UNU WIDER

3
24

Tax research for development

Un guide pratique pour l'estimation des écarts fiscaux selon une approche *bottom-up*

Mostafa Bahbah,[1] Sebastián Castillo,[2] Kwabena Adu-Ababio,[3] and Amina Ebrahim[3]

Décembre 2024

WIDER TECHNICAL NOTE

wider.unu.edu

United Nations University
World Institute for Development Economics Research

Résumé: Cette note technique concerne le guide pratique sur les écarts fiscaux, qui comprend un code (fichier Stata) et un fichier README (comment exécuter le code). La note décrit la littérature et la méthodologie derrière le développement du guide. Ce guide pratique des écarts fiscaux est lié à l'estimation des écarts en matière de taxe sur la valeur ajoutée (TVA), d'impôt sur les sociétés (IS) et d'impôt sur le revenu des personnes (IRP) à l'aide de l'approche *bottom-up*, où les audits opérationnels et l'impôt potentiel sont calculés à l'aide du *machine learning*.

Mots-clés: guide pratique, écart fiscal, approche *bottom-up*, audits opérationnels, *machine learning*

JEL classification: H25, H26, H32

Acknowledgements: Les auteurs remercient Jukka Pirttilä, Maria Jouste, Gerald Agaba et Hilja-Maria Takala pour leurs contributions. Les auteurs remercient l'International Tax Compact (ITC) pour son financement. L'ITC facilite le secrétariat de l'Addis Tax Initiative (ATI), qui a soutenu le développement du guide pratique sur les écarts fiscaux. L'ITC est financé par le ministère fédéral allemand de la coopération économique et du développement, cofinancé par l'Union Européenne, et implémenté par le Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH. Ce travail fait partie du programme Domestic Revenue Mobilization financé au travers de contributions spécifiques par Norad.

Matériel supplémentaire : Le code correspondant et les fichiers README peuvent être téléchargés gratuitement à partir de la page web du guide pratique.

Publications connexes :

- Estimating tax gaps in Zambia: WIDER Working Paper 2023/25
- Estimating the value-added tax gap in Tanzania: WIDER Working Paper 2024/66

Cette note technique est disponible en anglais (original), espagnol et portugais.

[1] Tampere University, Finland; [2] University of Helsinki, Finland, and Finnish Centre of Excellence in Tax Systems Research (FIT); [3] UNU-WIDER, Helsinki, Finland; corresponding author: amina@wider.unu.edu

This study has been prepared within the UNU-WIDER project Tax research for development (phase 3), which is part of the research area Creating the fiscal space for development. The project is part of the Domestic Revenue Mobilization programme, which is financed through specific contributions by the Norwegian Agency for Development Cooperation (Norad). The Tax Gap toolkit received financial support from the International Tax Compact (ITC), which facilitates the Secretariat of the Addis Tax Initiative (ATI). The ITC is funded by the German Federal Ministry of Economic Cooperation and Development, co-funded by the European Union, and implemented by the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH.

https://doi.org/10.35188/UNU-WIDER/WTN/2024-2

United Nations University World Institute for Development Economics Research – UNU-WIDER

Katajanokanlaituri 6 B, 00160 Helsinki, Finland

1 Introduction

La mobilisation des recettes nationales pour financer les dépenses publiques est un objectif essentiel pour de nombreux gouvernements. Par conséquent, les autorités fiscales jouent un rôle crucial dans l'élaboration de systèmes fiscaux efficaces et efficients qui peuvent garantir le montant optimal de recettes avec un minimum de fuites. Cependant, les recettes optimales ou maximales réalisables ne sont pas au niveau du résultat souhaité, car les personnes et les entreprises redevables vont jusqu'aux limites extrêmes pour éviter ou se soustraire à ces obligations. Partant de ce constat, la recherche cherche à mesurer l'écart entre les recettes fiscales réalisées et potentielles, afin de quantifier l'ampleur de la perte de recettes qui se produirait, si toutes les personnes et toutes les entreprises respectaient pleinement les règles de la politique fiscale.

Cette note technique illustre la manière dont les écarts fiscaux, définis comme la différence entre les recettes fiscales réalisées et potentielles, sont calculés sur la base d'une méthodologie *bottom-up* en utilisant diverses formes de déclarations et de données statistiques. En outre, cette note accompagne le guide pratique sur les écarts fiscaux en tant que précurseur des concepts généraux des écarts fiscaux, en mettant l'accent sur l'approche *bottom-up* pour estimer ces écarts. Le guide pratique est développé par l'Institut mondial de recherche sur l'économie du développement de l'Université des Nations unies (UNU-WIDER) et tente de simplifier les concepts complexes de cette approche d'estimation en guidant les utilisateurs à travers des méthodes de mesure systématiques avec des paramètres disponibles à leur portée.

La définition des écarts fiscaux étant large, elle s'accompagne d'une grande variété d'approches d'estimation du concept. Néanmoins, elles s'intéressent toutes aux raisons pour lesquelles les recettes fiscales réalisées s'écartent des recettes potentielles. Certaines méthodes générales et couramment utilisées se concentrent sur des indicateurs macroéconomiques agrégés, comme référence pour l'écart, tandis que les méthodes moins utilisées s'appuient sur des données microéconomiques pour l'estimation de l'écart. Le choix dépend de l'accès et de la disponibilité des données administratives, selon les juridictions. Bien que cette note technique présente les principales approches d'estimation des écarts fiscaux, l'accent est mis sur l'utilisation de données microéconomiques provenant d'audits opérationnels (le plus souvent) et lorsque les audits aléatoires sont rares.

Dans la note qui suit, nous commencerons par définir les écarts fiscaux et nous expliquerons comment l'idée se transforme en concept, en fonction de variables d'intérêt spécifiques liées à des paramètres de politique et de respect des règles (section 2). Ensuite, les méthodes générales utilisées pour estimer les écarts fiscaux sont examinées dans la section 3, où l'approche *bottom-up* constitue l'approche de base, décrite sous ses différentes formes et méthodes. Dans la section 4, nous décrivons les composants du guide pratique, qui comprend deux étapes principales : le nettoyage des données et une approche de façon *machine learning* pour l'estimation des écarts fiscaux.

2 Définition de l'écart fiscal

Les autorités fiscales sont souvent confrontées à une différence notable entre les recettes fiscales attendues et celles qui sont réellement perçues. Cette différence, connue sous le nom de perte de recettes, survient principalement lorsque des impôts dus au cours d'une certaine période restent

impayés. Cet impôt dû par les contribuables représente le montant de l'impôt qui pourrait théoriquement être collecté. D'où le concept d'écart fiscal, défini comme la différence entre les recettes effectivement perçues et les recettes fiscales théoriques dans l'hypothèse d'un respect total du code des impôts.

D'un point de vue politique, l'écart fiscal peut être caractérisé plus largement par deux composantes principales : l'écart de conformité et l'écart en matière de politique fiscale. L'écart de conformité correspond à la différence entre les recettes effectivement perçues au cours d'une année donnée et les recettes maximales qui auraient pu être obtenues sur la base des activités économiques qui se sont déroulées au cours de cette période. L'écart en matière de politique fiscale est le résultat de décisions législatives visant à modifier les réglementations fiscales standard, en introduisant des exonérations spécifiques, des déductions ou des taux réduits pour certains cas (Hutton 2017). Les changements dans le cadre politique peuvent entraîner une augmentation ou une diminution de l'écart en matière de politique fiscale. Par exemple, si le seuil d'imposition zéro est relevé, ce qui permet à une plus grande partie du revenu d'être exonérée d'impôt, ou si un taux d'imposition réduit est introduit pour un groupe spécifique de contribuables, tels que les petites entreprises ou les personnes à faible revenu, l'écart en matière de politique fiscale s'accroît, car moins de recettes sont collectées par rapport au maximum potentiel en vertu des règles fiscales standard. D'autre part, l'écart en matière de politique fiscale pourrait également se creuser sans modification du cadre politique, en raison de changements dans la composition de l'assiette fiscale, qui soumettraient une plus grande partie du revenu net au taux d'imposition normal (Barra et al. 2023).

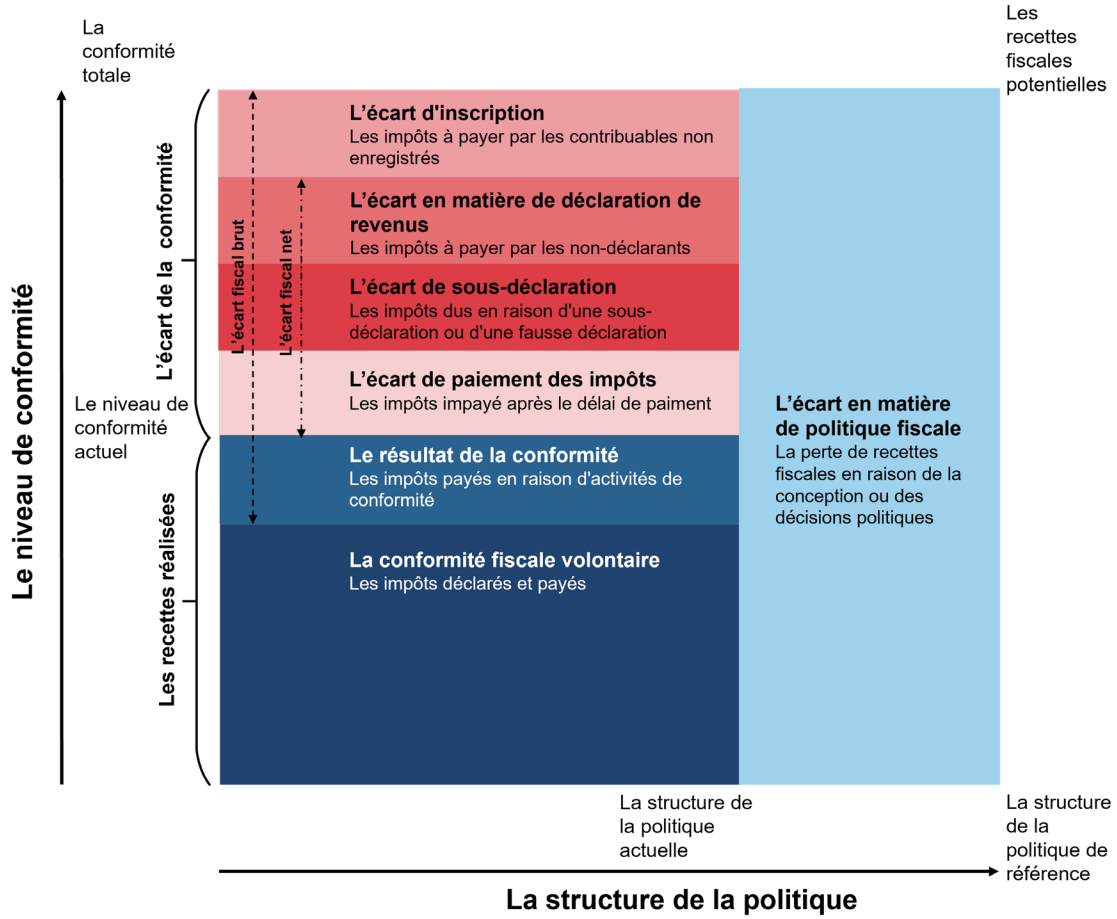
L'écart de conformité se compose de deux éléments : l'écart d'évaluation et l'écart de recouvrement. L'écart d'évaluation résulte principalement des activités économiques que les autorités fiscales ne connaissent pas ou ne sont pas en mesure d'atteindre, y compris les activités des entités qui ne sont pas enregistrées, qui ne déclarent pas, qui sous-déclarent ou qui déclarent mal leurs impôts, comme le montrent les juridictions où l'informalité est élevée. L'écart de recouvrement correspond à l'écart entre les obligations fiscales calculées, en tenant compte des remboursements et des retenues à la source, et les impôts qui ont été effectivement payés. Il s'agit des montants d'impôts impayés, dont les autorités fiscales ont connaissance, mais qu'elles n'ont pas réussi à recouvrer, généralement parce qu'ils sont liés à des litiges ou parce qu'ils sont considérés comme trop coûteux à poursuivre ou impossibles à recouvrer par des moyens légaux.

La littérature distingue également trois composantes de l'écart de conformité qui complètent les écarts d'évaluation et de recouvrement susmentionnés (Gemmell et Hasseldine 2014; Durán-Cabré et al. 2019).

1. *Composante de sous-déclaration :* les contribuables déclarent moins de revenus qu'ils n'en ont réellement gagnés ou demandent plus de déductions, de crédits ou d'autres avantages fiscaux que la loi ne le permet, ou une combinaison des deux. Il en résulte une différence entre l'impôt réellement dû par le contribuable et le montant déclaré.
2. *Composante de non-remplissage :* indique l'écart entre les contribuables potentiels et les contribuables réels, révélant l'ampleur de la non-déclaration et de l'évasion fiscale.
3. *Composante de non-paiement :* différence entre les recettes fiscales potentielles et les recettes fiscales réalisées, reflétant la fraction de l'impôt éludée par la non-déclaration ou la sous-déclaration à l'administration fiscale.
4. *Composante non-enregistrement* : la différence entre le nombre d'entreprises ou de personnes qui devraient être enregistrées fiscalement (comme les entreprises, les indépendants ou les propriétaires fonciers) et celles qui sont déjà enregistrées. Également classé comme la lacune d'enregistrement.

Enfin, du point de vue de la collecte, certaines autorités fiscales définissent l'écart fiscal en deux catégories : l'écart fiscal brut et l'écart fiscal net.[11] Par exemple, le Service des recettes fédérales (IRS) du Gouvernement des Etats-Unis définit l'écart fiscal brut comme la différence entre l'obligation fiscale réelle totale imposée par la loi pour une année fiscale spécifique et le montant de l'impôt que les contribuables paient volontairement et dans les délais pour cette année. D'autre part, l'écart fiscal net fait référence au montant restant dû de l'obligation fiscale totale après avoir pris en compte tous les paiements effectués dans le cadre de mesures d'exécution et les paiements volontaires tardifs pour une année fiscale spécifique (Plumley 2005). La Figure 1 met en évidence les principales composantes de l'écart fiscal global et le chevauchement entre les différentes définitions de ses composantes.

Figure 1: Les concepts d'écart fiscal



Note : illustration simplifiée des concepts d'écart fiscal.

Source : illustration des auteurs.

[1] Il est important de noter que les définitions de l'écart fiscal brut et net peuvent présenter des différences mineures entre les différentes autorités fiscales, reflétant les environnements uniques de mise en œuvre de l'impôt et les priorités administratives de chaque pays.

3 Méthodes de calcul de l'écart fiscal

Il existe deux approches générales pour estimer l'écart fiscal : l'approche *top-down* et l'approche *bottom-up*. L'approche *top-down* utilise des données agrégées, telles que des indicateurs macroéconomiques ou des données de comptabilité nationale, pour évaluer de manière exhaustive toutes les pertes fiscales, en mesurant l'écart comme la différence entre les recettes potentielles estimées et les recettes réalisées. Toutefois, elle ne permet pas de déterminer les origines de l'écart fiscal ni d'expliquer pourquoi certains domaines ou activités restent non imposés. En revanche, l'approche *bottom-up* utilise des données microéconomiques provenant des administrations fiscales, notamment les résultats d'audits aléatoires ou opérationnels portant sur des critères particuliers ou d'autres données administratives générales provenant des autorités fiscales, afin d'évaluer le degré de non-conformité de certains segments du système fiscal, de certains groupes de contribuables ou de certains types de non-conformité (Hutton 2017).

3.1 Avantages et inconvénients de l'approche *bottom-up*

Avantages

L'approche *bottom-up* de l'estimation des écarts fiscaux présente plusieurs avantages par rapport à d'autres méthodologies, en particulier sa capacité à fournir des informations détaillées (estimations granulaires) sur la base d'audits fiscaux. En voici les principaux avantages :

- *Amélioration de la précision grâce à des données détaillées :* La méthode *bottom-up* s'appuie sur des données granulaires provenant d'audits financiers, ce qui permet des estimations plus précises de l'écart fiscal. Cette technique s'oppose aux stratégies descendantes, qui dépendent d'indicateurs économiques généraux et peuvent négliger des subtilités dans les actions de contribuables individuels ou d'industries particulières.

- *Des informations détaillées pour des actions précises :* Les tactiques *bottom-up* permettent de comprendre en détail le respect de la législation fiscale au niveau de l'individu ou de l'entreprise. Ce niveau de détail permet aux autorités fiscales d'élaborer des interventions précises pour des secteurs d'activité, des catégories de contribuables ou des cas de non-conformité particuliers, ce qui renforce l'efficacité et l'impact des mesures d'exécution (Hutton 2017).

- *Aborder le biais de sélection dans l'estimation de l'écart fiscal :* Le biais de sélection constitue un obstacle majeur à l'estimation précise de l'écart fiscal, en raison de la nature non représentative des contribuables sélectionnés pour les contrôles fiscaux. L'utilisation de l'approche *bottom-up*, en particulier lorsqu'elle est intégrée à des techniques de façon *machine learning*, peut atténuer efficacement ce biais. Cette méthode ne dépend pas de présomptions concernant la distribution des données, ce qui permet d'éviter tout biais susceptible de fausser l'estimation de l'écart fiscal (Alaimo Di Loro et al. 2023).

- *Analyse secteur par secteur :* L'approche *bottom-up* permet une analyse détaillée, secteur par secteur, des écarts de conformité fiscale. Ces informations détaillées permettent aux autorités fiscales d'orienter leurs stratégies de conformité de manière plus précise, en se concentrant sur les secteurs présentant les écarts les plus importantes. Cette approche ciblée pourrait permettre d'améliorer l'efficacité de la collecte de l'impôt sans qu'il soit nécessaire d'augmenter de manière générale les activités d'audit ou d'application de la loi (Barra et al. 2023; Hutton 2017).

- *Adaptabilité à différents types d'impôts :* La flexibilité de l'approche *bottom-up* permet de l'adapter à l'estimation des écarts dans différents types d'impôts, y compris la taxe sur la valeur ajoutée (TVA), l'impôt sur les sociétés (IS) et l'impôt sur le revenu des personnes (IRP). Cette adaptabilité est cruciale, car les différents impôts sont confrontés à différents types de défis de conformité et de tactiques d'évasion fiscale.

- *Amélioration de la conformité fiscale :* Une approche *bottom-up* peut fournir des informations sur le comportement des contribuables, permettant de vérifier ou d'affiner les modèles existants d'identification et de gestion des risques. Elle permet également de mettre en évidence des erreurs spécifiques qui pourraient être gérées plus efficacement par d'autres approches, telles que l'amélioration de l'éducation des contribuables, l'amélioration des services ou la réalisation d'audits et de réévaluations supplémentaires (Barra et al. 2023).

- *Prévoir des limites supérieures et inférieures dans les estimations :* L'approche *bottom-up* permet l'application de plusieurs techniques à la même unité de contribuables, en plus de permettre l'analyse de sensibilité statistique des résultats (Barra et al. 2023).

Inconvénients

Malgré les points forts de l'approche *bottom-up*, la littérature indique qu'elle présente les limites suivantes (Warren 2018; Fiscalis Tax Gap Project Group 2018).

- *Endogénéité :* Cette méthode s'appuie fortement sur les connaissances et les données existantes au sein de l'administration fiscale, ce qui la rend moins efficace pour capturer les facteurs inconnus ou les problèmes non observés.

- *Difficultés à prendre en compte les inconnues :* Comme la méthode est basée sur des données et des résultats opérationnels connus, elle peine à prendre en compte des facteurs qui ne sont pas facilement observables, tels que les revenus sous-déclarés. Elle ne couvre pas non plus l'ensemble de l'économie souterraine, puisque seuls les contribuables enregistrés sont généralement sélectionnés pour les contrôles. Par conséquent, les estimations de ces inconnues impliquent souvent des ajustements approximatifs, ce qui peut réduire la précision.

- *Un champ d'application étroit :* Cette approche va du spécifique au général, en se concentrant sur les contribuables individuels. Bien que cette approche permette d'obtenir des informations détaillées, elle peut négliger les tendances ou les modèles macroéconomiques.

- *Risque d'agrégation :* Les approches *bottom-up* ne calculent que les composantes de l'écart fiscal, ce qui nécessite une agrégation pour estimer l'écart total. Cependant, ce processus comporte un risque de double comptage et de surestimation de l'écart fiscal total, en particulier lorsqu'il existe des chevauchements entre les différentes composantes de l'écart.

3.2 Type d'audit

Les autorités fiscales s'appuient généralement sur des informations d'audit pour prévoir la fraude fiscale et estimer l'écart fiscal. Ces audits peuvent être classés en deux catégories principales : les audits aléatoires et les audits opérationnels. Ces deux types d'audit ont des objectifs différents et des méthodologies uniques qui permettent de mieux comprendre le respect des règles par les contribuables.

Audits aléatoires

Les initiatives d'audit aléatoire impliquent la sélection d'échantillons de contribuables par le biais d'un processus aléatoire, visant à refléter fidèlement la population plus large qu'ils sont censés représenter. Lors de ces contrôles, tous les contribuables sélectionnés font l'objet d'un examen approfondi, afin d'identifier toute divergence entre ce qu'ils ont déclaré sur leurs impôts et ce qu'ils sont légalement tenus de déclarer. Les résultats de ces audits fournissent une mesure fiable du niveau global de conformité au sein du groupe de l'échantillon. Pour extrapoler les résultats de l'échantillon à l'ensemble de la population, nous devons nous assurer que le processus de sélection est totalement aléatoire et n'implique aucun critère de sélection (Barra et al. 2023).

Les contrôles aléatoires, bien qu'approfondis, présentent des inconvénients, selon Feinstein (1999), notamment des coûts élevés tant pour les bureaux fiscaux que pour les contribuables, en particulier ceux qui respectent la législation fiscale. Il existe également un délai entre la période couverte par les données et le moment où les résultats sont disponibles. Le rendement financier est généralement inférieur à celui des contrôles ciblés, car ils examinent à la fois les contribuables en règle et ceux qui ne le sont pas, contrairement aux contrôles ciblés, qui se concentrent sur ceux qui sont les plus susceptibles d'échapper à l'impôt. En outre, ils ne peuvent pas détecter les contribuables non enregistrés, ce qui entraîne une sous-estimation de certaines lacunes fiscales.

Enfin, les autorités fiscales peuvent être réticentes à effectuer des contrôles aléatoires pour des raisons liées à l'image publique de l'autorité auprès des contribuables. Les contrôles aléatoires pourraient être perçus, par les contribuables respectueux des règles, comme un contrôle excessif ou injuste, ce qui susciterait un sentiment négatif dans l'opinion publique et une diminution de la confiance dans l'administration fiscale.

Audits opérationnels

Les audits opérationnels reposent sur une évaluation des risques et ils ciblent des contribuables spécifiques choisis en fonction de critères définis par l'analyse des risques de l'administration fiscale. Ces contrôles peuvent porter sur un ou plusieurs types d'impôts et, pour chaque type d'impôt, ils peuvent couvrir l'ensemble du champ d'application de l'impôt ou seulement un segment spécifique. Par conséquent, ce type d'audit peut ne pas être représentatif de l'ensemble de la population en raison des critères de sélection, car tous les contribuables n'ont pas la chance d'être sélectionnés, comme dans un audit aléatoire. C'est pourquoi les administrations fiscales mettent en œuvre une estimation *bottom-up* des écarts, en utilisant des données d'audit non aléatoires à l'aide de techniques visant à déduire les caractéristiques de la population plus large à partir de l'échantillon non représentatif.

3.3 Procédures d'estimation *bottom-up*

Plusieurs procédures peuvent être utilisées pour réaliser une approche *bottom-up*. Elles utilisent toutes des informations d'audit pour prédire le comportement d'entreprises ou de périodes non auditées. Dans cette section, nous passons en revue les estimations les plus courantes et nous soulignons leurs principales caractéristiques (avantages et inconvénients).

Techniques de régression

Les techniques de régression sont considérées comme courantes dans la littérature *bottom-up* et peuvent corriger le biais de sélection. Elles peuvent également aider à déterminer les caractéristiques qui permettent de prédire, si un contribuable se conformera ou non à la loi, et

d'estimer l'ampleur de la non-conformité. Ces techniques de régression comprennent l'approche de Heckman et l'approche de l'appariement des scores de propension.

Approche de Heckman. L'approche de Heckman tient compte du biais de sélection qui se produit au cours du processus d'audit opérationnel, ce qui entraîne une endogénéité dans le sous-ensemble des contribuables contrôlés. Cette méthode, fondée sur les travaux de Heckman (1979), implique un processus d'estimation en deux étapes. La première étape identifie la probabilité qu'une observation soit incluse dans l'échantillon, en calculant essentiellement la probabilité qu'un contribuable soit sélectionné pour un contrôle, à l'aide d'une équation de régression probit. La deuxième étape se concentre sur l'estimation de la variable d'intérêt qui, dans ce cas, est le montant recouvré à la suite du contrôle. Pour ce faire, on prend en compte des variables explicatives et un régresseur spécifique qui corrige le biais de sélection. Ce régresseur particulier, connu sous le nom de ratio inverse de Mills, est dérivé des paramètres estimés dans l'équation de sélection. L'équation des résultats est ensuite calculée à l'aide de la régression des moindres carrés ordinaires (OLS), en incorporant un facteur de l'équation de la première étape.

Le Fiscalis Tax Gap Project Group (2018) souligne que deux considérations importantes doivent être prises en compte lors de l'estimation de l'écart fiscal à l'aide de la méthode de Heckman. Premièrement, l'équation de sélection doit être puissante pour expliquer les résultats, puisque la méthode repose fortement sur la capacité de l'équation à prédire le non-respect des règles. Deuxièmement, l'équation doit inclure au moins une variable qui influence la sélection pour l'audit, mais qui n'a pas d'impact sur le non-respect lui-même. Cela permet d'éviter les problèmes d'estimations inexactes dus à la multi-colinéarité. Essentiellement, pour une estimation précise de l'écart fiscal, il est nécessaire de disposer de données sur les facteurs qui conduisent à être contrôlé, mais qui ne sont pas directement liés au niveau de non-conformité, et, dans la pratique, cette restriction d'exclusion est difficile à satisfaire.

Approche de l'appariement par score de propension. La méthode de l'appariement des scores de propension est utilisée pour corriger le biais de sélection sur la base des pondérations des données. Cette méthode commence par le calcul d'un « score de propension » pour chaque entité, en utilisant des modèles statistiques pour prédire leur probabilité d'être non conformes ou non audités. Un modèle de sélection binaire calcule les propensions à l'aide de probits ou de logits. Une fois ces scores estimés, la méthode associe les entités qui ont été contrôlées à celles qui ne l'ont pas été mais qui partagent des scores de propension similaires. L'approche utilisée pour faire correspondre les observations peut être le plus proche du voisin, de l'étrier, du noyau ou de la méthode linéaire locale. Après l'appariement, la dernière étape consiste à attribuer une valeur aux déclarations non auditées. Cette valeur, appelée N, est une valeur imputée ou estimée de ce que la déclaration non auditée aurait déclaré si elle avait été auditée. L'imputation est basée sur les valeurs réelles observées dans les déclarations auditées appariées. Cette étape est nécessaire pour estimer ce qu'aurait été la conformité fiscale du groupe non audité, s'il avait fait l'objet d'un audit.

Approche par regroupement

Cette approche consiste à classer les contribuables contrôlés et non contrôlés en groupes sur la base des variables significatives utilisées pour sélectionner l'entreprise à contrôler, telles que la taille de l'entreprise, la région géographique et le secteur d'activité. Elle permet de calculer l'écart fiscal global, en additionnant les écarts estimés pour chaque groupe. Ces estimations sont obtenues en appliquant un facteur d'échelle aux résultats de l'audit des contribuables contrôlés, projetant ainsi ces résultats sur l'ensemble de la population au sein de chaque groupe. Bien que simple à appliquer et facile à mettre en œuvre, cette méthode ne corrige que partiellement le biais de sélection, ce qui se traduit par des résultats qui ne sont pas entièrement fiables.

Approche des valeurs extrêmes

L'approche des valeurs extrêmes exploite le biais de sélection de l'audit opérationnel en faveur des contribuables dont le niveau de non-conformité est le plus élevé. Elle traite du comportement des valeurs extrêmes (maximales ou minimales) d'un ensemble de données, plutôt que des valeurs moyennes, en suggérant que, quelle que soit la distribution globale des données, les valeurs extrêmes suivent souvent une distribution de Pareto généralisée. Cela signifie qu'il est possible d'obtenir des informations sur le taux global de non-respect des obligations fiscales par les grandes entreprises à partir d'un nombre limité de cas extrêmes (à savoir les fraudeurs fiscaux les plus importants). Cette approche est applicable lorsque les données présentent des caractéristiques de la distribution de Pareto - une forme de distribution en loi de puissance indiquant qu'une petite fraction de cas contribue de manière disproportionnée à la valeur totale observée dans les données, comme lorsque la sous-déclaration fiscale est fortement asymétrique (quelques grandes entreprises représentant la majeure partie de l'écart) (Bloomquist et al. 2014).

Approches de façon machine learning

L'application des approches de façon *machine learning* (ML) aux études économiques, bien que récente, connaît une augmentation progressive, en particulier dans la recherche liée à la fiscalité, telle que la prédiction de l'évasion fiscale, de la fraude et de la conformité, ainsi que l'amélioration du contrôle fiscal et de l'estimation de l'écart d'imposition. Alors que la recherche dans ce domaine s'appuie généralement sur des méthodes traditionnelles pour faire des prédictions, ces méthodes souffrent de limitations liées à la dépendance aux méthodes de régression linéaire et aux hypothèses de distribution strictes qu'elles comportent. En réalité, les données présentent souvent des schémas plus complexes, ce qui fait que ces méthodes ne sont pas assez flexibles pour la prédiction. C'est pourquoi certaines études ont commencé à adopter des méthodes de façon *machine learning* (ML) pour améliorer les résultats des prédictions.

Pour illustrer l'utilisation de *machine learning*, Pérez López et al. (2019) a employé des modèles de réseaux neuronaux à perceptron multicouche (MLP) pour prédire la fraude fiscale, en utilisant les données complètes des déclarations de l'impôt sur le revenu des personnes (IRP), en Espagne. Cette méthode de façon *machine learning* (ML) a permis de prédire la probabilité de fraude fiscale et la probabilité d'implication dans des pratiques liées à la fraude pour chaque contribuable. Zumaya et al. (2021) a utilisé deux algorithmes de *machine learning* (ML), dont la méthode Artificial Neural Network (ANNs) et la méthode *random forests* (RF), en plus de l'analyse de réseaux complexes pour prédire la fraude à la TVA au Mexique, en analysant les données transactionnelles et les réseaux d'interaction des contribuables. L'article a montré que la combinaison de ces trois méthodes permettait d'identifier de nouveaux suspects potentiels, en apprenant des schémas à partir de fraudeurs connus. Ioana-Florina et Mare (2021) a tenté de prédire la propension des contribuables à frauder le fisc en fonction de leur confiance dans le système fiscal à l'aide d'un modèle de réseau neuronal à perceptron multicouche (MLP). Cette approche a démontré une performance prédictive supérieure, surpassant celle du modèle de régression logistique binaire.[2]

D'autre part, les méthodes de façon *machine learning* (ML) sont également utilisées pour améliorer les efforts de contrôle fiscal. Par exemple, Howard et al. (2020) a évalué le potentiel des techniques de *machine learning* pour améliorer le processus de sélection des cas d'audit par correspondance par

[2] Voir aussi Alsadhan (2023); Baghdasaryan et al. (2022); Holtzblatt et Engler (2022); Murorunkwere et al. (2022, 2023); Raikov (2021); Savic´ et al. (2022) pour d'autres exemples d'utilisation de méthodes de façon *machine learning* (ML) pour prédire les comportements de fraude et d'évasion fiscales.

le Service des recettes fédérales (IRS) du Gouvernement des Etats-Unis. L'étude a révélé que, pour certaines catégories d'audit, les méthodes de façon *machine learning* (ML) sont plus performantes que les approches traditionnelles pour le classement et la sélection des déclarations de revenus en vue d'un contrôle par correspondance. Ces méthodes permettent non seulement d'augmenter les recettes, mais aussi de réduire le ratio de non-changement, ce qui signifie que moins d'audits n'aboutissent à aucun ajustement par rapport à d'autres méthodes. De même, Battaglini et al. (2022) a utilisé des données fiscales administratives italiennes pour explorer le potentiel des techniques de *machine learning* telles que la méthode *random forests* dans l'amélioration de la détection de l'évasion fiscale et du recouvrement, en améliorant le processus de sélection des contribuables à contrôler. L'article indique que dans certains scénarios, la *machine learning* (ML) pourrait améliorer la prédiction de la détection de la fraude jusqu'à 83 pourcentage et récupérer jusqu'à 65 pourcentage de la fraude détectée.

La recherche sur l'estimation des écarts fiscaux n'était pas très éloignée de ces nouveaux développements. Compte tenu des limites des approches d'estimation de l'écart fiscal mentionnées précédemment, qui reposent sur des méthodes de régression traditionnelles, certains chercheurs et autorités fiscales ont commencé à intégrer l'utilisation de techniques semi-paramétriques dans les méthodes traditionnelles et à utiliser la *machine learning* (ML) pour améliorer les résultats de la prédiction. Si la *machine learning* (ML) est plus performante dans les tâches de prédiction que les approches traditionnelles, il est également efficace pour remédier au biais de sélection dans le cas des estimations de l'écart fiscal basées sur des audits opérationnels.

Pour aborder la question du biais de sélection dans le contexte des estimations des écarts fiscaux, il est essentiel de faire la distinction entre les deux principaux types de biais de sélection : le biais de causalité et le biais de sélection de l'échantillon. Le biais de sélection causal affecte l'estimation de paramètres non biaisés dans l'analyse causale, par exemple lorsque les groupes traités et les groupes de contrôle ne sont pas assignés de manière aléatoire, ce qui conduit à des estimations biaisées des effets du traitement. Cependant, nous nous concentrons sur le biais de sélection de l'échantillon, qui se produit lorsque l'échantillon d'entraînement utilisé pour construire un modèle prédictif diffère de l'échantillon de prédiction. Dans le cas des estimations de l'écart fiscal basées sur des audits opérationnels, ce biais est dû au fait que l'échantillon d'entraînement est constitué de contribuables audités sélectionnés sur la base de certains critères connus des autorités fiscales et non représentatifs de l'ensemble de la population des contribuables, alors que l'échantillon de prédiction comprend des contribuables non audités. Cette divergence peut conduire à des prédictions biaisées, si elle n'est pas correctement traitée.

Un aspect crucial du traitement du biais de sélection de l'échantillon est la distinction entre les biais découlant de facteurs observables et non observables. Le biais de sélection observable se produit lorsque le processus de sélection, tel que la décision de contrôler, est basé sur des variables connues et mesurables. Dans ce cas, si la probabilité d'être contrôlée peut être estimée avec précision à l'aide de ces co-variables observables, le biais peut être corrigé en incluant ces co-variables dans le modèle de *machine learning* (ML). Cette méthodologie correspond aux stratégies décrites par Brewer et Carlson (2024), qui préconisent de contrôler le biais de sélection en ajustant les facteurs observables. En calculant et en intégrant la probabilité de sélection dans le modèle, il est possible d'atténuer le biais de sélection, en supposant que les décisions d'audit sont principalement motivées par des données observables.[3]

[3] On peut supposer que les autorités fiscales disposent d'informations sur la manière de décider qui contrôler. Ces informations sont généralement réservées, mais elles peuvent être utilisées dans le modèle de *machine learning* (ML) pour prédire avec précision

Dans les scénarios où le processus de sélection est régi par des facteurs inobservables qui ne sont pas pris en compte dans l'ensemble des données, la complexité du biais augmente. Les méthodes traditionnelles peuvent ne pas être suffisantes pour contrer cette forme de biais. Dans de tels cas, des techniques plus avancées sont nécessaires, telles que l'incorporation d'une fonction de contrôle dans le modèle de *machine learning* (ML) basé sur la méthode de Heckman, pour traiter le biais de sélection basé sur des facteurs inobservables qui ne sont pas capturés dans l'ensemble des données (Brewer et Carlson 2024). Dans la littérature récente, on trouve des exemples notables d'intégration d'approches de façon *machine learning* (ML) dans des méthodes traditionnelles, ainsi que des études qui estiment les écarts fiscaux en utilisant principalement des techniques de *machine learning* (ML).

Alaimo Di Loro et al. (2023) a proposé une méthode basée sur la *machine learning* (ML) qui consiste en deux étapes de l'algorithme de renforcement du gradient. Cette méthode corrige le biais de sélection découlant de l'utilisation de données d'audit non aléatoires et fournit des prédictions précises. Tout d'abord, la méthode estime les scores de propension de la probabilité qu'un contribuable soit contrôlé à l'aide d'un modèle de classification basé sur le gradient boosting avec des arbres de classification et de régression (CART), en tant qu'apprenants de base. Pour ce faire, les données sont divisées en ensembles de formation et de test et, au cours du processus de formation, les co-variables importantes sont sélectionnées. Cette étape permet de prédire les probabilités d'audit de chaque entreprise, en fonction de ses co-variables.

Deuxièmement, la méthode emploie un modèle de régression utilisant le gradient boosting avec des arbres de classification et de régression (CART), comme base d'apprentissage pour prédire l'assiette fiscale potentielle, y compris la TVA non déclarée, et donc les montants éludés pour chaque entreprise. Dans cette étape, les scores de propension précédemment obtenus sont utilisés pour créer des pondérations pour chaque contribuable, en corrigeant toute surreprésentation ou sous-représentation dans l'échantillon contrôlé. La comparaison de cette approche de façon *machine learning* (ML) avec le modèle traditionnel de Heckman révèle la supériorité de la *machine learning* (ML) pour saisir la variabilité de l'assiette fiscale potentielle et fournir des prédictions plus précises pour l'estimation de l'écart fiscal.

Adu-Ababio et al. (2024) a utilisé des algorithmes de *machine learning* supervisé avec des déclarations fiscales et des données d'audit pour estimer les écarts fiscaux en Zambie. Le principal algorithme de *machine learning* (ML) utilisé dans cette étude est le réseau neuronal artificiel (ANN) en deux étapes. La première étape s'appuie uniquement sur les données vérifiées pour créer des itérations de versions multiples des ensembles de données de formation et de test de manière aléatoire, 90 pourcentage des données étant utilisées pour la formation du modèle. À chaque itération, l'algorithme apprend à partir de l'ensemble de données de formation en analysant divers paramètres liés à la fiscalité. L'algorithme utilise ensuite ce qu'il a appris pour prédire les taux d'évasion fiscale à l'aide des données de test. Ensuite, l'algorithme compare les taux d'évasion fiscale réels et prédits et, si ces prédictions ne correspondent pas étroitement aux taux réels, des améliorations peuvent être apportées au modèle et ce processus est répété jusqu'à ce qu'il atteigne une performance satisfaisante. Pour la deuxième étape, le modèle est déployé, en utilisant l'échantillon complet où les données auditées sont utilisées dans l'ensemble de formation et les données non auditées forment l'ensemble de test. Une fois que le modèle a appris à partir des variables explicatives sélectionnées, il prédit la fraude fiscale à partir des données de test et utilise ensuite la fraude fiscale prédite et réelle pour estimer les écarts fiscaux. Les auteurs ont également

---

les résultats. Nous conseillons de ne pas partager la pertinence des co-variables dans la prédiction, car ces informations sont liées au processus d'audit. Cependant, ces résultats peuvent également être utilisés pour améliorer le processus de prise de décision en matière d'audit.

utilisé d'autres algorithmes de *machine learning* (ML), tels que la forêt aléatoire, pour vérifier la stabilité et la fiabilité de la méthode principale et les résultats étaient légèrement proches.

Dans le même ordre d'idées, l'étude de Ebrahim et al. (2024) a utilisé les données de l'administration fiscale et de l'audit pour estimer l'écart de TVA, en Tanzanie, à l'aide de façon *machine learning* (ML) et, plus particulièrement, de l'algorithme de la forêt aléatoire. Cette approche visait à prédire les montants de l'évasion fiscale pour les entreprises non auditées et auditées dans les périodes où aucun audit n'a été effectué. Les auteurs ont comparé les performances de l'approche de façon *machine learning* (ML) avec celles de la régression traditionnelle par les moindres carrés ordinaires (OLS). Ils ont constaté une réduction significative de l'erreur quadratique moyenne (RMSE) et des valeurs R-carré plus élevées lors de l'utilisation de l'algorithme de la forêt aléatoire, ce qui indique une performance de prédiction plus précise. Les résultats révèlent un écart de TVA d'environ 62 pourcentage en moyenne, avec des différences considérables entre les divers secteurs économiques. Le secteur de l'agriculture, en particulier, présente l'écart de TVA le plus important, ce qui met en évidence une évasion fiscale significative dans ce domaine.

D'autres avancées dans les techniques de *machine learning* (ML) impliquent l'utilisation d'anciennes approches de régression. Chudý et al. (2020) a appliqué une sélection d'échantillon semi-paramétrique du modèle de Heckman pour estimer l'écart de l'impôt sur les sociétés (IRP), en Slovaquie. Cette extension du modèle de Heckman est plus performante que le modèle de Heckman traditionnel, car elle permet une hypothèse de normalité plus souple et une meilleure modélisation des structures de données complexes, ainsi que le traitement des relations non linéaires et de l'hétéroscédasticité inhérente aux données. Dans la première étape de ce modèle, l'équation de sélection a été estimée à l'aide d'une méthode non paramétrique, telle que le lissage par noyau, qui fournit des approximations flexibles des distributions. Ensuite, dans la deuxième étape du modèle, l'équation de résultat a incorporé ces estimations de la première étape pour fournir une correction plus robuste du biais de sélection et capturer les relations plus complexes qu'un modèle de régression linéaire pourrait négliger. L'article a montré que cette approche était plus performante que d'autres approches alternatives, telles que l'appariement des scores de propension et la régression linéaire pondérée par les moindres carrés ordinaires (OLS), pour ce qui est de la correction du biais de sélection et de l'obtention de meilleures prédictions.

Les autorités fiscales ont également commencé à utiliser des techniques de *machine learning* (ML) pour améliorer leurs estimations des écarts fiscaux ou leurs processus d'audit. L'Administration fiscale italienne (n.d.) a utilisé la façon *machine learning* (ML), ainsi que d'autres méthodes traditionnelles pour estimer l'écart de TVA dans le cadre de l'approche dite assistée par la *machine learning* (ML). L'étape initiale de cette approche vise à remédier le biais de sélection qui découle de l'utilisation d'audits non aléatoires, en utilisant la régression logistique pour diviser la population en groupes, chaque groupe ayant une probabilité similaire d'être contrôlé. Ensuite, la population est stratifiée en quintiles sur la base de ces probabilités, ce qui permet aux contribuables contrôlés d'être représentatifs de l'ensemble de la population dans chaque groupe. Dans un deuxième temps, la *machine learning*, et plus précisément les arbres de régression à sac, est utilisé pour faire des prédictions au sein de chaque strate. La dernière étape vise à améliorer la précision des prédictions, en utilisant le modèle d'appariement moyen prédictif (PMM) qui utilise les prédictions initiales pour apparier chaque contribuable non contrôlé (appelé bénéficiaire) avec un contribuable contrôlé (appelé donneur) sur la base de la similarité de leurs valeurs prédites. Ce processus garantit que les valeurs imputées reflètent la véritable distribution de la variable cible, ce qui permet de tirer des conclusions précises sur diverses caractéristiques de distribution au-delà des seules moyennes.

L'Administration fiscale canadienne (2019) utilise une technique de *machine learning* (ML) non supervisée pour identifier les groupements au sein d'une population, similaire à la première étape mentionnée précédemment pour l'Italie, où les éléments de chaque groupement sont plus

similaires les uns aux autres qu'à ceux des autres groupements. Cet algorithme de *machine learning* (ML) classe automatiquement les entreprises en groupements sur la base de caractéristiques spécifiques, en supposant que les entreprises non auditées de chaque groupement partagent le même ratio de non-conformité par rapport aux recettes brutes déclarées que les entreprises auditées. Cette approche a été utilisée pour fournir une estimation de la limite supérieure et elle a été combinée avec l'approche des valeurs extrêmes pour fournir une estimation de la limite inférieure de l'écart fiscal.

Résumé

L'estimation de l'écart fiscal à l'aide d'approches *bottom-up* peut être réalisée à l'aide de différentes méthodes d'estimation. Cependant, chaque méthode peut être plus appropriée, en fonction du contexte et des données utilisées. En règle générale, l'utilisation d'une approche *bottom-up* pourrait se fonder sur des données d'audit aléatoires ou fondées sur le risque. De nombreux chercheurs considèrent que l'utilisation de données d'audit aléatoires est la méthode idéale pour réaliser une estimation *bottom-up* de l'écart fiscal. Toutefois, dans de nombreux cas, les autorités fiscales ont tendance à préférer effectuer un audit basé sur le risque, ce qui pose certains problèmes d'estimation, étant donné que les contribuables sélectionnés pour l'audit peuvent être très différents des autres contribuables, ce qui fait que les résultats de l'audit ne sont pas représentatifs de l'ensemble de la population en situation de non-conformité. Dans ce cas, l'estimation traditionnelle par les moindres carrés ordinaires (OLS) pourrait ne pas être le meilleur choix pour les chercheurs, en raison du biais de sélection du processus d'audit. C'est pourquoi les chercheurs utilisent d'autres méthodes pour obtenir des estimations non biaisées. Dans ce qui suit, nous résumons les principaux enseignements des méthodes mentionnées dans cette section.

Bien que l'approche de Heckman, en deux étapes, soit considérée comme l'une des méthodes les plus couramment utilisées pour tenir compte du biais de sélection, sa restriction d'exclusion est parfois difficile à satisfaire, ce qui peut conduire à des erreurs types gonflées, en raison de la multi-colinéarité, et elle a tendance à sous-estimer l'écart fiscal, étant donné que l'évasion fiscale et la non-conformité non détectée sont souvent négligées. L'appariement par score de propension permet d'éliminer le biais de sélection, en créant des groupes appariés de contribuables respectueux et non respectueux de la loi sur la base de caractéristiques observables, ce qui permet d'attribuer plus précisément les différences de résultats en matière de respect de la loi fiscale au non-respect de la loi plutôt qu'à des facteurs non observés. Certaines administrations fiscales utilisent l'approche par groupement pour détecter les comportements anormaux et mettre au jour la sous-déclaration fiscale au sein de grappes spécifiques, puis estiment l'écart fiscal pour chaque groupement, en extrapolant les résultats de l'audit des contribuables contrôlés à l'ensemble de la population au sein de cette groupement spécifique. D'autre part, l'approche des valeurs extrêmes est plus simple et plus économique en termes de temps et de ressources que les autres approches. Néanmoins, elle nécessite davantage d'hypothèses, notamment en ce qui concerne la définition de la queue de la distribution de Pareto, sur laquelle elle s'appuie pour la modélisation.

En revanche, les techniques d'estimation basées sur la *machine learning* (ML) offrent des avantages significatifs par rapport aux méthodes susmentionnées, en particulier lorsqu'il s'agit de traiter des relations complexes et non linéaires et des facteurs non observés influençant le biais de sélection. Les méthodes de façon *machine learning* (ML) peuvent être préférées pour leur flexibilité et leur performance prédictive supérieure.

4 Le guide pratique

Dans cette section, nous expliquons les composants du guide pratique. L'objectif de ce guide pratique est d'estimer l'écart fiscal sur la taxe sur la valeur ajoutée (TVA), l'impôt sur le revenu des sociétés (IS) ou l'impôt sur le revenu des personnes (IRP). Le guide pratique comporte deux éléments principaux : le nettoyage des données et l'estimation. Le processus de nettoyage des données vise à garantir l'harmonisation et la cohérence des fichiers de données nécessaires à l'estimation *bottom-up*. En outre, il permet d'aligner les exigences générales de l'estimation par *machine learning* (ML). Ceci est important, car les données proviennent de sources et de périodes différentes et leur standardisation simplifie le processus d'estimation. L'estimation est basée sur la méthodologie de la forêt aléatoire, une technique de *machine learning* (ML). Une explication technique de cette méthode est fournie à l'Annexe A.

4.1 Nettoyage des données

Le processus de nettoyage des données peut être divisé en trois étapes principales : les deux premières traitent des déclarations fiscales administratives (TVA, IS et IRP) et des données d'audit et la dernière consiste à combiner ces fichiers de données en vue d'une analyse ultérieure. Cette étape vise à traiter les différentes sources de données, à les harmoniser et à construire une structure unique qui combine les informations sur les contribuables, les déclarations fiscales et les résultats des audits ou des évaluations.

En général, les informations relatives aux déclarations fiscales (TVA, IS ou IRP) sont contenues dans des fichiers différents de ceux des audits, car ces derniers sont réalisés après que les entreprises ou les personnes ont déposé leurs déclarations. Toutefois, les déclarations fiscales peuvent contenir au moins deux séries de déclarations pour le même contribuable. Cela peut être dû au fait que le contribuable a mis à jour sa déclaration à un moment donné pendant ou en dehors de la période de dépôt. Il s'agit d'un problème courant de duplication qui se pose dans les bases de données administratives fiscales. Dans ce cas, le même élément d'information est répliqué pour le même contribuable. En d'autres termes, pour un contribuable au cours d'une année de déclaration donnée, il existe deux ou plusieurs réplications de la même information. L'un des principaux objectifs de la section de nettoyage des données est de s'assurer que chaque contribuable est identifié de manière unique par ses identifiants et l'année de déclaration. Dans la première étape du guide pratique, nous fournissons des scénarios possibles, en créant de telles erreurs de duplication, et nous montrons comment l'utilisateur peut les traiter individuellement. Il est important de traiter les doublons dans tous les fichiers de données de déclaration de revenus et d'audit requis, qu'il s'agisse de fichiers uniques ou multiples. Dans le cas de fichiers multiples, l'approche consiste à traiter d'abord les doublons, puis à ajouter les ensembles de données respectifs dans un seul fichier.

Dans cette étape du guide pratique, nous abordons également les problèmes observés dans les données d'audit concernant les périodes d'audit et la manière dont elles sont liées à des périodes de déclaration spécifiques. Dans certains cas, les données d'audit sont identifiées par le fait que l'année d'évaluation double l'année de déclaration pour le dépôt. Parfois, c'est plutôt l'année d'audit qui fait office d'année de déclaration. Quoi qu'il en soit, il est nécessaire d'identifier l'année spécifique dans les données d'audit qui correspond à l'année de déclaration et de les joindre pour obtenir un fichier unique, si les données se trouvent dans plusieurs fichiers. Cela permet de s'assurer que chaque évaluation d'audit est correctement liée à une période de déclaration spécifique.

À la fin de ces deux premières étapes, nous agrégeons les données des déclarations fiscales au niveau annuel. Étant donné que nous estimons les écarts fiscaux chaque année, nous nous assurons également que les évaluations de contrôle sont liées aux résultats annuels, même si des contrôles ont été effectués pour plusieurs années de déclaration. L'agrégation se fait généralement pour la

taxe sur la valeur ajoutée (TVA) et l'impôt sur le revenu des personnes (IRP), mais pas pour l'impôt sur le revenu des sociétés (IS), qui fait toujours l'objet d'une déclaration annuelle. Cette procédure garantit que nous disposons d'une déclaration de revenus ou d'un résultat de contrôle (si le contribuable est contrôlé) par contribuable et par an.

Enfin, nous combinons les fichiers de données requis en gardant à l'esprit que les variables des déclarations fiscales et des audits se trouvent dans deux fichiers distincts. Il est important de comprendre le processus de fusion, car il montre à quel point nous avons nettoyé et traité les doublons dans tous les fichiers de données. L'objectif est de fusionner les informations concernant la même unité (contribuable) au cours de la même période (année-mois). En outre, nous voulons que les informations fournies par les données d'audit, telles que le résultat de l'audit pour une année de déclaration donnée, soient fusionnées avec l'enregistrement fiscal pour la période de dépôt de la déclaration correspondante. Par exemple, nous fusionnons l'enregistrement fiscal de l'année de déclaration 2018 avec le résultat de l'audit concernant les déclarations fiscales erronées en 2018, si l'entreprise a fait l'objet d'un audit. Il n'y aura pas d'informations sur le résultat de l'audit, si l'entreprise n'a pas fait l'objet d'un audit. Il s'agit d'un problème habituel auquel l'utilisateur est confronté, car les contrôles sont effectués rétroactivement sur un nombre limité de contribuables, sur la base de déclarations antérieures. Nous expliquons comment obtenir les résultats de l'audit pour ces entreprises non auditées dans l'étape suivante du guide pratique.

4.2 Estimation par *machine learning*

L'approche *bottom-up* est suivie pour estimer l'écart fiscal. Cette approche requiert, en entrée, les résultats de l'audit des contribuables, que nous considérons comme des déclarations fiscales erronées. Cette variable est obtenue et trouvée dans les données d'audit après le processus d'audit. Cependant, cette variable est visible pour les entreprises qui ont fait l'objet d'un contrôle. Il est donc nécessaire d'estimer ou de prédire les résultats pour les contribuables et les périodes qui n'ont pas fait l'objet d'un contrôle. Les prédictions concernant les contribuables et les périodes non audités sont nécessaires, parce que les informations sur les déclarations erronées provenant des audits dépendent de périodes et d'unités spécifiques. Ainsi, un contribuable contrôlé au cours de l'année de déclaration 3 n'est pas contrôlé au cours de l'année de déclaration 2, ce qui signifie que nous devons inclure une prévision pour les périodes non contrôlées afin de nous assurer que nous disposons de toutes les informations nécessaires. Une procédure d'estimation est nécessaire pour obtenir des prédictions précises sur les fausses déclarations fiscales.

Dans le guide pratique, nous suivons la méthode *random forests* pour prédire les fausses déclarations fiscales dans les entreprises et les périodes non auditées. Cette méthodologie permet une estimation granulaire, en capturant mieux les valeurs aberrantes ou atypiques potentielles que la prédiction linéaire. La méthode *random forests* doit être affinée, en choisissant deux paramètres essentiels : le nombre d'itérations (ou d'arbres) et le nombre d'utilisations à prédire dans chaque fraction. Pour ce faire, il est nécessaire d'utiliser des données qui contiennent la variable à prédire en cas de déclaration fiscale erronée. Par conséquent, l'ensemble de données est d'abord divisé en données auditées et non auditées. Les premières seront utilisées pour ajuster le modèle et les secondes seront utilisées pour prédire.

La division des données d'audit en échantillons de formation et de test est nécessaire pour le processus d'ajustement. Cela permet d'améliorer la précision de l'estimation, puisque la méthodologie utilise les données d'entraînement pour apprendre à connaître les variables et elle confronte ensuite la prédiction à la valeur réelle dans les données de test. En agissant ainsi, les deux paramètres critiques sont obtenus. En outre, ces paramètres garantissent que l'erreur de prédiction, c'est-à-dire la différence entre la prédiction et la valeur réelle, est la plus faible possible.

Avec le modèle optimal, le guide pratique effectue un contraste avec un modèle de régression. Cela permet de montrer la précision de la prédiction et de valider le modèle de prédiction.

Enfin, l'écart fiscal est obtenu. Tout d'abord, le modèle n'est exécuté que pour les données auditées, car ces observations contiennent des informations erronées. Dans cette étape, le modèle estime l'indice (ou le poids) que doit avoir chaque variable auxiliaire (ou covariante). Ensuite, le modèle prédit les données non auditées à l'aide de l'indice optimal, et les prédictions de déclarations erronées sont obtenues. L'écart fiscal est obtenu en additionnant la variable de déclaration erronée (prédite ou découverte par l'audit) et la déclaration fiscale, ce qui donne l'impôt potentiel. L'écart fiscal est le taux entre la déclaration erronée et l'impôt potentiel, indiquant le pourcentage de l'impôt potentiel non perçu en raison de la déclaration erronée. Cette variable est obtenue en fonction de la période du groupe (comme l'industrie), ce qui montre la granularité de la méthodologie.

5 Remarques finales

Dans ce guide pratique *bottom-up* d'estimation des écarts fiscaux, nous avons cherché à développer un cadre pratique pour l'estimation des écarts fiscaux dans la taxe sur la valeur ajoutée (TVA), l'impôt sur les sociétés (IS) et l'impôt sur le revenu des personnes physiques (IRP) à l'aide d'une méthodologie *bottom-up*. Le guide pratique est conçu pour permettre aux autorités fiscales et aux décideurs politiques d'estimer la différence entre les recettes fiscales effectivement perçues et les recettes potentielles qui auraient pu être perçues dans le cadre d'un respect total des réglementations fiscales. Il fournit un cadre normalisé applicable aux pays en développement, compte tenu de leur contexte et des ressources dont ils disposent. Notre approche est basée sur l'application d'un algorithme de *machine learning* (ML) utilisant les micro-déclarations fiscales administratives et les données d'audit pour prédire les déclarations fiscales erronées et le non-respect des règles fiscales, puis pour estimer les écarts fiscaux à la fois au niveau global et par secteur ou région spécifique.

Dans cette note technique, nous avons passé en revue les définitions des écarts fiscaux afin d'en comprendre les composantes, puisque notre méthode vise à estimer les écarts fiscaux liés à la sous-déclaration, à la déclaration erronée et à la non-conformité parmi les contribuables enregistrés. Ensuite, nous passons en revue les procédures traditionnellement employées en soulignant les avantages de l'utilisation de l'estimation par *machine learning* (ML)dans le cadre d'une approche *bottom-up* par rapport à d'autres estimations alternatives

Le guide pratique peut être divisé en deux étapes principales : la gestion des données et l'analyse par *machine learning* (ML). Au cours de la phase de gestion des données, les ensembles de données fiscales et d'audit sont préparés pour une procédure d'analyse par le biais du nettoyage, du traitement des doublons et de la fusion, afin d'assurer l'harmonisation des données fiscales et d'audit et de permettre un passage sans heurts aux étapes de la façon *machine learning* (ML). La *machine learning* (ML) prédit les fausses déclarations fiscales pour les contribuables ou les périodes non couvertes par les audits grâce à l'application d'algorithmes de forêt aléatoire. Ces modèles ont la capacité de fournir une estimation correcte puisqu'ils sont formés sur des données auditées et peuvent estimer avec précision la fraude fiscale pour les cas non audités, ce qui permet d'estimer l'écart fiscal de manière exhaustive. Par rapport aux modèles de régression traditionnels, le guide pratique a également comparé les performances des modèles de façon *machine learning* (ML), afin de mettre en évidence l'amélioration du pouvoir prédictif.

Enfin, certaines suggestions de travaux futurs impliquent d'étendre et d'améliorer le guide pratique actuel de plusieurs manières. Il serait possible d'utiliser d'autres algorithmes de *machine learning* (ML), tels que les réseaux neuronaux, et de comparer la précision des prédictions entre les différentes méthodes. La généralisation du guide pratique dans d'autres langages de programmation que STATA, élargissant ainsi sa portée, est un autre domaine qui pourrait être envisagé. Il est nécessaire de poursuivre les travaux concernant la manière dont le guide pratique pourrait être mis en œuvre dans différents contextes nationaux. Enfin, le guide pratique peut lui-même servir de point de départ à de futures recherches sur le comportement des contribuables, afin d'aider les autorités à concevoir et à mettre en œuvre de meilleures stratégies de contrôle et de meilleures mesures de conformité.

Références

Adu-Ababio, K., Koivisto, A., and Mwale, E. (2024). *Estimating tax gaps in Zambia.* (Preprint)

Alaimo Di Loro, P., Scacciatelli, D., and Tagliaferri, G. (2023). '2-step Gradient Boosting approach to selectivity bias correction in tax audit: an application to the VAT gap in Italy'. *Statistical Methods & Applications*, 32(1): 237–270. https://doi.org/10.1007/s10260-022-00643-4

Alsadhan, N. (2023). 'A Multi-Module Machine Learning Approach to Detect Tax Fraud'. *Computer Systems Science and Engineering*, 46(1): 241–253. https://doi.org/10.32604/csse.2023.033375

Athey, S., and Imbens, G. W. (2019). 'Machine learning methods that economists should know about'. *Annual Review of Economics*, 11(1): 685–725. https://doi.org/10.1146/annurev-economics-080217-053433

Baghdasaryan, V., Davtyan, H., Sarikyan, A., and Navasardyan, Z. (2022). 'Improving tax audit efficiency using machine learning: The role of taxpayer's network data in fraud detection'. *Applied Artificial Intelligence*, 36(1): 2012002. https://doi.org/10.1080/08839514.2021.2012002

Barra, P. A., Hutton, M. E., and Prokof'yeva, P. (2023). *Corporate Income Tax Gap Estimation by using Bottom-Up Techniques in Selected Countries: Revenue Administration Gap Analysis Program*. Washington, DC: International Monetary Fund. https://doi.org/10.5089/9798400246265.005

Battaglini, M., Guiso, L., Lacava, C., Miller, D. L., and Patacchini, E. (2024). 'Refining public policies with machine learning: The case of tax auditing'. *Journal of Econometrics*, 105847. https://doi.org/10.1016/j.jeconom.2024.105847

Békés, G., and Kézdi, G. (2021). 'Regression Trees'. In *Data analysis for business, economics, and policy* (pp. 417–437). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108591102.015

Bloomquist, K. M., Hamilton, S., and Pope, J. (2014). 'Estimating Corporation Income Tax Under-Reporting Using Extreme Values from Operational Audit Data'. *Fiscal Studies*, 35(4): 401–419. https://doi.org/10.1111/j.1475-5890.2014.12036.x

Brewer, D., and Carlson, A. (2024). 'Addressing sample selection bias for machine learning methods'. *Journal of Applied Econometrics*, 39(3): 383–400. https://doi.org/10.1002/jae.3029

Canada Revenue Agency (2019). *Tax gap and compliance results for the federal corporate income tax system.*

Chudý, M., Gábik, R., Bukovina, J., and Šrámková, L. (2020). *Searching for gaps: Bottom-up approach for Slovakia.* Institute for Financial Policy (IFP).

Cutler, A., Cutler, D. R., and Stevens, J. R. (2012). 'Random forests'. In C. Zhang and Y. Ma (eds), *Ensemble ma- chine learning: Methods and applications* (pp. 157–175). New York: Springer. ttps://doi.org/10.1007/978- 1-4419-9326-7_5

Durán-Cabré, J. M., Esteller Moré, A., Mas-Montserrat, M., and Salvadori, L. (2019). 'The tax gap as a public management instrument: application to wealth taxes'. *Applied Economic Analysis*, 27(81): 207–225. https://doi.org/10.1108/AEA-09-2019-0028

Ebrahim, A., Castillo, S., Leyaro, V., Swema, E., and Haule, O. (2024). *Estimating the Value-Added Tax Gap for SMMEs in Tanzania: An Empirical Analysis.* (Manuscript)

Feinstein, J. S. (1999). 'Approaches for estimating noncompliance: examples from federal taxation in the United States'. *The Economic Journal*, 109(456): 360–369. https://doi.org/10.1111/1468-0297.00439

FISCALIS Tax Gap Project Group (2018). 'The Concept of Tax Gaps: Corporate Income Tax Gap Estimation Methodologies'. Working paper 73 – 2018. Luxembourg: Publications Office of the European Union. (European Commission's Directorate-General Taxation and Customs Union) https://doi.org/10.2778/83206

Gemmell, N., and Hasseldine, J. (2014). 'Taxpayers' behavioural responses and measures of tax compliance 'gaps': A critique and a new measure'. *Fiscal Studies*, 35(3): 275–296. https://doi.org/10.1111/j.1475- 5890.2014.12031.x

Hartshorn, S. (2016). *Machine learning with random forests and decision trees: A Visual guide for beginners*. Kindle edition.

Heckman, J. J. (1979). 'Sample selection bias as a specification error'. *Econometrica*, 47(1): 153–161. https://doi.org/10.2307/1912352

Holtzblatt, J., and Engler, A. (2022). *Machine Learning and Tax Enforcement.* Tax Policy Center, Urban Institute & Brookings Institution.

Howard, B., Lykke, L., Pinski, D., and Plumley, A. (2020). 'Can Machine Learning Improve Correspondence Audit Case Selection? Considerations for Algorithm Selection, Validation, and Experimentation'. In A. Plumley (ed.), *The IRS Research Bulletin: Proceedings of the 2020 IRS / TPC Research Conference* (pp. 147–169). Internal Revenue Service.

Hutton, M. E. (2017). *The Revenue Administration–Gap Analysis Program: Model and Methodology for Value-Added Tax Gap Estimation.* International Monetary Fund. https://doi.org/10.5089/9781475583618.005

Ioana-Florina, C., and Mare, C. (2021). 'The utility of neural model in predicting tax avoidance behavior'. In I. Czarnowski, R. Howlett, and L. Jain (eds), *Intelligent Decision Technologies: Proceedings of the 13th KES-IDT 2021 Conference* (pp. 71–81). https://doi.org/10.1007/978-981-16-2765-1_6

Italian Revenue Agency (n.d.). *Italy: VAT gap estimation via bottom up approach.*

Murorunkwere, B. F., Haughton, D., Nzabanita, J., Kipkogei, F., and Kabano, I. (2023). 'Predicting tax fraud using supervised machine learning approach'. *African Journal of Science, Technology, Innovation and Development*, 15(6): 731–742. https://doi.org/10.1080/20421338.2023.2187930

Murorunkwere, B. F., Tuyishimire, O., Haughton, D., and Nzabanita, J. (2022). 'Fraud detection using neural networks: A case study of income tax'. *Future Internet*, 14(6): 168. https://doi.org/10.3390/fi14060168

Pérez López, C., Delgado Rodríguez, M. J., and de Lucas Santos, S. (2019). 'Tax fraud detection through neural networks: An application using a sample of personal income taxpayers'. *Future Internet*, 11(4): 86. https://doi.org/10.3390/fi11040086

Plumley, A. (2005). 'Preliminary update of the tax year 2001 individual income tax underreporting gap estimates'. In *Proceedings. Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association* (Vol. 98, pp. 19–25).

Raikov, A. (2021). 'Decreasing tax evasion by artificial intelligence'. *IFAC-PapersOnLine*, 54(13): 172–177.

Savic´, M., Atanasijevic´, J., Jakovetic´, D., and Krejic´, N. (2022). 'Tax evasion risk management using a Hybrid Unsupervised Outlier Detection method'. *Expert Systems with Applications*, 193(May): 116409. https://doi.org/10.1016/j.eswa.2021.116409

Schonlau, M., and Zou, R. Y. (2020). 'The random forest algorithm for statistical learning'. *The Stata Journal*, 20(1): 3–29. https://doi.org/10.1177/1536867X20909688

Varian, H. R. (2014). 'Big data: New tricks for econometrics'. *Journal of Economic Perspectives*, 28(2): 3–28.

Warren, N. (2018, April). 'Estimating Tax Gap is Everything to an Informed Response to the Digital Era'. In *13th International Revenue Administration Conference on Tax System Integrity in a Digital Age* (p. 1-41). Disponible sur: https://ssrn.com/abstract=3200838 (dernière révision: 23 juin 2019)

Zacharis, N. Z. (2018). 'Classification and regression trees (CART) for predictive modeling in blended learning'. *IJ Intelligent Systems and Applications*, 3(1): 9. https://doi.org/10.5815/ijisa.2018.03.01

Zumaya, M., Guerrero, R., Islas, E., Pineda, O., Gershenson, C., Iñiguez, G., and Pineda, C. (2021). 'Identi- fying tax evasion in Mexico with tools from network science and machine learning'. In O. Granados and J. Nicolás-Carlock (eds), *Corruption networks: Concepts and applications* (pp. 89–113). Cham: Springer. https://doi.org/10.1007/978-3-030-81484-7_6

Annex A : Algorithme de *random forests*

La méthode *random forests* est considérée comme l'un des algorithmes de *machine learning* (ML) les plus utilisés et les plus performants pour les tâches de prédiction (Athey et Imbens 2019).[4] Contrairement aux modèles de régression traditionnels, qui supposent la linéarité et luttent lorsque le nombre d'observations est inférieur aux variables indépendantes, la méthode *random forests* peut gérer des relations non linéaires dans les données et elle évite le problème de l'estimation d'un nombre de paramètres supérieur à celui que les points de données peuvent supporter. En outre, elle saisit mieux l'existence de valeurs aberrantes et atypiques, ce qui permet d'obtenir des prédictions plus précises dans de tels cas (Athey et Imbens 2019). Elle y parvient, en n'utilisant pas toutes les variables prédictives en même temps, ce qui permet d'obtenir de meilleures prédictions que la régression traditionnelle (Schonlau et Zou 2020). Outre sa simplicité d'utilisation, la méthode *random forests* est facile à comprendre et rapide à mettre en œuvre. De plus, elle donne de bons résultats par rapport à d'autres algorithmes de *machine learning* (ML) (Varian 2014).

Par essence, une méthode *random forests* peut nous permettre de prédire la variable cible (y), à l'aide de variables d'entrée (x). Il s'agit essentiellement d'une collection d'arbres de décision créés à partir de sous-ensembles aléatoires de données. Mais que sont les arbres de décision et comment sont-ils utilisés pour créer un modèle de *random forests*? Pour répondre à cette question, le document commence par expliquer les concepts des arbres de décision et leur fonctionnement, puis il explique comment construire un modèle de *random forests* et l'utiliser pour accomplir des tâches de prédiction.

A1 Arbres de décisions

Les arbres de décision sont un type d'algorithme d'apprentissage supervisé utilisé pour les tâches de régression et de classification. Ils divisent les données en sous-ensembles basés sur les valeurs des variables d'entrées (x) pour prédire les valeurs (y). Ce processus de division se poursuit jusqu'à ce que les données de chaque sous-ensemble soient aussi homogènes que possible en ce qui concerne la variable cible. Il est également connu sous le nom d'algorithme des arbres de classification et de régression (CART), qui permet de trouver le meilleur découpage à chaque étape afin de maximiser la précision de la prédiction.

Algorithme CART

Types de CART :

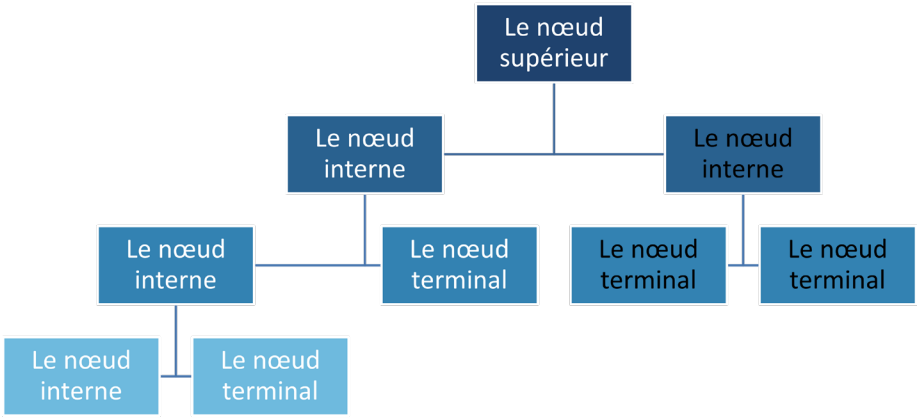
- Les arbres de classification sont un type d'algorithme d'arbre de décision utilisé pour classer des variables cibles catégoriques. Ils fonctionnent en segmentant l'espace des prédicteurs en régions distinctes, chaque région correspondant à une étiquette de classe spécifique. L'objectif est de déterminer la catégorie à laquelle appartient la variable cible en fonction des caractéristiques d'entrée.
- Les arbres de régression sont un type d'algorithme d'arbre de décision conçu pour prédire des variables cibles continues. Ils divisent l'espace de prédiction en régions et fournissent une valeur continue en sortie pour chaque région.

[4] Une méthode d'ensemble combine plusieurs modèles simples, connus sous le nom d'apprenants faibles, pour créer un modèle prédictif unique et plus puissant.

Comment fonctionne l'algorithme CART ?

La structure de construction d'un arbre de décision à l'aide de CART commence par le nœud supérieur, qui représente l'ensemble des données. Ce nœud supérieur est le point de départ de l'arbre. À partir de ce point, l'algorithme identifie le meilleur attribut pour diviser l'ensemble de données et étiquette le nœud avec cet attribut. Cela crée des branches qui mènent à des nœuds internes, où chaque nœud interne représente une décision basée sur la valeur de l'attribut choisi. Les données sont à nouveau divisées à chaque nœud interne, ce qui permet de générer davantage de branches et de nœuds. Ce processus se répète, créant une structure hiérarchique. Les extrémités de ces branches sont les nœuds terminaux qui fournissent la prédiction finale, comme le montre la Figure A1. Dans les tâches de classification, la prédiction à un nœud terminal est la classe majoritaire des observations dans ce nœud et, dans les tâches de régression, il s'agit de la valeur moyenne des observations.

Figure A1 : Structure de l'arbre de décision



Source : illustration des auteurs.

Dans les tâches de régression, CART utilise la réduction des résidus comme critère de division. Cela implique de partitionner les données à chaque nœud pour minimiser la différence quadratique moyenne entre les valeurs prédites et les valeurs réelles, afin d'obtenir l'erreur résiduelle la plus faible. Pour les tâches de classification, CART utilise l'impureté de Gini pour évaluer toutes les divisions potentielles, en choisissant celle qui réduit le plus efficacement l'impureté et augmente ainsi la pureté des sous-ensembles résultants entre les valeurs prédites et les valeurs réelles, afin d'obtenir l'erreur résiduelle la plus faible. L'impureté de Gini quantifie la probabilité de mal classer une instance aléatoire sur la base de la classe majoritaire au sein d'un sous-ensemble. Ce processus de division est récursif et il se poursuit jusqu'à ce que certains critères d'arrêt soient remplis. Ces critères comprennent l'atteinte d'un nœud où tous les enregistrements partagent la même valeur cible, la taille du nœud inférieure à un seuil défini par l'utilisateur, l'atteinte de la profondeur maximale prédéfinie de l'arbre, la présence d'un nombre de cas inférieur à un nombre minimal dans un nœud ou le fait qu'un fractionnement supplémentaire n'améliore pas de manière significative la pureté (Zacharis 2018).

Un risque essentiel lors de l'utilisation d'arbres de décision est le surajustement du modèle. Cela peut se produire, si le modèle croît sans contraintes, par exemple lorsqu'un arbre de régression continue à se diviser jusqu'à ce que chaque nœud terminal ne contienne qu'une seule observation. Bien qu'il puisse en résulter un ajustement presque parfait aux données d'apprentissage (voir la

Note ci-dessous pour une définition), cela a un impact négatif sur la capacité du modèle à se généraliser à de nouvelles données inédites. Les modèles surajustés donnent généralement de bons résultats sur les données d'apprentissage, mais de piètres résultats sur les données de validation ou de test, parce qu'ils ont appris le bruit plutôt que le signal.

Pour résoudre les problèmes de surajustement, CART utilise une technique d'élagage une fois que l'arbre a atteint sa taille maximale. L'élagage consiste à couper l'arbre pour éliminer les nœuds qui n'apportent qu'une valeur prédictive minimale, ce qui permet de simplifier le modèle et d'améliorer sa généralisation. Une technique largement utilisée est l'élagage de la complexité des coûts, où un grand arbre est initialement développé, en utilisant un très petit paramètre de complexité pour s'assurer que toutes les divisions potentielles sont évaluées. Ensuite, les divisions sont supprimées séquentiellement et les performances du modèle sont réévaluées à l'aide de la validation croisée. Ce processus se poursuit jusqu'à ce qu'un élagage supplémentaire n'améliore pas l'adéquation du modèle (Békés et Kézdi 2021).

La Figure A2 présente un exemple de pseudocode pour un algorithme de croissance d'arbre qui explique les tâches de classification et de régression. Considérons un scénario dans lequel nous cherchons à construire un arbre de décision pour prédire une variable cible à l'aide d'un ensemble de données *X,* qui contient plusieurs co-variables *A* et la variable cible *y*. Le paramètre de tâche indique s'il s'agit d'une classification ou d'une régression.

L'algorithme commence par initialiser un arbre unique *T* avec un nœud supérieur. Si tous les critères d'arrêt sont remplis, l'algorithme procède à l'étiquetage du nœud. Pour les tâches de classification, le nœud est étiqueté avec la classe la plus courante parmi les échantillons de *X*. Pour les tâches de régression, le nœud est étiqueté avec la valeur moyenne de *y*.

Si les critères d'arrêt n'ont pas été remplis, l'algorithme recherche le meilleur attribut $a \in A$ qui divise l'ensemble de données *X* le plus efficacement possible. Les tâches de classification sont effectuées à l'aide d'une fonction d'impureté telle que l'impureté de Gini. Pour les tâches de régression, l'algorithme vise à minimiser la variance au sein des nœuds. Le nœud est alors étiqueté avec l'attribut *a*.

Figure A2 : Pseudocode de l'algorithme de croissance de l'arbre pour les tâches de classification et de régression

Algorithme 1 Algorithme de croissance arborescente `arbre de croissance`(*X, A, y,* tâche)

Exigence : Ensemble de données d'apprentissage *X,* ensemble d'attributs *A*, variable de sortie *y*, tâche (classification ou régression)

Assurer : Arbre de décisions

1: Commencez par un arbre unique *T* avec un nœud supérieur

2: si tous les critères d'arrêt sont remplis, **alors**

3: **s'il** y a tâche == classification **alors**

4: *T* a un nœud avec la classe la plus courante dans *X* comme étiquette

5: **plus**

6: *T* a un nœud avec la moyenne de *y* in *X* comme étiquette

7: **la fin si=**

8: **plus**

9: trouver $a \in A$, qui divise au mieux *X* en utilisant la fonction d'impureté (pour la classification) ou en minimisant la variance (pour la régression)

10: Étiqueter le nœud avec *a*

11: **pour** une valeur possible *v* de *a* **fait**

12: X_v = le sous-ensemble de *X* qui a $a = v$

13: $A_v = A - a$

14: `arbre de croissance`(X_v, A_v, *y,* tâche)

15: connecter le nouveau nœud au nœud supérieur avec l'étiquette *v*

16: **la fin pour**

17: **la fin si**

18: **retour** de `l'arbre d'élagage`(*X , A, y,* tâche)

Note

Dans la façon *machine learning* (ML), nous divisons les données en deux sous-ensembles principaux :

L'ensemble d'apprentissage : Ce sous-ensemble est utilisé pour construire des modèles tels que les arbres de régression et les forêts aléatoires. Il comprend des caractéristiques d'entrée (variables indépendantes) et la variable cible (variable dépendante). Le système apprend des modèles et des relations à partir de ces données.

Ensemble de test : Ce sous-ensemble est utilisé pour évaluer les performances du modèle. L'ensemble de test n'est pas vu par le modèle pendant la phase de formation, ce qui permet une évaluation impartiale de la manière dont le modèle se généralise à de nouvelles données non vues.

Ensuite, l'algorithme itère sur toutes les valeurs possibles *v* de l'attribut choisi *a*. Pour chaque valeur *v*, il crée un sous-ensemble de *X* où l'attribut *a* prend la valeur *v*. Il met également à jour l'ensemble d'attributs *A* en supprimant l'attribut *a*. L'algorithme s'appelle ensuite récursivement pour faire croître l'arbre davantage, en utilisant le sous-ensemble de *X* et l'ensemble d'attributs *A* mis à jour. Ce processus récursif se poursuit, connectant de nouveaux nœuds au nœud supérieur avec des étiquettes correspondant aux valeurs *v*.

Une fois que l'arbre a atteint sa taille maximale sur la base des critères initiaux, l'algorithme procède à l'élagage de l'arbre. Le processus d'élagage implique l'utilisation d'une fonction d'élagage distincte qui évalue si la suppression de certains nœuds et branches améliore les performances de l'arbre sur un ensemble de données de test. Pour ce faire, on utilise des techniques de validation croisée afin de s'assurer que l'arbre se généralise bien à des données inédites.

En répétant ce processus, l'algorithme d'élagage construit un arbre de décision qui divise l'ensemble de données *X* en régions de plus en plus petites. Chaque nœud terminal (feuille) de l'arbre correspond à une région spécifique de l'espace des caractéristiques. Dans les tâches de classification, le nœud de la feuille représente la classe majoritaire dans cette région, tandis que dans les tâches de régression, il représente la valeur moyenne de *y*.

A2 La méthode *random forests*

Les arbres de décision, bien qu'utiles, présentent des limites notables, en particulier leur tendance à surajuster les données malgré l'élagage. Dans le monde réel, les données peuvent être désordonnées et contenir des anomalies qui ne se généralisent pas bien. Les arbres de décision peuvent créer des divisions très spécifiques qui s'adaptent aux données d'apprentissage, mais qui ne donnent pas de bons résultats sur de nouvelles données inédites. Les forêts aléatoires résolvent ce problème, en utilisant plusieurs arbres décisionnels et en calculant la moyenne de leurs résultats. La simple génération de plusieurs arbres à partir du même ensemble de données ne résout pas le problème, car elle produirait des arbres similaires. Au lieu de cela, les forêts aléatoires créent des arbres en utilisant des sous-ensembles aléatoires de données. Ce processus d'utilisation de sous-ensembles variés garantit que les arbres sont différents, ce qui permet d'atténuer les anomalies et d'améliorer la précision globale des prédictions, en combinant les divers arbres en un modèle plus robuste.

Agrégation bootstrap et critères de sélection

Dans la méthode *random forests*, le caractère aléatoire est introduit de deux manières principales. Premièrement, en sélectionnant un sous-ensemble aléatoire de données pour chaque arbre, et deuxièmement, en choisissant un sous-ensemble aléatoire de variables prédictives pour chaque division de l'arbre. Chaque arbre d'une *random forests* est construit à l'aide d'une technique appelée agrégation bootstrap, ou bagging. L'algorithme de bagging fonctionne en prélevant d'abord de multiples échantillons aléatoires dans l'ensemble de données d'origine. Supposons que nous prélevions *B* échantillons, où *B* est un grand nombre, généralement de l'ordre de la centaine. Pour chaque échantillon, un grand arbre de décision est créé sans aucune simplification. Ces arbres sont ensuite utilisés pour faire des prédictions. L'algorithme crée *B* règles de prédiction à partir de ces arbres et les combine. Dans une configuration où nous testons la précision du modèle, *B* prédictions sont faites pour chaque point de données sur la base des résultats de chacun des arbres *B*. L'étape finale consiste à faire la moyenne de ces prédictions *B* pour obtenir la valeur prédite finale.

Les forêts aléatoires introduisent également un caractère aléatoire, en limitant les caractéristiques prises en compte à chaque division. Plutôt que d'évaluer toutes les variables prédictives (variables *x*) à chaque point de ramification, l'algorithme ne sélectionne au hasard qu'un sous-ensemble de ces variables à prendre en considération. La taille de ce sous-ensemble est généralement prédéterminée, souvent autour de la racine carrée du nombre total de prédicteurs, avec un minimum communément fixé à 4. Cette approche est appliquée à chaque échantillon bootstrap, ce qui entraîne la construction d'arbres *B*. La prédiction finale est réalisée, en faisant la moyenne des sorties de ces arbres *B*.

La raison d'être de l'utilisation d'un nombre limité de variables prédictives à chaque division est de minimiser la probabilité que tous les arbres deviennent trop similaires, en particulier si un prédicteur fort est dominant. En restreignant l'ensemble des variables à chaque point de décision, l'algorithme permet une contribution plus équilibrée de tous les prédicteurs, y compris les plus faibles qui pourraient fournir des informations précieuses lorsqu'ils sont considérés ensemble. Sans cette sélection aléatoire, les arbres résultants favoriseraient fortement les prédicteurs les plus forts, ce qui conduirait à des prédictions fortement corrélées et moins diversifiées.

Ajustement du modèle

Lors de l'exécution d'une forêt aléatoire, plusieurs paramètres de réglage clés doivent être pris en compte pour garantir une performance optimale du modèle. Les principaux paramètres comprennent le nombre d'arbres, le nombre de prédicteurs évalués à chaque division et la règle d'arrêt pour la croissance des arbres.

- Nombre d'arbres (B) :
  - Ce paramètre contrôle le nombre d'échantillons bootstrap utilisés pour construire la forêt. Un plus grand nombre d'arbres augmente généralement la précision du modèle, mais aussi le temps de calcul.
- Nombre de prédicteurs par division (x) :
  - À chaque nœud, seul un sous-ensemble de prédicteurs est sélectionné pour la division. Une bonne règle consiste à utiliser la racine carrée du nombre total de prédicteurs. Par exemple, avec 64 prédicteurs, il faut en utiliser environ 8 pour chaque fractionnement. Au moins quatre prédicteurs doivent être utilisés.
- Règle d'arrêt pour la croissance de l'arbre :
  - Déterminer le moment où il faut cesser de diviser les nœuds d'un arbre. Une règle simple consiste à fixer un nombre minimum d'observations par nœud terminal. En général, on utilise de 5 à 20 observations.

La méthode examine ensuite la combinaison de ces trois paramètres de réglage qui produit l'erreur de prédiction la plus faible. Cette erreur est mesurée par l'erreur quadratique moyenne (RMSE), qui nous indique à quel point nos prédictions sont éloignées des valeurs réelles.

Une autre mesure importante est l'erreur hors du sac, abrégée en OOB. Cette mesure permet d'évaluer les performances du modèle. Lors de la construction de chaque arbre de la forêt, l'algorithme échantillonne au hasard environ 63,2 pourcentage des données, laissant les 36,8 pourcentage restants inutilisés ou « hors du sac ». Ces données hors du sac ne sont pas utilisées dans la construction d'un arbre particulier, mais elles peuvent être utilisées pour estimer la précision de cet arbre, en testant la façon dont l'arbre prédit les données « hors du sac » (OOB). La moyenne de ces erreurs « hors du sac » (OOB pour tous les arbres de la forêt fournit une estimation fiable de la performance du modèle, connue sous le nom de taux d'erreur « hors du sac » (OOB). Cette technique garantit que tous les points de données sont évalués dans le cadre de l'évaluation des performances du modèle, ce qui permet d'obtenir une mesure solide de la précision sans avoir besoin d'un ensemble de tests distinct (Hartshorn 2016).

Importance des variables

Dans la forêt aléatoire, il est essentiel de comprendre l'importance de chaque variable prédictive pour interpréter le modèle et affiner sa précision prédictive. La méthode utilise une méthode directe connue sous le nom d'importance de permutation, qui évalue l'importance des variables, en observant les changements dans la précision de la prédiction, lorsque les valeurs de chaque prédicteur sont mélangées de manière aléatoire. La performance de prédiction du modèle est

ensuite comparée en utilisant à la fois les valeurs originales et permutées de la variable, en particulier en utilisant des données hors sac (OOB). L'importance de la permutation est calculée en mesurant l'augmentation de l'erreur de prédiction - telle que l'erreur quadratique moyenne (RMSE) pour les tâches de régression ou le taux d'erreur pour les tâches de classification - lorsque les valeurs d'une variable sont permutées dans les données « hors du sac » (OOB). Une augmentation significative de l'erreur indique l'importance de la variable. Cette technique permet non seulement d'identifier les prédicteurs clés, mais aussi de saisir les interactions complexes entre les variables. Étant donné que l'algorithme de *random forests* sélectionne des sous-ensembles aléatoires de prédicteurs pour chaque division, l'algorithme peut identifier tous les prédicteurs corrélés, comme étant importants, si l'un d'entre eux contribue de manière significative au résultat (Cutler et al. 2012).

A3 Exemple

Cette section développe un exemple simple pour clarifier les caractéristiques de la forêt aléatoire. À cette fin, nous nous concentrerons sur le développement du modèle et de la prédiction, en expliquant chaque étape mais sans fournir d'exemples empiriques.

Imaginons une population de contribuables égale à 100. Chaque contribuable remplit une déclaration fiscale comprenant la base d'imposition (montant sur lequel les impôts sont prélevés) et des informations complémentaires. Supposons que l'information complémentaire soit composée de dix variables. Par exemple, il peut s'agir du montant payé au titre des salaires des employés et des coûts de production, entre autres. Il est important de noter que ces informations complémentaires ne font pas directement partie de la base d'imposition, mais qu'elles peuvent être utiles pour comprendre comment atteindre le niveau de la base d'imposition.

Sur les 100 contribuables, 50 ont été contrôlés. Cela signifie que pour 50 contribuables, nous disposons également d'informations sur les écarts (potentiels) entre la déclaration de l'assiette fiscale et le montant réel. Pour clarifier ce point, imaginons que les 50 contribuables fraudent le fisc et que, grâce aux contrôles, nous collectons (au moins) le montant faussement déclaré et l'assiette fiscale réelle.

La première étape consiste à comprendre que nous ne disposons d'informations sur les montants mal déclarés que pour 50 contribuables. Cela signifie que ce n'est que dans ce sous-échantillon que nous pouvons comparer les prédictions avec les variables réelles afin de tester la précision du modèle de prédiction. C'est pourquoi nous allons diviser l'ensemble de l'échantillon et nous concentrer sur les contribuables contrôlés.

L'échantillon de contribuables contrôlés est divisé en deux sous-échantillons. Pour des raisons de simplicité, nous conserverons 25 contribuables, comme échantillon de formation et le reste comme échantillon de test. Dans l'échantillon de formation, nous exécuterons le modèle de *random forests* et nous utiliserons l'échantillon de test pour l'ajuster. Nous devons choisir deux nombres critiques : le nombre d'itérations (ou le nombre d'arbres) et le prédicteur par fractionnement. Le modèle se concentre sur l'estimation de la quantité de déclarations erronées à l'aide des co-variables (les dix variables supplémentaires déclarées par les entreprises). Nous utiliserons donc toutes les variables pour deux raisons principales. Premièrement, ces variables aident à caractériser la base d'imposition, étant pertinentes pour déterminer ce niveau. Deuxièmement, étant donné que ces informations sont disponibles, elles sont également essentielles pour décider quel contribuable sera contrôlé. L'intégration de toutes les variables nous permet d'éviter un biais de sélection de l'échantillon par des facteurs observables.

Voyons d'abord le nombre d'arbres à utiliser. Pour cela, nous conservons le nombre de prédicteurs utilisés dans chaque fractionnement (variables utilisées pour estimer les déclarations erronées). Pour simplifier, supposons que nous utiliserons l'une des dix variables disponibles. Pour décider du nombre d'arbres, nous devons exécuter le modèle dans l'échantillon d'apprentissage et comparer la prédiction dans l'échantillon de test, en utilisant différents nombres d'arbres. En d'autres termes, nous exécutons N fois N forêts aléatoires différentes en changeant uniquement le nombre d'arbres que nous utilisons. Lors de l'exécution du modèle de forêt aléatoire, nous faisons des prédictions dans l'échantillon de test et comparons la valeur prédite avec les déclarations erronées réelles constatées lors du processus d'audit. À la fin, nous aurons N valeurs de l'erreur quadratique moyenne (RMSE) (une par exécution du modèle). Nous choisissons la valeur minimale et nous voyons le nombre d'arbres associés, B. Ce nombre est optimal car il minimise l'erreur de prédiction mesurée par l'erreur quadratique moyenne (RMSE), produisant ainsi l'estimation la plus précise de l'assiette fiscale ayant fait l'objet d'une déclaration erronée.

Nous passons maintenant à l'estimation du prédicteur utilisé dans chaque fractionnement. Dans ce cas, nous utilisons le nombre optimal d'arbres, B, obtenu précédemment. Nous répétons la même procédure itérative, mais dans ce cas, nous exécutons dix modèles de *random forests* différents, en obtenant la prédiction dans l'échantillon d'apprentissage pour chacun d'entre eux et en la comparant à la valeur réelle de la déclaration erronée. Nous exécutons dix modèles parce que nous avons dix variables à utiliser. En effet, le nombre total de variables correspond au nombre maximum de prédicteurs pour chaque fractionnement. Il est important de noter que si vous disposez de dix variables, mais que vous décidez d'en utiliser huit pour le modèle de prédiction, vous devez exécuter huit modèles. Le nombre de modèles à exécuter dans cette étape doit toujours être égal aux variables que vous avez décidé d'utiliser dans le modèle de prédiction. Enfin, nous répétons le processus, en choisissant l'erreur quadratique moyenne (RMSE) minimum et en observant le nombre de prédicteurs utilisés, x. Ce nombre de prédicteurs x est optimal pour minimiser l'erreur de prédiction.

Ces deux étapes nous permettent de trouver le nombre optimal d'arbres (B) et de prédicteurs par division (x) à utiliser dans la forêt aléatoire. Rappelons que pour les estimer, nous utilisons les 50 contribuables contrôlés, en divisant cet échantillon en deux ensembles, l'un pour la formation et l'autre pour le test. Nous pouvons maintenant faire les prédictions pour les 50 autres contribuables qui n'ont pas été contrôlés. La procédure est la suivante. Tout d'abord, nous exécutons la méthode *random forests* avec les paramètres optimaux sur l'ensemble des 50 contribuables contrôlés. Ensuite, prédire les valeurs dans l'ensemble des 50 contribuables non contrôlés. Enfin, vous pouvez créer une variable composée de la déclaration erronée découverte pour les 50 contribuables contrôlés et de la déclaration erronée prédite pour les 50 contribuables non contrôlés.