UNU WIDER

5
24

Tax research for development

Conjunto de ferramentas para a estimativa das lacunas fiscais, utilizando uma abordagem ascendente

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

Dezembro 2024

WIDER TECHNICAL NOTE

wider.unu.edu

United Nations University
World Institute for Development Economics Research

Resumo: Esta nota técnica diz respeito ao conjunto de ferramentas sobre as disparidades fiscais, que inclui código (o do-file do Stata) e um arquivo README (como executar o código). A nota técnica descreve a literatura e a metodologia subjacentes ao desenvolvimento do conjunto de ferramentas. O conjunto de ferramentas de lacunas fiscais está relacionado com a estimativa dos desvios do imposto sobre o valor acrescentado (IVA), do imposto sobre o rendimento das pessoas colectivas (IRPC) e do imposto sobre o rendimento das pessoas singulares (IRPS), utilizando a abordagem ascendente, em que as auditorias operacionais e o imposto potencial são calculados utilizando a aprendizagem automática.

Palavras-chave: conjunto de ferramentas, lacuna fiscal, abordagem ascendente, auditorias operacionais, aprendizagem automática

Códigos de classificação do *Journal of Economic Literature* (JEL): H25, H26, H32

Agradecimentos: Os autores agradecem os contributos de Jukka Pirttilä, Maria Jouste, Gerald Agaba e Hilja-Maria Takala. Os autores agradecem o financiamento do International Tax Compact (ITC). O ITC facilita o Secretariado da Addis Tax Initiative (ATI), que apoiou o desenvolvimento do conjunto de ferramentas Tax Gap. O ITC é financiado pelo Ministério Federal Alemão da Cooperação Económica e do Desenvolvimento, cofinanciado pela União Europeia e implementado pela Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH. Este trabalho faz parte do programa de Domestic Revenue Mobilization do UNU-WIDER, que é financiado pela Norad.

Material suplementar: O código relacionado e os arquivos README podem ser descarregados gratuitamente a partir da página Web do conjunto de ferramentas.

Publicações relacionadas:

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

Esta nota técnica está disponível em inglês (original), francês e espanhol.

[1] Universidade de Tampere, Finlândia; [2] Universidade de Helsínquia, Finlândia, e Finnish Centre of Excellence in Tax Systems Research (FIT); [3] UNU-WIDER, Helsínquia, Finlândia; correspondencia: amina@wider.unu.edu

Este estudo foi preparado no âmbito do projeto UNU-WIDER Tax research for development (phase 3), que faz parte da área de investigaçãoCreating the fiscal space for development. O projeto faz parte do programa de Domestic Revenue Mobilization, que é financiado através de contribuições específicas da Agência Norueguesa de Cooperação para o Desenvolvimento (Norad). O conjunto de ferramentas Tax Gap recebeu apoio financeiro do International Tax Compact (ITC), que facilita o Secretariado da Addis Tax Initiative (ATI). O ITC é financiado pelo Ministério Federal Alemão da Cooperação Económica e do Desenvolvimento, cofinanciado pela União Europeia e implementado pela Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH.

https://doi.org/10.35188/UNU-WIDER/WTN/2023-5

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

Katajanokanlaituri 6 B, 00160 Helsinki, Finland

1 Introdução

A mobilização de receitas internas para financiar a despesa pública é um objectivo essencial para muitos governos. Consequentemente, as autoridades fiscais desempenham um papel crucial na criação de sistemas fiscais eficientes e eficazes que possam garantir o montante ideal de receitas com um mínimo de fuga. No entanto, as receitas ideais ou máximas alcançáveis ficam aquém do resultado desejado, porque os indivíduos e as empresas responsáveis chegam a extremos para evitar ou sonegar essas obrigações. Com este facto conhecido, esta pesquisa procura medir a diferença entre as receitas fiscais efectivas e as receitas fiscais potenciais, com o objectivo de quantificar o grau de perda de receitas que ocorreria, se todos os indivíduos e as empresas aderissem plenamente às regras da política fiscal.

Esta nota técnica ilustra a maneira como as lacunas fiscais, definidas como a diferença entre as receitas fiscais efectivas e as receitas fiscais potenciais, são calculados com base numa metodologia ascendente que emprega várias formas de retorno e dados estatísticos. Além disso, esta nota acompanha o conjunto de ferramentas sobre as lacunas fiscais, como um precursor dos conceitos gerais de lacunas fiscais, com ênfase na chamada abordagem ascendente para estimar esses diferenças. O conjunto de ferramentas foi desenvolvido pelo Instituto Mundial de Pesquisa em Economia do Desenvolvimento da Universidade das Nações-Unidas (UNU-WIDER) e procura simplificar os conceitos complexos desta abordagem de estimativa, orientando os utilizadores, através de formas sistemáticas de medição, utilizando parâmetros disponíveis ao seu alcance.

Com uma definição alargada de lacunas fiscais, existe uma variedade de abordagens de estimativa do conceito. Contudo, todas elas abordam e investigam a razão pela qual as receitas fiscais efectivas se desviam das receitas fiscais potenciais. Alguns métodos gerais e habitualmente utilizados concentram-se em indicadores macroeconómicos agregados, como referência para o desvio, enquanto os métodos menos utilizados se baseiam em dados microeconómicos para a estimativa do desvio. A escolha depende do acesso e da disponibilidade de dados administrativos, consoante as jurisdições. Embora esta nota técnica apresente as principais abordagens para estimar as lacunas fiscais, a tónica é colocada na utilização de dados a nível microeconómico provenientes de auditorias operacionais (o caso mais frequente) e nos casos em que as auditorias aleatórias são raras.

Na nota que se segue, começamos por apresentar as definições de lacunas fiscais e a forma como a ideia se transforma num conceito, em função de variáveis de interesse específicas relacionadas com parâmetros de política e de conformidadena Secção 2. Em seguida, na Secção 3, são discutidos os métodos gerais utilizados para estimar os desvios de tributação, sendo a abordagem ascendente a abordagem de base, descrita nas suas várias formas e métodos. Na Secção 4, descrevemos os componentes do conjunto de ferramentas, que inclui duas fases principais: limpeza dos dados e uma abordagem de aprendizagem automática para a estimativa das lacunas fiscais.

2 Definição de Lacuna Fiscal

As autoridades fiscais deparam-se frequentemente com uma diferença notável entre as receitas fiscais previstas e o que é efectivamente cobrado. Esta diferença, conhecida como perda de receitas, surge principalmente quando os impostos devidos num determinado período não são pagos. Este imposto devido pelos contribuintes representa o montante de imposto que teoricamente poderia ser cobrado. Daí resulta o conceito de lacuna fiscal, definido como a

diferença entre as receitas fiscais efectivamente cobradas e as cobranças fiscais teóricas em caso de conformidade integral do código fiscal.

De um ponto de vista político, a lacuna fiscal pode ser caracterizada de forma mais ampla por duas componentes principais: a lacuna de conformidade e a lacuna de política fiscal. A lacuna de conformidade refere-se à diferença entre as receitas efectivamente cobradas num determinado ano e as receitas máximas possíveis que poderiam ter sido obtidas com base nas actividades económicas que ocorreram nesse período. A lacuna de políticas fiscais é o resultado de decisões legislativas destinadas a alterar a regulamentação fiscal normal, introduzindo isenções específicas, deduções ou taxas reduzidas para determinados casos (Hutton 2017). As alterações no quadro de políticas podem fazer com que a lacuna de políticas fiscais aumente ou diminua. Por exemplo, se o limiar de tributação zero for aumentado, permitindo que uma parte maior do rendimento fique isenta de impostos, ou se for introduzida uma taxa de imposto reduzida para um grupo específico de contribuintes, como as pequenas empresas ou as pessoas com baixos rendimentos, a lacuna de políticas fiscais aumentaria, uma vez que são cobradas menos receitas em comparação com o potencial máximo ao abrigo das regras fiscais normais. Por outro lado, a lacuna de políticas fiscais poderia também aumentar sem qualquer alteração do quadro de políticas, devido a alterações na composição da base tributável que tornassem uma parte maior do rendimento líquido sujeita à taxa de imposto normal (Barra et al. 2023).

A lacuna de conformidade é composta por dois elementos: a lacuna de avaliação e a lacuna de arrecadação. A lacuna de avaliação resulta principalmente de actividades económicas que as autoridades fiscais desconhecem ou que não conseguem alcançar, incluindo actividades de entidades que não estão registadas, que não apresentam declarações, que não declaram ou que declaram erradamente os seus impostos, tal como evidenciado em jurisdições com elevada informalidade. A lacuna de arrecadação de cobrança refere-se à diferença entre as obrigações fiscais calculadas, tendo em conta eventuais reembolsos e retenções na fonte, e os impostos efectivamente pagos. Engloba os montantes de impostos pendentes de que as autoridades fiscais têm conhecimento, mas que não foram recuperados com êxito, porque normalmente estão ligados a litígios ou são considerados demasiado dispendiosos para serem perseguidos ou impossíveis de cobrar por meios legais.

A literatura também distingue três componentes da lacuna de conformidade que complementam as lacunas de avaliação e arrecadação acima referidas (Gemmell e Hasseldine 2014; Durán-Cabré et al. 2019).

1. *Componente de sub-avaliação:* os contribuintes declaram menos rendimentos do que efectivamente auferiram ou reclamam mais deduções, créditos ou outros benefícios fiscais do que os permitidos por lei ou uma combinação de ambos. Isto cria uma diferença entre a verdadeira obrigação fiscal do contribuinte e o montante declarado.

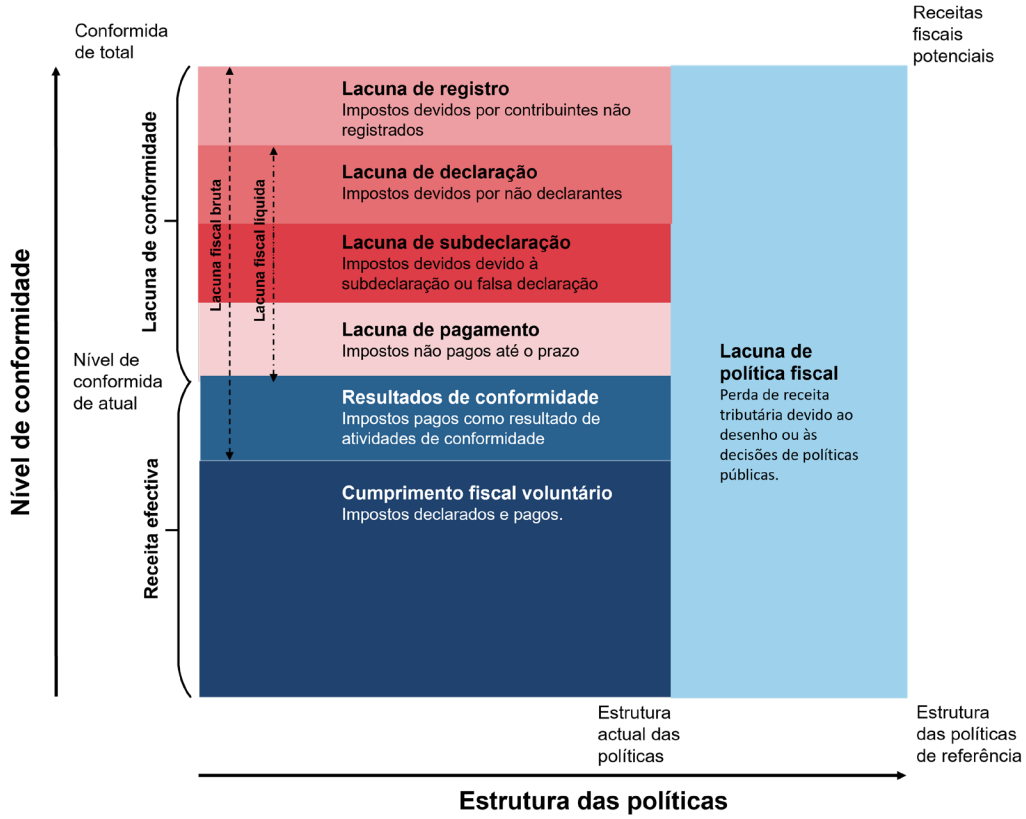
2. *Componente de não-preenchimento:* indica a diferença entre os declarantes potenciais e os declarantes efectivos, revelando a extensão da não declaração e da evasão fiscal.

3. *Componente de não-conformidade:* indica a diferença entre as receitas fiscais potenciais e as receitas fiscais efectivas, reflectindo a fracção de imposto evadida através da não declaração ou da declaração insuficiente à autoridade fiscal.

4. *Componente de não-registro:* refere-se à diferença entre o número de entidades ou indivíduos que deveriam estar registrados para fins fiscais (como empresas, trabalhadores autônomos ou proprietários de imóveis) e aqueles que estão efetivamente registrados. Também conhecido como Lacuna de Registro.

Por último, do ponto de vista da cobrança, algumas autoridades fiscais definem a lacuna fiscal em duas categorias: a lacuna fiscal bruta e a lacuna fiscal líquido.[11] Por exemplo, o Serviço de Receita Federal do Governo dos Estados Unidos (IRS) define a lacuna fiscal bruta, como a discrepância entre o total das obrigações fiscais efectivas impostas por lei para um determinado ano fiscal e o montante do imposto que os contribuintes pagam voluntária e atempadamente para esse ano. Por outro lado, a lacuna fiscal líquida refere-se ao montante remanescente da obrigação fiscal total, após a contabilização de todos os pagamentos efectuados através de acções de execução e de pagamentos voluntários em atraso para um determinado ano fiscal (Plumley 2005). A Figura 1 destaca as principais componentes da lacuna fiscal global e a sobreposição entre as diferentes definições das suas componente.

Figura 1: Conceitos de Lacuna Fiscal



Nota: ilustração simplificada dos conceitos de lacuna fiscal.

Fonte: ilustração dos autores.

3 Metodologias de Lacuna Fiscal

Existem duas abordagens gerais para estimar a lacuna fiscal - a abordagem descendente e a abordagem ascendente. A abordagem descendente utiliza dados a nível agregado, tais como indicadores macroeconómicos ou dados das contas nacionais, para avaliar de forma abrangente todas as perdas fiscais, medindo a lacuna como a diferença entre as receitas potenciais estimadas e

[1] É importante notar que as definições de lacuna fiscal bruta e líquida podem apresentar pequenas diferenças entre as várias autoridades fiscais, reflectindo os contextos únicos de aplicação da lei fiscal e as prioridades administrativas de cada país.

as receitas efectivas. No entanto, não pode determinar as origens da lacuna fiscal nem explicar por que razão certas áreas ou actividades continuam a não ser tributadas. Em contrapartida, a abordagem ascendente utiliza dados microeconómicos das administrações fiscais, incluindo os resultados de auditorias aleatórias ou operacionais que visam critérios específicos ou outros dados administrativos gerais das autoridades fiscais, para avaliar o grau de não-conformidade de determinados segmentos do sistema fiscal, grupos específicos de contribuintes ou tipos de não-conformidade (Hutton 2017).

3.1 Vantagens e desvantagens da abordagem ascendente

Vantagens

A abordagem ascendente na estimativa das lacunas fiscais oferece várias vantagens em relação a outras metodologias, em particular a sua capacidade de fornecer informações pormenorizadas (estimativas granulares) com base em auditorias fiscais. Eis as principais vantagens:

- *Maior precisão através de dados pormenorizados:* O método ascendente utiliza dados granulares de auditorias financeiras, permitindo estimativas mais precisas da lacuna fiscal. Esta técnica contrasta com as estratégias de cima para baixo, que dependem de indicadores económicos gerais e podem ignorar subtilezas nas acções de contribuintes individuais ou de indústrias específicas.

- *Informações detalhadas para acções precisas:* As tácticas ascendentes permitem uma compreensão detalhada da conformidade fiscal a nível individual ou empresarial. Este nível de pormenor permite que as autoridades fiscais elaborem intervenções precisas para determinados sectores, categorias de contribuintes ou casos de não-conformidade, aumentando a eficiência e o impacto das medidas de execução (Hutton 2017).

- *Abordar o viés de selecção na estimativa das lacunas fiscais:* O viés de selecção constitui um grande obstáculo à estimativa exacta da lacuna fiscal devido à natureza não representativa dos contribuintes escolhidos para as auditorias fiscais. A utilização da abordagem ascendente, especialmente quando integrada em técnicas de aprendizagem automática, pode atenuar eficazmente este enviesamento. Este método não depende de pressupostos relativos à distribuição dos dados, proporcionando assim robustez contra quaisquer enviesamentos que possam distorcer a estimativa da lacuna fiscal (Alaimo Di Loro et al. 2023).

- *Análise sectorial:* A abordagem ascendente permite uma análise sectorial detalhada das lacunas na conformidade das obrigações fiscais. Esta análise detalhada permite que as autoridades fiscais orientem as suas estratégias de conformidadede forma mais precisa, concentrando-se nos sectores com as lacunas mais significativas. Esta abordagem orientada poderia conduzir a uma maior eficiência na cobrança de impostos sem a necessidade de um aumento generalizado das actividades de auditoria ou de aplicação da leis (Barra et al. 2023; Hutton 2017).

- *Adaptabilidade a diferentes tipos de impostos:* A flexibilidade da abordagem ascendente permite-lhe ser adaptada para estimar as lacunas em vários tipos de impostos, incluindo o imposto sobre o valor acrescentado (IVA), o imposto sobre o rendimento das pessoas colectivas (IRPC) e o imposto sobre o rendimento das pessoas singulares (IRPS). Esta adaptabilidade é crucial, uma vez que diferentes impostos enfrentam diferentes tipos de desafios de conformidadee tácticas de evasão fiscal.

- *Conformidade fiscal melhorado:* Uma abordagem ascendente pode fornecer informações sobre o comportamento dos contribuintes, permitindo a verificação ou o aperfeiçoamento dos modelos existentes para identificar e gerir o risco. Ajudam também a identificar erros específicos que poderiam ser geridos de forma mais eficaz, através de abordagens alternativas, como o reforço da educação dos contribuintes, a melhoria dos serviços ou a realização de novas auditorias e reavaliações (Barra et al. 2023).

- *Permitir limites superiores e inferiores nas estimativas:* A abordagem ascendente permite a aplicação de várias técnicas à mesma unidade de contribuinte, para além de permitir a análise de sensibilidade estatística dos resultados (Barra et al. 2023).

Desvantagens

Apesar dos pontos fortes da abordagem ascendente, a literatura indica que esta tem as seguintes limitações (Warren 2018; Fiscalis Tax Gap Project Group 2018).

- *Endogeneidade:* Este método baseia-se fortemente nos conhecimentos e nos dados existentes na administração fiscal, o que o torna menos eficaz na captação de factores desconhecidos ou questões não observadas.

- *Desafios na contabilização de factores desconhecidos:* Uma vez que o método se baseia em dados conhecidos e em resultados operacionais, existem dificuldades em ter em conta os factores que não são facilmente observados, como os rendimentos não declarados. Também não abrange a economia completamente oculta, uma vez que apenas os contribuintes registados são normalmente seleccionados para auditorias. Consequentemente, as estimativas para estas incógnitas envolvem frequentemente ajustamentos grosseiros, o que pode reduzir a exactidão.

- *Âmbito restrito:* Esta abordagem vai do específico para o geral, centrando-se nos contribuintes individuais. Embora forneça informações detalhadas, pode ignorar tendências ou padrões macroeconómicos.

- *Risco de agregação:* As abordagens ascendentes apenas estimam componentes da lacuna fiscal, exigindo uma agregação para estimar o desvio total. No entanto, este processo acarreta o risco de dupla contagem e de sobrestimação do desvio total, especialmente quando existem sobreposições entre diferentes componentes do mesmo.

3.2 Tipo de Auditoria

As autoridades fiscais baseiam-se geralmente em informações de auditoria para prever a evasão fiscal e estimar as lacunas fiscais. Estas auditorias podem ser classificadas em dois tipos principais: auditorias aleatórias e operacionais. Ambos os tipos servem objectivos diferentes e têm metodologias únicas que permitem obter informações sobre a conformidadedas obrigações por parte dos contribuintes.

Auditorias aleatórias

As iniciativas de auditoria aleatória envolvem a selecção de amostras de contribuintes, através de um processo aleatório, com o objetivo de reflectir com precisão a população mais vasta que pretendem representar. Ao realizar estas auditorias, todos os contribuintes seleccionados são submetidos a um exame minucioso para identificar quaisquer discrepâncias entre o que declararam nos seus impostos e o que são legalmente obrigados a declarar. Os resultados destas auditorias

fornecem uma medida fiável do nível geral de conformidade no grupo da amostra. Para extrapolar os resultados da amostra para a população total, temos de garantir que o processo de selecção é completamente aleatório e não envolve critérios de selecção (Barra et al. 2023).

As auditorias aleatórias, embora exaustivas, têm desvantagens, segundo Feinstein (1999), incluindo custos elevados, tanto para as repartições de finanças, como para os contribuintes, especialmente os que cumprem com a legislação fiscal. Há também um desfasamento entre o período abrangido pelos dados e o momento em que os resultados estão disponíveis. Os resultados financeiros são geralmente inferiores aos das auditorias específicas, uma vez que examinam, tanto os contribuintes cumpridores, como os não cumpridores, ao contrário das auditorias específicas, que se concentram nos contribuintes mais susceptíveis de fugir aos impostos. Além disso, não podem detectar contribuintes não registados, o que leva a uma sub-estimação de algumas lacunas fiscais.

Por último, as autoridades fiscais podem ter relutância em efectuar auditorias aleatórias por razões relacionadas com a imagem pública da autoridade perante os contribuintes. As auditorias aleatórias podem ser entendidas como um controlo excessivo ou injusto por parte dos contribuintes cumpridores, o que leva a um sentimento público negativo e à diminuição da confiança na autoridade fiscal.

Auditorias operacionais

As auditorias operacionais baseiam-se na avaliação dos riscos e visam contribuintes específicos seleccionados de acordo com critérios definidos pela análise de risco das autoridades fiscais. Estas auditorias podem incidir sobre um ou vários tipos de impostos e, para cada tipo de imposto, podem abranger todo o seu âmbito ou apenas um segmento específico. Consequentemente, este tipo de auditoria pode não ser representativo de toda a população devido aos critérios de selecção, uma vez que nem todos os contribuintes têm a possibilidade de serem seleccionados como numa auditoria aleatória. Por conseguinte, as administrações fiscais efectuam uma estimativa ascendente das diferenças utilizando dados de auditoria não aleatórios com a ajuda de técnicas destinadas a inferir as características da população, em geral, a partir da amostra não representativa.

3.2 Procedimentos de estimativa ascendente

Podem ser utilizados vários procedimentos para efectuar uma abordagem ascendente. Todos eles utilizam informações de auditoria para prever o comportamento de empresas ou períodos não auditados. Nesta secção, passamos em revista as estimativas mais comuns e destacamos as suas principais caraterísticas (prós e contras).

Técnicas de regressão

As técnicas de regressão são consideradas comuns na literatura ascendente e podem ajustar o viés de selecção. Podem também ajudar a determinar as características que podem prever se um contribuinte será cumpridor e estimar o grau de não-conformidade. Estas técnicas de regressão incluem a abordagem de Heckman e a abordagem de correspondência da pontuação de propensão[2].

Abordagem de Heckman. A abordagem de Heckman aborda o viés de selecção, que ocorre durante o processo de auditoria operacional, levando à endogeneidade no sub-conjunto de contribuintes auditados. Este método, baseado no trabalho de Heckman (1979), envolve um

[2] "Propensity Score Matching", em Inglês.

processo de estimação em duas fases. A primeira fase identifica a probabilidade de uma observação ser incluída na amostra, calculando essencialmente a probabilidade de um contribuinte ser seleccionado para uma auditoria, utilizando uma equação de regressão probit. A segunda fase centra-se na estimativa da variável de interesse, que neste caso é o montante recuperado da auditoria. Para tal, são consideradas variáveis explicativas e um regressor específico que ajusta o viés de selecção. Este regressor específico, conhecido como o rácio de Mills inverso, é derivado dos parâmetros estimados na equação de selecção. A equação de resultados é então calculada utilizando a regressão por mínimos quadrados ordinários (OLS), incorporando um factor da equação da primeira fase.

O Grupo de Projecto de Lacunas Fiscais da União Europeia FISCALIS (2018) salienta que devem ser tomadas em consideração dois aspectos importantes ao estimar a lacuna fiscal, utilizando o método de Heckman. Em primeiro lugar, a equação de selecção tem de ser forte para explicar os resultados, uma vez que o método se baseia fortemente na capacidade da equação para prever a não-conformidade. Em segundo lugar, a equação deve incluir, pelo menos, uma variável que influencie a selecção para auditoria, mas que não tenha impacto na não-conformidade propriamente dito. Isto ajuda a evitar problemas de estimativas inexactas devido à multicolinearidade. Essencialmente, para uma estimativa precisa da lacuna fiscal, é necessário dispor de dados sobre os factores que levam à realização de auditorias, que não estão directamente relacionados com o nível de não-conformidade e, na prática, esta restrição de exclusão é difícil de satisfazer.

Abordagem de correspondência da pontuação de propensão O método de correspondência da pontuação de propensão é utilizado para corrigir o viés de selecção com base em ponderações dos dados. Este método começa por calcular uma "pontuação de propensão" para cada entidade, utilizando modelos estatísticos para prever a sua probabilidade de não estar em conformidade ou de ser objecto de auditoria. Um modelo de seleção binária calcula as propensões utilizando probit ou logit. Uma vez estimadas estas pontuações, o método emparelha as entidades que foram auditadas com as que não o foram, mas partilham pontuações de propensão semelhantes. A abordagem utilizada para fazer corresponder as observações pode ser a do vizinho mais próximo, a do calibre, a do núcleo ou a da linearidade local. Após o emparelhamento, o passo final consiste em atribuir um valor às devoluções não auditadas. Este valor, designado por N, é um valor imputado ou estimado do que a declaração não auditada teria comunicado, se tivesse sido auditada. A imputação baseia-se nos valores reais observados nas declarações auditadas correspondentes. Esta etapa é necessária para estimar qual teria sido a conformidade fiscal do grupo não auditado, se este tivesse sido objeto de uma auditoria.

Abordagem de agregação

Esta abordagem categoriza os contribuintes auditados e não auditados em grupos com base em variáveis significativas utilizadas para seleccionar a empresa para auditoria, tais como a dimensão da empresa, a região geográfica e o sector industrial. Permite o cálculo da lacuna fiscal global, através da soma dos desvios estimados para cada agrupamento. Estas estimativas são obtidas, através da aplicação de um factor de escala aos resultados da auditoria dos contribuintes auditados, projectando assim estes resultados para a população mais vasta dentro de cada agrupamento. Embora simples de aplicar e fácil de implementar, este método corrige apenas parcialmente o viés de selecção, resultando em conclusões que não são totalmente fiáveis.

Abordagem de valores extremos.

A abordagem de valores extremos aproveita o enviesamento de seleção na auditoria operacional para os contribuintes com níveis mais elevados de não-conformidade. Trata do comportamento

dos valores extremos (máximos ou mínimos) num conjunto de dados, em vez dos valores médios, sugerindo que, independentemente da distribuição geral dos dados, os valores extremos seguem frequentemente uma distribuição de Pareto generalizada. Isto significa que é possível obter informações sobre a taxa global de não-conformidade fiscal entre as grandes empresas a partir de um número limitado de casos extremos (nomeadamente, os evasores fiscais mais significativos). Esta abordagem é aplicável quando os dados apresentam caraterísticas da distribuição de Pareto - uma forma de distribuição de lei de potência que indica que uma pequena fracção de casos contribui de forma desproporcionada para o valor total observado nos dados, como acontece quando a sub-declaração de impostos está fortemente enviesada (com algumas grandes empresas a representarem a maior parte da lacuna) (Bloomquist et al. 2014).

Abordagens de aprendizagem automática

A aplicação de abordagens de aprendizagem automática (ML) aos estudos económicos, embora seja bastante recente, está a registar um aumento gradual, em especial na investigação relacionada com a fiscalidade, como a evasão fiscal, a fraude e a previsão da conformidade das obrigações fiscais, bem como a melhoria da auditoria fiscal e da estimativa das lacunas fiscais. Embora a investigação neste domínio se baseie geralmente em métodos tradicionais para fazer previsões, estes métodos sofrem limitações relacionadas com a dependência de métodos de regressão linear e com os pressupostos de distribuição rigorosos que têm. Na realidade, os dados apresentam frequentemente padrões mais complexos, o que faz com que estes métodos não sejam suficientemente flexíveis para efectuar previsões. Por conseguinte, alguns estudos começaram a adoptar métodos de aprendizagem automática para melhorar os resultados da previsão.

A título de exemplo da utilização da aprendizagem automática, Pérez López et al. (2019) foram utilizados modelos de redes neuronais de perceptron multicamadas (MLP) para prever a fraude fiscal, utilizando dados abrangentes das declarações do imposto sobre o rendimento das pessoas singulares (IRPS) em Espanha. Este método de aprendizagem automática (ML) foi capaz de prever a probabilidade de fraude fiscal e a probabilidade de envolvimento em práticas relacionadas com a fraude para cada contribuinte. Zumaya et al. (2021) utilizaram dois algoritmos de aprendizagem automática (ML), incluindo redes neurais artificiais profundas (ANNs) e floresta aleatória (RF), além de uma análise de rede complexa para prever a evasão ao imposto sobre o valor acrescentado (IVA) no México, analisando os dados transaccionais e as redes de interacção dos contribuintes. O documento concluiu que a combinação destes três métodos permitiu a identificação de novos potenciais suspeitos através da aprendizagem de padrões de evasores conhecidos. Ioana-Florina e Mare (2021) tentou prever a propensão dos contribuintes para a evasão fiscal com base na sua confiança no sistema fiscal, utilizando um modelo de rede neural de perceptron multicamada (MLP). Esta abordagem demonstrou um desempenho de previsão superior, ultrapassando o do modelo de regressão logística binária.[3]

Por outro lado, os métodos de aprendizagem automática são também utilizados para melhorar os esforços de auditoria fiscal. Por exemplo, Howard et al. (2020) avaliou o potencial das técnicas de aprendizagem automática para melhorar o processo de seleção de casos de auditoria por correspondência pelo Serviço de Receita Federal do Governo dos Estados Unidos (IRS). O estudo descobriu que, para algumas categorias de auditoria, os métodos de aprendizagem automática (ML) superam as abordagens tradicionais na classificação e seleção de declarações fiscais para auditorias

[3] Vide também Alsadhan (2023); Baghdasaryan et al. (2022); Holtzblatt e Engler (2022); Murorunkwere et al. (2022, 2023); Raikov (2021); Savic´ et al. (2022) para outros exemplos de utilização de métodos de aprendizagem automática na previsão de comportamentos de fraude e de evasão fiscais.

por correspondência. Estes métodos não só produzem receitas mais elevadas, como também reduzem o rácio de não alteração, o que significa que menos auditorias não resultam em ajustamentos, em comparação com outros métodos. Do mesmo modo, Battaglini et al. (2022) utilizou dados fiscais administrativos italianos para explorar o potencial das técnicas de aprendizagem automática, como a floresta aleatória, para melhorar a detecção e a recuperação da evasão fiscal, melhorando o processo de selecção dos contribuintes para auditoria. O documento indica que, em alguns cenários, a aprendizagem automática (ML) pode melhorar a previsão da detecção de evasão até 83 por cento e recuperar até 65 por cento da evasão detectada.

A pesquisa sobre a estimativa das lacunas fiscais não esteve longe destes novos desenvolvimentos. Dadas as limitações das abordagens de estimativa de lacunas fiscais anteriormente mencionadas, que se baseiam em métodos de regressão tradicionais para fazer previsões, alguns investigadores e autoridades fiscais começaram a incorporar a utilização de técnicas semi-paramétricas nos métodos tradicionais e começaram a utilizar a aprendizagem automática para melhorar os resultados das previsões. Embora a aprendizagem automática seja superior nas tarefas de previsão em comparação com as abordagens tradicionais, é também eficaz para resolver o problema do viés de selecção nas estimativas do diferencial de tributação que se baseiam em auditorias operacionais.

Para abordar a questão do viés de selecção no contexto das estimativas das diferenças fiscais, é fundamental distinguir entre os dois principais tipos de viés de selecção: o viés de selecção causal e o viés de selecção por amostragem. O viés de selecção causal afecta a estimativa de parâmetros não enviesados na análise causal, como quando os grupos tratados e de controlo não são atribuídos aleatoriamente, levando a estimativas enviesadas dos efeitos do tratamento. No entanto, o nosso foco é o viés de selecção da amostra, que ocorre quando a amostra de treino utilizada para construir um modelo de previsão difere da amostra de previsão. No caso de estimativas de desvios fiscais baseadas em auditorias operacionais, este enviesamento surge porque a amostra de treino consiste em contribuintes auditados seleccionados com base em alguns critérios conhecidos das autoridades fiscais e não representativos de toda a população de contribuintes, enquanto a amostra de previsão inclui contribuintes não auditados. Esta discrepância pode dar origem a previsões incorrectas se não for devidamente corrigida.

Um aspecto crucial do tratamento do viés de selecção da amostra é a distinção entre enviesamentos decorrentes de factores observáveis e não observáveis. O viés de selecção observável ocorre quando o processo de selecção, tal como a decisão de auditar, se baseia em variáveis conhecidas e mensuráveis. Nesses casos, se a probabilidade de ser auditado puder ser estimada com exactidão, utilizando essas co-variáveis observáveis, o enviesamento pode ser corrigido incluindo essas co-variáveis no modelo de aprendizagem automática. Esta metodologia corresponde às estratégias delineadas por Brewer e Carlson (2024), que defendem o controlo do viés de selecção, através do ajustamento dos factores observáveis. Calculando e integrando a probabilidade de selecção no modelo, é possível atenuar o viés de selecção, partindo do princípio de que as decisões de auditoria são determinadas principalmente por dados observáveis.[4]

Em cenários em que o processo de selecção é regido por factores não observáveis que não são captados no conjunto de dados, a complexidade do enviesamento aumenta. Os métodos tradicionais podem não ser suficientes para contrariar esta forma de enviesamento. Nesses casos,

[4] Presumivelmente, as autoridades fiscais dispõem de informações sobre a forma de decidir quem deve ser objecto de auditoria. Esta informação é normalmente reservada, mas pode ser utilizada no modelo de aprendizagem automática para prever com exactidão os resultados. O nosso conselho é não partilhar a relevância das co-variáveis na previsão, uma vez que esta informação está relacionada com o processo de auditoria. No entanto, esses resultados também podem ser utilizados para melhorar o processo de tomada de decisões de auditoria.

são necessárias técnicas mais avançadas, como a incorporação de uma função de controlo no modelo de aprendizagem automática (ML) baseado no método de Heckman, para resolver o viés de selecção baseado em factores não observáveis (Brewer e Carlson 2024) Na literatura recente, há exemplos notáveis de integração de abordagens de aprendizagem automática em métodos tradicionais, bem como estudos que estimam as lacunas fiscais, utilizando principalmente técnicas de aprendizagem automática.

Alaimo Di Loro et al. (2023) propôs um método baseado na aprendizagem automática que consiste em duas etapas do algoritmo de aumento do gradiente. Este método resolve o problema do enviesamento de seleção resultante da dependência de dados de auditoria não aleatórios e fornece previsões precisas. Em primeiro lugar, o método calcula as pontuações de propensão da probabilidade de um contribuinte ser objecto de uma auditoria, utilizando um modelo de classificação baseado no aumento de gradiente com árvores de classificação e regressão (CART), como aprendentes de base. Para o efeito, os dados são divididos em conjuntos de treino e de teste e, durante o processo de treino, são seleccionadas as co-variáveis importantes. Este passo conduzirá à previsão das probabilidades de cada empresa ser auditada com base nas suas co-variáveis.

Em segundo lugar, o método emprega um modelo de regressão, utilizando aumento de gradiente com árvores de classificação e regressão (CART), como aprendentes de base para prever a base fiscal potencial, incluindo o imposto sobre o valor acrescentado (IVA) não declarado, e, por conseguinte, os montantes evadidos para cada empresa. Nesta etapa, as pontuações de propensão obtidas anteriormente são utilizadas para criar ponderações para cada contribuinte, corrigindo qualquer sobre ou sub-representação na amostra auditada. A comparação desta abordagem de aprendizagem automática (ML) com o modelo tradicional de Heckman revela a superioridade da abordagem de aprendizagem automática (ML) na captação da variabilidade da base tributária potencial e na obtenção de previsões mais exactas para a estimativa da lacuna fiscal

Adu-Ababio et al. (2024) utilizou algoritmos de aprendizagem automática supervisionada com dados de declarações fiscais e de auditorias para estimar as lacunas fiscais na Zâmbia. O principal algoritmo de aprendizagem automática utilizado neste estudo foi a rede neural artificial (ANN) em duas fases. A primeira fase baseia-se apenas nos dados auditados para criar iteracções de múltiplas versões de conjuntos de dados de treino e teste de forma aleatória, com 90 por cento dos dados utilizados para treinar o modelo. Em cada iteracção, o algoritmo aprende com o conjunto de treino, analisando vários parâmetros relacionados com a fiscalidade. Em seguida, o algoritmo utiliza o que foi aprendido para prever as taxas de evasão fiscal utilizando dados de teste. Depois disso, o algoritmo compara as taxas de evasão fiscal reais e previstas e, se estas previsões não corresponderem exactamente às taxas reais, podem ser introduzidas algumas melhorias no modelo, repetindo-se este processo até se atingir um desempenho satisfatório. Na segunda fase, o modelo é implementado, utilizando a amostra completa, em que os dados auditados são utilizados no conjunto de treino e os dados não auditados constituem o conjunto de teste. Uma vez que o modelo aprendeu com as variáveis explicativas seleccionadas, o mesmo prevê a evasão fiscal a partir dos dados de teste e, em seguida, utiliza a evasão fiscal prevista e real para estimar as lacunas fiscais. Os autores também utilizaram outros algoritmos de aprendizagem automática, como a floresta aleatória, para verificar a estabilidade e a fiabilidade do método principal e os resultados foram ligeiramente próximos.

Seguindo a mesma linha, o estudo de Ebrahim et al. (2024) utilizou dados fiscais e de auditoria da administração para estimar o desvio do imposto sobre o valor acrescentado (IVA) na Tanzânia utilizando a aprendizagem automática, especificamente o algoritmo de floresta aleatória. Esta abordagem tinha como objectivo prever os montantes da evasão fiscal para as empresas não auditadas e auditadas nos períodos em que não foi realizada qualquer auditoria. Os autores

compararam o desempenho da abordagem de aprendizagem automática (ML) com a regressão por mínimos quadrados ordinários (OLS) tradicional. Verificaram uma redução significativa da raiz do erro quadrático médio (RMSE) e valores de R-quadrado mais elevados, quando se utilizou o algoritmo de floresta aleatória, indicando um desempenho de previsão mais exacto. Os resultados revelam uma diferença média do imposto sobre o valor acrescentado (IVA) de cerca de 62 por cento, com diferenças consideráveis entre os vários sectores económicos. O sector agrário, em particular, apresentou a maior diferença do imposto sobre o valor acrescentado (IVA), o que revela uma evasão fiscal significativa nesta área.

Outros avanços nas técnicas de aprendizagem automática (ML) envolvem a utilização de antigas abordagens de regressão. Chudý et al. (2020) aplicaram uma seleção de amostra semi-paramétrica do modelo de Heckman para estimar o desvio do imposto sobre o rendimento das pessoas colectivas (IRPC) na Eslováquia. Esta extensão do modelo de Heckman supera o modelo de Heckman tradicional, uma vez que permite um pressuposto de normalidade mais flexível e uma melhor modelação das estruturas de dados complexas e o tratamento das relações não lineares e da heterocedasticidade inerentes aos dados. Na primeira fase deste modelo, a equação de selecção foi estimada, utilizando um método não paramétrico, como o alisamento do núcleo, que permite aproximações flexíveis das distribuições. Em seguida, na segunda fase do modelo, a equação de resultados incorporou estas estimativas da primeira fase para proporcionar uma correcção mais robusta do viés de selecção e captar as relações mais complexas que um modelo de regressão linear poderia ignorar. O documento concluiu que esta abordagem teve um melhor desempenho em comparação com outras abordagens alternativas, como a correspondência da pontuação de propensão e a regressão linear por mínimos quadrados ordinários (OLS) ponderada, no que diz respeito ao viés de selecção e à obtenção de melhores previsões.

As autoridades fiscais também começaram a utilizar técnicas de aprendizagem automática (ML) para melhorar as suas estimativas de lacunas fiscais ou processos de auditoria. A Autoridade Tributária Italiana (n.d.) utilizou a aprendizagem automática juntamente com outros métodos tradicionais para estimar o diferencial do IVA, na chamada abordagem assistida por aprendizagem automática. O passo inicial desta abordagem visa resolver o viés de selecção que decorre da utilização de auditorias não aleatórias, recorrendo à regressão logística para dividir a população em grupos, tendo cada grupo uma probabilidade semelhante de ser auditado. De seguida, a população é estratificada em quintis com base nestas probabilidades, o que permite que os contribuintes auditados sejam representativos de toda a população em cada grupo. Na segunda etapa, a aprendizagem automática, especificamente as árvores de regressão por agregação de bootstrap ou bagging, é empregue para fazer previsões dentro de cada extracto. A última etapa tem como objectivo melhorar a precisão das previsões, utilizando o modelo de correspondência média preditiva (PMM) que utiliza as previsões iniciais para fazer corresponder cada contribuinte não auditado (designado por destinatário) a um contribuinte auditado (designado por dador) com base na semelhança dos seus valores previstos. Este processo garante que os valores imputados reflectem a verdadeira distribuição da variável-alvo, permitindo inferências precisas sobre várias características de distribuição para além das médias.

A Autoridade Tributária Canadiana (2019) utiliza uma técnica de aprendizagem automática não supervisionada para identificar agrupamentos numa população, semelhante à primeira etapa mencionada anteriormente para a Itália, em que os elementos de cada agrupamento são mais semelhantes entre si do que com os de outros agrupamentos. Este algoritmo de aprendizagem automática categoriza automaticamente as empresas em agrupamentos com base em caraterísticas específicas, assumindo que as empresas não auditadas em cada agrupamento partilham o mesmo rácio de não-conformidade em relação à receita bruta declarada que as empresas auditadas. Esta abordagem foi utilizada para fornecer uma estimativa de limite superior e foi combinada com a abordagem de valor extremo para fornecer uma estimativa de limite inferior da lacuna fiscal.

Resumo

A estimativa das lacunas fiscais, através de abordagens ascendentes, pode ser efectuada utilizando diferentes métodos de estimações. Porém, cada método pode ser mais adequado em função do contexto e dos dados utilizados. De um modo geral, a utilização de uma abordagem ascendente pode basear-se em dados de auditoria aleatórios ou baseados no risco. Muitos pesquisadores consideram que o recurso a dados de auditoria aleatórios é a forma ideal de efectuar uma estimativa ascendente da lacuna fiscal. No entanto, em muitos casos, as autoridades fiscais tendem a preferir fazer uma auditoria baseada no risco, o que introduz alguns desafios à estimativa, uma vez que os contribuintes seleccionados para auditoria podem diferir significativamente dos outros contribuintes, fazendo com que os resultados da auditoria não representem a população global em situação de não-conformidade. Neste caso, a estimativa por mínimos quadrados ordinários (OLS) tradicional pode não ser a escolha ideal para os pesquisadores devido ao viés de selecção do processo de auditoria. Por conseguinte, os pesquisadores estão a utilizar outros métodos para obter estimativas imparciais. A seguir, resumimos as principais ideias sobre os métodos mencionados nesta secção.

Embora a abordagem de Heckman em duas fases seja considerada um dos métodos mais utilizados para ter em conta o viés de selecção, por vezes a sua restrição de exclusão é difícil de satisfazer, o que pode levar a erros-padrão inflacionados devido à multicolinearidade, e tende a subestimar a lacuna fiscal, uma vez que a evasão fiscal e o não-conformidade não detectado são frequentemente ignorados. O método da pontuação de propensão ajuda a eliminar o viés de selecção, criando grupos de contribuintes cumpridores e não cumpridores com base em caraterísticas observáveis, o que permite uma atribuição mais precisa das diferenças nos resultados da conformidade das obrigações fiscais à não-conformidade e não a factores não observados. Algumas autoridades fiscais utilizam a abordagem de agrupamento para detectar comportamentos anómalos e revelar a sub-declaração de impostos em grupos específicos, estimando depois o diferencial de imposto para cada grupo, extrapolando os resultados da auditoria dos contribuintes auditados para toda a população desse grupo específico. Por outro lado, a abordagem do valor extremo é mais simples e mais económica em termos de tempo e de utilização de recursos do que as outras abordagens. Contudo, requer mais pressupostos, especialmente no que diz respeito à definição da cauda da distribuição de Pareto, na qual se baseia para a modelação.

Em contrapartida, as técnicas de estimação baseadas na aprendizagem automática oferecem vantagens significativas em relação aos métodos supramencionados, nomeadamente quando se trata de relações complexas e não lineares e de factores não observados que influenciam o viés de selecção. Os métodos de aprendizagem automática podem ser preferidos pela sua flexibilidade e desempenho preditivo superior.

4 O conjunto de ferramentas

Nesta secção, explicamos os componentes do conjunto de ferramentas. O objetivo do conjunto de ferramentas é estimar o diferencial de tributação do imposto sobre o valor acrescentado (IVA), do imposto sobre o rendimento das pessoas colectivas (IRPC) ou do imposto sobre o rendimento das pessoas singulares (IRPS). O conjunto de ferramentas tem dois elementos principais: limpeza de dados e estimativa. O processo de limpeza dos dados destina-se a garantir a harmonização e a coerência dos ficheiros de dados necessários para a estimativa ascendente. Além disso, ajuda a alinhar os requisitos gerais na estimativa de aprendizagem automática (ML). Isto é importante, uma vez que os dados provêm de diferentes fontes e períodos e a sua normalização simplifica o

processo de estimativa. A estimativa é baseada na metodologia de floresta aleatória, uma técnica de aprendizagem automática. Uma explicação técnica sobre isto é apresentada no Anexo A.

4.1 Limpeza de dados

O processo de limpeza dos dados pode ser dividido em três fases principais: as duas primeiras tratam dos dados das declarações administrativas (IVA, IRPC e IRPS) e dos dados de auditoria e a última demonstra a combinação destes ficheiros de dados para análise posterior. Esta etapa tem por objectivo processar diferentes fontes de dados, harmonizá-las e construir uma estrutura única que combine informações sobre os contribuintes, as declarações fiscais e os resultados das auditorias ou das avaliações.

Normalmente, a informação sobre as declarações fiscais (IVA, IRC ou IRS) está contida em ficheiros diferentes dos das auditorias, uma vez que estas últimas são realizadas depois de as empresas ou os indivíduos apresentarem as suas declarações. Contudo, as declarações de impostos podem conter pelo menos dois conjuntos de ficheiros para o mesmo contribuinte. Isto pode dever-se ao facto de o contribuinte ter actualizado a declaração num determinado momento, dentro ou fora do período de apresentação. Este é um problema comum de duplicação que surge nas bases de dados administrativas fiscais. Nestes casos, o mesmo elemento de informação é reproduzido para o mesmo contribuinte. Por outras palavras, para um contribuinte num determinado ano de declaração, existem duas ou mais réplicas da mesma informação. Um dos principais objectivos da secção de limpeza de dados é garantir que cada contribuinte é identificado de forma única pelos seus identificadores e pelo ano de declaração. Na primeira fase do conjunto de ferramentas, fornecemos cenários possíveis que criam esses erros de duplicação e demonstramos como o utilizador pode lidar com eles individualmente. É importante resolver quaisquer duplicações em todos os ficheiros de dados da declaração de impostos e de auditoria necessários, independentemente de se apresentarem em ficheiros únicos ou múltiplos. No caso de vários ficheiros, a abordagem consiste em tratar primeiro os duplicados e, em seguida, anexar os respectivos conjuntos de dados a um único ficheiro.

Nesta fase do conjunto de ferramentas, abordamos também os problemas observados nos dados de auditoria relativos aos períodos de auditoria e à forma como estes se relacionam com períodos de declaração específicos. Nalguns casos, os dados de auditoria são identificados pelo facto de o ano de avaliação ser o mesmo que o ano de apresentação da declaração. Por vezes, é sobretudo o ano de auditoria que duplica o ano de declaração. Seja qual for o caso, é necessário identificar o ano específico nos dados de auditoria que corresponde ao ano de declaração e anexá-los para obter um ficheiro único, se os dados estiverem em vários ficheiros. Isto garante que cada avaliação de auditoria está corretamente ligada a um período de retorno específico.

No final destas duas primeiras etapas, agregamos os dados das declarações fiscais ao nível anual. Como estimamos as lacunas fiscais anualmente, também garantimos que as avaliações de auditoria se relacionam com os resultados anuais, mesmo que as auditorias tenham sido efectuadas em vários anos de declaração. A agregação ocorre normalmente para o imposto sobre o valor acrescentado (IVA) e o imposto sobre o rendimento das pessoas singulares (IRPS), mas não para o imposto sobre o rendimento das pessoas colectivas (IRPC), uma vez que este é sempre comunicado anualmente. Este procedimento garante que temos uma declaração de imposto ou um resultado de auditoria (se o contribuinte for auditado) por contribuinte e por ano.

Finalmente, combinamos os ficheiros de dados necessários, tendo em conta que as variáveis das declarações fiscais e das auditorias estão em dois ficheiros separados. É importante compreender o processo de fusão, uma vez que mostra até que ponto limpámos e tratámos os duplicados em todos os ficheiros de dados. O objectivo é fundir a informação para a mesma unidade

(contribuinte) no mesmo período (ano-mês). Além disso, pretendemos que as informações fornecidas pelos dados de auditoria, tais como o resultado da auditoria para um determinado ano de declaração, sejam fundidas com o registo fiscal no período de apresentação da declaração correspondente. Por exemplo, fundimos o registo fiscal do ano de declaração de 2018 com o resultado da auditoria sobre as declarações fiscais incorrectas em 2018, se a empresa tiver sido auditada. Não haverá informações sobre o resultado da auditoria, se a empresa não tiver sido auditada. Este é um problema habitual que o utilizador terá de enfrentar, uma vez que as auditorias são realizadas retroactivamente a um número limitado de contribuintes com base em declarações anteriores. Explicamos como obter os resultados da auditoria para estas empresas não auditadas na fase seguinte do conjunto de ferramentas.

4.2 Estimativa da aprendizagem automática

A abordagem ascendente é seguida para estimar a lacuna fiscal. Esta abordagem requer, como ponto de entrada, o resultado da auditoria dos contribuintes, que representamos como a declaração incorrecta de impostos. Esta variável é obtida e encontrada nos dados de auditoria, após o processo de auditoria. No entanto, esta variável é visível para as empresas que foram objecto de auditoria. Este facto cria a necessidade de estimar ou prever resultados para contribuintes e períodos não auditados. As previsões sobre contribuintes e períodos não auditados são necessárias, porque a informação sobre declarações incorrectas das auditorias depende de momentos e unidades específicos. Assim, um contribuinte auditado no ano de retorno 3 não é auditado no ano de retorno 2, o que significa que temos de incluir uma previsão para períodos não auditados para garantir que temos toda a informação necessária. É necessário um procedimento de estimação para obter previsões exactas sobre declarações fiscais incorrectas.

No conjunto de ferramentas, seguimos o método de floresta aleatória para prever a declaração incorrecta de impostos em empresas e em períodos não auditados. Esta metodologia permite uma estimativa granular, captando melhor os potenciais valores atípicos ou anómalos do que a previsão linear. A floresta aleatória deve ser ajustada, através da escolha de dois parâmetros críticos: o número de iteracções (ou árvores) e o número de utilizações a prever em cada divisão. Para o efeito, é necessário utilizar dados que contenham a variável a prever em caso de declarações fiscais incorrectas. Assim, em primeiro lugar, o conjunto de dados é dividido em dados auditados e não auditados. Os primeiros serão utilizados para afinar o modelo e os segundos serão utilizados para efectuar previsões.

A divisão dos dados de auditoria em amostras de treino e de teste é necessária no processo de afinação. O objectivo é melhorar a precisão da estimativa, uma vez que a metodologia utiliza os dados de treino para aprender sobre as variáveis e, posteriormente, contrasta a previsão com o valor real nos dados de teste. Ao efectuar este processo, obtêm-se os dois parâmetros críticos. Para além disso, os parâmetros garantem que o erro de previsão, ou seja, a diferença entre a previsão e o valor real, é o mínimo possível. Com o modelo ideal, o conjunto de ferramentas efectua um contraste com um modelo de regressão. Isto serve para mostrar a exactidão da previsão e é útil para validar o modelo de previsão.

Por fim, obtém-se a lacuna fiscal. Em primeiro lugar, o modelo é executado apenas nos dados auditados, uma vez que essas observações contêm informações incorrectas. Nesta etapa, o modelo estima o índice (ou peso) que cada variável auxiliar (ou co-variante) deve ter. Posteriormente, o modelo prevê os dados não auditados, utilizando o índice ideal, e obtêm-se as previsões de declarações incorrectas. A lacuna fiscal é obtida somando a variável de declaração incorrecta (prevista ou descoberta pela auditoria) com a declaração de imposto, obtendo-se o imposto potencial. A lacuna fiscal é a taxa entre as declarações incorrectas e o imposto potencial, indicando a percentagem do imposto potencial não cobrado devido a declarações incorrectas. Esta variável

é obtida pelo tempo do grupo (como o sector da indústria), mostrando a granularidade da metodologia.

5 Considerações finais

Neste conjunto de ferramentas ascendentes para a estimativa das lacunas fiscais, o nosso objectivo foi desenvolver um quadro prático para estimar as lacunas fiscais no imposto sobre o valor acrescentado (IVA), no imposto sobre o rendimento das pessoas colectivas (IRPC) e no imposto sobre o rendimento das pessoas singulares (IRPS), utilizando uma metodologia ascendente. O conjunto de ferramentas foi concebido para que as autoridades fiscais e os decisores políticos possam estimar a diferença entre as receitas fiscais efectivamente cobradas e as receitas potenciais que poderiam ter sido cobradas, se a regulamentação fiscal fosse integralmente cumprida. Fornece uma definição padronizada aplicável aos países em desenvolvimento, tendo em conta o seu contexto e os recursos disponíveis. A nossa abordagem baseia-se na aplicação de um algoritmo de aprendizagem automática que utiliza declarações administrativas de micro-fiscalidade e dados de auditoria para prever declarações fiscais incorrectas e não-conformidade e, em seguida, estimar as lacunas fiscais, tanto a nível agregado, como por regiões ou sectores específicos.

Nesta nota técnica, passamos em revista as definições de lacuna fiscal para compreender as suas componentes, uma vez que o objectivo do nosso método é estimar os lacuna de sub-declaração, declaração incorrecta e não-conformidade fiscal entre os contribuintes registados. Em seguida, passamos em revista os procedimentos tradicionalmente utilizados, salientando as vantagens da utilização de estimativas de aprendizagem automática, utilizando uma abordagem ascendente em relação a outras estimativas alternativas.

O conjunto de ferramentas pode ser dividido em duas fases principais: gestão de dados e análise de aprendizagem automática. Na fase de gestão de dados, os conjuntos de dados fiscais e de auditoria são preparados para um procedimento de análise, através da limpeza, do tratamento de duplicações e da fusão para garantir a harmonização dos dados fiscais e de auditoria, de modo a permitir uma passagem sem problemas para as fases de aprendizagem automática. A aprendizagem automática prevê declarações fiscais incorrectas para contribuintes ou períodos não abrangidos por auditorias, através da aplicação de algoritmos de floresta aleatória. Estes modelos têm a capacidade de fornecer uma estimativa correcta, uma vez que são treinados com base em dados auditados e podem estimar com precisão a evasão fiscal em casos não auditados, permitindo assim estimar a lacuna fiscal de uma forma abrangente. Em comparação com os modelos de regressão tradicionais, o conjunto de ferramentas também comparou o desempenho dos modelos de aprendizagem automática para salientar a melhoria do poder de previsão.

Por último, algumas sugestões para trabalhos futuros envolvem a expansão e a melhoria do actual conjunto de ferramentas das seguintes formas. Seria possível utilizar outros algoritmos de aprendizagem automática, como as redes neuronais, e comparar a exactidão da previsão entre os diferentes métodos. A generalização do conjunto de ferramentas para outras linguagens de programação para além do STATA, alargando assim o seu alcance, é outra área a considerar. É necessário continuar a trabalhar na forma como o conjunto de ferramentas pode ser implementado em diferentes contextos nacionais. Por último, o conjunto de ferramentas pode, ele próprio, constituir um ponto de partida para futuras pesquisas sobre o comportamento dos contribuintes, com o objectivo de ajudar as autoridades a conceber e a aplicar melhores estratégias de controlo e medidas de conformidade.

Referências

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). Disponível em: https://ssrn.com/abstract=3200838 (última revisão: 23 de Junho de 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

A ANEXO - Algoritmo de floresta aleatória

A floresta aleatória é considerada um dos algoritmos de aprendizagem automática em conjunto mais utilizados e com melhor desempenho para tarefas de previsão (Athey e Imbens 2019).[5] Ao contrário dos modelos de regressão tradicionais, que assumem a linearidade e têm dificuldades quando o número de observações é inferior ao das variáveis independentes, a floresta aleatória pode lidar com relações não lineares nos dados e evita o problema de estimar mais parâmetros do que o que os pontos de dados podem suportar. Além disso, capta melhor a existência de valores atípicos e anómalos, produzindo previsões mais precisas nesses casos (Athey e Imbens 2019). Consegue-o não utilizando todas as variáveis de previsão ao mesmo tempo, o que resulta em melhores previsões do que a regressão tradicional (Schonlau e Zou 2020). Para além de ser simples de utilizar, a floresta aleatória é fácil de compreender e rápida de implementar. Além disso, tem um bom desempenho quando comparado com outros algoritmos de aprendizagem automática (Varian 2014).

Essencialmente, uma floresta aleatória pode permitir-nos prever a variável alvo (y), utilizando variáveis de entrada (x). Trata-se essencialmente de uma colecção de árvores de decisão criadas, utilizando subconjuntos aleatórios de dados. Mas o que são árvores de decisão e como são usadas para criar um modelo de floresta aleatória? Para responder a esta questão, o documento começa por explicar os conceitos das árvores de decisão e como funcionam e, em seguida, explica como podemos criar um modelo de floresta aleatória e utilizá-lo para realizar tarefas de previsão.

A1 Árvores de decisão

As árvores de decisão são um tipo de algoritmo de aprendizagem supervisionada, utilizado para tarefas de regressão e de classificação. Funcionam dividindo os dados em subconjuntos, com base nos valores das características das variáveis de entrada (x) para prever valores (y). Este processo de divisão continua até que os dados dentro de cada subconjunto sejam tão homogéneos quanto possível em relação à variável-alvo. É também conhecido como algoritmo de árvores de classificação e regressão (CART), que é uma forma de encontrar a melhor divisão em cada passo para maximizar a precisão da previsão.

Algoritmo CART

Tipos de CART:

- As árvores de classificação são um tipo de algoritmo de árvore de decisão utilizado para classificar variáveis-alvo categóricas. Funcionam segmentando o espaço do preditor em regiões distintas, com cada região correspondendo a um rótulo de classe específico. O objectivo é determinar a que categoria pertence a variável-alvo com base nas características de entrada.

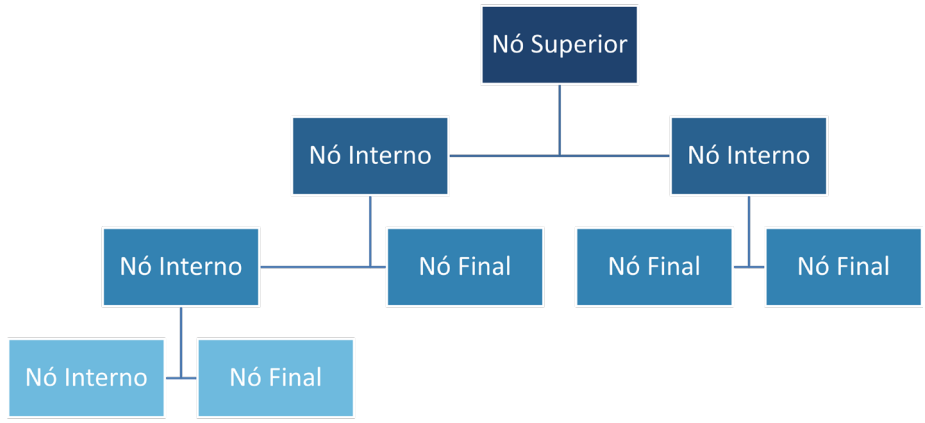
- As árvores de classificação são um tipo de algoritmo de árvore de decisão concebido para prever variáveis-alvo contínuas. Dividem o espaço de previsão em regiões e fornecem um valor contínuo, como resultado para cada região.

[5] Um método de conjunto combina vários modelos simples, conhecidos como aprendentes fracos, para criar um modelo preditivo único e mais forte.

Como é que o algoritmo CART funciona?

A estrutura de construção de uma árvore de decisão, utilizando o CART, começa com o nó superior, que representa todo o conjunto de dados. Esse nó superior é o ponto de partida da árvore. A partir desse ponto, o algoritmo identifica o melhor atributo para dividir o conjunto de dados e rotula o nó com esse atributo. Isto cria ramos que conduzem a nós internos, em que cada nó interno representa uma decisão baseada no valor do atributo escolhido. Os dados são ainda divididos em cada nó interno, continuando a gerar mais ramos e nós. Este processo repete-se, criando uma estrutura hierárquica. Os pontos finais destes ramos são os nós terminais que fornecem a previsão final, como se mostra na Figura A1. Nas tarefas de classificação, a previsão num nó terminal é a classe maioritária das observações nesse nó e, nas tarefas de regressão, é o valor médio das observações.

Figura A1: Estrutura da árvore de decisão



Fonte: ilustração dos autores.

Nas tarefas de regressão, o CART utiliza a redução residual, como critério de divisão. Isto envolve a partição dos dados em cada nó para minimizar a diferença média ao quadrado entre os valores previstos e reais, com o objetivo de obter o menor erro residual. Para tarefas de classificação, o CART utiliza a impureza de Gini para avaliar todas as divisões potenciais, escolhendo a que reduz mais eficazmente a impureza e, assim, aumenta a pureza dos subconjuntos resultantes. A impureza de Gini quantifica a probabilidade de classificar incorrectamente uma instância aleatória com base na classe maioritária dentro de um subconjunto. Este processo de divisão é recursivo, continuando até que determinados critérios de paragem sejam cumpridos. Estes critérios incluem chegar a um nó em que todos os registos partilham o mesmo valor alvo, o tamanho do nó ser inferior a um limiar definido pelo utilizador, a árvore atingir a sua profundidade máxima predefinida, ter menos do que um número mínimo de casos num nó ou quando a divisão adicional não aumenta significativamente a pureza (Zacharis 2018).

Um risco essencial na utilização de árvores de decisão é o sobre-ajuste do modelo. Isto pode acontecer se o modelo crescer sem restrições, por exemplo, quando uma árvore de regressão continua a dividir-se até que cada nó terminal contenha apenas uma única observação. Embora isso possa resultar em um ajuste quase perfeito aos dados de treinamento (vide Nota abaixo para obter uma definição), isso afecta negativamente a capacidade do modelo de generalizar para dados novos e não vistos. Os modelos sobre-ajustados têm normalmente um bom desempenho nos dados de treino, mas um fraco desempenho nos dados de validação ou de teste, porque aprenderam o ruído em vez do sinal

Para resolver os problemas de sobre-ajuste, o CART utiliza uma técnica de poda, quando a árvore está totalmente desenvolvida. A poda envolve o corte da árvore para eliminar os nós que adicionam valor preditivo mínimo, simplificando assim o modelo e melhorando sua generalização. Uma técnica amplamente usada é a poda de complexidade de custo, em que uma árvore grande é inicialmente cultivada, usando um parâmetro de complexidade muito pequeno, para garantir que todas as divisões potenciais sejam avaliadas. Em seguida, as divisões são removidas sequencialmente e o desempenho do modelo é reavaliado, usando validação cruzada. Este processo continua até que a poda adicional não melhore o ajuste do modelo (Békés e Kézdi 2021).

A Figura A2 apresenta um exemplo de pseudocódigo para um algoritmo de crescimento de árvore que explica as tarefas de classificação e de regressão. Consideremos um cenário em que o objectivo é construir uma árvore de decisão para prever uma variável-alvo utilizando um conjunto de dados X , que contém múltiplas co-variáveis A e a variável-alvo y. O parâmetro de tarefa indica se estamos a lidar com classificação ou regressão.

O algoritmo começa por inicializar uma única árvore T com um nó superior. Se todos os critérios de paragem tiverem sido cumpridos, o algoritmo procede à etiquetagem do nó. Para tarefas de classificação, o nó é rotulado com a classe mais comum entre as amostras em X. Para tarefas de regressão, o nó é rotulado com o valor médio de y.

Se os critérios de paragem não tiverem sido cumpridos, o algoritmo procura o melhor atributo $a \in A$ que divide o conjunto de dados X de forma mais eficaz. As tarefas de classificação são feitas, usando uma função de impureza, como a impureza de Gini. Para tarefas de regressão, o algoritmo tem como objectivo minimizar a variação dentro dos nós. O nó é então rotulado com o atributo a.

Figura A2: Pseudocódigo do algoritmo de crescimento da árvore para as tarefas de classificação e regressão

Algoritmo 1 Algoritmo de crescimento da árvore `crescimento da árvore`(*X , A, y,* tarefa)

Requer: Conjunto de dados de treino *X*, conjunto de atributos *A*, variável de saída *y*, tarefa (classificação ou regressão)

Garantir: Árvore de decisão

1: Começar uma única árvore *T* com um nó superior

2: se todos os critérios de paragem forem satisfeitos, então

3: se a tarefa == classificação, então

4: *T* tem um nó com a classe mais comum em *X* como etiqueta

5: além disso

6: *T* tem um nó com a média de *y* em *X* como etiqueta

7: fim, se

8: além disso

9: encontra $a \in A$, que melhor divide *X* usando a função de impureza (para classificação) ou minimizando a variância (para regressão)

10: Etiquetar nó com *a*

11: para o possível valor *v* de *a* fazer

12: X_v = o subconjunto de *X* que tem $a = v$

13: $A_v = A - a$

14: `crescimento da árvore`(X_v, A_v, *y,* tarefa)

15: ligar o novo nó ao nó de topo com a etiqueta *v*

16: fim para

17: fim, se

18: devolver `árvore de poda`(*X , A, y,* tarefa)

Nota:

Na aprendizagem automática, dividimos os dados em dois subconjuntos principais:

Conjunto de treino: Este subconjunto é utilizado para construir modelos como as árvores de regressão e as florestas aleatórias. Inclui características de entrada (variáveis independentes) e a variável-alvo (variável dependente). O modelo aprende padrões e relações a partir destes dados.

Conjunto de teste: Este subconjunto é utilizado para avaliar o desempenho do modelo. O conjunto de teste não é visto pelo modelo durante a fase de aprendizagem, permitindo uma avaliação imparcial da capacidade de generalização do modelo para dados novos e não vistos.

Em seguida, o algoritmo itera sobre todos os valores possíveis *v* do atributo escolhido *a*. Para cada valor *v*, cria um subconjunto de *X* em que o atributo *a* assume o valor *v*. Também actualiza o conjunto de atributos *A*, removendo o atributo *a*. O algoritmo chama-se a si próprio recursivamente para fazer crescer a árvore, utilizando o subconjunto de *X* e o conjunto de atributos actualizado *A*. Este processo recursivo continua, ao ligar novos nós ao nó superior com etiquetas correspondentes aos valores *v*.

Quando a árvore tiver crescido até à sua extensão máxima com base nos critérios iniciais, o algoritmo procede à poda da árvore. O processo de poda envolve a utilização de uma função de

poda separada que avalia se a remoção de determinados nós e ramos melhora o desempenho da árvore num conjunto de dados de teste. Isto é feito utilizando técnicas de validação cruzada para garantir que a árvore se generaliza bem para dados não vistos.

Ao repetir este processo, o algoritmo de crescimento da árvore constrói uma árvore de decisão que divide o conjunto de dados X em regiões cada vez mais pequenas. Cada nó terminal (folha) da árvore corresponde a uma região específica no espaço de características. Em tarefas de classificação, o nó da folha representa a classe maioritária dentro dessa região, enquanto que em tarefas de regressão, representa o valor médio de y.

A2 Florestas aleatórias

As árvores de decisão, embora úteis, têm limitações notáveis, em particular a sua tendência para sobre-ajustar os dados apesar da poda. Em cenários do mundo real, os dados podem ser confusos e conter anomalias que não se generalizam bem. As árvores de decisão podem criar divisões muito específicas que se ajustam aos dados de treino, mas não têm um desempenho exacto em dados novos e não vistos. As florestas aleatórias resolvem este problema utilizando várias árvores de decisão e calculando a média dos seus resultados. A simples geração de várias árvores a partir do mesmo conjunto de dados não resolve o problema, pois produziria árvores semelhantes. Em vez disso, as florestas aleatórias criam árvores, usando subconjuntos aleatórios dos dados. Esse processo de usar subconjuntos variados garante que as árvores sejam diferentes, o que ajuda a suavizar anomalias e melhorar a precisão geral da previsão, combinando as diversas árvores num modelo mais robusto.

Agregação de bootstrap e critérios de selecção

Nas florestas aleatórias, a aleatoriedade é introduzida de duas formas principais. Primeiro, selecionando um subconjunto aleatório de dados para cada árvore e, segundo, escolhendo um subconjunto aleatório de variáveis de previsão para cada divisão na árvore. Cada árvore numa floresta aleatória é construída utilizando uma técnica denominada agregação de bootstrap ou bagging. O algoritmo de bagging funciona recolhendo primeiro várias amostras aleatórias do conjunto de dados original. Digamos que retiramos B amostras, onde B é um número grande, geralmente na casa das centenas. Para cada amostra, é criada uma grande árvore de decisão sem qualquer simplificação. Estas árvores são depois utilizadas para fazer previsões. O algoritmo cria B regras de previsão a partir destas árvores e combina-as. Numa configuração em que testamos a precisão do modelo, são efectuadas B previsões para cada ponto de dados com base nos resultados de cada uma das árvores B. O passo final é calcular a média dessas previsões B para obter o valor final previsto.

As florestas aleatórias também introduzem aleatoriedade ao limitar as características consideradas em cada divisão. Em vez de avaliar todas as variáveis de previsão (variáveis x), em cada ponto de ramificação, o algoritmo selecciona aleatoriamente apenas um subconjunto destas variáveis para consideração. O tamanho deste subconjunto é normalmente pré-determinado, muitas vezes em torno da raiz quadrada do número total de preditores, com um mínimo comum fixado em 4. Esta abordagem é aplicada a cada amostra bootstrap, resultando na construção de árvores B. A previsão final é efectuada, através da média dos resultados destas árvores B.

A lógica subjacente à utilização de um número limitado de variáveis preditoras em cada divisão é minimizar a probabilidade de todas as árvores se tornarem demasiado semelhantes, especialmente se um preditor forte for dominante. Ao restringir o conjunto de variáveis em cada ponto de decisão, o algoritmo permite uma contribuição mais equilibrada de todos os preditores, incluindo os mais fracos, que podem fornecer informações valiosas quando considerados em conjunto. Sem

esta seleção aleatória, as árvores resultantes favoreceriam fortemente os preditores mais fortes, levando a previsões altamente correlacionadas e menos diversificadas.

Afinar o modelo

Ao executar uma floresta aleatória, há vários parâmetros de afinação importantes a serem considerados para garantir o desempenho ideal do modelo. Os parâmetros principais incluem o número de árvores, o número de preditores avaliados em cada divisão e a regra de paragem para o crescimento da árvore.

- Número de árvores (B):
  - Este parâmetro controla quantas amostras bootstrap são usadas para construir a floresta. Mais árvores geralmente aumentam a precisão do modelo, mas também o tempo de computação.
- Número de preditores por divisão (x):
  - Em cada nó, apenas um subconjunto de preditores é seleccionado para divisão. Uma boa regra é usar a raiz quadrada do número total de preditores. Por exemplo, com 64 preditores, usamos cerca de oito para cada divisão. Pelo menos quatro preditores devem ser usados.
- Regra de paragem para o crescimento da árvore:
  - Determina quando parar de dividir os nós de uma árvore. Uma regra simples é definir um número mínimo de observações por nó terminal. Normalmente, são utilizadas de 5 a 20 observações.

Em seguida, o método analisa a combinação destes três parâmetros de afinação que produz o menor erro de previsão. Este erro é medido pela raiz quadrada do erro médio (**RMSE**), que nos indica a distância entre as nossas previsões e os valores reais.

Outra métrica importante é o erro fora do saco, abreviado como **OOB**. Esta métrica estima o desempenho do modelo. Ao construir cada árvore na floresta, o algoritmo recolhe aleatoriamente amostras de aproximadamente 63,2 por cento dos dados, deixando os restantes 36,8 por cento como não utilizados ou "fora do saco". Estes dados fora do saco (OOB) não são utilizados na construção de uma determinada árvore, mas podem ser utilizados para estimar a exactidão dessa árvore, testando a forma como a árvore prevê os dados fora do saco (OOB). A média destes erros fora do saco (OOB) em todas as árvores da floresta fornece uma estimativa fiável do desempenho do modelo, conhecida como a taxa de erro fora do saco (OOB). Esta técnica garante que todos os pontos de dados são analisados na avaliação de desempenho do modelo, oferecendo assim uma medida robusta de precisão sem necessidade de um conjunto de teste separado (Hartshorn 2016).

Importância da variável

Na floresta aleatória, entender a importância de cada variável preditora é essencial para interpretar o modelo e refinar sua precisão preditiva. O modelo utiliza um método directo conhecido, como a importância da permutação, que avalia a importância da variável, observando as alterações na precisão da previsão, quando os valores de cada preditor são baralhados aleatoriamente. O desempenho de previsão do modelo é então comparado, usando os valores originais e permutados da variável, utilizando especificamente dados fora do saco (OOB). A importância da permutação é calculada, medindo o aumento no erro de previsão - como o erro quadrático médio (MSE) para

tarefas de regressão ou a taxa de erro para tarefas de classificação - quando os valores de uma variável são permutados nos dados fora do saco (OOB). Um aumento significativo no erro indica a importância da variável. Esta técnica não só identifica os principais factores de previsão, como também capta interacções complexas entre variáveis. Uma vez que o algoritmo de floresta aleatória selecciona subconjuntos aleatórios de preditores para cada divisão, o algoritmo pode identificar todos os preditores correlacionados, como importantes, se qualquer um deles contribuir significativamente para o resultado (Cutler et al. 2012).

A3 Exemplo

Esta secção desenvolve um exemplo simples para clarificar as caraterísticas da floresta aleatória. Para o efeito, centrar-nos-emos no desenvolvimento do modelo e da previsão, explicando cada passo, mas não fornecendo exemplos empíricos.

Imaginemos uma população de contribuintes igual a 100. Cada contribuinte preenche uma declaração de impostos que inclui a matéria colectável (montante sobre o qual incidem os impostos) e informação complementar. Suponhamos que a informação complementar é composta por dez variáveis. Por exemplo, pode ser o valor pago a título de salários de empregados e custos de produção, entre outros. É importante ressaltar que essas informações complementares não fazem parte directamente da base tributária, mas podem ser úteis para entender como chegar ao nível da base tributária.

Dos 100 contribuintes, 50 foram objeto de auditoria. Isto significa que, para 50 contribuintes, também temos informações sobre (potenciais) discrepâncias entre a declaração de base tributária e o montante efectivo. Para clarificar este ponto, imagine que todos os 50 contribuintes fogem aos impostos produzidos e que, através de auditorias, recolhemos (pelo menos) o montante incorrectamente declarado e a base tributária real.

O primeiro passo é compreender que só para 50 contribuintes é que temos informação sobre os montantes incorrectamente declarados. Isto significa que apenas nesta sub-amostra podemos contrastar as previsões com as variáveis reais para testar a exactidão do modelo de previsão. Por esta razão, vamos dividir a amostra inteira e concentrarmo-nos nos contribuintes auditados.

A amostra de contribuintes auditados é dividida em duas sub-amostras. Para simplificar, manteremos 25 contribuintes como amostra de treino e os restantes como amostra de teste. Na amostra de treino, executamos o modelo de floresta aleatória e utilizamos a amostra de teste para o afinar. Temos de escolher dois números críticos: o número de iteracções (ou o número de árvores) e o preditor por divisão. O modelo centra-se na estimativa da quantidade de declarações incorrectas, utilizando as co-variáveis (as dez variáveis adicionais que as empresas declaram). Por conseguinte, utilizaremos todas as variáveis por duas razões principais. Em primeiro lugar, essas variáveis ajudam a caracterizar a base tributária, sendo relevantes na determinação deste nível. Em segundo lugar, uma vez que esta informação está disponível, são também fundamentais para decidir o contribuinte que será objecto de auditoria. A incorporação de todas as variáveis permite-nos evitar um viés de selecção da amostra por factores observáveis.

Vejamos primeiro o número de árvores a utilizar. Para isso, mantemos o número de preditores utilizados em cada divisão (variáveis utilizadas para estimar as declarações incorrectas). Para simplificar, vamos assumir que vamos utilizar uma das dez variáveis disponíveis. Para decidir o número de árvores, é necessário executar o modelo na amostra de treino e comparar a previsão na amostra de teste, utilizando diferentes números de árvores. Por outras palavras, executamos N vezes N florestas aleatórias diferentes, alterando apenas o número de árvores que estamos a utilizar. Ao executar o modelo de floresta aleatória, fazemos previsões na amostra de teste e

comparamos o valor previsto com as declarações incorrectas reais encontradas no processo de auditoria. No final, teremos N valores da raiz quadrada do erro médio (RMSE) (um por cada execução do modelo). Escolhemos o valor mínimo e verificamos o número de árvores associadas, B. Este número é ideal, porque minimiza o erro de previsão medido pela raiz quadrada do erro médio (RMSE), produzindo a estimativa mais exacta da base tributável das declarações incorrectas.

Agora, passamos a estimar o preditor usado em cada divisão. Neste caso, usamos o número ideal de árvores, B, que foi obtido anteriormente. Repetimos o mesmo procedimento de iteracção, mas, neste caso, executamos dez modelos diferentes de floresta aleatória, obtendo a previsão na amostra de treino para cada um deles e comparando-a com o valor real de declarações incorrectas. Executamos dez modelos, porque temos dez variáveis para usar. Isto deve-se ao facto de o número total de variáveis ser o número máximo de preditores para cada divisão. É importante notar que, se tiver dez variáveis, mas decidir utilizar oito para o modelo de previsão, terá de executar oito modelos. O número de modelos a serem executados nesta etapa deve ser sempre igual às variáveis decididas a serem usadas no modelo de previsão. Finalmente, repetimos o processo, escolhendo a raiz quadrada do erro médio (RMSE) mínima e verificando o número de preditores utilizados, *x*. Este número de preditor *x* é ideal para minimizar o erro de previsão.

Com estes dois passos, encontramos o número ideal de árvores (B) e de preditores por divisão (*x*) a utilizar na floresta aleatória. Recorde-se que, para os estimar, utilizamos os 50 contribuintes auditados, dividindo esta amostra em conjuntos de treino e de teste. Agora, podemos fazer as previsões para os outros 50 contribuintes que não foram auditados. O procedimento é o seguinte. Primeiro, executar a floresta aleatória com os parâmetros ideais no conjunto de 50 contribuintes auditados. Posteriormente, prever os valores no conjunto de 50 contribuintes não auditados. Por último, é possível criar uma variável composta pela declaração incorrecta descoberta para os 50 contribuintes auditados e a declaração incorrecta prevista para os 50 contribuintes não auditados.