Estimation des montants manquants de versements de TVA

Exploitation des données du contrôle fiscal

Simon Quantin (Insee), Cécile Welter-Médée (Insee, CREST)

Mercredi 30 mars 2022

SESSION 6 - Non-réponse, imputation et machine learning

Journées de méthodologie statistique de l'Insee 2022

Plan

Contexte et remarques liminaires

Cette étude mentionne des **montants manquants de versements de TVA**, et n'estime pas la *fraude à la TVA* à proprement parler.

- Pas connaissance de l'intention du contribuable (erreur ou comportement frauduleux).
- Prise en compte des montants notifiés positifs uniquement : l'existence de notifications en faveur du contribuable (montants négatifs) confirme la possibilité d'erreurs de bonne foi, il existe certainement des montants positifs notifiés du même ordre de grandeur qui correspondent à des erreurs de bonne foi.

Par ailleurs, nous estimons les **montants notifiés par l'administration fiscale**, mais :

- ni les montants recouvrés par l'administration fiscale,
- l'activité dissimulée sous-jacente.

Contexte

La **méthodologie d'extrapolation** en quelques lignes et analogie avec le redressement de la non-réponse :

- on dispose des données issues des **contrôles** menés par l'administration fiscale, sans connaître le processus de sélection des entreprises contrôlées,
- or parmi les entreprises contrôlées, plus de 60 % se voient notifier un montant de redressement positif dans certains secteurs d'activités : a priori les contrôleurs fiscaux savent **cibler** les entreprises fraudeuses : il y a un **biais de sélection**.

  => *C'est ici qu'intervient l'analogie avec le redressement de la non-réponse :*

- on considère les entreprises contrôlées comme des entreprises enquêtées **qui sont répondantes**. Les entreprises **contrôlables et non contrôlées** sont assimilées à des entreprises enquêtées et non répondantes : on va par suite estimer les probabilités de contrôle (comme on estimerait des probabilités de réponse à une enquête) pour essayer de corriger du biais de sélection.
- On estimera enfin à partir des entreprises contrôlées un montant total de TVA "manquante" *par mois d'activité*, et *par domaine* (section Naf), en tenant compte des probabilités de contrôle.

La base de gestion ALPAGE

La base de gestion ALPAGE recense les dossiers de contrôles fiscaux effectués sur place (et non sur pièces), réalisés par la DGFiP. Un dossier correspond au contrôle (i) d'une entreprise, (ii) sur une période comptable (iii) par un contrôleur fiscal.

Chaque dossier précise :

- la **nature de l'opération**, permet de savoir quel type de contrôle est effectué (examen de comptabilité, ou vérification ponctuelle d'un impôt),
- les **motifs de redressement**, appelés « codes thesaurus »,
- les **montants notifiés** associés,
- la **période comptable** contrôlée,
- le **siren** de l'entreprise (si existence légale).

En revanche, ne sont pas connus :

- le **motif** du contrôle,
- les **impôts effectivement contrôlés**,
- les **corrections comptables** correspondantes (par ex. sous déclaration de base imposable, taux erroné, etc.).

De la période contrôlée à l'exercice comptable - Choix de l'année d'estimation

Lors d'un contrôle fiscal, les entreprises sont possiblement contrôlées au titre de plusieurs années (en moyenne 3 ans) et peuvent se voir notifier un montant total de redressement pour l'ensemble de la période.

1. Une entreprise est considérée comme **contrôlée une année donnée**, dès lors qu'au moins un mois de cette année a été contrôlé.
2. Si plusieurs redressements sont notifiés à différentes années pour une même période contrôlée, ceux-ci sont agrégés.
3. **Ventilation du montant total notifié au prorata du nombre de mois contrôlé pour l'année considérée**, dans la période totale contrôlée (en mois). En répartissant les montants redressés sur toute la période contrôlée, les montants/comportements de fraude sont lissés au cours du temps $\rightarrow$ il n'est pas possible de comparer rigoureusement des années.

- Environ 46 000 dossiers de contrôles instruits par an,
- environ 140 000 dossiers contrôlés au titre d'une année d'exercice donnée.

On réalise des estimations au titre de l'année d'exercice 2012, année la plus récente pour laquelle on considère ne plus avoir de contrôle en cours.

Entreprises contrôlées au titre de la TVA et entreprises contrôlables

I - Entreprises contrôlées au titre de la TVA

Certains contrôles n'aboutissent pas à la notification d'un redressement : *comment identifier une entreprise contrôlée au titre de la TVA, mais non redressée ?*

=> via la **nature de l'opération** (vérification générale, vérification simple ou ponctuelle d'un impôt spécifique, examen de comptabilité, etc...).

Sont considérées comme contrôlées à la TVA, les entreprises contrôlées au titre :

- d'une **vérification simple de TVA**,
- d'une **vérification ponctuelle de TVA**,
- d'une **vérification générale** (hypothèse peu restrictive mais raisonnable ; en pratique, une VG n'implique pas que tous les impôts soient effectivement vérifiés).

II - Entreprises contrôlables

Chaque année, le champ des entreprises **contrôlables au titre de la TVA** correspond aux **entreprises ayant effectué une déclaration de TVA**.

En 2012, parmi les entreprises contrôlables, 3,7 % sont contrôlées au titre de la TVA par l'administration fiscale, parmi lesquelles, 61,8 % se voient notifier un montant de redressement positif.

- Faible proportion d'entreprises contrôlées,
- Ciblage des contrôles vers cas avec forte suspicion de redressement : c'est un **biais de sélection**.

L'organisation du contrôle fiscal

Le contrôle fiscal est assuré par trois niveaux de contrôle, **national**, **interrégional**, et **local**, qui correspondent à une segmentation du tissu fiscal des entreprises selon leur CA (en distinguant ventes de biens ou prestations de services) et leur actif brut.

- 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, ainsi que par leurs filiales.
- Au niveau interrégional, les **Directions spécialisées de contrôle fiscal** (Dircofi) contrôlent les entreprises de taille moyenne relevant de leur ressort territorial.
- Au niveau départemental, le contrôle fiscal des petites entreprises est assuré par les **Directions Départementales des Finances Publiques** et les **Directions Régionales des Finances Publiques**.

Tableau 1 – Entreprises contrôlées et redressées sur leur exercice comptable de l'année 2012

	Nb. d'ent. redevables	Contrôlées (%)	Redressées (%)	Redressées Montant moyen (€)
DVNI	92 416	13,1	41,8	2 689
Dircofi	226 257	13,3	58,3	1 505
Directions locales	3 141 081	2,1	68,0	1 183

Note : 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

Plan

Vue d'ensemble de la méthodologie

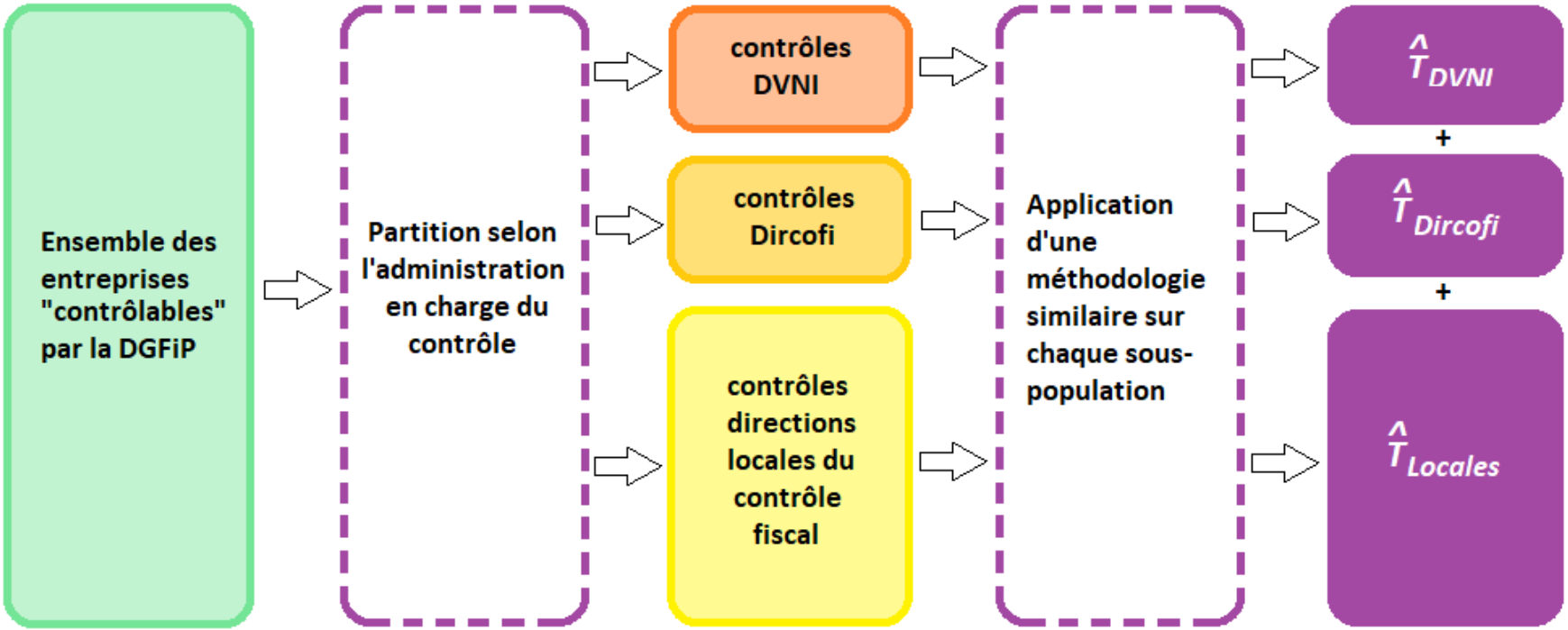
Reconstitution du processus de sélection des entreprises contrôlées - Correction du biais de sélection :

1. découpage de la population en trois sous-populations selon l'administration en charge du contrôle,
2. estimation d'une probabilité d'être contrôlée à partir des informations issues des déclarations de TVA à l'aide d'un algorithme de *boosting*,
3. constitution de 50 (DVNI et Dirccofi) à 100 (Directions locales) *groupes de contrôle homogène* (GCH),
4. attribution à chaque entreprise du taux de contrôle observé dans son GCH comme pondération.

Extrapolation :

1. partition en domaines selon la NAF pour chacune des trois sous-populations,
2. application de deux estimateurs classiques en théorie des sondages : un estimateur par le ratio (qui mobilise le montant de TVA brute déclarée) et un estimateur par la moyenne (qui s'appuie sur le nombre de mois d'activité),
3. calcul du total estimé par somme des estimateurs par sous-population.

Partition de la population en trois sous-populations



Estimation des probabilités de contrôle et constitution des GCH

Il est d'usage à l'INSEE de déterminer des *groupes de réponse homogène* par des méthodes de *machine learning* basées sur des arbres de décision en redressement de la non-réponse par repondération dans les enquêtes auprès des entreprises (Sarndal, 2003[5] ; Deroyon, 2018[6]).

On procède de façon similaire : on détermine des GCH à partir de **probabilités de contrôle** prédites par des méthodes de *machine learning* basées sur des arbres de décision.

- Découpage par étapes successives de la population 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.

- Permet d'intégrer de façon sous-jacente des facteurs économiques importants (taille de l'entreprise) permettant de retracer (partiellement) l'hétérogénéité de nature des entreprises, en plus de leur **propension à être contrôlées** par l'administration fiscale.

Utilisation d'algorithmes de *machine learning* particulièrement adaptée à notre cas :

- décision d'un contrôle partiellement alimentée par un algorithme de détection de situations potentiellement fraudogènes + difficulté de formaliser une règle de décision à partir de l'expertise des contrôleurs fiscaux.
- Faire appel à des méthodes de *machine learning* plutôt qu'à des approches économétriques permet de tenir compte au mieux de la complexité du processus de sélection, sans que ses déterminants ne soient interprétés ou même explicités.

.

I - Entraînement et calibrage des algorithmes

On découpe la population en deux : un **échantillon d'entraînement** et un **échantillon de validation** [1].

- **Échantillon d'entraînement** : on calibre par **validation croisée** puis on entraîne plusieurs algorithmes (arbre de classification, méta-algorithmes de *bagging*, forêts aléatoires et *boosting*, ainsi qu'une régression pénalisée elastic-net).
- **Échantillon de validation** : les performances (basées sur plusieurs indicateurs cette fois-ci) de l'algorithme ainsi calibré sur le sous-échantillon d'entraînement sont ensuite évaluées.

Situation d'**échantillon déséquilibré** : le contrôle d'une entreprise est rare dans la population (3,1 % des entreprises « contrôlables » au titre de 2012) $\longrightarrow$ qualités prédictives réduites des algorithmes.

- **Rééquilibrage des échantillons d'entraînement** à l'aide de l'algorithme **SMOTE** (Synthetic Minority Over-sampling TEchnique, voir Chawla et al. (2002)[1]) : accroissement de la classe minoritaire à partir d'entreprises synthétiques, créées à partir des entreprises contrôlées existantes par combinaison linéaire des caractéristiques des k-plus-proches-voisins (individus réels et synthétiques plus nombreux et mieux répartis dans la région à prédire, permet une meilleure généralisation).

1. L'échantillon d'entraînement représente 60 % de la population totale pour les DVNI et les Dircofi et 20 % pour les directions locales.

II - Estimation des probabilités de contrôle

Les variables explicatives retenues doivent être corrélées au fait d'être contrôlée et aux montants éventuels de redressement (Haziza et al . (2020)[2]). On sélectionne des variables issues des déclarations fiscales de TVA déposées sur les exercices comptables 2011, 2012 et 2013, i.e. aux années $N-1$, N et $N+1$ considérées [2].

III - Constitution des GCH

Les performances des algorithmes testés sont comparées pour sélectionner le plus performant : on retient **l'algorithme de boosting**.

Les GCH sont construits à partir des quantiles de la distribution des probabilités d'être contrôlée prédites des entreprises contrôlées (permet de s'assurer que chaque GCH comporte un minimum d'entreprises de l'échantillon). En pratique sont constitués :

- **50 GCH** pour les entreprises dépendant de la **DVNI** et des **Dircofi**,
- **100 GCH** pour les entreprises dépendant des **directions locales**.

Enfin, on attribue à chaque entreprise le **taux de contrôle observé** dans son GCH comme pondération individuelle (permet de se départir de la forme fonctionnelle de l'algorithme retenu, et d'éviter en pratique les probabilités de contrôle trop faibles qui pourraient constituer des points influents au moment de l'extrapolation).

2. Variables continues (montants moyens sur les 3 années) : TVA brute totale déclarée, acquisitions intracommunautaires, livraisons intracommunautaires, TVA déductible, CA réalisé à l'exportation, TVA déductible sur les biens constituant des immobilisations, crédits antérieurs non imputés et non remboursés, TVA à déduire "autre" (dont régularisation sur de la TVA collectée ou déductible) ; Variables discrètes : indicatrice de TVA brute totale déclarée nulle sur les 3 années, indicatrice de TVA déductible déclarée nulle sur les 3 années, secteur d'activité (21 positions), CJ.

Extrapolation (I) - Choix des domaines

Estimations par domaine

Réaliser une estimation sur un **domaine** permet de tenir compte de la taille de ce domaine dans la population totale dans les calculs de variance (taille de l'échantillon déterministe, mais taille du domaine dans l'échantillon aléatoire).

$\Longrightarrow$ Toutes nos estimations seront réalisées sur le **domaine** constitué de l'ensemble des entreprises *contrôlables* par la DGFiP. Nous réalisons également une estimation sur une partition en sous-domaines selon les **secteurs d'activité** au niveau section de la NAF.

- **Intérêt** : choix de variables positivement corrélée(s) à la variable d'intérêt à extrapoler pour créer des domaines permet une **différenciation importante des comportements de fraude entre eux**, et une **homogénéité de ces comportements en leur sein** $\longrightarrow$ les secteurs d'activité sont bien **corrélés à des déterminants du fait de se faire notifier des montants positifs re rectifications** (conditionnellement au fait d'être contrôlé),

- l'estimation d'une régression logistique avec pénalisation LASSO pour prédire la notification d'un montant positif parmi les *entreprises contrôlées* met en évidence comme principaux déterminants les **secteurs d'activité**,

Extrapolation (II) - Montants exceptionnels

Considérations autour des montants de rectifications exceptionnels

Les extrapolations par domaine reposent sur les montants de rectification en base prononcés à l'encontre des entreprises contrôlées, or **certains montants exceptionnellement élevés de VA dissimulée traduisent un comportement de fraude "particulier"** à une entreprise, qu'on ne souhaite pas extrapoler à d'autres.
Ces rectifications exceptionnelles sont détectées à partir d'un score (Rousseeuw 2011[3]) défini par secteur d'activité comme suit :

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

où MAD correspond à la médiane des valeurs absolues des écarts à la médiane (pour *median of all absolute deviations from the median*).

Les entreprises correspondant **aux dix scores les plus élevés parmi toutes les entreprises contrôlés sont écartées**.

Extrapolation (III) - Estimateurs

I - Estimateur par le ratio

- estimateur qui s'appuie sur de l'**information auxiliaire**, et est particulièrement performant quand la variable auxiliaire (qui doit être parfaitement connue dans la population) est *a priori* corrélée avec la variable d'intérêt à extrapoler. On retient comme variable auxiliaire le **montant de TVA brute déclarée** par les entreprises : connue pour l'ensemble des entreprises qui font une déclaration de TVA (sauf dans le cas d'une activité non déclarée), et très corrélée aux montants de redressement notifiés.

- L'estimateur par le ratio (cf. note méthodologique INSEE [4]) est 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.

II - Estimateur par la moyenne

- Estimateur qui s'appuie sur la crédibilité de l'extrapolation du **montant moyen de redressement notifié**.

- Un contrôle fiscal porte généralement sur plusieurs exercices comptables. Le montant de rectification éventuellement notifié à l'issue du contrôle n'est pas rattaché à l'/aux exercice(s) comptable(s) qui l'a/ont occasionné. On calcule alors un montant de rectification *mensuel* moyen. On multiplie le montant mensuel moyen de rectification estimé dans le domaine par la durée des exercices comptables déclarés par les entreprises.

- estimateur asymptotiquement sans biais : nous l'appliquons à des domaines suffisamment larges pour pouvoir le considérer comme sans biais.

Estimateur par le ratio

- Soit une **sous-populations U**, partitionnée en H **domaines** $U_1, ..., U_H$
- Soit **S** l'échantillon correspondant, également partitionné en H $(S_1, ..., S_H)$.
- y_k le montant notifié de **chiffre d'affaire éludé** d'une entreprise k (potentiellement nul) sur la période contrôlée de l'année considérée (au plus 12 mois).
- x_k la **TVA brute déclarée** correspondant à la même période. 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 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 (asymptotiquement sans biais).

L'estimateur par le ratio du total des y_k dans le domaine U_h s'écrit comme suit :

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

On estime le total de la variable y dans la sous-population considérée par :

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

▸ Précision et intervalle de confiance

Estimateur par la moyenne

- Soit une **sous-populations U**, partitionnée en H **domaines** $U_1, ..., U_H$
- Soit **S** l'échantillon correspondant, également partitionné en H $(S_1, ..., S_H)$.
- y'_k le montant notifié de **chiffre d'affaire éludé** d'une entreprise k (potentiellement nul) sur la période contrôlée de l'année considérée **mensualisé**.

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

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

- Soit d_i la durée d'exercice qui figure dans la déclaration de TVA de l'entreprise i de l'année considérée, en mois et $\mathbf{D}_h$ le total de ces durées d'exercice (sans interprétation économique).

Alors l'estimateur par la moyenne du total des y_k sur un domaine U_h est :

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

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 :

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

▸ Précision et intervalle de confiance

Plan

Performances des algorithmes de *Boosting*

▸ Mesures d'ajustement des prédictions

Tableau 2 – Performances des algorithmes de *boosting* de prédiction des probabilités de contrôle sur l'échantillon de validation

	DVNI	Dircofi	Directions Locales
Précision de la prédiction (*Accuracy*)	72.7	70.8	79.9
Rappel/Sensibilité	61.9	47.1	60.8
Spécificité	74.3	74.5	80.3
Précision	26.8	22.1	6.0

Précision de la prédiction/Accuracy : l'algorithme prédit correctement respectivement 73 %, 61 % et 80 % des entreprises dépendant de la DVNI, des Dircofi et des directions locales *(attention, mesure peu adaptée aux échantillons déséquilibrés)*.

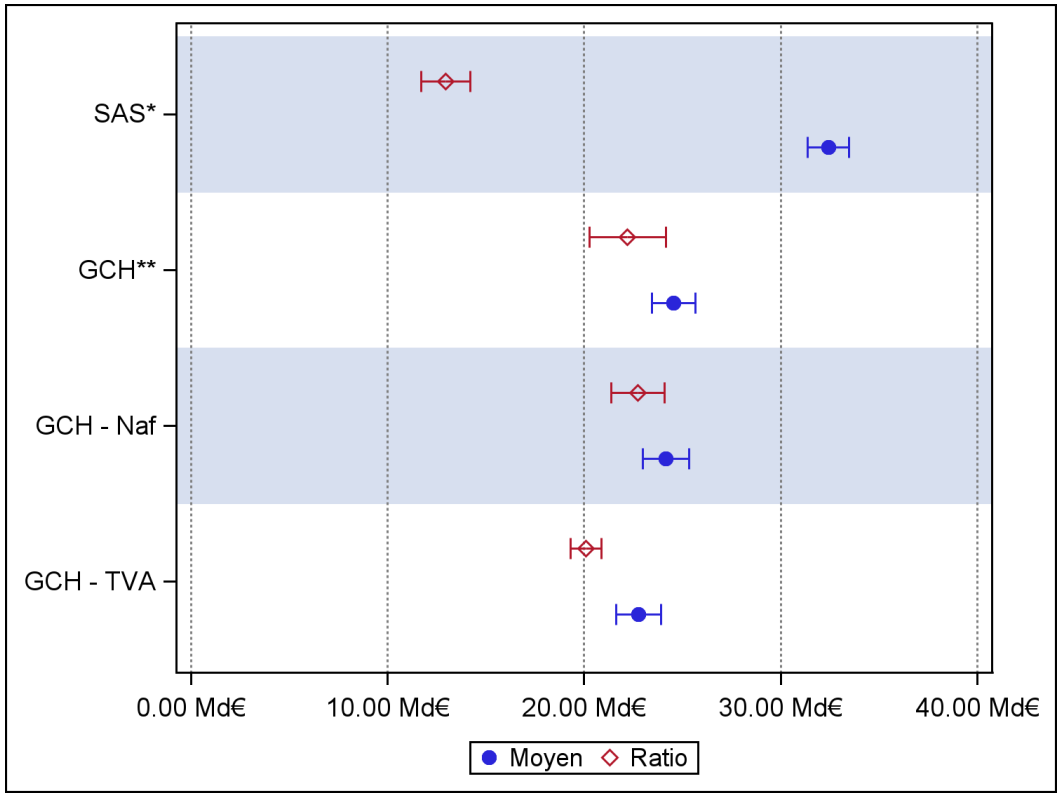
Rappel : l'algorithme prédit un contrôle fiscal pour respectivement 62 %, 47 % et 61 % des entreprises qui seront effectivement contrôlées par la DVNI, les Dircofi et les directions locales.

Spécificité : l'algorithme prédit l'absence de contrôle fiscal pour respectivement 74 %, 74 % et 80 % des entreprises qui ne seront effectivement pas contrôlées par la DVNI, les Dircofi et les directions locales.

Précision : 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.

▸ Bonus : distributions des probabilités prédites

Graphique 1 – 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), on fait l'hypothèse que les entreprises contrôlées sont issues d'un SAS. Dans les lignes (GCH), la probabilité d'inclusion de chaque entreprise correspond à la proportion d'entreprises contrôlées de son GCH. 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.

Tableau 3 – Résultats avec l'estimateur par le ratio (en Md€)

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

Conclusion

- Quel que soit l'estimateur retenu, le montant total de TVA non recouvré serait compris **entre 20 et 26 Md€**.
- Persistent des **limites** : l'algorithme de *boosting* retenu pour prédire les probabilités de contrôle (et *a fortiori* l'ensemble des algorithmes entraînés) montre des performances limitées, en particulier en termes de précision. La modélisation du processus de sélection à partir des caractéristiques observées est incomplète.
- 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 permettra de répondre à cette difficulté (mise en place courant 2022).
- En s'appuyant sur les résultats des contrôles menés par la DGFiP, notre estimation fait l'hypothèse que les comportements de fraude des entreprises sont *tous détectés* par les services fiscaux => persistance d'un biais de détection.

Bibliographie

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

[2] David Haziza, Sixia Chen, and Yimeng Gao.
Targeting Key Survey Variables at the Unit Nonresponse Treatment Stage.
Journal of Survey Statistics and Methodology, 11 2020.

[3] Peter J. Rousseeuw and Mia Hubert.
Robust statistics for outlier detection.
WIREs Data Mining and Knowledge Discovery, 1(1) :73–79, 2011.

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

[5] Carl-Erik Särndal, Bengt Swensson, and Jan Wretman.
Model assisted survey sampling.
Springer Science and Business Media, 2003.

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

Précision et intervalle de confiance de l'estimateur par le ratio

- Soit N la taille de la sous-population considérée et n la taille de l'échantillon associé,
- Soit $f = \frac{n}{N}$ le taux de sondage dans la population,
- Soit $v_k = w_k I_{U_h}(k) = w_k I(k \in U_h)$ le poids de l'entreprise k relatif au domaine U_h.

Dans un domaine U_h d'une sous-population donnée, on a comme estimateur de la variance :

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

$$g_k = \frac{v_k(y_k - x_k \hat{\mathbf{R}}_h)}{\sum_{k \in U} v_k x_k} \quad \text{et} \quad \bar{g} = \frac{1}{N} \sum_{k \in U} g_k$$

On note par ailleurs : $\hat{\sigma}(\hat{\mathbf{R}}_h) = \sqrt{\hat{V}(\hat{\mathbf{R}}_h)}$

On considère que le ratio suit une loi normale. 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 est ainsi défini comme suit :

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

Précision et intervalle de confiance de l'estimateur par le ratio

A partir du déterminisme du total de la variable x_k dans chaque domaine et en s'appuyant sur le fait que les montants x_k sont positifs ou nuls, on a :

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

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

On en déduit enfin un estimateur de la variance du total estimé sur la sous-population considérée, ainsi que l'IC 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 par somme des trois estimateurs de variances de chaque sous-population (estimateurs indépendants), puis l'IC associé est recalculé à partir de l'estimateur du total (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. ▸ Retour à l'estimateur par le ratio

Précision et intervalle de confiance de l'estimateur par la moyenne

- Soit N la taille de la sous-population considérée et n la taille de l'échantillon associé,
- Soit $f = \frac{n}{N}$ le taux de sondage dans la population,
- Soit $v_k = w_k I_{U_h}(k) = w_k I(k \in U_h)$ le poids de l'entreprise k relatif au domaine U_h.

Dans un domaine U_h d'une sous-population donnée, on a comme estimateur de la variance :

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

$$r_k = \frac{v_k(y'_k - \hat{\bar{Y}}'_h)}{\sum_{k \in U} v_k} \quad \text{et} \quad \bar{r} = \frac{1}{N} \sum_{k \in U} r_k$$

On note par ailleurs $\hat{\sigma}(\hat{\bar{Y}}'_h) = \sqrt{\hat{V}(\hat{\bar{Y}}'_h)}$
On considère que le ratio suit une loi normale. 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 :

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

Précision et intervalle de confiance de l'estimateur par la moyenne A partir du déterminisme de la durée d'exercice d_k, connue pour l'ensemble des entreprises ayant fait une déclaration de TVA et en s'appuyant sur le fait que les durées d_k sont positives ou nulles, on a :

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

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

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

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

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

Finalement, la variance s'obtient par somme des trois estimateurs de variances de chaque sous-population (estimateurs indépendants), puis l'IC associé est recalculé à partir de l'estimateur du total (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. ▸ Retour à l'estimateur par la moyenne

Mesures d'ajustement des prédictions

Les performances de l'algorithme entraîné sont mesurées à partir de l'**échantillon test**.

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

Le **rappel** (ou sensibilité) mesure la proportion de contrôles prédits parmi les contrôles effectivement réalisés, tandis que la **spécificité** mesure à l'inverse la proportion d'absence de contrôles prédits parmi les entreprises non contrôlées :

$$\text{rappel/sensibilité} = \frac{\text{VP}}{\text{VP} + \text{FN}} \quad \text{et} \quad \text{spécificité} = \frac{\text{VN}}{\text{VN} + \text{FP}}$$

La **précision** quantifie la proportion d'entreprises effectivement contrôlées parmi celles qui se sont vues prédire un contrôle fiscal :

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

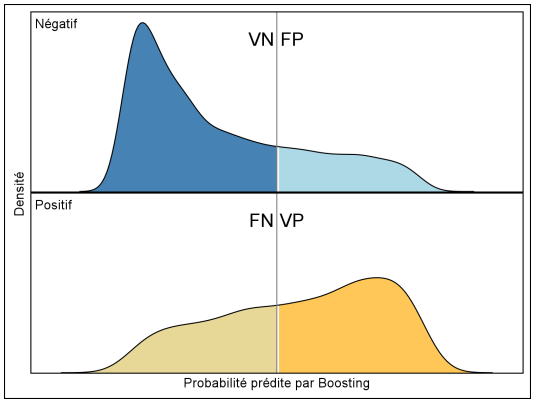
la **précision de la prédiction** (ou **Accuracy**) reflète la proportion de prédictions correctes (mesure très sensible à la proportion d'entreprises qui connaissent un contrôle) :

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

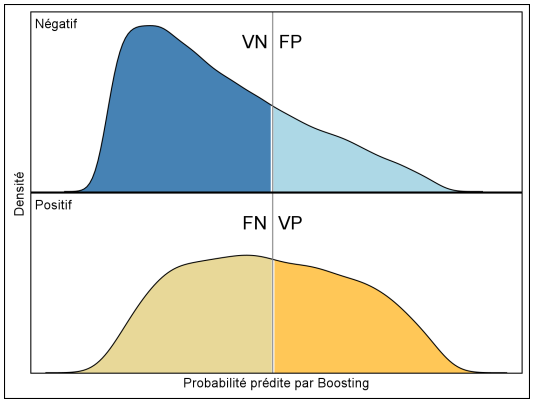
▸ Retour aux résultats du Boosting

Distributions des probabilités prédites par Boosting

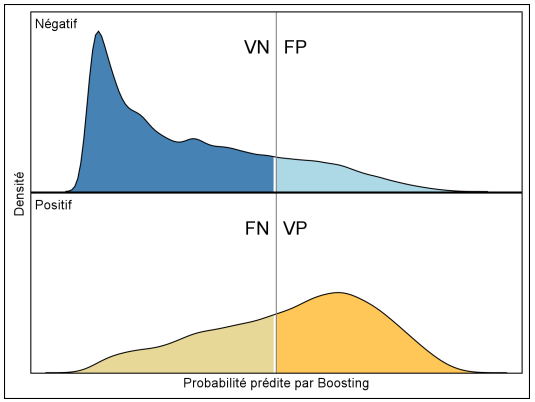
Graphique 2 – Probabilités prédites par *boosting* sur l'échantillon de validation - Vrais/Faux positifs et négatifs



(a) DVNI



(b) Dircofi



(c) Directions locales

▸ Retour aux résultats du Boosting