PREVISÃO DE COMPORTAMENTO E CLASSIFICAÇÃO DE CONTRIBUINTES TRIBUTÁRIOS: UMA ABORDAGEM POR MODELOS LINEARES GENERALIZADOS HIERÁRQUICOS

Alexandre S. Barreto (autor responsável), Pesquisador, Ministério da Fazenda, Esplanada dos Ministérios, Bloco P, 9 andar, Brasília-DF, Brazil, CEP: 70000-000, fone: +51(61)34122951, e-mail: alexsbdr@yahoo.com.br

Aran B. T. Morales, Diretor Técnico-Científico, Instituto Stela, R. Prof. Ayrton Roberto de Oliveira, 32, Itacorubi, Florianópolis-SC, Brazil, CEP 88034-050, +51(48)32392529, e-mail: aran@stela.org.br

Dalton F. Andrade, Professor Adjunto, Universidade Federal da Santa Catarina, Depto. de Informática e Estatística, Campus Universitário, Florianópolis-SC, CEP: 88000-000, +51(48)33317545, e-mail: dandrade@inf.ufsc.br

ABSTRACT: A crucial task to tributary agencies is to select taxpayers for audit purposes and this is challenged by the unavoidable time gap between the occurrence of a taxpayer's infraction and its effective detection by the agency. There are studies addressing tax compliance by linear regression, but have several limitations due to the possible dependence among observations within the same economic sector. We propose a novel approach, apparently not yet applied in tributary studies, relying in a hierarchical generalized linear model to predict tax compliance behavior and establish a subsequent classification among corporate taxpayers based in their predicted probabilities. The model account for within-cluster correlation where clusters refer to economic sectors. The taxpayer's classification into compliant/non-compliant is made via cut-points and we assess the predictive performance of the model by estimating internal validation measures. The focused database belongs to Brazilian taxes agency and is related to its regular taxpayer selection method and audits. The main results allow important tributary interpretations and showed that the model can classify faster, provided the observation to some prescriptions, and is more predictive than the mentioned regular selection method. Commands from a free statistical computing software utilized to produce subject-specific inferences are accompanying this article.

Key-Words: prediction, classification, taxes, hierarchical generalized linear models.

1 Introdução

Uma das maiores preocupações das agências tributárias tem sido a manutenção e melhoria do processo de formulação e emissão de políticas de fiscalização tributária. O processo de fiscalização tributária pressupõe a existência de um processo de seleção de contribuintes, que pode ser intrinsecamente aleatório, ou então não aleatório - este último procura elevar o valor esperado das auditorias executadas. Quando as auditorias envolvem os contribuintes pessoas jurídicas (PJ), a complexidade da tributação envolvida e da possível auditoria aumenta sensivelmente devido à ampla gama da operações comerciais e de tributos envolvidos em um contexto empresarial e isso, consequentemente, eleva da mesma maneira a complexidade da avaliação da capacidade contributiva do contribuinte. De qualquer forma, a tarefa de selecionar contribuintes para fins de fiscalização é desafiada pelo inevitável período de tempo decorrido entre o cometimento de uma infração por parte do contribuinte e a efetiva detecção desta infração por parte da agência tributária. A existência deste aspecto geral tributário restará mais claro após uma breve descrição do principal processo de seleção de contribuintes para fiscalização utilizado pela Secretaria da Receita Federal (SRF), conforme segue.

Anualmente os contribuintes pessoas jurídicas enviam suas declarações de informações econômico-fiscais das pessoas jurídicas (DIPJ) à SRF pela internet. Ao lado disso, a agência brasileira recebe diversas outras informações por meio de convênio, ou então compulsoriamente, de fontes externas como Ministérios, INSS, estados, municípios e bancos. Dessa forma, as DIPJ, as informações de fontes externas e os registros de pagamentos/retenções formam os subsídios à consolidação de uma base de dados que permite cruzar todas as informações simultâneamente com o objetivo de detectar indícios de infração tributária e, ato contínuo, selecionar contribuintes para fiscalização. Destaque-se que um indício está geralmente associado a uma inconsistência material em relação às informações do contribuinte. Uma vez selecionada uma empresa (contribuinte PJ), a ação fiscal é finalmente programada e realizada, podendo durar meses e até anos, a depender da complexidade das infrações efetivamente encontradas, dos tributos envolvidos e da relevância fiscal da ação. Ao final da fiscalização duas ações são possíveis: pode ser lavrado um auto de infração discriminando os tributos e as respectivas infrações fiscais cometidas, assim como sua capitulação legal, e o valor em espécie dos tributos e multas a pagar; ou então a ação fiscal pode ser encerrada sem resultado.

O fato relevante aqui é que para deflagrar as auditorias do exercício, a agência tem antes de

receber e consolidar uma ampla gama de informações econômico-fiscais das pessoas jurídicas, externas e de pagamentos/retenções, informações estas no mais da vezes protegidas por sigilo fiscal. Todo esse processamento, na prática, implica inevitável tempo decorrido entre o cometimento de uma infração fiscal por parte do contribuinte e sua efetiva detecção por parte da agência. Contudo, é importante ressaltar que esse espaço de tempo para operacionalização das auditorias não é uma exclusividade da SRF, mas antes, uma característica do domínio tributário, como argumentado em Andreoni (1992).

Pode-se dizer que o processo de seleção acima descrito é bastante moderno e atual e vem funcionando a contento, não fosse isso não se verificariam os sucessivos recordes anuais de arrecadação federal vivenciados pela SRF. Contudo ele possui certas limitações. A primeira diz respeito a seu volumoso custo de implementação, mas este aspecto não será aprofundado neste trabalho. A outra relaciona-se ao mencionado tempo decorrido entre uma infração fiscal e sua efetiva detecção por parte da agência. Quanto maior esse tempo decorrido, maior o risco de embaraço contábil e cadastral à constituição e cobrança do crédito tributário na esfera administrativa, já que muitas vezes quando se vai notificar um contribuinte após decorrido um longo tempo do cometimento de uma determinada infração, este pode ter incorrido em fatos contábeis relevantes - falência, cisão, incorporação, fusão, ou ainda não ser localizado em seu domicílio fiscal. Assim, qualquer esforço no sentido de agilizar a ação do sistema de fiscalização é considerado positivo e relevante no contexto de órgãos tributários como a SRF.

É importante notar que nem todas as pessoas jurídicas que possivelmente apresentem indícios de infração fiscal serão efetivamente fiscalizadas. Na verdade, principalmente devido ao instituto jurídico tributário da decadência e, também, à carência de servidores públicos, as agências devem por ofício atuar nos casos mais relevantes e isso certamente implica a adoção de um determinado nível tolerável de evasão fiscal por parte dos órgãos tributários, como declarado em Reinganum and Wilde (1988).

Uma constatação relevante nesse momento é que as pessoas jurídicas (empresas) subdividem-se em diversos setores de atividades econômicas que são de naturezas distintas (v.g., indústria, comércio, serviços, extrativas etc). Isso acarreta que os setores da economia devem ser levados em consideração no estabelecimento de padrões de comportamento econômico-fiscais de empresas, pois é intuitivo que empresas de um mesmo setor econômico tendem a ser mais parecidas entre si do que empresas participantes de distintos setores da economia, haja vista que aquelas partilham certas características contábeis individuais e de desempenho econômico-setorial em comum que acabam por ocasionar a correlação de

medidas intra-setores. Além disso, acredita-se que as próprias infrações à legislação tributária tendem a ser mais parecidas intra-setores, pois o fenômeno da concorrência de mercado talvez faça com que sejam rapidamente disseminadas entre empresas que são diretamente concorrentes.

Na verdade, esse tipo de correlação é bastante comum em estruturas de dados correlacionados, ou estruturas hierárquicas de dados, tais como contribuintes agrupados em setores econômicos, e, portanto, deve ser avaliada pelo pesquisador ao propor métodos não aleatórios de seleção de contribuintes para fiscalização. Dessa forma, no que concerne aos modelos estatísticos, é importante dispor daqueles que, em sua formulação, levem em consideração a correlação de medidas intragrupos.

A partir do exposto, o reconhecimento da possível correlação de medidas intragrupos no contexto tributário e o fato de que essa correlação de medidas não vem sendo considerada de forma sistematizada (abrangendo, de forma estruturada, todos os setores de atividades econômicas das PJ) pelos estudos e trabalhos já realizados envolvendo a seleção de contribuintes para fiscalização, conforme o exposto na próxima seção, este artigo - fundamentado na pesquisa documentada em Barreto (2005) - apresenta um novo método para, agilmente, classificar e selecionar contribuintes para fiscalização, com base na previsão de seus comportamentos tributários, levando em consideração e avaliando, de forma sistematizada, a existência de correlação de medidas intra-setores econômicos, por meio de modelos estatísticos apropriados para esta finalidade.

2 Modelando efeitos dos Setores Econômicos

2.1 Estratégia Desagregada

A maneira mais simples de modelar a evasão tributária, por exemplo o relacionamento entre o Lucro Real (LR) do contribuinte e um índice econômico-fiscal específico, poderia se dar por uma regressão linear (nos parâmetros e na variável preditora), como segue:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{2.1}$$

$i = 1, 2, \ldots, n$; onde Y_i: lucro real do i-ésimo contribuinte, X_i: índice econômico-fiscal para o i-ésimo contribuinte, β_0 e β_1: coeficientes de regressão, ϵ_i: termo de erro aleatório associado ao i-ésimo contribuinte. Nesse modelo assumem-se ϵ_i's independentes e $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$.

Rememorando o teorema de Gauss-Markov, vemos que este modelo produz estimativas de maior eficiência, mas sob a suposição de termos de erros não correlacionados (DAVIDSON e MACKINNON, 1993). Ocorre que essa suposição pode não ser realista quando se está lidando com estruras de dados correlacionados como, por exemplo, contribuintes aninhados ou agrupados em setores econômicos (neste momento não importando se se trata de contribuintes pessoas físicas ou jurídicas), como já explanado.

Em estudos envolvendo métodos de seleção de contribuintes para fiscalização tem sido usual modelar o comportamento do contribuinte por meio da equação de regressão (2.1), assumindo termos de erros não correlacionados, e levando em consideração os efeitos setoriais por meio de algumas variáveis dummy. Esses estudos, em geral, consideram e consequentemente são válidos apenas para alguns poucos setores econômicos. Possivelmente, tais trabalhos vivenciaram e foram desafiados pelo risco crescente de viés nas estimativas à medida que mais e mais variáveis dummies (setores econômicos) são incorporadas aos modelos, não obstante sustentarem ainda a tradicional suposição de termos de erros não correlacionados. E mais, indo mais adiante nessa estratégia e considerando todos os setores de uma economia em um modelo de regressão por meio de variáveis dummies, não seria garantida a existência de graus de liberdade suficientes para a produção de dezenas ou até centenas de estimativas, a depender do caso em estudo.

Como exemplos dessas abordagens tributárias, podem-se mencionar os estudos de Mills (1996), Murray (1995), Joulfaian e Rider (1998) entre outros. Uma revisão desses trabalhos pode ser vista em Barreto (2005), no qual foi identificada, até o momento, a falta de uma abordagem sistematizada, em termos de setores econômicos, no que se refere aos métodos de seleção de contribuintes para fiscalização.

2.2 Estratégia Agregada

Outra possibilidade de modelar a evasão tributária é trabalhar de maneira agregada, segundo o seguinte modelo:

$$\bar{Y}_{.j} = \beta_0 + \beta_1 \bar{X}_{.j} + \bar{\epsilon}_{.j} \tag{2.2}$$

$j = 1, 2, \ldots, J$; onde $\bar{Y}_{.j}$: lucro real médio do j-ésimo setor, $\bar{X}_{.j}$: índice econômico-fiscal médio para o j-ésimo setor, β_0 e β_1: coeficientes de regressão, $\bar{\epsilon}_{.j}$: erro aleatório médio associado ao j-ésimo setor. Suposições: $\bar{\epsilon}_{.j}$'s independentes e $\bar{\epsilon}_{.j} \sim \mathrm{N}(0, \sigma^2)$.

Pode-se certamente proceder como modelado na equação (2.2), mas estar-se-ia perdendo bastante

informação, como afirmam Bryk e Raudenbush (1992), porque em estruturas hierárquicas de dados, em geral, a maior parte da variabilidade da resposta é devida à variabilidade dos indivíduos.

2.3 Uma regressão linear designada a cada setor econômico

Finalmente, pode-se abordar o problema de modelar o comportamento do contribuinte utilizando uma regressão específica, como mostrado na equação (2.2), para cada um dos setores econômicos. Essa abordagem pode lidar eficientemente com as naturais diferenças entre interceptos e inclinações de setores distintos, e a suposição de termos de erros não correlacionados pode ser sustentada, na medida em que há uma equação de regressão (e um termo de erro específico) designada para cada setor. O problema aqui está mais associado ao elevado número de equações a serem estimadas, a depender do caso, e a possível existência de tamanhos de amostra insuficientes para a estimação dos coeficientes de regressão.

A partir de todo o exposto, vê-se que a correlação intra-setores deve ser incorporada e avaliada de alguma maneira ao se modelar o comportamento tributário de contribuintes PJ, e mais, ela deve ser abordada de forma sistematizada (envolvendo todos os setores da economia). A proposta da pesquisa aqui enfocada é, portanto, modelar tal comportamento por Modelos Lineares Generalizados Hierárquicos (MLGH), um desenvolvimento a partir do Modelos Lineares Hierárquicos (MLH), estes últimos especificamente desenvolvidos para os casos em que a suposição de termos de erros não correlacionados não se sustenta.

2.4 Modelando efeitos de setores econômicos pr MLGH

Como pode ser observado em Raudenbush e Bryk (2002) e também em Goldstein (2003), um MLH é projetado para lidar com estruturas de dados correlacionados dispostas em uma estrutura hierárquica, como estudantes (nível 1) agrupados em escolas (nível 2) e assim por diante. A contabilização da correlação intragrupos nesta classe de modelos está diretamente associada a termos de efeitos aleatórios atribuídos a cada um dos níveis hierárquicos, o que resulta na indução de uma estrutura de variância/covariância para a resposta. Portanto, o MLH permite avaliar a variabilidade da resposta entre os grupos, além de fornecer estimativas pontuais para todos os grupos ou clusters considerados no estudo. A grande potencialidade aqui está em obter todos esses resultados em um único modelo.

Já o MLGH é uma especificação ulterior dos MLH aplicável quando a variável aleatória resposta distribui-se de acordo com a família exponencial de distribuições. Os MLGH são os requisitados nesta

pesquisa, haja vista ser utilizada uma variável aleatória binária como resposta resposta de trabalho. Portanto, nesta seção é introduzido o MLGH por meio da teoria desenvolvoda em Raudenbush e Bryk (2002).

Inicialmente, deve-se mencionar que o MLGH está, identicamente ao MLH, estruturado em níveis hierárquicos distintos. O nível 1 de um MLGH é formado por três partes distintas, a saber: um modelo amostral não Normal, uma função de ligação e um modelo estrutural.

Para que seja especificado o modelo amostral não Normal, seja E um experimento no qual A denota um evento binário a ele associado. Assume-se que as probabilidades de ocorrência para este evento são P(A) = ϕ e P($\bar{A}$) = 1- ϕ, e também que essas probabilidades mantêm-se constantes ao longo das n repetições de E. O espaço amostral deste experimento é dado por todos os conjuntos possíveis de $\{a_1, a_2, \ldots, a_n\}$, sendo que, nele, a_i é A ou $\bar{A}$ a depender da ocorrência ou não do evento A na i-ésima repetição de E. A partir do definido, seja Y uma variável aleatória (VA) representando o número de vezes em que A ocorre em n repetições de E. Os valores assumidos por Y são, claramente, 0, 1, 2,..., n; e a cada um desses valores está associada uma probabilidade P(Y). Essas definições do experimento levam diretamente a que Y siga a conhecida distribuição Binomial de probabilidades com parâmetros n e ϕ, em relação ao seu espaço amostral. Como firmamos aqui o compromisso de modelar uma estrutura hierárquica ou aninhada de dados, devemos introduzir neste momento os subscritos i e j responsáveis por identificar indivíduos e setores econômicos. Deflui disso que nossa VA de interesse se torna Y_{ij}, contabilizando agora o número de ocorrências de A em n_{ij} repetições independentes de E. Adotando essa notação, Y_{ij} pode assumir os valores 0, 1, 2,..., n_{ij}; a probabilidade de sucesso (ocorrência de A) mantém-se ϕ_{ij} para todas as repetições e Y_{ij} agora segue $Y_{ij}|\phi_{ij} \sim B(n_{ij}, \phi_{ij})$.

É bem conhecido que no caso de uma única repetição para E, a VA Y_{ij} segue uma distribuição Bernoulli e ϕ_{ij} significa então a probabilidade associada à ocorrência de A, ou, em outras palavras, o valor esperado para Y_{ij}. Como ϕ_{ij} necessita estar constrangido ao intervalo [0,1], a função de resposta do modelo tem de ser adaptada a isso, e é frequentemente usada nestes casos a função de resposta logística, uma função não linear relacionando ϕ_{ij} a uma parâmetro, normalmente denotado por η_{ij}, na prática a ser predito pelo modelo, conforme segue:

$$E(Y_{ij}) = \phi_{ij} = \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})} \tag{2.3}$$

Assim, η_{ij} pode ser relacionado a um modelo linear (o modelo estrutural de nível 1) contando com Q variáveis explicativas (ou preditoras) e p parâmetros:

$$\eta_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + ... + \beta_{Qj}X_{Qij} = \beta_{0j} + \sum_{q=1}^{Q} \beta_{qj}X_{qij} \tag{2.4}$$

q= 0,1,..., Q; onde β_{qj}: coeficientes de nível 1, X_{qij}:preditoras de nível 1.

Com o advento do modelo estrutural de nível 1, o valor esperado para a equação (2.3) agora significa a probabilidade de sucesso dados os níveis das preditoras em (2.4). Resolvendo (2.3) em termos de η_{ij}, obtém-se a função de ligação de nível 1 envolvendo uma transformação do valor esperado para Y_{ij}:

$$\eta_{ij} = \log(\frac{\phi_{ij}}{1-\phi_{ij}}) \tag{2.5}$$

Pela equação (2.5) pode-se observar que o MLH é, na verdade, um caso particular do MLGH onde a função de ligação não envolve uma transformação da resposta esperada, mas antes uma função de ligação identidade $\eta_{ij} = \mu_{ij}$.

Em MLGH o modelo estrutural de nível 2 apresenta a seguinte forma geral:

$$\beta_{qj} = \gamma_{q0} + \gamma_{q1}W_{1j} + ... + \gamma_{qS_q}W_{S_qj} + u_{qj} = \gamma_{q0} + \sum_{s=1}^{S_q} \gamma_{qs}W_{sj} + u_{qj} \tag{2.6}$$

q = 0, 1,..., Q; s = 0,1,..., S_q; onde γ_{qs}: coeficientes de nível 2 (efeitos fixos), W_{sj}: preditoras de nível 2, u_{qj}: efeito aleatório de nível 2 relacionado ao coeficiente β_{qj}. As suposições do modelo relacionam-se ao fato de que para cada grupo j existe um vetor de efeitos aleatórios $\mathbf{U}_j = (u_{0j}, u_{1j}, \ldots, u_{Qj})^T$ distribuído segundo a Normal Multivariada e contando com elementos $u_{qj} \sim \mathrm{N}(0, \tau_{qq})$. Também, para cada dupla de efeitos aleatórios pertencentes ao mesmo grupo, a covariância entre eles equivale a $\mathrm{Cov}(u_{qj}, u_{q'j}) = \tau_{qq'}$. A equação (2.6) formaliza o caso mais completo do MLGH, mas não necessariamente, nestes modelos, todos os coeficientes de nível 1 têm de ser modelados como aleatórios.

Por meio de (2.6) pode-se observar uma outra potencialidade dos MLH/MLGH, que é não só contabilizar adequadamente a correlação intragrupos, mas, ao lado disso, proporcionar a tentativa de explicação da variabilidade entre os grupos, ou unidades de nível 2, por meio de variáveis explicativas de nível 2, W_{sj}; ressalte-se que este último aspecto não pode ser atendido por abordagens como a dos Modelos Lineares Generalizados.

O já mencionado nível 1 de um MLGH pode ser denotado genericamente por:

$$Y_{ij} = \phi_{ij} + \epsilon_{ij} \tag{2.7}$$

A partir da equação (2.3) observa-se que esse modelo é não linear, mas pode ser aproximadamente linearizado, como mostrado em Raudenbush e Bryk (2002), por meio da expansão de taylor, até a primeira ordem, centrada em $\eta_{ij}^{(s)}$, que significa η_{ij} avaliado para $\phi_{ij}^{(s)}$ (ϕ_{ij} avaliado no s-ésimo passo) fornecendo:

$$\phi_{ij} \approx \phi_{ij}^{(s)} + \frac{d\phi_{ij}}{d\eta_{ij}}(\eta_{ij} - \eta_{ij}^{(s)}) \tag{2.8}$$

Também, pode-se calcular a seguinte derivada:

$$\frac{d\phi_{ij}}{d\eta_{ij}} = w_{ij} = \phi_{ij}(1 - \phi_{ij}) \tag{2.9}$$

Avaliando (2.9) para $\phi_{ij}^{(s)}$ e substituindo-a em (2.7), obtém-se:

$$Y_{ij} = \phi_{ij}^{(s)} + w_{ij}^{(s)}(\eta_{ij} - \eta_{ij}^{(s)}) + \epsilon_{ij} \tag{2.10}$$

O aspecto chave agora é posicionar todos os termos observáveis no lado esquerdo da igualdade, resultando em:

$$\frac{Y_{ij} - \phi_{ij}^{(s)}}{w_{ij}^{(s)}} + \eta_{ij}^{(s)} = \eta_{ij} + \frac{\epsilon_{ij}}{w_{ij}^{(s)}} \tag{2.11}$$

sob as seguintes suposições: $\frac{\epsilon_{ij}}{w_{ij}^{(s)}} \sim N(0, w_{ij}^{(s)-1})$; $u_{0j} \sim N(0, \tau_{00})$ e u_{0j}'s independentes; $u_{1j} \sim N(0, \tau_{11})$ e u_{1j}'s independentes.

A título ilustrativo, considere-se agora o caso dos contribuintes PJ aninhados em setores econômicos e assumindo que se está sob o conhecido modelo de coeficientes aleatórios (todos os coeficientes de nível 1 são considerados aleatórios) contando com apenas uma preditora de nível 1 (um índice econômico-fiscal individual, por exemplo "IND") e uma única preditora de nível 2 (que poderia ser uma variável indicadora binária sinalizando se o setor econômico é ou não exportador, nomeada "Export"). Ou seja, esse modelo ilustrativo possui Q =1 na equação (2.4), e Q =1 e S_q =1 na equação (2.6). Para esse caso particular, substituindo (2.6) em (2.4) obtém-se o seguinte modelo estrutural combinado:

$$\eta_{ij} = \gamma_{00} + \gamma_{01}Export_j + \gamma_{10}IND_{ij} + \gamma_{11}Export_jIND_{ij} + u_{0j} + u_{1j}IND_{ij} \tag{2.12}$$

E substituindo-o em (2.11) obtém-se:

$$\begin{aligned}\frac{Y_{ij}-\phi_{ij}^{(s)}}{w_{ij}^{(s)}}+\eta_{ij}^{(s)} &= \gamma_{00}+\gamma_{01}Export_j+\gamma_{10}IND_{ij}+\gamma_{11}Export_jIND_{ij}+ \\ & \quad u_{0j}+u_{1j}IND_{ij}+\frac{\epsilon_{ij}}{w_{ij}^{(s)}} \qquad (2.13)\end{aligned}$$

A partir desta última equação, pode-se ver que a matriz de variância/covariância da resposta induzida, pelo modelo, para um setor econômico j possuindo dois contribuintes PJ é dada por:

$$\mathbf{VC}_j=\begin{pmatrix} A & B \\ B & C \end{pmatrix}$$

onde

$A=(\tau_{00}+2\tau_{01}IND_{1j}+\tau_{11}IND_{1j}^2+w_{ij}^{(s)-1})$,

$B=\Big(\tau_{00}+\tau_{01}(IND_{1j}+IND_{2j})+\tau_{11}IND_{1j}IND_{2j}\Big)$,

$C=(\tau_{00}+2\tau_{01}IND_{2j}+\tau_{11}IND_{2j}^2+w_{ij}^{(s)-1})$.

Pelo exposto, o objetivo aqui de contabilizar adequadamente (de forma sistematizada) a correlação de medidas intra-setores econômicos em uma modelagem estatística é atingido. Assim, o padrão, em termos de variáveis exploratórias, demonstrado em (2.13), pode ser explorado nas próximas seções.

Finalmente, no que concerne inferência estatística, em MLH podem-se obter três tipos de estimativas: para efeitos fixos (γ_{qs}'s), coeficientes aleatórios de nível 1 (β_{qj}'s) e componentes de variância/covariância (τ_{qq}'s and $\tau_{qq'}$'s). Os efeitos fixos não variam entre os grupos e podem ser estimados por Mínimos Quadrados Generalizados ou Máxima Verossimilhança.

Os coeficientes aleatórios de nível 1 podem ser obtidos por meio dos estimadores empíricos de Bayes. Estes últimos são, na verdade, uma composição ótima da seguinte forma:

$$\beta_j^*=\Lambda_j\hat{\beta}_j+(\mathbf{I}-\Lambda_j)\mathbf{W}_j\hat{\gamma} \qquad (2.14)$$

onde $\hat{\beta}_j$: vetor de estimadores de mínimos quadrados ordinários para β_j, $\mathbf{I}$: matriz identidade (n_j x n_j), $\mathbf{W}_j$: matriz [$(Q+1)$ x F] de variáveis explicativas de nível 2, $\hat{\gamma}$: um vetor (F x 1) de efeitos fixos, Λ_j= $\mathbf{T}(\mathbf{T}+\mathbf{V}_j)^{-1}$, na qual $\mathbf{T}$ é uma matriz [$(Q+1)$x$(Q+1)$] de componentes de variância/covariância (τ_{qq}'s and $\tau_{qq'}$'s) significando a dispersão paramétrica de β_j e $\mathbf{V}_j$ representa a dispersão do erro de $\hat{\beta}_j$ como

um estimador de β_j. Assim, o estimador empírico de Bayes da equação (2.14) compõe uma combinação ótima dos estimadores de mínimos quadrados ($\hat{\beta}_j$) e da grande média estimada ($\mathbf{W}_j\hat{\gamma}$).

Já os componentes de variância/covariância da matriz $\mathbf{T}$ são geralmente estimados por algoritmos de máxima verossimilhança.

No caso do MLGH, η_{ij} na equação linearizada (2.11) pode ser estimado por Penalized Quasi-Likelihood (PQL), um algoritmo numérico que encapsula um processo máxima verossimilhança internamente em seus passos iterativos com o objetivo de estimar os referidos parâmetros de MLH. O leitor pode acessar Raudenbush e Bryk (2002) ou Goldstein (2003) para maiores detalhes em estimação de MLGH e também sobre algoritmos disponíveis para este fim.

3 Novo método de classificação de contribuintes para fiscalização

Neste artigo é apresentada uma nova abordagem para a seleção de contribuintes pessoas jurídicas para fiscalização. Esta se fundamenta em um MLGH para prever o comportamento tributário dos contribuintes e em sua subsequente classificação em termos de infratores/não infratores com base nas respectivas probabilidades preditas. A utilização de MLGH é justificada pelo exposto nas seções anteriores, e isso possibilita levar em consideração a correlação de medidas intra-setores de forma sistematizada, ou, em outras palavras, considerando ao mesmo tempo, e em um único modelo, os efeitos de todos os setores da economia.

Os dados de trabalho permitem o conhecimento sobre as auditorias de pessoas jurídicas conduzidas pela SRF e seus respectivos resultados. Assim, pode-se definir a variável binária resposta em termos do comportamento tributário infrator/não infrator em relação à legislação tributária Federal - envolvendo todos os impostos e contribuições administrados pela SRF e incidentes sobre as PJ (IRPJ, CSLL, PIS/PASEP, COFINS e IPI)- comportamento este a ser predito por um MLGH. Em resumo, a variável binária resposta de trabalho indica se o contribuinte é ou não um infrator, e a característica de interesse a ser modelada aqui é "ser um infrator das normas tributárias".

Para que seja atingida a almejada agilidade da fiscalização tributária, como mencionado na introdução, a proposta é que MLGH utilize como variáveis de nível 1 tão-somente as informações prestadas

pelo próprio declarante em sua DIPJ (indicadores econômico-fiscais individuais e mais alguns valores contábeis brutos), de forma a que o MLGH seja independente de dados individuais de contribuintes prestados por fontes externas de informação e também de dados de retenção, que tipicamente requisitam um intervalo de tempo precioso para serem consolidados e disponibilizados ao processo de seleção de contribuintes. No caso do nível 2, serão pesquisadas variáveis setoriais não sigilosas de fácil e rápido acesso, que possam explicar a variabilidade nas respostas devida exclusivamente aos setores de atividades econômicas. Estas características de especificação e utilização do modelo possibilitam uma maior agilidade de ação por parte da fiscalização tributária, na medida em que a classificação dos contribuintes e sua subsequente seleção podem ser executadas tão logo as declarações de informação das PJ sejam recebidas pela agência tributária.

Sintetizando, o novo método de classificação se fundamenta em uma tríade: utilização de MLGH de forma a levar em consideração a correlação intragrupos, onde os grupos relacionam-se aos setores econômicos; utilização, como variáveis preditoras de nível 1, tão-somente das informações compulsoriamente prestadas pelo próprio declarante em sua DIPJ; consideração, como base para a formação de variáveis explicativas de nível 2, apenas de informações não sigilosas e de fácil e rápido acesso.

Deve-se salientar neste momento que este novo método de classificação, uma vez materializado, vem a corresponder, na prática, a um processo de seleção de contribuintes para fiscalização, fundamentando-se na previsão de comportamento e classificação dos contribuintes em dois grupos distintos, os de interesse à fiscalização tributária (que comporão a programação de ações fiscais) e os demais (controlados nos dossiês de contribuintes). Nas próximas seções, os resultados da aplicação deste novo método são apresentadas, a partir dos subsídios da análise exploratória de dados descrita na próxima seção.

4 Análise exploratória de dados

4.1 Descrição da base de dados

Os dados disponíveis para esta pesquisa são relacionados às informações prestadas à SRF pelas próprias pessoas jurídicas brasileiras (164.466 contribuintes) em suas DIPJ referentes ao exercício 2000, ano-calendário 1999, que apuraram o imposto de renda pelo regime de apuração do Lucro Real, excluindo-se as instituições financeiras. São as grandes empresas brasileiras, uma vez que o Lucro Real foi o regime

obrigatório para os contribuintes cuja receita total fosse superior ao limite anual de 24 milhões de reais, para o ano-calendário de 1999. Esse regime de apuração exige completa e detalhada escrituração contábil dos contribuintes e não adota nenhum processo estimado ou de presunção de lucro.

Nesse contexto, os dados apresentam o seguinte conteúdo:

a) 33 índices econômico-fiscais de resultado, mistos e de balanço; (ver Apêndice F de Barreto (2005);

b) 10 valores contábeis brutos, quais sejam: - Lucro Operacional (LOPER), - Lucro Bruto (LBRUTO), - Receita Líquida (RECLIQ), - Lucro Real (LREAL), - Lucro antes da Contribuição Social (LACSLL), - Base de Cálculo da Contribuição Social sobre o Lucro Líquido (BCCSLL), - Receita Bruta do PIS/PASEP (RBPISP), - Base de Cálculo do PIS/PASEP (BCPISP), - Receita Bruta da COFINS (RBCOF), - Base de Cálculo da COFINS (BCCOF).

c) A classificação nacional de atividade econômico-fiscal do contribuinte (CNAE-Fiscal);

d) A informação sobre o preenchimento do Anexo de Atividade Rural na DIPJ do contribuinte;

e) A informação sobre os contribuintes PJ (9.757 do total) selecionados para fiscalização em função de indício de infração fiscal, associado ao ano-calendário 1999, exercício 2000, detectado com base no método atual de seleção da SRF;

f) A informação sobre 12.227 auditorias concluídas relacionadas às operações praticadas no ano-calendário 1999, exercício 2000 (englobando toda e qualquer ação fiscal que inclua o ano-calendário 1999, mesmo que os indícios deflagradores da ação se refiram a outros anos-calendário que não o de 1999), sendo que, dentre estas auditorias, 8.073 foram concluídas com a apuração de uma ou mais infrações e, portanto, implicaram autuação do contribuinte PJ por parte da SRF.

g) A informação de que dentre os 9.757 contribuintes PJ selecionados para fiscalização, em relação ao ano-calendário 1999, exercício 2000, em razão de indício de infração detectado com base no cruzamento de informações internas e externas executado pelo método atual de seleção da SRF, 8.728 deles constam como fiscalizados entre 12.227 auditorias gerais concluídas mencionadas no item "f". Além disso, a informação de que 5.180 contribuintes deste último subconjunto de 8.728 fiscalizados foram autuados em relação às operações efetuadas no ano-calendário 1999, exercício 2000, ou seja, 59,35 por cento apresentaram infrações à legislação fiscal.

Deve-se mencionar que embora nesta pesquisa se trabalhe com toda a população de declarantes

PJ (após a exclusão de 8.263 registros de contribuintes PJ apresentando dados faltantes em termos de informações econômico-setoriais, e que aqui são considerados "missing completely at random"), cientificamente deseja-se considerar a população atual como uma realização amostral obtida a partir de uma população conceitualmente infinita estendendo-se no tempo e, também, possivelmente no espaço, na verdade uma superpopulação, conforme conceituado em (GOLDSTEIN, 2003), a qual admite generalizações e predições além da população atual.

Inicialmente, todas as estatísticas descritivas, histogramas, box-plots e stem and leaf plots foram avaliados para todos os índices econômico-fiscais e, ao lado disso, foi analisado também se cada um desses índices pode ou não assumir valor negativo de acordo com a teoria contábil, o que é relevante no contexto das infrações fiscais - ver apêndice F em Barreto(2005).

Uma necessidade de trabalho foi a de, na apresentação de resultados, denotar os índices econômico-fiscais de maneira genérica, por meio de um "I" maiúsculo seguido de um número sequencial, ou seja, na verdade eles estão descaracterizados confome foi solicitado pelos proprietários dos dados, mormente porque eles foram utilizados na pesquisa como sedimento para um MLGH voltado a prever o comportamento tributário infrator, e claramente a SRF não possui interesse em revelar publicamente uma fórmula de tal escopo.

As análises exploratórias permitiram a definição de três diretrizes gerais de transformação aplicáveis aos índices, conforme segue (onde X denota um número identificador sequencial genérico para o índice):

1) Índices que podem assumir, teoricamente, qualquer valor e que apresentam padrão de histograma similar ao de I04 - ver Apêndice G em Barreto (2005): uma variável assumindo log(IX + 1) para valor maior ou igual a 0 (zero) e -log(—IX— + 1) para valor negativo.

2) Índices que não podem assumir, pela teoria contábil, valor negativo e que apresentam um padrão de histograma como o de I08 - ver Apêndice G em Barreto (2005): uma variável assumindo valor igual a log (IX + 1), no caso do valor do índice maior ou igual a 0 (zero), e assumindo valor 0 (zero) no caso de valor negativo, associada a uma outra variável binária assumindo valor 1 (um) se o valor do índice é negativo e 0 (zero) caso contrário.

3) Índices que possuem mais de 80% dos valores iguais a 0 (zero), apresentando padrão de histograma como o de I11 -m ver Apêndice G em Barreto (2005): duas variáveis binárias assumindo valores 0 (zero) e 1 (um), uma para captar valores positivos e outra para valores negativos.

A opção pelo logaritmo natural no que concerne as transformações é bastante natural, como ponderado em Barbetta (2001, p. 299): "A transformação logarítmica aumenta as distâncias entre os valores pequenos e reduz as distâncias entre os valores grandes, tornando distribuições assimétricas de cauda longa à direita em distribuições aproximadamente simétricas". Ademais, segundo Arino e Frenses (2000): "Utilizar a transformação logaritmo natural (log) antes da formulação de um modelo econométrico auxilia na redução do impacto dos outliers, faz com que as primeiras diferenças tornem-se taxas de crescimento e reduz a frequentemente observada variância crescente em séries temporais".

Dessa forma, as três diretrizes de tranformação relacionadas acima foram designadas e aplicadas aos índices - ver Apêndice F em Barreto (2005). Para um melhor aproveitamento das informações disponíveis, no caso dos índices abrangidos pelas diretrizes "2" e "3" acima, foi-lhes também testada a transformação mais geral prescrita pela diretriz "1". Adicionalmente, para o caso dos índices enquadrados na transformação "3" e que se caracterizam concomitantemente por não assumirem, em teoria, valores negativos, testou-se ainda a diretriz de transformação "2", estando claro que apenas uma das transformações, no caso a de melhor desempenho em termos da contribuição do índice transformado para o ajuste dos modelos, foi adotada nos modelos finais.

Assim, por convenção, denota-se a atribuição e aplicação da diretriz "1" em I04 por LI04A. A aplicação da segunda diretriz em I05 é denotada por uma variável codificada como LI05B e mais uma variável indicadora binária LI05NEG. Já a aplicação da diretriz "3" em I29 resulta em duas variáveis indicadoras binárias, I29NEG e I29POS. Este padrão de codificação foi o utilizado para todos os índices econômico-fiscais de trabalho.

Quanto aos valores contábeis brutos, como eles tendem a ser altamente correlacionados entre si, optou-se por uma análise em componentes principais (ACP) visando à redução de dimensionalidade de variáveis. O resumo dos resultados dessa análise (dez autovetores, assim como a variância explicada pelas componentes principais) está apresentado no Apêndice H em Barreto (2005). Foram utilizadas na modelagem as quatro primeiras componentes principais, denotadas por ZPC1,...,ZPC4, que explicam 97,22% da inércia da nuvem de pontos contábeis.

4.2 Estrutura hierárquica em unidades

A hierarquia dos MLGH foi estruturada de acordo com a mencionada classificação CNAE-Fiscal

de atividades econômicas de contribuintes PJ. A CNAE-Fiscal é um instrumento de padronização nacional dos códigos de atividade econômica utilizados pelos diversos órgãos de administração tributária no Brasil, tanto Federal quanto estaduais, cuja estrutura hierárquica pode ser visualizada pela Tabela 1:

Table 1: Estrutura Hierárquica da CNAE-Fiscal

Unidades	Nível	Agrupamentos	Identificação
Seção	Primeiro	17	Código Alfabético - 1 Digito
Divisão	Segundo	59	Código Numérico - 2 Digitos
Grupo	Terceiro	217	Código Numérico - 3 Digitos (*)
Classe	Quarto	563	Código Numérico - 4 Digitos (*)
Subclasse	Quinto	1.094	Código Numérico - 7 Digitos (*)

Obs: Os códigos assinalados com (*) estão integrados no código imediatamente anterior.

Fonte: SRF (2004).

Seguindo a Tabela 1, neste trabalho o MLGH teve seu nível 2 estruturado e identificado pelos Grupos CNAE-Fiscal, na medida em que, para os dados de trabalho, a utilização das Seções agregaria contribuintes bastante distintos na mesma unidade, enquanto que a estruturação em Subclasses reduziria drasticamente os tamanhos de amostra intra-unidades.

4.3 Procedimentos de Amostragem

De forma a avaliar a capacidade preditiva do MLGH estimado para prever o comportamento tributário dos contribuintes, deve-se ser extraída uma amostra de dados específica para validação, e seu tamanho de amostra foi definido à luz de alguns conceitos importantes em termos de avaliação da capacidade preditiva de modelos estatísticos: Sensitividade (S) e Especificidade (E). Tanto a sensitividade quanto a especificidade podem, então, ser abstraídas como duas proporções populacionais (P) a serem estimadas com base em amostras de validação extraídas a partir da divisão da população em dois estratos demarcando os elementos que possuem e os que não possuem a característica de interesse.

Como é desejável que após a extração das amostras de validação seja preservada na remanescente amostra de estimação a proporção populacional original de ocorrência da característica de interesse (de forma a se dispor, após o processo de extração, de amostras aleatórias estratificadas proporcionalmente

em relação à população objetivo, em ambos os subconjuntos), deve-se inicialmente proceder a uma estratificação da população objetivo em termos da ocorrência/não ocorrência da característica de interesse para que, subsequentemente, seja definida a fração amostral (f) a ser aplicada aos estratos.

Para definir tal fração, deve ser lembrado que o objetivo principal aqui é estimar um MLGH, a partir de 12.227 registros de contribuintes auditados (dentre os quais 8.073 deles foram identificados como infratores), de forma a proporcionar a classificaÇão futura de contribuintes como infratores/não infratores. É claro que o tamanho da futura população não é conhecido, e, portanto, o mais cauteloso neste momento é considerar a população futura como infinita. Sabe-se que, na prática, podem ser utilizadas as equações que regem a amostragem aleatória simples com reposição para o caso de uma amostragem aleatória simples sem reposição aplicada a grandes populações, como é o caso. Com isso, um tamanho de amostra de validação conservativo, para um erro de estimação de no máximo 5 pontos percentuais para mais ou para menos, e considerando-se um intervalo de confiança de 95% (neste ponto se está utilizando do resultado proporcionado pelo Teorema do Limite Central, qual seja, de que a distribuição amostral de $\hat{P}$ aproxima-se da Normal, à medida que o tamanho de amostra aumenta), é dado por:

$$n = \frac{PQ \cdot 1,96^2}{E^2} = \frac{0,25 \cdot 1,96^2}{0,05^2} = 384,16 \approx 385 \tag{4.1}$$

where n: tamanho de amostra, $Q = (1 - P)$.

Com o resultado em (4.1) f pode ser prontamente obtida a partir do menor estrato (4.154 contribuintes não apresentando a característica de interesse), o qual, naturalmente, rege a definição de f:

$$f = \frac{385}{4.154} = 0,0927 \tag{4.2}$$

Com o advento de f, as amostras de validação podem ser completamente especificadas: 385 contribuntes para estimar a especificidade e $751 = (f \cdot 8.073)$ contribuintes para fins da estimação da sensitividade do modelo. Consequentemente, ao final do processo de amostragem do conjunto de validação, terminou-se dispondo de conjunto (amostra) de estimação contando com 11.091 registros de contribuintes PJ.

5 Estimação do modelo e aplicação à seleção de contribuintes para fiscalização

5.1 Modelo nulo

O modelo nulo é o modelo linear hierárquico mais simples que pode ser estimado, na medida em que ele não possui preditoras em nenhum de seus níveis. Não obstante essa simplicidade estrutural, ele é útil para a avaliação de uma importante informação preliminar: a avaliação da variabilidade da resposta, em relação à variabilidade total, devida a cada um dos níveis hierárquicos. Em MLH, o modelo nulo permite calcular o coeficiente de correlação intra-classe, que representa de forma exata a proporção da variabilidade total devida ao nível 2. Contudo, em MLGH, em razão de uma heterocedasticidade de variância no nível 1, este coeficiente não produz mais um resultado único em termos de proporção da variabilidade total. Entretanto, Goldstein (2003) apresenta alguns procedimentos que permitem a avaliação aproximada desta proporção para MLGH. Portanto, por meio da utilização do Método da Linearização, pôde ser avaliado que aproximadamente 2.75% da variabilidade total da resposta, ou seja, probabilidade do contribuinte ser infrator, deve-se exclusivamente ao nível 2, grupos praticando distintas atividades econômicas, e isso confirma a hipótese prévia de existência de correlação de medidas internamento ao mesmo setor econômico.

5.2 Variáveis candidatas

Foram utilizadas como variáveis candidatas de nível 1 os 33 índices econômico-fiscais transformados, a variável indicadora "rural" (com o objetivo de captar o efeito do contribuinte ter preenchido o Anexo de Atividade Rural em sua DIPJ) e as 4 primeiras componentes principais advindas da ACP procedida a partir dos 10 valores contábeis brutos.

Em relação às variáveis candidatas de nível 2, o objetivo em vista é estimar um modelo a partir de informações não sigilosas e de fácil acesso, conforme já explanado.

Assim, a primeira considerada representa uma agregação geral para a natureza principal da atividade praticada pelos distintos grupos CNAE-Fiscal, sendo parametrizada como uma variável categorizada com 4 classes possíveis, representada por três variáveis indicadoras, sendo o setor agregado de serviços a referência em termos de parametrização. A codificação dessas indicadoras pode ser visualizada pela

tabela a seguir.

Table 2: Natureza da Atividade Agregada

Variável Indicadora	Descrição da Natureza da Atividade Agregada	Valores
agropecagr	agricultura, pecuária, silvicultura e exploração florestal agregadas	zero ou um
indagr	indústria agregada	zero ou um
comagr	comércio agregado	zero ou um

Importa dizer que como esta última variável é eminentemente estática no tempo, ela foi utilizada isoladamente em um modelo estimado especificamente com objetivo descritivo instantâneo (ano de 1999), e não para futura classificação de contribuintes, haja vista que neste último caso estar-se-ia acolhendo um determinismo setorial econômico não garantido ao longo do tempo.

Outra variável de nível 2 considerada é a ausência ou presença de exposição à economia informal. É natural que algumas atividades econômicas estejam mais expostas à informalidade que outras e, para avaliar este efeito nas respostas esperadas para os modelos, utilizou-se como fonte de informação a pesquisa realizada pelo IBGE sobre o assunto, qual seja, a pesquisa Economia Informal Urbana, de 1997. O IBGE detectou que, naquele ano, as seguintes divisões setoriais da economia eram afetadas pela concorrência representada pela economia informal: Comércio de mercadorias; Serviços de reparação, pessoais, domiciliares e de diversões; Indústrias da construção; Indústrias de transformação e extrativa mineral; Serviços técnicos e auxiliares; Serviços de alojamento e alimentação; e Serviços de transporte (IBGE, 1997). Assim, pode-se definir uma variável binária codificada como "inform" assumindo valores 1 (one) ou 0 (zero) respectivamante para os casos de exposição e não exposição do grupo CNAE-Fiscal à economia informal.

Quanto à situação de crescimento ou retração anual das atividades econômicas praticadas pelos diversos setores econômicos brasileiros, o IBGE detectou taxas acumuladas de variação percentual do PIB para os anos 1998 e 1999, que são apresentadas na Tabela 3.

Com o subsídio da Tabela 3, é possível construir duas variáveis numéricas contínuas codificadas por "VarPerc98" e "VarPerc99" que sintetizam a variação percentual anual acumulada para o PIB dos diversos grupos CNAE-Fiscal.

Table 3: Variação percentual do PIB (1998 e 1999)

Subsetor	1998	1999	Subsetor	1998	1999
AGROPECUÁRIA			SERVIÇOS		
Lavouras	-0.23	11.26	Comércio	-3.39	0.50
Extrativa Vegetal	-7.27	1.45	Transporte	7.18	-0.13
Produção Animal	3.86	5.73	Comunicações	6.38	8.64
INDÚSTRIA			Instituições Financeiras	0.15	0.82
Extrativa Mineral	9.04	0.85	Outros Serviços	-1.10	-0.34
Transformação	-3.29	-1.25	Aluguel de Imóveis	2.10	1.95
Construção	1.70	-3.61	Administração Pública	1.28	0.67
SIUP	4.16	1.97			

Obs: SIUP são Serviços Industriais de Utilidade Pública.

Fonte: IBGE (1998) e IBGE (1999).

Neste trabalho também foram consideradas as exportações, e a Fundação Estudos de Comércio Exterior (FUNCEX) registrou atividades de exportação, em 1999, para os setores econômicos relacionados na Tabela 4.

Com isso, pode-se codificar uma variável binária denominada de "export" que assume valores 0 (zero) ou 1 (um), este no caso do grupo CNAE-Fiscal estar contido em um dos subsetores relacionados na Tabela 4.

Finalmente, considerou-se também como candidatas de nível 2 as médias de grupos CNAE-Fiscal padronizadas para os 10 valores contábeis brutos, denotadas por: ZLOPER, ZLBRUTO, ZRECLIQ, ZLREAL, ZLACSLL, ZBCCSLL, ZRBPISP,ZBCPISP, ZRBCOF, ZBCCOF.

5.3 Variable selection process

Para selecionar as variáveis de nível 1, o procedimento adotado foi o de inclusão uma a uma das variáveis no modelo (método stepwise forward), em diversas rodadas, até que todas fossem testadas, de forma a avaliar suas contribuições individuais para a parcela de variância explicada. Isso se deu pelo teste para efeito fixo, o general linear hypothesis test (GLHT). Na prática o procedimento adotado

Table 4: Setores Exportadores em 1999

Subsetor	Subsetor	Subsetor
Agropecuário	Extrativo Mineral	Minerais não Metálicos
Siderurgia	Metalurgia de Não Ferrosos	Outros Produtos Metalúrgicos
Máquinas e Tratores	Material Elétrico	Equipamentos Eletrônicos
Veículos Automotores	Peças e outros Veículos	Madeira e Mobiliário
Celulose, Papel e Gráfica	Borracha	Elementos Químicos
Refino de Petróleo e Petroquímicos	Químicos Diversos	Têxtil
Calçados Couros e Peles	Café	Beneficiamento de Vegetais
Abate Animais	Açúcar	Óleos Vegetais
Outros Produtos Alimentares	Indústrias Diversas	

Source: (FUNCEX, 1999).

busca avaliar as variáveis candidatas, com base no GLHT, considerando-as isoladamente no modelo. Ao final da primeira rodada forma-se um modelo provisório inicial incluindo as de melhor desempenho no GLHT. Em uma segunda rodada, testam-se individualmente novas variáveis candidatas, mas desta feita no modelo provisório inicial, e, ao final da segunda rodada, incluem-se novamente as de melhor desempenho, formando-se, então, um segundo modelo provisório. As rodadas se sucedem até que se convirja para um modelo considerado final em que não há mais variáveis candidatas, mas sim incluídas ou descartadas do modelo.

Já no caso das variáveis de nível 2 (grupos CNAE-Fiscal), o processo adotado envolveu o mesmo procedimento stepwise forward, mas partindo-se de um modelo de nível 1 já especificado. Ressalte-se que como o número de grupos CNAE-Fiscal de trabalho foi de 202, utilizou-se, em adição ao GLHT, também o teste t para verificar a significância estatística em relação à inclusão de cada uma destas variáveis no modelo.

É importante mencionar que em todo esse processo de seleção de variáveis, cuidou-se pela análise da estabilidade (desejável) dos sinais dos coeficientes de regressão a cada novo ingresso de variáveis. Além disso, é preciso que se registre que o processo de seleção de variáveis utilizado, na verdade, é um dos que se demonstra factível, já que, de posse das 50 variáveis disponíveis para os níveis 1 e 2, o número de

possíveis MLGH eleva-se a incríveis $2^{50} = 1{,}1259$ x 10^{15} modelos, e a estimação e avaliação de todo esse conjunto de possibilidades, em termos de suas possíveis previsões, está claramente fora do escopo desta pesquisa.

5.4 Estimativas, métricas, algoritmos e programas utilizados

Em se tratando de executar previsões futuras individuais (predições), a partir de informações relativas a observações não participantes da amostra de estimação, Afshartous e de Leeuw (2003) demonstraram que, em MLH, a regra de predição a partir da utilização o vetor ótimo de estimativas empíricas de Bayes da equação (2.14) é a de melhor desempenho em termos do Erro Quadrático Médio Preditivo, suplantando as predições obtidas a partir das estimativas de mínimos quadrados ordinários (MQO) e as obtidas por meio da grande média estimada com base no vetor de efeitos fixos, respectivamente $\hat{\boldsymbol{\beta}}_j$ e $\mathbf{W}_j\hat{\boldsymbol{\gamma}}$. Portanto, todas as estimativas apresentadas nas próximas seções são unit-specific, haja vista serem estas as requisitadas para a obtenção do vetor ótimo em (2.14).

Para produzir essas estimativas foi adotado o algoritmo Penalized Quasi-Likelihood (PQL) - o qual envolveu um processo de máxima verossimilhança internamente em seus passos iterativos - via programa HLM 5.0 e também pelo R Program for Statistical Computing (ver apêndice L em Barreto (2005), para acessar a lista de todos os comandos do R utilizados para obtenção das inferencias e estimativas de interesse). A partir dos resultados obtidos, pôde-se perceber que as estimativas entre os dois programas só diferem a partir da terceira casa decimal para efeitos fixos e componentes de variância.

Por fim, cabe ainda registrar que as estimativas foram obtidas a partir da utilização das variáveis em sua métrica natural, a menos dos 10 valores contábeis brutos utilizados na ACP, já que todas as variáveis podem em teoria assumir, como de fato se verifica, valores iguais a 0 (zero).

5.5 Modelo interpretativo

O modelo final, de cunho exclusivamente interpretativo, obtido a partir do processo stepwise forward é formalizado a seguir.

Nível 1:

$$\begin{aligned}\eta_{ij} &= \beta_{0j} + \beta_{1j} \cdot LI02A_{ij} + \beta_{2j} \cdot LI04A_{ij} + \beta_{3j} \cdot LI10A_{ij} + \\ &\quad \beta_{4j} \cdot LI11A_{ij} + \beta_{5j} \cdot LI13B_{ij} + \beta_{6j} \cdot LI05A_{ij} +\end{aligned}$$

$$\beta_{7j} \cdot LI24A_{ij} + \beta_{8j} \cdot LI22A_{ij} + \beta_{9j} \cdot LI26B_{ij} + \tag{5.1}$$

$$\beta_{10j} \cdot LI25A_{ij} + \beta_{11j} \cdot LI30A_{ij} + \beta_{12j} \cdot ZPC1_{ij} +$$

$$\beta_{13j} \cdot ZPC2_{ij}$$

Nível 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} \cdot agropecagr_j + \gamma_{02} \cdot indagr_j + u_{0j} \tag{5.2}$$

$$\beta_{pj} = \gamma_{p0}, \quad p = 1, ..., 13 \tag{5.3}$$

onde η_{ij}: log-odds ou logito de presença (ocorrência) de infração fiscal para o i-ésimo contribuinte do grupo j; β_{0j}: log-odds de presença (ocorrência) de infração fiscal esperado para um contribuinte do grupo j apresentando valor zero para todas as covariáveis de nível 1; β_{pj}: mudança esperada no log-odds de presença (ocorrência) de infração fiscal, dado um incremento unitário na p-ésima covariável de nível 1, *ceteris paribus*; γ_{00}: valor esperado de β_{0j} (intercepto) para os grupos pertencentes aos setores de comércio e serviços agregados; γ_{01}: diferença entre os valores esperados dos β_{0j} (interceptos) de grupos agropecuários, silvicultores e florestais agregados e de comércio e serviços agregados; γ_{02}: diferença entre os valores esperados dos β_{0j} (interceptos) de grupos industriais e de comércio e serviços agregados; u_{0j}: efeito aleatório associado ao grupo j; τ_{00}: variância dos interceptos β_{0j} entre os grupos, corrigida por $agropecagr_j$ and $indagr_j$. Suposições do modelo: $u_{0j} \sim N(0, \tau_{00})$ e u_{0j}'s independentes.

Todas as iterações e estatísticas obtidas para este modelo interpretativo podem ser acessadas no Apêndice K em Barreto (2005). É importante ressaltar que, para esse modelo, testes de coeficientes aleatórios individuais não rejeitaram a hipótese nula para todos os coeficientes de nível 1, à exceção do intercepto aleatório. Ao lado disso, por meio de um GLHT verificou-se que o coeficiente para "comagr" não possui efeito significativamente distinto da referência em termos de parametrização, o setor de serviços agregado, e com isso pôde ser retirado do modelo. Na Tabela 5 podem-se observar as estimativas finais para o modelo interpretativo:

Como o componente de variância estimado para o modelo final de nível 1 (ainda sem o ingresso de variáveis de nível 2) equivaleu a $\tau_{00} = 0,03887$, a partir do componente de variância estimado apresentado na Tabela 5 pode-se avaliar que o modelo interpretativo explica 46.82% da variabilidade de nível 2 devida aos grupos CNAE-Fiscal dedicando-se a distintas atividades econômicas. Esse resultado reconfirma a hipótese prévia de trabalho relacionada à existência de observações correlacionadas internamente em um

Table 5: Estimativas para o Modelo Interpretativo

Parâmetro	Estimativa	Erro Padrão	N. Descritivo	Parâmetro	Estimativa	Erro Padrão	N. Descritivo
γ_{00}	0,602977	0,091643	$< 0{,}001$	γ_{01}	-0,709302	0,136248	$< 0{,}001$
γ_{02}	-0,150765	0,057158	0,009	γ_{10}	-0,064246	0,014663	$< 0{,}001$
γ_{20}	-0,046725	0,014980	0,002	γ_{30}	0,034447	0,012501	0,006
γ_{40}	0,031471	0,015038	0,036	γ_{50}	-0,122959	0,026396	$< 0{,}001$
γ_{60}	0,148317	0,033053	$< 0{,}001$	γ_{70}	0,167816	0,025241	$< 0{,}001$
γ_{80}	-0,029533	0,009699	0,003	γ_{90}	-0,067432	0,021778	0,002
γ_{100}	-0,069782	0,025502	0,007	γ_{110}	0,015905	0,007760	0,040
γ_{120}	0,077412	0,018593	$<0{,}001$	γ_{130}	-0,029637	0,013990	0,034
τ_{00}	0,020670		$< 0{,}001$				

mesmo setor econômico.

A partir da Tabela 5, pode-se inicialmente perceber que o log-odds de presença de infração fiscal para um contribuinte pertencente ao grupo típico (grupo em que $u_{0j} = 0$), contando com $agropecagr_j = indagr_j = 0$, o que, segundo a parametrização adotada o insere internamente aos grupos de comércio e serviços agregados, e também apresentando valor 0 (zero) para todas as covariáveis de nível 1, situação esta nomeada a partir deste ponto, visando a simplicidade de explanação, por "situação nula", equivale a $0,602977$, implicando numa probabilidade de ser infrator de 64,63%.

O fato do grupo CNAE-Fiscal ser pertencente ao setor industrial agregado associa-se a um log-odds de infração fiscal inferior, já que $\hat{\gamma}_{02} = -0,150765$, à razão entre odds de 0,86, ceteris paribus; o que indica que, partindo-se da situação nula, a probabilidade do contribuinte ser infrator reduz-se para 61,11%. Já a situação do grupo estar englobado nos setores de agricultura, pecuária, silvicultura e exploração florestal agregadas relaciona-se a um log-odds ainda menor, pois $\hat{\gamma}_{01} = -0,709302$, à razão entre odds de 0,49, ceteris paribus; e, partindo-se da situação nula, a probabilidade do contribuinte ser infrator passa a ser de 47,34%, portanto 26,75% inferior.

Estes resultados indicam que três contribuintes absolutamente idênticos em tudo, a menos de pertencerem a grupos CNAE-Fiscal de setores agregados distintos, apresentam log-odds de presença de infração e respectivas probabilidades de serem infratores de magnitudes diferentes, por ordem decrescente

de probabilidades, os contribuintes de: comércio e serviços agregados; indústria; e agricultura, pecuária, silvicultura e exploração florestal agregadas; ou seja, em 1999, um contribuinte interno a um grupo participante dos setores de comércio e serviços agregados possuía mais chances de ser infrator do que contribuintes que lhe eram idênticos individualmente, mas que pertenciam a grupos dos demais setores agregados.

5.6 Modelo preditivo

O modelo final preditivo obtido a partir do mencionado processo stepwise forward assumiu a formalização a seguir.

Nível 1:

$$\begin{aligned} \eta_{ij} \quad = \quad & \beta_{0j} + \beta_{1j} \cdot LI02A_{ij} + \beta_{2j} \cdot LI04A_{ij} + \beta_{3j} \cdot LI10A_{ij} + \\ & \beta_{4j} \cdot LI11A_{ij} + \beta_{5j} \cdot LI13B_{ij} + \beta_{6j} \cdot LI05A_{ij} + \\ & \beta_{7j} \cdot LI24A_{ij} + \beta_{8j} \cdot LI22A_{ij} + \beta_{9j} \cdot LI26B_{ij} + \\ & \beta_{10j} \cdot LI25A_{ij} + \beta_{11j} \cdot LI30A_{ij} + \beta_{12j} \cdot ZPC1_{ij} + \\ & \beta_{13j} \cdot ZPC2_{ij} \end{aligned} \tag{5.4}$$

Nível 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} \cdot export_j + \gamma_{02} \cdot VRPERC99_j + u_{0j} \tag{5.5}$$

$$\beta_{pj} = \gamma_{p0}, \quad p = 1, ..., 13 \tag{5.6}$$

onde η_{ij}: log-odds ou logito de presença (ocorrência) de infração fiscal para o i-ésimo contribuinte do grupo j; β_{0j}: log-odds de presença (ocorrência) de infração fiscal esperado para um contribuinte do grupo j apresentando valor 0 (zero) para todas as covariáveis de nível 1; β_{pj}: mudança esperada no log-odds de presença (ocorrência) de infração fiscal, dado um incremento unitário na p-ésima covariável de nível 1, *ceteris paribus*; γ_{00}: valor esperado de β_{0j} (intercepto) para um grupo j contando com $export_j = VRPERC99_j = 0$; γ_{01}: diferença entre os valores esperados dos β_{0j} (interceptos) de grupos exportadores e não exportadores, *ceteris paribus*; γ_{02}: diferença esperada no valor de β_{0j} (intercepto) para um incremento unitário em $VRPERC99_j$, *ceteris paribus*; u_{0j}: efeito aleatório associado ao grupo j; τ_{00}: variância dos β_{0j} (interceptos) entre os grupos, corrigida por $export_j$ e $VRPERC99_j$. Suposições do modelo: $u_{0j} \sim N(0, \tau_{00})$ e u_{0j}'s independentes.

Todas as iterações e estatísticas obtidas para este modelo preditivo podem ser acessadas no Apêndice K em Barreto (2005). Além disso, para esse modelo, testes de coeficientes aleatórios individuais não rejeitaram a hipótese nula para todos os coeficientes de nível 1, à exceção do intercepto aleatório. Na Tabela 6 podem-se observar as estimativas finais para o modelo preditivo.

Table 6: Estimativas para Modelo Preditivo

Parâmetro	Estimativa	Erro Padrão	N. Descritivo	Parâmetro	Estimativa	Erro Padrão	N. Descritivo
γ_{00}	0,546603	0,089167	$<$ 0,001	γ_{01}	-0,202106	0,057097	0,001
γ_{02}	-0,038157	0,012655	0,003	γ_{10}	-0,063924	0,014643	$<$ 0,001
γ_{20}	-0,049270	0,014937	0,001	γ_{30}	0,035916	0,012475	0,004
γ_{40}	0,031971	0,015043	0,033	γ_{50}	-0,125664	0,026420	$<$ 0,001
γ_{60}	0,147289	0,033066	$<$ 0,001	γ_{70}	0,167257	0,025315	$<$ 0,001
γ_{80}	-0,029840	0,009714	0,003	γ_{90}	-0,063283	0,021819	0,004
γ_{100}	-0,062478	0,025307	0,014	γ_{110}	0,018934	0,007812	0,016
γ_{120}	0,079879	0,018748	$<$ 0,001	γ_{130}	-0,029927	0,013984	0,032
τ_{00}	0,024860		$<$ 0,001				

Como o componente de variância estimado para o modelo final de nível 1 (ainda sem o ingresso de variáveis de nível 2) equivaleu a $\tau_{00} = 0,03887$, a partir do componente de variância estimado da Tabela 6 pode-se avaliar que o modelo preditivo explica 36,04% da variabilidade de nível 2 devida aos grupos CNAE-Fiscal dedicando-se a distintas atividades econômicas.

A Tabela 6 mostra que um contribuinte estando na situação nula, ou seja, pertencendo ao grupo típico (grupo em que $u_{0j} = 0$), contando com $export_j = VRPERC_j = 0$, e também apresentando valor 0 (zero) para todas as covariáveis de nível 1, apresenta um log-odds de presença de infração de $0,546603$, implicando numa probabilidade de ser infrator de 63,33%.

Não obstante o enfoque preditivo, este último modelo também fornece interpretações a partir de suas variáveis de nível 2. Pode-se observar que um incremento unitário positivo em VrPerc99 está associado a um log-odds de presença de infração fiscal inferior, pois $\hat{\gamma}_{02} = -0,03857$, à razão entre odds de $exp(-0,03857) = 0,96$, *ceteris paribus*, ou seja comparando-se dois contribuintes similares em tudo, mas pertencendo a grupos CNAE-Fiscal diferindo uma unidade em termos de VrPerc99, pode-se esperar

que o odds de presença de infração para o contribuinte do grupo de maior VrPerc99 equivalha a 0,96 vezes o odds de presença de infração para o contribuinte do grupo de menor VrPerc99. Com o advento da mencionada variação percentual unitária positiva em VrPerc99, e partindo-se da situação nula, a probabilidade do contribuinte ser infrator reduz-se a 62,43%. Um incremento adicional de nove unidades percentuais a VrPerc99 - correspondendo a um grupo CNAE-Fiscal apresentando variação anual positiva de 10% no PIB - e a razão entre odds passa a 0,68, ceteris paribus, e a probabilidade do contribuinte ser infrator passa a ser de 54,01%, ou seja, 14,72% inferior, *ceteris paribus*.

Já para o caso de export, esta assumindo valor unitário acarreta redução no log-odds de presença de infração, uma vez que $\hat{\gamma}_{01} = -0,202106$, à razão entre odds de 0,82, *ceteris paribus*; o que implica que, partindo-se da situação nula, a probabilidade de um contribuinte pertencente a um grupo CNAE-Fiscal exportador ser infrator reduz-se a 58,53%, tudo o mais mantido constante. Em resumo, em 1999, um contribuinte de um grupo CNAE-Fiscal exportador apresentava menor probabilidade de ser infrator, odds 18% inferior em relação à referência (grupos não exportadores), *ceteris paribus*.

5.7 Avaliação, aplicação e validação dos modelos estimados

Em termos de avaliação dos pressupostos do modelo, deve-se mencionar que o teste de Shapiro-Wilk, apresentando nível descritivo de 0,0902, não rejeitou formalmente a suposição de normalidade dos resíduos de nível 2, cuja ordenação em relação aos quantis Normais teóricos pode ser observada na Figura 1.

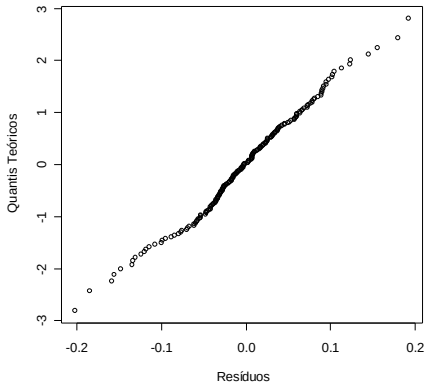

Figura 1: Gráfico de probabilidade Normal para resíduos de nível 2 do modelo preditivo

5.7.1 *Desempenho das previsões para observações conhecidas*

No que concerne a previsão de uma resposta binária em modelos estatísticos, isso envolve basicamente a utilização de uma determinada regra de classificação operacional, e.g., probabilidade predita maior ou igual a 50% classificada como 1, caso contrário classificada como 0, o que claramente requisita a definição de um ponto de corte probabilístico por parte do analista - no caso exemplificado ele seria de 50%. Um útil instrumento de análise da capacidade preditiva destes modelos é a curva Receiver Operating Characteristic (ROC), que, segundo Agresti (2002), relaciona a sensitividade em função de (1 - especificidade) para todos os possíveis pontos de corte. No caso da pesquisa enfocada neste artigo, a área sob a curva para o modelo preditivo foi de 0,61, o que indica que o modelo fornece previsões superiores à mera aleatoriedade na decisão fiscal.

Portanto, para prever a partir de observações desconhecidas, uma regra de classificação tem de ser definida com base na amostra de estimação (observações conhecidas). Para tanto, seguindo em linhas gerais o prescrito por Neter et al. (1996) para os casos de amostras aleatórias estratificadas proporcionalmente à população, como é o caso, haja vista os procedimentos de amostragem descritos na seção 4.3, a abordagem aqui adotada para a escolha da regra de classificação envolve, utilizando-se da amostra de estimação, avaliar o desempenho de diversos pontos de corte probabilísticos possíveis à luz de uma ou mais medidas de validade interna do modelo, por exemplo, sensitividade, especificidade ou taxa de acertos, sendo que uma boa aproximação para os valores de pontos de corte a serem testados pode ser obtida entre os que se situam próximos à proporção populacional de positivos. Após isso, a regra considerada como a de melhor desempenho pelo analista é, então, submetida à amostra de validação para avaliação da capacidade preditiva do modelo em relação às novas observações.

Lembrando que a amostra de estimação é formada por 11.091 contribuintes, dentre os quais 7.322 deles foram autuados como infratores (casos positivos), a Tabela 7 apresenta os resultados, em termos de Sensitividade (S), Especificidade (E) e Valor Preditivo Positivo (VPP), obtidos a partir da adoção de diversos pontos de corte às probabilidades preditas pelo modelo para a amostra de estimação.

A partir da Tabela 7, o ponto de corte probabilístico escolhido para o caso do modelo preditivo aplicado às novas observações é 0,620, pois pontos de corte inferiores elevam a sensitividade, mas levam o VPP a ultrapassar o limiar abaixo de 70%, enquanto que pontos de corte superiores reduzem rapidamente a sensitividade.

Table 7: Pontos de Corte para o Modelo Preditivo (Amostra de Estimação)

Ponto de Corte	S (%)	E (%)	VPP (%)	Ponto de Corte	S (%)	E (%)	VPP (%)
0,615 (61,5%)	84,69	29,18	69,90	0,636 (63,6%)	76,78	38,47	70,80
0,620 (62,0%)	83,13	31,04	70,08	0,642 (64,2%)	74,49	41,22	71,11
0,622 (62,2%)	82,24	31,92	70,12	0,650 (65,0%)	70,49	45,48	71,52
0,627 (62,7%)	80,55	33,80	70,27	0,655 (65,5%)	67,97	48,53	71,95
0,630 (63,0%)	79,28	35,23	70,40				

5.7.2 *Desempenho das previsões para novas observações*

A aplicação do ponto de corte escolhido na subseção anterior (0,620) como regra de classificação das probabilidades preditas pelo modelo preditivo para a amostra de validação - 1.136 registros de contribuintes dentre os quais 751 são positivos - implica nos seguintes valores para a sensitividade (S), especificidade (E) e VPP:

Table 8: Ponto de Corte para o Modelo Preditivo (Amostra de Validação)

Ponto de Corte	$\hat{S}(\%)$	$\hat{E}(\%)$	$V\hat{P}P$ (%)
0,620 (62,0%)	80,16	33,51	70,16

A Tabela 8 mostra que o modelo preditivo, em média, detecta 80,16% dos casos de infração; e que a cada 1.000 contribuintes por ele classificados como infratores, 702 confirmarão tal previsão.

Os procedimentos de amostragem apresentados na seção 4.3 permitem estabelecer um intervalo de 95% de confiança para uma determinada medida genérica (M) de validade interna relacionada ao modelo preditivo, estimada a partir das amostras de validação, o qual é dado por:

$$\left(\hat{M} - 1,96\sqrt{\frac{\hat{M}(1-\hat{M})}{n-1}}; \hat{M} + 1,96\sqrt{\frac{\hat{M}(1-\hat{M})}{n-1}}\right) \tag{5.7}$$

Aplicando (5.7), obtêm-se os seguintes intervalos de 95% de confiança para as medidas de validade interna do modelo preditivo:

$$\left(\hat{S} \pm 1.96 \cdot 0.014562\right) \Longrightarrow 77.30\% \leq S \leq 83.02\% \tag{5.8}$$

$$\left(\hat{E} \pm 1.96 \cdot 0.024088\right) \Longrightarrow 28.78\% \leq E \leq 38.24\% \tag{5.9}$$

$$\left(\hat{VPP} \pm 1.96 \cdot 0.013581\right) \Longrightarrow 67.49\% \leq VPP \leq 72.83\% \tag{5.10}$$

Os resultados do processo de validação indicam a estabilidade dos coeficientes estimados frente às novas observações, já que as medidas pontuais de validade interna estimadas para a amostras de validação (Tabela 8) não diferem sensivelmente em relação às medidas obtidas a partir da amostra de estimação (Tabela 7).

5.8 Comparação entre os desempenhos do método proposto e do atual da SRF

Este é o momento adequado para que seja efetuada uma comparação entre o desempenho da seleção de contribuintes efetuada com base no novo método de classificação proposto (tendo por base o modelo preditivo) frente ao método de seleção por cruzamento de informações utilizado pela SRF e descrito na introdução.

Inicialmente, deve-se salientar que, quanto ao método atual da SRF, não há informações disponíveis nas bases de dados de trabalho que possam aclarar os valores de Sensitividade e Especificidade obtidos por meio de sua utilização, isso porque os 157.709 (164.466 - 9.757 = 157.709) contribuintes que não foram selecionados para fiscalização não foram efetivamente fiscalizados e, portanto, não há como se apurar os falsos negativos e verdadeiros negativos proporcionados pela seleção efetuada com base nesse método. Contudo, na seção 4.1 foi mencionado que dentre os 9.757 contribuintes selecionados para fiscalização a partir da utilização da base de dados de cruzamento de informações da SRF, 8.728 deles encontram-se como auditados entre as 12.227 auditorias concluídas em relação às operações referentes ao ano-calendário 1999, exercício 2000. Sendo que, destes 8.728 contribuintes, 5.180 foram efetivamente autuados como infratores. Estes quantitativos fornecem exatamente os subsídios necessários para o cálculo do VPP do

método atual, qual seja, VPPatual = (5.180/8.728) = 59,35%.

Sintetizando, o modelo preditivo das equações (5.4), (5.5) e (5.6) apresenta uma estimativa pontual do VPP para novas observações cerca de 11 pontos percentuais superior à mesma medida de validade interna calculada para o método atual da SRF. Além disso, o próprio intervalo de 95% de confiança para o VPP do modelo preditivo apresenta limite inferior (67,49%) mais de oito pontos percentuais superior aos 59,35% obtidos como VPP do método atual. Portanto, não há dúvidas sobre a superioridade das previsões do modelo preditivo em relação às do método atual da SRF.

6 Conclusões

O objetivo geral da pesquisa enfocada neste artigo foi o de desenvolver e especificar um novo método de classificação de contribuintes pessoas jurídicas brasileiras, com base na previsão de seus comportamentos tributários, objetivando classificá-las como prioritárias ou não à fiscalização, representando na prática um processo de seleção de contribuintes para fiscalização. A opção por operacionalizar o classificador por meio dos MLGH deu-se à luz do reconhecimento de que qualquer estudo tributário que envolva a classificação de empresas como de relevante interesse fiscal deve levar em consideração e avaliar a correlação de medidas existente entre empresas de um mesmo setor de atividades econômicas; e de que essa correlação de medidas deve ser contabilizada adequadamente (de forma sistematizada).

Não obstante tais necessidades, a seção 2.1 demonstrou que essa correlação não vinha sendo considerada, até o momento, de forma sistematizada pelos estudos científicos abordando a seleção de contribuintes para fiscalização, mas sim de forma pontual e limitada, o que os torna válidos apenas para poucos setores da economia e, consequentemente, para apenas para alguns subgrupos de contribuintes.

De forma a enfrentar o problema do tempo de espera para deflagração das auditorias fiscais e privilegiar a agilidade de ação das agências tributárias, o novo método de classificação se fundamentou em uma tríade: utilização de MLGH de forma a levar em consideração a correlação intra-grupos, onde os grupos relacionam-se aos setores econômicos; utilização, como variáveis preditoras de nível 1, tão-somente das informações compulsoriamente prestadas pelo próprio declarante em sua DIPJ; consideração, como base para a formação de variáveis explicativas de nível 2, apenas de informações não sigilosas e de fácil e rápido acesso.

Essas diretrizes de modelagem que nortearam a especificação do modelo preditivo conferem uma maior agilidade de ação ao órgão tributário e proporcionam uma redução substancial no tempo de espera para a deflagração das ações fiscais, na medida em que a classificação de contribuintes, e consequentemente sua seleção, pode se dar tão logo as declarações de informação (no caso da SRF as DIPJ) sejam recebidas pela agência tributária.

As estimativas obtidas e apresentadas nas seções 5.1, 5.5 e 5.6 confirmaram a efetiva existência de medidas correlacionadas entre observações pertencentes ao mesmo setor econômico. E mais, a existência de variáveis de nível 2 (grupos CNAE-Fiscal) significativas para todos os modelos estimados mostrou que os MLGH levaram em consideração, ineditamente de forma sistematizada, a correlação de medidas existente entre contribuintes pertencentes a um mesmo grupo de atividades econômicas, assim como explicaram, também de forma inédita, parte da variabilidade (entre grupos econômicos) aferida a partir dessa sistematização.

A partir do modelo interpretativo da seção 5.5, viu-se que três contribuintes absolutamente idênticos em tudo, a menos de pertencerem a grupos CNAE-Fiscal de setores agregados distintos, apresentaram log-odds de presença de infração e respectivas probabilidades de serem infratores de magnitudes diferentes, por ordem decrescente de probabilidades, os contribuintes de: comércio e serviços agregados; indústria; e agricultura, pecuária, silvicultura e exploração florestal agregadas; ou seja, em 1999, um contribuinte interno a um grupo participante dos setores de comércio e serviços agregados possuía mais chances de ser infrator do que contribuintes que lhe eram idênticos individualmente, mas que pertenciam a grupos dos demais setores agregados.

A partir do modelo preditivo da seção 5.6, viu-se que um incremento unitário positivo em Vr-Perc99 está associado a um log-odds de presença de infração fiscal inferior; e com respeito à variável export, esta assumindo um valor unitário provoca uma redução no log-odds de presença de infração fiscal, ou seja, em 1999, um contribuinte de um setor exportador possuía menor probabilidade de ser infrator do que um contribuinte que lhe era idêntico em tudo, a menos de pertencer a um setor não exportador da economia.

Finalmente, em relação às previsões para novas observações, foi visto que o modelo preditivo detecta 80,16% dos infratores. Além disso, o valor estimado para o seu VPP pontual, (70,16%), é cerca de 11 pontos percentuais superior à mesma medida calculada para o método atual da SRF, e sua

superioridade em termos de desempenho é ratificada pelo fato de que o limite inferior do intervalo de 95% de confiança para o VPP do modelo preditivo (67,49%) é também superior ao VPP calculado para o método da SRF (59,35%). Em se tratando da fiscalização de grandes contribuintes que apuram o Lucro Real, esta diferença em termos de VPP entre os dois métodos pode representar até bilhões de reais a serem recuperados por meio das fiscalizações.

7 Referência Bibliográficas

Afshartous, D; Leeuw, J. de. (2003) Prediction in multilevel models. *Journal of Educational and Behavioral Statistics*, in Press.

Agresti, A. (2002) *Categorical data analysis.* New Jersey: John Wiley.

Andreoni, J. (1992) IRS as loan shark: tax compliance with borrowing constraints. *Journal of Public Economics, North-Holland*, **49**, 35-46.

Arino, M.A.; Frenses, P.H. (2000) Forecasting the Levels of vector autoregressive log-transformed time series. *International Journal of Forecasting*, **16**, 111- 116.

Barbetta, P.A. (1999) *Estatística Aplicada às Ciências Sociais.* Florianópolis: Editora da UFSC.

Barreto, A.S. (2005) Previsão de comportamento e classificação de contribuintes tributários: uma abordagem por modelos lineares generalizados hierárquicos (Tese de Doutorado em Engenharia de Produção) - Universidade Federal de Santa Catarina, UFSC, Florianópolis, Brasil. Disponível em: http://br.geocities.com/alexsbdr/

Bryk, A.S.; Raudenbush, S.W. (1992) *Hierarchical linear models: applications and data analysis methods.* Newbury Park: Sage Publications.

Fundação Estudos de Comércio Exterior, FUNCEX. (2002) Boletim Setorial Funcex, Ano VI, n. 3, jul./ago./set. Disposable in: http://www.funcex.com.br. Acesso em: the 5th of may, 2005.

Goldstein, H. (2003) *Multilevel statistical models.* London: Arnold.

Instituto Brasileiro de Geografia e Estatística, IBGE. (1997) Economia Informal Urbana. Disponível em: http://www.ibge.gov.br/home/estatistica/economia/ecinf/default.shtm Acesso em: 12 dez. 2004.

Instituto Brasileiro de Geografia e Estatística, IBGE. (1998) Indicadores IBGE - Produto Interno Bruto -

4º Trimeste de 1998. Disponível em: http://www.ibge.gov.br/home/estatistica/indicadores/pib/defaultcnt.shtm. Acesso em: 12 dez. 2004.

Instituto Brasileiro de Geografia e Estatística, IBGE. (1999) Indicadores IBGE - Produto Interno Bruto - 4º Trimeste de 1999. Disponível em: http://www.ibge.gov.br/home/estatistica/indicadores/pib/defaultcnt.shtm. Acesso em: the 12 december, 2004.

Joulfaian, D.; Hider, M. (1998) Differential taxation and tax evasion by small business. *National Tax Journal*, **51**, 675-687.

Mills, L.F. (1996) Corporate tax compliance and financial reporting. *National Tax Journal*, **49**, 421-435.

Murray, M.N. (1995) Sales tax compliance and audit selection. *National Tax Journal*, **48**, 515-530.

Neter, J.; Kutner, M.H.; Nashtsheim, C.J.; Wasserman, W. (1996) *Applied linear regression models.* Times Mirror Higher Education Group.

Raudenbush, S.W.; Bryk, A.S. (2002) *Hierarchical linear models: applications and data analysis methods.* Thousand Oaks: Sage Publications.

Davidson, R.; Mackinnon, J. G. (1993) *Estimation and inference in econometrics.* Oxford: Oxford University Press.

Reinganum, J.F.; Wilde, L.L. (1988) A note on enforcement uncertainty and taxpayer compliance. *The Quarterly Journal of Economics*, **103**, 793-798.

Secretaria da Receita Federal.(2004) Classificação Nacional de Atividade Econômico-Fiscal do Contribuinte. Disponível em: http://www.receita.fazenda.gov.br/PessoaJuridica/CNAEFiscal/cnaef.htm Acesso em: 12 dez. 2004.