Rapport de stage de troisième année

Lorenzo Mathieu

30 août 2024

Table des matières

Table des figures

List of Algorithms

Liste des tableaux

Remerciements

Je tiens à exprimer ma profonde gratitude à plusieurs personnes qui ont contribué de manière significative à la réussite de ce stage, de ma scolarité et à l'aboutissement de ce mémoire.

Tout d'abord, je souhaite remercier chaleureusement mon maître de stage, Phillipe Clarenc, pour son aide précieuse dans l'apprentissage des indicateurs comptables et pour les conseils avisés qu'il m'a prodigués pour ma future carrière.

Je remercie également Maurice Billionaire pour ses conseils pertinents concernant la rédaction et la méthodologie, ainsi que pour le partage de sa passion pour les musiques créoles.

Un grand merci à Zinaïda Salibekyan-Rosain pour ses précieux conseils rédactionnels.

Je tiens à exprimer ma reconnaissance à Clément Guillo, dont l'aide sur les modèles avancés m'a permis d'avancer sereinement dans ce stage, bien que je ne sois pas directement son stagiaire. Son enseignement du taro a également enrichi mon expérience.

Merci à Fernando Zavala pour ses remarques toujours pertinentes sur mon travail et pour avoir partagé son bureau, rendant ainsi mes journées plus agréables.

Je souhaite remercier Aurélie Duchesne pour l'incroyable accompagnement dont elle a fait preuve tout au long de ma scolarité à l'ENSAI.

Mes sincères remerciements vont à Laurent Costa, mon tuteur, qui m'a accompagné avec beaucoup de bienveillance et de gentillesse.

Un grand merci à mon ami Gaëtan Carrère pour son aide précieuse dans la rédaction de ce mémoire et pour sa compagnie très appréciée durant ce stage.

Je remercie également ma cousine Maude Mathieu, qui m'a rejoint en fin de stage.

Un énorme merci à Laura Mitterand qui est mon amie depuis sa naissance et ce n'est pas rien.

Merci à mon ami Dylan Corazza pour m'avoir initié aux préceptes de la mode et du streetworkout, et pour m'avoir aidé à cultiver un esprit sain dans un corps sain.

Grazie ai miei amici italiani, Gabriella, Mauro e Francesco, che sono il mio incontro più bello dell'anno.

Un grand remerciement à mon ami Nassim Khelifi, avec qui j'ai découvert mon intérêt pour l'économie et grâce à qui j'ai fait de nombreuses belles rencontres (La DT je parle bien de vous).

Un remerciement particulier à mon ami Corentin Duval, qui a toujours été à mes côtés depuis le début de l'ENSAI, ainsi qu'à mon ami Carl Moreteau, qui m'accompagne depuis de nombreuses années et j'espère pour de nombreuses autres encore.

Enfin, je souhaite exprimer toute ma gratitude à ma famille : mes parents pour leur amour et leur soutien indéfectible depuis toujours, ma soeur pour me supporter depuis tant d'années, ainsi qu'à mes beaux-parents qui m'ont toujours traité comme leur propre enfant.

1 Introduction

Les départements et régions d'outre-mer (DROM) sont caractérisés par une part élevée des transferts publics dans le PIB. En effet, en 2019, cette part est de 26% pour la Guyane, 27% pour La Réunion et la Martinique et 29% pour la Guadeloupe contre 20% pour la France métropolitaine. Le poids élevé de ces transferts publics peut être assimilée à une rente administrative (Poirine 1993) qui pourrait contribuer au développement du secteur non marchand. En lien avec ce phénomène, le poids de l'emploi public dans l'emploi total est plus élevé dans les DROM qu'en France hexagonale. Cette part est de 34% aux Antilles, 45% en Guyane et 33% à La Réunion contre 22% en France métropolitaine. Ces transferts publics pourraient augmenter le pouvoir d'achat des habitants et contribuer à alimenter l'activité économique des entreprises dans ces territoires afin d'ajuster la faible demande à l'offre (Sudrie 2021). La politique budgétaire mise en place pour les DROM pourrait créer un dynamisme économique au sein de ces territoires. Les travaux existants montrent que la rentabilité financière des entreprises dromiennes, qui est un indicateur de leur performance, est supérieure à celles de la France métropolitaine du fait d'une capitalisation plus faible (Caupin et Savoye 2012, Dreyer et Savoye 2013). L'objectif de ce papier est de tester la validité de ce résultat en intégrant d'autres méthodes de mesure de la performance financière des entreprises qui sont en lien avec le taux d'intérêt apparent, le levier financier ou encore la capacité de remboursement. De plus, les travaux précédents considèrent les DROM dans leur globalité (Dreyer et Savoye 2013), alors que ce papier prendra en compte l'hétérogénéité du tissu productif au sein des DROM et aussi en France métropolitaine.

Le tissu productif dromien est différent de celui de la France métropolitaine. La taille des entreprises dromiennes en termes d'effectifs en Equivalent Temps Plein (EQTP) est en moyenne plus petite que celle des entreprises en France métropolitaine (4,0 salariés en France métropolitaine contre 2,0 en Guadeloupe, 1,8 en Martinique, 2,6 en Guyane et 3,1 à La Réunion en 2019) (Annexe A, tableau 17). L'absence de GE et TGE se manifeste avec une taille moyenne des entreprises plus faible en DROM.

Concernant l'activité économique des différents territoires, Les contributions au chiffre d'affaires observé sont différentes. Alors que les TPE constituent en moyenne 85% du nombre total d'entreprises des territoires, elles ne génèrent que 16% du chiffre d'affaires total en France métropolitaine contre 36% en Guadeloupe, 26% à La Réunion, 30% en Martinique et 41% en Guyane (Annexe A, tableau 19). En France métropolitaine, les GE et TGE, bien que peu nombreuses, engendrent une part importante du chiffre d'affaires total avec respectivement 32% et 13% indiquant une concentration économique très forte. La concentration des entreprises en terme de chiffre d'affaires est plus forte en France métropolitaine que dans les DROM. L'indice de gini qui permet de mesurer la concentration du chiffre d'affaires est de 0,9 en France métropolitaine, bien plus que la Martinique qui est en seconde place avec 0,81 (Annexe A, tableau 20). Les diversité des activités est plus restreinte dans les DROM qu'en France métropolitaine : 338 à 487 activités différentes pour les DROM contre 615 pour la france métropolitaine (Annexe A, tableau 17). En classant les 100 départements français par ordre décroissant de leur niveau de diversification (1 très diversifié et 100 très peu diversifié), La Réunion est 35eme, la Guadeloupe est 45eme, la Martinique 62eme et la Guyane 91eme. Le manque d'activités relativement à la France métropolitaine est lié à la taille des territoires et à leur éloignement géographique.

L'appareil productif dromien est plus orienté vers les consommateurs avec une forte présence des entreprises BtoC (REF). En effet, dans les Droms la part des BtoC dans le chiffre d'affaires total varie de 32,4% en Guyane à 37,3% en Guadeloupe contre 18,0% en France hexagonale. Ces entreprises sont présentes dans le commerce de détail, l'hébergement et restauration, l'immobilier ou encore les services aux personnes. À l'inverse, les entreprises métropolitaines sont plus souvent en relation commerciale avec d'autres entreprises – BtoB (REF). En France métropolitaine, la part des BtoB dans le chiffre d'affaires total est de 82,0%. Ces entreprises sont plus particulièrement dans les secteurs d'industrie, commerce de gros, services aux entreprises et construction

Ces différences tant au niveau de la répartition par taille que du secteur d'activité sont en lien avec le caractère insulaire inhérent aux différents DROM, la taille limitée du marché, leur éloignement géographique, ainsi que la taille de la population qui handicapent la capacité de développement des entreprises. Dans quelle mesure ces caractéristiques spécifiques aux DROM pourraient réduire la performance financière des entreprises dromiennes ? Est ce que les entreprises dromiennes se distinguent réellement de celles des autres départements de la France métropolitaine en terme de performance financière ? Enfin, est ce que l'effet de levier financier serait un facteur favorable aux entreprises des DROM par rapport aux entreprises métropolitaines ?

La contribution de cette étude est d'enrichir la littérature empirique sur les performances des entreprises non seulement au niveau national mais également à une échelle territoriale. L'étude permettra d'établir un lien, s'il existe, entre la rentabilité économique et financière. Nous mobilisons les Fichiers Approché des Résultats d'Esane (FARE-ESANE) pour les quatre DROM et la France métropolitaine pour l'année 2019. La structure de l'article est la suivante : la section 2 fournit un aperçu de la littérature pertinente ; la section 3 décrit les données utilisées ; la section 4 décrit la méthodologie retenue ; la section 5 détaille les résultats descriptifs et empiriques et la section 6 présente les conclusions de l'étude.

2 Contextualisation du stage

Le stage s'inscrit dans le cadre d'une étude visant à comparer la performance financière des entreprises implantées dans les Départements et Régions d'Outre-mer (DROM) à celles opérant en France métropolitaine. Les entreprises situées dans les DROM sont confrontées à des contraintes spécifiques, qui diffèrent de celles auxquelles font face les entreprises en métropole. L'objectif principal du stage est de mettre à jour les travaux de (Dreyer et Savoye 2013), publiés dans la revue **Économie et Statistique**, en construisant une base de données retraçant la performance des entreprises sur la période 2013-2020, à partir des bases FARE-ESANE.

Ce projet implique l'analyse de plusieurs aspects cruciaux, tels que les similarités ou différences dans la distribution des performances des entreprises (rentabilité financière, taux de marge, gains de productivité), les modes de croissance des entreprises (évolution des effectifs, valeur ajoutée), les contraintes d'accessibilité et l'étroitesse des marchés (stocks, charges de personnel), ainsi que le partage de la valeur ajoutée. En complément, une analyse départementale sera effectuée à l'aide d'une méthode de classification ascendante hiérarchique, permettant de mettre en évidence les profils des départements en termes de performance des entreprises.

Le champ d'étude se concentre sur les entreprises en tant qu'unités légales, et les résultats du stage seront matérialisés sous deux formes : une étude publiée dans la série « Insee Analyses » et un article académique qui sera soumis à la revue « Économie et Statistique ».

Ce stage est réalisé au sein de la Direction interrégionale Antilles-Guyane de l'Insee, plus précisément au Service Études et Diffusion (SED), situé au 11 parc d'activités de Jabrun, à Baie-Mahault en Guadeloupe. Le cadre géographique et institutionnel de ce stage offre une opportunité unique d'explorer les spécificités économiques des DROM en lien avec les enjeux nationaux.

3 Revue de littérature

Peu d'articles de la littérature économique s'intéressent aux performances financières des entreprises dromiennes relativement à celles de la France métropolitaine. Les travaux existants (Caupin et Savoye 2012, Dreyer et Savoye 2013) n'abordent pas le syndrome hollandais causé par la rente administrative dont pourraient être victimes les DROM. Les références à notre disposition pour l'analyse de ce syndrôme sont des modèles théoriques au niveau macro-économique. Le terme du syndrome hollandais a été indroduit pour la première fois par le journal The Economist ("The Dutch Disease" 1977) afin de décrire le déclin du secteur manufacturier aux Pays-Bas après la découverte du grand gisement de gaz naturel de Groningue en 1959. Suite à cet article, le terme a été repris pour décrire des situations où l'expansion économique d'un secteur entraîne le déclin d'un autre, favorisant ainsi une hausse des importations comme dans le papier de Corden et Neary 1982 qui fait une analyse théorique de ce phénomène économique. Cette étude a été menée comme mentionné par son auteur sous différentes contraintes assez fortes comme le souci des grandeurs réelles et non nominales, le maintien de l'équilibre de la balance commerciale, l'absence de mobilité internationale des capitaux et le plein emploi continu. Concernant les économies insulaires, (Poirine 1993) se base sur un modèle macro-économique théorique afin d'expliquer l'existence de cette maladie au sein de ces territoires. Il suppose pour son étude que les petites économies n'ont pas d'impact sur les prix internationaux, que le marché du travail est ouvert, qu'il y a une mobilité parfaite du travail entre le secteur insulaire et l'extérieur, et l'existence de trois secteurs (public, privé, traditionnel). Il est important de noter que l'hypothèse de mobilité parfaite des travailleurs reste forte vis-à-vis du coût de la mobilité vers l'extérieur qui peut être élevé. En prenant en compte les différentes hypothèses, les résultats montrent qu'il existe un effet d'éviction de la rente administrative sur les activités productives en faveur du secteur tertiaire non marchand pour les économies insulaires.

Dans le cas spécifique des DROM, la rente administrative pourrait s'expliquer, comme mentionné dans le

papier de (De Miras 1988), par la politique de rattrapage économique des DROM dans les années 1960, qui avait pour but d'améliorer le niveau de vie moyen des Dromiens et de désamorcer les tensions coloniales. Cette mesure a progressivement amené les DROM vers des économies de transfert. Le papier de (Mehoumoud Issop 2016) montre d'une part que le syndrome hollandais provoque un déséquilibre macro-économique, mais d'autre part qu'il s'amenuise grâce aux transferts publics qui permettent des gains de productivité.

Concernant l'analyse financière des entreprises des différents territoires, des travaux comme celui de (Mathouraparsad et Decaluwé 2018) utilisant les Matrices de Comptabilité Sociales (MCS) ont été réalisé au niveau macro-économique. Au niveau micro-économique, les papiers les plus récents sont ceux de (Dreyer et Savoye 2013) et (Caupin et Savoye 2012), où les auteurs montrent que la rentabilité financière est plus faible en France métropolitaine que dans les DROM en raison d'un phénomène de sous-capitalisation. D'autres indicateurs sont également utilisés pour comparer les entreprises, tels que la rotation des stocks, qui est plus élevée dans les DROM en raison de l'étroitesse du marché, ou encore la valeur ajoutée par salarié, plus faible en raison d'une activité moins capitalistique et d'une possibilité d'économies d'échelle plus faible.

Notre étude cherche à compléter la littérature en proposant une analyse micro-économique de chaque DROM indépendamment les uns des autres. Les papiers mentionnés précédemment qui effectuent une analyse micro-économique considèrent les DROM comme un même territoire, puis effectuent une comparaison avec la France métropolitaine. Chaque territoire ayant des caractéristiques qui lui sont propres, tant au niveau du tissu productif que de la localisation, il est pertinent de les observer individuellement. Ce travail a également pour but de tester la validité des résultats de (Dreyer et Savoye 2013) en intégrant différentes approches de mesure de la performance financière des entreprises, en lien avec le taux d'intérêt apparent, le levier financier ou encore la capacité de remboursement.

4 Données et Méthodologie

4.1 Données

Les données mobilisées pour cette etude proviennent du Fichier Approché des Résultats d'Esane (FARE)(Brion 2011), qui fait appel aux données financières des entreprises, notamment celles permettant le calcul de la valeur ajoutée, du chiffre d'affaires, des rentabilités economiques et financières. En effet, c'est l'unique source qui permet d'avoir une vue d'ensemble de la santé financière des entreprises dromiennes et métropolitaine (Clarenc, Autran, Arnault). La force de cette source est de combiner des liasses fiscales avec d'autres sources, comme l'enquête sectorielle annuelle (ESA). Le champ d'Esane est celui des entreprises marchandes, à l'exception du secteur financier et des exploitations agricoles. Ce champ est défini à partir des codes de la nomenclature d'activité NAF. Le fichier FARE mobilisé correspond au millésime 2019. Au total, l'étude porte sur 1 057 039 entreprises, dont 1 032 377 situées en France métropolitaine, 6062 en Guadeloupe, 2509 en Guyane, 10 505 à La Réunion, et 55586 en Martinique (tableau 1). Le champ de l'étude correspond aux unités légales employeuses en EQTP. Ces dernières appartiennent toutes au régime fiscal normal ou simplifié des bénéfices industriels et commerciaux. Pour simplifier les unités légales seront assimilées aux "entreprise". La taille des entreprises est définie en fonction du nombre de salariés en EQTP, avec des intervalles spécifiques pour chaque type d'entreprise (inférieur à 10 TPE, entre 10 et 249 PME, entre 250 et 4999 GE, et TGE pour les autres). Un filtre sectoriel est appliqué, incluant les industries, le commerce, les services marchands et la construction, tout en écartant les sections Agriculture, Activités financières et d'assurance, Santé humaine et action sociale et P Enseignement selon la NAF. Les administrations publiques et les services non marchands sont exclus de l'étude. De plus, seules les entreprises avec une acivité certaine sont retenues (CA au moins égal à 1000€).

Table 1 – Nombre d'entreprises et chiffre d'affaires par zone

Zone	Nombre d'entreprises	Chiffre d'affaires (en milliers d'euros)
France	1 032 377	3 747 909
Guadeloupe	6 062	10 633
Guyane	2 509	4 138
La Réunion	10 505	22 278
Martinique	5 586	11 883
Total	1 057 039	3 796 842

Note de lecture : En 2019 la Guadeloupe produit 10 633 099€ de chiffre d'affaires avec 6 062 entreprises sur son territoire.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude
Source : calculs des auteurs à partir de la base de données ESANE.(Brion 2011)

4.2 Méthodes

Dans un premier temps, afin de mettre en évidence les similarités et les différences de performances financières des départements de la France métropolitaine avec les DROM, la méthode d'analyse en composante principale (Pearson 1901) suivi d'une classification sera mise en oeuvre . Pour cela, les indicateurs financiers (Annexe C, **??**) sont calculés en agrégeant au niveau départemental les données comptables présentent dans la base ESANE.

L'ACP est particlièrement adapté pour le traitement des variables quantitatives continues. En effet elle permet de réduire le nombre de dimensions tout en conservant un maximum d'information (de variance), ce qui permet de visualiser des données complexe afin de les analyser. De plus, elle permet de mettre en évidence les tendances linéaires qui seraient moins évidentes à comprendre sur les données brutes. Une fois les principales composantes identifiées grâce à l'ACP, la classification ascendante hierarchique (CAH) (Székely et Rizzo 2005) permettra de regrouper les départements dans différents clusters selon leurs caractéristiques financières. La limite de ces méthodes est qu'elles supposent que les relations entre les variables et les départements soient linéaires. L'hypothèse de linéarité étant très forte, une deuxième méthode sera mise en place : un algorithme d'apprentissage non supervisé t-SNE (Maaten et Hinton 2008). Cette méthode présente l'avantage d'établir les clusters en prenant en compte la non linéarité. La limite de cette méthode est la variabilité des résultats en fonction des paramètres choisis tels que le taux d'apprentissage et le nombre de dimensions de l'espace de sortie.

4.3 Stratégie empirique

Dans la deuxième partie de l'étude, un modèle de régression linéaire sera mis en place pour commencer afin d'étudier si les liens entre de la rentabilité économique et financière existe et sont les même entre les entreprises dans les DROM et la France métropolitaine. Cette analyse sera réalisé au niveau des entreprises.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_j X_j + \epsilon \tag{1}$$

où :
— Y est la rentabilité financière, (comment le mesurer)
— X_1 est la rentabilité économique (EBE/(immobilisation corp incorp BFR))
— X_j est le vecteur des caractéristique. représente le vecteur des caractéristiques moyennes des entreprises (taille, secteur d'activités),
— ϵ correspond au résidu

Pour confirmer ou non les résultats obtenus, deux modèles non paramétriques seront utilisés afin de s'affranchir des hypothèses plus strictes de la regression linéaire telle l'homoscédasticité des résidus ou encore pour capturer des relations non linéaires qui pourraient être présente. Le premier sera le modèle de Random Forest (Ho 1995, Breiman 2000) et le second XGBoost (GitHub 2022). La regression linéaire aura tout de même l'avantage d'avoir des coefficients interprétables afin de pouvoir estimer, s'il existe, la puissance de la relation entre la rentabilité économique et financière des entreprises.

4.3.1 Le modèle Random Forest

La Random Forest, introduite par (Breiman 2000), est une version améliorée du bagging (voir algorithme en annexe D). Les "weak learners" utilisés dans ce modèle sont des arbres de décision, comme l'indique le nom du modèle. Contrairement au bagging classique, où chaque arbre est construit en utilisant l'ensemble complet des variables disponibles d, la Random Forest introduit une diversification supplémentaire en sélectionnant un sous-ensemble aléatoire de variables $mtry < d$ à chaque nœud pour chaque arbre. Ce processus de sélection aléatoire garantit une plus grande diversité parmi les arbres générés, ce qui réduit leur corrélation et améliore la robustesse du modèle. Le graphique 1 illustre ce mécanisme. Le nombre de variables sélectionnées est un

hyperparamètre à définir, qui peut être optimisé par la suite. Lors du processus d'optimisation, on peut choisir de maximiser la variance expliquée par le modèle ou de minimiser l'erreur quadratique moyenne (Root Mean Square Error, RMSE). De plus, la pureté des nœuds peut aussi être un indicateur de performance du modèle. Dans un arbre de décision de régression, un nœud est considéré comme pur si les valeurs qu'il contient ont une faible variance, en d'autres termes, si les observations classées dans un nœud sont proches les unes des autres.

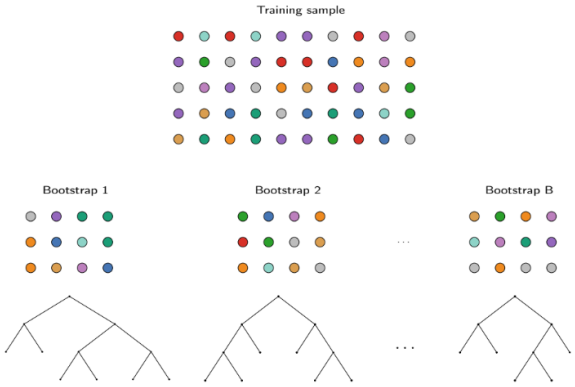
$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$\text{Pureté d'un noeud} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

où :

— n est le nombre d'observations dans le nœud,
— y_i est la valeur cible de l'observation i,
— $\bar{y}$ est la moyenne des valeurs cibles dans ce nœud.

FIGURE 1 – Exemple de fonctionnement de la Random Forest



Note de lecture : Les bootstrap utilise un échantillon du jeu de données d'entraînement afin d'effectuer un entraînement.
Champ : Aucun champ
Source : extrait de (VEIGA 2023–2024)

Les prédictions des nouveaux échantillons seront fait en utilisant la moyenne des prédictions de tout les arbres de régression individuels. L'algorithme du Bagging est disponible en annexe D.

$$\hat{f}(x') = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

où :

— B est le nombre total d'arbres,
— x' est le nouvel échantillon
Ci dessous le détail de l'algorithme :

Algorithm 1 Algorithme de Random Forest

Input: Ensemble d'entraînement $\{(x_i, y_i)\}_{i=1}^{N}$,
nombre d'arbres B,
nombre de variables $mtry$ sélectionnées à chaque nœud,
taux d'échantillonnage α
Output: Fonction de prédiction $\hat{f}(x)$

for $b = 1$ **to** B **do**

Échantillonnage bootstrap :
Échantillonner aléatoirement avec remplacement un sous-ensemble de l'ensemble d'entraînement $\{(x_i^{(b)}, y_i^{(b)})\}_{i=1}^{\alpha N}$.

Construction de l'arbre :
Pour chaque nœud de l'arbre, sélectionner aléatoirement un sous-ensemble de $mtry$ variables parmi les d variables disponibles.

$$\text{Choisir } mtry \text{ variables } \{x_{j_1}, x_{j_2}, \ldots, x_{j_{mtry}}\} \subseteq \{x_1, x_2, \ldots, x_d\}.$$

À chaque nœud, déterminer le meilleur critère de division parmi les $mtry$ variables sélectionnées :

$$\theta = \arg\min_{\theta} \sum_{i} \mathbb{I}(x_i \leq \theta) L(y_i, x_i),$$

où $L(y_i, x_i)$ est la fonction de perte.

Construire un arbre de décision $f_b(x)$ en utilisant les observations échantillonnées et les variables sélectionnées.

$$f_b(x) = \hat{y}_{\text{feuille}(x)},$$

L'arbre est construit jusqu'à un certain critère d'arrêt (profondeur maximale, nombre minimal d'observations par feuille, etc).

end

Prédiction finale :
La prédiction pour un nouvel échantillon x' est obtenue en moyennant les prédictions de tous les arbres construits :

$$\hat{f}(x') = \frac{1}{B} \sum_{b=1}^{B} f_b(x').$$

4.3.2 Le modèle XGBoost

Le modèle XGBoost est l'un des algorithmes de boosting les plus populaires. Comme cela a été évoqué précédemment, dans le bagging, les "weak learners" apprennent de manière indépendante les uns des autres, souvent en utilisant des échantillons différents de données d'entraînement. Dans le boosting, le principe reste d'agréger les "weak learners", mais cette fois-ci en les construisant de manière séquentielle. Chaque nouvel arbre est entraîné pour corriger les erreurs commises par les arbres précédents, rendant ainsi le modèle global plus performant.

De plus, l'algorithme XGBoost utilise une descente de gradient améliorée, non seulement en optimisant sur la base des dérivées de premier ordre (comme le gradient classique), mais également en intégrant les dérivées de second ordre pour une optimisation plus précise. XGBoost applique également une régularisation supplémentaire sur les arbres pour éviter le surapprentissage, en pénalisant les modèles trop complexes via des termes de régularisation L_1 et L_2.

Un autre aspect clé de XGBoost est l'utilisation d'un taux d'apprentissage (α), qui contrôle la contribution de chaque arbre au modèle final. Cela permet de réduire la vitesse d'apprentissage et d'améliorer la robustesse du modèle. Enfin, XGBoost est conçu pour tirer pleinement parti de la parallélisation, ce qui le rend particulièrement rapide et adapté pour travailler sur de grands ensembles de données. L'algorithme ci-dessous permet de détailler le fonctionnement du XGBoost.

Algorithm 2 Algorithme de Boosting avec Gradient

Input: Ensemble d'entraînement $\{(x_i, y_i)\}_{i=1}^N$,
fonction de perte différentiable $L(y, F(x))$,
nombre de "weak learners" M,
taux d'apprentissage α
Output: Fonction de prédiction $\hat{f}(x)$

Initialiser le modèle avec une valeur constante :

$$\hat{f}_{(0)}(x) = \arg\min_{\theta} \sum_{i=1}^{N} L(y_i, \theta).$$

for $m = 1$ **to** M **do**

Calculer les gradients et les fonctions hessiennes :

$$\hat{g}_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\hat{f}_{(m-1)}(x)}.$$

$$\hat{h}_m(x_i) = \left[\frac{\partial^2 L(y_i, f(x_i))}{\partial f(x_i)^2}\right]_{f(x)=\hat{f}_{(m-1)}(x)}.$$

Ajuster un "weak learner" (par exemple un arbre) en utilisant l'ensemble d'entraînement $\left\{x_i, \frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)}\right\}_{i=1}^N$ en résolvant le problème d'optimisation suivant :

$$\hat{\phi}_m = \arg\min_{\phi \in \mathbf{\Phi}} \sum_{i=1}^{N} \frac{1}{2} \hat{h}_m(x_i) \left[\phi(x_i) - \frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)}\right]^2.$$

Mettre à jour le modèle :

$$\hat{f}_m(x) = \alpha \hat{\phi}_m(x).$$
$$\hat{f}_{(m)}(x) = \hat{f}_{(m-1)}(x) + \hat{f}_m(x).$$

end

Sortie :

$$\hat{f}(x) = \hat{f}_{(M)}(x) = \sum_{m=0}^{M} \hat{f}_m(x).$$

5 Gestion des données entreprises

ESANE est une base de données complexe contenant des indicateurs comptables qui peuvent varier considérablement et inclure des valeurs extrêmes. En effet, il existe des cas particuliers d'entreprises, telles que celles en liquidation ou rencontrant des problèmes financiers, qui pourraient présenter des indicateurs atypiques. Par exemple, le taux d'intérêt apparent, qui représente les intérêts payés divisés par les dettes, doit être à 0 si l'ensemble des dettes sont remboursées. Cependant, un problème survient lors du calcul de ce ratio lorsque les intérêts payés sont positifs et les dettes sont nulles ou proches de 0. Cette situation, où un flux positif est divisé par 0, entraîne une valeur extrêmement grande et peut créer des outliers ou des valeurs infinies. Afin d'obtenir des résultats optimaux, il est essentiel de nettoyer la base de données avant d'effectuer l'ACP ou tout autre modèle. Cela permet d'éviter que les valeurs extrêmes n'influencent les résultats et d'assurer la fiabilité des analyses. Pour ce faire, une analyse des valeurs manquantes et des outliers sera réalisée.

5.1 Analyse et traitement des valeurs manquantes

Les valeurs manquantes (NA) sont l'ensemble des valeurs qui sont absentes au sein de la base de données. Au sein de cette analyse, les valeurs infini (inf) seront aussi analysées. Avant toute chose, il est essentiel de

connaître le nombre de valeurs manquantes. Pour traiter les valeurs manquantes, (Makarov et Namiot 2023) proposent plusieurs méthodes : la suppression des lignes contenant des NA, le remplissage avec des valeurs comme la moyenne ou la médiane, et l'utilisation d'algorithmes de machine learning comme la régression linéaire ou le k-plus proches voisins (k-NN) pour imputer les valeurs manquantes. Le tableau 2 donne le résumé des valeurs manquantes et infinies par variable. Il y a au total 199 valeurs manquantes et 1470 valeurs infinies, soit un total de 1669 valeurs non utilisables. La base de données contient au total plus d'un million d'observations. Le nombre de valeurs manquantes étant très faible, elles seront supprimées

Table 2 – Valeurs manquantes et infinies de la base de données

Variable	NA	Inf	Total
siren	0	0	0
departement	0	0	0
ratio charges de personnel sur valeur ajoutée	2	2	4
taux de stocks	35	0	35
taux de BFR et autres actifs	35	0	35
disponibilites	35	0	35
taux d'immobilisation	35	0	35
taux de BFR passif	9	27	36
taux de valeur ajoutée	0	0	0
taux de marge brute	0	0	0
taux de resultat courant	0	0	0
rentabilité économique	1	1319	1320
rentabilité financiere	1	12	13
ratio chiffre d'affaires sur capitaux propres	0	13	13
autonomie financiere	8	28	36
taux d'endettement	0	13	13
taux de prélevement financier	8	28	36
intenstité capitalistique	28	8	36
solde commercial	2	0	2
Total	199	1470	1669

Note de lecture : La variable de la rentabilité économique contient 1 valeur manquante et 1319 valeurs infinies, soit un total de 1320 valeurs non traitables.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude
Source : calculs des auteurs à partir de la base de données ESANE.

5.2 Analyse et traitement des valeurs aberrantes

Les valeurs aberrantes peuvent affecter significativement des modèles comme la régression et l'analyse en composantes principales (ACP) car elles peuvent distordre les résultats. Dans la régression, les valeurs aberrantes peuvent entraîner des estimations biaisées des coefficients, rendant le modèle moins fiable. Pour l'ACP, elles peuvent fausser les directions des composantes principales, ce qui impacte l'interprétation des données et la réduction de dimension. C'est pourquoi il est essentiel de porter une attention particulière au traitement de celles-ci. (Makarov et Namiot 2023) et (Kremp 1993) proposent différentes méthodes, chacune ayant des avantages et inconvénients. Le tableau 3 montre les résultats du test d'Anderson-Darling (Anderson et Darling 1952) qui permet de mesurer la normalité d'une distribution. Aucune variable ne suivant une distribution normale (l'ensemble des p-values sont inférieures à 0,05), les méthodes reposant sur l'hypothèse de normalité telles que le z-score ou la règle des 3 sigmas sont écartées. Les méthodes retenues sont la mesure de l'ecart inter-quartile, la winsorisation et l'isolation forest.

Table 3 – Résultats des tests de normalité (Anderson-Darling)

Variable	Statistique AD	p value
ratio charges de personnel sur valeur ajoutée	397011.5	3.7×10^{-24}
aux de stocks	408139.2	3.7×10^{-24}
taux de BFR et autres actifs	320704.0	3.7×10^{-24}
disponibilites	331024.3	3.7×10^{-24}
taux d'immobilisation	340706.1	3.7×10^{-24}
taux de BFR passif	408038.7	3.7×10^{-24}
taux de valeur ajoutée	403530.6	3.7×10^{-24}
taux de marge brute	404208.0	3.7×10^{-24}
taux de resultat courant	404023.9	3.7×10^{-24}
rentabilité économique	408140.0	3.7×10^{-24}
rentabilité financière	408130.9	3.7×10^{-24}
ratio chiffre d'affaires sur capitaux propres	408129.7	3.7×10^{-24}
autonomie financiere	408037.8	3.7×10^{-24}
taux d'endettement	408113.3	3.7×10^{-24}
taux de prélevement financier	408139.5	3.7×10^{-24}
intenstité capitalistique	387703.9	3.7×10^{-24}
solde commercial	400293.1	3.7×10^{-24}

Note de lecture : Le résultat du test d'Anderson-Darling pour la variable "taux de stocks" est de 408 139, avec une p-value associée de 3.7×10^{-24}. Étant donné que la p-value est inférieure à 0,05, l'hypothèse nulle selon laquelle cette variable suit une distribution normale est rejetée.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude
Source : calculs des auteurs à partir de la base de données ESANE.

5.3 Choix de la méthode de nettoyage de la base de données

Après avoir testé différentes méthodes pour nettoyer la base de données ainsi que leurs avantages et limites, il est nécessaire de choisir laquelle conserver pour effectuer l'ACP. Afin d'effectuer ce choix, les tableaux 4 et 5 répertorient les statistiques des différents jeux de données et le nombre d'observations considérées comme outliers, comme cela a été fait précédemment pour l'isolation forest.

La méthode de l'écart inter-quartile, bien que produisant de bons résultats statistiques, comporte un biais trop important en raison du nombre excessif de données considérées comme aberrantes : 695 333 (pour un seuil de 1,5) et 464 528 (pour un seuil de 3) observations. La méthode de la winsorisation est intéressante car elle permet de conserver l'ensemble des données, maximisant ainsi les informations disponibles. Comme vu précédemment, le seuil de 2,5 est plus acceptable et produit des valeurs extrêmes plus interprétables, avec pour la rentabilité financière un maximum de 180 56 contre 1028 09 pour l'isolation forest. Cependant, la winsorisation pose le problème de déformer la structure des données, ce qui pourrait nuire aux modèles futurs tels que la régression linéaire, compromettant notamment l'hypothèse d'hétéroscédasticité des données.

Pour conclure, la méthode choisie sera celle de l'isolation forest. Cette méthode permet d'éliminer les outliers de manière plus raisonnable que l'écart inter-quartile, avec 202 211 observations considérées comme outliers (tableau 5, et ne déforme pas la structure des données comme le fait la winsorisation, ce qui pourrait être problématique pour les analyses futures.

Table 4 – Statistiques des différents jeux de données pour la rentabilité financière

méthode	moyenne	variance	P1	médiane	P99	maximum
Jeu de données initial	-4.28×10^{11}	3.64×10^{14}	-394.19	16.91	436.78	9.63×10^{16}
Winsorized 5	22.91	38.90	-61.36	16.91	106.28	106.28
Winsorized 2.5	22.57	52.94	-144.63	16.91	180.56	180.56
Ecart Inter-Quartile avec seuil de 1.5	20.40	22.83	-33.76	16.00	91.91	100.90
Ecart Inter-Quartile avec seuil de 3	21.68	30.21	-66.37	16.75	103.04	159.45
Isolation Forest	22.10	46.64	-124.23	16.78	158.65	1028.09

Note de lecture : L'application de la winsorisation pour repérer les outliers permet de réduire significativement les indicateurs statistiques de la base de données avec par exemple la moyenne qui passe de -4.28×10^{11} pour la rentabilité financière à 22,57 avec un seuil de 2,5.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude
Source : calculs des auteurs à partir de la base de données ESANE.

Table 5 – Nombre d'observations pour chaque méthode

Méthode	Observations retirées
Winsorized 5	0
Winsorized 2.5	0
Ecart Inter-Quartile avec seuil de 1.5	695 333
Ecart Inter-Quartile avec seuil de 3	464 528
Isolation Forest	202 211

Note de lecture : L'algorithme isolation forest a considéré 202 211 observations comme aberrantes avec un seuil sur le score d'anomalie à 0,53
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude
Source : calculs des auteurs à partir de la base de données ESANE.

Le tableau 6 présente le nombre d'unités légales conservées en DROM et dans la france métropolitaine. La proportion de données conservée pour les DROM est d'environ 75% soit un peu moins que pour la france métropolitaine qui a un total de 80%. Au total, 80,85% des unités légales initiales ont été conservées après le nettoyage des données. Cela suggère que malgré les disparités régionales, le processus de nettoyage a été efficace et a permis de conserver une majorité substantielle des données originelles.

Table 6 – Nombre d'unités légales dans le jeu de données nettoyé et originel

Département	Nombre d'unités légales conservées	Nombre d'unités légales	Pourcentage (%)
Guadeloupe	4590	6 062	75.73
Martinique	4053	5 586	72.55
Guyane	1880	2 509	74.93
La Réunion	7998	10 505	76.13
France métropolitaine	835827	1 032 377	80.95
Total	**854348**	**1 057 039**	**80.85**

Note de lecture : Le jeu de donnée nettoyé conserve 4590 observations pour la Guadeloupe sur les 6062 présentes initialement, soit 75,73% des observations de ce département.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude
Source : calculs des auteurs à partir de la base de données ESANE.

6 Une rentabilité financière plus élevé dans les DROM dû à une sous capitalisation des entreprises

L'article de Dreyer et Savoye 2013 montre que les DROM ont une rentabilité financière plus élevée, notamment en raison de la sous-capitalisation de ces entreprises. Cette conclusion est corroborée ici avec l'analyse des différents DROM de manière indépendante. En effet, la rentabilité financière est de 11,8 pour la Martinique, 12,1 pour la Guadeloupe, 12,6 pour La Réunion et 16,2 pour la Guyane, ce qui est bien plus élevé que la France métropolitaine, qui se situe à 10,8 (tableau 7). Les DROM présentent également une rentabilité économique notable, allant de 11,1 pour la Martinique à 15,3 pour la Guyane. (ici chercher pourquoi la Guyane a une renta éco et fi bien plus élevé montagne or et spatial)).

Table 7 – Moyennes de la rentabilité par département

Département	Rentabilité économique	Rentabilité financière
Guadeloupe	13.5	12.1
Martinique	11.1	11.8
Guyane	15.3	16.2
La Réunion	12.8	12.6
France métropolitaine	10.8	10.8

Note de lecture : La rentabilité économique de la Guadeloupe en 2018 est de 13,5%
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

Le tableau 27 présente la décomposition de la rentabilité financière, qui peut être exprimée comme le produit du ratio chiffre d'affaires sur capitaux propres et du ratio résultat sur chiffre d'affaires (le taux de profit). D'un côté, le taux de profit des entreprises d'outre-mer est plus élevé qu'en France métropolitaine, avec 3,2% pour la Martinique, 3,3% pour la Guadeloupe, 3,7% pour La Réunion et 4,9% pour la Guyane. Le ratio chiffre d'affaires sur capitaux propres est également plus élevé, avec 3,3% pour la Guyane, 3,4% pour La Réunion, et 3,7% pour la Guadeloupe et la Martinique, contre 3,3 pour la France métropolitaine. Ce ratio plus élevé est dû à la plus faible capitalisation des entreprises dromiennes par rapport aux entreprises métropolitaines. En effet, l'intensité capitalistique des DROM varie de 60,9 % pour la Guadeloupe à 73,3 % pour la Guyane, contre 82,3 % pour la France métropolitaine.

L'intensité capitalistique, qui mesure le rapport entre les immobilisations corporelles (comme les machines et les bâtiments) et le nombre d'employés, influence directement les capitaux propres. Les capitaux propres représentent les ressources financières apportées par les actionnaires et les bénéfices non distribués, servant à financer les actifs de l'entreprise. Une entreprise avec une forte intensité capitalistique investit beaucoup dans des actifs physiques, nécessitant souvent plus de financement, ce qui augmente les capitaux propres de l'entreprise. En revanche, une intensité capitalistique plus faible signifie moins d'investissement en actifs physiques par employé, nécessitant souvent moins de capitaux propres. Ainsi, plus une entreprise investit en actifs physiques par employé, plus elle a besoin de capitaux propres pour financer ces investissements.

Table 8 – Décomposition de la rentabilité financière

Département	Intensité capitalistique	Ratio CAHT / CP	Taux de profit
Guadeloupe	60.9	3.7	3.3
Martinique	70.9	3.7	3.2
Guyane	73.3	3.3	4.9
La Réunion	65.5	3.4	3.7
France métropolitaine	82.3	3.3	3.1

Note de lecture : Le taux de profit de la Guadeloupe en 2018 est de 3,3%
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

En conclusion, les entreprises des DROM ont une rentabilité financière plus élevée que les entreprises métropolitaines, confirmant ainsi les conclusions de l'article de (Dreyer et Savoye 2013). Maintenant, il s'agit de déterminer si les entreprises des DROM ont un comportement similaire à celles de la France hexagonale ou si elles se distinguent par des particularités spécifiques. Pour ce faire, un profilage des entreprises sera effectué en utilisant les variables présentées dans le tableau 23.

7 Des entreprises dromiennes similaires aux entreprises métropolitaines

7.1 L'analyse en composantes principales (ACP)

L'analyse en composantes principales (ACP) (Hotelling 1933,Pearson 1901) est utilisée pour obtenir une vue comparative des différents indicateurs comptables des entreprises des départements français, en tenant compte de leurs différentes dimensions. L'ACP identifie un nombre limité de facteurs ou de composantes principales qui expliquent la matrice de corrélation des variables considérées. Cette analyse est suivie d'une classification ascendante hiérarchique (CAH) (Székely et Rizzo 2005,Ward 1963). L'objectif de cette double approche est d'identifier les variables qui contribuent le plus aux différentes composantes principales, puis de regrouper les départements en clusters distincts.

L'ACP est une technique permettant de décrire les combinaisons linéaires entre différentes variables. Son principal intérêt réside dans sa capacité à réduire un grand nombre de dimensions en quelques composantes principales. Les combinaisons linéaires des composantes principales doivent expliquer une proportion élevée de la variance totale des données initiales.

Deux critères principaux sont ici utilisés pour décider du nombre de composantes principales à retenir. Le premier est le critère de Kaiser, qui consiste à retenir autant de facteurs que le nombre de valeurs propres supérieures à 1 dans la matrice de corrélation. (Hair Jr et al. 1998) indiquent que cette règle est efficace lorsqu'il y a entre 20 et 50 variables, mais elle a tendance à en retenir trop peu si le nombre de variables est inférieur à 20, et trop si le nombre de variables dépasse 50.(Stevens 2002) souligne que cette méthode tend à sélectionner trop de facteurs lorsque le nombre de variables est supérieur à 40. En revanche, elle est précise avec 10 à 30 variables. La selection de variable a mené à en garder 19, ce qui est idéal pour le critère de kaiser selon (Stevens 2002) mais qui pourrait être trop faible selon (Hair Jr et al. 1998). C'est pourquoi un deuxième critère utilisé est la méthode du coude, qui consiste à examiner un graphique des valeurs propres et à retenir le nombre de facteurs avant que la diminution des valeurs propres ne devienne moins prononcée, formant ainsi un "coude". Le nombre de composantes principales retenu pour cette ACP en tenant compte des deux critères précedement mentionnés (tableau 9, figure 2b) est de 5.

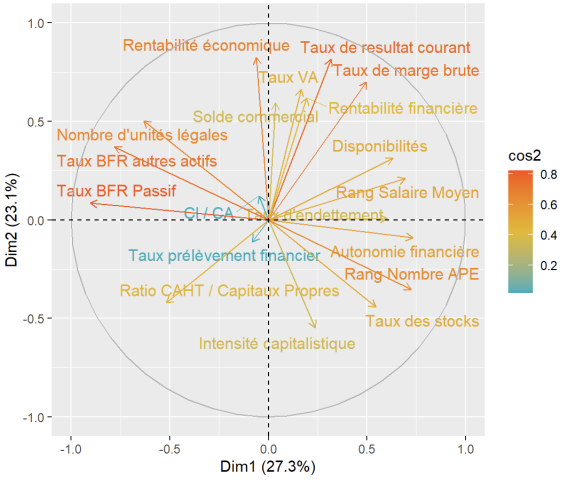
Table 9 – Valeurs propres et variance expliquée par chaque dimension

Dimension	Valeur propre	Variance (%)	Variance cumulée (%)
Dim.1	5.2	27.3	27.3
Dim.2	4.4	23.1	50.4
Dim.3	2.3	12.1	62.4
Dim.4	1.8	9.6	72.1
Dim.5	1.4	7.6	79.6
Dim.6	0.9	4.8	84.4
Dim.7	0.5	2.8	87.2
Dim.8	0.5	2.6	89.8
Dim.9	0.5	2.5	92.4
Dim.10	0.4	2.0	94.3
Dim.11	0.3	1.6	95.9
Dim.12	0.2	1.0	97.0
Dim.13	0.2	0.9	97.8
Dim.14	0.1	0.8	98.6
Dim.15	0.1	0.5	99.2
Dim.16	0.1	0.4	99.5
Dim.17	0.0	0.2	99.8
Dim.18	0.0	0.2	99.9
Dim.19	0.0	0.1	100.0

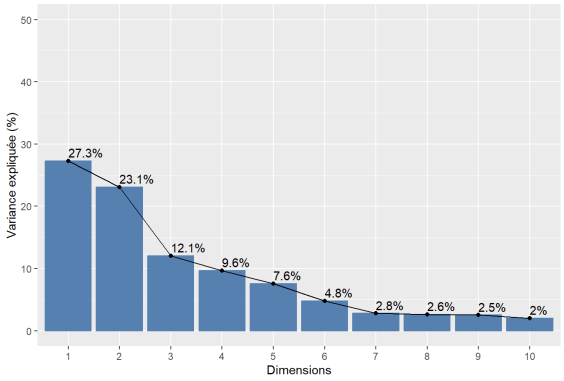
Note de lecture : La dimension 5 explique 7,6% de la variance totale et est la dernière qui possède une valeur propre supérieur à 1.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

Une propriété utile de l'ACP est que les composantes principales sont indépendantes les unes des autres, ce qui signifie qu'elles représentent des dimensions statistiques différentes du jeu de données initial. Toutefois, il est important de noter que l'ACP ne peut pas toujours réduire un grand nombre de variables à un petit nombre de variables transformées. Une réduction significative de la dimensionnalité ne peut être obtenue que lorsque les variables initiales sont fortement corrélées (positivement ou négativement). Enfin, plus la proportion de la variation des données expliquée par les deux premiers axes est élevée, meilleure est la représentation graphique des résultats.

FIGURE 2 – Analyse en composante principale sur les variables sélectionnées



(a) Corrélations des variables sur les deux premières dimensions de l'ACP



(b) Contribution de chaque dimension de l'ACP

Note de lecture : La figure 2a illustre les corrélations des variables sur les deux premières dimensions de l'ACP, expliquant respectivement 27,3% et 23,1% de la variance totale. La figure 2b montre la décroissance des contributions des différentes dimensions, indiquant que les deux premières dimensions capturent 27,3% et 23,1% de la variance totale des données.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

La figure 5 présente les résultats de l'ACP basé sur l'ensemble des données financières des départements avec les variables selectionnées comme mentionné précedemment. Le premier axe (vertical) explique 27,3% de la variance totale (de la matrice de corrélation) et le deuxième axe (horizontal), 23,1%. Le cercle de corrélation (2a) des variables sur les deux premières dimensions de l'ACP révèle que la première dimension est principalement influencée par des variables telles que le taux de BFR passif, le taux de BFR autres actifs, l'autonomie financière, le rang du nombre de salarié EQTP moyen, le rang du nombre d'activités et le nombre d'unités légales. Ces variables, ayant des coefficients de contribution élevés, représentent des aspects clés de la structure financière et de la taille des entreprises. Le graphique des contributions des variables pour la première dimension (voir annexe E, figure 22) confirme que les variables précédemment citées sont les plus déterminantes, expliquant une part significative de la variance totale.

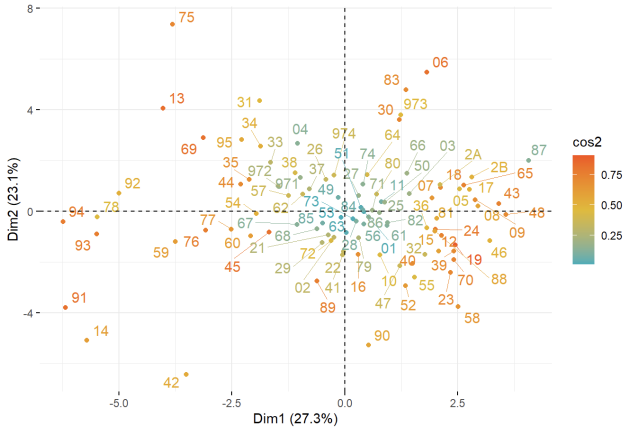
Pour la deuxième dimension, le cercle de corrélation montre que les variables ayant le plus grand impact sont la rentabilité économique, le taux de résultat courant, le taux de marge brute, le taux de VA, la rentabilité financière et le solde commercial. Ces variables sont des indicateurs de la performance économique et financière des entreprises. Le graphique des contributions des variables pour la deuxième dimension (voir annexe E, figure 23) montre que ces variables expliquent ensemble une proportion importante de la variance totale, mettant en évidence leur importance dans cette dimension.

En somme, la première dimension de l'ACP peut être interprétée comme représentant la structure financière et l'effet taille des entreprises, tandis que la deuxième dimension représente les performances économiques et financières des entreprises. Ces interprétations sont cohérentes avec les contributions des variables respectives aux deux premières dimensions.

Les différents départements sont représentés dans la figure 3 sur les deux premières dimensions de l'ACP. Les DROM se situent dans la partie supérieure du graphique, indiquant une meilleure performance économique et financière. Toutefois, la Guyane se distingue en étant positionnée sur la droite du graphique, tandis que les trois autres DROM se trouvent à gauche. Cette disposition montre que la Guyane est moins diversifiée, avec un nombre inférieur d'unités légales et de salariés, corroborant les résultats présentés précédemment (voir tableaux x et x). De plus, les DROM ne semblent pas montrer de specificités par rapport aux autres départements de la France métropolitaine en se placant à l'intérieur du nuage de points.

La prochaine étape consiste à examiner comment les différents DROM et les autres départements se regroupent au sein des clusters. La spécificité de la Guyane, en grande partie composée de territoires sauvages comparée aux autres DROM, explique cette divergence. Les analyses de clustering permettront de mieux appréhender ces variations et d'identifier des regroupements cohérents.

FIGURE 3 – Dimension 5

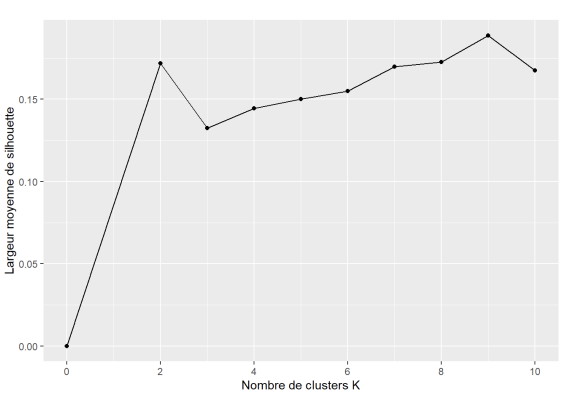


Note de lecture :
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

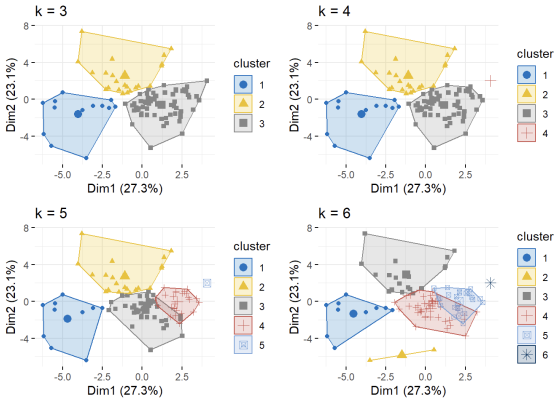
7.2 Regroupement des départements en fonction de leur caractéristiques communes

Une fois l'ACP effectuée, il est intéressant de réaliser une CAH afin de regrouper les départements en différents groupes. Cela permet d'avoir une vision d'ensemble des différents types de départements en fonction de leurs représentations sur les dimensions de l'ACP et de déterminer dans quel type de département se trouvent les différents DROM. Il est également pertinent de savoir si les DROM se retrouvent dans le même cluster ou non, étant donné que la Guyane possède une structure financière différente des autres DROM.

FIGURE 4 – Choix du nombre de cluster pour la CAH



(a) Largeur moyenne de silhouette

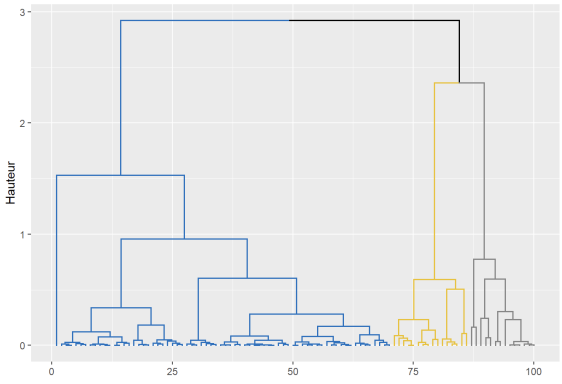


(b) Évolution de la statistique du gap

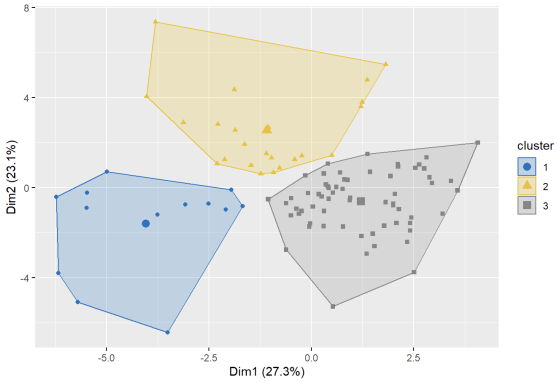
Note de lecture : La première figure présente la largeur moyenne de silhouette pour différents nombres de clusters k, indiquant également que 2 ou 3 clusters offrent une bonne séparation avec une largeur de silhouette maximale. La deuxième figure montre l'évolution de la statistique du gap en fonction du nombre de clusters k, suggérant que le nombre optimal de clusters se situe autour de 2 ou 3, là où la statistique du gap présente un saut significatif.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

Afin de déterminer le nombre optimal de clusters, l'indice silhouette (figure 4a) peut être utilisé, ainsi que les représentations graphiques (figure 4b). Ces représentations aident à identifier la valeur pour laquelle les résultats sont les meilleurs. Pour maximiser l'indice silhouette, il faudrait choisir un nombre élevé de clusters. Cependant, en observant la figure 4b, on constate qu'au-delà de 3 clusters, certains ne contiennent qu'une seule observation ou que les clusters se superposent. Pour permettre une analyse efficiente des différents clusters, la valeur choisie est donc de 3 clusters.

FIGURE 5 – Choix du nombre de cluster pour la CAH



(a) Dendrogramme de la classification ascendante hiérarchique



(b) Répartition des données en trois clusters

Note de lecture : La première figure présente le dendrogramme de la classification ascendante hiérarchique en trois clusters. La deuxième figure illustre la répartition des données en trois clusters distincts sur les sept premières dimensions de l'ACP.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

Les trois clusters sont représentés sur les deux premiers axes de l'ACP dans la figure 5b. D'après les résultats obtenus pour les deux premières dimensions, le cluster 1 correspondrait aux départements diversifiés, ayant de nombreuses unités légales et salariés, mais affichant une performance économique et financière plus faible. Le cluster 2 représenterait les départements ayant également de nombreuses unités légales et salariés, mais avec de meilleures performances économiques et financières. Le cluster 3 serait intermédiaire, regroupant principalement des départements avec des performances moins bonnes et comportant moins d'unités légales et de diversité que les autres clusters.
Le tableau 10 confirme les lectures graphiques précédemment faites. Le cluster 1 possède en moyenne 14 503 unités légales par département, contre 8 543 pour la moyenne globale. Il présente également un nombre moyen de salariés beaucoup plus élevé, se classant en moyenne 14ème sur 100. Le taux de BFR autres actifs et passifs y est plus élevé que la moyenne, avec des valeurs de 31,7 % et 45,4 % respectivement. En ce qui concerne les performances financières, elles sont inférieures à celles des autres clusters, avec une rentabilité financière de 9,6 %, une rentabilité économique de 11,6 % et des disponibilités de 8,8 %, contre une moyenne globale de 12,3 %, 12,6 % et 11,5 % respectivement.

Le cluster 2 possède en moyenne 15 543 unités légales par département, contre 8 543 pour la moyenne globale. Le taux de valeur ajoutée y est nettement supérieur, s'établissant à 30,8 % par rapport à la moyenne globale de 27,8 %. En termes de performance financière, le cluster 2 affiche une rentabilité financière de 13,9 %, surpassant la moyenne globale de 12,3 %, et un taux de résultat courant de 5,2 %, contre 4,5 % pour l'ensemble des clusters. Le classement par nombre d'activités place le cluster 2 en moyenne au 26ème rang. Enfin, l'intensité capitalistique du cluster 2 est de 61,0, ce qui est inférieur à la moyenne globale de 74,8. Ces indicateurs montrent que le cluster 2 est caractérisé par un nombre d'unités légales élevé et une intensité capitalistique plus faible, tout en affichant des performances économiques et financières supérieures à la moyenne.

Le cluster 3 possède en moyenne 4 817 unités légales par département, ce qui est nettement inférieur à la moyenne globale de 8 543 unités légales. En termes de classement par nombre d'activités, le cluster 3 se situe en moyenne au 64ème rang, bien au-delà de la moyenne globale, indiquant un niveau de diversification inférieur. La productivité du travail dans ce cluster est de 65,7, en dessous de la moyenne globale de 69,4. L'intensité capitalistique du cluster 3 est de 79,7, ce qui est supérieur à la moyenne globale de 74,8, indiquant une utilisation plus élevée de capital par rapport aux autres clusters. En ce qui concerne le taux d'endettement, le cluster 3 affiche un taux de 18,7 %, supérieur à la moyenne globale de 17,3 %. Enfin, le levier financier est également plus élevé dans le cluster 3, s'établissant à 42,3 contre 40,5 pour la moyenne globale. Ces indicateurs montrent que le cluster 3 est caractérisé par un nombre d'unités légales plus faible, une intensité

capitalistique et un levier financier plus élevés, ainsi qu'un taux d'endettement supérieur à la moyenne, mais avec une productivité du travail inférieure. La figure 6 présente les différents clusters selon les départements sur la carte de la France métropolitaine avec les DROM. Il est à noter que les DROM se rapproche des départements avec les plus grandes villes de France comme Paris (75), Lyon (69) ou encore Rennes (35). Ce constat suggère que les DROM pourraient avoir un dynamisme économique comparable, se traduisant par une plus grande diversité d'activités économiques et une présence accrue d'entreprises. Les indicateurs comptables similaires à ceux des grandes villes indiquent également un niveau de développement économique plus avancé dans les DROM, avec des entreprises présentant des structures financières et des performances économiques et financières meilleures.

Table 10 – Données moyennes par cluster

Variable	**1**	**2**	**3**	**Global**
Nombre d'unités légales	14503	15543	4817	8543
Charges personnel sur VA	0.8	0.7	0.7	0.7
CI/CA	0.4	0.4	0.4	0.4
Taux des stocks	9.1	9.8	12.0	11.1
Taux de BFR autres actifs	31.7	31.6	24.3	26.9
Disponibilités	8.8	11.7	12.0	11.5
Taux d'immobilisation	50.3	46.7	51.6	50.3
Taux de BFR passif	45.4	40.3	34.4	37.2
Taux de VA	24.5	30.8	27.3	27.8
Taux de marge brute	4.6	6.5	6.0	5.9
Taux de résultat courant	3.2	5.2	4.5	4.5
Rentabilité économique	11.6	14.9	12.0	12.6
Rentabilité financière	9.6	13.9	12.3	12.3
Ratio CAHT / Capitaux propres	4.5	3.5	3.6	3.7
Autonomie financière	36.9	41.3	44.9	43.0
Taux d'endettement	13.0	15.8	18.7	17.3
Levier financier	35.4	38.4	42.3	40.5
Taux de prélèvement financier	5.0	4.0	4.2	4.2
Taux d'intérêt apparent	2.8	2.3	2.1	2.2
Capacité de remboursement	46.5	50.8	43.9	45.9
Intensité capitalistique	75.2	61.0	79.7	74.8
Productivité du travail	76.3	76.0	65.7	69.4
Délais de paiement client	44.1	51.0	39.8	42.9
Délais de paiement fournisseur	55.5	56.6	46.4	49.9
Solde commercial	2.3	11.7	5.8	6.7
Rang nombre APE	22.6	26.4	64.8	50.5
Rang nombre de salarié moyen	14.2	54.2	56.5	50.5
% de la population	13.0	23.0	64.0	100.0

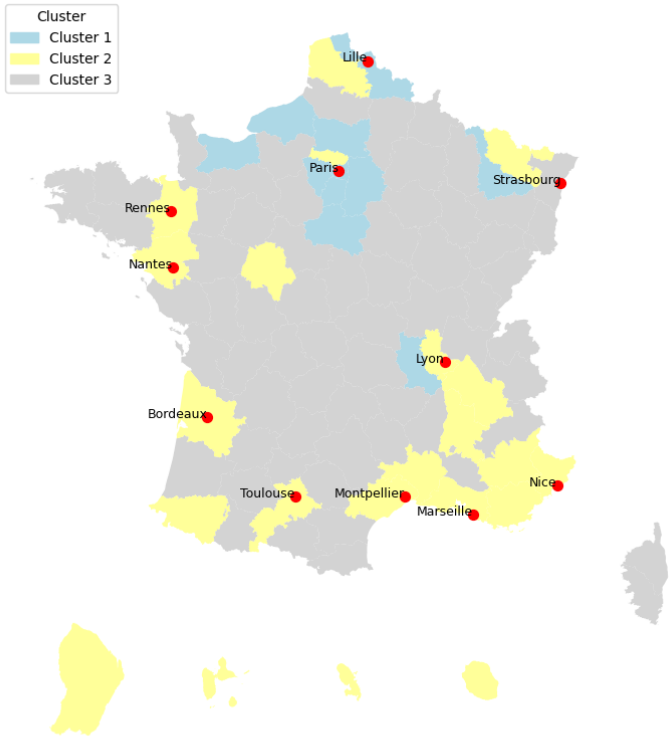
Note de lecture : Le cluster numéro 1 possède 14503 unités légales
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

Table 11 – Répartition des secteurs par cluster

Cluster	**Services Marchands (%)**	**Industrie (%)**	**Commerces (%)**	**Construction (%)**	**ME (%)**	**PME (%)**	**GE (%)**
1	47.1	9.2	24.1	19.6	83.2	16.1	0.7
2	49.6	9.1	24.6	16.6	84.4	15.2	0.4
3	38.8	13.4	27.8	20.0	83.7	16.0	0.3

Note de lecture : Le cluster numéro 1 possède 47,1% d'unités légales qui sont des services marchands.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

FIGURE 6 – Carte de la France avec les différents clusters



Note de lecture : Paris et son département se situe dans le cluster numéro 2.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

8 Une relation positive entre la rentabilité économique et financière

Dans les faits, une entreprise peut présenter une bonne rentabilité financière sans pour autant afficher une rentabilité économique satisfaisante. Cela peut s'expliquer par le fait qu'une entreprise peut bénéficier de conditions de financement favorables, comme des taux d'intérêt bas ou des délais de paiement avantageux, qui gonflent artificiellement sa rentabilité financière. Cependant, cette performance peut masquer une inefficience au niveau de l'exploitation de ses actifs, ce qui impacte négativement sa rentabilité économique. Inversement, une entreprise peut être très performante sur le plan économique, en optimisant l'utilisation de ses ressources et en générant des marges élevées, mais afficher une faible rentabilité financière en raison d'un endettement excessif ou de charges financières trop importantes.

C'est pourquoi il est intéressant de vérifier s'il existe un lien ou non entre la rentabilité économique et financière en utilisant différents modèles pour s'assurer de la robustesse des résultats. Dans un premier temps, un modèle de régression linéaire sera mis en place pour examiner les relations linéaires simples. Ensuite, un modèle de Random Forest sera appliqué pour explorer les interactions non linéaires et complexes entre les variables. Enfin, un modèle XGBoost sera utilisé pour affiner les prédictions et capter les subtilités des relations entre ces deux types de rentabilité.

8.1 La regression linéaire qui présente des premiers résultats, mais insuffisante

Pour commencer cette analyse, une régression linéaire a été utilisée afin de déterminer le lien entre la rentabilité financière et la rentabilité économique, ainsi que l'importance de ce lien.

Table 12 – Comparaison des coefficients entre la Métropole et les DROM

Variable	Coefficient Métropole	Importance Métropole	Coefficient DROM	Importance DROM
Intercept	1.044	significatif	1.038	significatif
Rentabilité économique	0.557	significatif	0.574	significatif
Commerces	0.005	non significatif	-0.082	significatif
Construction	-0.102	significatif	0.009	non significatif
Industrie	0.072	significatif	0.032	non significatif
Grande entreprise (GE)	0.054	significatif	0.219	non significatif
Petite et moyenne entreprise (PME)	0.047	significatif	0.040	non significatif
Très grande entreprise (TGE)	0.078	non significatif	.	non significatif
ul de 5 à 10ans	-0.006	non significatif	0.003	non significatif
ul de 10 à 15ans	-0.072	significatif	-0.019	non significatif
ul de moins de 5ans	0.291	significatif	0.277	significatif

Note de lecture : La rentabilité économique est une variable signigicative pour les DROM et possède un coefficient de 0,557
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

Le tableau des coefficients (tableau 12) compare les résultats pour la métropole et les DROM. Un lien positif est observé entre la rentabilité économique et la rentabilité financière dans les deux contextes, avec un coefficient de régression de 0,557 pour la métropole et de 0,574 pour les DROM. Ces coefficients, tous deux significatifs, suggèrent que la rentabilité économique est un bon indicateur de la rentabilité financière des entreprises dans les deux zones géographiques.

Dans le cadre de cette régression log-log, les coefficients peuvent être interprétés comme des élasticités, indiquant le pourcentage de variation de la rentabilité financière en réponse à une variation de 1% de la rentabilité économique. Par exemple, un coefficient de 0,557 en métropole signifie qu'une augmentation de 1% de la rentabilité économique entraînera une augmentation de 0,557% de la rentabilité financière, toutes choses égales par ailleurs. De manière similaire, pour les DROM, une augmentation de 1% de la rentabilité économique se traduit par une augmentation de 0,574% de la rentabilité financière. Toutefois, certaines variables montrent des impacts différents selon les régions. Par exemple, le coefficient associé aux grandes entreprises est significatif pour la métropole, mais non significatif pour les DROM, illustrant ainsi des variations dans les facteurs influençant la rentabilité en fonction des zones géographiques.

Table 13 – Comparaison des métriques entre la Métropole et les DROM

Métrique	DROM	Métropole
Erreur standard résiduelle	0.987	0.938
R-carré ajusté	0.344	0.378

Note de lecture : L'erreur standart résiduelle du modèle de regression dans les DROM est de 0,987 contre 0,938 dans le modèle en métropole.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

L'évaluation des performances des modèles repose sur le coefficient de détermination (R^2) et l'erreur standard résiduelle, qui mesurent respectivement la proportion de la variance expliquée par le modèle et la dispersion des résidus autour des valeurs prédites. Un R^2 plus élevé et une erreur standard résiduelle plus faible indiquent un meilleur ajustement du modèle. Le tableau (13) montre que les performances des modèles sont relativement similaires entre les deux régions, avec une erreur standard résiduelle de 0,987 pour les DROM et de 0,938 pour la métropole. Le R^2 est légèrement supérieur pour la métropole, atteignant 0,378 contre 0,344 pour les DROM. Cette différence pourrait s'expliquer par une plus grande quantité de données en métropole, offrant ainsi un meilleur ajustement du modèle.

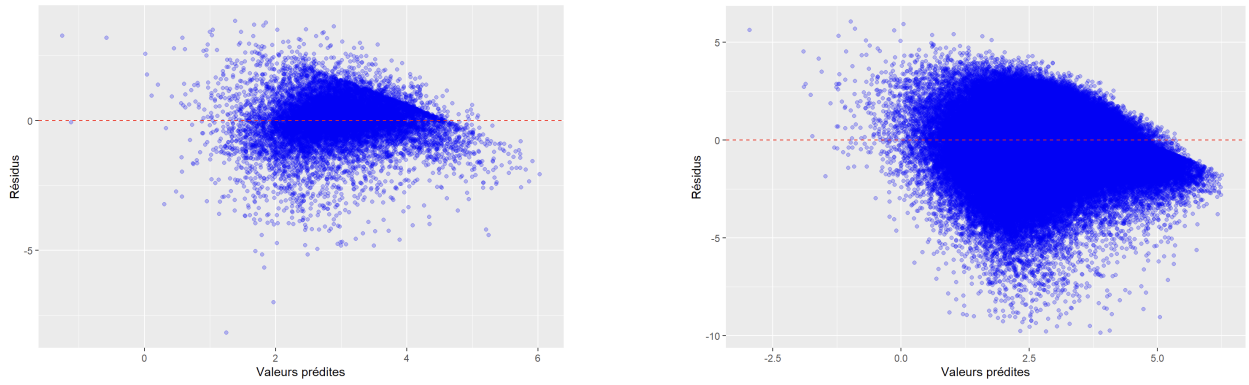
Table 14 – Comparaison des VIF entre la Métropole et les DROM

Variable	DROM	Métropole
Rentabilité économique	1.179	1.158
Commerces	1.248	1.215
Construction	1.182	1.176
Industrie	1.173	1.157
Grande entreprise (GE)	1.003	1.009
Petite et moyenne entreprise (PME)	1.079	1.093
Très grande entreprise (TGE)	.	1.000
Jeunes entreprises	1.356	1.322
Entreprises matures	1.252	1.234
Entreprises naissantes	1.548	1.485

Note de lecture : Le VIF de la rentabilité économique est de 1,179 ce qui signifie que cette variable ne pose pas de problème de multicolinéarité.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

En ce qui concerne les Variance Inflation Factors (VIF), les valeurs faibles observées indiquent l'absence de multicolinéarité dans les modèles, ce qui renforce la fiabilité des coefficients estimés comme présenté dans le tableau 14). Les VIF pour les DROM et la métropole sont tous inférieurs à 2, ce qui montre que la colinéarité entre les variables explicatives n'est pas préoccupante. Par exemple, la rentabilité économique affiche un VIF de 1,179 pour les DROM et de 1,158 pour la métropole, indiquant une faible corrélation avec les autres variables du modèle. Les secteurs tels que le commerce et la construction présentent également des VIF faibles, ce qui confirme la robustesse des estimations. De manière notable, les grandes entreprises montrent des VIF très bas, suggérant une indépendance quasi totale par rapport aux autres variables. Les VIF les plus élevés se retrouvent chez les entreprises naissantes, mais ces valeurs restent en dessous des seuils critiques généralement acceptés.

Figure 7 – Résidus de la regression linéaire



(a) Résidus de la regression linéaire pour les DROM

(b) Résidus de la regression linéaire pour la métropole

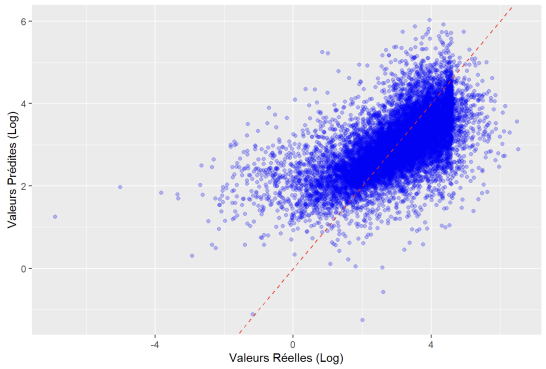
Note de lecture : Les résidus présentent une forme d'éventail, qui peut être un signe d'hétéroscédasticité.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

L'analyse des résidus du graphique 7 révèle plusieurs aspects importants. Tout d'abord, une forme en éventail est visible dans la dispersion des résidus, qui s'élargit à mesure que les valeurs prédites augmentent, indiquant ainsi la présence d'hétéroscédasticité. Ce phénomène peut affecter l'efficacité des estimations des

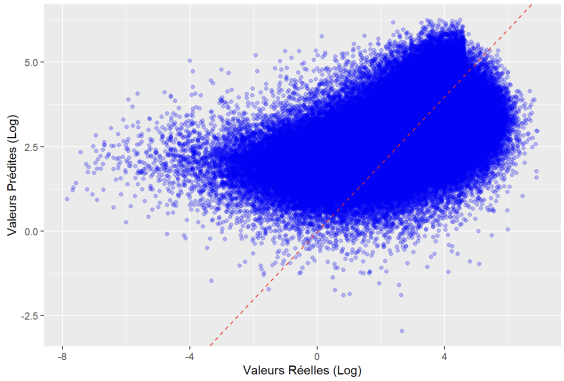
coefficients de régression et les tests statistiques associés, ce qui suggère qu'un ajustement du modèle pourrait être nécessaire. Néanmoins, les résidus sont globalement centrés autour de zéro, ce qui est un bon signe d'absence de biais systématique dans les prédictions. Cependant, quelques résidus extrêmes se distinguent nettement de la ligne zéro, indiquant des erreurs particulièrement élevées pour certaines observations, méritant une investigation plus approfondie.

Enfin, bien qu'il n'y ait pas de structure apparente dans les résidus en dehors de l'hétéroscédasticité, il pourrait être pertinent d'explorer d'autres types de modèles pour améliorer les prédictions. Des modèles d'ensemble tels que le Random Forest ou le XGBoost, qui reposent sur des techniques non paramétriques, sont particulièrement efficaces pour capturer des relations complexes et non linéaires entre les variables. Contrairement aux régressions linéaires, ces modèles ne nécessitent pas d'hypothèses strictes sur la distribution des résidus ou l'homoscédasticité des erreurs. Le Random Forest, par exemple, peut mieux gérer l'hétéroscédasticité en agrégant les résultats de multiples arbres de décision, tandis que le XGBoost permet de minimiser les erreurs résiduelles de manière plus fine, en ajustant les prédictions de manière itérative. Ces approches pourraient donc offrir des prédictions plus robustes et mieux adaptées aux données observées, réduisant ainsi les problèmes liés à l'hétéroscédasticité et aux résidus extrêmes.

FIGURE 8 – Graphique des valeurs prédites versus les valeurs réelles



(a) Graphique des valeurs prédites versus les valeurs réelles pour les DROM



(b) Graphique des valeurs prédites versus les valeurs réelles pour la métropole

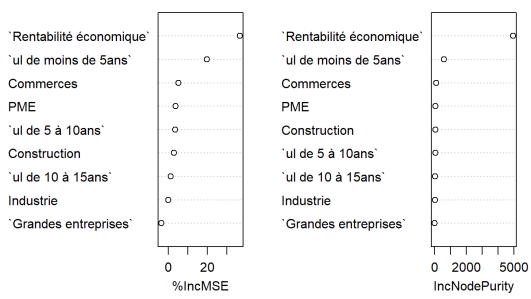
Note de lecture : Les regressions présentent des correspondances entre les valeurs prédites et réelles bien que la regression en métropole présente quelques dispersion sur les valeurs extrêmes.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

Les graphiques des valeurs prédites versus les valeurs réelles (figure 8) pour les DROM et la métropole révèlent des tendances similaires. Dans les deux cas, les points se concentrent autour de la diagonale, indiquant une correspondance raisonnable entre les valeurs prédites et réelles. Toutefois, une dispersion notable apparaît pour les valeurs extrêmes. Le modèle a tendance à sous-estimer les valeurs élevées et à montrer une grande variabilité des prédictions pour les petites valeurs, ce qui suggère des difficultés à capturer ces observations avec précision. Bien que le modèle soit relativement performant pour les valeurs moyennes, ces observations suggèrent qu'il pourrait être avantageux d'explorer des modèles plus complexes, tels que le Random Forest ou le XGBoost, pour améliorer la précision des prédictions, en particulier pour les valeurs extrêmes, et mieux gérer la variabilité observée dans les données.

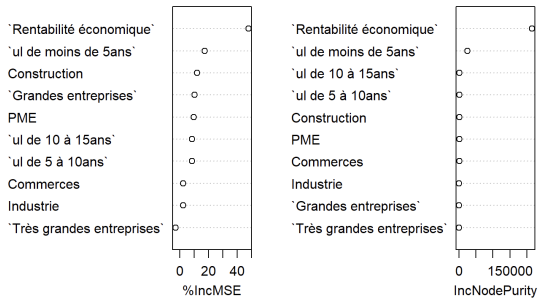
8.2 Une Random Forest qui corrobore les résultats de la régression linéaire

Pour vérifier les résultats de la régression linéaire, qui présente certaines faiblesses comme mentionné précédemment, un modèle de Random Forest a été mis en place afin de confirmer la justesse des résultats obtenus. La Random Forest à l'avantage de capturer des relations non linéaires s'adapte plus à la complexité des données par rapport à la régression linéaire, offrant ainsi des prédictions plus robustes sans les contraintes liées aux hypothèses de linéarité.

FIGURE 9 – Importance des variables pour la Random Forest



(a) Importance des variables pour la Random Forest pour les DROM



(b) Importance des variables pour la Random Forest pour la métropole

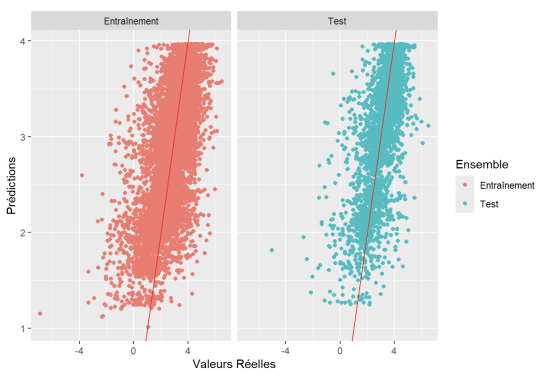
Note de lecture : La variable la plus contributrice aux modèles est la rentabilité économique, suivi des ul de moins de 5ans.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

La figure 9 présente l'importance des variables dans la construction du modèle de Random Forest, en utilisant deux indicateurs de performance : la Mean Squared Error (MSE) et la pureté des nœuds. Dans ce contexte, plus une variable est élevée dans la figure, plus son retrait augmente l'erreur de prédiction du modèle (MSE), ou plus elle contribue à améliorer la pureté des nœuds. Le détail chiffré de l'importance des variables peut être retrouvé en annexe F.

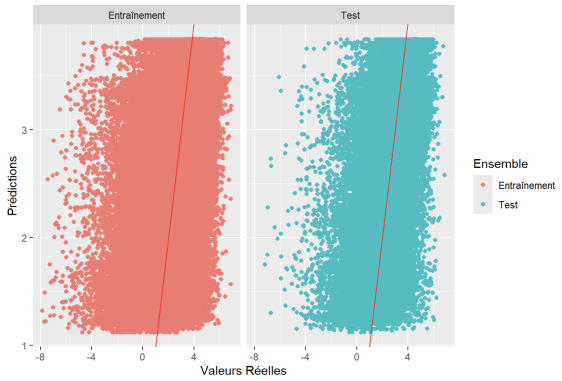
La rentabilité économique, comme attendu, occupe une place centrale dans la construction du modèle en raison de son influence significative sur les résultats. La variable de contrôle "ul de moins de 5 ans" se révèle également cruciale pour le modèle, jouant un rôle important dans la performance du modèle, quel que soit le territoire. Pour les DROM, le secteur du commerce, significatif dans la régression linéaire, se positionne juste après les "ul de moins de 5 ans" en termes d'importance pour le modèle. Les autres variables de contrôle, bien que moins influentes, restent importantes pour réduire l'erreur du modèle, à l'exception des grandes entreprises, dont l'importance est presque négligeable.

En métropole, les variables "Commerces", "Très Grandes Entreprises" et "ul de 5 à 10 ans", qui ne sont pas significatives dans la régression linéaire, apparaissent également parmi les dernières en termes d'importance pour la construction du modèle. Cependant, elles jouent un rôle non négligeable dans la réduction de l'erreur, à l'exception de "Très Grandes Entreprises", dont l'importance est quasiment nulle.

FIGURE 10 – Graphique des valeurs prédites versus les valeurs réelles pour le jeu d'entraînement et de test pour la Random Forest



(a) Graphique des valeurs prédites versus les valeurs réelles pour les DROM



(b) Graphique des valeurs prédites versus les valeurs réelles pour la métropole

Note de lecture : Les modèles ont la même tendance pour le jeu d'entraînement et de test. Les modèles ont donc réussi à généraliser sur un nouvel ensemble de données.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

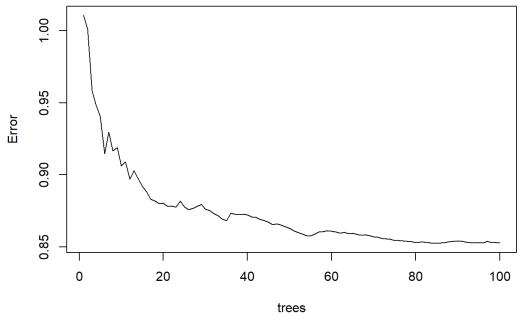
La figure 10 montre le graphique des valeurs prédites versus les valeurs réelles pour le modèle de Random Forest. Que ce soit pour les DROM ou pour la métropole, le modèle montre un meilleur ajustement par rapport à la régression linéaire. Les observations sont plus proches de la ligne de régression, ce qui indique une précision accrue des prédictions, sans sous-estimation des valeurs élevées comme dans le modèle précédent.

TABLE 15 – Comparaison des MSE et de la Variance Expliquée entre la Métropole et les DROM sur les jeux d'entraînement et de test

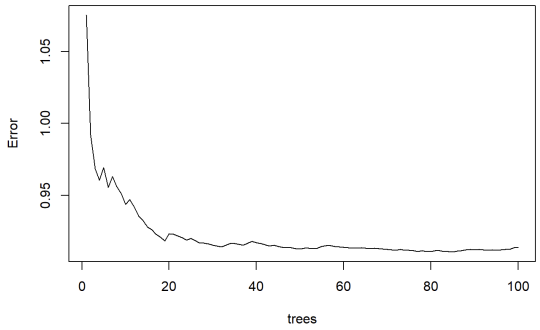
Métrique	**DROM**		**Métropole**	
	Entraînement	**Test**	**Entraînement**	**Test**
RMSE	0.919	0.853	0.957	0.914
Variance Expliquée (en %)	40.97	39.74	39.02	38.29

Comme le montre le tableau 15, les erreurs des modèles sont plus faibles avec une erreur de 0,853 pour les DROM et de 0,914 pour la métropole. De plus, la variance expliquée par le modèle est également plus élevée, avec 39,74% pour les DROM et 38,29% pour la métropole, comme indiqué dans le tableau X. Ce modèle confirme ainsi l'existence d'une corrélation entre la rentabilité économique et la rentabilité financière, comme observé dans la régression linéaire. De plus les performances sont similaires du jeu de d'entraînement au jeu de test ce qui indique une bonne généralisation du modèle.

FIGURE 11 – Erreur du modèle en fonction du nombre d'arbre



(a) Erreur du modèle en fonction du nombre d'arbre pour les DROM



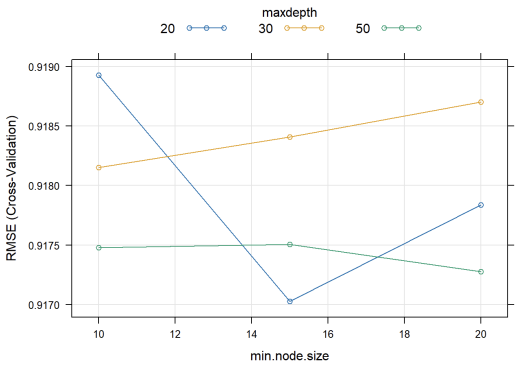
(b) Erreur du modèle en fonction du nombre d'arbre pour la métropole

Note de lecture : L'erreur du modèle converge à 0,85 quand le nombre d'abre augmente.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

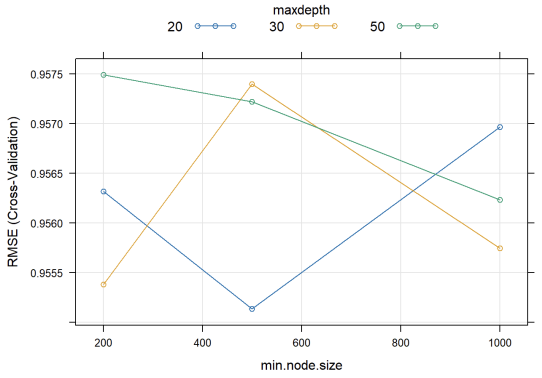
Afin d'assurer que le modèle soit le plus performant possible, une recherche en grille a été effectuée pour optimiser certains hyperparamètres. Dans ce cas, la profondeur de l'arbre et la taille minimale des nœuds ont été ajustées. Le graphique 12 présente les résultats de cette recherche en grille, montrant la combinaison des meilleurs paramètres pour les deux modèles.

Pour les DROM, le modèle est optimal avec une profondeur maximale de 20 et une taille minimale des nœuds de 10 observations. En ce qui concerne la métropole, le modèle atteint une performance optimale avec une profondeur maximale de 50 et une taille minimale des nœuds de 200 observations avant division. La figure 11 illustre que l'erreur des deux modèles converge, indiquant une stabilité dans les performances des deux modèles.

FIGURE 12 – Optimisation par recherche en grille pour le modèle de Random Forest



(a) Résultat de la recherche en grille pour les DROM



(b) Résultat de la recherche en grille pour la métropole

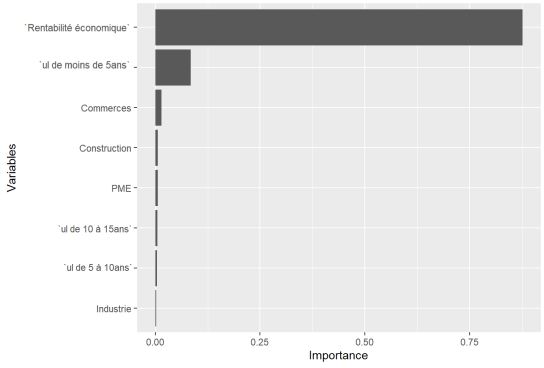
Note de lecture : Les hyperparamètres optimaux pour le modèle de Random Forest dans les DROM sont 20 en profondeur maximale et 10 comme taille de noeud minimal.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

Afin de garantir la robustesse de l'analyse, un modèle de XGBoost sera utilisé pour corroborer ou non les résultats obtenus avec le modèle de Random Forest et la régression linéaire.

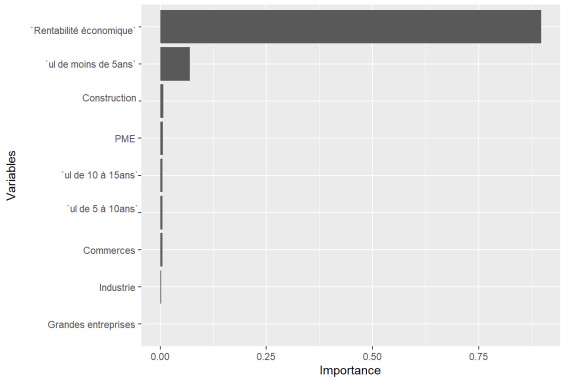
8.3 Un modèle XGBoost plus performant mais avec des résultats similaires

Le XGBoost apporte une approche complémentaire en optimisant l'ajustement des prédictions de manière itérative, ce qui permet de mieux gérer les biais et la variance par rapport à la Random Forest et à la régression linéaire, tout en offrant une flexibilité accrue pour capturer des relations complexes dans les données.

FIGURE 13 – Importance des variables pour les modèles XGBoost



(a) Importance des variables pour le modèle dans les DROM



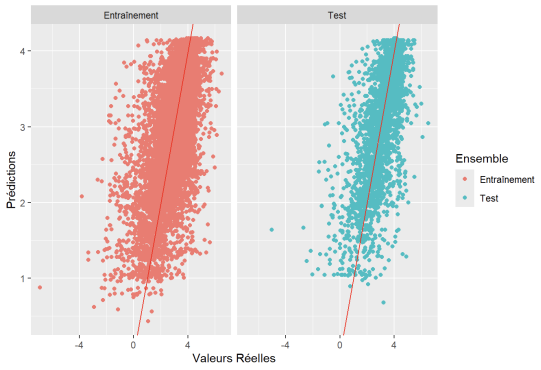
(b) Importance des variables pour le modèle en métropole

Note de lecture : La rentabilité économique est la variable la plus importante pour les deux modèles.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

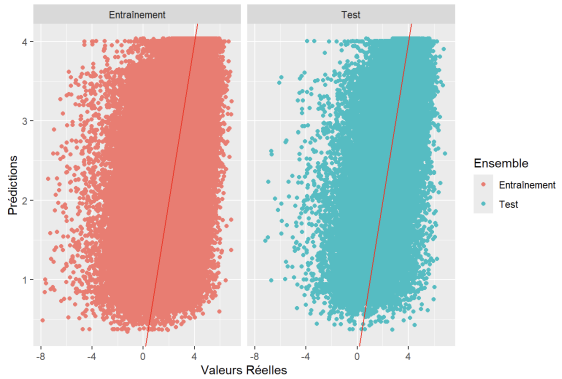
La figure 13 présente les variables les plus importantes pour la construction du modèle XGBoost. Les variables les plus déterminantes sont celles qui réduisent le plus l'erreur de prédiction du modèle, ou, de manière équivalente, celles qui améliorent le plus la précision du modèle. Pour le modèle XGBoost, que ce soit pour les DROM ou pour la métropole, les variables les plus importantes restent les mêmes que celles identifiées précédemment.

La rentabilité économique est la variable la plus influente dans la construction du modèle, suivie des unités légales de moins de 5 ans. Pour les DROM, la variable "Commerces" occupe de nouveau la troisième position en termes d'importance, tandis que pour la métropole, une contribution plus équilibrée des autres variables est observée, sans qu'aucune ne se distingue de manière évidente.

FIGURE 14 – Graphique des valeurs prédites versus les valeurs réelles pour le jeu d'entraînement et de test pour le modèle XGBoost



(a) Graphique des valeurs prédites versus les valeurs réelles pour le jeu d'entraînement et de test pour les DROM



(b) Graphique des valeurs prédites versus les valeurs réelles pour le jeu d'entraînement et de test pour la métropole

Note de lecture : Les modèles ont la même tendance pour le jeu d'entraînement et de test. Les modèles ont donc réussi à généraliser sur un nouvel ensemble de données.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

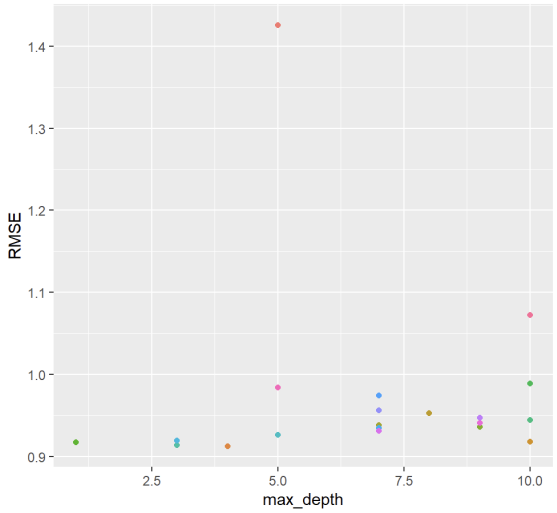
La Figure 14 présente le graphique des valeurs prédites versus les valeurs réelles pour le modèle de XGBoost. Les graphiques obtenus, tant pour les jeux d'entraînement que pour les jeux de test, sont très similaires aux résultats précédemment obtenus avec la Random Forest. La différence se manifeste dans les performances du modèle, comme illustré dans le tableau 16, où il apparaît que le modèle a non seulement réussi à généraliser comme auparavant, mais aussi à obtenir de meilleures performances. La variance expliquée par les modèles est en effet plus élevée, atteignant 41,93% pour les DROM et 39,67% pour la métropole, contre respectivement 39,74% et 38,29% dans le modèle précédent. Ces résultats montrent une amélioration de la précision par rapport aux modèles précédents, tout en conservant une cohérence avec les observations antérieures.

TABLE 16 – Comparaison des RMSE et de la Variance Expliquée (R^2) entre la Métropole et les DROM sur les jeux d'entraînement et de test (XGBoost)

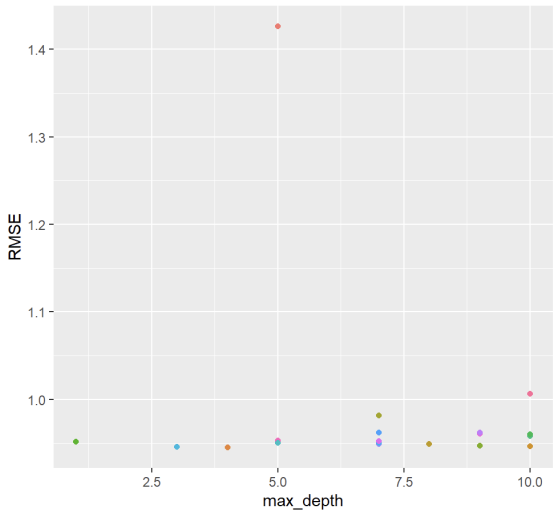
Métrique	**DROM**		**Métropole**	
	Entraînement	**Test**	**Entraînement**	**Test**
RMSE	0.913	0.905	0.945	0.953
Variance Expliquée (en %)	41.20	41.93	39.65	39.67

Afin d'obtenir les résultas précédents une recherche en grille a été effectué optimisant le nombre d'arbre, la profondeur maximale des arbres, le taux d'apprentissage, la régularisation, la proportion de variable selectionnée, ainsi que le poids minimum des enfants. Le résultats de ces optimisations sont représentés dans la figure 15 avec pour exemple le maximum de profondeur comme hyperparamètre. Les points les plus bas représentent les meilleures combinaisons d'hyperparamètres.

FIGURE 15 – Recherche en grille pour le modèle XGBoost



(a) Recherche en grille pour le modèle XGBoost pour les DROM



(b) Recherche en grille pour le modèle XGBoost pour la métropole

Note de lecture : Les points les plus bas sont les combinaisons d'hyperparamètres pour lesquelles les modèles font le moins d'erreur.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

8.4 Conclusion

À travers les différents modèles, une relation positive a été mise en évidence entre la rentabilité économique et la rentabilité financière. La régression linéaire a d'abord permis d'obtenir des coefficients interprétables, tandis que les modèles de Random Forest et XGBoost ont corroboré ces résultats tout en offrant de meilleures performances. Il apparaît ainsi que la rentabilité économique est un facteur clé de la rentabilité financière, bien que la régression linéaire seule ne soit pas suffisante pour capturer cette relation de manière efficace. En effet, les modèles de Random Forest et XGBoost permettent de contourner les hypothèses strictes de la régression linéaire, telles que l'homoscédasticité des résidus ou la gestion des relations non linéaires complexes.

Les performances des modèles dans les DROM et en métropole sont globalement similaires, bien qu'elles soient légèrement supérieures en métropole, ce qui peut s'expliquer par une quantité de données beaucoup plus importante. Ces résultats soulignent l'importance de recourir à des modèles plus sophistiqués pour une meilleure capture des dynamiques entre rentabilité économique et financière.

A Statistiques sur le tissu productif des entreprises

Table 17 – Taille moyenne des entreprises en 2019

Indice	**Guadeloupe**	**Guyane**	**La Réunion**	**Martinique**	**France**
Taille moyenne des entreprises	2.0	2.6	3.1	1.8	4.0
Diversification des entreprises	474	388	487	448	615

Note de lecture : En 2019, la Guadeloupe comptait 474 types d'activités différents pour les entreprises, avec une taille moyenne de 2 salariés par entreprise.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude
Source : calculs des auteurs à partir de la base de données ESANE.(Brion 2011)

Table 18 – Répartition des entreprises et de la proportion de chiffre d'affaires par secteur d'activité en 2019

Zone	Commerces		Construction		Industrie		Services marchands	
	nb d'ul	CA (%)	nb d'ul	CA (%)	nb d'ul	CA (%)	nb d'ul	CA (%)
France	26.3	38.7	17.2	6.98	10.1	29.3	46.4	25.1
Guadeloupe	32.2	52.0	12.6	8.65	14.2	13.3	41.1	26.0
Guyane	28.9	46.8	16.3	13.5	15.2	15.4	39.5	24.3
La Réunion	30.2	51.2	16.8	9.36	13.1	15.8	39.9	23.6
Martinique	30.4	50.1	13.5	6.27	12.1	18.5	43.9	25.1

Note de lecture : En 2019, le secteur du commerce représente 26,3% des entreprises en France métropolitaine et produisent 38,7% du chiffre d'affaires total.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude
Source : calculs des auteurs à partir de la base de données ESANE.(Brion 2011)

Table 19 – Répartition des entreprises et de la proportion de chiffre d'affaires par taille en 2019

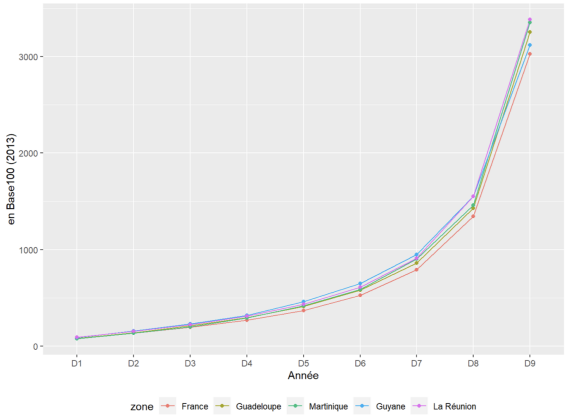
Zone	TPE		PME		GE		TGE	
	nb d'ul	CA (%)	nb d'ul	CA (%)	nb d'ul	CA (%)	nb d'ul	CA (%)
France	84.9	15.9	14.6	38.0	0.42	32.8	0.01	13.3
Guadeloupe	86.9	35.9	13.1	58.6	0.07	5.50	-	-
Guyane	85.1	41.3	14.9	56.9	0.04	1.84	-	-
La Réunion	82.7	25.7	17.1	60.0	0.2	14.3	-	-
Martinique	-	-	-	-	-	-	-	-

Note de lecture : En 2019, les TPE représentent 84,9% des entreprises en France métropolitaine et produisent 15,9% du chiffre d'affaires total.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude
Source : calculs des auteurs à partir de la base de données ESANE.

FIGURE 16 – Distribution du CA par décile par zone en 2019

TABLE 20 – Indice de Gini du CA par zone en 2019

Zone	Indice de Gini
France	0.90
Guadeloupe	0.78
Guyane	0.75
La Réunion	0.80
Martinique	0.81



Note de lecture : En 2019, l'indice de Gini du chiffre d'affaires des entreprises est de 0,9 pour la France métropolitaine.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude
Source : calculs des auteurs à partir de la base de données ESANE.

B Nettoyage de la base de données

B.1 Méthode de l'écart inter-quartile

Dans la méthode de l'écart interquartile, la différence entre le premier et le troisième quartile est utilisée pour identifier les valeurs aberrantes. Graphiquement représentée par un diagramme en boîte à moustaches, la ligne médiane à l'intérieur de la boîte correspond à la médiane des données, tandis que les extrémités de la boîte représentent les premier et troisième quartiles. Les « moustaches » s'étendent jusqu'à la valeur maximale qui ne dépasse pas 1,5 fois l'écart interquartile au-dessus du troisième quartile ou en dessous du premier quartile, les valeurs au-delà de ces limites étant considérées comme des anomalies. On appelle les "moustaches" valeurs inférieures ou supérieures adjacentes. Mathématiquement on définit le calcul ainsi :

$$\text{IQR} = Q3 - Q1$$

Soit un seuil s,

$$VAI = Q1 - s \times \text{IQR} \quad \text{et} \quad VAS = Q3 + s \times \text{IQR}$$

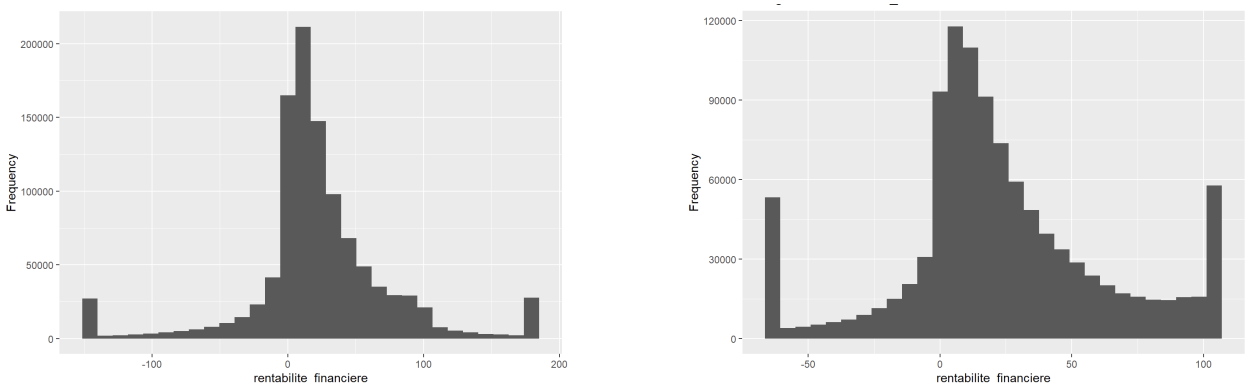
Le seuil s est généralement établi à 1,5 comme on l'a vu précédemment mais peut être étendu à 3 selon la distribution de nos données pour obtenir des valeurs plus extrêmes comme outlier.

(ici faire un boxplot de la rentabilité financière pour montrer l'avant après avec le seuil de 1.5 et 3)

B.2 Winsorisation

La winsorisation (HASTINGS et al. 1947) est une méthode de seuillage qui permet de réduire l'influence des outliers tout en conservant l'ensemble de nos données. Si X est une variable aléatoire avec des valeurs extrêmes X_{low} et X_{high}, la winsorisation à 5% transformerait X_{low} en $P5$ et X_{high} en $P95$.

FIGURE 17 – Winsorisation avec un seuil de 2,5 et un seuil de 5 pour la rentabilité financière



Note de lecture :
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude
Source : calculs des auteurs à partir de la base de données ESANE.

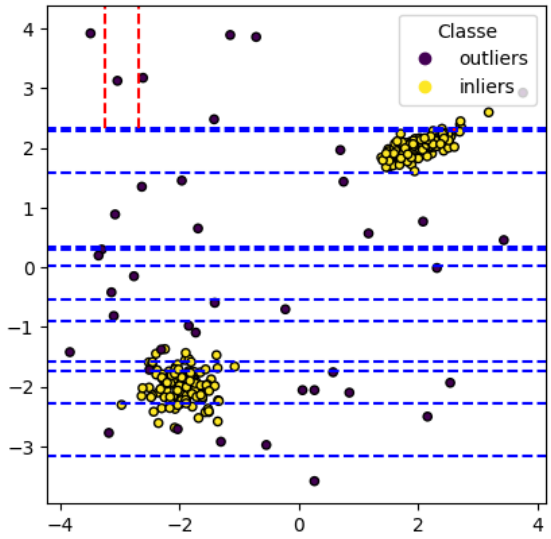
La figure 17 illustre l'impact de la winsorisation sur la distribution de la rentabilité financière du jeu de données. Une winsorisation avec un seuil de 2,5 modifie la structure de plus de 50 000 données (plus de 25 000 à droite et à gauche). Avec un seuil de 5, l'altération des données est encore plus importante, affectant plus de 100 000 données (50 000 à gauche et à droite). En comparant les deux distributions, le seuillage est plus acceptable pour un seuil de 2,5 car il perturbe moins les extrémités de la distribution, réduisant ainsi la variance tout en limitant le biais.

B.3 L'isolation forest

L'isolation forest (LIU, TING et ZHOU 2008) est une technique de détection d'anomalies basé sur des arbres de décisions. Le principe de cet algorithme repose sur un processus itératif : Pour commencer une variable est sélectionnée de manière aléatoire. Ensuite, le jeu de données est partitionné aléatoirement selon cette variable, créant ainsi deux sous-ensembles de données. Ce processus est répété jusqu'à ce qu'une donnée soit isolée (figure 18).

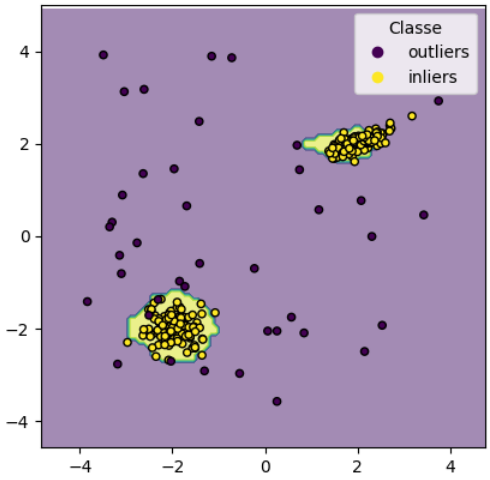
De manière récursive, les étapes de sélection aléatoire et de partitionnement sont répétées pour isoler d'autres données. Des dizaines ou centaines d'arbres sont créés puis combinés afin d'obtenir un résultat optimal. Les anomalies seront alors l'ensemble des points qui seront isolées dans les feuilles les plus proches de la racine de l'arbre. Cette distance par rapport à la racine permet d'attribuer un score d'anomalie à l'observation. Plus ce score est élevé, plus l'observation est susceptible d'être un outlier. La figure 20 représente les frontières de décisions à la suite de l'application de l'algorithme. Cet algorithme possède plusieurs hyperparamètres qui sont ajustables tel que la taille de l'échantillon ou le nombre d'arbre construit. Les hyperparamètres utilisés sont ceux conseillés par l'article de référence (LIU, TING et ZHOU 2008). Les figures 19b et 19a représentent les frontières de décisions de l'isolation forest pour un jeu de données simulé aléatoirement selon une loi normale avec des outliers générés selon une loi uniforme.

FIGURE 18 – Génération des droites de décisions avec l'isolation forest

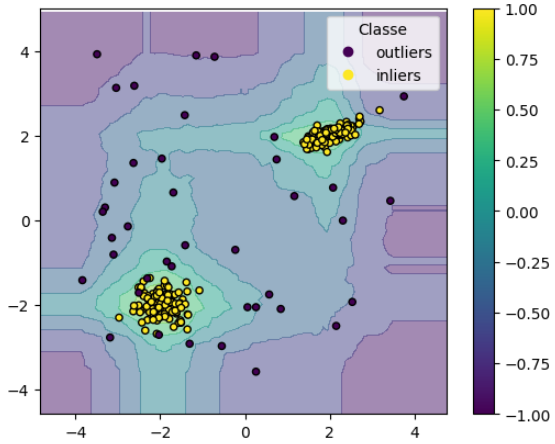


Note de lecture : Les outliers représentées en violet sont progressivement isolées grâce aux droites de décisions tracées aléatoirement.
Champ : Données générées aléatoirement selon une loi normale pour les inliers et une loi uniforme pour les outliers
Source : calculs des auteurs à partir de données simulées.

FIGURE 19 – Représentation des frontières de décision de l'isolation forest



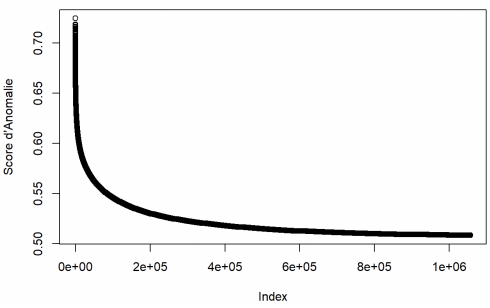
(a) Frontières de décision de l'isolation forest



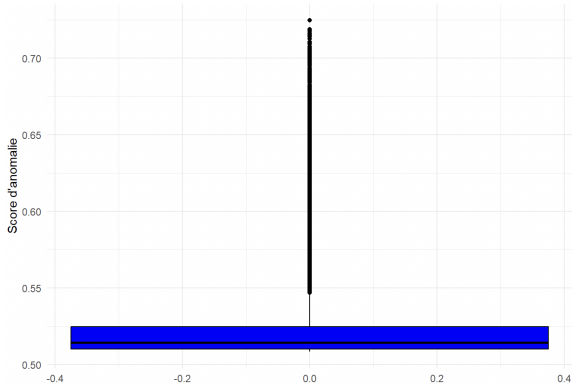
(b) Frontières de décision de l'isolation forest en terme de longueur de chemin

Note de lecture : La figure 19a représente les frontières de décisions générées par l'algorithme pour décider si une observation est considérée comme outlier ou inlier. Les couleurs et contours de la figure 19b représentent les niveaux de scores d'anomalie, avec des valeurs allant de -1 (fortes probabilité d'anomalies) à 1 (faibles probabilité d'anomalies).
Champ : Données générées aléatoirement selon une loi normale pour les inliers et une loi uniforme pour les outliers
Source : calculs des auteurs à partir de données simulées.

FIGURE 20 – Représentation des scores d'anomalies



(a) Scores d'anomalies de l'isolation forest par ordre décroissant



(b) Boxplot des scores d'anomalies

Note de lecture La figure 20a représente les observations triées par ordre décroissant avec un coude qui se situe avec un score d'anomalie d'environ 0,53. La figure 20b
Champ : Données générées aléatoirement selon une loi normale pour les inliers et une loi uniforme pour les outliers
Source : calculs des auteurs à partir de données simulées.

Une fois les scores d'anomalies calculés, il est nécessaire de choisir un seuil au-delà duquel les observations seront considérées comme outliers. Les figures 20a et 20b illustrent la distribution des scores d'anomalies, où les outliers apparaissent comme les observations avec un score d'anomalie se distinguant nettement du reste du jeu de données, notamment entre 0,52 et 0,55.

Le tableau 21 présente les statistiques descriptives du jeu de données initial pour la rentabilité financière, comparées aux jeux de données résultant de l'isolation forest selon différents seuils. Le tableau 22 indique le nombre d'observations considérées comme outliers par l'isolation forest en fonction du seuil choisi. Étant donné la complexité du jeu de données ESANE, le nombre d'observations considérées comme outliers peut être significatif. Retirer trop d'observations pourrait introduire un biais important dans les analyses futures. Il est donc essentiel de réaliser un compromis biais-variance pour conserver un maximum d'informations tout en produisant des modèles interprétables.

Les statistiques obtenues après l'application de l'isolation forest sont nettement plus acceptables que celles du jeu de données initial, bien qu'elles restent parfois élevées. La moyenne du jeu de données est considérablement réduite, passant de -4.28×10^{11} à 22,98 pour un seuil de 0,55, tout en conservant une médiane équivalente. La variance est également réduite de manière significative, passant de 3.64×10^{14} pour le jeu de données initial à 77,42 après l'isolation forest avec un seuil de 0,55. Contrairement à la moyenne et à la médiane, la variance, les percentiles et les valeurs extrêmes varient considérablement en fonction du seuil choisi, avec par exemple une variance de 34,82 pour un seuil de 0,52, soit 42,6 de moins que pour le seuil le plus permissif.

En considérant les différentes statistiques, le seuil de 0,52 semble le plus intéressant, car il réduit les valeurs extrêmes à des niveaux plus acceptables que les seuils plus élevés. Cependant, ce seuil entraîne la suppression de 351 245 observations, soit un tiers du jeu de données, ce qui peut introduire un biais très important. Par conséquent, le seuil de 0,53 sera retenu pour l'ACP. Bien que plus permissif concernant les valeurs extrêmes, ce seuil permet de conserver 80% du jeu de données et de minimiser le biais (tableau 22).

Table 21 – Statistiques des différents jeux de données pour la rentabilité financière

Jeux de données	moyenne	Variance	P1	médiane	P99	minimum	maximum
Jeu de données initial	-4.28×10^{11}	3.64×10^{14}	-394.19	16.91	436.78	-2.91×10^{17}	9.63×10^{16}
Isolation forest 0.52	21.70	34.82	-81.90	16.63	112.94	-362.50	382.64
Isolation forest 0.53	22.10	46.64	-124.23	16.78	158.65	-602.76	1028.09
Isolation forest 0.54	22.61	60.76	-164.13	16.83	200.89	-1074.41	2396.28
Isolation forest 0.55	22.98	77.42	-201.63	16.83	242.90	-3895.58	6603.17

Note de lecture : L'application de l'algorithme isolation forest pour repérer les outliers permet de réduire significativement les indicateurs statistiques de la base de données avec par exemple la moyenne qui passe de -4.28×10^{11} pour la rentabilité financière à 21,7 pour un seuil sur le score d'anomalie de 0,52.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude
Source : calculs des auteurs à partir de la base de données ESANE.

Table 22 – Nombre d'observations retirés pour chaque seuil de l'isolation Forest

Méthode	Observations
Isolation forest 0.52	351 245
Isolation forest 0.53	202 211
Isolation forest 0.54	128 218
Isolation forest 0.55	83 881

Note de lecture : L'algorithme isolation forest a considéré 202 211 observations comme aberrantes avec un seuil sur le score d'anomalie à 0,53
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude
Source : calculs des auteurs à partir de la base de données ESANE.

C Définitions d'indicateurs financiers

Table 23 – Définitions des indicateurs financiers

Variable	Définition
taux marge commerciale	Rapport de la marge commerciale au coût d'achat des marchandises.
taux de valeur ajoutée	Taux mesurant le taux d'intégration de la société dans le processus de production ainsi que le poids des charges externes - Calcul : Valeur ajoutée hors taxes/Chiffre d'affaires.
taux de marge	Le taux de marge est le rapport entre le résultat net et le chiffre d'affaires. Il mesure la rentabilité d'une entreprise en montrant la part du chiffre d'affaires qui reste après déduction de toutes les charges.
taux de marge brute	Taux mesurant la capacité de l'entreprise à générer une rentabilité à partir du chiffre d'affaires - Calcul : Excédent brut d'exploitation (EBE)/Chiffre d'affaires (CA).
taux de résultat d'exploitation	Le taux de résultat d'exploitation est le rapport entre le résultat d'exploitation et le chiffre d'affaires. Il mesure la rentabilité opérationnelle de l'entreprise, c'est-à-dire la capacité de l'entreprise à générer des bénéfices à partir de ses activités principales avant la prise en compte des éléments financiers et exceptionnels.

TABLE 23 – (suite)

Variable	Définition
taux de résultat courant	Le taux de résultat courant est le rapport entre le résultat courant avant impôt et le chiffre d'affaires. Il permet de mesurer la performance économique et financière de l'entreprise en incluant les résultats d'exploitation et les résultats financiers.
rentabilité économique	Excédent brut d'exploitation / (immobilisations corporelles et incorporelles + besoin en fonds de roulement).
rentabilité financière	Résultat net comptable / Capitaux propres.
taux de profit	Le taux de profit est le rapport entre le bénéfice net et le chiffre d'affaires. Il indique la proportion de profit réalisé par rapport au chiffre d'affaires total.
ratio CAHT/Capitaux propres	Ce ratio compare le chiffre d'affaires hors taxes (CAHT) aux capitaux propres. Il permet de mesurer la capacité d'une entreprise à générer des ventes par rapport à ses fonds propres.
autonomie financière	L'autonomie financière est le rapport entre les capitaux propres et le total du passif. Elle mesure la part des ressources propres de l'entreprise par rapport à l'ensemble de ses ressources, indiquant ainsi la capacité de l'entreprise à financer ses activités sans recourir à des dettes.
taux d'endettement	Taux mesurant la part des dettes dans les ressources totales de l'entreprise - Calcul : Emprunts et dettes assimilées / Total passif.
levier financier	Rapport entre les apports extérieurs et stables et les apports internes - Le ratio de levier se calcule en rapportant les emprunts et dettes assimilées sur les capitaux propres et les autres fonds propres.
poids des charges financières	Le poids des charges financières est le rapport entre les charges financières (comme les intérêts sur les emprunts) et le chiffre d'affaires. Il mesure l'impact des coûts de financement sur la performance de l'entreprise.
taux d'intérêt apparent	Taux d'intérêt moyen payé sur les dettes financières - Calcul : Intérêts et charges assimilées / Emprunts et dettes assimilées.
capacité de remboursement	Capacité d'autofinancement / Emprunts et dettes assimilées.
marge d'autofinancement	Part de la valeur ajoutée disponible pour le financement des investissements, le remboursement des emprunts, l'augmentation du fonds de roulement et la rémunération des associés.
stocks totaux en JCA	Le stock total en jours de chiffre d'affaires (JCA) est le ratio des stocks totaux divisés par le chiffre d'affaires moyen journalier. Il mesure le nombre de jours de vente que représentent les stocks.
BFR en JCA	Actif circulant - Passif circulant.
intensité capitalistique	Rapport entre les immobilisations corporelles et l'effectif en fin d'exercice et l'effectif salarié en équivalent temps plein.
productivité du travail	La productivité du travail est généralement mesurée par le rapport entre la valeur ajoutée et l'effectif salarié en équivalent temps plein (ETP). Elle indique la quantité de valeur ajoutée produite par chaque employé.
taux d'investissement	Part de la valeur ajoutée consacrée à l'investissement corporel (hors apports reçus des autres sociétés) / Valeur ajoutée hors taxes.
taux d'autofinancement	Capacité d'autofinancement / investissements corporels bruts hors apport.
délais de paiement clients	Rapport entre le total des créances clients sur l'ensemble du secteur et le total des chiffres d'affaires annuels TTC divisé par 360, pour être exprimé en jours de chiffre d'affaires.

Table 23 – (suite)

Variable	Définition
délais de paiement fournisseurs	Rapport entre le total des dettes fournisseurs sur l'ensemble du secteur et le total des achats et charges externes annuels TTC divisé par 360, pour être exprimé en jours d'achats.
solde commercial	le solde commercial d'une entreprise est la différence entre ce qu'elle gagne en vendant ses produits ou services et ce qu'elle dépense pour les acheter ou les produire.

Table 24 – Données moyennes par variable avant le nettoyage de la base de données

Variable	**Min**	**1er Quartile**	**Médiane**	**Moyenne**	**3e Quartile**	**Max**
Nombre d'unités légales	1197.0	3888.3	6608.0	9682.0	11057.8	77696.0
Charges personnel sur VA	0.2	0.4	0.4	0.4	0.4	0.6
Taux des stocks	4.4	10.6	12.1	12.3	13.8	24.3
Taux de BFR autres actifs	16.0	23.3	25.4	26.4	29.1	38.0
Disponibilités	4.4	9.3	10.6	10.6	11.7	16.6
Taux de BFR passif	23.4	33.9	36.4	37.5	40.4	63.2
Taux de VA	15.0	23.6	25.8	25.7	27.8	46.9
Taux de marge brute	0.5	4.8	5.3	5.3	5.8	8.3
Taux de résultat courant	-1.7	3.6	4.0	4.0	4.5	8.7
Rentabilité économique	1.5	10.0	11.1	11.0	12.1	18.0
Rentabilité financière	-21.8	9.9	11.3	10.9	12.5	20.0
Ratio CAHT / Capitaux propres	177.6	330.7	361.5	375.0	403.6	780.4
Autonomie financière	18.8	37.9	41.5	40.7	43.7	51.2
Taux d'endettement	7.0	17.0	19.1	19.1	21.1	41.7
Taux de prélèvement financier	3.0	4.6	5.2	6.4	6.2	51.3
Intensité capitalistique	50.1	72.1	81.0	81.7	89.6	115.0
Solde commercial	-22.5	0.1	3.1	2.9	5.9	22.2

Note de lecture : Le taux de marge brut moyen des unités légales en 2018 est de 5,3%
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

Table 25 – Données moyennes par variable après le nettoyage de la base de données

Variable	Min	1er Quartile	Médiane	Moyenne	3e Quartile	Max
Nombre d'unités légales	1099.0	3434.5	5831.0	8543.5	10027.8	65444.0
Charges personnel sur VA	0.2	0.4	0.4	0.4	0.4	0.6
Taux des stocks	4.6	9.9	11.3	11.1	12.5	17.1
Taux de BFR autres actifs	15.9	23.7	25.8	26.9	29.2	42.3
Disponibilités	4.8	10.6	11.4	11.5	12.7	17.8
Taux de BFR passif	20.2	34.2	36.5	37.2	39.5	51.9
Taux de VA	15.5	25.7	27.6	27.8	29.5	47.7
Taux de marge brut	2.9	5.3	5.9	5.9	6.4	9.2
Taux de résultat courant	-1.8	4.1	4.5	4.5	4.9	8.2
Rentabilité économique	7.5	11.5	12.4	12.6	13.4	20.5
Rentabilité financière	-10.3	11.1	12.5	12.3	13.6	21.0
Ratio CAHT / Capitaux propres	255.9	329.4	363.4	368.7	398.5	776.6
Autonomie financière	30.1	41.3	43.5	43.0	45.3	52.7
Taux d'endettement	8.3	15.1	17.0	17.3	19.0	45.2
Taux de prélèvement financier	2.4	3.5	3.8	4.2	4.5	17.0
Intensité capitalistique	38.3	65.3	73.6	74.8	81.8	113.3
Solde commercial	-10.9	4.2	6.4	6.7	9.4	23.4

Note de lecture : Le taux de marge brut moyen des unités légales en 2018 est de 5,9% après le nettoyage de la base de données
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

Table 26 – Moyenne de la rentabilité économique et financière par département

Département	Rentabilité économique	Rentabilité financière
Guadeloupe	15.3	13.2
Martinique	13.5	14.0
Guyane	16.9	17.7
La Réunion	14.7	13.8
France métropolitaine	12.5	12.2

Note de lecture : La rentabilité économique moyenne de la Guadeloupe en 2018 est de 15,3%
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

Table 27 – Décomposition de la rentabilité financière par département

Département	Intensité capitalistique	Ratio CAHT / CP	Taux de profit
Guadeloupe	55.7	3.6	3.7
Martinique	61.8	3.9	3.6
Guyane	69.3	3.1	5.7
La Réunion	59.7	3,3	4.2
France métropolitaine	75.3	3,7	3.4

Note de lecture : Le taux de profit moyen de la Guadeloupe en 2018 est de 3,7%
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

D Bagging

Algorithm 3 Algorithme de Bagging

Input: Ensemble d'entraînement $\{(x_i, y_i)\}_{i=1}^{N}$,
nombre d'arbres B,
taux d'échantillonnage α
Output: Fonction de prédiction $\hat{f}(x)$

1 **for** $b = 1$ **to** B **do**
2 **Échantillonnage bootstrap** :
Échantillonner aléatoirement avec remplacement un sous-ensemble de l'ensemble d'entraînement $\{(x_i^{(b)}, y_i^{(b)})\}_{i=1}^{\alpha N}$.
3 **Construction de l'arbre** :
Construire un arbre de décision $f_b(x)$ en utilisant les observations échantillonnées et toutes les variables disponibles.
4 **end**
5 **Prédiction finale** :
La prédiction pour un nouvel échantillon x' est obtenue en moyennant les prédictions de tous les arbres construits :

$$\hat{f}(x') = \frac{1}{B} \sum_{b=1}^{B} f_b(x').$$

E ACP

Pour commencer, une analyse de la distribution et de la linéarité des données sera effectuée. En effet, l'Analyse en Composantes Principales (ACP) fonctionne mieux sur un ensemble de données linéaires avec une distribution normale (Pearson 1901).

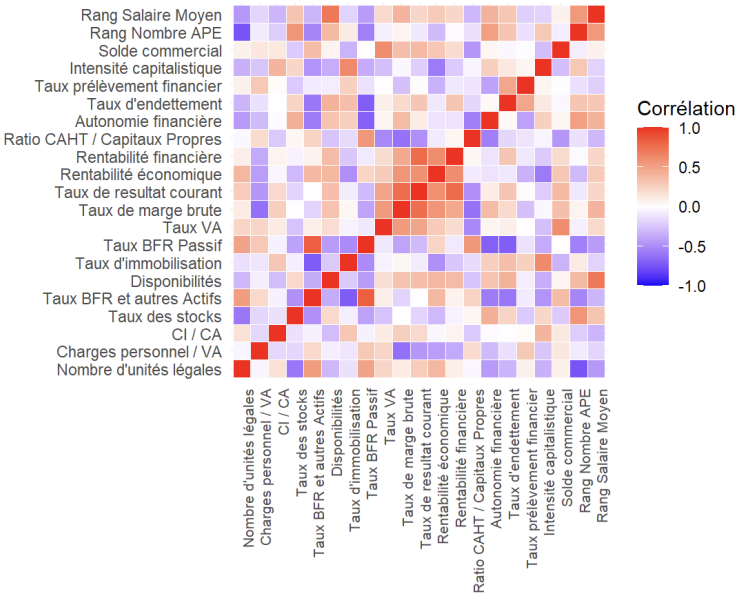
E.1 Préparation et adéquation des données entreprises pour l'ACP

Afin de determiner si les données et variables sont adéquates pour l'ACP, le test de Barlett (Bartlett 1950) et le test de Kaiser-Meyer-Olkin (Kaiser 1970) seront effectués. Le premier permet de determiner si la matrice de corrélation utilisée pour les variables est significativement différente de la matrice identité. Le second permet de mesurer la proportion de variance parmi les variables qui est due à une variance commune. Avant d'effectuer les tests et l'ACP, le jeu de données est centré réduit afin d'exprimer les variables à la même échelle et qu'elles aient le même poids pour l'ACP. L'ACP érant une technique linéaire, elle ne tient pas compte des relations non linéaires et présente une forte sensibilité aux valeurs extrêmes. C'est pourquoi le jeu de données a été nettoyé et que les tests d'adéquations sont effectués.

E.1.1 Etude de la multicolinéarité

L'étude de la multicolinéarité est cruciale avant d'effectuer l'ACP afin d'obtenir des résultats fiables. En effet, deux variables trop fortement corrélées peuvent apporter la même information et ainsi perturber le résultat final. Pour éviter la redondance de l'information, il est nécessaire d'examiner la matrice de corrélation et d'éliminer les variables présentant une corrélation trop élevée. Ici, le seuil retenu est de 0,9 en valeur absolue. Après calcul (figure 21), aucune variable ne dépasse ce seuil, elles seront donc toutes conservées pour la suite.

FIGURE 21 – Dimension 5



Note de lecture : La variable du rang du nombre d'APE est fortement corrélée négativement au nombre d'unités légales.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

Le coefficient de corrélation de Pearson utlisé dans la figure 21 se calcule à l'aide de la formule suivante :

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2}$$

où r est le coefficient de corrélation de Pearson, x_i et y_i sont les valeurs individuelles des deux variables, $\bar{x}$ et $\bar{y}$ sont les moyennes des variables x et y, et n est le nombre de paires de valeurs.

E.1.2 Test de Kaiser-Meyer-Olkin (KMO)

Le test de Kaiser-Meyer-Olkin (KAISER 1970) est utilisé pour mesurer l'adéquation de l'échantillonnage pour les techniques de réduction de dimensionalités. Il évalue la proportion de la variance parmi les variables qui pourrait être causée par des facteurs sous-jacents communs. Le test KMO compare les corrélations simples entre les variables avec les corrélations partielles. Une valeur KMO élevée indique que les données sont bien adaptées à l'ACP. Le KMO global est calculé comme suit :

$$KMO = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} p_{ij}^2} \tag{3}$$

où r_{ij} est la corrélation simple entre les variables i et j, et p_{ij} est la corrélation partielle entre les variables i et j. Une valeur KMO supérieure à 0.8 est considérée comme excellente, entre 0.7 et 0.8 comme bonne, entre 0.6 et 0.7 comme moyenne, entre 0.5 et 0.6 comme médiocre, et inférieure à 0.5 comme inacceptable (KAISER et RICE 1974).

La mesure d'adéquation de l'échantillonnage (MSA) pour chaque variable est calculée de manière similaire et permet d'identifier les variables qui contribuent le plus ou le moins à la structure globale des données. Une MSA faible pour une variable peut indiquer qu'elle ne s'intègre pas bien dans le modèle.

Le tableau 28 présente les valeurs des MSA avant le retrait des variables "charges de personnels sur la valeur ajoutée" et "taux d'immobilisation". Ces variables ont été retirées en raison de leurs très faibles MSA, ce qui pourrait nuire à la qualité de l'ACP. Une fois ces variables retirées, le MSA global passe de

0,56 à 0,67, améliorant ainsi la qualité de médiocre à moyenne. Le retrait d'autres variables n'augmente pas significativement ce score et entraîne une perte d'information importante. Par conséquent, seules ces deux variables seront retirées.

TABLE 28 – Valeurs MSA avant et après retrait de deux variables pour chaque variable

Variable	**MSA avant retrait**	**MSA**
Nombre d'unités légales	0.71	0.80
Charges personnel / VA	0.36	-
Capacité de remboursement (CI / CA)	0.59	0.59
Taux des stocks	0.40	0.77
Taux BFR autres actifs	0.55	0.77
Disponibilités	0.45	0.73
Taux d'immobilisation	0.42	-
Taux BFR passif	0.64	0.66
Taux VA	0.42	0.65
Taux de marge brute	0.62	0.66
Taux de résultat courant	0.67	0.67
Rentabilité économique	0.75	0.64
Rentabilité financière	0.55	0.57
Ratio CAHT / Capitaux propres	0.66	0.62
Autonomie financière	0.51	0.55
Taux d'endettement	0.46	0.51
Taux de prélèvement financier	0.58	0.52
Intensité capitalistique	0.55	0.70
Solde commercial	0.79	0.62
Rang nombre APE	0.79	0.90
Rang salaire moyen	0.85	0.79
Moyenne (KMO)	0,56	0.67

Note de lecture : La MSA (Measure of Sampling Adequacy) moyenne après le retrait des variables taux d'immobilisation et charges de personnel sur valeur ajoutée passe de 0,56 à 0,67.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

E.1.3 Test de Barlett

Le test de Bartlett (BARTLETT 1950) est utilisé pour vérifier si les variables dans un ensemble de données sont suffisamment corrélées pour justifier l'utilisation de méthodes de réduction de dimensions tel que l'ACP. Ce test évalue l'hypothèse nulle selon laquelle la matrice de corrélation est une matrice d'identité, ce qui signifierait que les variables ne sont pas corrélées entre elles. Rejeter cette hypothèse indique que les variables sont corrélées de manière significative, rendant l'ACP appropriée. La statistique de test de Bartlett est calculée ainsi :

$$\chi^2 = -\left(n - \frac{2p+5}{6}\right) \ln |\mathbf{R}| \tag{4}$$

où n est le nombre d'observations, p est le nombre de variables, et $|\mathbf{R}|$ est le déterminant de la matrice de corrélation. Cette statistique suit une distribution χ^2 avec $\frac{p(p-1)}{2}$ degrés de liberté. Une p-valeur faible (généralement inférieure à 0,05) permet de rejeter l'hypothèse nulle, indiquant que les données sont adaptées à l'ACP. Le tableau 29 présente les résultats de ce test, avec une valeur du chi-carré très élevée de 1896 et une p-value extrêmement faible de 1.96×10^{-288}. Cela permet de rejeter l'hypothèse nulle selon laquelle la matrice de corrélation est semblable à la matrice identité.

Table 29 – Résultats du test de Bartlett

Variable	Valeur
Chi-carré	1896.5
p-value	1.96×10^{-288}
Degrés de liberté	171

Note de lecture : La p-value est égale à 1.96×10^{-288} et est inférieur à 0,05, l'hypothèse nulle est rejetée et la matrice de corrélation est différente de la matrice identité.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

E.2 Distribution des données entreprises

Les résultats des tests de normalité de Shapiro-Wilk (Shapiro et Wilk 1965) (tableau 30) pour chaque variable offrent une vision claire de la distribution des données. Les résultats de ces tests montrent que plusieurs variables, telles que le nombre d'unités légales, le taux de prélèvement financier, et la rentabilité financière, présentent des p-values très faibles, indiquant des écarts significatifs par rapport à une distribution normale. Pour qu'une variable ait une distribution normale, la p-value doit être supérieure à un seuil conventionnel (ici 0.05). La majorité des variables ne suivent pas une distribution normale, comme le confirment ces p-values.

L'ACP peut tout de même être appliquée, mais elle aurait été plus performante avec des variables suivant une distribution normale. En effet, l'ACP repose sur le calcul de la covariance entre les variables, et une distribution normale des données permet de mieux capturer la variance et les relations linéaires entre les variables. Cela se traduit par une meilleure identification des composantes principales qui expliquent la plus grande part de la variance totale, rendant l'analyse plus précise et fiable.

TABLE 30 – Résultats du test de Shapiro-Wilk pour la normalité des variables

Variable	**W**	**p-value**
Nombre d'unités légales	0.69	3×10^{-13}
Capacité de remboursement	0.98	0.08
Taux des stocks	0.98	0.19
Taux de BFR autres actifs	0.94	3×10^{-5}
Disponibilités	0.98	0.15
Taux de BFR passif	0.96	5×10^{-3}
Taux de VA	0.89	$6,8 \times 10^{-7}$
Taux de marge brute	0.94	2×10^{-4}
Taux de résultat courant	0.83	$2,1 \times 10^{-9}$
Rentabilité économique	0.96	5×10^{-3}
Rentabilité financière	0.76	$1,3 \times 10^{-11}$
Ratio CAHT / Capitaux propres	0.83	$2,7 \times 10^{-9}$
Autonomie financière	0.96	3×10^{-3}
Taux d'endettement	0.78	5×10^{-11}
Taux de prélèvement financier	0.56	$9,2 \times 10^{-16}$
Intensité capitalistique	0.99	0.37
Solde commercial	0.94	1×10^{-4}
Rang nombre APE	0.95	2×10^{-3}
Rang nombre de salarié moyen	0.95	2×10^{-3}

Note de lecture : La variable nombre d'unités légales ne suit pas une distribution normale car la p-value du test de Shapiro-Wilk égale à 3×10^{-13} qui est inférieure à 0.05.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

La statistique de test du test de Shapiro-Wilk est définie comme suit :

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{5}$$

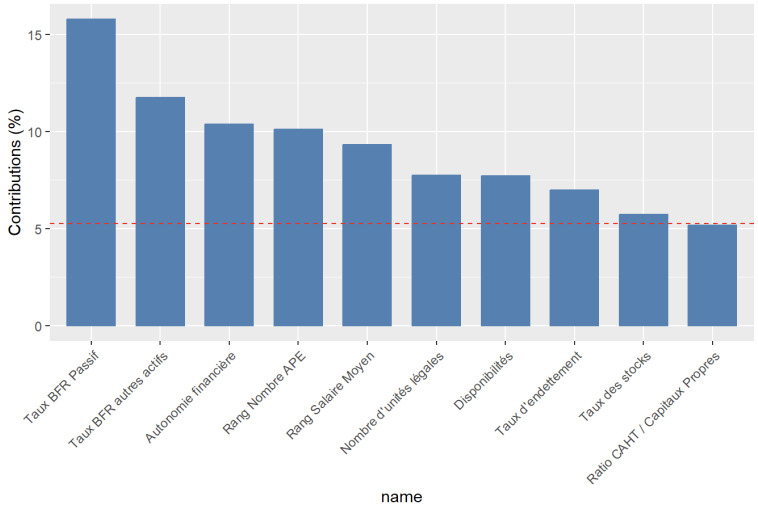
où W est la statistique du test de Shapiro-Wilk, $x_{(i)}$ est la i-ième plus petite valeur de l'échantillon, $\bar{x}$ est la moyenne des valeurs de l'échantillon, et a_i sont des coefficients constants calculés à partir de la moyenne, de la variance et de la covariance de l'échantillon, et dépendent de la taille de l'échantillon.

E.3 Conclusion

Au vu des résultats des tests précédemment effectués, l'ACP peut être réalisée et devrait produire des résultats interprétables. Bien qu'une distribution normale des variables aurait été idéale, les résultats obtenus jusqu'à présent indiquent que l'analyse pourra néanmoins fournir des informations significatives. Les ajustements et nettoyages de la base de données, tels que le traitement des valeurs extrêmes et la vérification de la multicolinéarité, ont contribué à améliorer la qualité des données, permettant ainsi une ACP robuste et fiable.

E.4 Graphiques des valeurs propres

FIGURE 22 – Dimension 1



Note de lecture : Le taux de profiy moyen de la Guadeloupe en 2018 est de 3,7%
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

FIGURE 23 – Dimension 2

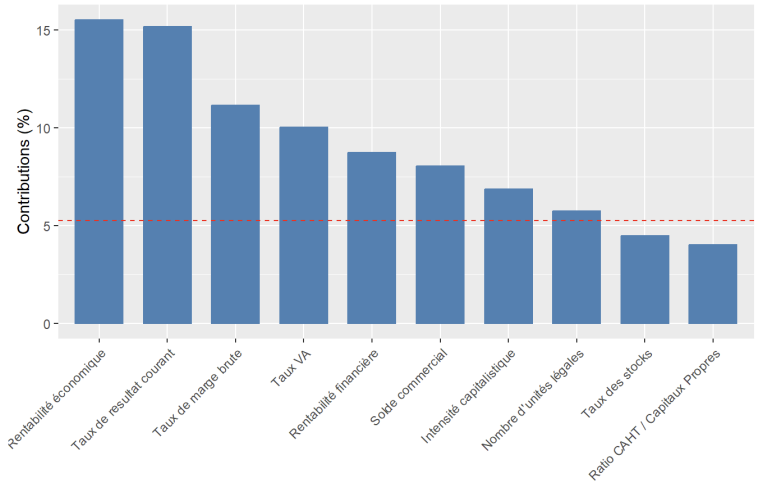


FIGURE 24 – Dimension 3

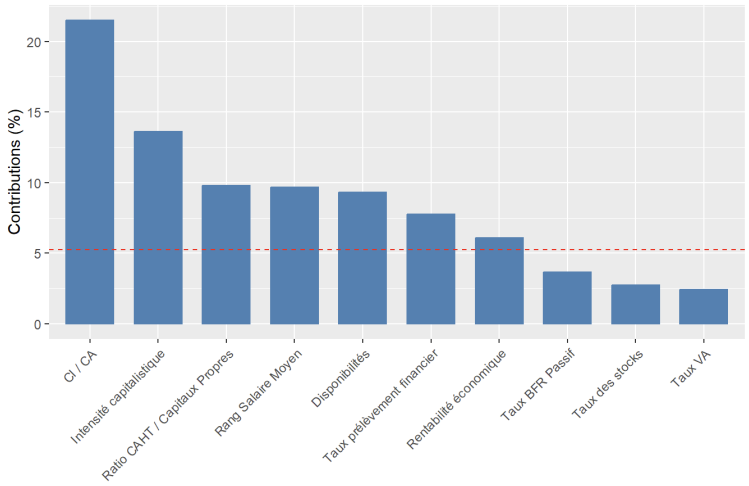


FIGURE 25 – Dimension 4

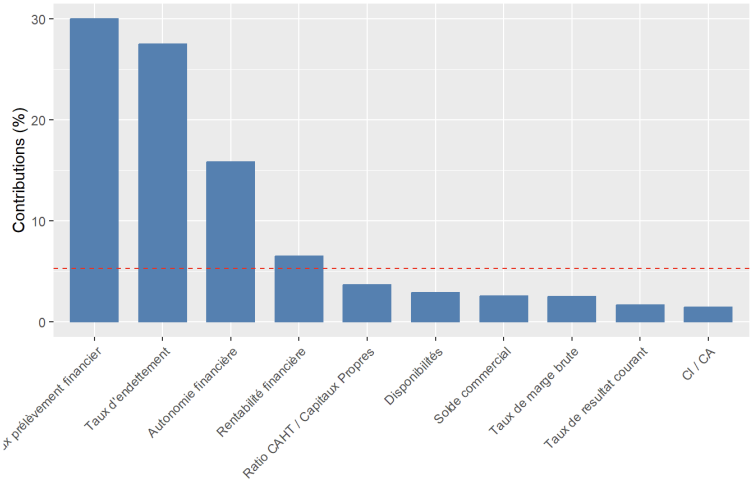
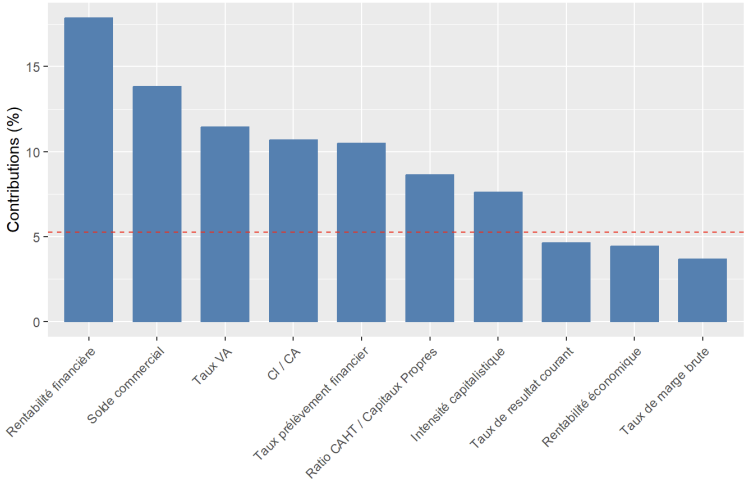


FIGURE 26 – Dimension 5



F Random Forest

Table 31 – Importance des variables pour la Random Forest

Variable	**DROM**		**Métropole**	
	%IncMSE	**IncNodePurity**	**%IncMSE**	**IncNodePurity**
Rentabilité économique	36.733	4966.429	48.004	214935.7
Commerces	5.172	103.567	2.341	1051.860
Construction	2.749	56.404	12.241	1314.383
Industrie	-0.072	28.096	2.337	401.633
Grandes entreprises	-3.691	2.443	10.479	123.318
PME	3.688	62.051	9.865	1251.308
Très grandes entreprises			-2.919	4.341
ul de 5 à 10 ans	3.432	47.818	8.697	1829.235
ul de 10 à 15 ans	1.154	35.716	8.727	2051.343
ul de moins de 5 ans	19.709	608.149	17.507	25381.30

Note de lecture : La rentabilité économique augmente la MSE de 36,733%.
Champ : entreprises comptant au moins un salarié et soumises au régime réel de l'impôt sur les BIC dans le périmètre de l'étude.
Source : calculs des auteurs à partir de la base de données ESANE.

Références

[1] Anderson, T. W. et D. A. Darling (1952). "Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes". In : *Annals of Mathematical Statistics* 23.2, p. 193-212. doi : 10.1214/aoms/117772943.

[2] Bartlett, M. S. (1950). "Tests of significance in factor analysis". In : *British Journal of Mathematical and Statistical Psychology* 3.2, p. 77-85. doi : 10.1111/j.2044-8317.1950.tb00285.x.

[3] Breiman, Leo (2000). *Some infinity theory for predictor ensembles*. Technical Report 579. Statistics Dept., University of California, Berkeley.

[4] Brion, Pascal (2011). "ESANE, le dispositif rénové de production des statistiques structurelles d'entreprises". In : *Courrier des statistiques* 130.

[5] Caupin, Vincent et Bertrand Savoye (2012). "Une entreprise dans un DOM Est-ce que cela change la donne ?" In : *AFD, Focales* 15.

[6] Corden, W Max et J Peter Neary (1982). "Booming sector and de-industrialisation in a small open economy". In : *The economic journal* 92.368, p. 825-848.

[7] De Miras, Claude (1988). "L'économie martiniquaise : croissance ou excroissance ?" In : *Revue tiers monde*, p. 365-383.

[8] Dreyer, Antoine et Bertrand Savoye (2013). "Une analyse comparative des entreprises des DOM et de la métropole". In : *Economie et Statistique* 462.1, p. 99-123. doi : 10.3406/estat.2013.10218.

[9] GitHub (juin 2022). *GitHub project webpage*. https://github.com. Archived from the original on 2021-04-01. Retrieved 2016-04-05.

[10] Hair Jr, Joseph F et al. (1998). *Multivariate Data Analysis*. 4th. London : Prentice Hall.

[11] Hastings Jr., Cecil et al. (1947). "Low moments for small samples : a comparative study of order statistics". In : *Annals of Mathematical Statistics* 18.3, p. 413-426.

[12] Ho, Tin Kam (août 1995). "Random Decision Forests". In : *Proceedings of the 3rd International Conference on Document Analysis and Recognition*. IEEE. Montreal, QC, p. 278-282.

[13] Hotelling, Harold (1933). "Analysis of a Complex of Statistical Variables into Principal Components". In : *Journal of Educational Psychology* 24, p. 417-441.

[14] Kaiser, Henry F. (1970). "A second generation Little Jiffy". In : *Psychometrika* 35.4, p. 401-415. doi : 10.1007/BF02291817.

[15] Kaiser, Henry F. et John Rice (1974). "Little Jiffy, Mark IV". In : *Educational and Psychological Measurement* 34, p. 111-117.

[16] Kremp, Elizabeth (1993). "Nettoyage de fichiers dans le cas de données d'entreprises : recherche de la cohérence transversale". In : *Insee Méthodes*.

[17] Liu, Fei Tony, Kai Ming Ting et Zhi-Hua Zhou (2008). "Isolation Forest". In : *2008 Eighth IEEE International Conference on Data Mining*. IEEE, p. 413-422.

[18] Maaten, L.J.P. van der et G.E. Hinton (nov. 2008). "Visualizing High-Dimensional Data Using t-SNE". In : *Journal of Machine Learning Research* 9, p. 2579-2605.

[19] Makarov, Artem et Dmitry Namiot (2023). " " In : *International Journal of Open Information Technologies* 11.10.

[20] Mathouraparsad, Sébastien et Bernard Decaluwé (2018). "Une analyse comparative des économies des Dom à travers les matrices de comptabilité sociale". In : *Économie Régionale et Urbaine* 1, p. 61-90.

[21] Mehoumoud Issop, Zoulfikar (déc. 2016). *Le Syndrome hollandais dans les DOM est-il toujours d'actualité ?* Note d'analyse ATOM, n° 7.

[22] Pearson, K. (1901). "On Lines and Planes of Closest Fit to Systems of Points in Space". In : *Philosophical Magazine* 2.6, p. 559-572.

[23] Poirine, Bernard (1993). "Le développement par la rente dans les petites économies insulaires". In : *Revue économique*, p. 1169-1199.

[24] Shapiro, Samuel Sanford et Martin Bradbury Wilk (1965). "An analysis of variance test for normality (complete samples)". In : *Biometrika* 52.3-4, p. 591-611.

[25] Stevens, James P. (2002). *Applied Multivariate Statistics for the Social Sciences*. University of Cincinnati. Mahwah, NJ : Lawrence Erlbaum Associates.

[26] Sudrie, Olivier (2021). "Une modélisation des trajectoires de croissance à long terme des Outre-mer". In : *Une modélisation des trajectoires de croissance à long terme des Outre-mer*. Sous la dir. de Bertrand Savoye. Éditions AFD, p. 1-51.

[27] Székely, Gabor J. et Maria L. Rizzo (sept. 2005). "Hierarchical Clustering via Joint Between-Within Distances : Extending Ward's Minimum Variance Method". In : *Journal of Classification* 22.2, p. 151-183.
[28] "The Dutch Disease" (nov. 1977). In : *The Economist*, p. 82-83.
[29] VEIGA, Sébastien DA (2023–2024). *Lecture Notes on Machine Learning*. ENSAI - CREST.
[30] Ward, Joe H. (1963). "Hierarchical Grouping to Optimize an Objective Function". In : *Journal of the American Statistical Association* 58.301, p. 236-244.