, Asa, Elisseeff, Andre et Guyon., Isabelle. 2002. A stability based method for discovering structure in clustered data. *Biocomputing.* 2002, 6-17.

Berger, Vance W. et Zhou, YanYan. 2014. Kolmogorov–smirnov test: Overview. *Wiley statsref: Statistics reference online.* 2014.

Blashfield, Roger K. et Aldenderfer, Mark S. . 1978. The literature on cluster analysis. *Multivariate behavioral research.* 1978, Vol. 13.3, 271-295.

Blashfield, Roger. 1977. The equivalence of three statistical packages for performing hierarchical cluster analysis. *Psychometrika.* 1977, Vol. 42.3, 429-431.

Boumezoued, Alexandre. 2022. *Cours ENSAE - Asssurance-Vie : Tables et modèles pour les risques biométriques.* 2022.

Breiman, Leo. 2001. Random forests. *Machine learning.* 2001, Vol. 45, 5-32.

Breiman, Leo, Friedman, Jerome et Olshen, Richard. 1984. *Classification and regression trees.* 1984.

Brown, Robert L. et Leon, R. Gottlieb. 2007. *Introduction to ratemaking and loss reserving for property and casualty insurance.* s.l. : Actex Publications, 2007.

Burduk, Robert. 2020. Classification Performance Metric for Imbalance Data Based on Recall and Selectivity Normalized in Class Labels. *arXiv preprint arXiv : 2006.* 2020, 13319.

Caponnetto, Andrea et Alexander, Rakhlin. 2006. Stability Properties of Empirical Risk Minimization over Donsker Classes. *Journal of Machine Learning Research.* 2006, Vol. 7.12.

Charbonneau, Alexis. 2003. *La mise en place d'un modèle d'évaluation du risque de crédit dans le cadre de la réforme Solvabilité 2.* s.l. : Université d'Orleans, 2003.

Chen, Tianqi et Guestrin, Carlos. 2016. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.* 2016.

Chevalier, Fabien et Le Bellac, Jérôme. 2013. *La classification.* s.l. : Université de Rennes 1, 2013.

Choi, Seung-Seok, Sung-Hyuk, Cha et Charles, C. Ta. 2010. A survey of binary similarity and distance measures. *Journal of systemics, cybernetics and informatics.* 2010, Vol. 8.1, 43-48.

COFACE. 2023. Our mission. [En ligne] 2023. [Citation : 19 06 2023.] https://www.coface.com/Group/Our-mission.

—. 2022. *Rapport financier, Premier semestre 2022.* 2022.

Cormack, Richard M. 1971. A review of classification. *Journal of the Royal Statistical Society.* Series A (General) , 1971, Vol. 134.3, 321-353.

Cutler, Adele, Cutler, D. Richard et Stev, John R. 2012. Random forests. *Ensemble machine learning: Methods and applications.* 2012, 157-175.

Das, Sanjiv Ranjan, Freed, Laurence et Geng, Gary. 2002. *Correlated default risk.* s.l. : EFA 2003 Annual Conference Paper, 2002.

De Jong, Valentijn et Eijkemans, Marinus. 2019. Sample size considerations and predictive performance of multinomial logistic prediction models. *Statistics in medicine.* 9, 2019, Vol. 38, 1601-1619.

Deng, Jiayi, Ding, Yi et Zhu, Yingqiu. 2021. Subsampling Spectral Clustering for Large-Scale Social Networks. *arXiv preprint arXiv.* 2021, Vol. 2110, 13613.

Dettling, Marcel et Bühlmann, Peter. 2002. Supervised clustering of genes. *Genome biology.* 2002, Vol. 3, 1-15.

Dhiman, Paula, Ma, Jie et Qi, Cathy. 2023. Sample size requirements are not being considered in studies developing prediction models for binary outcomes: a systematic review. *BMC Medical Research Methodology.* 1, 2023, Vol. 23, 1-11.

Diday, Edwin et Simon, J. C. 1976. Clustering analysis : Digital pattern recognition. *Springer Berlin.* 1976, 47-94.

Dolnicar, Sara et Katie, Lazarevski. 2009. Methodological reasons for the theory/practice divide in market segmentation. *Journal of marketing management.* 2009, Vol. 25.3-4, 357-373.

Dorman, Karin S. et Ranjan, Maitra. 2022. An efficient k-modes algorithm for clustering categorical datasets. *Statistical Analysis and Data Mining: The ASA Data Science Journal.* 2022, Vol. 15.1, 83-97.

Dreyfuss, Marie-Laure. 2015. *Les grands principes de Solvabilité 2.* s.l. : Argus Editions, 2015. 978-2-35474-209-6.

Dudoit, Sandrine et Jane , Fridlyand. 2002. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome biology.* 2002, Vol. 3, 1-21.

—. 2003. Bagging to improve the accuracy of a clustering procedure. *Bioinformatics.* 2003, Vol. 19.9, 1090-1099.

Duffie, Darrell. 2011. *Measuring corporate default risk.* s.l. : Oxford University Press, 2011.

Everitt, Brian S. 1980. *Cluster analysis.* London : Heinemann, 1980.

Friedland, Jacqueline. 2013. *Fundamentals of general insurance actuarial analysis.* s.l. : Society of Actuaries, 2013.

Gaid, Ouafa, Halilou, Zouleikha et Sliman, Aouatef. 2015. *Stabilité au sens de Lyapunov.* 2015.

Gastwirth, Joseph L. 1971. A general definition of the Lorenz curve. *Econometrica: Journal of the Econometric Society.* 1971, 1037-1039.

—. 1972. The estimation of the Lorenz curve and Gini index. *The review of economics and statistics.* 1972, 306-316.

Genuer, Robin et Poggi, Jean-Michel. 2017. *Arbres CART et Forêts aléatoires : Importance et sélection de variables.* 2017.

Germain, Valentin. 2022. *Mémoire d'actuariat - Prise en compte du changement climatique dans la modélisation des risques biométriques et financiers.* 2022.

Gower, John C. 1971. A general coefficient of similarity and some of its properties. *Biometrics.* 1971, 857-871.

Hancock, T. P., Coomans, D. H. et Everingha, Y. L. . 2003. Supervised hierarchical clustering using CART. *Proceedings of MODSIM 2003 International Congress on Modelling and Simulation, Townsville, QLD, Australia.* 2003.

Hastie, Trevor, Tibshirani, Robert et Friedman, Jero. 2009. Unsupervised learning. *The elements of statistical learning: Data mining, inference, and prediction.* 2009, 485-585.

Herve, Fabrice. 2002. La persistance de la performance des fonds de pension individuels britanniques: une étude empirique sur des fonds investis en actions et des fonds obligataires. *Documento de trabajo de el Laboratoire Orleáns de Gestion.* 2002, 3.

Homa, Hajibaba, Grün, Bettina et Dolnicar, Sara. 2020. Improving the stability of market segmentation analysis. *International Journal of Contemporary Hospitality Management.* 2020, Vol. 32.4, 1393-1411.

Huang, Zhexue. 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery.* 2.3, 1998, 283-304.

Jain, Anil K. et Dubes., Richard C. 1988. *Algorithms for clustering data.* s.l. : Prentice-Hall, Inc., 1988.

Jain, Anil K. et Moreau, J. V. . 1987. Bootstrap technique in cluster analysis. *Pattern Recognition.* 1987, Vol. 20.5, 547-568.

Jancey, R. C. 1966. Multidimensional group analysis. *Australian Journal of Botany.* 1966, Vol. 14.1, 127-130.

Kassambara, A. 2018. *CART model: Decision tree essentials.* 2018.

Kerr, M. Kathleen et Gary, A. Churchill. 2001. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proceedings of the national academy of sciences.* 2001, Vol. 98.16, 8961-8965.

Le Vallois, Franck. 2021. *Cours ENSAE - Réglementation prudentielle en Assurance : Le régime prudentiel Solvabilité 2.* 2021.

Lefebvre, Benoît. 2017. *Mémoire - Modélisation des défauts totaux, une application à la tarification des traités en Excédent de Sinistre en Assurance-Crédit.* 2017.

Levine, Erel et Eytan, Domany. 2001. Resampling method for unsupervised estimation of cluster validity. *Neural computation.* 2001, Vol. 13.11, 2573-2593.

Liaw, Andy et Wiener, Matthew. 2002. Classification and regression by randomForest. *R news.* 2002, Vol. 2.3, 18-22.

Liu, Tianmou, Han, Yu et Blair, Rachael Hageman. 2022. Stability estimation for unsupervised clustering: A review. *Wiley Interdisciplinary Reviews: Computational Statistics.* 2022, Vol. 14.6, e1575.

Lorr, M. 1983. Cluster Analysis for Social Scientists. Techniques for Analyzing and Simplifying Complex Blocks of Data. *Jossey-Bass.* 1983, 233.

Madhulatha, T. Soni. 2012. An overview on clustering methods. *arXiv preprint arXiv.* 1205, 2012, 1117.

McKight, Patrick E. et Najab, Julius. 2010. Kruskal-wallis test. *The corsini encyclopedia of psychology.* 2010, 1-1.

Mehalla, Sophian. 2021. *Taux d'intérêt pour l'assurance : approximations et calibrages de modèles.* 2021.

Merton, Robert C. 1974. On the pricing of corporate debt: The risk structure of interest rates. *The Journal of finance.* 29, 1974, Vol. 2, 449-470.

Milligan, Glenn W et Martha, C. Cooper. 1987. Methodology review: Clustering methods. *Applied psychological measurement.* 1987, Vol. 11.4, 329-354.

Müller, Henriette et Ulrich, Hamm. 2014. Stability of market segmentation with cluster analysis–A methodological approach. *Food Quality and Preference.* 2014, Vol. 34, 70-78.

Ndoye, Babacar. 2019. *Risque de Souscription pour une branche Assurance-Crédit: Comparaison entre l'approche assurantielle et l'approche risque de crédit.* 2019.

Nielsen, Didrik. 2016. *Tree boosting with xgboost-why does xgboost win "every" machine learning competition?* s.l. : NTNU, 2016. MS thesis.

Ogundimu, Emmanuel et Douglas, Altman. 2016. Adequate sample size for developing prediction models is not simply related to events per variable. *Journal of clinical epidemiology.* 76, 2016, 175-182.

Ooreka. 2023. Pourcentage garantie par l'assurance crédit. [En ligne] 2023. [Citation : 19 06 2023.] https://assurance-

professionnelle.ooreka.fr/astuce/voir/747081/pourcentage-garantie-par-l-assurance-credit.

Pádraig, Cunningham, Matthieu, Cord et Sarah, Jane. 2008. Supervised learning : Machine learning techniques for multimedia: case studies on organization and retrieval. *Springer Berlin Heidelberg.* 2008, 21-49.

Rand, William M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association.* 1971, Vol. 66.336, 846-850.

Richardson, John T. 2011. Eta squared and partial eta squared as measures of effect size in educational research. *Educational research review.* 2011, Vol. 6.2, 135-147.

Salazar, Jose J., Garland, Lean et Ochoa , Jesus. 2022. Fair train-test split in machine learning: Mitigating spatial autocorrelation for improved prediction accuracy. *Journal of Petroleum Science and Engineering.* 2022, Vol. 209, 109885.

Schapire, Robert E. 2003. The boosting approach to machine learning: An overview. *Nonlinear estimation and classification.* 2003, 149-171.

Shai, Ben-David, Von Luxburg, Ulrike et Pál, Dávid. 2006. A sober look at clustering stability. *Learning Theory: 19th Annual Conference on Learning Theory, COLT 2006, Pittsburgh, PA, USA, June 22-25, 2006.* Springer Berlin Heidelberg, 2006, 19.

Shalev, Ben-David et Lev, Reyzin. 2014. Data stability in clustering: A closer look. *Theoretical Computer Science.* 2014, Vol. 558, 51-61.

Somnath, Chatterjee. 2015. *Centre for Central Banking Studies : Modelling credit risk.* s.l. : Bank of England, 2015.

Sonagara, Darshan et Badheka, Soham. 2014. Comparison of basic clustering algorithms. *Int. J. Comput. Sci. Mob. Comput.* 2014, Vol. 3.10, 58-61.

Sun, Shuyan, Wei, Pan et Lihshing, Leigh Wang. 2010. A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology.* 2010, Vol. 102.4, 989.

Tallarida, Ronald J. et Murray, Rodney B. 1987. Chi-square test. *Manual of pharmacologic calculations: With computer programs.* 1987, 140-142.

Vasicek, Oldrich Alfons. 1987. *Probability of loss on loan portfolio.* s.l. : KMV, 1987.

Vergnes, Jean. 1980. Détermination d'un pas optimum d'intégration pour la méthode de Simpson. *Mathematics and Computers in Simulation.* 1980, Vol. 22.3, 177-188.

Von Luxburg, Ulrike. 2010. Clustering stability: an overview. *Foundations and Trends in Machine Learning.* 2010, Vol. 2.3, 235-274.

Liste des figures

Liste des tableaux

Annexes

Sommaire

Annexe 1 : Compléments des statistiques descriptives

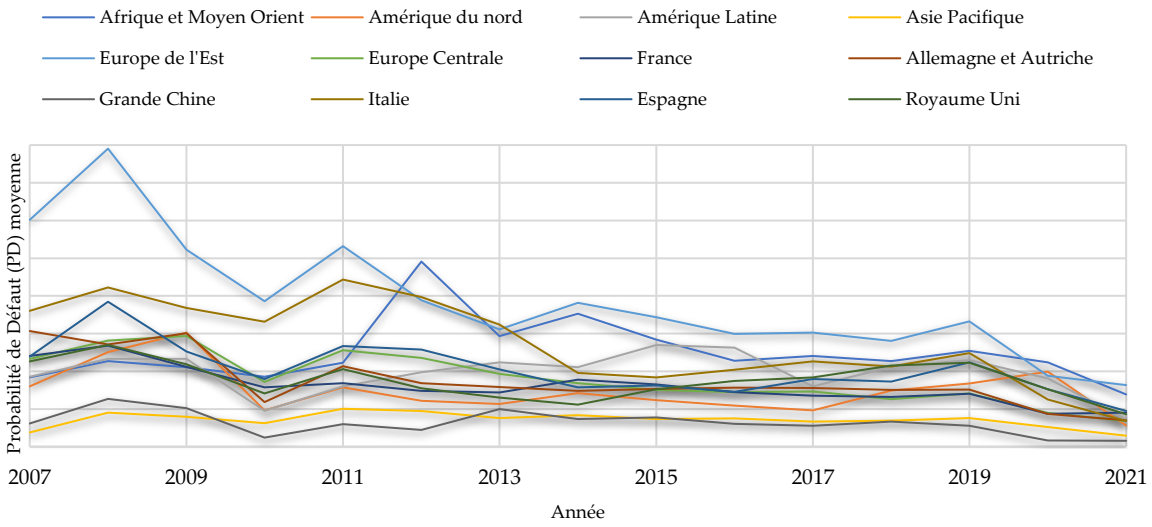


Figure 49 : Evolution de la probabilité de défaut moyenne par zone géographique des acheteurs

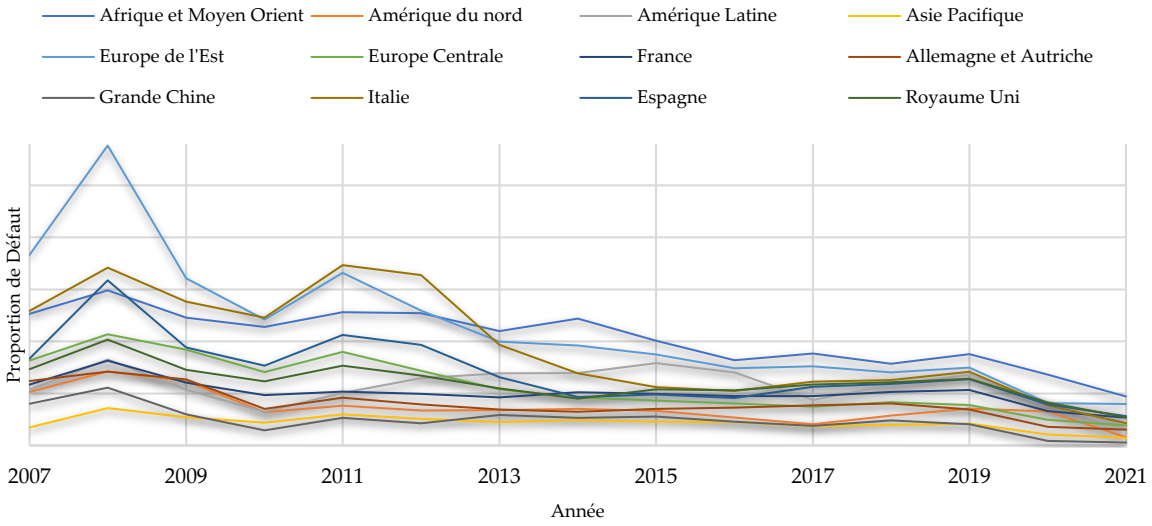


Figure 50 : Evolution de la proportion de défaut par zone géographique des acheteurs

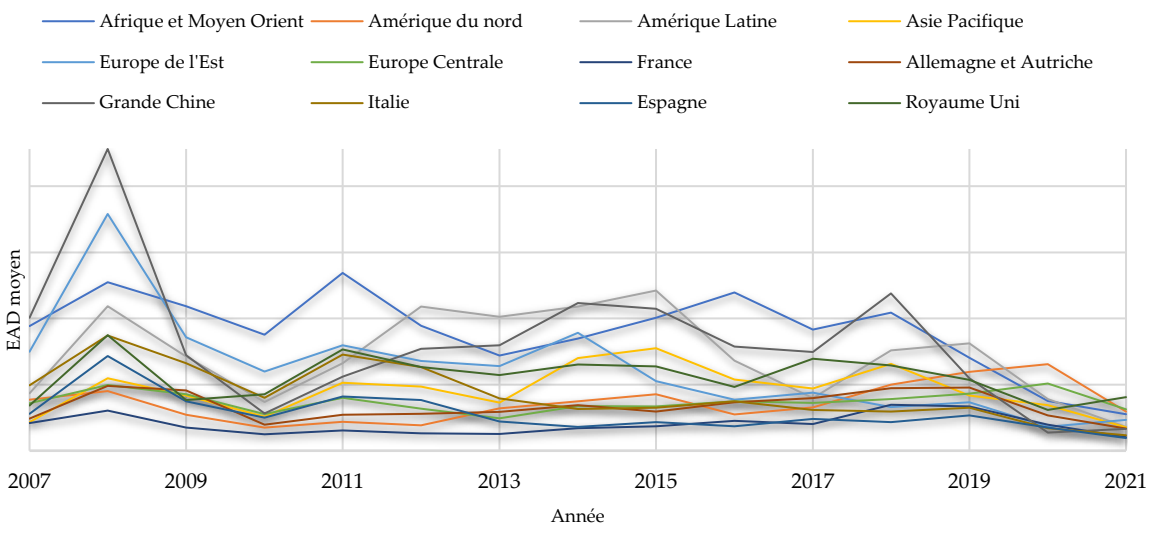


Figure 51 : Evolution de l'EAD moyen par zone géographique des acheteurs

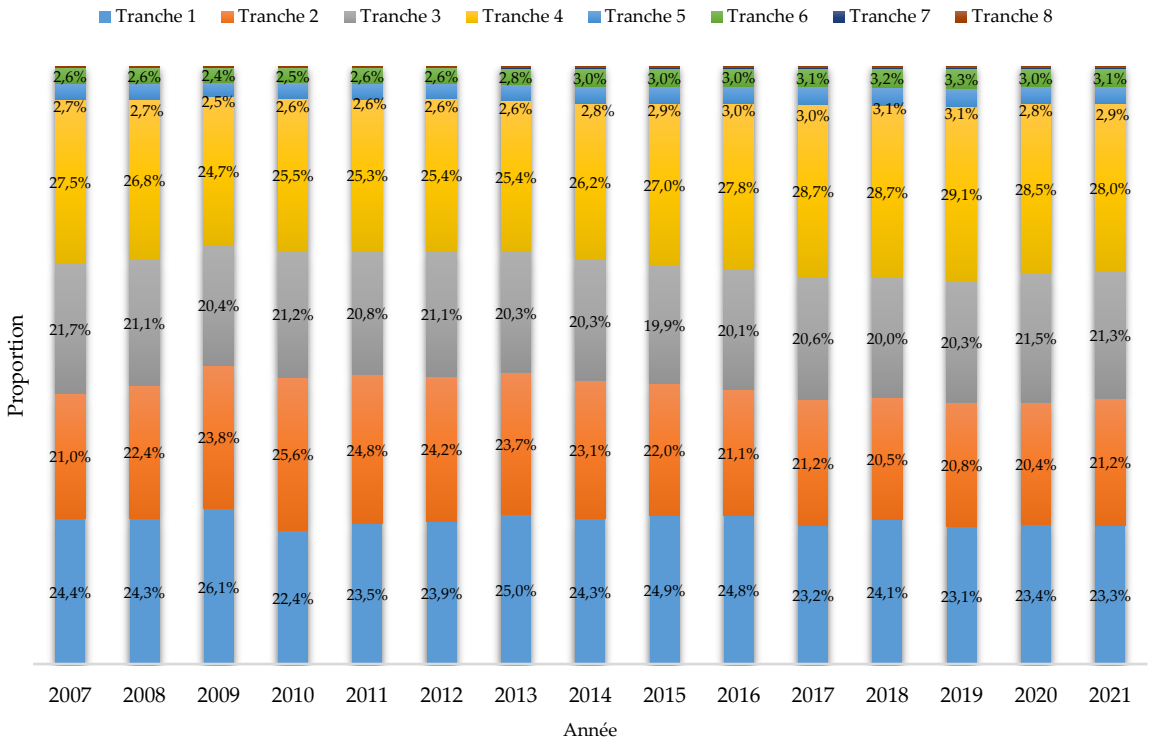


Figure 52 : Evolution de la répartition des acheteurs par tranche d'exposition

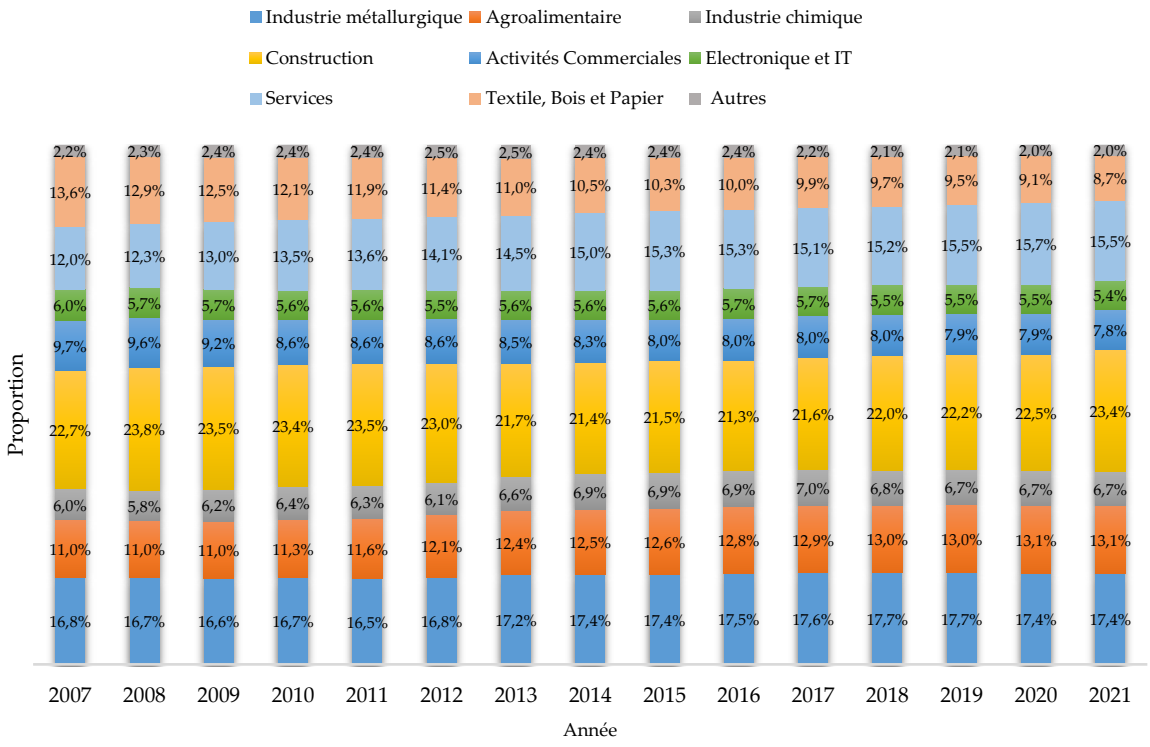


Figure 53 : Evolution de la répartition des acheteurs par secteur d'activité

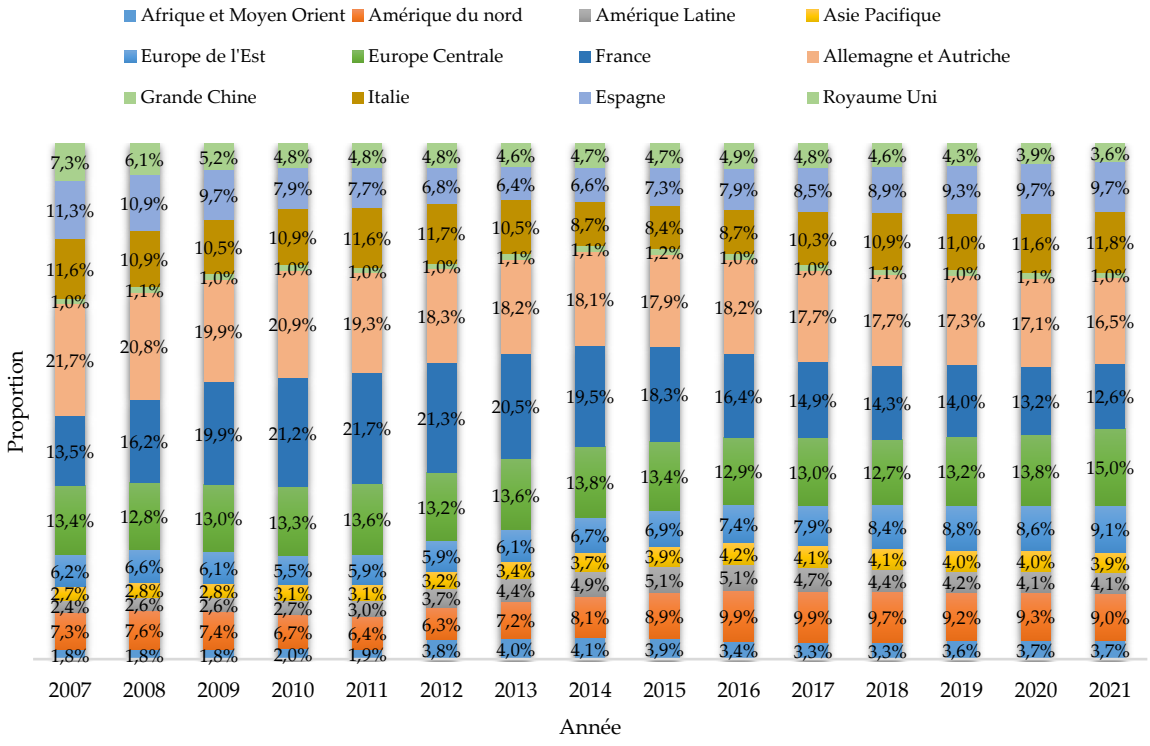


Figure 54 : Evolution de la répartition des acheteurs par zone géographique des acheteurs

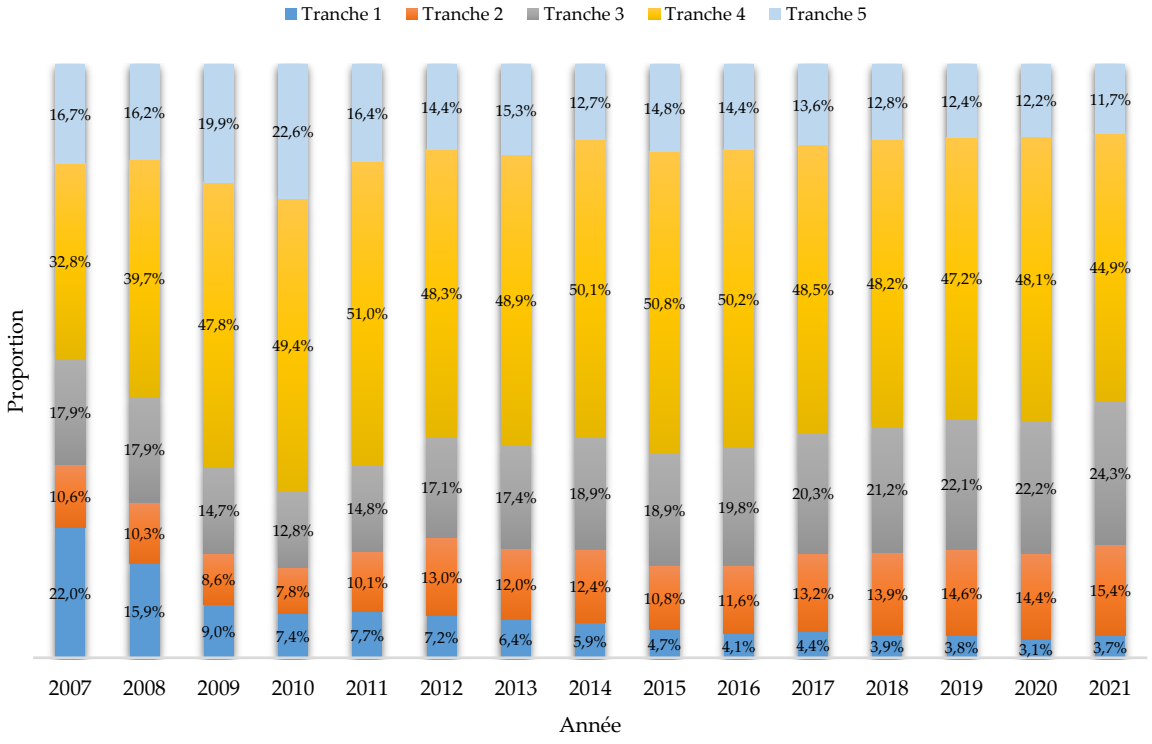


Figure 55 : Evolution de la répartition des acheteurs par tranche de rating

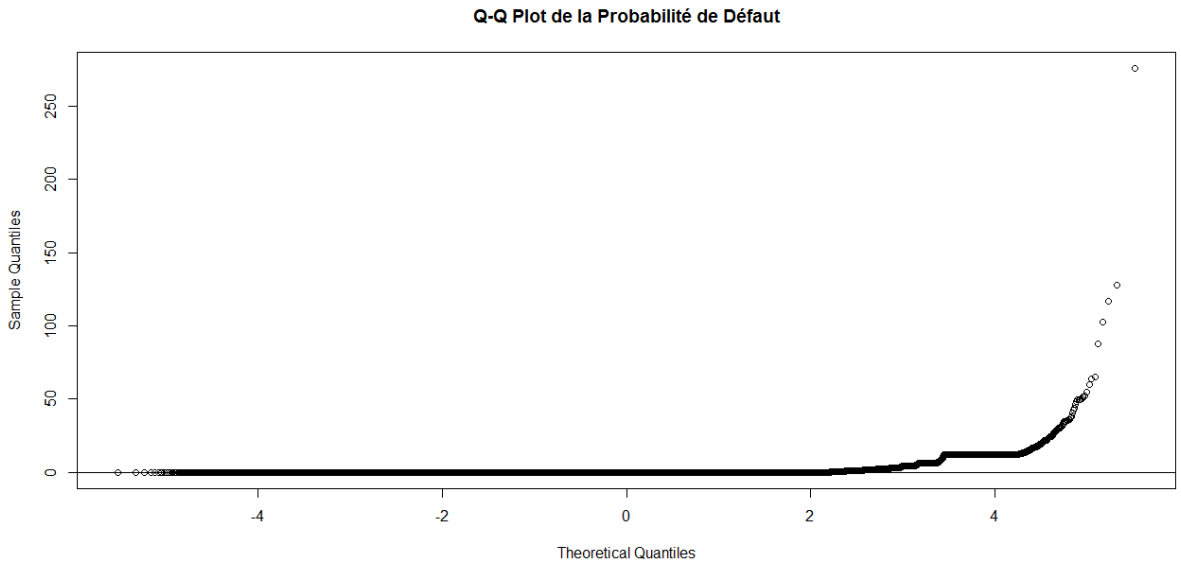


Figure 56 : QQ plot de la probabilité de défaut

Annexe 2 : Compléments de l'étude de la stabilité

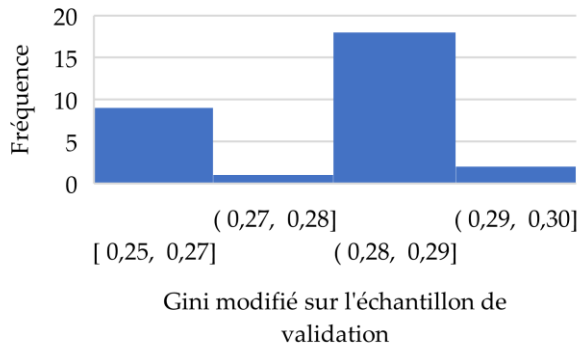


Figure 57 : Histogramme des valeurs de l'indice de Gini modifié sur l'échantillon de validation pour le modèle CART

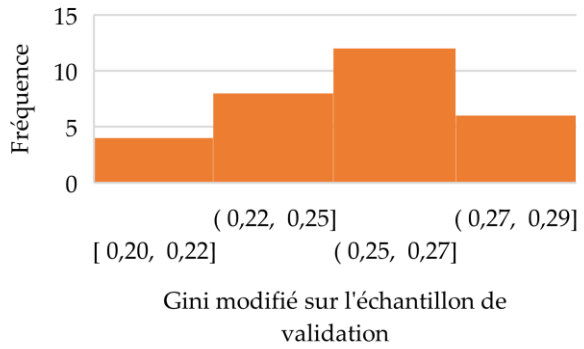


Figure 58 : Histogramme des valeurs de l'indice de Gini modifié sur l'échantillon de validation pour le modèle de k-prototypes

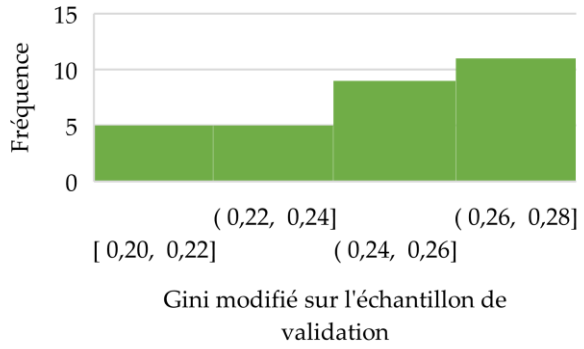


Figure 59 : Histogramme des valeurs de l'indice de Gini modifié sur l'échantillon de validation pour le modèle CAH

Annexe 3 : Compléments de l'analyse des sources d'instabilité

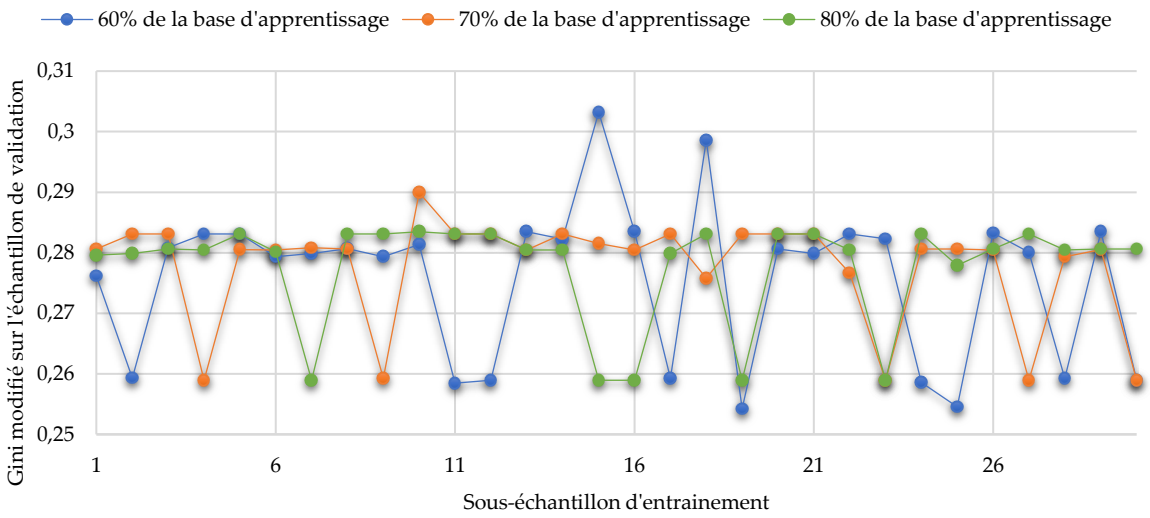


Figure 60 : Indice de Gini modifié par sous-échantillon et par taille d'échantillon

Proportion de la base d'apprentissage	Pentes (*m*)	Ordonnés à l'origine (*c*)	Coefficients d'instabilité
60%	-8,1751	0,0451	6,3413
70%	-0,4480	0,0212	1,2833
80%	-4,3849	0,0243	3,4071

Tableau 27 : Coefficients d'instabilité par taille de sous-échantillon

	Pentes (*m*)	Ordonnés à l'origine (*c*)	Coefficients d'instabilité
Scénario 0	-8,1751	0,0451	6,3413
Scénario 1	-8,1429	0,0451	6,3265
Scénario 2	1,9577	0,0109	1,5241
Scénario 3	1,9395	0,0107	1,5059

Tableau 28 : Coefficients d'instabilité par choix de variables

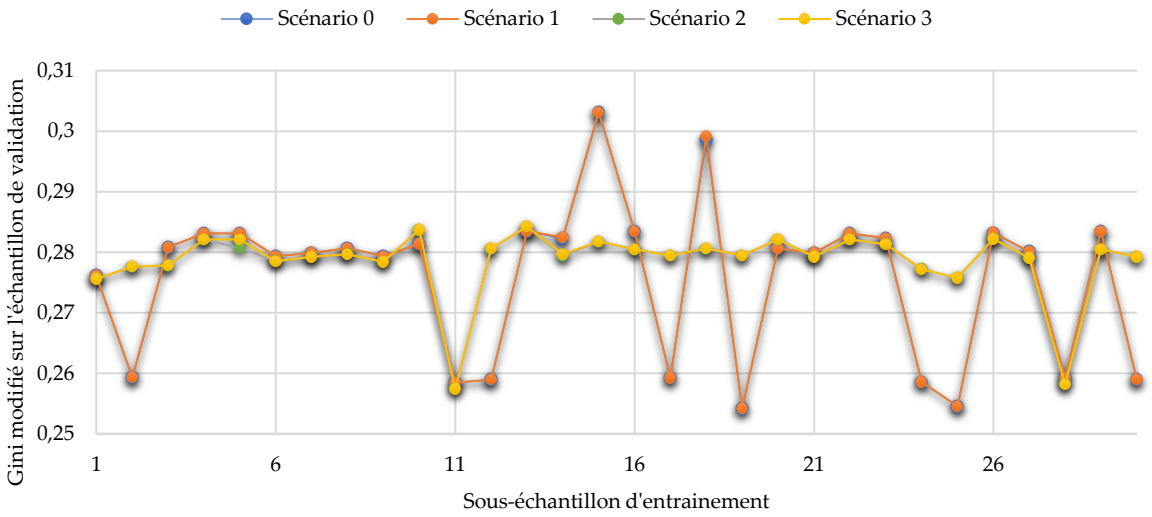


Figure 61 : Indice de Gini modifié par sous-échantillon et par choix de variables explicatives

	Pentes (*m*)	**Ordonnés à l'origine (*c*)**	**Coefficients d'instabilité**
Occurrence de défaut	-8,1754	0,0451	6,3413
Probabilité de défaut	-0,1503	0,0205	1,0984

Tableau 29 : Coefficients d'instabilité par variable cible

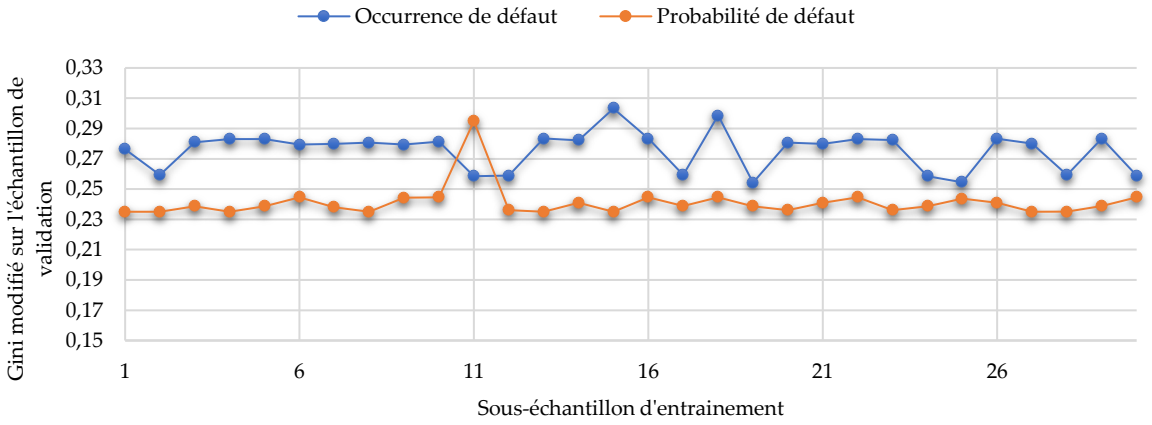


Figure 62 : Indice de Gini modifié par sous-échantillon et par variable cible