Maiores contribuintes: uma abordagem setorial baseada na análise de *clusters*

3º Lugar

CHRISTIAN MONTENEGRO JARDIM*
FRANCISCO ASSIS CORRÊA BARBOSA JÚNIOR**

* Graduado em Engenharia Eletrônica e Direito – Ita e Univap
Auditor-Fiscal da Receita Federal do Brasil
Delegacia da Receita Federal
São José dos Campos – São Paulo

** Mestre em Economia Empresarial – Universidade Cândido Mendes
Auditor-Fiscal da Receita Federal do Brasil
Delegacia da Receita Federal
São José dos Campos – São Paulo

Maiores contribuintes: uma abordagem setorial baseada na análise de *clusters*

Resumo

O objetivo principal da presente monografia é apresentar um modelo de análise setorial para aplicação no acompanhamento dos maiores contribuintes. A pretensão é que a abordagem utilizada viabilize a construção de uma base de conhecimento sobre os perfis tributários de setores específicos da economia sob a ótica dos maiores contribuintes, propiciando uma definição mais clara e objetiva de futuras estratégias de atuação fiscal.

A abordagem setorial proposta é baseada na técnica estatística multivariada denominada Análise de *Clusters* (ou Análise de Agrupamentos). Trata-se de técnica que objetiva a reunião de indivíduos (ou objetos) em grupos tais que os objetos do mesmo grupo são mais parecidos uns com os outros que com os objetos de outros grupos.

A título de aplicação prática, é analisado um setor econômico específico (vendas de veículos automotores) do Estado de São Paulo, no ano-calendário 2008.

Os resultados obtidos permitiram a propositura de uma tipologia, ou rótulo, para descrever o perfil tributário do setor econômico sob foco. Com base nos perfis traçados, a Administração Tributária poderá decidir sobre os aspectos tributários do comportamento dos contribuintes de

determinado setor merecedores de um aprofundamento de investigação, o que contribui para a melhora da qualidade das seleções e fiscalizações efetuadas, para o aumento da presença fiscal e, via de consequência, alavanca o cumprimento espontâneo das obrigações tributárias dos contribuintes.

Maiores contribuintes: uma abordagem setorial baseada na análise de *clusters*

1 Introdução

Sabe-se que o Estado Brasileiro necessita de recursos para promover o bem-estar social. Esses recursos são indispensáveis para que o Governo cumpra com as suas funções principais (GIAMBIAGI; ALÉM, 1999): 1. Função alocativa; 2. Função distributiva e; 3. Função estabilizadora.

A função alocativa[1] está relacionada ao papel do Estado na provisão de bens públicos (exemplos: iluminação pública, justiça, segurança pública). Por outro lado, a função distributiva caracteriza-se pela utilização de instrumentos pelo Estado para redistribuir a renda e reduzir as desigualdades. Para isso, o Governo apresenta como principais formas de distribuição de renda as transferências, os impostos e os subsídios. Cite-se, no caso da União, como exemplos, os programas Bolsa Família (transferência) e Minha Casa, Minha Vida (subsídio). Por último, a função

1 Para mais detalhes sobre as funções alocativas do governo, vide Giambiagi e Além (1999), Rezende (2001) e Viol (2005).

estabilizadora está relacionada à noção do Governo como garantidor da estabilidade de preços e incentivador do crescimento econômico e de um elevado nível de emprego.

O ingresso de recursos públicos ocorre primordialmente de três formas: 1. Empréstimos; 2. Exploração do patrimônio; e 3. Arrecadação de tributos. Destaca-se que a arrecadação de tributos é o tipo de ingresso de recursos que responde pela maior parte do montante da receita da União. Conforme dados da Secretaria do Tesouro Nacional (STN)[2] sobre as receitas da União no ano de 2009, a Receita Tributária e a de Contribuições representam conjuntamente 42,30% da receita total. De outra forma, desconsiderando a Receita com o Refinanciamento (rolagem da Dívida Pública Federal), dada a sua irrelevância econômica segundo Rezende (2001), a participação da Receita Tributária e de Contribuições aumenta para 55,38%.

Como a Receita Tributária e de Contribuições são as mais relevantes para a União, e como este é o principal Ente da Federação em termos de arrecadação tributária, sobreleva-se o papel que a Secretaria da Receita Federal do Brasil (RFB) tem para o Estado Brasileiro como instituição provedora dos recursos públicos indispensáveis para a promoção do desenvolvimento econômico e social do País. Para reforçar esta ideia, destaca-se, de acordo com o Mapa Estratégico da RFB para o período de 2009-2011,[3] os dois grandes objetivos da instituição: a provisão de recursos públicos e a promoção do desenvolvimento econômico e social.

À guisa do alcance de tais objetivos institucionais, a presente monografia apresenta uma abordagem setorial dos maiores contribuintes com intuito de contribuir para a construção do perfil integral do setor a partir da identificação da existência de grupos heterogêneos no que se refere ao comportamento tributário.

A abordagem setorial proposta tem base na técnica estatística multivariada denominada “Análise de *Clusters*”. A partir dos resultados

2 Portaria nº 365, de 29 de junho de 2010.

3 Acesso em: 6 set. 2010. Disponível em: <http://www.receita.fazenda.gov.br/Historico/SRF/PlanejAdminTribAduaneira/Default.htm>.

obtidos, a Administração Tributária poderá definir com mais clareza e objetividade as linhas de atuação sobre os maiores contribuintes de um setor específico.

Diante disso, esta monografia apresenta mais uma ferramenta de análise dos maiores contribuintes com a finalidade de promover um aumento da presença fiscal e, consequentemente, elevar o nível de arrecadação espontânea de setores específicos da economia.

2 A importância da análise setorial para o acompanhamento tributário dos maiores contribuintes

2.1 Por que alguém paga espontaneamente um determinado tributo?

Esta subseção suscita o questionamento acerca dos motivos que conduzem uma pessoa, física ou jurídica, a obedecer a uma determinada norma tributária, mesmo quando esta se opõe aos seus objetivos. No caso das empresas, tal objetivo consiste em maximizar lucros. Sugerem-se três fatores básicos que impulsionam o contribuinte à adesão a normas tributárias ou, como denominam Siqueira e Ramos (2005), à "obediência tributária". São eles: a identificação, a internalização e a submissão (GIANNETTI, 2004).

A identificação é caracterizada pelo cumprimento da legislação em razão da construção ou manutenção de uma boa imagem perante o grupo social do qual a pessoa faz parte ou anseia fazer. Nas palavras de Giannetti (2004, p. 89): "Pago o imposto x porque as pessoas que mais prezo e admiro assim fazem e sinto-me bem pelo fato de, como elas, também fazê-lo." Os pontos em destaque são o incentivo a se comportar como o bom exemplo e a conquista do respeito das pessoas que são admiradas e veneradas.

No caso da internalização, o que ocorre é uma reflexão ética por parte do contribuinte. Este decide pagar o imposto independentemente do que os outros pensam ou fazem. O contribuinte paga o imposto por ser um cidadão, sujeito dotado de direitos e obrigações perante a sociedade da qual faz parte.

Por último, aborda-se a submissão ou, como denominado por Siqueira e Ramos (2005), a "política tributária de imposição", que tem como fulcro uma análise de custo-benefício realizada pelo contribuinte. Trata-se de cálculo racional que adota variáveis como o tamanho e a natureza da sanção (por exemplo, multa de 150% sobre o valor do imposto sonegado e a restrição de liberdade) e a probabilidade de detecção da sonegação.

Preliminarmente, é analisado o tamanho da sanção. No caso do Brasil, o sonegador não se depara com uma penalidade elevada. Pelo contrário: desde a ciência do auto de infração, o contribuinte já começa a receber benefícios como, por exemplo, um desconto de 50% no valor da multa de ofício se o pagamento ocorrer dentro do prazo (art. 44, §3º, da Lei nº 9.430/96, combinado com o art. 6º da Lei nº 8.218/91). Os parcelamentos especiais que, frequentemente, são instituídos pela União, como o Paes, o Paex e os milhares de Refis, também reduzem o custo do sonegador com vantajosas reduções das multas, dos juros e dos demais encargos incidentes sobre o débito, além da concessão de prazos dilargados para pagamento.

Além disso, a extinção da punibilidade pelo pagamento e a suspensão da pretensão punitiva do Estado pelo parcelamento do crédito tributário são outros fatores que contribuem para a minimização do custo de sonegação brasileiro.

A segunda variável do custo da sonegação é a probabilidade de detecção da evasão fiscal. Quanto maior a probabilidade de detecção, maior é a percepção do risco de sonegar. Elevando-se a percepção de risco do contribuinte, aumenta-se a presença fiscal e, consequentemente, a arrecadação espontânea.

2.2 Foco nos maiores contribuintes e na análise setorial

Considerando que a RFB é o órgão responsável pelo provimento dos recursos necessários ao cumprimento das funções estatais e que um aumento na percepção de risco do contribuinte eleva o recolhimento espontâneo, indaga-se: quais são os contribuintes de maior interesse para a RFB?

Contribuindo para a resposta, obtempera De Lima (2007) que a arrecadação está concentrada em poucos e grandes contribuintes, ou seja, uma parcela mínima do total de contribuintes responde pela maior parte da arrecadação federal. Bogéa e Cunha (1999, p. 11) ressaltam que os maiores contribuintes "determinam o sucesso ou fracasso das metas arrecadatórias da administração tributária. Exigem acompanhamento permanente e integral".

Nessa tessitura, a RFB editou a Portaria SRF nº 578, de 11/6/2001, que tratava do acompanhamento da arrecadação dos maiores contribuintes. A partir de então, a RFB passou a acompanhar o comportamento tributário dos maiores contribuintes. Para tanto, foi criada na estrutura administrativa do órgão a Coordenação Especial de Acompanhamento dos Maiores Contribuintes (Comac) e as Delegacias Especiais de Maiores Contribuintes em São Paulo (8º Região Fiscal – RF) e Rio de Janeiro (7º Região Fiscal – RF).

Ademais, foram promovidos avanços tecnológicos direcionados precipuamente aos maiores contribuintes, como por exemplo, os sistemas informatizados *DataWarehouse* Maiores Contribuintes (DWMaco), *WEB*-Análise, Maco *Explorer*, Contágil, Escrituração Contábil Digital (ECD) e Escrituração Fiscal Digital (EFD).

Diante do exposto, não mais se duvida que, hodiernamente, a RFB tem como um de seus focos principais de atuação o trabalho dedicado aos maiores contribuintes. Nesse sentido, destaca-se a seguir o inciso III do art. 3º da Portaria RFB/GAB/Sufis nº 1.317, de 16/6/2010, que trata, dentre outras coisas, das diretrizes para seleção de contribuintes a serem fiscalizados:

> Art. 3º A seleção deverá **priorizar** os contribuintes:
>
> [...]
>
> III – Sujeitos ao **acompanhamento econômico–tributário diferenciado**, no caso de pessoas jurídicas, considerando, sem prejuízo de outros critérios, os casos de indícios de evasão tributária identificados pelas **equipes de acompanhamento dos contribuintes diferenciados**. (negritos nossos)

O principal pilar do acompanhamento dos maiores contribuintes é a construção de seu perfil integral. A construção desse perfil corresponde a um dos procedimentos internos necessários ao cumprimento dos objetivos da RFB, conforme o mapa estratégico delineado para o biênio 2009-2011.

Nesse acompanhamento, o contribuinte é analisado sob múltiplas dimensões tributárias, como a arrecadação, a compensação, o parcelamento e a suspensão de recolhimento diante da propositura de demandas judiciais, bem como o perfil de suas operações de Comércio Exterior. Esse acompanhamento do comportamento tributário perfaz-se tanto sob o enfoque fazendário como sob o previdenciário. Outro aspecto que caracteriza o acompanhamento dos maiores contribuintes é a possibilidade de execução multisetorial das providências sugeridas pelas equipes responsáveis, o que demanda esforço conjunto das projeções de Arrecadação (X-cat), Orientação e Análise Tributária (X-ort), Fiscalização de Tributos Internos e de Comércio Exterior (X-fis e EFA), e Programação Fiscal (X-pac).

A construção do perfil integral do contribuinte parte de uma perspectiva temporal. Analisa-se o comportamento tributário ao longo do tempo com o objetivo de identificar possíveis distorções das quais resultem ações por parte da Unidade Jurisdicionante. Essa análise temporal, no entanto, se mostra insuficiente, pois o contribuinte pode não apresentar nenhuma variação de desvio comportamental nos períodos analisados. Dito de outra forma: a série histórica da dimensão tributária sob estudo (arrecadação, por exemplo) pode apresentar um comportamento regular, se for analisada de forma individualizada. Entretanto, exsurge a seguinte pergunta: será que a evolução do nível de arrecadação tributária do contribuinte analisado é compatível com a do setor econômico no qual este se insere? Ou será que o contribuinte está arrecadando relativamente menos em relação aos demais contribuintes do setor?

Além da análise temporal individualizada, portanto, reputa-se necessário um estudo do contribuinte em relação aos seus pares do setor, ou seja, faz-se mister situar o contribuinte no setor econômico do qual faz parte. É precisamente este o objetivo da análise setorial.

A análise sob o prisma setorial tem fundamento na premissa básica de que os contribuintes de um determinado setor têm comportamento tributário semelhante, visto que estão sujeitos ao mesmo ambiente de mercado e à mesma legislação tributária.

Subsiste, no entanto, a seguinte indagação: qual a abordagem apropriada para a execução da análise setorial que tenha a capacidade de delinear o comportamento tributário dos maiores contribuintes em determinado período com base nas suas múltiplas dimensões tributárias (por exemplo, compensação, arrecadação e débitos declarados)? Tal abordagem permite o traçado do perfil do correspondente setor econômico?

Com o objetivo de solucionar o problema proposto, é apresentada no capítulo seguinte uma abordagem setorial para os maiores contribuintes baseada na técnica estatística multivariada denominada "Análise de *Clusters*".

3 Análise de *clusters*

3.1 Considerações preliminares

A análise de agrupamentos (ou análise de *clusters*) é uma técnica exploratória pertencente ao campo da Estatística Multivariada que objetiva a reunião de indivíduos ou objetos em grupos tais que os objetos do mesmo grupo são mais parecidos uns com os outros do que com os objetos de outros grupos. A ideia é maximizar a homogeneidade de objetos dentro do mesmo grupo ao mesmo tempo em que se maximiza a heterogeneidade entre os grupos (HAIR et al., 2009). Desta forma, se a classificação for bem-sucedida, os objetos dentro dos agrupamentos estarão próximos quando representados graficamente, e diferentes agrupamentos estarão distantes.

Hodiernamente, a análise de agrupamentos é utilizada nas mais diversas áreas do conhecimento humano. Variando da obtenção de taxonomias em biologia para agregar todos os organismos vivos (McGARIGAL; CUSHMAN; STAFFORD, 2000) e de classificações psicológicas baseadas em traços de personalidade (SWANSON; HARRIS; GRAHAM, 2006) até a análise de segmentação de mercados (BERRY;

LINOFF, 2004), a análise de agrupamentos sempre teve forte tradição de agrupar indivíduos e classificar objetos.

3.1.1 Metodologia básica da análise de clusters

Uma das questões-chave da análise de agrupamentos reside na escolha das denominadas "variáveis estatísticas de agrupamento", que correspondem ao conjunto das variáveis que representam as características usadas para comparar os objetos e que, portanto, determinam o seu caráter (HAIR et al., 2009).

Escolhidas as variáveis e obtidas, a matriz numérica que representa os objetos, tem-se que o objetivo principal da análise de agrupamentos é definir a estrutura dos dados colocando os objetos mais parecidos no mesmo grupo.

Para alcançar tal desiderato, três questões básicas devem ser tratadas:

3.1.1.1 Quantificação da similaridade entre os objetos

A primeira questão a ser resolvida, e que se constitui no foco da análise de agrupamentos, envolve a decisão quanto à medida de similaridade existente entre os objetos, de modo a quantificar objetivamente o nível de semelhança entre eles.

Existem várias medidas de similaridade cuja utilização implica algum conhecimento da matriz de dados, ou, nomeadamente, das escalas de medida das variáveis de agrupamento (MAROCO, 2007).

Matematicamente, se for considerado que para cada objeto da amostra foram escolhidas e se encontram disponíveis *p* variáveis de agrupamento, os objetos poderão ser representados como um vetor *p*-dimensional, e a comparação entre diferentes objetos poderá ser efetuada por meio de métricas de distância (ou dessemelhança).

As métricas de distância de utilização mais frequente em análise de *clusters* são:

1. **DISTÂNCIA EUCLIDIANA:** esta métrica quantifica o comprimento da reta que une dois objetos num espaço

p-dimensional. Para *p* variáveis, a distância euclidiana entre os objetos **x**=[x_1 x_2 ... x_p] e **y**=[y_1 y_2 ... y_p] é dada por:

$$d(\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x}-\boldsymbol{y})^T(\boldsymbol{x}-\boldsymbol{y}) = \left[\sum_{i=1}^{p}(x_i - y_i)^2\right]^{1/2} \quad (1)$$

onde x_i e y_i correspondem à *i*-ésima variável de agrupamento (*i*=1, 2,..., *p*) dos objetos **x** e **y** da amostra utilizada na análise.

2. **DISTÂNCIA DE MINKOWSKI:** esta métrica é uma generalização da distância euclidiana e é dada por:

$$d(\boldsymbol{x},\boldsymbol{y}) = \left[(\boldsymbol{x}-\boldsymbol{y})^m\right]^T(\boldsymbol{x}-\boldsymbol{y})^m = \left[\sum_{i=1}^{p}|x_i - y_i|^m\right]^{1/m} \quad (2)$$

note que, para *m* = 1, d(**x**,**y**) é o módulo da distância absoluta entre os objetos **x** e **y** relativamente às *p* variáveis medidas (conhecida por "distância de Manhattan"); para *m* = 2, obtém-se a distância euclidiana.

3. **DISTÂNCIA DE MAHALANOBIS:** trata-se de medida de dessemelhança baseada na correlação estatística entre as variáveis de agrupamento (YOUNIS, 1996):

$$d(\boldsymbol{x},\boldsymbol{y}) = \left[(\boldsymbol{x}-\boldsymbol{y})^T\boldsymbol{S}^{-1}(\boldsymbol{x}-\boldsymbol{y})\right]^{1/2} \quad (3)$$

onde **S** é a matriz de covariâncias amostrais entre os objetos **x** e **y**. Observe que a métrica de Mahalanobis pondera as distâncias entre os objetos nas variáveis de agrupamento em função da correlação entre elas. Ademais, se **S** = **I** (matriz identidade), a métrica de Mahalanobis reduz-se à métrica Euclidiana.

3.1.1.2 Formação dos agrupamentos

Calculada a medida de similaridade de cada par de objetos da amostra, deverá ser desenvolvido um procedimento para a formação dos agrupamentos.

Uma vasta gama de procedimentos tem sido desenvolvida nas disciplinas em que a análise de agrupamentos se aplica. Os algoritmos mais comumente usados podem ser classificados como: (1) hierárquicos e (2)

não hierárquicos. Outros procedimentos têm sido propostos, como, por exemplo, agrupamentos nebulosos (*fuzzy*), métodos grafo-teóricos, redes neurais e modelos evolucionários (JAIN; MURTY; FLYNN, 1999). Todavia, em razão da popularidade das abordagens hierárquicas e não hierárquicas de partição, a discussão aqui travada será limitada às mesmas.

3.1.1.2.1 Técnicas hierárquicas

As técnicas hierárquicas recorrem a passos sucessivos de agregação dos objetos em uma estrutura do tipo árvore. Os procedimentos hierárquicos podem ser aglomerativos (no caso em que cada objeto constitui um agregado isolado na etapa inicial do algoritmo e, à medida que este prossegue, os agregados vão sendo agrupados de acordo com as suas proximidades) ou divisivos (caso em que o algoritmo inicia com um único agrupamento, que vai sendo sucessivamente dividido até que cada objeto constitua um agregado isolado).

Os métodos aglomerativos são utilizados com mais frequência e o algoritmo que explica o seu funcionamento pode ser evidenciado como segue (MAROCO, 2007):

1. começa-se com todos os objetos formando agrupamentos isolados (ou seja, cada objeto forma um agrupamento unitário) de forma que, sendo *n* o número de objetos da amostra, tem-se de início *n* agregados;
2. usando a medida de similaridade, deve-se combinar os dois agrupamentos mais "parecidos" em um novo (agora contendo dos objetos), reduzindo assim a quantidade de agrupamentos em uma unidade. Para tanto, devem ser procurados na matriz de distâncias ($\mathbf{D}_{nxn}$, na qual cada elemento da matriz – d_{ij} – corresponde à distância entre os objetos *i* e *j*) os objetos mais semelhantes, ou seja, com menor valor d_{ij}. Caso existam vários grupos com d_{ij} iguais, deve ser priorizado o agregado que possuir o sujeito de menor valor alfanumérico (d_{29} e não d_{78}, por exemplo);
3. após a combinação dos agregados efetuada conforme o passo 2, a matriz de distâncias **D** deve ser atualizada eliminando a linha e a coluna correspondentes aos agregados combinados

e adicionando uma nova linha e coluna contendo as distâncias entre o novo agregado formado e os restantes;

4. os passos 2 e 3 devem ser repetidos *n-1* vezes, combinando em cada passo os dois agrupamentos mais semelhantes. Ao fim do algoritmo, todos os objetos estarão contidos em um só agrupamento.

Da observação minuciosa do algoritmo acima apresentado percebe-se que, a partir do primeiro passo, existe a necessidade de se definir uma medida de similaridade não entre objetos, mas entre agregados (*clusters*). Tal medida é comumente denominada "método de aglomeração".

Matematicamente, a questão proposta no parágrafo anterior pode ser formalizada da seguinte maneira (TOLEDO, 2005):

Considere dois *clusters*, C_i e C_j, onde $|C_i|$ e $|C_j|$ correspondem ao número de elementos de cada *cluster*, respectivamente. Se d(**x**,**y**) é a medida de distância entre os objetos **x** e **y**, onde $\mathbf{x} \in C_i$ e $\mathbf{y} \in C_j$, definir uma função δ tal que:

$$d(C_i, C_j) = \delta\left[d(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{x}, \boldsymbol{y}, |C_i|, |C_j|\right] \quad (4)$$

onde $d(C_i,C_j)$ caracteriza a medida de distância entre os clusters C_i e C_j.

Da mesma forma que existem várias métricas para quantificar o nível de semelhança (ou dessemelhança) entre objetos, várias abordagens têm sido propostas para medir a similaridade entre agrupamentos compostos de múltiplos objetos. Entre elas, as mais populares são (ANDERBERG, 1973):

1 – LIGAÇÃO SIMPLES: o método de ligação simples (também conhecido como *single linkage* ou *nearest neighbor*) define a semelhança entre agrupamentos como a menor distância de qualquer objeto de um agrupamento a qualquer objeto do outro:

$$d(C_i, C_j) = \min_{\substack{x \in C_i \\ y \in C_j}} d(\boldsymbol{x}, \boldsymbol{y}) \quad (5)$$

2 – LIGAÇÃO COMPLETA: o método de ligação completa (também conhecido como *complete linkage* ou *farthest*

neighbor) define a semelhança entre agrupamentos como a maior distância de qualquer objeto de um agrupamento a qualquer objeto do outro:

$$d(C_i, C_j) = \max_{\substack{x \in C_i \\ y \in C_j}} d(\mathbf{x}, \mathbf{y}) \quad (6)$$

3 – LIGAÇÃO MÉDIA: o método de ligação média (também conhecido como *average linkage*) define a semelhança entre agrupamentos como a média das distâncias dos objetos de um agrupamento aos objetos do outros:

$$d(C_i, C_j) = \frac{1}{|C_i| \cdot |C_j|} \sum_{\substack{x \in C_i \\ y \in C_j}} d(\mathbf{x}, \mathbf{y}) \quad (7)$$

4 – MÉTODO DO CENTRÓIDE: no método centróide, a semelhança entre agrupamentos corresponde à distância entre seus centróides. Centróides são os valores médios das observações sobre as variáveis estatísticas de agrupamento:

$$d(C_i, C_j) = d(\mathbf{x_c}, \mathbf{y_c}) \quad (8)$$

$$\mathbf{x_c} = \frac{1}{|C_i|} \sum_{x \in C_i} \mathbf{x} \qquad \mathbf{y_c} = \frac{1}{|C_j|} \sum_{y \in C_j} \mathbf{y} \quad (9)$$

onde $\mathbf{x_c}$ e $\mathbf{y_c}$ são os centróides dos agrupamentos C_i e C_j, respectivamente.

5 – MÉTODO DE WARD: no método de Ward as distâncias calculadas são versões ponderadas das distâncias euclidianas quadradas entre os centróides dos clusters:

$$d(C_i, C_j) = \frac{|C_i| \cdot |C_j|}{|C_i| + |C_j|} (\mathbf{x_c} - \mathbf{y_c})^T (\mathbf{x_c} - \mathbf{y_c}) \quad (10)$$

onde $\mathbf{x_c}$ e $\mathbf{y_c}$ são os centróides dos agrupamentos C_i e C_j, respectivamente.

3.1.1.2.2 Técnicas não hierárquicas

Diferentemente dos métodos hierárquicos, os procedimentos não hierárquicos não envolvem o processo de construção em árvore. Ao invés disso, designam objetos a agrupamentos assim que o número de

agregados a serem formados tenha sido especificado.

Tal característica das técnicas não hierárquicas, qual seja, a de que o número de agregados a serem formados é pré-especificado, atribui às duas vantagens significativas em relação às técnicas hierárquicas (MAROCO, 2007): 1 – a facilidade com que são aplicadas a matrizes de dados de tamanho considerável, uma vez que, pela estrutura dos algoritmos, não são necessários o cálculo e o armazenamento de uma nova matriz de distâncias a cada passo; 2 – a capacidade de reagrupamento de objetos em um agregado diferente daquele em que foram inicialmente incluídos (o que não acontece nas técnicas hierárquicas, nas quais a inclusão de um objeto em um determinado agregado é definitiva), o que diminui a probabilidade de ocorrência de equívocos de classificação.

Existem inúmeros métodos não hierárquicos que diferem essencialmente no modo como se efetua a primeira designação dos objetos. Como exemplo, poderiam ser citados o método das k-médias ou "*k-means*" (HARTIGAN; WONG, 1979), o método "*fuzzy*" c-médias (BEZDEK, 1981) e métodos baseados em redes neurais artificiais (MANGIAMELI; CHEN; WEST, 1996).

Um dos métodos de aplicação prática frequente é o **método das k-médias**, cujo algoritmo que explica o seu funcionamento é descrito a seguir (MAROCO, 2007):

1 – escolha inicial do número de agregados a serem formados;

2 – especificação dos pontos de partida, conhecidos como "sementes de agrupamento", em número equivalente ao de agregados. Entre as possibilidades de escolha das sementes iniciais, tem-se: a) os centróides dos clusters formados com a utilização das técnicas hierárquicas; b) escolha aleatória dentro do conjunto de objetos; c) os "k" primeiros objetos da matriz de dados, onde "k" corresponde ao número de clusters definido no passo 1;

3 – designação dos objetos aos agregados de cujos centróides se encontram mais próximos;

4 – repetição dos passos 2 e 3 até que: a) não ocorra modificação nas designações efetuadas; b) não ocorra variação significativa na distância mínima de cada objeto a cada um dos centróides dos agregados formados, ou; c) um número máximo de iterações seja alcançado.

A grande dificuldade das técnicas não hierárquicas, como se observa da análise do algoritmo acima, reside na escolha inicial do número de agrupamentos, que é base para a especificação dos pontos sementes.

Uma solução aconselhável para o problema é o emprego de um algoritmo hierárquico de agrupamento para estabelecer o número de agregados, gerando-se os pontos sementes a partir dos resultados obtidos (MILLIGAN, 1980).

Com isso, as vantagens do método hierárquico são complementadas pela habilidade dos métodos não hierárquicos de refinar os resultados e promover alterações de pertinência dos objetos aos agrupamentos.

3.1.1.3 Determinação do número de agrupamentos na solução final

Uma questão de grande importância na análise hierárquica de agrupamentos repousa na escolha do número de agregados mais representativo da estrutura de dados da amostra e que define a partição do conjunto de objetos sob estudo.

Infelizmente, não existe um procedimento de seleção padrão e objetivo (BOCK, 1985; HARTIGAN, 1985). Existem, no entanto, alguns métodos heurísticos (regras de parada) que permitem avaliar a solução escolhida e o número de agregados formados como forma de subsidiar o processo decisório (MINGOTI, 2005).

As regras de parada utilizadas na determinação da solução hierárquica são usualmente baseadas em medidas de heterogeneidade dos agrupamentos em cada passo do algoritmo, definindo a solução quando a medida de heterogeneidade excede um valor pré-especificado ou quando o valor obtido em uma etapa é consideravelmente superior ao obtido na etapa anterior. Nesse último caso, quando o passo *k* do algoritmo exibe um aumento substancial de heterogeneidade, a solução obtida no passo anterior (*k-1*) é a escolhida.

Uma medida de heterogeneidade bastante usual é o denominado "coeficiente de aglomeração", cuja definição é baseada na distância entre os agregados, definida na subseção 3.1.1.2.1 – Técnicas hierárquicas. O coeficiente de aglomeração nada mais é que a menor distância encontrada entre os pares de *clusters* formados.

Outra medida de quão diferentes são os agregados em cada passo do algoritmo é o R^2 (R-Quadrado), que corresponde à razão entre a soma dos quadrados entre os agregados e a soma dos quadrados totais.

As grandezas acima discriminadas são obtidas a partir do seguinte procedimento de cálculo:

Soma de quadrados totais:

$$SQT = \sum_{i=1}^{N} \sum_{j=1}^{N_i} (\boldsymbol{x}_{ij} - \boldsymbol{x}_c)^T (\boldsymbol{x}_{ij} - \boldsymbol{x}_c) \quad (11)$$

onde $\mathbf{x}_{ij}$: j-ésimo objeto pertencente ao i-ésimo agregado;

$\mathbf{x}_c$: centróide da amostra como um todo;

N : número total de agregados;

N_i : número de objetos pertencentes ao i-ésimo agregado.

Soma de quadrados entre os agregados:

$$SQE = \sum_{i=1}^{N} N_i (\boldsymbol{x}_{ci} - \boldsymbol{x}_c)^T (\boldsymbol{x}_{ci} - \boldsymbol{x}_c) \quad (12)$$

onde $\mathbf{x}_{ci}$: centróide do i-ésimo agregado;

R^2:

$$R^2 = \frac{SQE}{SQT} \quad (13)$$

Se for elaborado um gráfico que dispõe o valor de R^2 em função do número de agregados (em contraposição aos passos do algoritmo de aglomeração), perceber-se-á que, à medida que o número de agregados aumenta, o valor de R^2 também aumenta. O valor máximo de R^2 é alcançado na etapa inicial do algoritmo, quando o número de agregados formados é igual ao número de objetos ($R^2 = 1$), pois, nesse caso, a

soma de quadrados entre os agregados (associada de forma direta à heterogeneidade) é igual à soma de quadrados totais, visto que cada objeto constitui um agregado.

Uma estratégia utilizada com bastante frequência (MAROCO, 2007; MINGOTI, 2005) consiste em definir, com base no coeficiente de aglomeração, uma região de possibilidades para o número de agregados em função das variações bruscas de valor à medida que o número de agregados aumenta. Em seguida, escolhe-se a região que contenha um mínimo número de agrupamentos e, cumulativamente, apresente valor de R^2 superior a uma certa percentagem mínima (por exemplo, $R^2 > 80\%$).

3.1.2 Processo de decisão em análise de agrupamentos

Os procedimentos que devem ser desenvolvidos para a obtenção da solução final em análise de agrupamentos, por certo, passam pela execução da metodologia básica descrita na subseção 3.1.1, mas aí não se esgotam.

De fato, começando com a definição dos objetivos, que podem ser exploratórios ou confirmatórios, o delineamento de uma análise de agrupamentos lida com o seguinte processo decisório:

- Partição do conjunto de dados para formar agrupamentos e a seleção de uma solução;
- Interpretação dos agrupamentos para compreender as suas características, descrevendo-as para explicar como os mesmos podem diferir quanto a dimensões relevantes, sendo desenvolvido, ao fim, um rótulo que defina apropriadamente a sua natureza;
- Validação dos resultados da solução final, relacionada à sua estabilidade e capacidade de generalização.

A multiplicidade de procedimentos a serem empregados para a obtenção da solução final faz com que a análise de agrupamentos, como qualquer técnica multivariada, possa ser melhor compreendida a partir de uma abordagem de construção de modelo em seis estágios (HAIR et al., 2009).

É este, precisamente, o objetivo da próxima subseção: a obtenção da solução final de uma análise de agrupamentos com base na execução sequencial de todos os estágios que compõem o processo decisório.

4 Aplicação prática: avaliação do perfil dos maiores contribuintes do setor de venda de veículos automotores sediados na 8ª região fiscal

Para ilustrar a aplicação das técnicas de análise de agrupamentos, foram obtidas, a partir das bases de dados da Secretaria da Receita Federal do Brasil, informações fazendárias e previdenciárias de uma classe particular de contribuintes.

As empresas foram selecionadas em função de um CNAE específico (4511 – comércio de veículos automotores) e são todas jurisdicionadas pela SRRF/8ª RF. O número total de empresas é 197 (cento e noventa e sete).

A obtenção dos dados foi realizada com auxílio dos sistemas informatizados DW (*DataWarehouse*) e SIF (Sistema de Inteligência Fiscal).

Os dados, organizados em colunas com identificação do CNPJ, da razão social da empresa e das dimensões analisadas (variáveis de agrupamento), referem-se ao ano-calendário 2008 e foram armazenados em uma planilha que permite posterior exportação para o programa responsável pela execução dos algoritmos de agrupamento (MATLAB©). A planilha não é aqui reproduzida por questões associadas ao sigilo fiscal dos dados.

4.1 Estágio 1: objetivos da análise de agrupamentos

O principal objetivo da análise é desenvolver uma taxonomia que particione os objetos (contribuintes do setor de venda de veículos automotores sediados na 8ª Região Fiscal) em grupos com perfis tributários similares nas variáveis de agrupamento selecionadas. A pretensão imediata é a de propiciar maior conhecimento sobre tais perfis, viabilizando o desenvolvimento de futuras estratégias de atuação fiscal.

No caso sob foco, no qual o objetivo é delinear perfis de índole tributária, tanto na área fazendária como na previdenciária, foram selecionadas as seguintes variáveis estatísticas de agrupamento:

Tabela 1 – Descrição das variáveis estatísticas de agrupamento

Variável	Descrição
X_1	Arrecadação Fazendária[b] /Receita Declarada[a]
X_2	Débitos Fazendários Declarados[b] / Receita Declarada
X_3	Crédito Total Declarado em DCOMP / Receita Declarada
X_4	Arrecadação Previdenciária[b] / Receita Declarada

[a] a receita declarada refere -se à receita total informada no Demonstrativo de Apuração de Contribuições Sociais (Dacon);
[b] a arrecadação, fazendária ou previdenciária, refere-se ao montante efetivamente recolhido pelo contribuinte, enquanto os débitos declarados referem-se ao montante consignado nas Declarações de Débitos e Créditos Tributários Federais (DCTF), não necessariamente recolhido de forma integral.

4.2 Estágio 2: projeto de pesquisa em análise de agrupamentos

Com os objetivos definidos e as variáveis selecionadas, devem ser abordadas mais quatro questões antes do início do procedimento de partição, relacionadas ao (à):

1. Tamanho da amostra;
2. Padronização dos dados;
3. Presença de observações atípicas;
4. Medida de similaridade entre os objetos;

Considerando que, no caso, foi selecionada a universalidade de contribuintes pertencentes ao setor de mercado e à região analisados, a questão do tamanho amostral dispensa maiores discussões.

A padronização consiste na conversão das variáveis de agrupamento para escores Z (HARTIGAN, 1985; JAIN; MURTY; FLYNN, 1999), que compreende a subtração do valor original da variável pela respectiva média e divisão do resultado pelo desvio padrão. Justifica-se tal procedimento uma vez que as variáveis apresentam níveis de dispersão muito altos, como se observa da Tabela 2. Com isso, fica

minimizado o viés introduzido pelas diferenças nas escalas das variáveis utilizadas na análise, que poderia afetar de forma negativa o processo de agrupamento.

Tabela 2 – Estatística descritiva das variáveis de agrupamento

Variável	Média	Desvio padrão	Coef de variação (%)
X_1	0,0297	0,0613	197,53
X_2	0,0259	0,0485	181,64
X_3	0,0107	0,0947	884,40
X_4	0,0193	0,0364	188,86

A terceira questão diz respeito à presença de observações atípicas na amostra, que podem representar: 1 – observações verdadeiramente aberrantes que ao são representativas da população geral; 2 – observações representativas de segmentos pequenos ou insignificantes da população geral, ou; 3 – grupos reais na população mal representados pela amostragem realizada.

No primeiro caso, as observações atípicas distorcem a verdadeira estrutura dos dados e tornam os agrupamentos obtidos não representativos da população geral. No segundo caso, a observação atípica deve ser removida, de forma que os agrupamentos resultantes passam a representar com maior precisão os segmentos relevantes da população. No terceiro caso, todavia, as observações atípicas devem ser incluídas nas soluções, pois representam grupos válidos e relevantes.

Para a detecção de eventuais observações atípicas, foi utilizada a métrica D^2 de *Mahalanobis*.

A Tabela 3 apresenta os valores D^2 em ordem decrescente para os 10 (dez) objetos mais distantes da média da amostra. Observe que uma observação (a de nº 37) apresenta valor D^2 substancialmente mais elevado que os demais.

Tabela 3 – Identificação de observações atípicas com a métrica D^2 de *Mahalanobis*

Objeto	D^2 Mahalanobis	Objeto	D^2 Mahalanobis
37	189,1	83	38,5
8	82,5	6	33,4
193	79,1	180	30,1
11	65,3	133	29,1
135	55,1	15	22,0

Nota: D^2 de *Mahalanobis* baseada nos valores padronizados das variáveis de agrupamento

Considerando, no entanto, que o universo de contribuintes objeto da análise foi incluído na amostra, fica afastada a ocorrência da situação na qual a observação atípica pode representar um pequeno grupo da população geral. Dessa forma, a observação atípica deve ser removida, de forma que os agrupamentos resultantes passem a representar com maior precisão os segmentos relevantes.

Ressalte-se, contudo, que a eliminação da observação atípica jamais pode significar a ausência de análise sobre o correspondente objeto.

Comparando as variáveis de agrupamento associadas à observação atípica (contribuinte de nº 37), dispostas na Tabela 4, com as médias discriminadas na Tabela 2, verifica-se que a variável $\mathbf{X_3}$ apresenta valor cerca de cem vezes superior à média da amostra.

Tabela 4 – Variáveis de agrupamento associadas às observações atípicas

Observação	X_1	X_2	X_3	X_4
37	0,0499	0,0250	1,2453	0,0815

Justifica-se, portanto, a análise em separado do contribuinte nº 37 sob a ótica da variável X_3, associada à compensação tributária.

A próxima questão envolve a escolha de uma medida de similaridade. Como já visto na subseção 3.1.1.1, o conceito de "*similaridade*" é fundamental na análise de agrupamentos, e consiste na mensuração empírica de correspondência, ou semelhança, entre os objetos a serem agrupados.

Considerando que o conjunto das quatro variáveis é de natureza métrica, foi adotada a medida de distância **euclidiana**.

4.3 Estágio 3: suposições em análise de agrupamentos

A análise de agrupamentos não é uma técnica de inferência estatística, na qual os parâmetros a partir de uma amostra são avaliados com base na probabilidade de serem representativos da população. Em vez disso, a análise de agrupamentos é uma técnica que busca quantificar as características estruturais de um conjunto de observações e, como tal, tem profundo apelo matemático, mas sem possuir fundamentos estatísticos.

Não obstante, duas questões críticas devem ser focalizadas na análise de agrupamentos, pertencentes ao campo estatístico: 1 – a representatividade da amostra; 2 – a existência de multicolinearidade entre as variáveis estatísticas de agrupamento.

Uma exigência no uso da análise de agrupamentos para que sejam atendidos quaisquer dos objetivos apontados no estágio 1 é que a amostra seja representativa da população de interesse. Seja o objetivo exploratório ou confirmatório, os resultados da análise de agrupamentos não são generalizáveis se a representatividade não for garantida.

Considerando que a representatividade amostral já foi garantida pela seleção da totalidade dos contribuintes, o objetivo desta etapa é verificar se existe multicolinearidade entre as variáveis de agrupamento (X_1 a X_4) e inferir sobre o impacto desta na solução final obtida.

A "multicolinearidade" pode ser definida como o grau de dependência linear existente entre as variáveis independentes. A multicolinearidade pode alterar os padrões de agrupamento, pelo fato de as variáveis colineares serem implicitamente ponderadas com maior peso (CORRAR; PAULO; DIAS, 2007).

Várias técnicas têm sido propostas para detectar a presença de multicolinearidade. As mais utilizadas são: os fatores de inflação da variância, ou VIF (*variance inflation factors*) (MARQUARDT, 1970); o exame do número de condição e dos autovalores da matriz de correlações (FADEN, 1978; MONTGOMERY; PECK; VINING, 2001).

O "Fator de Inflação da Variância" (*variance inflation factor*), ou VIF, é definido pela equação:

$$VIF(k) = \frac{1}{1 - R_k^2} \quad (14)$$

onde R_k^2 corresponde ao coeficiente de determinação da regressão linear que tem a variável de agrupamento x_k como função das demais.

O VIF é uma medida do grau em que cada variável de agrupamento (independente) é explicada pelas demais. Quanto maior for o VIF, mais severa é a multicolinearidade. Uma regra prática aceitável é a de que, se $\mathbf{VIF_k > 10}$, a colinearidade existente entre a variável "*k*" e as demais é significativa (JOHNSON; WICHERN, 1988; HAIR et al., 2009, p. 192).

Outra maneira de se quantificar a multicolinearidade deriva da decomposição singular da matriz de correlações, $\mathbf{X}^T\mathbf{X}$, onde $\mathbf{X}$ é a matriz de dados da análise de agrupamentos com a primeira coluna preenchida por 1's (MAROCO, 2007, p. 566). Neste caso, o número de condição (NC) da matriz de correlações, definido por:

$$NC = \frac{\lambda_{max}(\mathbf{X}^T\mathbf{X})}{\lambda_{min}(\mathbf{X}^T\mathbf{X})} \quad (15)$$

é utilizado como medida do nível de multicolinearidade. Na equação acima, $\lambda_{max}(\mathbf{A})$ e $\lambda_{min}(\mathbf{A})$ correspondem ao maior e ao menor autovalor da matriz $\mathbf{A}$, respectivamente. Não é motivo de preocupação um número de condição menor que 100. No entanto, valores acima de 1000 indicam a existência de severa multicolinearidade entre as variáveis de agrupamento (MONTGOMERY; PECK; VINING, 2001). Também se admite a determinação do número de condição associado a cada autovalor da matriz de correlações (SOUZA, 1998). Neste caso, problemas de multicolinearidade podem surgir quando um autovalor menor que 0,01 possuir um número de condição maior que 100.

Observe-se que o critério acima evidenciado não identifica quais as variáveis de agrupamento afetadas pela multicolinearidade. A fim de detectar quais variáveis são dependentes entre si, devem ser calculadas

as proporções de variância decompostas associadas a cada autovalor:

$$PV(\lambda_i) = \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j} \quad (16)$$

As variáveis com proporções de variância decompostas maiores que **0,80** para cada um desses autovalores são candidatas à dependência linear (SOUZA, 1998).

Para o caso analisado, o diagnóstico de multicolinearidade foi realizado com base em 03 (três) critérios: 1 – os fatores de inflação de variância (VIF) devem ser superiores a **10**; 2 – número de condição da matriz de correlações ($\mathbf{X}^T\mathbf{X}$) superior a **1000**; 3 – inexistência de proporções de variância decompostas maiores que **0,80** para cada um dos autovalores da matriz de correlações.

A Tabela 5 discrimina os resultados obtidos da aplicação do triplo critério acima descrito à matriz de dados padronizados.

Tabela 5 – Diagnóstico de multicolinearidade aplicado à matriz de dados padronizados

Variável	VIF	Autovalor	Valor	NC	PV
X_1	5,4886	λ_1	17,55	28,66	0,0180
X_2	6,6081	λ_2	67,84	7,42	0,0695
X_3	1,0241	λ_3	191,54	2,62	0,1963
X_4	2,2677	λ_4	196,00	2,57	0,2008
		λ_5	503,06	1,00	0,5154

Nota 1: a matriz de dados padronizados não contém a observação atípica
Nota 2: a matriz de correlações compree nde a matriz de dados com acréscimo de uma coluna de 1's à esquerda, o que explica a existência de 5 (cinco) autovalores.

Da análise da tabela acima, depreende-se o seguinte: 1 – nenhum dos VIF's associados às variáveis de agrupamento é superior a 10; 2 – o número de condição de matriz de correlações (28,66) é inferior a 1000; 3 – a maior proporção de variância decomposta (0,5154) é inferior ao limite máximo de 0,80.

Conclui-se, assim, que o grau de multicolinearidade existente entre as variáveis de agrupamento é extremamente reduzido, de modo que não se espera que tenha um impacto relevante sobre a solução final de agrupamento.

4.4 Estágio 4: determinação de agrupamentos e avaliação do ajuste geral

Os procedimentos utilizados nesse estágio da análise já foram descritos com detalhe nas subseções 3.1.1.2 – Formação dos agrupamentos e 3.1.1.3 – Determinação do número de agrupamentos na solução final.

A execução dos algoritmos de agrupamento depende da seleção prévia dos seguintes parâmetros: 1 – métrica (selecionada no Estágio 2); 2 – técnica de agrupamento; 3 – método de aglomeração; 4 – critério de parada.

A Tabela 6 discrimina os parâmetros aplicados ao caso sob foco.

Tabela 6 – Parâmetros utilizados na execução do algoritmo de agrupamento

Métrica	Euclidiana
Técnica de agrupamento	Mista (hierárquica seguida da não hierárquica)
Método de aglomeração	Ligação Média (*average linkage*)
Critério de parada	Avaliação da variação de heterogeneidade entre grupos, associada ao valor do coeficiente R^2

A escolha da técnica mista auxilia no desenvolvimento de uma solução ótima para cada número de agregados. Inicialmente é utilizada a técnica hierárquica de agrupamento, que tem por objetivo identificar um conjunto preliminar de soluções como base para estabelecer o número apropriado de agrupamentos. Em seguida, o algoritmo não hierárquico é utilizado para refinar a solução, gerando agregados utilizando os pontos iniciais de busca (pontos sementes) a partir dos resultados da análise hierárquica.

4.4.1 Análise hierárquica de agrupamentos

Na fase hierárquica, assumem especial importância a escolha do método de aglomeração e a fixação do critério de parada.

O método de ligação média (*average linkage*), que define a dessemelhança (ou distância) entre dois agregados como a média das distâncias entre os pares de elementos pertencentes a ambos, foi escolhido por corresponder a uma solução intermediária entre os métodos

de ligação simples (*single linkage*) e completa (*complete linkage*). Além disso, o método escolhido tende a formar agregados com pequena variabilidade interna, ou seja, mais homogêneos.

Como critério de parada do algoritmo hierárquico, foi empregada uma estratégia baseada na variação de heterogeneidade entre os agregados formados. O raciocínio básico é o seguinte: a cada passo do algoritmo de agrupamento, a combinação entre diferentes agregados faz com que a heterogeneidade entre eles aumente. A ocorrência de variações percentuais bruscas na medida de heterogeneidade permite identificar estágios de combinação de agrupamentos que são sensivelmente distintos. Deve ser escolhida, assim, a solução obtida no passo anterior àquele em que foi verificado aumento significativo de heterogeneidade.

Ora, se o coeficiente de aglomeração, que quantifica a heterogeneidade entre os *clusters*, for tabulado em função do número de agregados formados, e se o resultado for disposto em um gráfico, as considerações acima tecidas quanto à variação de heterogeneidade podem ser resumidas pela seguinte regra (MAROCO, 2007, p. 438): **as soluções para o número de agregados (*clusters*) são aquelas onde a inclinação do gráfico "*Heterogeneidade x nº de agregados*" torna-se próxima de zero**.

Como forma de evitar soluções contendo agrupamentos extremamente heterogêneos entre si (o que pode acontecer se a solução escolhida contiver um número pequeno de agregados – 1 ou 2 – ou seja, quando o algoritmo for interrompido nos estágios finais) é utilizado o coeficiente R^2, que é uma medida de percentagem da variabilidade total que é retida em cada uma das soluções.

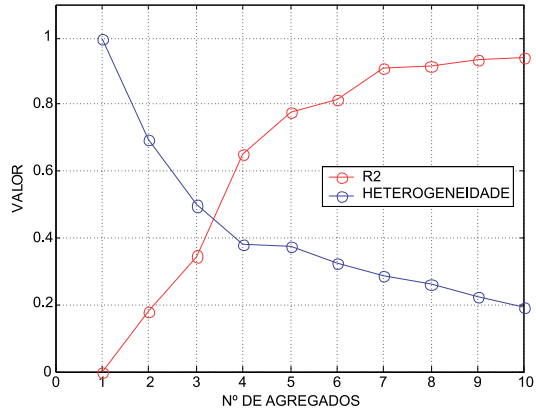
No caso analisado, portanto, foi utilizado um critério de parada composto de 02 (duas) regras:

1 – inclinação pequena do gráfico que representa a heterogeneidade normalizada entre os agregados em função do número de *clusters* formados;

2 – coeficiente $R^2 >= 0,75$ (75%).

A Figura 1 abaixo apresenta os resultados do procedimento descrito nos parágrafos anteriores aplicado à matriz de dados padronizados.

Figura 1 – Critério de parada do algoritmo hierárquico de agrupamento



Da observação do gráfico acima é possível identificar a solução mínima de **cinco agrupamentos** como uma candidata a ser examinada posteriormente pela análise não hierárquica. Nesta situação, o valor do coeficiente R^2 é igual a 0,7721 (77,21%).

Os cinco agrupamentos estão dispostos na Tabela 7.

Tabela 7 – Descrição dos agrupamentos formados pela análise hierárquica

Agrupamento	Nº de Objetos	Objetos
1	180	Todos os demais
2	5	8, 15, 20, 105, 193
3	8	6, 30, 75, 83, 133, 135, 163, 180
4	2	35, 73
5	1	11

4.4.2 Análise não hierárquica de agrupamentos

A análise não hierárquica tem a vantagem de ser capaz de otimizar a solução encontrada na etapa hierárquica pela redesignação de observações até que seja conseguida uma heterogeneidade mínima dentro dos conglomerados.

A primeira tarefa é a seleção do método para especificação dos pontos iniciais para cada um dos agrupamentos obtidos na análise hierárquica (pontos sementes). No caso sob foco, foram escolhidos os **centróides** da solução de agrupamento obtida com base na análise hierárquica.

Com as sementes especificadas, o próximo consiste na escolha do algoritmo a ser utilizado para formar os agrupamentos.

Pelo fato de ser um dos mais conhecidos e utilizados em problemas práticos, o algoritmo das **k-médias** (*k-means*) foi o escolhido na etapa não hierárquica.

A solução obtida com a utilização desse algoritmo, considerando-se como sementes os **centróides** dos agrupamentos da análise hierárquica, é descrita na Tabela 8.

Tabela 8 – Descrição dos agrupamentos formados pela análise não hierárquica

Agrupamento	Nº de Objetos	Objetos
1	174	todos os demais
2	5	8, 15, 20, 105, 193
3	7	6, 30, 42, 75, 133, 135, 163
4	7	14, 35, 73, 126, 132, 149, 171
5	3	11, 83, 180

As partições obtidas nas duas etapas podem também ser comparadas objetivamente por meio do cálculo do valor do R^2 e da variabilidade residual média (média dos valores das somas de quadrados dentro dos grupos, calculada em relação ao número de grupos) (MINGOTI, 2007, p. 195).

A indicação que se tem pelos valores da Tabela 9 é a de que a solução final não hierárquica, obtida pelo algoritmo das k-médias executado com as sementes iniciais posicionadas nos centróides dos agregados (técnica mista), são melhores do que a solução encontrada pelo algoritmo hierárquico, pois resultam em maior valor de R^2 e menor variação residual média.

Tabela 9 – Comparação entre os resultados das técnicas hierárquica e mista

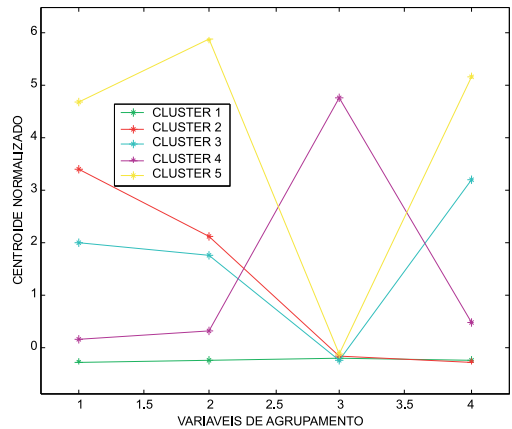
Técnica	Soma de Quadrados Residual	Variação Residual Média	R^2
Hierárquica	177,77	44,44	0,7721
Mista	128,72	32,18	0,8350

Antes de iniciar o próximo estágio, todavia, recomenda-se a confecção do perfil da solução obtida sobre as variáveis estatísticas de agrupamento para garantir que as diferenças entre os agrupamentos formados sejam significativas.

A elaboração do perfil das variáveis pode ser realizada por meio de abordagem gráfica, listando-se as variáveis de agrupamento ao longo do eixo horizontal e seus valores padronizados no eixo vertical. Cada ponto do gráfico representa o valor da respectiva variável no centróide do agrupamento, e os pontos são conectados para facilitar a interpretação visual. Cada linha do gráfico representa, pois, um agregado.

O gráfico abaixo fornece subsídios para a análise de perfil das variáveis padronizadas para a solução de cinco agrupamentos.

Figura 2 – Análise de perfil das variáveis padronizadas para a solução de cinco agrupamentos



Note que cada um dos agrupamentos é relativamente distinto dos demais quanto à magnitude das variáveis, se estas forem consideradas

no seu conjunto. As diferenças são menos distintas na variável X_3, onde somente o agrupamento de nº 4 diferencia-se dos demais.

Outra forma de visualização gráfica dos agrupamentos formados perfaz-se mediante o escalonamento multidimensional, ou MDS (*Multidimensional Scaling*). Trata-se de técnica exploratória multivariada que permite representar de forma parcimoniosa e num sistema dimensional reduzido as proximidades entre os objetos de uma análise de agrupamento (HAIR et. al. 2009).

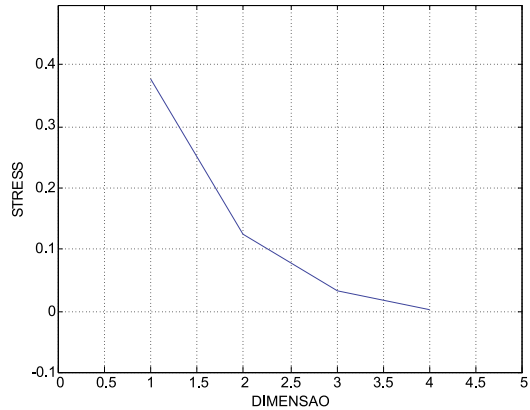
O MDS é uma técnica matemática apropriada para representar graficamente os objetos em um espaço de dimensão menor que o original, levando-se em consideração a distância ou similaridade que os objetos têm entre si. A informação obtida dos elementos amostrais é sintetizada em q dimensões ($q \leq n - 1$; n: número de objetos), o que possibilita a construção de um gráfico no espaço euclidiano (conhecido pela denominação de "*mapa perceptual*") que permite a visualização dos objetos, preservando-se, contudo, a medida de distância entre eles (MAROCO, 2007, p. 458).

No caso específico em que as variáveis de agrupamento são quantificadas, o MDS é chamado "*métrico*" e envolve, para a obtenção da solução, o emprego de ferramentas de Álgebra Linear.

A solução clássica do problema do escalonamento multidimensional métrico encontra-se descrita com detalhes no Anexo A.

Aplicando-se o método de escalonamento multidimensional à solução de agrupamentos obtida, obtém-se o perfil de STRESS (*Standardized Residual Sum of Squares*), conforme a Figura 3.

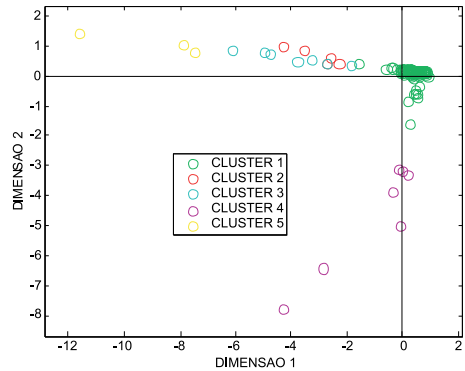
Figura 3 – Nível de STRESS associado ao escalonamento multidimensional



Como se observa do gráfico, se a dimensionalidade for reduzida para 2 (o que permite a representação gráfica dos objetos num plano cartesiano), o nível de STRESS é 0,1249 (abaixo de 0,2), que é um resultado satisfatório.

O mapa perceptual resultante da aplicação do MDS clássico ao caso em testilha encontra-se representado no gráfico a seguir.

Figura 4 – Mapa perceptual correspondente à solução de agrupamentos obtida.

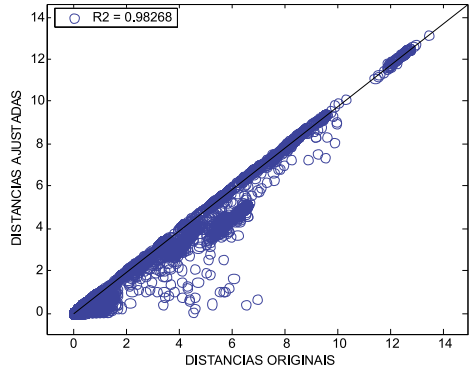


A qualidade da solução final pode também ser avaliada graficamente através da comparação entre as distâncias medidas nas novas

coordenadas *q*-dimensionais ($d_{ij}(\mathbf{X})$) e as distâncias originais (δ_{ij}). O gráfico assim obtido é denominado "gráfico de dispersão" ou "diagrama de Shepard". Ajustando-se um modelo de regressão linear simples entre as variáveis do gráfico, é possível calcular o coeficiente de regressão R^2.

Observe-se, a partir da análise do gráfico de dispersão abaixo, que foi obtido um valor para R^2 igual a 0,982, indicando um excelente nível de ajuste.

Figura 5 – Gráfico de dispersão entre as distâncias originais e as distâncias ajustadas pelo método MDS



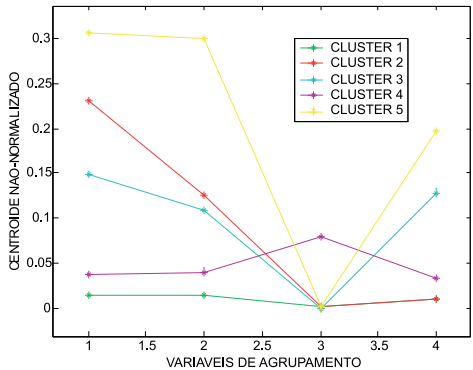
4.5 Estágio 5: interpretação dos agrupamentos

O estágio de interpretação envolve o exame de cada agrupamento em termos das variáveis estatísticas e tem por objetivo a nomeação ou designação de um rótulo que descreva precisamente a natureza dos agregados.

Uma medida frequentemente usada como referência do processo de interpretação é o centróide do agrupamento (HAIR et al., 2009). Se o algoritmo de formação dos grupos foi executado sobre os dados padronizados ou se houve prévia eliminação de variáveis de agrupamento em decorrência da multicolinearidade, a conversão para os dados originais será necessária para que sejam computados os perfis de cada agregado.

A Figura 6 apresenta o diagrama de perfis associados aos agregados formados (dados originais).

Figura 6 – Análise de perfil das variáveis originais para a solução de cinco agrupamentos



Uma observação cuidadosa da Figura 6 permite que se chegue às seguintes conclusões:

1 – o agregado 5 é o que apresenta os melhores níveis de arrecadação fazendária e previdenciária;

2 – o agregado 2 mostra um desequilíbrio entre as arrecadações fazendária e previdenciária. Esta última apresenta níveis comparáveis aos do agregado 1, que é o agregado cujo nível de arrecadação é o mais baixo de todos;

3 – o agregado 3 tem comportamento arrecadatório fazendário e previdenciário médio;

4 – o agregado 4 diferencia-se dos demais por apresentar elevados níveis de compensação tributária. A arrecadação fazendária e a previdenciária deste agregado são baixas em relação ao setor;

5 – o agregado 1, que contém cerca de 80% dos contribuintes, é o que apresenta os níveis mais baixos de arrecadação fazendária e previdenciária.

Na Tabela 10, são descritas as principais características de cada agregado com relação às variáveis de agrupamento utilizadas.

Tabela 10 – Perfil dos agregados formados

Agregado	X_1	X_2	X_3	X_4
1	Baixa arrecadação fazendária	Compatibilidade entre débitos declarados e recolhidos	Baixos níveis de compensação tributária	Baixa arrecadação previdenciária
2	Alta arrecadação fazendária	Débitos declarados em montante inferior aos recolhidos	Baixos níveis de compensação tributária	Baixa arrecadação previdenciária
3	Arrecadação fazendária média	Compatibilidade entre débitos declarados e recolhidos	Baixos níveis de compensação tributária	Arrecadação previdenciária média
4	Baixa arrecadação fazendária	Compatibilidade entre débitos declarados e recolhidos	Alto nível de compensação tributária	Baixa arrecadação previdenciária
5	Alta arrecadação fazendária	Compatibilidade entre débitos declarados e recolhidos	Baixos níveis de compensação tributária	Alta arrecadação previdenciária

A partir de tais características, e com intuito de complementar a descrição, é possível atribuir um rótulo a cada um dos agregados, como o que segue[4]:

- **Agregado 1:** “Contribuintes com baixa arrecadação fazendária e previdenciária”;
- **Agregado 2:** “Contribuintes com alta arrecadação fazendária, mas baixa arrecadação previdenciária, com débitos declarados em montante inferior aos recolhidos”;
- **Agregado 3:** “Contribuintes com arrecadação fazendária e previdenciária média”;
- **Agregado 4:** “Contribuintes com baixa arrecadação fazendária e previdenciária e utilização expressiva do instituto da compensação tributária”;

4 Com exceção do agregado 2, todos os demais agregados apresentam compatibilidade entre os débitos declarados e os recolhidos.

- **Agregado 5:** “Contribuintes com alta arrecadação fazendária e previdenciária”.

4.6 Estágio 6: validação do perfil dos agrupamentos

No estágio final, o processo de validação de perfil envolve a utilização de critérios para garantir que a solução é representativa da população geral e, assim, generalizável para outros objetos e estável ao longo do tempo.

A abordagem mais direta de validação é a análise de amostras separadas (*split sample*), comparando as soluções de agrupamento e avaliando a correspondência dos resultados (HAIR et. al., 2009).

Nesse caso, um método comum é particionar a amostra em dois grupos. Sobre cada um deles é executado o algoritmo de agrupamento e os resultados obtidos são então comparados com o resultado obtido da amostra inicial. A avaliação do resultado, aqui, é efetuada com base em uma métrica que permita verificar o nível de similaridade existente entre duas partições efetuadas sobre um mesmo conjunto de objetos.

Diversas medidas têm sido propostas para avaliar o nível de similaridade entre duas partições (TOLEDO, 2005). Uma delas é o chamado “*Rand Index*”, que mensura a fração do número total de pares de objetos que estão num mesmo *cluster* e na mesma partição, ou em diferentes *clusters* e em diferentes partições (RAND, 1971). Os valores do índice estão situados entre 0 e 1, e valores próximos de 1 indicam um alto nível de similaridade entre as partições comparadas. A formulação matemática empregada na definição do *Rand Index* encontra-se descrita no Anexo B.

O procedimento utilizado na validação da solução de agrupamento, acima evidenciado, pode ser resumido da seguinte forma:

1 – divisão da amostra **A** com a solução de agrupamentos inicial (**C**) em duas amostras de tamanho similar (designadas por $\mathbf{A}_1$ e $\mathbf{A}_2$), escolhidas aleatoriamente;

2 – execução do algoritmo de agrupamento sobre as amostras, obtendo-se, com isso, as partições $\mathbf{P}_1$ e $\mathbf{P}_2$;

3 – cálculo do *Rand Index* entre as partições $\mathbf{P_1}$ e $\mathbf{P_2}$, individualmente, e a partição inicial **C**.

Se os passos 1 a 3 anteriores forem repetidos para um número considerável de amostras escolhidas aleatoriamente (simulação de Monte Carlo), será possível a obtenção da distribuição estatística dos *Rand Index* calculados.

Com isso, afigura-se possível a inferência quanto à estabilidade e à robustez da solução final encontrada.

O procedimento anteriormente descrito foi executado diante da solução final de agrupamentos obtida, sendo elaborados os gráficos de distribuição em frequência do *Rand Index* para cem amostras aleatórias de tamanho similar designadas por $\mathbf{P_1}$ e $\mathbf{P_2}$ (complementar de $\mathbf{P_1}$).

Conclui-se, da análise dos gráficos, pela excelente qualidade dos resultados (média em $\mathbf{P_1}$ igual a 0,9785 e desvio-padrão igual a 0,0224; média em $\mathbf{P_2}$ igual a 0,9779 e desvio-padrão igual a 0,0207), e que a solução de agrupamentos encontrada é consideravelmente estável e robusta.

Figura 7 – Distribuição em frequência do *Rand Index* associado à partição $\mathbf{P_1}$

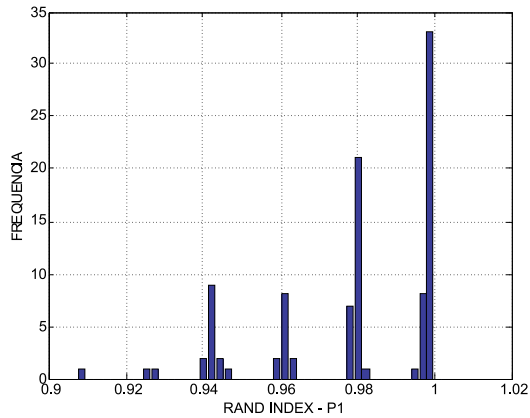
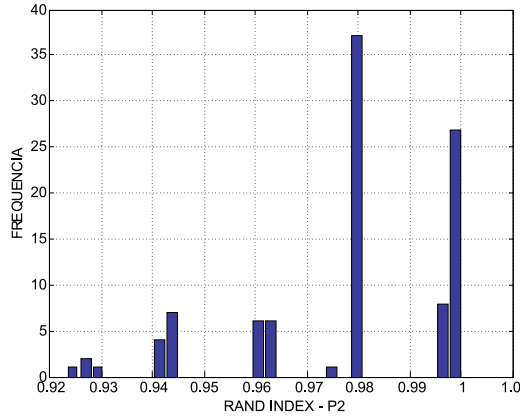


Figura 8 – Distribuição em frequência do *Rand Index* associado à partição $\mathbf{P_2}$



5 Conclusões e recomendações

À guisa de atender a um dos objetivos institucionais da RFB – qual seja, o do provimento dos recursos necessários ao cumprimento das funções estatais – e, mais especificamente, de contribuir para o desenvolvimento de futuras estratégias de atuação dedicadas aos maiores contribuintes, foi apresentada uma abordagem diferenciada com a finalidade de promover um aumento da presença fiscal e, consequentemente, elevar o nível de arrecadação espontânea de setores específicos da economia.

A abordagem proposta objetivou contribuir para a construção do perfil integral a nível setorial, em complementação à metodologia clássica de análise individualizada. Para tanto, foi utilizada a técnica estatística multivariada denominada "Análise de *Clusters*", descrita com o devido formalismo matemático na seção 3.

As técnicas de Análise de *Clusters* foram escolhidas diante de sua eficiência direcionada à redução de complexidade diante de bases de dados extensas e não padronizadas, o que contribui para uma melhor construção dos processos decisórios.

Para ilustrar a aplicação das técnicas de Análise de *Clusters*, foram

obtidas, a partir das bases de dados da Secretaria da Receita Federal do Brasil, informações fazendárias e previdenciárias de uma classe particular de contribuintes (CNAE 4511 – comércio de veículos automotores – com sede na 8ª Região Fiscal, num total de 197 empresas).

A solução final foi obtida a partir da execução de todos os seis estágios do processo de decisão descritos em Hair et al. (2009, p. 436-458), merecendo destaque as seguintes observações: 1 – as variáveis estatísticas foram escolhidas com base em dimensões tributárias relevantes: arrecadação fazendária e previdenciária, débitos declarados e compensação; 2 – foi removida uma observação atípica (objeto de nº 37), que deve ser analisada de forma individualizada sob a ótica da variável associada à compensação; 3 – não foi detectada a existência de multicolinearidade substancial entre as variáveis estatísticas de agrupamento; 4 – o agrupamento dos objetos sob estudo foi efetuado com emprego inicial da técnica hierárquica, sendo utilizada a distância euclidiana como medida de similaridade e a ligação média como método de aglomeração. Como critério de decisão sobre o número de *clusters* a reter, foi usado o critério do R^2 (≥ 75%) associado à declividade do gráfico que representa a distância entre os agregados. Foi identificada a solução de **cinco agrupamentos** como candidata à solução final. Em seguida, a classificação dos objetos nos *clusters* formados foi refinada com o procedimento não hierárquico das k-médias.

Os resultados obtidos da Análise de *Clusters* permitiram a propositura de uma tipologia, ou rótulo, para descrever o perfil tributário do setor econômico sob foco. Os perfis tributários, evidenciados na subseção 4.5, estão associados ao nível de arrecadação fazendária e previdenciária (alta, média ou baixa), bem como ao nível de compensação, tendo sido verificado que, à exceção do agregado 2, todos os demais agregados apresentam compatibilidade entre os débitos declarados e os recolhidos. Nesse sentido: 1 – o **agregado 1** é formado por contribuintes com baixa arrecadação fazendária e previdenciária; 2 – o **agregado 2** é formado por contribuintes com alta arrecadação fazendária, mas baixa arrecadação previdenciária, com débitos declarados em montante inferior aos recolhidos; 3 – o **agregado 3** é formado por contribuintes com arrecadação fazendária e previdenciária média; 4 – o **agregado 4**

é formado por contribuintes com baixa arrecadação fazendária e previdenciária e utilização expressiva do instituto da compensação tributária; 5 – o **agregado 5** é formado por contribuintes com alta arrecadação fazendária e previdenciária.

Por último, foi enfrentado o problema da validação dos resultados a partir de uma avaliação da estabilidade da solução. Da utilização de técnicas de simulação de Monte Carlo ficou demonstrado que, sem embargo de dúvida, a robustez é atributo que pode ser conferido à solução final encontrada.

Por todo o exposto, depreende-se que, tendo em vista o conjunto de variáveis utilizado, existem *clusters* distintos dentre os maiores contribuintes pertencentes ao setor de comércio de veículos automotores, sediados na 8ª Região Fiscal. Esses *clusters* são significativamente distintos uns dos outros e apresentam alta homogeneidade interna.

Diante de tais considerações, tem-se por atingido o principal objetivo do trabalho, e a maior ou menor utilidade da tipologia proposta ficará a cargo dos possíveis usuários dessas informações.

Um aspecto a ser ressaltado é o de que a qualidade da solução encontrada por meio da Análise de *Clusters* depende fundamentalmente da pertinência e qualidade das variáveis incluídas no processamento. A omissão de qualquer variável relevante, ou a inclusão de outra irrelevante, conduzirá a diferentes resultados. Assim, merece ser considerada a possibilidade de que exista um conjunto de variáveis mais propício para as finalidades propostas.

Diversos estudos podem ser desenvolvidos com o intuito de explorar o tema com mais detalhe e sob outros prismas. Podem ser coletadas, por exemplo, informações relativas a empresas sediadas em outras Regiões Fiscais, bem como incluídas outras variáveis estatísticas julgadas relevantes. Da mesma forma, estudos específicos podem ser realizados com foco em outros setores econômicos.

A partir da reunião de todos os resultados obtidos, a Administração Tributária estará munida de uma eficiente ferramenta de seleção para definir com mais clareza e objetividade as linhas de atuação sobre os maiores contribuintes de um setor específico.

Anexo A – Escalonamento multidimensional (MDS) métrico: solução clássica

O MDS "*métrico*" envolve a operação matemática de obtenção da forma canônica, ou fatoração espectral, da matriz **B** dada por (MARO-CO, 2007, p. 464):

$$\boldsymbol{B} = -\frac{1}{2}(\boldsymbol{I}_{nxn} - \frac{1}{n}\boldsymbol{1}_{nx1}\boldsymbol{1}_{1xn})\boldsymbol{D}^2(\boldsymbol{I}_{nxn} - \frac{1}{n}\boldsymbol{1}_{nx1}\boldsymbol{1}_{1xn}) \quad \text{(A.1)}$$

onde *n* é o número de objetos e $\mathbf{D}^2$ (*n x n*) corresponde à matriz de distâncias entre os objetos com todos os seus elementos elevados ao quadrado.

Pode ser matematicamente demonstrado que, se $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$ são os autovalores positivos da matriz **B** e $\mathbf{v_1}$, $\mathbf{v_2}$, ..., $\mathbf{v_p}$ os correspondentes autovetores, então **B** pode ser expressa sob a forma:

$$\boldsymbol{B} = \Gamma_{nxp}\ \Lambda_{pxp}\ \Gamma^T_{pxn} = \boldsymbol{X}_{nxp}\boldsymbol{X}^T_{pxn} \quad \text{(A.2)}$$

onde $\Gamma = [\mathbf{v_1}\ \mathbf{v_2}\ ...\ \mathbf{v_p}]$ é a matriz dos autovetores de **B** associados aos autovalores positivos, $\Lambda = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_p)$ é a matriz diagonal dos autovalores positivos de **B**, e **X** é a matriz das novas coordenadas dos objetos no gráfico de dimensionalidade reduzida (mapa perceptual). As equações acima permitem inferir que é possível escrever **X** sob a forma:

$$\boldsymbol{X} = \Gamma\Lambda^{1/2} \quad \text{(A.3)}$$

Desta forma, torna-se possível a representação dos *n* objetos num gráfico de dimensão *p*, sendo que a cada objeto *i* corresponde uma coordenada $(v_{i1}\lambda_1^{1/2}, v_{i2}\lambda_2^{1/2}, ..., v_{ip}\lambda_p^{1/2})$.

Pode ser demonstrado, ademais, que as distâncias entre os objetos nas novas coordenadas são equivalentes às distâncias originais, ou seja, $\delta_{ij} = d_{ij}(\mathbf{X})$, onde δ_{ij} é a distância entre os objetos *i* e *j* nas coordenadas originais e $d_{ij}(\mathbf{X})$ é a distância entre objetos *i* e *j* nas novas coordenadas, especificadas pela matriz **X**.

Em grande parte das aplicações práticas, no entanto, o número de dimensões obtido a partir das expressões matemáticas acima resulta maior que 3 ($p > 3$), o que inviabiliza a utilização do MDS.

Uma solução para o problema é reter apenas os *q* maiores autovalores ($q < p$), em função de uma medida chamada STRESS (*Standardized Residual Sum of Squares*), que indica o quanto as distâncias nas novas coordenadas estão em conformidade com as distâncias nas coordenadas originais. O coeficiente de STRESS é definido por:

$$STRESS(\mathbf{X}) = \left[\frac{\sum_{i=2}^{n} \sum_{j=1}^{i-1} (\delta_{ij} - d_{ij}(\mathbf{X}))^2}{\sum_{i=2}^{n} \sum_{j=1}^{i-1} \delta_{ij}^{\ 2}} \right]^{1/2} \qquad \text{(A.4)}$$

A escolha do número de dimensões a reter para explicar de forma apropriada as proximidades multidimensionais entre os objetos é geralmente avaliada com um gráfico do tipo *scree-plot* onde o eixo das abscissas representa o número de dimensões (*q*) enquanto o eixo das ordenadas representa o valor do STRESS.

À semelhança do caso da escolha do número de agregados formados pela técnica hierárquica, deve-se reter o número mínimo de dimensões para o qual a curva estabiliza com um declive reduzido. Preferencialmente, com um STRESS abaixo de 0,2.

Anexo B – *Rand Index*: definição

Considere $\mathbf{C} = \{\mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_m\}$ uma solução de agrupamentos obtida a partir da base de dados especificada pela matriz $\mathbf{X}_{nxp}$ (*n*: número de objetos; *p*: número de variáveis), e $\mathbf{P} = \{\mathbf{P}_1, \mathbf{P}_2, ..., \mathbf{P}_s\}$ uma partição extraída da matriz **X**. O número de clusters em **C** e em **P** não precisa ser o mesmo.

Considere agora o par de objetos de **X** dado por $(\mathbf{x}_u, \mathbf{x}_v)$, onde $\mathbf{x}_u$ e $\mathbf{x}_v$ são vetores de dimensão *p*.

Definindo agora a=SS o número de pares de objetos em **X** que pertencem ao mesmo *cluster* **C** e ao grupo na partição **P**; b=SD o número de pares de objetos em **X** que pertencem ao mesmo *cluster* **C** e a grupos diferentes na partição **P**; c=DS o número de pares de objetos em **X** que pertencem a *clusters* diferentes em **C** e ao mesmo grupo na partição **P**; d=DD o número de pares de objetos em **X** que pertencem a *clusters* diferentes em **C** e a grupos diferentes na partição **P**. Assim, $a+b+c+d = M$, onde *M* corresponde ao número total de pares em **X**.

Com base na notação anterior, um índice estatístico criado para quantificar a similaridade entre as partições **C** e **P** é o ***Rand Index***, que mede a fração do número total de pares de objetos em **X** que está no mesmo *cluster* em **C** e no mesmo grupo em **P**, ou em *clusters* diferentes em **C** e em grupos diferentes em **P**:

$$RI = \frac{a + d}{M} \quad \text{(B.1)}$$

O índice assume valores entre 0 (zero) e 1 (um), e valores próximos de 1 indicam um elevado grau de similaridade entre as partições.

Referências

ALLINGHAM, M. G.; SANDMO, A. *Income tax evasion*: a theoretical analysis. Journal of Public Economics, Amsterdam, v. 1, p. 323-338, 1972.

ANDERBERG, M. R. *Cluster analysis for applications*. New York: Academic Press, 1973. 359 p. ISBN-10 0120576503.

ARANHA, F.; ZAMBALDI, F. *Análise fatorial em administração*. São Paulo: CENGAGE Learning, 2008.

BARROSO, L. P.; ARTES, R. *Análise multivariada*. Lavras: UFLA, 2003. 151p.

BERRY, M. J. A.; LINOFF, G. *Data mining techniques*: for marketing, sales, and customer support. 2nd ed. [S.l.]: Wiley Computer Publishing, 2004. 672 p.

BEZDEK, J. C. *Pattern recognition with fuzzy objective function algorithms*. 1st ed. New York: Springer, 1981. 272 p.

BOCK, H. H. On some significance tests in cluster analysis. *Communication in Statistics*, v. 3, p. 1-27, 1985.

BOGÉA, M. S.; CUNHA, M. R. S. *A fiscalização sob o enfoque da Administração Tributária*. Trabalho apresentado no Seminário Internacional de Fiscalização. Vitória, 1999.

CORRAR, L. J.; PAULO, E.; DIAS, J. M. F. *Análise multivariada para os cursos de administração, ciências contábeis e economia*. 1. ed. São Paulo: Atlas, 2007. 544 p.

CRUZ, C. D.; CARNEIRO, P. C. S. *Modelos biométricos aplicados ao melhoramento genético*. Viçosa: UFV, 2003. 585 p.

DA MATTA, R. *Em torno de alguns aspectos sócio-culturais da fiscalização*. Texto apresentado no Seminário Internacional da Fiscalização. Vitória, 1999.

DE LIMA, S. L. M. *O Acompanhamento Tributário* – um novo paradigma em fiscalização para a Receita Federal do Brasil. Brasília: ESAF, 2006. 45 p. Monografia premiada em 1º lugar no 6º Schöntag/2007. nov. 2007. Disponível em <www.esaf.fazenda.gov.br/esafsite/premios/schontag/Monografias_premiadas_arquivos/monografia/monografias6/1LUGAR.pdf>. Acesso em: 8 set. 2010.

FADEN, V. B. *Shrinkage in ridge regression and ordinary least squares multiple regression estimators*. 1978. PhD Dissertation – University of Maryland.

GIAMBIAGI, F.; ALÉM, A. C. *Finanças Públicas*: teoria e prática no Brasil. Rio de Janeiro: Campus, 1999. 380 p.

GIANNETTI, E. *Vícios Privados, Benefícios Públicos?* São Paulo: Companhia das Letras, 2004. 244 p.

HAIR, J. F. Jr. et al. *Análise multivariada de dados*. Tradução: Adonai Schlup Sant'Anna. 6. ed. Porto Alegre: Bookman, 2009. 688 p.

HARTIGAN, J. A. Statistical theory in clustering. *Journal of Classification*, v. 2, p. 63-76, 1985.

HARTIGAN, J. A.; WONG, M. A. Algorithm AS36: a k-means clustering algorithm. *Journal of the Royal Statistical Society, Series C (Applied Statistics),* v. 28 n. 1, p. 100-108, 1979.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. *ACM Computing Surveys*, New York, NY, USA, v. 31, n. 3, p. 264-323, set. 1999.

JOHNSON, R.; WICHERN, D. W. *Applied multivariate statistical analysis*. New Jersey: Prentice Hall International, Inc. 1988. 642 p.

MANGIAMELI, P.; CHEN, S. K.; WEST, D. A comparison of SOM neural network and hierarchical clustering methods. *European Journal of Operational Research*, v. 93, n. 2, p. 402-417, 1996.

MAROCO, J. *Análise estatística* – com utilização do SPSS. 3. ed. Lisboa: Edições Sílabo, 2007. 822 p.

MARQUARDT, D. W. Generalized inverse, ridge regression, biased linear estimation and nonlinear estimation. *Technometrics*, v. 12, p. 591-612, 1970.

MAS-COLLEL, A.; WHINSTON, M. D.; GREEN, J. R. *Microeconomic Theory*. NewYork: Oxford University Press, 1995. 981 p.

McGARIGAL, K; CUSHMAN, S.; STAFFORD, S. *Multivariate statistics for wildlife and ecology research*. New York: Springer-Verlag, 2000. 283 p. ISBN 0-387-98891-2.

MILLIGAN, G. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, v. 45, p. 325-342, 1980.

_____. A monte carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, v. 46, n. 2, 1981.

MINGOTI, S. A. *Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada*. 1. ed. Belo Horizonte: Editora UFMG, 2005. 297 p.

MONTGOMERY, D. C.; PECK, L. A.; VINING, G. G. *Introduction to linear regression analysis*. 3rd ed. New York: Wiley-Interscience, 2001. 672 p.

RAND, W. M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, v. 66, n. 336, p. 846-850, 1971.

REZENDE, F. *Finanças Públicas*. 2ª ed. São Paulo: Atlas, 2001. 382 p.

SIQUEIRA, M. L. *Um modelo econômico para análise da evasão fiscal do imposto sobre a renda no Brasil*. 2004. 117f. Tese (Doutorado em Teoria Econômica) – Departamento de Economia, PIMES, Universidade Federal de Pernambuco, Recife.

SIQUEIRA, M. L.; RAMOS, F. S. A economia da sonegação: teorias e evidências empíricas. *Revista de Economia Contemporânea*, Rio de Janeiro, v. 9, n. 3, p. 555-581, set./dez. 2005.

_____. Evasão fiscal do imposto sobre a renda: uma análise do comportamento do contribuinte ante o sistema impositivo brasileiro. *Economia Aplicada*, v. 10, n. 3, p. 399-424, jul./set. 2006.

SOUZA, G. S. *Introdução aos modelos de regressão linear e não-linear*. Brasília: Embrapa-SPI/Embrapa-SEA, 1998. 505 p.

SWANSON, H. L.; HARRIS, K. R.; GRAHAM, S. *Handbook of learning disabilities*. 1st ed. New York: The Guilford Press, 2006, 30, p. 501-511.

TOLEDO, M. D. G. *A comparison in cluster validation techniques*. 2005. 106 f. Thesis (Master of Science in Mathematics) – Escuela Graduada de la Universidad de Puerto Rico, Puerto Rico.

VARIAN, H. R. *Microeconomia:* princípios básicos. Tradução da 4ª edição americana. Rio de Janeiro: Campus, 1999. 740 p.

VARSANO, R. *A fiscalização tributária sob um enfoque econômico*. Texto apresentado no Seminário Internacional da Fiscalização. Vitória, 1999.

VIEIRA, A. C. M. Análise matemática do risco da sonegação. *Revista Tributação & Desenvolvimento*. Ano 4, n.1. Disponível em: <www.sefaz.pe.gov.br/flexpub/versao1/filesdirectory/sessions581>. Acesso em: 8 set. 2010.

VIOL, A. L. *As Finalidades da Tributação e sua Difusão na Sociedade.* Trabalho apresentado no II Seminário de Política Tributária. Brasília, 2005.

YOUNIS, K. S. *Weighted Mahalanobis distance for hyper-ellipsoidal clustering.* 1996. 98 f. Thesis (Master of Science in Electrical Engineering) – Faculty of the Graduate School of Engineering, Air Force Institute of Technology, Air University, Ohio.