Journées de méthodologie statistique de l'Insee 2022

# Estimation des montants manquants de versements de TVA : exploitation des données du contrôle fiscal

*Cécile WELTER-MÉDÉE (*), Simon QUANTIN (**)*

*(*) Insee, Département des Études Économiques, Division Marchés et Entreprises, CREST*

*(**) Insee, Département des Études Économiques, Division Marchés et Entreprises*

simon.quantin@insee.fr
cecile.welter-medee@insee.fr

**Mots-clés.** : *Machine Learning*, biais de sélection, écart de TVA, contrôles fiscaux.

**Domaines.** Théorie des sondages aval, *Machine learning* / Apprentissage statistique

## Résumé

L'estimation du manque à gagner de l'administration fiscale dans son activité de recouvrement des impôts est un enjeu important, mais est difficile à effectuer. L'extrapolation à l'ensemble des entreprises redevables d'un impôt en s'appuyant sur des informations issues des contrôles fiscaux, nécessite de prendre en compte le processus qui a conduit à sélectionner des entreprises contrôlées. Or ce processus est particulièrement complexe et non formalisé : il est engendré par un programme de contrôle qui concerne l'ensemble des secteurs d'activité, qui est lui-même issu d'un travail d'expertise fouillé conduit par les contrôleurs fiscaux sur la base de nombreuses informations. En découle un biais de sélection certainement important, qui empêche toute extrapolation hâtive.

Ce travail s'appuie sur les données de gestion du contrôle fiscal transmises par la DGFiP pour estimer les montants manquants de TVA. Nous adoptons une méthodologie en deux étapes (appliquée également par Tagliaferri dans le même contexte d'estimation de l'écart de TVA [26]), appliquée sur une partition des entreprises redevables de la TVA selon l'administration en charge du contrôle afin de prendre en compte la différence structurante d'organisation des contrôles fiscaux.

La première étape, qui s'inspire des méthodes de redressement de la non réponse par repondération, consiste à attribuer à l'ensemble des entreprises faisant une déclaration de TVA l'année considérée, une probabilité de contrôle afin de repondérer leurs poids de sondage pour mener l'extrapolation. Les probabilités de contrôle sont inconnues (comme les probabilités de réponse à une enquête le sont) et sont estimées à partir des informations contenues dans les déclarations de TVA adressées par les entreprises à la DGFiP. Une telle démarche est fréquente dans la correction du biais induit par le phénomène de non-réponse aux enquêtes, par exemple en prenant en compte la probabilité de réponse pour redresser les probabilités d'inclusion (Sarndal 2003 [25],

Deroyon 2018[28]), mais aussi dans les méthodes économétriques d'évaluation comme dans le cas de l'ajustement par pondération de l'inverse du score de propension (Austin 2015[1], Imbens 2003[14]). La correction de la non-réponse par repondération dans les enquêtes s'appuie sur une repondération des probabilités d'inclusion initiales par les probabilités de réponse estimées. Par analogie, on considère ici que les entreprises qui font une déclaration de TVA sont un échantillon (en l'occurrence exhaustif) dont une partie seulement a été contrôlée. On repondère alors chaque entreprise qui réalise une déclaration de TVA par une estimation de sa probabilité d'être contrôlée. Comme il est d'usage en redressement de la non-réponse dans les enquêtes menées auprès des entreprises, des probabilités de contrôle sont estimées à l'aide d'algorithmes de machine learning, de la famille des arbres de classification (les algorithmes sont entraînés sur des échantillons rééquilibrés à l'aide de la méthode SMOTE [4]). Ce type de méthode permet d'intégrer des effets non linéaires des variables explicatives et permet par ailleurs de prendre en compte des caractéristiques structurantes des entreprises, permettant en plus de tenir compte, en plus de leur propension à être contrôlée, de leur hétérogénéité (en termes de taille notamment). Par suite, et comme préconisé par Haziza et Beaumont (Haziza et Beaumont 2007[10]) des « groupes de contrôle homogènes » sont constitués, comme le seraient des groupes de réponse homogène dans le cas du redressement de la non-réponse par repondération, à partir des quantiles de la distribution des probabilités prédites. Les groupes d'entreprises ainsi constitués sont donc considérés comme homogènes au sens de la probabilité d'être contrôlée. Chaque entreprise se voit enfin attribuer la probabilité empirique d'être contrôlée observée dans son groupe de contrôle homogène. Cette dernière étape permet de se départir de la forme fonctionnelle sous-jacente de l'algorithme retenu pour la prédiction des probabilités de contrôle et de limiter l'apparition de points influents.

Dans une seconde étape, les montants de manque à gagner sont extrapolés par un estimateur par le ratio qui est appliqué à des domaines d'entreprises considérées cette fois-ci comme homogènes au sens de la fraude. La partition de la population d'entreprises faisant une déclaration de TVA en domaines permet de tenir compte de l'hétérogénéité des comportements de fraude, au sens des redressements notifiés semblables de par leur motif et/ou le montant éludé. Ce découpage en domaines permet d'extrapoler directement les montants de fraude en leur sein, en s'appuyant sur une hypothèse de comportement de fraude homogène des entreprises qui les constituent. Les déterminants de la fraude ont été identifiés à l'aide de l'estimation d'une régression logistique avec pénalisation LASSO sur la notification d'un montant strictement positif par l'administration fiscale, parmi les seules entreprises contrôlées. Un estimateur classique de la théorie des sondages est enfin utilisé : l'estimateur par le ratio, qui s'appuie sur de l'information auxiliaire corrélée aux montants d'impôts éludés.

Dans l'ensemble, les différentes estimations obtenues à partir des exercices comptables de 2012 semblent assez peu dépendantes des variations méthodologiques adoptées et sont globalement comprises entre 20 et 26 milliards d'euros.

# Abstract

Estimating the revenue loss from tax avoidance is an important issue, but a difficult one. Extrapolating it from tax audits to all taxable companies requires taking into account the process which led to the selection of the firms to be audited. However, this selection process is particularly complex and not formalized : it is the result of an audit program concerning all sectors of activity, and of in-depth expertise work carried out by the tax auditors based on many private informations. This results in a selection bias that is certainly significant, which should prevent from any hasty extrapolation.

This article is focused on the estimation of the French Value Added Tax (VAT) gap on firms

subject to this tax. This exercise is based on the fiscal audits conducted by the fiscal administration. We adopt a two-step methodology that relies on a partition of VAT-taxable firms based on the administration in charge of the control, in order to take into account the structural difference in their organization. The first step is inspired by reweighting methods to correct for the non-response bias. This involves, using machine learning techniques, regrouping VAT-taxable firms that are similar to each other with respect to their probability to be controlled, and use these groups to robustly compute a common probability to be controlled. In a second step, revenue loss from VAT avoidance can then be extrapolated using the group-specific probability to be controlled, and applying a ratio and/or mean estimator to specific sets of firms with similar fraud behavior.

Overall, the various estimates obtained from the 2012 fiscal exercises seem to be relatively little dependent on the methodological variations adopted, generally between 19 and 26 billion euros.

# Introduction

Ce travail s'inscrit dans la suite de celui effectué par l'INSEE à la sollicitation de la Cour des Comptes sur un chiffrage de la fraude à la TVA figurant dans le rapport sur la fraude aux prélèvements obligatoires présenté le 2 décembre 2019. Quantifier la « fraude fiscale » pose inévitablement comme le rappelle le rapport de la Cour des comptes (2019) [5], la question de sa définition. Il est possible de s'appuyer sur la définition de la fraude figurant dans l'article 1741 du code général des impôts : le non respect du droit fiscal. Il existe cependant un certain nombre de pratiques, qui sans être comprises dans le périmètre sus-mentionné, pourraient être considérées comme des activités frauduleuses : c'est par exemple le cas d'une partie de l'évasion fiscale et de l'optimisation fiscale, qui constituent tout de même un évitement fiscal, bien que licite. Intégrer ou non ces comportements d'évitement de l'impôt dans le périmètre de l'étude, peut notablement changer les montants en jeu. Au delà de la définition de la fraude, il est courant de chercher à mesurer l'écart fiscal qui s'entend, pour reprendre la définition de la Cour des Comptes, comme l'écart entre ce qui a été effectivement recouvré et ce qui aurait dû l'être si la loi fiscale avait été respectée. L'écart de TVA mesure ainsi la différence entre le total des recettes de TVA attendues et le montant total effectivement perçu. Il fournit donc une estimation du manque à gagner lié à la fraude et à l'évasion fiscale, mais aussi aux faillites, aux insolvabilités et aux erreurs de bonne foi. L'estimation de ce manque à gagner s'appuie généralement sur une approche macroéconomique faisant appel à des données des comptes nationaux et est menée annuellement par la Commission européenne.

Ce travail propose une approche alternative tant dans le périmètre retenu des « montants manquants de versements de TVA » que dans la méthodologie mise en œuvre. En effet, notre étude s'appuie sur les données individuelles de contrôle fiscal fournies par la DGFiP. En cela, notre travail se concentre donc sur le champ sur lequel s'exerce le contrôle fiscal, c'est-à-dire le non respect du droit fiscal [23] qu'il soit intentionnel ou non. Par ailleurs, en privilégiant l'exploitation de données individuelles (*bottom-up*) au détriment d'une méthodologie macroéconomique (*top-down*), la démarche adoptée permet en théorie d'obtenir une estimation plus robuste mais est plus complexe à mettre en œuvre [6, 19]. Le principe est d'extrapoler à l'ensemble des entreprises redevables de la TVA, les informations collectées sur les seules entreprises contrôlées, en prenant en compte le processus qui a conduit à la sélection des entreprises contrôlées. Or ce processus de sélection est particulièrement difficile à modéliser : il est le résultat à la fois d'une volonté de contrôle de l'ensemble des secteurs d'activité (finalité dissuasive du contrôle fiscal) et d'un travail d'expertise fouillé conduit sur la base de nombreuses informations par les contrôleurs

fiscaux (finalités budgétaire et répressive). C'est notamment du fait de ce second aspect qui mobilise des informations dont nous ne disposons pas dans les données qu'il existe un biais de sélection évident, et certainement important, dont doit nécessairement tenir compte l'estimation proposée.

En effet, les entreprises contrôlées le sont notamment sur la décision des contrôleurs fiscaux dont l'expertise permet de juger de la propension à frauder des entreprises sur la base d'informations dont ils disposent parmi lesquelles, les déclarations fiscales. Ainsi le fait qu'une entreprise soit contrôlée (et donc sélectionnée au sens de notre processus de sélection) est précisément corrélé positivement avec leur propension à frauder. Dans cette situation, on parle de sélection endogène : la sélection étant corrélée au phénomène d'intérêt, une estimation directe de celui-ci sans prise en compte de l'endogénéité de la sélection conduirait à des estimations biaisées. La méthode la plus répandue permettant de prendre en compte l'endogénéité de la sélection est le modèle de Heckman [12, 13]. Mais le principe même de cette méthodologie nécessite de disposer d'une variable dite d'exclusion, qui impacte le processus de sélection, et pas directement le phénomène d'intérêt, afin de corriger du biais d'endogénéité de la sélection. Dans le contexte du contrôle fiscal (et des données à notre disposition), il est difficile d'identifier une variable qui impacte la décision de contrôler une entreprise, mais pas sa propension à frauder. La mise en place de contrôles aléatoires pourrait par exemple permettre de construire une telle variable d'exclusion.

Sans variable d'exclusion à notre disposition, ni source d'aléa dans les données de contrôle qui permettrait de supprimer le biais de sélection endogène lié à l'établissement du programme de contrôle, cet article se propose d'utiliser la théorie des sondages pour proposer une estimation. L'idée est ici d'exploiter au mieux la richesse des données disponibles, et ainsi réduire au maximum le biais de sélection fondé sur des caractéristiques observables. La prise en compte du biais de sélection nécessite cependant de ne pas appliquer directement les méthodes classiques. La démarche adoptée propose une méthodologie en deux étapes [31], appliquée séparément à une partition des entreprises redevables de la TVA selon l'administration en charge du contrôle afin de prendre en compte la différence structurante d'organisation des contrôles. En effet, le contrôle fiscal des différentes entreprises est réparti entre plusieurs administrations de la DGFiP en fonction de leur chiffre d'affaires. La Direction des Vérifications Nationales et Internationales (DVNI) est ainsi en charge du contrôle fiscal des plus grandes entreprises implantées en France et de leurs filiales. Au niveau interrégional, les Directions spécialisées de Contrôle Fiscal (Dircofi) s'occupent des entreprises de taille intermédiaire. Enfin, les directions locales contrôlent les très petites entreprises. Si 10 % environ des entreprises relevant de leur champ de compétence sont contrôlées par la DVNI et les Dircofi, moins de 2 % des petites entreprises sont contrôlées par les directions locales de la DGFiP. Elles sont cependant plus ciblées et plus susceptibles de se voir notifier des montants positifs quand elles sont contrôlées. Ces éléments justifient donc une approche différenciée, bien que similaire, à chaque sous-population d'entreprises.

La première étape de l'estimation consiste à regrouper les entreprises au sein de groupes homogènes en terme de probabilité d'être contrôlée. Notre prise en compte du biais de sélection, dans une première étape, s'inspire des méthodes de repondération, comme dans le redressement de la non réponse dans les enquêtes, ou l'ajustement par l'inverse du score de propension dans les méthodes économétriques d'évaluation. Il s'agit d'estimer pour chaque entreprise redevable de la TVA une probabilité de contrôle. Cette estimation est obtenue dans cette étude par un algorithme de *machine learning* comme dans les estimations menées par [26] sur des données italiennes pour estimer l'écart de TVA ou comme il est d'usage dans le traitement de la non-réponse dans les enquêtes menées par l'Insee auprès des entreprises [28]. L'approche par *machine learning* se justifie par un souci d'aller au-delà de la linéarité fonctionnelle de la corrélation entre les déterminants du contrôle fiscal et la probabilité d'être contrôlé, sans effectuer d'hypothèses fortes sur leurs distributions. Elle permet aussi d'exploiter au mieux la grande quantité d'infor-

mation disponible et ainsi de profiter de la grande taille des bases de données mobilisées. Dans leur étude, Tagliaferri et al. [26] utilisent un algorithme de *Gradient Boosting* [9]. Dans cette étude, nous utilisons également un algorithme de *Gradient Boosting*, mais seulement après avoir comparé ses performances à celles d'autres algorithmes de classification. Une fois la probabilité de contrôle estimée, nous constituons des « groupes d'entreprises de contrôle homogène » [25, 24, 10].

La seconde étape consiste à quantifier le montant total manquant de TVA grâce une estimation par domaines, classique en théorie des sondages. Les domaines sont créés sur la base de variables corrélées au comportement de fraude, permettant de créer des domaines d'entreprises homogènes, au sens de la fraude cette fois-ci. Les montants positifs notifiés sont alors extrapolés de deux façons : par un estimateur par le ratio et par un estimateur par la moyenne, qui exploitent une information auxiliaire connue sur l'ensemble des entreprises redevables de la TVA, le montant de TVA brute déclarée pour l'estimateur par le ratio et le nombre de mois d'activité pour l'estimateur par la moyenne.

La première partie de ce document de travail est consacrée à la description des données de gestion du contrôle fiscal transmises par la DGFiP sur lesquelles s'appuie notre estimation, ainsi qu'à l'organisation du contrôle fiscal et le champ des entreprises retenu. La seconde partie est exclusivement méthodologique : y sont détaillés la méthodologie en deux étapes et notamment la forme des estimateurs retenus et de leurs intervalles de confiance. La troisième partie présente les résultats de chaque étape de l'estimation, depuis les pondérations des entreprises, jusqu'aux montants totaux de TVA manquants estimés. Enfin, la quatrième et dernière partie conclut.

# 1 Données, organisation du contrôle fiscal et périmètre de l'étude

## 1.1 La base de données Alpage de gestion des contrôles fiscaux

La base de données Alpage utilisée dans cette étude [1] est une base de données de gestion de la DGFiP. Elle recense l'ensemble des contrôles fiscaux ayant, à la date de l'extraction, abouti à une notification au contribuable de l'absence ou de la présence d'un redressement entre 2012 et 2018, l'année de notification correspondant à l'année à laquelle le comptable de la DGFiP notifie à l'entreprise redressée le montant dû (ou non). Son exploitation statistique est cependant rendue complexe par les principes qui régissent sa mise à jour.

Tout d'abord, pour chaque entreprise contrôlée, sur une période comptable donnée, par un contrôleur fiscal, est ouvert un dossier dans cette base de gestion. Ainsi, pour une année de notification donnée, pour une même période comptable, une même entreprise contrôlée par plusieurs contrôleurs fiscaux présente autant de dossiers différents dans la base Alpage. Dans ce cas particulier, on fait l'hypothèse qu'il s'agit d'un unique contrôle portant sur l'entreprise considérée [2]. Cela implique toutefois que l'absence d'identifiant de l'entreprise contrôlée (par exemple, en cas d'activité non déclarée) impose de faire l'hypothèse simplificatrice selon laquelle un numéro de

1. Il s'agit d'une extraction du logiciel de gestion qui date du 06/09/2019.

2. D'après la DGFiP, ces cas de figures sont très peu nombreux. Seules les très grosses entreprises peuvent avoir plusieurs numéros de dossier pour un même période contrôlée. Dans les publications budgétaires faites par la DGFiP, le nombre de dossiers traités une année comptabilise distinctement tous ces numéros de dossiers, sans suppression de ces « doublons ». Par ailleurs, il existe aussi des vrais doublons d'enregistrement dans la base Alpage, au sens où pour une même entreprise (SIREN) et une même période contrôlée, plusieurs identifiants dossier peuvent être associés à un montant redressé identique pour un même motif (code Thesaurus). Ces observations multiples ont été supprimées en accord avec la DGFiP.

dossier est associé à un unique contrôle.

De plus, lors de l'enregistrement automatique du dossier, autant de lignes que de motifs de redressement possibles (identifiés par des codes Thesaurus) sont automatiquement créées, sans lien avec les impôts effectivement contrôlés [3]. Dès lors, l'identification, par exemple à partir des codes thesaurus relatifs à la TVA, des entreprises effectivement contrôlées au titre de la TVA n'est pas immédiate. Cette liste des codes thesaurus établie par la Cour des Comptes avec la DGFiP [4] permet en effet de déterminer pour chaque dossier le montant de redressement prononcé à l'encontre de l'entreprise contrôlée [5] sur cet impôt, et donc le nombre de dossiers de contrôles fiscaux ayant conduit à un redressement de TVA [6].

Mais cet enregistrement systématique de multiples motifs de contrôle sans que ceux-ci ne soient nécessairement réalisés ne permet pas d'identifier les entreprises qui auraient effectivement été contrôlées à la TVA, et n'auraient pas été notifiées d'un montant de fraude à la TVA à l'issue du contrôle. Il est donc nécessaire de recourir à une autre information, afin de définir le champ des entreprises contrôlées à la TVA mais non redressées. Il convient alors de s'intéresser à la procédure de contrôle fiscal engagée pour chaque dossier.

## 1.2 Les différents types de procédure de contrôle fiscal

On distingue deux types de contrôles menés par l'administration fiscale auprès des entreprises : le contrôle sur pièces et le contrôle sur place. Le contrôle sur pièces désigne un contrôle fiscal effectué au sein des locaux de la DGFiP. Il consiste en une analyse critique des déclarations faites par le contribuable ainsi qu'en des recoupements avec l'ensemble des autres informations disponibles ou recueillies par l'administration dans le cadre des procédures légales, notamment le droit de communication. Le contrôle sur place ou « contrôle fiscal externe » consiste, lui, en la vérification de la comptabilité des entreprises, en général dans ses locaux, en vue de contrôler la sincérité et l'exactitude de ses déclarations.

Au-delà du lieu où est effectué le contrôle fiscal, il existe plusieurs types de vérification, les trois principales étant la vérification générale de comptabilité et la vérification ponctuelle ou la vérification simple d'un impôt donné. La vérification générale désigne ainsi un ensemble d'opérations ayant pour objet d'examiner sur place la comptabilité d'une entreprise et de la confronter à certaines données de fait ou matérielles afin de vérifier l'exactitude et la sincérité des déclarations souscrites et d'effectuer les rehaussements nécessaires. À l'inverse, une vérification de comptabilité est considérée comme ponctuelle lorsque son champ d'investigation porte soit (i) sur un

3. À l'inverse, pour les quelques inscriptions manuelles dans la base de gestion, seules les lignes effectivement contrôlées sont générées.

4. Sur les 111 codes thesaurus relatifs à la TVA existants, 10 ont été supprimés en accord avec la DGFiP pour limiter les codes relevant des taxes sur le chiffre d'affaires qui étaient possiblement remboursées au contribuable à la suite du contrôle, et donc neutres fiscalement (cf. rapport de la Cour des comptes publié le 2 décembre 2019 et disponible ici : https ://www.ccomptes.fr/fr/publications/la-fraude-aux-prelevements-obligatoires).

5. Celui-ci ne sera pas nécessairement recouvré. Des procédures d'appel, par exemple, peuvent en diminuer l'ampleur.

6. Selon les motifs de redressement, une entreprise peut se voir notifier un montant positif ou négatif. Les montants de redressements négatifs correspondent à des erreurs de déclaration manifestes, bénéficiant à l'entreprise lors du contrôle. Il existe certainement des erreurs de bonne foi des contribuables qui conduisent également à la notification de montants de TVA positifs, probablement de montants équivalents en valeur absolue aux montants négatifs notifiés. Mais n'ayant pas connaissance de l'intentionnalité, il est impossible de les distinguer des montants notifiés suite à des fraudes intentionnelles. Par convention, dans la suite de ce document, une entreprise dite « redressée » se sera donc vue notifiée d'un montant total strictement positif de redressement pour l'exercice comptable considéré.

point de la situation fiscale du contribuable (par exemple le contrôle de certains postes clairement individualisés sur une déclaration tels que les provisions, ou bien le contrôle des opérations ayant concouru au crédit de TVA dont le remboursement est demandé) soit (ii) sur un impôt déterminé sur toute la période non prescrite (au plan administratif, ces vérifications ponctuelles sont comptabilisées dans la catégorie des vérifications dites simples, par exemple : vérification d'un bénéfice non commercial non soumis à la TVA), soit (iii) sur une période plus courte que le délai normal de reprise.

La base de données Alpage ne recense que les résultats des contrôles sur place mais précise le type de vérifications effectuées (simples, ponctuelles ou générales) en distinguant pour les vérifications ponctuelles et simples l'impôt concerné, et notamment celles qui concernent la TVA. Par définition, cette information n'est pas disponible dans le cas d'une vérification générale, qui est par ailleurs le motif de vérification le plus répandu. Il est donc nécessaire dans ce cas de figure d'émettre une hypothèse sur la teneur des impôts et taxes effectivement contrôlés, dès lors que l'entreprise n'a pas fait l'objet d'un redressement fiscal au titre de la TVA. Nous faisons l'hypothèse dans la suite de l'étude qu'une vérification générale conduit au contrôle systématique de la TVA. Cette hypothèse nous permet d'avoir un échantillon d'entreprises contrôlées le plus large possible et plusieurs discussions lors des réunions du groupe d'experts mandatés par la Cour des Comptes ont confirmé sa crédibilité.

***Au final, dans cette étude, le champ des entreprises contrôlées au titre de la TVA regroupe les entreprises contrôlées sur place soit au titre d'une vérification générale soit d'une vérification simple ou ponctuelle de TVA***.

Afin de valider les choix précédents, nous utilisons comme référence les chiffres issus des documents des Finances publiques (évaluations des voies et moyens – Tome 1, les évaluations de recettes, annexe au Projet de loi de finances pour 2019, noté V&M dans la suite de cette note [7]). En effet, retrouver le nombre total d'opérations effectuées nous permet de nous assurer de la définition d'un dossier dans un premier temps ; estimer les montants redressés de TVA devrait alors nous conduire à nous approcher des montants de taxe sur le chiffre d'affaires et assimilés redressés publiés par la DGFiP.

Table 1 – Estimation du nombre de dossiers à partir de la base Alpage

| | Nombre de dossiers | | |
|---|---|---|---|
| Année de notification | Estimé | Voies et Moyens | Écart |
| 2012 | 48 134 | 48 178 | -44 |
| 2013 | 48 132 | 48 219 | -87 |
| 2014 | 47 719 | 47 776 | -57 |
| 2015 | 46 264 | 46 266 | -2 |
| 2016 | 45 256 | 45 314 | -58 |
| 2017 | 44 240 | 44 287 | -47 |

Source : données DGFiP, calculs Insee/Cour des Comptes.

En ce qui concerne le nombre de dossiers, notre exploitation de la base de gestion Alpage aboutit à des résultats très proches de ceux publiés dans V&M (cf. Tableau 1).

7. Le nombre de dossiers et les montants redressés publiés dans V&M sont aussi issus de la base Alpage. Cependant, ils proviennent d'onglets différents de la base de gestion, conduisant à l'existence de menus écarts.

Table 2 – Montants de redressements notifiés par année de notification

| | Montant total des redressements (en M€) | | |
|---|---|---|---|
| Année de notification | Thesaurus TVA | Voies et Moyens | Écart |
| 2012 | 3 139 | 2 987 | +152 |
| 2013 | 2 576 | 2 442 | +134 |
| 2014 | 2 205 | 2 084 | +121 |
| 2015 | 2 022 | 1 961 | +61 |
| 2016 | 2 021 | 1 992 | +29 |
| 2017 | 2 019 | 1 962 | +57 |

Source : données DGFiP, calculs Insee/Cour des Comptes.

Le Tableau 2 présente les résultats obtenus pour les montants de redressement au titre de la TVA notifiés une année donnée. Cette fois encore, les montants totaux calculés à partir de la base Alpage sont proches de ceux publiés dans V&M, bien que systématiquement supérieurs. Dans les deux cas, le montant de redressement total notifié par année décroît entre 2012 et 2017, probablement en partie lié au fait que le nombre de dossiers diminue. Au-delà du montant annuel total, il est d'ores et déjà intéressant de décrire comment évolue la distribution des montants de redressement notifiés chaque année. Celle-ci est globalement similaire d'une année sur l'autre (cf. Tableau 3), on remarque par ailleurs que chaque année, des montants de redressement très élevés, jusqu'à plus de 100 fois plus élevés que le 99ème centile, sont notifiés.

Table 3 – Répartition des montants redressés par année de notification (en milliers d'€)

| Année de notification | Médiane | Moyenne | $90^e$ décile | $99^e$ centile | Maximum |
|---|---|---|---|---|---|
| 2012 | 22 | 104 | 123 | 738 | 811 000 |
| 2013 | 22 | 76 | 130 | 764 | 93 000 |
| 2014 | 23 | 74 | 130 | 729 | 102 000 |
| 2015 | 23 | 70 | 126 | 765 | 19 000 |
| 2016 | 23 | 72 | 126 | 727 | 71 000 |
| 2017 | 22 | 77 | 128 | 728 | 66 000 |

Source : données DGFiP, calculs Insee/Cour des Comptes.

## 1.3 Exercices fiscaux contrôlés et choix de la période considérée

Lors d'un contrôle fiscal, les entreprises sont possiblement contrôlées au titre de plusieurs années (le plus souvent pour 3 années d'exercices comptables), avant de recevoir une notification. Dans le comptage du nombre de contrôles présenté précédemment, l'unité temporelle considérée était l'année de notification, qui est celle présente dans l'ensemble des documents budgétaires. En termes de comportement de fraude, il est en revanche plus naturel de rapporter le redressement constaté à l'année ou aux années de la période contrôlée [8].

8. On notera que la similarité des distributions des montants de redressement prononcés par année de notification évoquée dans la section précédente, suggère, sans le démontrer, que le passage d'une analyse par année de notification à une approche par année d'exercice fiscal contrôlé n'est pas impacté sur cette période par une évolution temporelle des pratiques de contrôle, notamment en termes de redressement.

La présence de plusieurs contrôles fiscaux notifiés la même année[9] (par exemple, si une vérification simple de TVA sur une période comptable donnée est suivie pour la même période ou pour une période en partie concomitante d'une vérification générale), nous conduit tout d'abord à allouer le montant total de redressement à une période d'exercice factice, débutant à la date de début de contrôle la plus précoce (parmi tous les contrôles relatifs à l'impôt considéré, qui ont conduit à une notification de redressement l'année considérée), et se terminant à la date de fin de contrôle la plus récente (idem). Ainsi, les périodes contrôlées se trouvent mécaniquement étendues relativement à ce qui est présent dans la base Alpage.

Par ailleurs, analyser les redressements prononcés pour une année civile donnée est plus pertinent pour les premières années de notification recensées dans la base Alpage utilisée. En effet, il est probable que celle-ci recense, dans la version utilisée dans cette étude, quasi exhaustivement les contrôles portant sur ces exercices fiscaux. Pour les années plus récentes, des contrôles sont en cours, et d'autres seront décidés plus tard ; le total des contrôles qui ont eu lieu au titre de ces exercices ne reflète pas le nombre total de contrôles qui auront porté sur ces années. Ainsi, le tableau 4 ci-dessous révèle que le nombre de contrôles figurant dans la base Alpage relatif à une année civile donnée est le plus élevé entre 2011 et 2014 (cf. Nb. dossiers contrôlés). ***Notre étude portera donc sur l'exercice fiscal de l'année 2012*** : c'est sur cette année d'exercice fiscal que l'on compte dans la base Alpage dont on dispose le plus grand nombre de contrôles fiscaux.

TABLE 4 – Répartition des contrôles selon leur année de notification, par année comptable contrôlée dans ces dossiers

| Exercices contrôlés | Année de notification (en %) | | | | | | | Nb. dossiers contrôlés |
|---|---|---|---|---|---|---|---|---|
| | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | |
| 2008 | **72** | 20 | 4 | 2 | 1 | 1 | 1 | 34 647 |
| 2009 | **54** | **33** | 9 | 2 | 1 | 0 | 0 | 75 487 |
| 2010 | **35** | **35** | **21** | 6 | 1 | 1 | 0 | 118 085 |
| 2011 | **17** | **30** | **30** | **18** | 5 | 1 | 1 | 140 285 |
| 2012 | 3 | **16** | **30** | **29** | **17** | 5 | 1 | 140 471 |
| 2013 | 0 | 2 | **16** | **30** | **30** | **18** | 5 | 135 479 |
| 2014 | 0 | 0 | 2 | **16** | **31** | **31** | **19** | 126 423 |
| 2015 | 0 | 0 | 0 | 3 | **20** | **38** | **39** | 101 576 |

Note : les pourcentages supérieurs à 10 % sont en gras.
Lecture : Dans la base Alpage disponible, 35 % des 118 085 dossiers qui concernent l'exercice comptable 2010 ont fait l'objet d'une notification de redressement en 2012.
Source : données DGFiP, calculs Insee/Cour des Comptes.

## 1.4 Déterminer le champ des entreprises redevables de la TVA

La TVA se caractérise essentiellement comme un impôt général sur la consommation qui s'applique aux livraisons de biens et prestations de services situées en France. L'assujettissement à la taxe est déterminé par la nature des opérations effectuées ou des produits concernés, indépendamment de la situation personnelle de l'assujetti ou de son client. Dès lors, une entreprise est « redevable de la TVA » en tant qu'assujettie qui réalise une opération imposable à la TVA. Il n'est pas aisé de déterminer dans plusieurs secteurs d'activité si la nature des opérations réalisées est imposable à la TVA, et une démarche usuelle consiste à définir le champ des entreprises redevables de la TVA à partir de leurs obligations déclaratives auprès des services fiscaux. Le

9. conduisant ou non à un redressement de TVA.

reversement de la TVA s'effectue en effet à l'aide de déclarations dont la forme et la fréquence dépendent du régime d'imposition.

Tout d'abord, l'article 293 B du code général des impôts (CGI) institue une franchise en base de taxe sur la valeur ajoutée pour les petites entreprises. Plus précisément, en deçà de seuils de chiffre d'affaires (actualisés tous les trois ans et différents selon la nature de l'opération), les activités de livraisons de biens et les prestations de services (principalement) sont exonérées de déclaration et de TVA. Cette franchise en base de TVA n'est cependant pas obligatoire et résulte d'un choix de l'entreprise. Le régime simplifié d'imposition (RSI) concerne les entreprises dont le chiffre d'affaires hors taxe est plus élevé que ceux qui délimitent les seuils de franchise de TVA [10]. Ce régime impose de payer deux acomptes en juillet et en décembre de chaque année et d'adresser une déclaration comptable CA12 récapitulant l'ensemble des opérations imposables de l'année civile précédente ; la TVA correspondante due s'entendant sous déduction des acomptes déjà versés. Pour les chiffres d'affaires encore plus élevés [11], les entreprises sont imposées sous le régime réel normal (RN). Ce régime impose de télétransmettre une déclaration CA3 chaque mois qui renseigne de la TVA due au cours du mois précédent, ou lorsque la TVA est inférieure à 4000 € par an une déclaration trimestrielle.

***Les entreprises redevables de la TVA retenues dans cette étude sont celles ayant effectué au moins une déclaration de TVA l'année civile considérée***, que le montant déclaré des opérations imposables soit nul ou non. Par définition, ce champ ne retient donc pas les entreprises sous le régime de la franchise de TVA. Il ne retient pas non plus celles qui ne seraient pas éligibles à la franchise de TVA et qui n'auraient pour autant pas rempli de déclarations de TVA, ce qui constitue précisément un motif de redressement fiscal.

Les données de TVA utilisées dans cette étude sont des fichiers annuels agrégeant les déclarations effectuées par une entreprise au cours d'une année civile. Même si les informations disponibles dépendent du type de déclaration adressée (CA3 ou CA12), plusieurs éléments à déclarer sont communs aux deux régimes d'imposition, RSI et RN, parmi lesquels, le montant total de TVA brute déclarée, le montant total de TVA déductible, la demande d'un report de crédit de TVA, une partition de la TVA brute en fonction des taux d'imposition ou de la nature des opérations (livraisons intra-communautaires, TVA sur immobilisations. . . ) entre autres. Outre la définition du champ des entreprises redevables, ces données seront utilisées pour estimer la probabilité d'être contrôlée par la DGFiP. Nous détaillerons dans la partie méthodologique consacrée à l'estimation de la probabilité d'être contrôlée (section 2.1.1) les caractéristiques retenues.

## 1.5 L'organisation administrative du contrôle fiscal

Le contrôle fiscal est assuré par trois niveaux de contrôle - national, interrégional, et local - correspondant à une segmentation du tissu fiscal des entreprises (grandes, moyennes, petites). Au niveau national, la Direction des Vérifications Nationales et Internationales (DVNI) contrôle tous les impôts, droits et taxes dûs par les grandes entreprises nationales et internationales, réalisant un chiffre d'affaires supérieur à 152,4 M€ pour les ventes et 76,2 M€ pour les prestations de services (ou dont l'actif brut est supérieur à 400 M€) ainsi que par leurs filiales. Au niveau interrégional, les Directions spécialisées de contrôle fiscal (Dircofi) sont spécialisées dans

10. Au moment de la rédaction de cet article, le chiffre d'affaires hors taxe devait être compris entre 85800 et 818000 € pour les activités de vente et de prestation de logement, et compris entre 34400 et 247000 € pour les activités de prestations de services.

11. C'est-à-dire au moment de la rédaction de cet article, le chiffre d'affaires hors taxe devait être supérieur à 818000 € pour les activités de vente et de prestation de logement, et à 247000 € pour les activités de prestations de services.

le contrôle fiscal des entreprises de taille moyenne relevant de leur ressort territorial. Elles ont en charge les entreprises dont le chiffre d'affaires est compris entre 1,5 M€ et 152,4 M€ pour les ventes, et entre 0,5 M€ et 76,2 M€ pour les services. Avant la réforme territoriale qui a abouti au passage à 13 régions [12], on dénombrait neuf Dircofi [13] dont le périmètre géographique d'intervention dépendait de la région ou du département du siège de l'entreprise. Les Dircofi comportent de dix à trente deux brigades [14] réparties par secteur(s) d'activité. Enfin, au niveau départemental, le contrôle fiscal des petites entreprises (chiffre d'affaires inférieur à 1,5 M€ pour les ventes et 0,5 M€ pour les prestations de services) est assuré par les directions locales (Directions Départementales des Finances Publiques et Directions Régionales des Finances Publiques). Celles-ci disposent chacune d'une à dix brigade(s) spécialisée(s) par secteur d'activité.

TABLE 5 – Entreprises contrôlées et redressées sur leur exercice comptable de l'année 2012, par direction fiscale

| | Nb. d'ent. redevables | Contrôlées (%) | Redressées (%) | Redressées Montant moyen (€) |
|---|---|---|---|---|
| DVNI | 92416 | 13,1 | 41,8 | 2689 |
| Dircofi | 226257 | 13,3 | 58,3 | 1505 |
| Directions locales | 3141081 | 2,1 | 68,0 | 1183 |

Note : Les entreprises contrôlées correspondent aux entreprises redevables de la TVA dont au moins un mois de l'exercice comptable 2012 a été contrôlé. La proportion d'entreprises redressées s'entend parmi les entreprises contrôlées. Le montant (moyen) de redressement prononcé correspond à un montant *par mois contrôlé* prononcé à l'encontre des entreprises contrôlées.
Champ : Entreprises redevables de la TVA en 2012.
Source : DGFiP, Insee, calcul des auteurs

Le tableau 5 décrit, pour les entreprises redevables de la TVA, la proportion d'entre elles qui sont contrôlées et/ou redressées au titre de l'année 2012, en distinguant les directions dont elles dépendent (directions nationale, interrégionales et locales). Les entreprises redevables sont déterminées à partir des déclarations fiscales adressées à la DGFiP en 2012. Pour distinguer le chiffre d'affaires réalisé par des ventes et des prestations de service mais aussi pour déterminer la valeur de l'actif brut, nécessaires à l'affectation à l'un des trois niveaux d'administration fiscale en charge du contrôle, nous exploitons les liasses fiscales (fichier Fare de l'Insee) sur la même année. Enfin, le fichier des liaisons financières de l'Insee permet de déterminer les filiales des groupes et ainsi les affecter aux groupes dont le contrôle fiscal est assuré par la DVNI. Au total, en 2012, nous dénombrons près de 3,5 millions d'entreprises redevables de la TVA. De par leur nombre particulièrement élevé, seules 2,1 % des entreprises dont le contrôle fiscal est assuré par des directions locales sont contrôlées, alors que 13 % environ des entreprises dépendant d'une Dircofi ou de la DVNI le sont. Cependant, le contrôle fiscal se révèle plus ciblé au sein des directions locales, puisque 68 % des entreprises contrôlées font l'objet d'un redressement contre 58 % lorsque le contrôle est réalisé par une Dircofi et 42 % dans le cas de la DVNI. Enfin, les montants de redressement prononcés sont bien évidemment plus importants lorsque le chiffre d'affaires est élevé, mais plus faible en proportion de la TVA déclarée, comme le montre le tableau 6. Ainsi,

12. Réforme territoriale promulguée le 7 août 2015 ; passage de 22 à 13 régions en janvier 2016. C'est principalement l'organisation du contrôle fiscal avant réforme qui nous intéresse, puisque 78 % des contrôles sur l'exercice fiscal 2012 retenu pour cette étude, ont été menés avant 2016.

13. Centre, Est, Ouest, Nord, Sud-Est, Sud-Ouest, Sud Pyrénées, Rhône-Alpes-Bourgogne et Île-de-France.

14. en fonction de la zone géographique couverte.

pour les entreprises redevables de la TVA en 2012 et contrôlées sur cet exercice comptable par les directions locales, 2,4 % ne déclarent pas de TVA brute et sont redressées alors qu'elles sont moins de 1 % parmi les entreprises contrôlées par les Dircofi ou la DVNI. De plus, le montant du redressement prononcé à l'encontre d'une entreprise contrôlée par une direction locale représente dans 50 % des cas plus de 18 % du montant de la TVA brute déclarée alors qu'il n'excède pas 14 % (respectivement 23 %) du montant de la TVA brute déclarée dans 90 % des redressements prononcés par la DVNI (respectivement par les Dircofi).

TABLE 6 – Distribution du taux de redressement des entreprises contrôlées sur leur exercice comptable de l'année 2012, par direction fiscale

| | Redressées avec TVA= 0 (%) | Taux de redressement (en %) Q1 | Médiane | Q3 | P90 |
|---|---|---|---|---|---|
| DVNI | 0,3 | 0,1 | 0,5 | 2,8 | 13,8 |
| Dircofi | 0,2 | 0,5 | 1,9 | 7,2 | 22,9 |
| Directions locales | 2,4 | 5,0 | 17,9 | 64,2 | 227,4 |

Note : Les entreprises contrôlées correspondent aux entreprises redevables de la TVA dont au moins un mois de l'exercice comptable 2012 a été contrôlé. Le taux de redressement correspond au montant de redressement prononcé rapporté à la TVA brute déclarée sur la période contrôlée. Il n'est donc défini que pour les entreprises contrôlées ayant déclaré un montant de TVA brute strictement positif.
Champ : Entreprises redevables de la TVA en 2012 et contrôlées sur cet exercice comptable.
Source : DGFiP, Insee, calcul des auteurs

Ces premiers résultats révèlent ainsi des différences notables tant dans la probabilité d'être contrôlée que dans la nature du comportement de fraude ayant conduit à un redressement. ***Il apparaît donc nécessaire de procéder à une estimation séparée pour les entreprises relevant de la DVNI, des Dircofi et des directions locales***.

# 2 Méthodologie

La méthodologie adoptée dans cette étude s'appuie sur la théorie des sondages. On considère les entreprises contrôlées comme les répondants à une enquête à partir desquels il est possible d'estimer les montants de TVA non recouvrés pour l'ensemble des entreprises redevables de la TVA, de la même façon que l'on estime le total d'une grandeur économique pour l'ensemble d'une population à partir des seuls répondants à l'enquête.

Le processus conduisant à l'établissement du programme de contrôle de l'administration fiscale est complexe. Il est effectivement décidé à partir d'informations variées (existence d'un contrôle récent, données de déclaration fiscale, informations locales, etc.) et s'appuie aussi sur l'expertise des contrôleurs, difficile à formaliser et à généraliser. L'administration fiscale cible les entreprises pour lesquelles les suspicions de fraude sont fortes, mais elle cherche aussi dans la mesure du possible à recouvrer des montants importants non perçus. Les contrôles dépendent enfin des moyens humains dévolus au contrôle fiscal sur la période considérée. Les probabilités de contrôle sont inconnues (comme les probabilités de réponse à une enquête le sont) et on se propose de les estimer à partir des informations contenues dans les déclarations de TVA adressées par les entreprises à la DGFiP. Une telle démarche est fréquente dans la correction du biais induit par le phénomène de non-réponse aux enquêtes, par exemple en prenant en compte la probabilité de réponse pour redresser les probabilités d'inclusion Sarndal2003, Deroyon2018, mais aussi dans les méthodes économétriques d'évaluation comme dans le cas de l'ajustement par pondération

de l'inverse du score de propension Austin2015,Imbens2003.

La correction de la non-réponse par repondération dans les enquêtes s'appuie sur une repondération des probabilités d'inclusion initiales par les probabilités de réponse estimées. Par analogie, on considère ici que les entreprises redevables de la TVA sont un échantillon (en l'occurrence exhaustif) dont une partie seulement a été contrôlée. On repondère alors chaque entreprise redevable de par une estimation de sa probabilité d'être contrôlée.

Au-delà de l'hypothèse forte que la probabilité d'être contrôlée ne dépend que des informations contenues dans les déclarations de TVA, la pertinence de la démarche suppose aussi d'obtenir sous cette hypothèse une estimation convergente à partir des algorithmes utilisés. Pour ce faire, nous appliquons la démarche proposée par Haziza et Beaumont [10] dans le redressement de la non-réponse. Les probabilités prédites ne sont pas exploitées telles qu'elles pour estimer les pondérations. Nous construisons des « groupes de contrôle homogènes », c'est-à-dire des groupes d'entreprises homogènes vis-à-vis de la probabilité d'être contrôlée. Concrètement, les groupes de contrôle homogènes de cette étude sont construits à partir des quantiles de la distribution des probabilités prédites. Chaque entreprise se voit alors attribuer la probabilité empirique d'être contrôlée observée dans son groupe de contrôle homogène.

Une fois estimées les pondérations, nous réalisons des estimations par domaines afin de tenir compte de l'hétérogénéité des comportements de fraude, au sens des redressements notifiés semblables de par leur motif et/ou le montant éludé. Ce découpage en domaines permet d'extrapoler directement les montants de fraude en leur sein, en s'appuyant sur une hypothèse de comportement de fraude homogène des entreprises qui les constituent. Deux estimateurs classiques de la théorie des sondages sont utilisés : l'estimateur par le ratio et l'estimateur par la moyenne. Afin de s'abstraire de la durée des contrôles, ceux-ci sont calculés sur la base d'informations mensualisées issues des contrôles, puis extrapolés en tenant compte de la durée réelle des exercices comptables des entreprises redevables de la TVA. Enfin, les estimations par domaine sont ensuite agrégées pour obtenir une estimation du montant total de TVA manquant.

Nous revenons dans les sous-parties suivantes plus en détail sur la méthodologie. Tout d'abord, nous précisons la pertinence et la construction des groupes de contrôle homogènes, puis notre estimation des probabilités d'inclusion à partir de la probabilité d'être contrôlée, avant d'expliciter la méthode de calibrage ainsi que les mesures de performance des algorithmes de *machine learning* utilisés, qui conduiront à la sélection d'un algorithme pour estimer la probabilité d'être contrôlée (parties 2.1.1 à 2.1.3). Nous détaillons ensuite comment sont effectuées nos estimations à partir de l'échantillon des entreprises contrôlées et des probabilités d'inclusion estimées. La partie 2.2.1 explicite ainsi le choix et la constitution des domaines, puis les parties 2.2.2 à 2.2.5 détaillent formellement les estimateurs associés et les intervalles de confiance correspondants. Enfin, une dernière partie (2.3) discute de la prise en compte des redressements d'ampleur exceptionnelle dans l'extrapolation.

Cette méthodologie est appliquée de façon analogue aux trois sous-populations correspondant aux trois sous-ensembles d'entreprises dont le contrôle dépend de chacun des trois niveaux d'administrations décrits précédemment. Cette implémentation distincte d'une méthodologie commune permet la mise en évidence de déterminants propres à chacune de ces sous-populations, tout en garantissant la lisibilité de la méthode en appliquant une approche identique.

## 2.1 Estimation des pondérations par groupes de contrôle homogènes

Comme le soulignent Haziza et Beaumont [10], s'appuyer sur le seul échantillon des entreprises contrôlées pour obtenir un estimateur convergent de la pondération expose au risque de surapprentissage. Les probabilités prédites d'être contrôlée risquent d'être trop adaptées aux seules entreprises effectivement contrôlées et non à la réelle probabilité de contrôle sous-jacente. Pour contourner cette difficulté, il est d'usage de créer des groupes de contrôle homogènes (GCH), au sein desquels les entreprises ont des probabilités d'être contrôlée par l'administration fiscale semblables, mais qu'on peut considérer comme indépendantes.

Les GCH sont ici déterminés à partir des probabilités d'être contrôlée prédites par des méthodes de *machine learning* basées sur des arbres de décision [15]. Ces méthodes de classification découpent par étapes successives l'échantillon considéré en groupes, sur la base des variables auxiliaires les plus corrélées au fait d'être contrôlée, par ordre d'intensité de la corrélation, et tant que les groupes obtenus sont de taille suffisante. Ces méthodes permettent également d'intégrer de façon sous-jacente des facteurs économiques importants tels que la taille de l'entreprise, permettant de retracer au moins partiellement l'hétérogénéité de nature des entreprises françaises, en plus de leur propension à être contrôlées par l'administration fiscale. L'utilisation d'algorithmes de *machine learning* apparaît dans notre cas particulièrement pertinente [16]. En effet, la décision de mener un contrôle sur une entreprise est prise sur la base d'un programme de contrôle qui est partiellement alimenté par un algorithme de détection de situations susceptibles d'être liées à un comportement de fraude, ainsi que sur l'expertise des contrôleurs fiscaux. Ces derniers sont en mesure de lister des déterminants du contrôle, mais insistent sur la difficulté de formaliser une règle de décision systématique. Faire appel à des méthodes de *machine learning* apparaît donc préférable à des approches économétriques plus traditionnelles afin de tenir compte au mieux de la complexité du processus de sélection, sans que ses déterminants ne soient interprétés ou même explicités.

La convergence asymptotique de l'estimateur utilisé ensuite dépend de la crédibilité, pour la population considérée, des hypothèses (i) d'indépendance des contrôles et (ii) d'une probabilité de contrôle sous-jacente identique au sein de chaque GCH. Parmi les entreprises dont le contrôle dépend de la DVNI ou des Dircofi, les contrôleurs qui se répartissent les contrôles sont assez nombreux pour contrôler quelques centaines de milliers d'entreprises (plus de 12 % des entreprises contrôlées en 2012, cf Tableau 5), mais les contrôles sont relativement peu ciblés (à peu près la moitié des entreprises contrôlées se voient notifier un montant positif de TVA) ; dans ce cas, il peut être crédible de considérer les probabilités de contrôle des entreprises comme semblables, mais plus discutable de les considérer comme indépendantes : vu le relativement faible nombre d'entreprises, et les moyens limités, le fait qu'un contrôleur contrôle une entreprise impacte directement et négativement la probabilité qu'il contrôle une entreprise analogue, faute de temps. À l'inverse, parmi les entreprises contrôlées par les directions locales, (à peine plus de 1 % des entreprises sont contrôlées en 2012, cf Tableau 5), le très grand nombre d'entreprises redevables de la TVA rapporté au nombre limité de contrôleurs rend l'hypothèse d'indépendance des contrôles crédible, mais leur similitude au sein d'un groupe de contrôle homogène plus discutable du fait que les contrôles sont mieux ciblés par l'administration fiscale.

15. En pratique, il est d'usage dans le cadre du redressement de la non-réponse par repondération dans les enquêtes auprès des entreprises de déterminer les groupes de réponse homogènes par des méthodes de *machine learning* basées sur des arbres de décision.

16. Même si l'utilisation d'algorithmes de *machine learning* est appropriée à notre problématique, elle ne permet en revanche pas de traiter des problèmes d'hétérogénéité inobservée.

La méthode des groupes de contrôle homogènes est cependant considérée comme relativement robuste en pratique. En effet, l'estimateur corrigé est approximativement sans biais même si les hypothèses sur lesquelles repose la méthode ne sont pas complètement valides. Il est possible de montrer que le biais de l'estimateur obtenu avec des GCH est nul si la corrélation entre la variable d'intérêt dont on estime le total et la probabilité de contrôle des unités est nulle dans chaque groupe. Il est également important de noter que chaque groupe doit contenir suffisamment d'unités (entreprises contrôlées et non contrôlées) pour que la probabilité de contrôle commune soit estimée correctement. Il n'existe pas de règle autre qu'empirique concernant la taille minimale des groupes : il est recommandé que chaque groupe contienne au moins 100 unités, et d'éviter dans tous les cas les groupes contenant moins de 50 unités.

Nous allons donc faire appel à des méthodes de *machine learning* plutôt qu'à des méthodes économétriques pour prédire la probabilité de contrôle, celle-ci ayant besoin d'être prédite au mieux, mais sans que les déterminants de la prédiction ne soient interprétés ou même connus. De cette façon, les probabilités de contrôle des entreprises seront déterminées de façon fine (plusieurs centaines de GCH selon les algorithmes) et indépendamment des domaines qui seront constitués ensuite afin de recréer des ensembles d'entreprises homogènes en termes de comportement de fraude à partir desquels seront extrapolés les montants de redressement notifiés.

### 2.1.1 Algorithmes d'apprentissage supervisé

Plusieurs algorithmes d'apprentissage supervisé seront implémentés dans chaque sous-population, et leurs performances comparées, afin de retenir le plus approprié [17]. Classiquement, chaque algorithme est d'abord entraîné sur un sous-échantillon de la population totale, appelé échantillon d'entraînement, et ses paramètres sont calibrés par validation croisée. Les performances de l'algorithme sont ensuite évaluées sur le reste de la population - où échantillon test. Dans notre étude, les sous-échantillons d'entraînement de chacune des trois sous-populations regroupent 60 % des entreprises de chaque sous-population considérée (entreprises contrôlées par la DVNI, les Dircofi et les directions locales). Il en résulte que les échantillons test comportent chacun 40 % des sous-populations étudiées. Pour chaque algorithme, les paramètres sont calibrés par validation croisée ; pour ce faire, le sous-échantillon d'entraînement est partitionné aléatoirement en dix sous-ensembles de taille identique. Pour une séquence de valeurs du paramètre à calibrer, l'algorithme est entraîné séparément sur une combinaison de neuf sous-ensembles, et ses performances évaluées sur le dixième sous-ensemble restant [18]. Pour chaque valeur d'un paramètre à calibrer, on retient sa performance moyenne sur les dix séquences d'entraînement-validation, puis on retient la valeur du paramètre qui conduit à la meilleure performance moyenne. Les performances (basées sur plusieurs indicateurs cette fois-ci) de l'algorithme ainsi calibré sur le sous-échantillon d'entraînement sont ensuite évaluées sur l'échantillon test.

Plusieurs algorithmes d'apprentissage sont testés séparément sur chaque sous-population : un algorithme d'apprentissage par arbre de classification, des méta-algorithmes de *bagging* d'arbres de classifications, de forêts aléatoires et de *boosting*, ainsi qu'une régression pénalisée elastic-net (cf. Annexe 4 pour une présentation succincte de chaque algorithme). Les variables retenues comme déterminants du contrôle fiscal sont principalement des ratios comptables issus des déclarations de TVA. Comme le soulignent Haziza et al. [11], il convient en effet à cette étape de privilégier des variables explicatives disponibles pour les entreprises contrôlées et non contrôlées,

17. L'ensemble de ces algorithmes a été implémenté à l'aide du package Caret du logiciel R, [15].

18. Le critère de performance retenu pour le choix d'un algorithme est l'aire sous la courbe ROC (*Receiver Operating Characteristic*). Une courbe ROC est un graphique représentant les performances d'un modèle de classification pour tous les seuils de classification. Plus précisément, cette courbe trace le taux de vrais positifs en fonction du taux de faux positifs, en fonction des seuils de classification.

mais aussi corrélées au fait d'être contrôlée et aux montants éventuels de redressement. Dans notre étude, les principaux déterminants du contrôle fiscal retenus sont :

— la TVA brute totale déclarée,
— les acquisitions intracommunautaires rapportées à la TVA brute totale,
— les livraisons intracommunautaires rapportées à la TVA brute totale,
— la TVA déductible rapportée à la TVA brute totale,
— le chiffre d'affaires réalisé à l'exportation rapporté à la TVA brute totale,
— la TVA déductible sur les biens constituant des immobilisations rapportée à la TVA déductible,
— les crédits antérieurs non imputés et non remboursés rapportés à la TVA déductible,
— la TVA à déduire "autre" [19] rapportée à la TVA déductible.

Plusieurs années de déclaration sont retenues pour ces différentes grandeurs : 2011, 2012 et 2013, soit l'année considérée pour notre estimation (2012) ainsi que les années précédente et suivante [20]. En ce qui concerne le montant de TVA brute déclaré la variable retenue correspond à un montant mensuel moyen sur chaque exercice comptable considéré, c'est à dire le montant total déclaré sur l'année (qui est le montant disponible dans nos bases de données issues de la DGFiP) rapporté à la durée de l'exercice comptable déclaré correspondant. Puis, pour chaque entreprise et pour chaque information comptable, la valeur moyenne déclarée sur la période 2011-2013 est retenue comme variable explicative dans les algorithmes testés. Toutes ces variables sont discrétisées à partir de leur distribution dans l'échantillon test. Plus précisément, pour chaque variable, une modalité regroupe les déclarations d'un montant nul de TVA brute totale et les valeurs manquantes si le ratio correspondant n'est pas défini. Les autres modalités sont définies à partir des quantiles de chaque distribution. Enfin, sont aussi intégrés aux algorithmes le secteur d'activité de l'entreprise (sur 21 positions) et sa catégorie juridique, et pour les entreprises dont le contrôle fiscal dépend des directions locales le régime d'imposition en distinguant le régime réel normal et le régime simplifié [21].

### 2.1.2 Problème d'échantillon déséquilibré : application de SMOTE (Synthetic Minority Over-sampling TEchnique)

Les qualités prédictives d'un algorithme sont cependant souvent réduites lorsque le nombre d'évènements étudiés, ici le contrôle, est rare dans la population considérée. Le nombre d'entreprises contrôlées au titre de la TVA est en effet faible au regard du nombre d'entreprises redevables, en particulier dans le cas des entreprises dont le contrôle fiscal dépend des directions locales. Les entreprises contrôlées constituent ainsi une *classe minoritaire*, au contraire des entreprises redevables de la TVA mais non contrôlées qui sont *majoritaires*. Il est d'usage d'utiliser dans ce cas de figure des techniques de rééquilibrage de l'échantillon d'entraînement, afin de rendre les déterminants du contrôle d'une entreprise plus détectables dans les données.

Les principales techniques de rééquilibrage d'échantillon sont le sur-échantillonnage de la classe minoritaire, et le sous-échantillonnage de la classe majoritaire. Le sur-échantillonnage consiste à accroître le nombre d'observations d'entreprises de la classe minoritaire (c'est-à-dire le nombre d'entreprises contrôlées dans notre cas), par un tirage aléatoire avec remise parmi les entreprises de cette classe. De façon analogue, le sous-échantillonnage de la classe majoritaire (dans notre cas les entreprises redevables de la TVA qui ne sont pas contrôlées) consiste en le

19. dont régularisation sur de la TVA collectée ou déductible.

20. L'extension de la période de déclaration considérée (par exemple en incluant aussi les déclarations de 2010 et 2014) ne modifie pas les résultats.

21. Les entreprises au régime de la franchise en base de TVA n'effectuent pas de déclaration et ne sont pas incluses dans la population d'entreprises considérées dans cette étude.

retrait aléatoire d'entreprises de cette classe, afin de réduire sa taille relativement à celle de la classe minoritaire. Sous-échantillonner la classe majoritaire permet souvent d'obtenir des algorithmes plus performants qu'en sur-échantillonnant la classe minoritaire (cf [4] pour une revue de littérature sur le sujet), mais présente l'inconvénient d'écarter une partie de l'information disponible. Par ailleurs, sur-échantillonner la classe minoritaire par tirage aléatoire avec remise conduit à répéter strictement des observations identiques, ce qui accroît mécaniquement le risque de surapprentissage. [22]

Afin de contourner cette difficulté, Chawla et al. [4] ont développé une méthode de sur-échantillonnage qui consiste à accroître le nombre d'observations de la classe minoritaire, non par un simple tirage aléatoire avec remise mais à partir d'entreprises *synthétiques*, c'est-à-dire d'observations créées à partir des entreprises contrôlées existantes par combinaison linéaire des caractéristiques des *k-plus-proches-voisins*. Contrairement au sur-échantillonnage, cela permet de donner d'autres contours à la région de positifs à prédire : elle est plus large, contient des individus (réels et synthétiques) mieux répartis dans la région à prédire, et permet ainsi une meilleure généralisation. Il s'agit de l'algorithme SMOTE [23]. Il donne à l'utilisateur la possibilité de sous-échantillonner la classe majoritaire, tout en créant des observations synthétiques pour la classe minoritaire : dans un soucis d'implémentation et pour ne pas avoir des temps de calculs trop longs, nous avons fait le choix de combiner les deux. Chaque échantillon d'entraînement sera rééquilibré avant de calibrer les paramètres des algorithmes testés.

### 2.1.3 Mesures d'ajustement adaptées à un échantillon déséquilibré

Une fois l'algorithme entraîné, ses performances sont calculées à partir de l'échantillon test correspondant à la population étudiée. Classiquement, dans le cadre d'une classification d'un évènement binaire, celles-ci sont mesurées en s'appuyant sur les mesures d'ajustement associées à la courbe ROC. En effet, chaque algorithme testé attribue une probabilité d'être contrôlée aux entreprises de l'échantillon test. Une règle de décision (une probabilité de survenue supérieure à 50 % par exemple) permet d'associer à chaque entreprise une prédiction binaire (oui ou non) sur un éventuel contrôle fiscal. Dès lors, les performances d'un algorithme sont issues de la comparaison des entreprises qui ont été effectivement contrôlées (« Positif ») et celles qui ne l'ont pas été (« Négatif »), avec leurs prédictions respectives par l'algorithme. On distingue alors les faux positifs (c'est-à-dire les entreprises prédites comme contrôlées, à tort) et les faux négatifs (c'est-à-dire les entreprises prédites comme non contrôlées, à tort) des prédictions correctes (vrais positif et négatif) au sein de chaque groupe.

22. On parle de surapprentissage lorsqu'un modèle se construit si précisément sur les données d'apprentissage qu'il ne se généralise pas correctement sur l'échantillon test. Par ailleurs, utiliser simultanément un sous-échantillonnage de la classe majoritaire et un sur-échantillonnage de la classe minoritaire ne semble pas améliorer les performances d'un simple sous-échantillonnage de la classe majoritaire.

23. L'algorithme SMOTE est disponible sous le logiciel R dans le package *DMwR* "Data Mining with R" [29].

TABLE 7 – Matrice de confusion

| | **Négatifs prédits** | **Positifs prédits** |
|---|---|---|
| **Négatifs observés** | VN (*Vrais Négatifs*) | FP (*Faux Positifs*) |
| **Positifs observés** | FN (*Faux Négatifs*) | VP (*Vrais Positifs*) |

Plusieurs mesures d'ajustement, calculées à partir de la matrice de confusion (cf. tableau 7) associée à la règle de décision choisie, existent. Le rappel (ou sensibilité) mesure la proportion de contrôles prédits parmi les contrôles effectivement réalisés :

$$\text{rappel/sensibilité} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

La spécificité mesure à l'inverse la proportion d'absence de contrôles prédits parmi les entreprises non contrôlées :

$$\text{spécificité} = \frac{\text{VN}}{\text{VN} + \text{FP}}$$

Ces deux grandeurs reflètent le nécessaire compromis qu'impose un problème de classification binaire et que la courbe ROC permet justement d'illustrer, en représentant ces deux grandeurs pour différentes valeurs de seuils. Une autre mesure d'ajustement est aussi reportée dans cette étude : la précision. Cette mesure quantifie la proportion d'entreprises effectivement contrôlées parmi celles qui se sont vues prédire un contrôle fiscal.

$$\text{précision} = \frac{\text{VP}}{\text{VP} + \text{FP}}$$

Ces différentes mesures d'ajustement sont adaptées à des échantillons déséquilibrés comme le sont nos échantillons test (qui ne sont pas rééquilibrés par l'algorithme SMOTE). Néanmoins, nous reporterons aussi la précision de la prédiction qui reflète la proportion de prédictions correctes, bien que cette mesure soit particulièrement sensible à la proportion d'entreprises qui connaissent un contrôle [24].

$$\text{précision de la prédiction} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{FP} + \text{VN} + \text{FN}}$$

Les performances des différents algorithmes testés sont ensuite comparées pour sélectionner le plus performant. Par suite, les groupes de contrôle homogène sont construits à partir des quantiles de la distribution, pour les entreprises contrôlées, des probabilités d'être contrôlée prédites par l'algorithme choisi. Se restreindre à la distribution obtenue sur les seules entreprises contrôlées assure que chaque GCH comporte un minimum d'entreprises de l'échantillon. En pratique, 50 GCH ont été constitués pour les entreprises dépendant de la DVNI et des Dircofi et 100 pour les entreprises dépendant des directions locales.

## 2.2 Partition en domaines et extrapolation

Une fois déterminées les pondérations pour toutes les entreprises à partir des GCH, il est possible d'estimer le montant total de TVA non recouvré. Afin de tenir compte de l'hétérogénéité

24. Par exemple, comme moins de 2 % d'entreprises sont contrôlées par les directions locales, une prédiction systématique de l'absence de contrôle pour toutes les entreprises de cette sous population conduit à obtenir une précision de 98 %, alors même que l'algorithme ne détecte aucun déterminant.

des comportements de fraude, notre estimation par sous-population s'appuie sur des estimations séparées par domaines. Dans cette partie, nous précisons tout d'abord comment sont construits ces domaines (partie 2.2.1), avant de détailler les estimateurs par le ratio et par la moyenne qui y sont appliqués (parties 2.2.2 à 2.2.5).

### 2.2.1 Domaines retenus pour l'estimation

Une fois estimées les pondérations de chaque entreprise, l'estimation totale des montants manquants de TVA est obtenue par somme des estimations réalisées sur une partition de la population en domaines. Passer par une partition de la population en domaines a un intérêt quand on choisit pour créer ces domaines une ou plusieurs variables positivement corrélée(s) à la variable d'intérêt à extrapoler. Plus la corrélation est importante, plus les domaines retenus conduisent à une différenciation importante des comportements de fraude entre eux, et à une homogénéité de ces comportements en leur sein. La constitution des domaines retenus ici s'appuiera sur des caractéristiques observées. Dès lors, rappelons qu'elle ne saurait tenir compte de l'hétérogénéité inobservée des comportements de fraude, et qu'elle ne permet pas non plus de connaître l'ampleur du biais en résultant dans nos estimations.

L'intérêt d'avoir estimé des probabilités de réponse sur des GCH assez fins, est justement de pouvoir procéder à l'extrapolation en utilisant ces probabilités de contrôle, mais sur des domaines parcimonieux qu'on choisira donc corrélés plutôt à des déterminants du fait de se faire notifier des montants de redressement de TVA (conditionnellement au fait d'être contrôlé). Nous retiendrons pour chaque sous-population un nombre limité de domaines, afin que ceux-ci soient suffisamment larges pour nous appuyer sur les propriétés asymptotiques des estimateurs considérés.

**Quantiles de TVA brute déclarée**

Comme le soulignaient déjà les tableaux 5 et 6, les montants et les taux de redressements prononcés à l'encontre des entreprises contrôlées sont corrélés aux montants de TVA brute déclarés. Si, en moyenne, le montant de redressement augmente avec le montant de TVA brute, le taux de redressement, à l'inverse, diminue. Ce constat suggère donc de considérer une première partition de la population en domaines construite à partir des quantiles de la distribution de TVA brute déclarée. Pour chaque sous population, nous constituons donc des domaines définis par les quantiles d'ordre cinquante de la distribution de TVA brute déclarée. Cette partition permet par ailleurs comme nous le verrons dans la section des résultats d'isoler dans un unique domaine les entreprises dépendant des directions locales qui déclarent un TVA brute nulle.

**Nomenclature d'Activités Française (Naf)**

Une autre approche mise en œuvre dans cette étude consiste à construire les domaines à partir des principaux déterminants d'un redressement. Pour les identifier, nous estimons, sur l'échantillon des seules entreprises contrôlées par l'administration fiscale, une régression logistique avec pénalisation LASSO sur l'indicatrice de notification d'un redressement positif. La pénalisation LASSO est fréquemment utilisée pour sélectionner des variables explicatives en présence d'un très grand nombre de variables potentielles. Il s'agit d'une régression (logistique dans notre cas, mais qui peut également être linéaire), dans l'estimation de laquelle une contrainte est ajoutée afin de limiter le nombre de variables explicatives dont les coefficients seront effectivement estimés. Cette contrainte permet de sélectionner parmi les variables explicatives celles dont le coefficient de régression associé est important, forçant les coefficients des paramètres ayant les plus faibles contributions à zéro. Plus la pénalité est élevée, plus la contrainte sera forte et moins un nombre important de variables sera retenu (voir annexe 5 pour la formalisation). Nos estimations mettent principalement en évidence l'importance du secteur d'activité [25] dans la probabilité

25. Parmi les autres déterminants, notons aussi la présence de quelques catégories juridiques.

d'être redressée une fois contrôlée, quelle que soit la sous-population considérée. Nous choisissons donc également de réaliser une extrapolation à partir de domaines définis par les sections de la nomenclature d'activité française (21 positions).

### 2.2.2 Estimateur par le ratio

Nous choisissons dans un premier temps d'implémenter un estimateur par le ratio pour réaliser nos estimations, pour plusieurs raisons. D'abord, cet estimateur s'appuie sur de l'information auxiliaire, et est particulièrement performant quand cette variable auxiliaire (qui doit être parfaitement connue dans la population) est *a priori* corrélée avec la variable d'intérêt à extrapoler. Dans notre cas, nous choisissons comme variable auxiliaire le **montant de TVA brute déclarée** par les entreprises : celle-ci est connue pour l'ensemble des redevables de la TVA (sauf évidemment dans le cas d'une activité non déclarée), et est très corrélée aux montants de redressement notifiés. Par ailleurs, cet estimateur avait déjà été utilisé dans la précédente estimation réalisée dans le document de travail de Claudie Louvot [18]. L'estimateur par le ratio [26] est un estimateur asymptotiquement sans biais sous certaines hypothèses et notamment l'absence d'endogénéité de la sélection. Puisque même sous cette hypothèse, l'estimateur par le ratio est biaisé à distance finie, nous l'appliquons à des domaines suffisamment larges.

Par souci de simplicité, on se place dans l'une des trois sous-populations qu'on appelle $\mathbf{U}$. On note $y_k$ le montant notifié de redressement d'une entreprise $k$ (potentiellement nul) sur la période contrôlée de l'année considérée (au plus 12 mois) et $x_k$ la base imposable correspondant à la même période. La sous-population $\mathbf{U}$ est partitionnée en $H$ domaines $U_1, ..., U_H$, et l'échantillon $\mathbf{S}$ également $(S_1, ..., S_H)$.

Soit $\mathbf{X}_h$ le total de la TVA brute déclarée pour l'année considérée dans la population du domaine $U_h$ (connu). On note également $\mathbf{Y}_h$ le total des montants notifiés au titre de cette même année dans la population du domaine $U_h$ (inconnu celui-ci, puisque seules les entreprises de l'échantillon (i.e. contrôlées) se voient éventuellement notifier des montants de redressement par l'administration fiscale).

On note respectivement $\widehat{T_{Y_h R}}$ et $\widehat{T_{X_h R}}$ les estimateurs d'Horvitz-Thompson repondérés des totaux dans le domaine $U_h$ des $y_k$ et des $x_k$, pour l'année considérée. Ces estimateurs sont des estimateurs asymptotiquement sans biais.

L'estimateur par le ratio du total des $y_k$ dans un domaine $U_h$, $\widehat{T_{Y_h ratio}}$, corrige l'estimateur d'Horvitz-Thompson repondéré du total $\widehat{T_{Y_h R}}$ par le ratio $\frac{\mathbf{X}_h}{\widehat{T_{X_h R}}}$ qui mesure l'écart entre l'estimation du total des $x_k$, et sa vraie valeur qui est connue. En revanche le ratio qui donne son nom à l'estimateur est le ratio du total des $y_k$ rapporté au total des $x_k$. L'estimateur par le ratio du total des $y_k$ dans le domaine $U_h$ s'écrit comme suit :

$$\widehat{T_{Y_h ratio}} = \widehat{T_{Y_h R}} \frac{\mathbf{X}_h}{\widehat{T_{X_h R}}} = \mathbf{X}_h \frac{\widehat{T_{Y_h R}}}{\widehat{T_{X_h R}}} = \mathbf{X}_h \hat{\mathbf{R}}_h$$

Le ratio $\hat{\mathbf{R}}_h$ défini comme étant le rapport de l'estimateur d'Horvitz-Thompson repondéré du total des $y_k$ sur l'estimateur d'Horvitz-Thompson repondéré du total des $x_k$ dans le domaine $U_h$ est l'estimateur du "vrai" ratio $\mathbf{R}_h$.

On estime enfin le total de la variable $y$ dans la sous-population considérée par la somme des estimateurs par le ratio estimés sur chaque domaine de cette sous-population :

26. cf. note méthodologique INSEE d'Olivier Sautory, 2018 [22].

$$\widehat{T_{Yratio}} = \sum_{h=1}^{H} \mathbf{X}_h \hat{\mathbf{R}}_h$$

L'estimateur par le ratio est un cas particulier d'estimateur par calage sur marge.

Le ratio de redressement dans un domaine est donc égal au quotient du montant total de redressement des entreprises contrôlées du domaine rapporté au total de TVA brute déclarée sur la période contrôlée. Puis, notre estimation du montant total est obtenue en multipliant le ratio estimé par la TVA brute totale déclarée par les entreprises du domaine considéré. Finalement, le montant total estimé s'obtient simplement par somme des trois estimateurs des totaux correspondants à chaque sous-population, ces trois estimateurs étant parfaitement indépendants.

### 2.2.3 Précision des estimateurs par le ratio et intervalles de confiance

Estimer la variance de l'estimateur du montant total manquant de versements de TVA nécessite plusieurs étapes. Dans un premier temps, pour chaque domaine, la variance de l'estimateur du ratio $\hat{\mathbf{R}}_h$ doit être déterminée. La variance de l'estimateur du total $\widehat{T_{Yratio}} = \sum_{h=1}^{H} \mathbf{X}_h \hat{\mathbf{R}}_h$ se calcule ensuite comme la variance d'une somme de variables aléatoires, indépendantes, car les domaines sont une partition de la population totale et que le total de la TVA brute déclarée $\mathbf{X}_h$ est connu.

La variance du ratio $\hat{\mathbf{R}}_h$ n'est pas d'expression connue, car il s'agit de la variance d'un ratio de variables aléatoires. Deux approches peuvent cependant être implémentées pour l'estimer [voir][pour des explications détaillées]Sarndal2003. La première consiste à s'appuyer sur une expression approchée de la variance théorique et d'en déterminer une estimation par des techniques de linéarisation de Taylor. La deuxième approche s'appuie sur une estimation obtenue par des méthodes empiriques faisant appel à des simulations (type *jacknife* ou *bootstrap*). Dans cette partie méthodologique, nous rappelons pour un domaine donné la forme de l'estimateur de la variance approximative par linéarisation de Taylor de l'estimateur par le ratio, et c'est de cet estimateur que découleront les estimations de variances qui figurent dans les résultats présentés dans la suite de ce document. Néanmoins, des estimations par *bootstrap* ont aussi été réalisées et fournissent des résultats très similaires.

On rappelle que $N$ est la taille de la sous-population considérée, (et $n$ la taille de l'échantillon correspondant) et on note $f = \frac{n}{N}$ le taux de sondage dans la population. On définit par ailleurs un poids de l'entreprise $k$ relatif au domaine $v_k = w_k I_{U_h}(k) = w_k I(k \in U_h)$. Dans un domaine $U_h$ d'une sous-population donnée, on a donc comme estimateur de la variance la formule suivante :

$$\hat{V}(\hat{\mathbf{R}}_h) = \frac{N(1-f)}{N-1} \sum_{k \in U} (g_k - \bar{g})^2$$

$$g_k = \frac{v_k(y_k - x_k \hat{\mathbf{R}}_h)}{\sum_{k \in U} v_k x_k}$$

$$\bar{g} = \frac{1}{N} \sum_{k \in U} g_k$$

On note par ailleurs $\hat{\sigma}$ l'estimateur de l'écart-type associé à l'estimateur de la variance défini ci-dessus :

$$\hat{\sigma}(\hat{\mathbf{R}}_h) = \sqrt{\hat{V}(\hat{\mathbf{R}}_h)}$$

Afin de calculer un intervalle de confiance associé à chaque ratio estimé, on considère que le ratio suit une loi normale [27]. On note $q_{\frac{\alpha}{2}}$ le quantile d'ordre $\alpha/2$ de la loi normale centrée réduite. Alors l'intervalle de confiance bilatéral de niveau $(1-\alpha)$ de l'estimateur du ratio dans un domaine $U_h$ au sein d'une sous-population donnée est ainsi défini comme suit :

$$IC(\hat{\mathbf{R}}_h)_{(1-\alpha)} = \left[\hat{\mathbf{R}}_h - \hat{\sigma}(\hat{\mathbf{R}}_h)q_{\frac{\alpha}{2}}; \hat{\mathbf{R}}_h + \hat{\sigma}(\hat{\mathbf{R}}_h)q_{\frac{\alpha}{2}}\right]$$

On en déduit dans un second temps un estimateur de la variance du total estimé sur le domaine ainsi que l'intervalle de confiance associé, en nous appuyant sur le déterminisme du total de la variable $x_k$ (connu) dans chaque domaine. On se passe également de valeur absolue dans l'écart-type estimé de l'estimateur par le ratio du total au sein d'un domaine en s'appuyant sur le fait que les montants $x_k$ sont positifs ou nuls.

$$\hat{V}(\widehat{T_{Y_h ratio}}) = \mathbf{X}_h^2 \hat{V}(\hat{\mathbf{R}}_h) \quad \text{et} \quad \hat{\sigma}(\widehat{T_{Y_h ratio}}) = \mathbf{X}_h \sqrt{\hat{V}(\hat{\mathbf{R}}_h)} = \mathbf{X}_h \hat{\sigma}(\hat{\mathbf{R}}_h)$$

$$IC(\widehat{T_{Y_h ratio}})_{(1-\alpha)} = \left[\mathbf{X}_h\hat{\mathbf{R}}_h - \mathbf{X}_h\hat{\sigma}(\hat{\mathbf{R}}_h)q_{\frac{\alpha}{2}}; \mathbf{X}_h\hat{\mathbf{R}}_h + \mathbf{X}_h\hat{\sigma}(\hat{\mathbf{R}}_h)q_{\frac{\alpha}{2}}\right]$$

On en déduit enfin un estimateur de la variance du total estimé sur la sous-population considérée, ainsi que l'intervalle de confiance associé, basé sur l'hypothèse d'indépendance entre domaines.

$$\hat{V}(\widehat{T_{Y ratio}}) = \sum_{h=1}^{H} \mathbf{X}_h^2 \hat{V}(\hat{\mathbf{R}}_h)$$

$$IC(\widehat{T_{Y ratio}})_{(1-\alpha)} = \left[\sum_{h=1}^{H} \mathbf{X}_h\hat{\mathbf{R}}_h - q_{\frac{\alpha}{2}}\sqrt{\sum_{h=1}^{H} \mathbf{X}_h^2\hat{V}(\hat{\mathbf{R}}_h)}; \sum_{h=1}^{H} \mathbf{X}_h\hat{\mathbf{R}}_h + q_{\frac{\alpha}{2}}\sqrt{\sum_{h=1}^{H} \mathbf{X}_h^2\hat{V}(\hat{\mathbf{R}}_h)}\right]$$

Finalement, la variance s'obtient simplement par somme des trois estimateurs de variances correspondants à chaque sous-population (ces estimateurs étant parfaitement indépendants), puis l'intervalle de confiance associé est recalculé à partir de l'estimateur du total comme une somme d'estimateurs indépendants qui suivent une loi normale ; utilisant l'écart-type qui découle de la variance et le quantile d'une loi normale centrée-réduite conforme au niveau de confiance retenu.

### 2.2.4 Estimateur par la moyenne

Un inconvénient de l'estimateur par le ratio est qu'il n'est pas défini sur un domaine constitué d'entreprises dont la TVA brute déclarée est nulle. Il est pourtant possible d'extrapoler à un tel domaine un montant de redressement positif simplement en utilisant un autre estimateur. Afin de pallier ce problème, nous choisissons d'implémenter également un estimateur par la moyenne. Il s'agit d'un estimateur asymptotiquement sans biais. Son biais pourra donc être considéré comme nul puisqu'il est appliqué à des domaines suffisamment larges.

Comme nous l'avons souligné dans la partie 1.3, un contrôle fiscal porte le plus souvent sur plusieurs exercices comptables. Le montant de redressement éventuellement notifié à l'issue du contrôle et enregistré dans la base Alpage est certes partitionné par motifs de redressement, mais il n'est nullement fait mention du ou des exercices comptables qui l'ont occasionné. Utiliser directement le montant de redressement notifié aux entreprises contrôlées reviendrait à tenir

27. Le logiciel SAS considère que l'estimateur suit une loi de Student, dont le nombre de degrés de liberté est égal au nombre d'observations de chacun des domaines, moins un. En pratique, le nombre de degrés de liberté est important compte tenu de la grande taille des domaines, la loi de Student peut donc être raisonnablement approchée par une loi normale.

compte dans nos estimations de l'hétérogénéité de la durée des contrôles fiscaux. Nous passerons donc pour cet estimateur par un montant de redressement *mensuel* moyen pour chaque domaine. Le montant manquant de TVA est ensuite calculé à partir des déclarations comptables des entreprises du domaine, c'est-à-dire, en multipliant le montant mensuel moyen de redressement estimé par la durée des exercices comptables déclarés par les entreprises du domaine considéré. L'interprétation du montant total estimé est alors identique à celle détaillée précédemment pour l'estimateur par le ratio.

Par souci de simplicité, on se place dans l'une des trois sous-populations qu'on appelle **U**. On note $y'_k$ le montant notifié mensualisé de redressement d'une entreprise $k$ (potentiellement nul), contrôlée au moins un mois l'année considérée. La sous-population **U** est partitionnée en $H$ domaines $U_1, ..., U_H$, et l'échantillon **S** également $(S_1, ..., S_H)$.

L'estimateur de la moyenne des $y'_k$ mensualisés dans un domaine $U_h$, noté $\widehat{\bar{Y}'_h}$, s'écrit comme suit :

$$\widehat{\bar{Y}'_h} = \frac{\sum_{k \in s_h} w_k y'_k}{\sum_{k \in s_h} w_k}$$

On note $d_i$ la durée d'exercice qui figure dans la déclaration de TVA de l'entreprise $i$ de l'année considérée (en mois, et connue pour l'ensemble des entreprises qui ont fait une déclaration de TVA). Dans un soucis de simplification des notations, on définit également $\mathbf{D}_h$ le total des durées d'exercice des entreprises du domaine $U_h$, pour l'année considérée (mais ce total n'a pas vraiment d'interprétation économique). Par suite, l'estimateur par la moyenne du total des $y_k$ pour l'année considérée sur un domaine $U_h$, noté $\widehat{T_{Y_h mean}}$, s'écrit comme suit :

$$\widehat{T_{Y_h mean}} = \sum_{i \in U_h} \frac{\sum_{k \in s_h} w_k y'_k}{\sum_{k \in s_h} w_k} d_i = \widehat{\bar{Y}'_h} \sum_{i \in U_h} d_i = \widehat{\bar{Y}'_h} \mathbf{D}_h$$

On estime enfin le total de la variable $y$ dans la sous-population considérée par la somme des estimateurs direct du total estimés sur chaque domaine de cette sous-population :

$$\widehat{T_{Y mean}} = \sum_{h=1}^{H} \widehat{T_{Y_h mean}} = \sum_{h=1}^{H} \widehat{\bar{Y}'_h} \mathbf{D}_h$$

Finalement, le montant total estimé s'obtient simplement par somme des trois estimateurs des totaux correspondants à chaque sous-population, ces trois estimateurs étant parfaitement indépendants.

### 2.2.5 Précision des estimateurs par la moyenne et intervalles de confiance

Comme pour la variance de l'estimateur d'un ratio, l'expression de la variance de l'estimateur d'une moyenne n'est pas connue dès lors que la taille de la population du domaine est aléatoire [28]. Les deux approches mentionnées dans la partie 2.2.3 pour estimer cette variance peuvent cependant à nouveau être implémentées. Cette fois encore, nous présentons dans cette partie méthodologique l'expression d'un estimateur de la variance approximative de $\widehat{\bar{Y}'_h}$ obtenue par la technique de linéarisation de Taylor. Les résultats correspondants à cette approche sont ceux qui seront discutés dans la partie consacrée. Des estimations par *bootstrap* ont aussi été implémentées et donnent des résultats similaires.

28. L'estimateur de la moyenne est en effet, dans ce cas de figure, à nouveau un ratio de variables aléatoires.

On rappelle que $N$ est la taille de la sous-population considérée, (et $n$ la taille de l'échantillon correspondant) et on note $f = \frac{n}{N}$ le taux de sondage dans la population. De même, le poids de l'entreprise $k$ relatif au domaine est défini par $v_k = w_k I_{U_h}(k) = w_k I(k \in U_h)$. Dans un domaine $U_h$ d'une sous-population donnée, on a comme estimateur de la variance la formule suivante :

$$\hat{V}(\hat{\bar{Y}}_h) = \frac{N(1-f)}{N-1} \sum_{k \in U} (r_k - \bar{r})^2$$

$$r_k = \frac{v_k(y'_k - \hat{\bar{Y}}'_h)}{\sum_{k \in U} v_k}$$

$$\bar{r} = \frac{1}{N} \sum_{k \in U} r_k$$

On note par ailleurs $\hat{\sigma}$ l'estimateur de l'écart-type associé à l'estimateur de la variance défini ci-dessus :

$$\hat{\sigma}(\hat{\bar{Y}}'_h) = \sqrt{\hat{V}(\hat{\bar{Y}}'_h)}$$

Afin de calculer un intervalle de confiance associé à chaque ratio estimé, on considère que le ratio suit une loi normale [29]. On note $q_{\frac{\alpha}{2}}$ le quantile d'ordre $\alpha/2$ de la loi normale centrée réduite. L'intervalle de confiance bilatéral de niveau $(1-\alpha)$ de l'estimateur de la moyenne dans un domaine $U_h$ au sein d'une sous-population donnée est ainsi défini comme suit :

$$IC(\hat{\bar{Y}}'_h)_{(1-\alpha)} = \left[\hat{\bar{Y}}'_h - \hat{\sigma}(\hat{\bar{Y}}'_h) q_{\frac{\alpha}{2}}; \hat{\bar{Y}}'_h + \hat{\sigma}(\hat{\bar{Y}}'_h) q_{\frac{\alpha}{2}}\right]$$

On en déduit dans un second temps un estimateur de la variance du total estimé sur le domaine $U_h$, ainsi que l'intervalle de confiance associé, en nous appuyant sur le déterminisme de la durée d'exercice $d_k$, connue pour l'ensemble des entreprises ayant fait une déclaration de TVA. On se passe également de valeur absolue dans l'écart-type estimé de l'estimateur par le ratio du total au sein d'un domaine en s'appuyant sur le fait que les durées $d_k$ sont positives ou nulles.

$$\hat{V}(\widehat{T_{Y_h mean}}) = \mathbf{D}_h^2 \hat{V}(\hat{\bar{Y}}'_h) \quad \text{et} \quad \hat{\sigma}(\widehat{T_{Y_h mean}}) = \mathbf{D}_h \sqrt{\hat{V}(\hat{\bar{Y}}'_h)} = \mathbf{D}_h \hat{\sigma}(\hat{\bar{Y}}'_h)$$

$$IC(\widehat{T_{Y_h mean}})_{(1-\alpha)} = \left[\mathbf{D}_h \hat{\bar{Y}}'_h - \mathbf{D}_h \hat{\sigma}(\hat{\bar{Y}}'_h) q_{\frac{\alpha}{2}}; \mathbf{D}_h \hat{\bar{Y}}'_h + \mathbf{D}_h \hat{\sigma}(\hat{\bar{Y}}'_h) q_{\frac{\alpha}{2}}\right]$$

On en déduit enfin un estimateur de la variance du total estimé sur le domaine ainsi que l'intervalle de confiance associé, basé sur l'hypothèse d'indépendance entre domaines.

$$\hat{V}(\widehat{T_{Y mean}}) = \sum_{h=1}^{H} \mathbf{D}_h^2 \hat{V}(\hat{\bar{Y}}'_h)$$

$$IC(\widehat{T_{Y mean}})_{(1-\alpha)} = \left[\sum_{h=1}^{H} \mathbf{D}_h \hat{\bar{Y}}'_h - q_{\frac{\alpha}{2}} \sqrt{\sum_{h=1}^{H} \mathbf{D}_h^2 \hat{V}(\hat{\bar{Y}}'_h)}; \sum_{h=1}^{H} \mathbf{D}_h^2 \hat{\bar{Y}}'_h + q_{\frac{\alpha}{2}} \sqrt{\sum_{h=1}^{H} \mathbf{D}_h \hat{V}(\hat{\bar{Y}}'_h)}\right]$$

Finalement, la variance s'obtient simplement par somme des trois estimateurs de variances correspondants à chaque sous-population (ces estimateurs étant parfaitement indépendants), puis

29. Le logiciel SAS considère que l'estimateur suit une loi de Student, dont le nombre de degrés de liberté est égal au nombre d'observations de chacun des domaines, moins un. En pratique, le nombre de degrés de liberté est important compte tenu de la grande taille des domaines, la loi de Student peut donc être raisonnablement approchée par une loi normale.

l'intervalle de confiance associé est recalculé à partir de l'estimateur du total comme une somme d'estimateurs indépendants qui suivent une loi normale ; utilisant l'écart-type qui découle de la variance et le quantile d'une loi normale centrée-réduite conforme au niveau de confiance retenu.

## 2.3 Tenir compte des redressements exceptionnels

Comme nous l'avons souligné (cf. tableau 3), des montants de redressements « exceptionnels » dans leur ampleur au regard de la distribution des montants notifiés chaque année, peuvent être prononcés à l'encontre de quelques entreprises. Intégrer ces observations dans l'échantillon des entreprises contrôlées utilisées pour mener nos estimations peut bien sûr impacter les résultats obtenus. Implicitement, leur prise en compte pose la question de la représentativité de telles observations. Les écarter de l'extrapolation revient à considérer que ces entreprises sont très spécifiques et non représentatives d'un comportement de fraude qui pourrait être extrapolé à d'autres entreprises. En d'autres termes, ces comportements particuliers auraient été tous détectés et redressés. À l'inverse, intégrer les montants de redressement exceptionnels à l'extrapolation revient à faire l'hypothèse selon laquelle frauder l'administration fiscale à hauteur d'un montant particulièrement important reste aléatoire.

Dans notre étude, la détection des « redressements exceptionnels » s'appuie sur une démarche statistique explicitée par Rousseeuw et al. [21]. Un score est calculé pour chaque entreprise s'appuyant notamment sur une estimation robuste de l'écart-type, à savoir la médiane des valeurs absolues des écarts à la médiane (MAD, pour *median of all absolute deviations from the median*) [30].

$$z_i = \frac{Y_i - \text{médiane}(Y_i)}{MAD(Y_i)}$$

Plus précisément, le score de chaque entreprise redressée est déterminé en fonction de la distribution des redressements prononcés [31] *au sein de son secteur d'activité*. Ce score nous permet ensuite d'identifier les « redressements exceptionnels » comme étant ceux situés dans le haut de la distribution de l'ensemble des scores [32], indépendemment du secteur d'activité. Ces observations sont alors écartées de l'estimation. Nous présenterons dans un premier temps les résultats obtenus *en écartant les dix redressements exceptionnels de chaque sous échantillon des entreprises contrôlées* [33], puis dans un second temps, nous discuterons l'impact du nombre de redressements exceptionnels retenus sur nos résultats, justifiant à cette occasion le choix effectué.

30. L'utilisation d'une règle qui s'appuie sur les valeurs d'un score pour détecter les observations particulières est fréquente en statistiques. Usuellement, le score retenu est le z-score qui se définit par $z_i = \frac{(x_i - \bar{x})}{s}$ où $s$ désigne l'écart-type. Une telle approche n'est cependant pas la plus robuste, puisque tant la moyenne que l'écart-type sont sensibles à la présence de valeurs extrêmes... qui pourront alors présenter un score malgré tout peu élevé.

31. Les contrôles qui n'ont pas abouti à un redressement sont donc exclus.

32. On notera que nous ne retenons pas la valeur absolue du score pour détecter les remboursements particuliers. Seuls les redressements particulièrement élevés, et non particulièrement faibles, au regard de la distribution des redressements prononcés dans un secteur d'activité seront donc considérés comme exceptionnels.

33. Si les scores sont déterminés en fonction du secteur d'activité de chaque entreprise redressée, les dix observations exclues qui présentent donc les scores les plus élevés, le sont indépendamment de leur secteur d'activité.

# 3 Résultats

## 3.1 Estimation des probabilités de contrôle

Nous présentons dans cette section les performances des algorithmes retenus pour prédire les probabilités de contrôle des entreprises pour chaque type d'administration en charge du contrôle fiscal (DVNI, Dircofi et directions locales), ainsi que les raisons qui nous ont poussés à retenir l'algorithme de *boosting*.

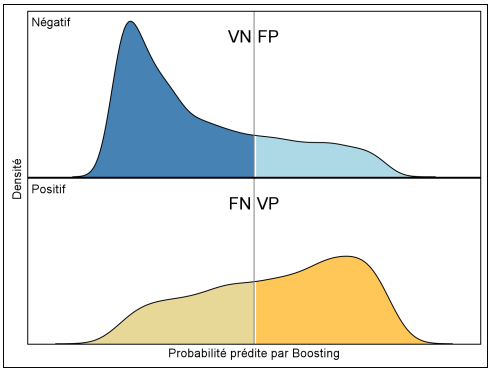
Pour chaque sous population, des algorithmes d'arbre de classification, de *bagging*, de forêts aléatoires, de *boosting* et une régression logistique pénalisée (elastic-net) ont été calibrés par validation croisée puis entraînés sur des échantillons rééquilibrés tirés aléatoirement (cf. sections méthodologiques 2.1.1 et 2.1.2). Les algorithmes calibrés sont ensuite appliqués à un échantillon test afin de quantifier leurs performances prédictives par différents critères d'ajustement (cf. section 2.1.3). Les résultats obtenus pour chaque algorithme, sur chaque sous population, sont présentés de manière exhaustive dans le tableau 9 situé en annexe.

Dans l'ensemble, pour une sous-population donnée, les performances des différents algorithmes sont assez semblables ; c'est la structure des données (notamment la nature déséquilibrée des données initiales) qui conditionne finalement ce que les algorithmes sont capables de détecter, et pas tant leurs différentes spécifications. Néanmoins, pour les entreprises dont le contrôle fiscal est assuré par les directions locales, l'algorithme de *boosting* présente des performances légèrement supérieures aux autres algorithmes testés sur cette sous-population. Nous retiendrons donc cet algorithme pour estimer l'ensemble des probabilités de contrôle quelle que soit la sous-population considérée, dont nous détaillons ci-dessous les performances.

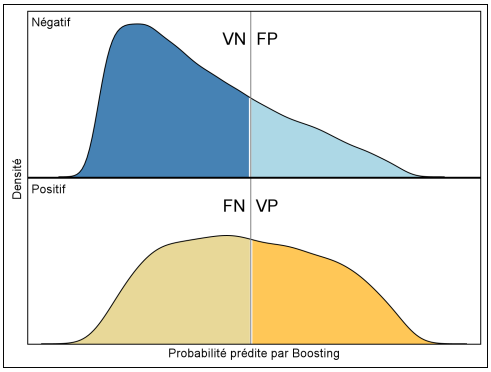
Les graphiques 1a, 1b et 1c représentent les distributions des probabilités estimées pour les entreprises des échantillons test de chaque sous population, en distinguant celles qui ont été effectivement contrôlées (« Positif »), et celles qui ne l'ont pas été (« Négatif »). Une règle de décision (ici, une probabilité prédite supérieure à 50 %) permet de distinguer les faux positifs, et les faux négatifs des prédictions correctes au sein de chaque groupe afin d'analyser les performances des algorithmes [34].

34. Les graphiques 7a, 7b et 7c en annexe représentent les courbes ROC pour chaque algorithme de *boosting*. Sur chaque courbe figurent aussi le rappel et la spécificité obtenus en retenant d'autres seuils pour la règle de décision.

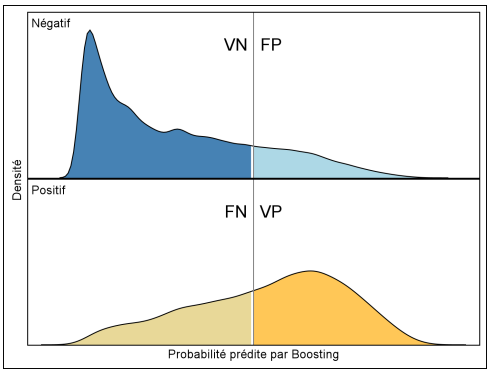
FIGURE 1 – Probabilités prédites par *boosting* - Vrais/Faux positifs et négatifs




(a) DVNI




(b) Dircofi




(c) Directions locales

Note : Pour chaque sous-direction, le graphique propose pour les entreprises effectivement contrôlées et non contrôlées, la probabilité de contrôle prédite par la méthode d'apprentissage de *boosting*.
Champ : Entreprises ayant effectué une déclaration de TVA en 2012.
Source : DGFiP, Insee, calcul des auteurs.

Comme explicité dans la section 2.1.3, ces grandeurs permettent de calculer plusieurs critères de performance qui sont présentés dans le tableau 8.

TABLE 8 – Performances des algorithmes de *boosting* de prédiction des probabilités de contrôle

| | DVNI | Dircofi | Directions Locales |
|---|---|---|---|
| Précision de la prédiction (*Accuracy*) | 0.727 | 0.708 | 0.799 |
| Précision | 0.268 | 0.221 | 0.060 |
| Rappel/Sensibilité | 0.619 | 0.471 | 0.608 |
| Spécificité | 0.743 | 0.745 | 0.803 |

Note : Chaque colonne présente, pour l'algorithme de *boosting*, les résultats obtenus sur différents critères de performance, pour les trois sous-échantillons de test des sous ensembles de directions fiscales.
Source : DGFiP, Insee, calcul des auteurs.

Les algorithmes de *boosting* retenus prédisent avec justesse (cf. *Accuracy*) l'absence ou la réalisation d'un contrôle fiscal respectivement pour 73 %, 71 % et 80 % des entreprises dépendant de la DVNI, des Dircofi et des directions locales. Plus précisément, ils envisagent, respectivement, un contrôle fiscal pour 62 %, 47 % et 61 % des entreprises qui seront effectivement contrôlées par la DVNI, les Dircofi et les directions locales (cf. Rappel/Sensibilité). De même, ils prédisent, respectivement, l'absence de contrôle fiscal pour 74 %, 75 % et 81 % des entreprises qui ne seront effectivement pas contrôlées par la DVNI, les Dircofi et les directions locales (cf. Spécificité). Cependant, nos algorithmes s'avèrent aussi peu « précis » : 27 % et 22 % des prédictions de contrôle fiscal par la DVNI et les Dircofi correspondent à des contrôles qui auront effectivement lieu, et seulement 6 % pour les directions locales.

Ces résultats soulignent que les algorithmes utilisés ont réussi à reproduire, à partir des variables retenues, une partie du processus de sélection des contrôles fiscaux mis en place par l'administration. Cependant les performances quelque peu limitées de nos algorithmes reflètent en partie les difficultés rencontrées par les méthodes de *machine learning* lorsque le nombre d'entreprises contrôlées dans la population totale est faible. De même elles témoignent aussi sans surprise que d'autres éléments connus de l'administration fiscale et non disponibles pour cette étude peuvent être mobilisés pour déterminer l'opportunité d'effectuer un contrôle fiscal, c'est-à-dire l'existence d'une sélection sur des caractéristiques inobservables. Mais surtout, la méthodologie mise en œuvre tout comme les critères de performance utilisés ne tiennent pas compte d'un élément déterminant dans la *réalisation effective* du contrôle, à savoir l'existence de contrôleurs fiscaux en nombre suffisant pour les mener à bien. En effet, quelle que soit la qualité prédictive du modèle retenu, le nombre de contrôles fiscaux effectivement menés sur une année civile dépendra directement des effectifs mobilisables par l'administration fiscale, comme le soulignent en partie, par exemple, les grandes différences de performances mesurées par le rappel et la spécificité. De même, si d'après ses dires, l'administration fiscale privilégie des contrôles fiscaux qui sont particulièrement susceptibles d'aboutir à un redressement fiscal (éventuellement conséquent), elle prend soin aussi chaque année de procéder à des contrôles fiscaux sur des entreprises moins susceptibles de frauder. Ce principe explique certainement aussi les faibles performances des algorithmes retenus au regard des standards associés à ce type de méthode.

Toutefois, l'utilisation d'algorithmes de *machine learning* visait ici, non à prédire avec justesse les contrôles fiscaux effectivement menés, mais à estimer au mieux des probabilités d'être contrôlée pour réduire autant que faire se peut, à partir des informations à notre disposition, le biais de sélection [35]. Il s'agissait, en effet, à partir de cette estimation de la probabilité d'être

35. En autorisant notamment des relations non linéaires entre le fait qu'une entreprise soit contrôlée et

contrôlée de définir des groupes d'entreprises pour lesquelles l'éventualité d'un contrôle fiscal est similaire, afin d'assurer que les entreprises effectivement contrôlées dans un groupe de contrôle homogène soient représentatives en termes de comportement déclaratif de toutes les entreprises du même groupe. Si les performances des algorithmes testés soulignent sans surprise la complexité du processus de sélection des entreprises contrôlées et donc de sa modélisation, la constitution de GCH et l'estimation de la probabilité de contrôle par la probabilité empirique de contrôle au sein des groupes ainsi constitués nous apparaît à même de pallier les faiblesses des algorithmes utilisés, mais pas le fait que les données ne contiennent pas toutes les informations permettant de corriger du biais de sélection.

Nous créons ensuite des GCH à partir de la distribution des probabilités de contrôle estimées par *Boosting*, comme explicité dans la partie 2.2. Cinquante GCH sont donc créés dans chacune des sous-populations dont le contrôle relève des DVNI et des DIRCOFI, sur la base d'un découpage sous forme de quantiles d'ordre 50, tandis que la dernière sous-population est partitionnée selon 100 quantiles de probabilité de contrôle. Les GCH ainsi créés sont assez nombreux tout en étant d'une taille suffisamment importante pour garantir la présence au sein de chaque GCH d'entreprises contrôlées en nombre suffisant, afin de pouvoir recalculer une probabilité empirique de contrôle associée au GCH, puis à attribuer celle-ci à l'ensemble des entreprises de ce GCH.

## 3.2 Estimations des montants manquants de versements de TVA

Nous présentons maintenant les résultats de nos estimations des montants manquants de TVA pour l'année 2012. Pour chaque sous ensemble de directions fiscales (DVNI, Dircofi et directions locales) et pour chaque domaine, nous réalisons une estimation par le ratio et une estimation par la moyenne.

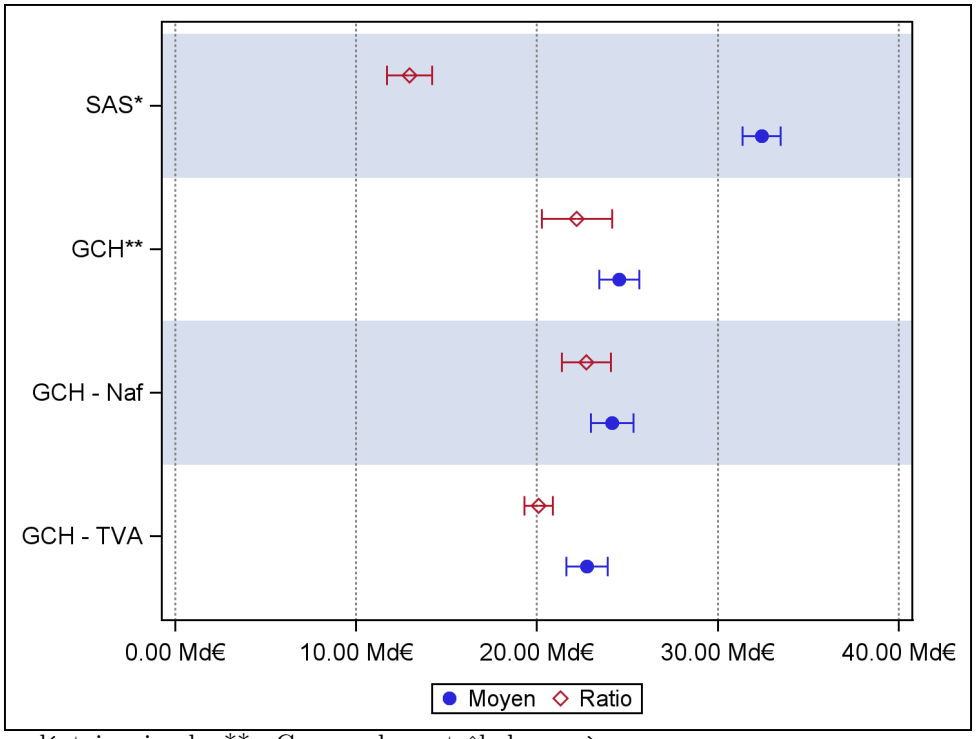
*Les estimations présentées ci-dessous diffèrent quelque peu des premières estimations annexées au rapport de la Cour des comptes de décembre 2019 qui mentionnait une quinzaine de Md€ : le ratio qui était alors utilisé pour calculer l'estimateur par le ratio du montant total n'était pas le même, il rapportait le montant de redressement, pouvant ne correspondre qu'à quelques mois dans l'année, au résultat annuel déclaré. Dans le cas où l'année considérée n'est que partiellement contrôlée, ce ratio était mécaniquement plus faible que celui aujourd'hui considéré, qui rapporte deux grandeurs rattachées à une même période (la période de contrôle). Sachant que près de 20% des entreprises sont contrôlées sur une partie seulement de l'année 2012, cette différence dans la définition du ratio explique en partie la légère hausse des estimations.*

Le graphique 2 représente les résultats obtenus après agrégation des estimations effectuées pour la DVNI, les Dircofi et les directions locales [36] pour différentes probabilités d'inclusion et/ou en retenant différentes définitions de domaines. Dans un premier temps, nous proposons une estimation « naïve » en considérant que les entreprises contrôlées sont simplement issues (à tort) d'un tirage aléatoire simple (ligne (SAS)). De plus, pour cette estimation, aucun domaine n'est retenu. Puis, nous affectons à chaque entreprise, comme probabilité d'inclusion, la probabilité corrigée de la probabilité de contrôle du groupe de contrôle homogène auquel elle appartient (cf. section 2.2). Plusieurs estimations sont alors effectuées, sans domaines (ligne (GCH)) ou en retenant deux partitions de la population, par secteur d'activité (ligne (GCH-Naf)) ou par quantiles de TVA brute déclarée en 2012 (ligne (GCH-TVA)).

les variables retenues dans les modèles.

36. Les résultats par type d'administrations en charge du contrôle fiscal sont détaillées dans les tableaux 10 et 11 situés en annexe.

FIGURE 2 – Estimations des montants manquants de TVA par la moyenne et par le ratio




* : Sondage aléatoire simple, ** : Groupe de contrôle homogène

Note : Chaque ligne présente l'estimation et l'intervalle de confiance à 95 % associé, obtenus en utilisant un estimateur par le ratio et un estimateur par la moyenne, sur l'ensemble de la population ou après agrégation des résultats par domaine. Dans la ligne (SAS), l'hypothèse retenue est que les entreprises contrôlées sont issues d'un tirage aléatoire simple. Dans les lignes (GCH), la probabilité d'inclusion de chaque entreprise correspond à la proportion d'entreprises contrôlées du groupe de contrôle homogène à laquelle elle appartient (cf. section 2.2). Les 50 quantiles de TVA qui définissent les domaines dans la quatrième ligne sont constitués à partir de la distribution, pour les seules entreprises contrôlées, de la TVA brute déclarée.

Champ : Entreprises ayant effectué une déclaration de TVA en 2012.

Source : DGFiP, Insee, calcul des auteurs.

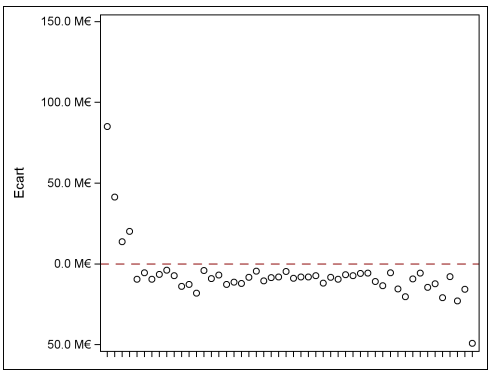
En ce qui concerne l'estimateur par le ratio, les estimations obtenues, en considérant que les entreprises contrôlées sont tirées au sort avec la même probabilité (sondage aléatoire simple), sont les plus faibles quelle que soit la sous population envisagée. Le montant manquant total de TVA est alors estimé à 13 Md€ mais cette estimation ne corrige pas du biais de sélection des entreprises contrôlées par les différentes administrations fiscales. Lorsque l'on utilise les pondérations issues de la correction apportées par les groupes de contrôle homogène (ligne (GCH)), le montant manquant de TVA estimé augmente pour chaque type d'administration (cf. tableau 10) pour s'établir au total à 22,2 Md€. Cette estimation est de plus significativement différente de celle obtenue en considérant un sondage aléatoire simple, témoignant aussi en cela de la meilleure prise en compte de la complexité du processus de sélection par la DGFiP des entreprises contrôlées (sans que l'on sache si on élimine effectivement la totalité du biais en résultant). Les estimations par Naf ou par quantiles de TVA brute déclarée conduisent à obtenir des estimations similaires, respectivement 22,7 Md€ et 20,1 Md€. Ces deux estimations ne sont pas statistiquement différentes de l'estimation sans domaine obtenue en retenant comme probabilités d'inclusion celles des GCH, mais elles sont plus précises.

À l'inverse, les résultats obtenus avec l'estimateur par la moyenne diminuent dès lors que l'on corrige du biais de sélection en retenant comme pondération les probabilités des GCH puis que l'on considère des montants moyens de redressement hétérogènes par secteur d'activité ou par quantiles de TVA brute déclarée. Ainsi, sous l'hypothèse « naïve » que les entreprises contrôlées sont issues d'un tirage aléatoire simple, nous estimons que le montant manquant total de recouvrement de TVA atteindrait 32,4 Md€, contre 24,6 Md€ si l'on considère les groupes de contrôle homogène (ligne GCH). Si l'on autorise en plus des montants moyens de redressement différents par secteur d'activités (GCH-Naf) ou par quantiles de TVA brute déclarée (ligne GCH-TVA), les montants manquants de TVA sont estimés respectivement à 24,2 Md€ et 22,8 Md€. Cependant, dès lors que l'on considère les groupes de contrôle homogène, les résultats obtenus par l'estimateur par la moyenne, avec ou sans domaine, ne sont pas statistiquement différents.

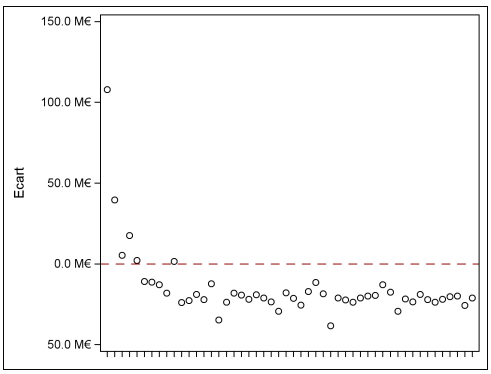
Les estimations par le ratio et par la moyenne des montants manquants de TVA ne sont pas statistiquement différentes, lorsque l'on tient compte du biais de sélection (lignes GCH), bien que les estimations par le ratio soient inférieures à celles obtenues par la moyenne. Pour comprendre ce résultat, une comparaison entre les deux estimateurs par quantiles de TVA est représentée pour chaque sous-population (DVNI, Dircofi, directions locales) dans les graphiques 3a, 3b et 3c. Sur ces graphiques, les quantiles sont ordonnés par ordre croissant de TVA brute déclarée de gauche à droite, de telle sorte que le premier quantile comporte notamment les entreprises ayant déclaré une TVA brute collectée nulle en 2012. Pour les directions locales, ce quantile ne regroupe d'ailleurs que des entreprises dans ce cas de figure. Quelle que soit la sous-population considérée, les estimations par le ratio sont plus faibles que les estimations par la moyenne pour les quatre/cinq premiers quantiles de la distribution, et particulièrement pour le quantile qui comporte les entreprises ayant déclaré une TVA brute nulle. Pour les autres quantiles, les estimations par le ratio sont toutes plus élevées que les estimations par la moyenne. Au total, pour les DVNI et les Dircofi, les estimations de TVA manquantes obtenues par l'estimateur par le ratio excèdent celles obtenues par l'estimateur par la moyenne, respectivement de 34 et 75 M€ (cf. tableaux 10 et 11). À l'inverse pour les directions locales, l'estimation par la moyenne dépasse l'estimation par le ratio de 3,8 Md€. L'ampleur de l'écart pour les directions locales s'explique principalement par la différence d'estimation obtenue pour les entreprises qui ne déclarent pas de TVA brute collectée, c'est-à-dire pour le premier quantile. En effet, pour ces entreprises, l'estimateur par le ratio n'est pas défini et de ce fait, aucun montant de fraude n'est extrapolé. À l'inverse, on obtient une estimation pour ces entreprises en utilisant l'estimateur par la moyenne, qui s'élève à 4 Md€. L'écart important observé sur ce quantile n'est donc que partiellement compensé par les estimations plus élevées obtenues par l'estimateur par le ratio sur les quantiles

correspondant à des déclarations de TVA brute plus élevées. Il ressort donc que la similitude des estimations obtenues par ces deux estimateurs masque en réalité des écarts plus contrastés par domaines.

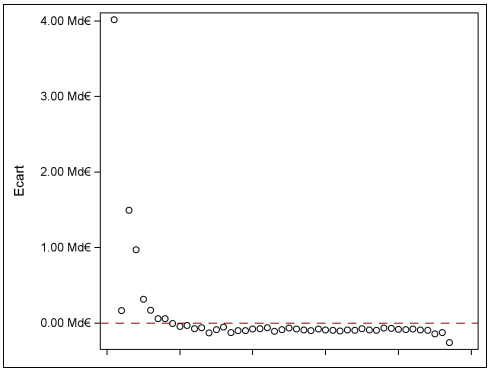
FIGURE 3 – Différence entre l'estimation par la moyenne et l'estimation par le ratio selon les quantiles de TVA brute déclarée




(a) DVNI




(b) Dircofi

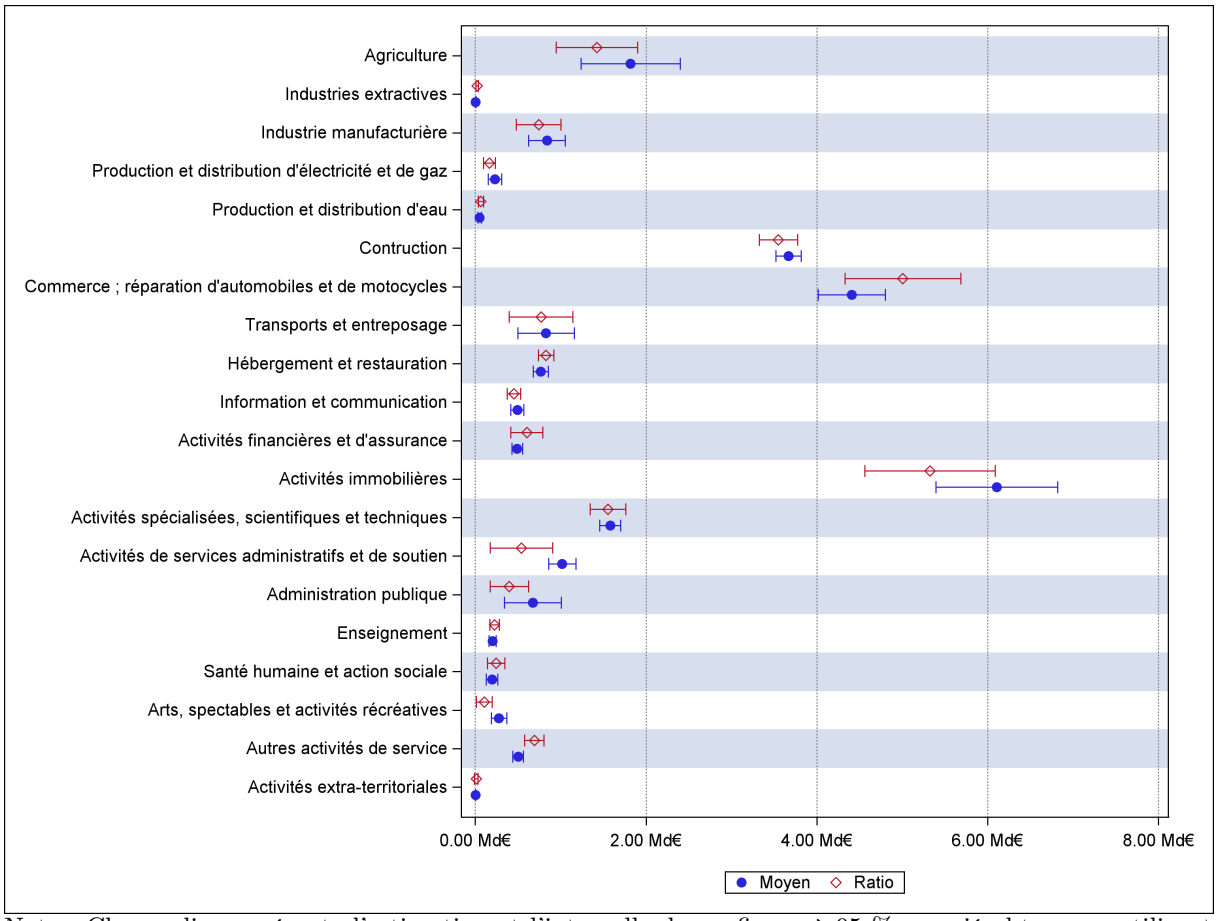



(c) Directions locales

Note : Pour chaque quantile, le graphique représente l'écart entre les estimations obtenues par l'estimateur par la moyenne et par le ratio. Une valeur située au dessus de la ligne indique une estimation par la moyenne supérieure à l'estimation obtenue par le ratio. Les quantiles sont constitués à partir de la distribution de la TVA brute déclarée. Ils sont ordonnés par valeurs croissantes de TVA brute déclarée. Le premier quantile comporte ainsi notamment les entreprises ayant déclaré en 2012 une TVA brute collectée nulle. Pour les directions locales, ce quantile ne regroupe que des entreprises ayant déclaré une TVA brute collectée nulle en 2012.
Champ : Entreprises ayant effectué une déclaration de TVA en 2012.
Source : DGFiP, Insee, calcul des auteurs.

Les résultats obtenus par secteur d'activité soulignent aussi la concentration dans un nombre restreint de secteurs des montants manquants de versement de TVA (cf. graphique 4). Seuls trois secteurs d'activité présentent un montant total manquant de versements de TVA supérieur à 2 Md€ ; le secteur de la construction (3,55 Md€ pour l'estimateur par le ratio et 3,67 Md€ pour l'estimateur par la moyenne), le secteur du commerce et de la réparation d'automobiles et de motocycles (respectivement 5,01 Md€ et 4,41 Md€) et celui des activités immobilières (respectivement 5,33 Md€ et 6,11 Md€). À eux seuls, ces trois secteurs représentent environ 55 % des montants totaux manquants de versement de TVA, quel que soit l'estimateur retenu, alors qu'ils ne représentent que 47 % du montant total de TVA brute déclarée.

FIGURE 4 – Estimations des montants manquants de TVA par la moyenne et par le ratio par secteurs d'activité

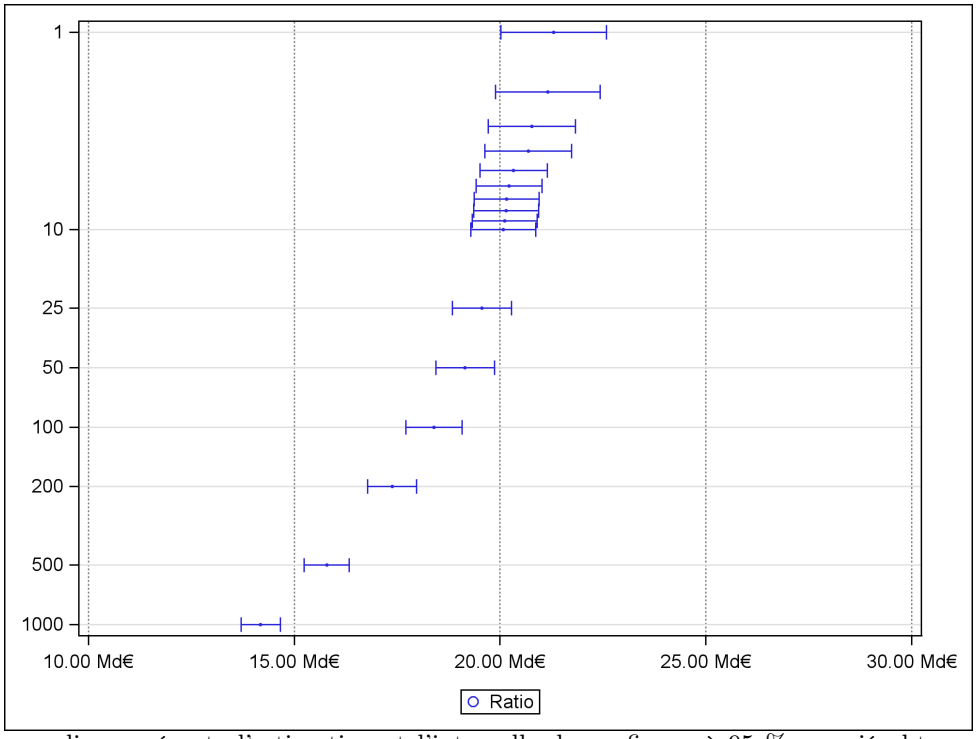



Note : Chaque ligne présente l'estimation et l'intervalle de confiance à 95 % associé obtenu en utilisant un estimateur par le ratio et un estimateur par la moyenne. La probabilité d'inclusion de chaque entreprise correspond à la proportion d'entreprises contrôlées du groupe de contrôle homogène à laquelle elle appartient (cf. section 2.2).
Champ : Entreprises ayant effectué une déclaration de TVA en 2012.
Source : DGFiP, Insee, calcul des auteurs.

## 3.3 L'impact des redressements exceptionnels

Les estimations précédentes ont été obtenues en ne retenant pas, pour chaque sous-population d'entreprises, dix redressements « exceptionnels [37] » prononcés à l'encontre des entreprises dont l'exercice fiscal de l'année 2012 a été contrôlé (cf. section 2.3). Dans cette partie, nous justifions ce choix en étudiant l'impact de l'intégration d'un nombre croissant de redressements « exceptionnels » sur nos estimations et leurs intervalles de confiance [38]. Plus précisément, nous réalisons à nouveau des estimations par le ratio par quantiles de TVA en retenant les probabilités d'inclusion corrigées de la probabilité de contrôle par groupes de contrôle homogène (ligne GCH-TVA dans les tableaux précédents). Pour chaque estimation, un nombre de redressements « exceptionnels », allant de 1000 à 1, n'est pas intégré à l'échantillon considéré pour chaque sous-population (DVNI, Dircofi et directions locales). Les résultats obtenus pour l'estimateur par le ratio sont présentés dans le graphique 5, ceux pour l'estimateur par la moyenne dans le graphique 6.

FIGURE 5 – Comparaison des estimations par le ratio obtenues par intégration successive de redressements exceptionnels




Note : Chaque ligne présente l'estimation et l'intervalle de confiance à 95 % associé obtenus (i) en ne retenant pas les 1000, 500, 200, 100 etc. plus grands redressements prononcés, (ii) en considérant comme domaines, les 50 quantiles de TVA constitués à partir de la distribution, pour les seules entreprises contrôlées retenues, de la TVA brute déclarée. La probabilité d'inclusion de chaque entreprise correspond à la proportion d'entreprises contrôlées du groupe de contrôle homogène à laquelle elle appartient (cf. section 2.2).
Champ : Entreprises ayant effectué une déclaration de TVA en 2012.
Source : DGFiP, Insee, calcul des auteurs.

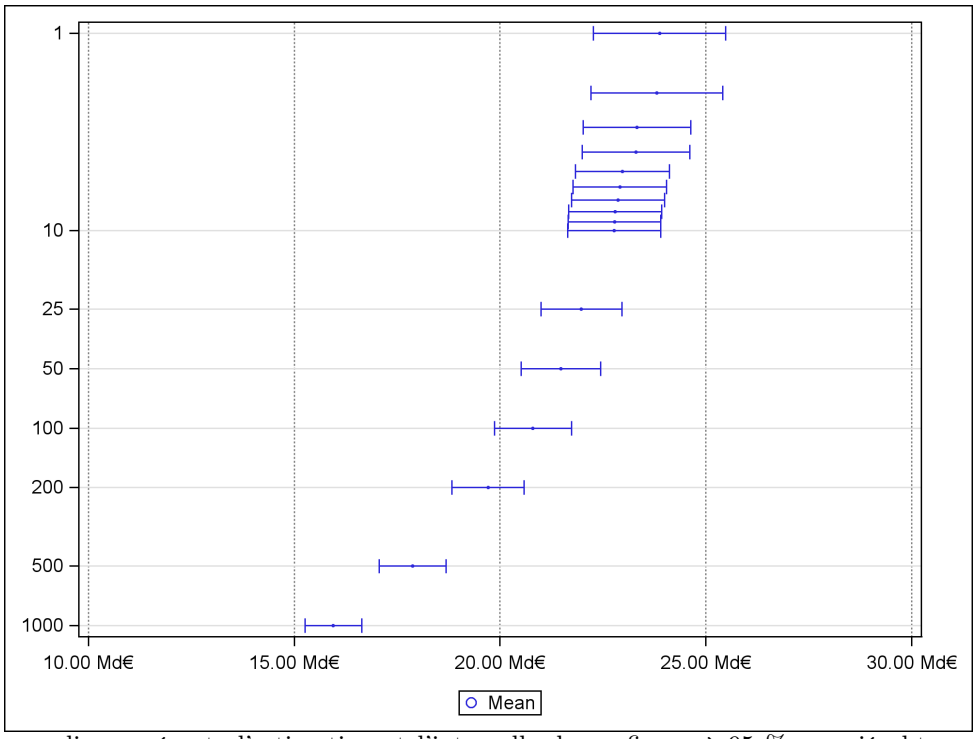
Quelle que soit la méthode, les estimations ainsi que les intervalles de confiance augmentent

37. soit 30 observations au total.

38. On trouvera en Annexe, une autre représentation graphique permettant de justifier ce choix. Pour chaque sous-population, nous déterminons le nombre de redressements « exceptionnels » à exclure de l'échantillon en fonction du seuil retenu pour le score (cf. graphique 8). Une partition par secteurs d'activité du nombre de redressements exclus en fonction du seuil retenu, par sous-population, est aussi proposé.

au fur et à mesure que le nombre de redressements « exceptionnels » intégrés dans l'échantillon des entreprises contrôlées augmente. Cependant, un plateau apparaît lorsque dix observations ne sont pas retenues, avant qu'une augmentation importante, tant du montant total estimé que de l'intervalle de confiance correspondant, ne survienne si l'on intègre les cinq redressements les plus « exceptionnels » prononcés dans chaque sous-population considérée. Cette sensibilité soudaine de nos résultats à l'intégration d'un nombre réduit d'observations supplémentaires ne démontre pas nécessairement que ces entreprises présentent un comportement de fraude qui serait singulier et non représentatif de celui d'autres entreprises, mais elle souligne qu'une estimation minimale et robuste des montants manquants de versements de TVA peut être obtenue en n'intégrant pas les dix redressements les plus « exceptionnels » prononcés dans chaque sous-population considérée.

FIGURE 6 – Comparaison des estimations par la moyenne obtenues par intégration successive de redressements exceptionnels




Note : Chaque ligne présente l'estimation et l'intervalle de confiance à 95 % associé obtenus (i) en ne retenant pas les 1000, 500, 200, 100 etc. plus grands redressements prononcés, (ii) en considérant comme domaines, les 50 quantiles de TVA constitués à partir de la distribution, pour les seules entreprises contrôlées retenues, de la TVA brute déclarée. La probabilité d'inclusion de chaque entreprise correspond à la proportion d'entreprises contrôlées du groupe de contrôle homogène à laquelle elle appartient (cf. section 2.2).
Champ : Entreprises ayant effectué une déclaration de TVA en 2012.
Source : DGFiP, Insee, calcul des auteurs.

# Conclusion

Cette étude vise à quantifier les montants manquants de versements de TVA que pourrait recouvrer la DGFiP en contrôlant l'ensemble des entreprises redevables de la TVA, c'est-à-dire les montants correspondant au non respect du droit fiscal. Notre méthodologie s'appuie sur la théorie des sondages en considérant que les entreprises contrôlées correspondent aux seules répondantes à une enquête, parmi un échantillon constitué de l'ensemble des entreprises redevables de la TVA. Cependant, les entreprises contrôlées par la DGFiP ne sont pas sélectionnées selon un plan de

sondage connu. En effet, la mission de contrôle fiscal de la DGFiP a trois objectifs : budgétaire, répressif et dissuasif. Recouvrer les montants non versés ou sanctionner les comportements de fraude délibérés par ses contrôles fiscaux conduit la DGFiP à cibler les entreprises pour lesquelles la suspicion de fraude est importante. Toutefois, la DGFiP contrôle aussi des entreprises exerçant dans des secteurs d'activité peu fraudogènes, de telle sorte que le processus de sélection des entreprises contrôlées est complexe à modéliser. Afin de tenter de s'abstraire au mieux de ce biais de sélection, nous estimons par des algorithmes de *machine learning* une probabilité de contrôle des entreprises redevables de la TVA ; il est alors possible de repondérer les poids de sondage par ces probabilités estimées afin de retracer au mieux le phénomène de sélection au moment de l'extrapolation. Enfin, le montant total de TVA non recouvrée est estimé sur différentes partitions en domaines des entreprises redevables de la TVA, par un estimateur par le ratio ou par la moyenne.

Nos résultats suggèrent que quel que soit l'estimateur retenu, le montant total de TVA non recouvré serait compris entre 20 et 26 Md€. Cette estimation n'intègre pas quelques montants de redressement exceptionnels dont la prise en compte nécessite de poser une hypothèse supplémentaire. En effet, les intégrer dans notre échantillon revient à considérer que ces comportements de fraude massive ne seraient pas isolés mais pourraient être représentatifs du comportement déclaratif d'autres entreprises, au même titre que le reste de l'échantillon des entreprises contrôlées.

Ces estimations sont cependant à manier avec prudence, car persistent quelques limites importantes qu'il convient de garder à l'esprit. D'abord, l'algorithme de *boosting* retenu pour prédire les probabilités de contrôle (et *a fortiori* l'ensemble des algorithmes entraînés et testés sur les données) montre des performances limitées, en particulier en termes de précision. Il s'avère donc que la modélisation du processus de sélection à partir des caractéristiques observées est incomplète. De plus, la démarche adoptée ne peut rendre compte de l'hétérogénéité inobservée sous-jacente des déterminants de la programmation d'un contrôle, par exemple l'expertise des contrôleurs fiscaux dans l'arbitrage entre deux entreprises à contrôler, faute de moyens. Cette partie non prise en compte du biais de sélection impacte directement l'extrapolation, *via* les poids de sondage repondérés. Il faut aussi souligner le fait qu'il n'est pas possible de savoir dans quelle mesure le biais de sélection est effectivement pris en compte. Seule la mise en place de contrôles aléatoires permettrait de répondre à cette difficulté.

Par ailleurs, par définition, les redressements prononcés à l'encontre des entreprises contrôlées résultent d'une détection par le contrôleur fiscal d'une irrégularité, ou d'un comportement frauduleux. En s'appuyant sur les résultats des contrôles menés par la DGFiP, notre estimation fait donc implicitement l'hypothèse que les comportements de fraude des entreprises sont *tous détectés* par les services fiscaux lorsque le contrôle a lieu. En cela, le champ de notre estimation ne retient pas les comportements frauduleux non connus à ce jour. De plus, comme nous l'avons souligné dans la partie consacrée à la méthodologie, nos résultats reposent implicitement sur l'hypothèse que le comportement frauduleux d'une entreprise s'étend sur toute la durée de son exercice fiscal. Une telle hypothèse ne tient pas compte des contraintes législatives qui encadrent la récurrence des contrôles d'une même entreprise, ni des effets dissuasifs sur le comportement déclaratif futur d'un contribuable suite à un redressement fiscal. Ainsi, si la mise en place de contrôles aléatoires améliorerait sensiblement la qualité des estimations de fraude fiscale à la TVA par la prise en compte rigoureuse du biais de sélection, une attention particulière doit demeurer dans l'analyse des résultats sur le champ et la nature de la fraude estimée.

# 4 Algorithmes de *machine learning* testés

Nous détaillons succinctement ci-dessous les cinq algorithmes testés pour chaque sous population.

L'apprentissage par **arbre de décision** est un algorithme d'apprentissage supervisé. Il vise à prédire la valeur d'une caractéristique catégorielle, ici le fait d'être contrôlé, à partir des valeurs de plusieurs variables définies comme étant des variables d'entrée. Les algorithmes comme l'alogrithme CART Breiman1984 utilisé dans cette étude vise à scinder l'espace des observations considérées, de manière récursive, en commençant par la population entière. Chaque partition d'ensembles est réalisée par un test sur la valeur d'une caractéristique d'entrée. Si la variable est discrète, les sous-ensembles sont constitués à partir de ses différentes modalités. Dans le cas d'une variable continue, c'est un seuil qui définit la partition en deux sous-ensembles de la population initiale.

Pour chaque sous-ensemble, la valeur prédite de la caractéristique cible est celle qui est la plus représentée. Ce processus est répété sur chaque sous-ensemble obtenu de manière récursive. Pour choisir la variable de séparation à chaque étape, l'algorithme teste les différentes variables d'entrée possibles et sélectionne celle qui maximise un critère donné, souvent l'indice de diversité de Gini. Ce processus est achevé, soit lorsque tous les sous-ensembles ont la même valeur de la caractéristique-cible, ou lorsque la séparation n'améliore plus la prédiction. Un tel processus peut conduire à la constitution d'un arbre de classification complexe même si chaque sous-ensemble est parfaitement homogène du point de vue de la variable-cible [39]. Or l'apprentissage est réalisé sur un échantillon que l'on espère représentatif d'une population. L'enjeu de toute technique d'apprentissage est d'arriver à saisir l'information utile sur la structure statistique de la population, en excluant les caractéristiques spécifiques au jeu de données étudié. Dans le cas contraire, le risque dit de sur-apprentissage conduit à voir le modèle incapable d'être extrapolé à de nouvelles données. La construction d'un arbre de classification s'appuie donc aussi sur un paramètre de complexité qu'il convient de calibrer au mieux.

Le méta-algorithme **bagging** (ou *Bootstrap AGGregation of classifiers*) d'arbres de classification consiste à appliquer une règle de décision (par exemple l'arbre de classification optimal déterminé comme décrit ci-dessus) pré-sélectionnée et donc déjà calibrée, à un ensemble de sous-échantillons aléatoires de même taille tirés avec remise dans l'échantillon de départ. De cette façon, les prédicteurs seront indépendants. La prédiction issue du *bagging* est la moyenne des prédictions faites sur l'ensemble des sous-échantillons par la règle de décision. En pratique les performances sont améliorées (la variance diminue toujours, le biais est au mieux réduit, au pire maintenu), en revanche la méthode perd en interprétabilité comme règle de décision. On calibre le nombre de sous-échantillons tirés dans l'échantillon ; avec l'intuition que plus ceux-ci sont nombreux, meilleure est la précision.

L'algorithme des **forêts aléatoires** est un algorithme de *bagging* prenant comme règle de décision un arbre de taille maximale.

**Boosting** sur des arbres de classification : parmi les algorithmes de boosting, on trouve en particulier le *Gradient Boosting* Dans le cas du *boosting*, contrairement au *bagging*, les algorithmes sous-jacents ne sont plus indépendants : à chaque itération de l'algorithme, un arbre de classification est entraîné, mais s'appuyant sur le précédent (il n'y a plus indépendance des prédicteurs). Le prédicteur d'initialisation du *boosting* prédit une unique étiquette : la plus fréquente dans les données. Ensuite, à chaque étape, un arbre est entraîné sur le résidu prédit à l'étape immédiatement précédente (à l'étape 1, pour chaque observation, est prédit l'écart entre la « moyenne » et

39. Le cas extrême étant lorsqu'il y a autant de sous-ensembles que d'observations

la vraie étiquette de l'observation considérée), puis ces prédictions sont multipliées par un facteur inférieur à un, afin que chaque itération apporte une correction à la prédiction précédente. Ainsi, chaque itération améliore les performances de la précédente en se rapprochant par « petits pas » successifs des vraies valeurs. La prédiction issue du *boosting* est la somme des prédictions faites par les arbres successifs.

La **régression régularisée elasticnet** est une régression logistique qui combine une pénalité sur la norme $L_1$ (LASSO) et sur la norme $L_2$ (Ridge) des coefficients. Elle comporte donc deux paramètres à calibrer : un paramètre de pénalité global, et un second paramètre qui pondère l'importance relative d'une des deux pénalités par rapport à l'autre.

# 5 Régression logistique avec pénalisation LASSO

Formellement, la variable d'intérêt $Y$ que l'on cherche à expliquer est une variable binaire, prenant la valeur 1 si l'entreprise se voit notifier un montant positif de TVA, et la valeur 0 sinon. Soit $X$ un ensemble de variables explicatives (contenant le vecteur unité) ; on s'intéresse alors à l'estimation des paramètres du modèle logistique suivant :

$$\mathbb{P}(Y_i = 1) = \frac{e^{X_i'\beta}}{1 + e^{X_i'\beta}}$$

Pour un paramètre $\lambda$ positif, l'estimateur LASSO correspondant est alors défini comme minimisant l'opposé de la log-vraisemblance du modèle augmenté d'une pénalité sur la norme-1 du vecteur de paramètres, comme suit :

$$\hat{\beta}_{LASSO} \in_{\beta \in \mathbb{R}^d} \sum_{i=1}^{n} \Big( \ln(1 + e^{X_i'\beta}) - Y_i(X_i'\beta) \Big) + \lambda ||\beta||_1$$

Les variables qu'il est pertinent d'envisager pour définir les domaines peuvent être définies comme celles dont les coefficients associés estimés sont non nuls pour un $\lambda$ donné, à savoir :

$$\hat{S}_Y^{\lambda} = \Big\{ X_{.p} : p = 1, ..., P, \hat{\beta}_{\lambda,p}^{LASSO} \neq 0 \Big\}$$

Enfin, il est important de noter que l'estimateur LASSO est uniquement utilisé à des fins de sélection de variables. En particulier, une variable ayant un effet bien présent mais modéré sur la probabilité d'être contrôlée peut ne pas être sélectionnée. C'est également le cas des variables explicatives très corrélées entre elles : le LASSO en sélectionnera une au détriment des autres. Finalement, sélectionner des variables explicatives et estimer leur effet sur la variable d'intérêt sont deux objectifs distincts qui ne peuvent être atteints simultanément.

# 6 Résultats complémentaires

TABLE 9 – Performances des algorithmes de prédiction des probabilités de contrôle

| | Arbre | Bagging | Forêts aléatoires | Boosting | Elasticnet |
|---|---|---|---|---|---|
| **DVNI** | | | | | |
| **Critère d'ajustement standard** | | | | | |
| Précision de la prédiction | 0.708 | 0.718 | 0.720 | 0.727 | 0.730 |
| **Critères basés sur l'AUC** | | | | | |
| AUC | 0.750 | 0.764 | 0.762 | 0.755 | 0.774 |
| Précision | 0.262 | 0.268 | 0.267 | 0.268 | 0.279 |
| Rappel | 0.672 | 0.660 | 0.646 | 0.619 | 0.662 |
| Spécificité | 0.713 | 0.727 | 0.731 | 0.743 | 0.741 |
| **DIRCOFI** | | | | | |
| **Critère d'ajustement standard** | | | | | |
| Précision de la prédiction | 0.680 | 0.699 | 0.701 | 0.708 | 0.701 |
| **Critères basés sur l'AUC** | | | | | |
| AUC | 0.648 | 0.675 | 0.670 | 0.667 | 0.671 |
| Précision | 0.207 | 0.221 | 0.220 | 0.221 | 0.222 |
| Rappel | 0.496 | 0.496 | 0.488 | 0.471 | 0.498 |
| Spécificité | 0.708 | 0.730 | 0.734 | 0.745 | 0.732 |
| **Directions locales** | | | | | |
| **Critère d'ajustement standard** | | | | | |
| Précision de la prédiction | 0.786 | 0.781 | 0.807 | 0.799 | 0.738 |
| **Critères basés sur l'AUC** | | | | | |
| AUC | 0.769 | 0.789 | 0.787 | 0.797 | 0.758 |
| Précision | 0.055 | 0.057 | 0.061 | 0.060 | 0.048 |
| Rappel | 0.587 | 0.625 | 0.592 | 0.608 | 0.626 |
| Spécificité | 0.790 | 0.784 | 0.811 | 0.803 | 0.740 |

AUC : *Area under the curve*, aire sous la courbe ROC.
Note : Chaque colonne présente, pour chaque algorithme testé, les résultats obtenus, sur différents critères de performance, pour les trois sous-échantillons de test des sous ensembles de directions fiscales.
Source : DGFiP, Insee, calcul des auteurs.

TABLE 10 – Résultats avec l'estimateur par le ratio (en Md€)

| Probabilités de sondage<br>Domaines | (I)<br>SAS* | (II)<br>GCH** | (III)<br>GCH<br>(Naf) | (IV)<br>GCH<br>(quantiles TVA) |
|---|---|---|---|---|
| DVNI | 0,66<br>[0,55;0,77] | 1,10<br>[0,91;1,28] | 1,36<br>[1,15;1,56] | 1,14<br>[1,01;1,26] |
| Dircofi | 1,79<br>[1,70;1,88] | 1,83<br>[1,68;1,99] | 1,84<br>[1,68;1,99] | 2,55<br>[2,37;2,72] |
| Locales | 10,50<br>[9,27;11,73] | 19,28<br>[17,34;21,23] | 19,52<br>[18,19;20,85] | 16,39<br>[15,63;17,15] |
| Total | 12,95<br>[11,72;14,19] | 22,21<br>[20,26;24,17] | 22,72<br>[21,36;24,07] | 20,08<br>[19,29;20,86] |

* : Sondage aléatoire simple, ** : Groupe de contrôle homogène

Note : Chaque colonne présente les résultats obtenus pour les trois sous ensembles de directions fiscales, en utilisant un estimateur par le ratio, sur l'ensemble de la population ou sur chaque domaine. Dans la colonne (I), l'hypothèse retenue est que les entreprises contrôlées sont issues d'un tirage aléatoire simple. Dans les colonnes (II), (III) et (IV), la probabilité d'inclusion de chaque entreprise correspond à la proportion d'entreprises contrôlées du groupe de contrôle homogène à laquelle elle appartient (cf. section 2.2). Les 50 quantiles de TVA qui définissent les domaines dans la quatrième colonne sont constitués à partir de la distribution, pour les seules entreprises contrôlées, de la TVA brute déclarée. Pour chaque estimation, l'intervalle de confiance à 95 % est donné.

Champ : Entreprises ayant effectué une déclaration de TVA en 2012.

Source : DGFiP, Insee, calcul des auteurs.

TABLE 11 – Résultats avec l'estimateur par la moyenne (en Md€)

| Probabilités de sondage<br>Domaines | (I)<br>SAS* | (II)<br>GCH** | (III)<br>GCH<br>(Naf) | (IV)<br>GCH<br>(quantiles TVA) |
|---|---|---|---|---|
| DVNI | 1, 27<br>[1,13;1,41] | 0, 92<br>[0,82;1,02] | 0, 95<br>[0,84;1,06] | 0, 80<br>[0,70;0,90] |
| Dircofi | 1, 79<br>[1,71;1,87] | 1, 84<br>[1,71;1,98] | 1, 85<br>[1,71;1,99] | 1, 80<br>[1,63;1,96] |
| Locales | 29, 36<br>[28,32;30,40] | 21, 78<br>[20,70;22,86] | 21, 35<br>[20,19;22,50] | 20, 17<br>[19,06;21,28] |
| Total | 32, 42<br>[31,37;33,47] | 24, 55<br>[23,45;25,64] | 24, 15<br>[22,98;25,32] | 22, 77<br>[21,64;23,90] |

* : Sondage aléatoire simple, ** : Groupe de contrôle homogène

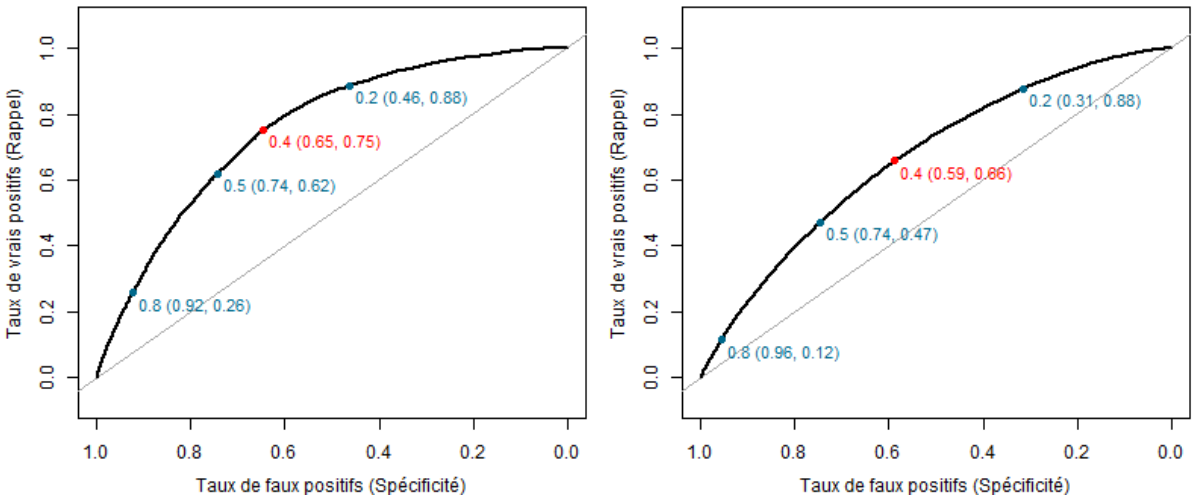
Note : Chaque colonne présente les résultats obtenus pour les trois sous ensembles de directions fiscales, en utilisant un estimateur par la moyenne, sur l'ensemble de la population ou sur chaque domaine. Dans la colonne (I), l'hypothèse retenue est que les entreprises contrôlées sont issues d'un tirage aléatoire simple. Dans les colonnes (II), (III) et (IV), la probabilité d'inclusion de chaque entreprise correspond à la proportion d'entreprises contrôlées du groupe de contrôle homogène à laquelle elle appartient (cf. section 2.2). Les 50 quantiles de TVA qui définissent les domaines dans la quatrième colonne sont constitués à partir de la distribution, pour les seules entreprises contrôlées, de la TVA brute déclarée. Pour chaque estimation, l'intervalle de confiance à 95 % est donné.

Champ : Entreprises ayant effectué une déclaration de TVA en 2012.

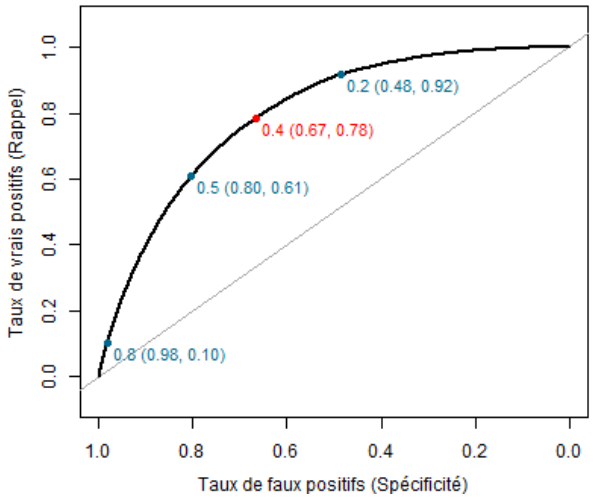
Source : DGFiP, Insee, calcul des auteurs.

# 7 Graphiques complémentaires

FIGURE 7 – Courbes ROC des algorithmes de *boosting*



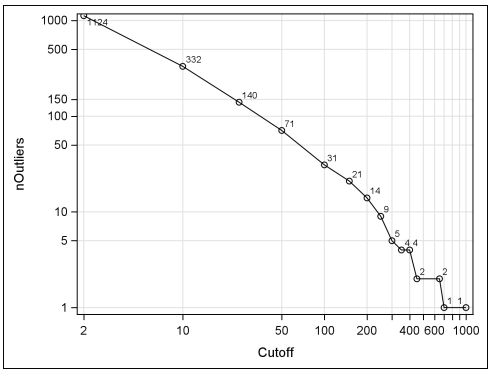

(a) DVNI (b) Dircofi



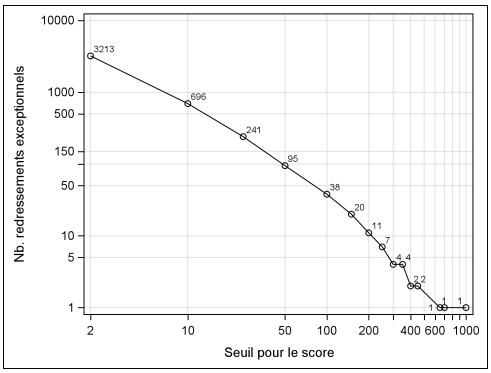

(c) Directions locales

Note : Chaque point correspond à un seuil de la règle de décision, et renseigne, entre parenthèses, les valeurs correspondantes, respectivement, pour les critères de spécificité et de rappel/sensibilité. À titre informatif, est aussi donné le point correspondant au seuil optimal pour la règle de décision (en rouge), au sens où il maximise la somme du rappel et de la spécificité.

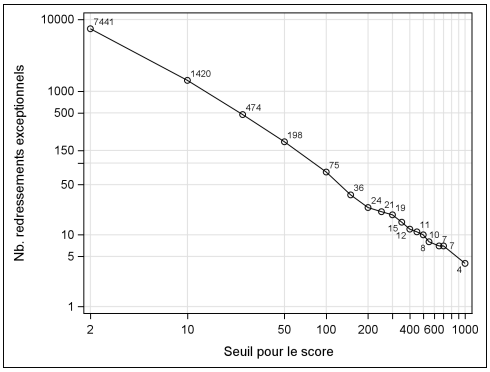
FIGURE 8 – Nombre de redressements exceptionnels en fonction du seuil du score retenu




(a) DVNI



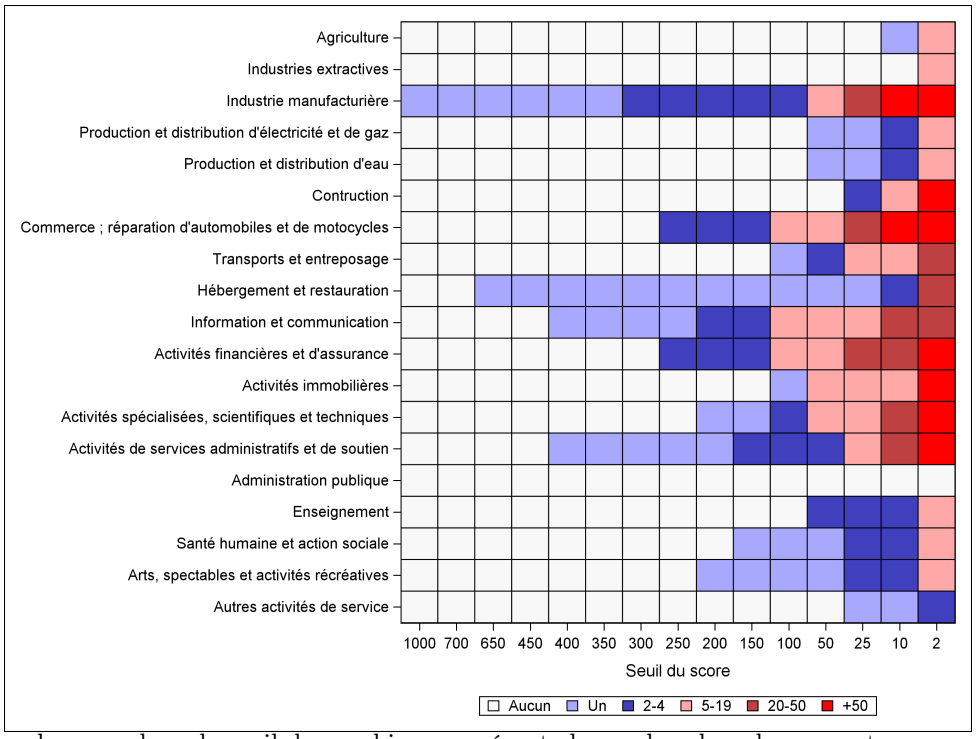

(b) Dircofi




(c) Directions locales

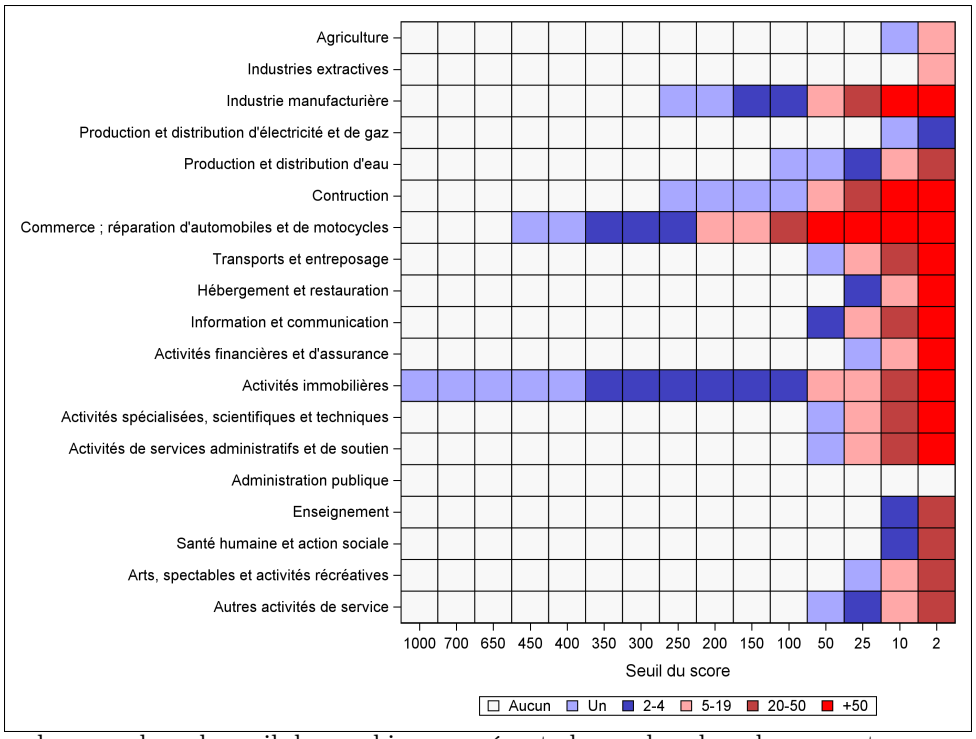
Note : Pour chaque valeur de seuil, le graphique représente le nombre de redressements « exceptionnels » écartés de l'échantillon des entreprises contrôlées utilisé pour l'estimation.
Champ : Entreprises redressées sur leur exercice comptable 2012.
Source : DGFiP, Insee, calcul des auteurs.

FIGURE 9 – Nombre de redressements exceptionnels en fonction du seuil du score retenu par Naf - DVNI




Note : Pour chaque valeur de seuil, le graphique représente le nombre de redressements « exceptionnels » écartés de l'échantillon des entreprises contrôlées utilisé pour l'estimation.
Champ : Entreprises redressées sur leur exercice comptable 2012.
Source : DGFiP, Insee, calcul des auteurs..

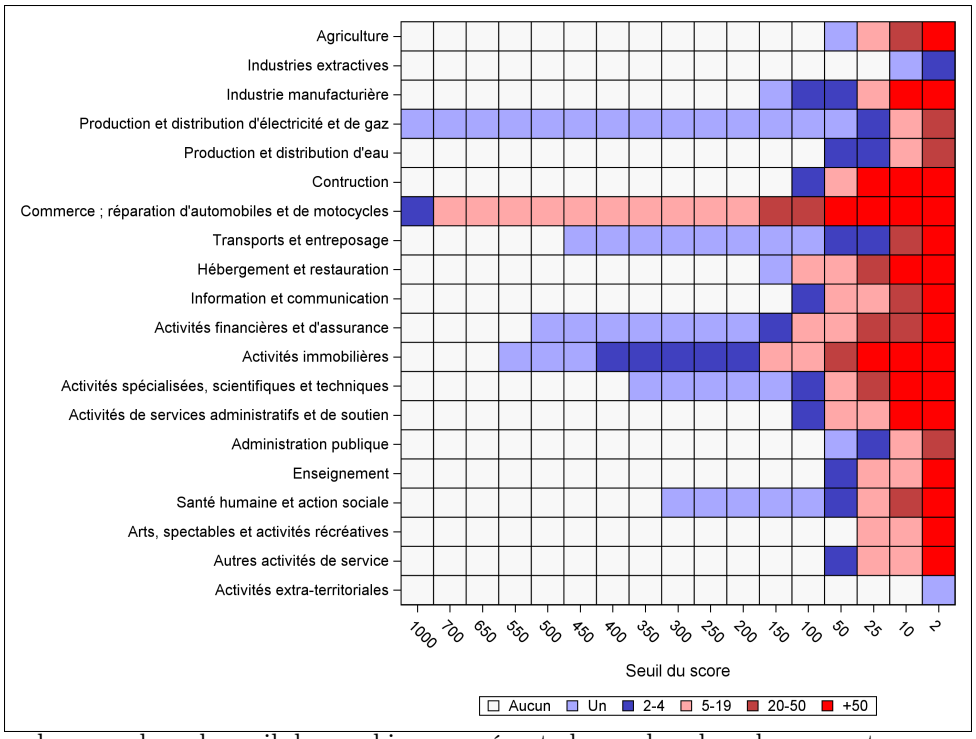
FIGURE 10 – Nombre de redressements exceptionnels en fonction du seuil du score retenu par Naf - Dircofi




Note : Pour chaque valeur de seuil, le graphique représente le nombre de redressements « exceptionnels » écartés de l'échantillon des entreprises contrôlées utilisé pour l'estimation.
Champ : Entreprises redressées sur leur exercice comptable 2012.

Source : DGFiP, Insee, calcul des auteurs..

FIGURE 11 – Nombre de redressements exceptionnels en fonction du seuil du score retenu par Naf - Directions locales




Note : Pour chaque valeur de seuil, le graphique représente le nombre de redressements « exceptionnels » écartés de l'échantillon des entreprises contrôlées utilisé pour l'estimation.
Champ : Entreprises redressées sur leur exercice comptable 2012.
Source : DGFiP, Insee, calcul des auteurs..

# Références


[1] Austin, P., and Stuart, E. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine 34* (08 2015).

[2] Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition 30*, 7 (1997), 1145–1159.

[3] Breiman, L., Friedman, J., Olsen, R., and Stone, C. *Classification and Regression Trees*. Wadsworth and Brooks, 1984.

[4] Chawla, N., Bowyer, K., Hall, L., and Kegelmeyer, W. SMOTE : Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research 16* (June 2002), 321–357.

[5] Cour des comptes. La fraude aux prélèvements obligatoires - Rapport, 2019.

[6] Dangerfield, B. J., and Morris, J. S. Top-down or bottom-up : Aggregate versus disaggregate extrapolations. *International journal of forecasting 8*, 2 (1992), 233–241.

[7] Drummond, C., and Holte, R. C. Explicitly Representing Expected Cost : An Alternative to ROC Representation. In *In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2000), ACM Press, pp. 198–207.

[8] Friedman, J., Hastie, T., and Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software 33*, 1 (2010), 1–22.

[9] Friedman, J. H. Greedy function approximation : A gradient boosting machine. *Annals of Statistics 29*, 5 (October 2001), 1189–1232.

[10] Haziza, D., and Beaumont, J.-F. On the Construction of Imputation Classes in Survey. *International Statistical Review 75*, 1 (2007), 25–43.

[11] Haziza, D., Chen, S., and Gao, Y. Targeting Key Survey Variables at the Unit Nonresponse Treatment Stage. *Journal of Survey Statistics and Methodology* (11 2020).

[12] Heckman, J. The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. *Annals of Economic and Social Measurement 5*, 4 (1976), 475–492.

[13] Heckman, J. Sample Selection Bias as a Specification Error. *Econometrica 47*, 1 (1979), 153–161.

[14] Imbens, G., Hirano, K., and Ridder, G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*, 4 (2003), 1161–1189.

[15] Kuhn, M. *caret : Classification and Regression Training*, 2020. R package version 6.0-86.

[16] Liaw, A., and Wiener, M. Classification and Regression by randomForest. *R News 2*, 3 (2002), 18–22.

[17] Ling, C., , Ling, C. X., and Li, C. Data Mining for Direct Marketing : Problems and Solutions. In *In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)* (1998), AAAI Press, pp. 73–79.

[18] Louvot, C. L'évaluation de l'activité dissimulée des entreprises sur la base des contrôles fiscaux et son insertion dans les comptes nationaux. *Documents de travail, Insee*, G2011/09 (2011).

[19] OECD. Tax administration 2017 : Comparative Information on OECD and Other Advanced and Emerging Economies. Tech. rep., OECD, 2017.

[20] Provost, F., and Fawcett, T. Robust Classification for Imprecise Environments. *Machine Learning 42* (01 2001), 203–231.

[21] Rousseeuw, P. J., and Hubert, M. Robust statistics for outlier detection. *WIREs Data Mining and Knowledge Discovery 1*, 1 (2011), 73–79.

[22] Sautory, Olivier. Les méthodes de calage, Note méthodologique INSEE, 2018.

[23] Syndicat National Solidaires Finances Publiques. Rapport du syndicat national Solidaires Finances Publiques : la fraude fiscale nuit gravement.... Tech. rep., Syndicat National Solidaires Finances Publiques, 2019.

[24] Särndal, C.-E., and Lundstrom, S. *Estimation in surveys with nonresponse.* John Wiley & Sons, 2005.

[25] Särndal, C.-E., Swensson, B., and Wretman, J. *Model assisted survey sampling.* Springer Science and Business Media, 2003.

[26] Tagliaferri, G., Scacciatelli, D., and Alaimo Di Loro, P. VAT tax gap prediction : a 2-steps Gradient Boosting approach, 12 2019.

[27] Therneau, T., and Atkinson, B. *rpart : Recursive Partitioning and Regression Trees*, 2019. R package version 4.1-15.

[28] Thomas Deroyon. La correction de la non-réponse par repondération, Note méthodologique INSEE, 2018.

[29] Torgo, L. *Data Mining with R, learning with case studies.* Chapman and Hall/CRC, 2010.

[30] Wendling, C., and Gomez, F. Sécurisation du recouvrement de la TVA. Tech. rep., Inspection Générale des Finances, 2019.

[31] Zadrozny, B. Learning and evaluating classifiers under sample selection bias. *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004 2004* (09 2004).