DETECÇÃO DE DADOS DISCREPANTES POR MEIO DE ANÁLISE DE AGRUPAMENTOS DE DADOS COMPOSICIONAIS SOBRE PRESENÇA DE DEFICIÊNCIA NA POPULAÇÃO DE MUNICÍPIOS DO ESTADO DE SÃO PAULO: COMPARAÇÃO DOS RESULTADOS SOB AS TRANSFORMAÇÕES ln E COMPONENTES PRINCIPAIS

Paulo Tadeu Meira e Silva de Oliveira1

poliver@usp.br

IME-USP

Resumo: A identificação de municípios com composição atípica da população segundo a presença de deficiência pode orientar a avaliação da infraestrutura, a destinação de recursos e a formulação de políticas públicas. Neste trabalho, as proporções de 16 categorias (15 tipos de deficiência, isolados ou combinados, e a categoria sem deficiência), obtidas da amostra do Censo Demográfico 2010 para os 645 municípios do Estado de São Paulo, foram tratadas como dados composicionais. Os dados foram transformados por logaritmo neperiano e por componentes principais e, em cada caso, agrupados pelo método de Ward; a análise discriminante, com elipses de 95%, foi usada para identificar municípios discrepantes em cada grupo. Os dendrogramas não indicaram discrepantes, mas a análise discriminante identificou 27 municípios com a transformação logarítmica e 22 com as componentes principais; Pirapora do Bom Jesus foi discrepante nas duas. A deficiência visual esteve presente em 77,9% das pessoas com deficiência.

Palavras-chave: dados discrepantes; pessoas com deficiência; análise exploratória de dados; análise de agrupamentos; análise discriminante; dados composicionais

Abstract: Identifying municipalities with an atypical population composition regarding disability can guide the assessment of infrastructure, the allocation of resources and the design of public policies. In this work, the proportions of 16 categories (15 types of disability, alone or combined, and the no-disability category), obtained from the 2010 Demographic Census sample for the 645 municipalities of São Paulo State, were treated as compositional data. The data were transformed by natural logarithm and by principal components and, in each case, clustered by Ward's method; discriminant analysis with 95% ellipses was used to identify outlying municipalities within each group. The dendrograms showed no outliers, but the discriminant analysis identified 27 municipalities under the logarithmic transformation and 22 under principal components; Pirapora do Bom Jesus was an outlier in both. Visual impairment was present in 77.9% of people with disabilities.

Keywords: outliers; people with disabilities; exploratory data analysis; cluster analysis; discriminant analysis; compositional data

1.INTRODUÇÃO

A identificação de dados discrepantes desempenha um papel importante na análise estatística, pois tais observações podem conter informações importantes em relação às hipóteses de estudo. Se modelos estatísticos são ajustados a dados que contêm valores atípicos, os resultados podem ser enganosos e decisões equivocadas podem ser tomadas (Barbosa et al., 2018).

Dados composicionais são dados que representam quantidades positivas das partes de um todo, por exemplo, proporções das diferentes deficiências em cada município. Em particular, estudam-se e aplicam-se técnicas estatísticas baseadas em transformações de log-razão e técnicas gráficas sobre os dados transformados na detecção das observações discrepantes, que correspondem às observações multivariadas que, por algum motivo, diferem das demais.

A análise de dados sob a perspectiva composicional é de grande interesse em diversas áreas, como: Geologia, em estudos de constituintes do solo; Economia, para a análise de componentes dos gastos das famílias; Medicina, ao avaliar componentes do corpo humano; e indústria alimentícia, na composição dos alimentos, entre outras.

Pessoa com deficiência é toda aquela que apresenta perda ou anormalidade de, no mínimo, uma estrutura ou função psicológica, fisiológica ou anatômica que gere incapacidade para o desempenho de uma ou mais atividades (Brasil, 1999). A Lei Brasileira de Inclusão atualizou esse conceito: pessoa com deficiência é aquela que tem impedimento de longo prazo de natureza física, mental, intelectual ou sensorial, o qual, em interação com uma ou mais barreiras, pode obstruir sua participação plena e efetiva na sociedade em igualdade de condições com as demais pessoas (Brasil, 2015).

Na Seção 2, são introduzidos os conceitos de valores discrepantes e dados composicionais. Além disso, são apresentados procedimentos para a detecção desses valores em conjuntos de dados univariados, bivariados e multivariados; na Seção 3, os resultados e a discussão; e, na Seção 4, as conclusões e propostas para trabalhos futuros.

2.MATERIAIS E MÉTODOS

2.1 MOTIVAÇÃO

Para avaliar o perfil dos municípios paulistas quanto à presença de deficiência na população, considerou-se o conjunto de dados das 3.681.111

pessoas que compuseram a amostra dos respondentes do Questionário Completo do Censo Demográfico do Instituto Brasileiro de Geografia e Estatística (IBGE) agregados nos 645 municípios do Estado de São Paulo (lista disponível em www.ibge.gov.br; IBGE, 2012).

Foram consideradas com deficiência as pessoas que declararam alguma dificuldade, grande dificuldade ou não conseguir de modo algum enxergar, ouvir ou caminhar/subir degraus, ou que declararam ter deficiência mental/intelectual. Esse é o critério da divulgação original do Censo 2010 (IBGE, 2012), segundo o qual 23,9% da população brasileira tinha ao menos uma deficiência. Com o critério recomendado pelo Grupo de Washington, adotado pelo IBGE na releitura desses dados, que considera apenas as respostas “grande dificuldade” e “não consegue de modo algum”, esse percentual cai para 6,7% (IBGE, 2018). Os resultados deste trabalho referem-se, portanto, ao critério mais amplo.

Em qualquer conjunto de dados, é importante fazer, antes da aplicação de qualquer técnica estatística, uma análise cuidadosa de seus componentes, pois podem existir pontos que se destacam quanto a uma ou mais variáveis do estudo que possam prejudicar a modelagem estatística dos dados. A detecção de valores discrepantes permite tratá-los de modo que não comprometam o ajuste.

Neste trabalho, propõe-se analisar, como dados composicionais, as proporções de pessoas com e sem deficiência na população de cada município do Estado de São Paulo, identificar os municípios considerados discrepantes e avaliá-los com mais detalhes.

2.2 DADOS COMPOSICIONAIS

Em estatística, dados composicionais são descrições quantitativas das partes de um todo, que comunicam exclusivamente informação de forma relativa. A característica mais marcante desse tipo de dados é que a sua soma sempre é igual a uma constante (1 para proporções e 100 para porcentagens). Tais dados são muito comuns em áreas de pesquisa como a geologia e a ciência do solo. Exemplos de dados composicionais são a distribuição do tamanho de partículas minerais (areia, silte e argila) de um solo ou a concentração de cátions na solução do solo.

Uma matriz de dados de dimensão $\boldsymbol{n \times p}$ é composicional se a soma das linhas é constante, e subcomposicional se as variáveis formam subconjuntos de uma matriz de dados composicionais. Esses dados ocorrem amplamente em arqueometria, em que é comum determinar a composição química do vidro,

cerâmica, metal ou outros artefatos utilizando técnicas como análise de ativação de nêutrons, análise de fluorescência de raios **X** (XRF), entre outras. O interesse, muitas vezes, gira em torno de saber se existem grupos químicos distintos dentro dos dados e se, por exemplo, estes podem ser associados com as diferentes origens ou tecnologias de fabricação (Baxter, 1999). O espaço amostral dos dados composicionais é o ***simplex,*** subconjunto (D – 1)-dimensional de $\mathbf{R^D}$. Métodos estatísticos usuais podem levar a resultados enganosos se forem aplicados diretamente aos dados originais fechados. Por essa razão, Aitchison (1986) introduziu a transformação log-razão centrada (clr). A transformação clr é uma transformação de S^D para $\mathbf{R^D}$, e o resultado de uma observação $\mathbf{x} \in \boldsymbol{R^D}$, como pode ser observada na expressão (1), são os dados transformados $y \in \boldsymbol{R^D}$ com

$$y = (y_1, \cdots, y_D)' = \left(\log \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \cdots, \log \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right) \qquad (1)$$

Dados composicionais têm propriedades particulares importantes que dificultam a aplicação de técnicas estatísticas usuais a tais dados de concentração. Essas técnicas foram desenvolvidas para dados que variam de $-\infty$ a ∞. Como a soma é constante, se uma componente aumenta, pelo menos uma outra deve diminuir. Isto significa que os resultados de análise estatística padrão do relacionamento entre componentes de dados de concentração ou partes em um conjunto de dados composicionais podem ser ofuscados por efeitos espúrios (Buccianti et al., 2006).

Neste trabalho, os dados composicionais são, para cada município do Estado de São Paulo, as proporções da população em cada uma de 16 categorias: os 15 tipos de deficiência (isolados ou combinados) e a categoria sem deficiência (SD). Essas proporções somam 1 em cada município; as siglas das categorias são apresentadas na Figura 2.

2.3 DETECÇÃO DE DADOS DISCREPANTES

Define-se como discrepante a observação cujo valor difere acentuadamente dos demais, podendo ser muito mais alto ou muito mais baixo.

A presença desses pontos pode causar distorções nos resultados de modelos e estimativas. Após esse diagnóstico, pode-se tomar uma decisão a respeito. Uma possibilidade é corrigi-lo, caso tenha ocorrido erro na transcrição

desses dados. Caso sejam pontos válidos, eles devem ser tratados de forma diferente dos demais, seja por meio de ponderação, seja com uma análise especial.

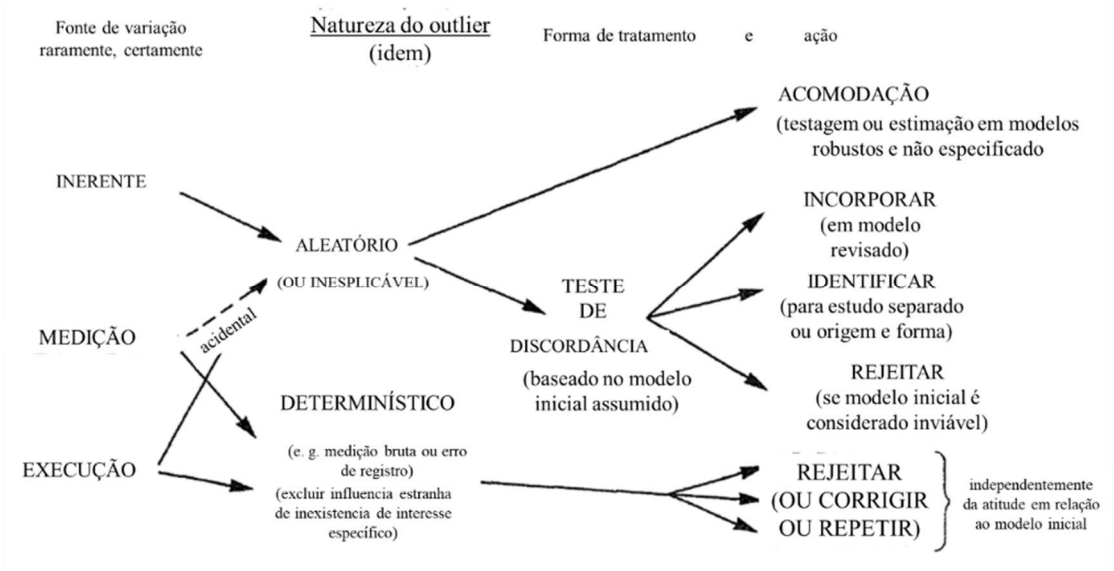
Nos casos univariados e bivariados, um dado discrepante pode ser detectado analisando-se o gráfico de dispersão que mostra o comportamento de cada observação no conjunto de dados de interesse. Se houver pontos distantes da massa de dados, eles devem ser considerados discrepantes. No caso multivariado, a detecção por meio de gráficos torna-se um pouco mais complexa porque a análise deveria ser feita observando-se duas variáveis por vez, o que tornaria o processo longo e pouco confiável, pois um ponto pode ser discrepante em relação a algumas variáveis e não ser em relação a outras, o que faria com que o resultado ficasse mascarado (Giroldo, 2008).

Entre as possíveis causas para a ocorrência de valores discrepantes podem-se citar: erros de medidas; erros de digitação ou transcrição; erros por considerar uma ou mais amostras que não pertencem à população de interesse, conforme apresentado na Figura 1. A importância desse estudo reside no fato de que a presença desses valores pode conduzir a interpretações equivocadas (Barnett & Lewis, 1994).

Na Figura 1, nota-se que os valores discrepantes podem ser classificados quanto à fonte de variação (inerente, de medida ou de execução); quanto à sua natureza (aleatória ou determinística); quanto à forma de tratamento (teste de discordância); e, finalmente, quanto à decisão, que pode ser acomodar, incorporar, identificar, rejeitar, corrigir ou repetir.

Sob o ponto de vista fenomenológico, podemos distinguir os seguintes tipos de dados discrepantes: observações extremas que podem acontecer apenas por acaso em observações aleatórias; amostras poluídas devido a um tratamento inadequado durante a amostragem ou no laboratório, entre outros.

Figura 1 - Tratamento de valores discrepantes



Fonte - Barnett & Lewis (1994)

Entre as diferentes técnicas para detecção de valores discrepantes é possível destacar:

i) dados univariados: diagrama de caixa e teste de Grubbs; ii) dados bivariados: elipse de confiança e bag-plot; iii) dados multivariados: distâncias euclidiana, de Pearson e de Mahalanobis; iv) dados multivariados robustos: BACON, MVE, MCD, MED, COMEDIANA e diagrama ternário; e, finalmente, v) dados particionados em grupos: dendrograma pelo método de Ward e análise discriminante, com a finalidade de avaliar dados discrepantes nos diferentes grupos.

Neste trabalho, utilizam-se o dendrograma pelo método de Ward (Seção 2.4) e a análise de componentes principais com elipse de predição (Seção 2.5), complementados pela análise discriminante (Seção 2.6).

2.4 DENDROGRAMA PELO MÉTODO DE WARD

Entre os diversos métodos hierárquicos de agrupamento, adotou-se o método de Ward (Ward, 1963), que utiliza a distância euclidiana ao quadrado e, a cada passo, une os dois grupos que resultam no menor aumento da soma de quadrados dentro dos grupos. O número de grupos é definido em função dos saltos dessa soma de quadrados a cada passo (Barroso & Artes, 2003).

O dendrograma é um recurso gráfico que também pode ser usado na detecção de dados discrepantes. Baxter (1999), por exemplo, considerou os métodos de ligação média e de Ward, com as distâncias euclidiana e euclidiana ao quadrado, aplicados às concentrações transformadas por logaritmo.

Esses e vários outros métodos de análise de agrupamentos estão implementados em aplicativos computacionais como SAS, Minitab, SPSS, R e Statistica. O método de detecção consiste em verificar, no dendrograma, as observações que ficam isoladas em um único grupo ou cuja medida de dissimilaridade é muito maior que a mesma medida para as demais (Hair Jr. et al., 1996).

O objetivo principal desta técnica é identificar padrões de agrupamento, bem como identificar possíveis valores discrepantes em cada um dos diferentes grupos formados. A escolha do método de Ward deveu-se ao forte apelo estatístico, por basear-se na partição da soma de quadrados totais de uma ANOVA e por gerar grupos que possuem uma alta homogeneidade interna e máxima heterogeneidade externa (Barroso & Artes, 2003).

2.5 ANÁLISE DE COMPONENTES PRINCIPAIS E ELIPSE DE PREDIÇÃO

A análise de componentes principais (Pearson, 1901; Barroso & Artes, 2003) foi realizada com o intuito de resumir o padrão de correlação entre as variáveis e, muitas vezes, é possível chegar a conjuntos de variáveis que sejam não correlacionadas umas com as outras, levando assim a um agrupamento delas.

Para a detecção de dados discrepantes adotaram-se os seguintes procedimentos:

i) Calcular os escores das componentes principais;

ii) Construir o dendrograma a partir desses escores;

iii) Avaliar o número de grupos formados;

iv) Ajustar uma análise discriminante tendo como classes os grupos formados e construir o diagrama de dispersão dos escores das duas primeiras funções discriminantes; e, finalmente,

v) Traçar a elipse de predição de 95% para cada um dos diferentes grupos formados.

Um dado será considerado discrepante em seu grupo quando se encontrar fora da respectiva elipse de predição de 95%, construída a partir do vetor de médias e da matriz de covariâncias dos escores do grupo e do quantil de 95% da distribuição F, de modo a conter, sob normalidade, cerca de 95% dos municípios do grupo. Cabe observar que, sob normalidade, espera-se que cerca de 5% das observações fiquem fora de uma elipse de 95% mesmo na ausência de valores discrepantes; por isso, os municípios assim identificados devem ser entendidos como candidatos a discrepantes, cuja atipicidade é examinada a partir de suas proporções (Tabela 3).

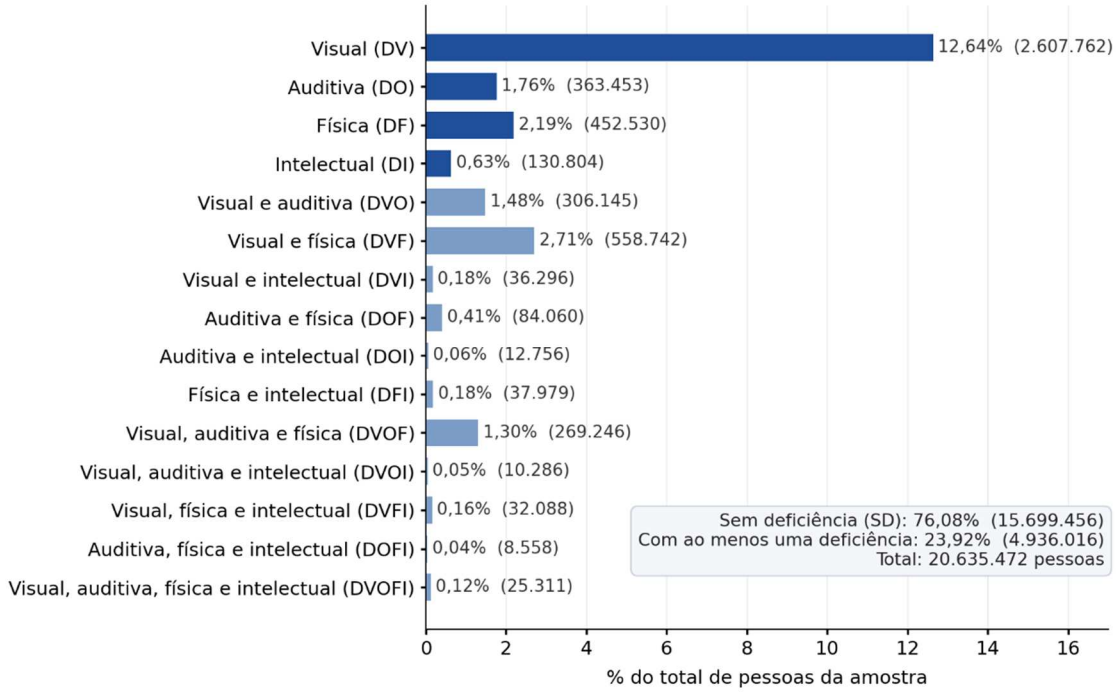
2.6 ANÁLISE DISCRIMINANTE

Nas duas abordagens, depois da formação dos grupos pelo método de Ward, ajustou-se uma análise discriminante linear, com funções discriminantes canônicas, tendo como classes os sete grupos obtidos (Hair Jr. et al., 1996). Os escores das duas primeiras funções discriminantes foram representados em gráficos de dispersão, um para cada grupo e outro para o conjunto dos municípios (Figuras 3 a 6). Para cada grupo, traçou-se a elipse de predição de 95% descrita na Seção 2.5, e os municípios situados fora dela foram considerados discrepantes. A porcentagem de municípios classificados no próprio grupo pela função discriminante é apresentada como medida de separação entre os grupos.

3.RESULTADOS E DISCUSSÕES

A Figura 2 apresenta as distribuições por tipo de deficiência na população formada pelos respondentes do questionário completo do Censo 2010 no Brasil (frequências da amostra, sem expansão pelos pesos amostrais) do IBGE. É possível observar nessa figura que a maior quantidade de pessoas entre as diferentes deficiências é formada por pessoas com deficiência visual que representam 77,9% das pessoas que têm ao menos uma deficiência e que 28% das pessoas com deficiência têm deficiência múltipla.

Figura 2 - Distribuição das pessoas por tipo de deficiência na amostra do Censo Demográfico 2010 – Brasil



Fonte - IBGE, Censo Demográfico 2010 – amostra (frequências não expandidas pelos pesos amostrais)

3.1 PARTE I: DADOS COMPOSICIONAIS TRANSFORMADOS POR LOGARITMO NEPERIANO

Os logaritmos neperianos dos dados composicionais foram agrupados pelo método de Ward, com a distância euclidiana ao quadrado, e os municípios foram alocados em sete grupos. Por inspeção do dendrograma, nenhum município ficou isolado em grupo próprio, ou seja, não foi detectado dado discrepante em nenhum dos grupos. O grupo com maior número de municípios foi o 1, e o de menor, o 7.

O método de Ward permitiu formar grupos o mais homogêneos possível internamente e o mais heterogêneos possível entre si. O uso desse método contribui para uma detecção muito menor de dados atípicos.

A Tabela 1 mostra as médias das proporções (%) das diferentes deficiências nos grupos formados a partir dos dados transformados por logaritmo neperiano, em que ***n*** é o número de municípios de cada grupo, SD indica sem deficiência e as demais siglas indicam as deficiências: DV visual, DO auditiva, DF física, DI intelectual, DVO visual e auditiva, DVF visual e física, DVI visual e intelectual, DOF auditiva e física, DOI auditiva e intelectual, DFI física e intelectual, DVOF visual, auditiva e física, DVOI visual, auditiva e intelectual, DVFI visual, física e intelectual, DOFI auditiva, física e intelectual e DVOFI visual, auditiva, física e intelectual.

Para cada um dos diferentes grupos é possível destacar:

- Grupo 1: maiores proporções de DV e DVOI e menores de DF e DOF;
- Grupo 2: maiores proporções de DI, DOI, DFI, DVFI, DOFI e DVOFI e menores de DO e DVOI;
- Grupo 3: maior proporção de DOI e menor de DOFI;
- Grupo 4: maiores proporções de DO e DOF e menores de DI, DVF e DVI;
- Grupo 5: maiores proporções de DVO, DVF, DVI e DVOF e menores de SD e DOI;
- Grupo 6: maiores proporções de DO e DF e menores de DVOF e DVOFI;
- Grupo 7: maior proporção de SD e menores de DV, DVO, DFI e DVFI.

Tabela 1 - Médias das proporções (%) das diferentes deficiências, por grupo formado a partir dos dados transformados por logaritmo neperiano

Nº de discrepantes	%	n	Grupo	SD	DV	DO	DF	DI	DVO	DVF	DVI	DOF	DOI	DFI	DVOF	DVOI	DVFI	DOFI	DVOFI
10	4,5	222	1	77,85	11,37	1,76	2,37	0,63	1,27	2,44	0,17	0,41	0,06	0,19	1,10	0,05	0,15	0,06	0,12
2	3,0	67	2	77,87	10,76	1,72	2,53	0,68	1,28	2,61	0,18	0,45	0,08	0,24	1,21	0,00	0,19	0,08	0,13
4	4,5	88	3	77,85	11,00	1,78	2,41	0,64	1,27	2,61	0,17	0,45	0,08	0,21	1,19	0,04	0,17	0,00	0,11
1	2,1	47	4	78,92	10,41	1,81	2,63	0,53	1,22	2,33	0,00	0,53	0,05	0,18	1,10	0,04	0,11	0,04	0,10
5	4,7	106	5	77,02	11,35	1,79	2,57	0,62	1,38	2,77	0,19	0,46	0,00	0,21	1,28	0,04	0,17	0,04	0,12
5	6,9	72	6	78,77	10,18	1,81	2,69	0,60	1,26	2,51	0,15	0,50	0,05	0,22	1,03	0,04	0,15	0,05	0,00
-	-	43	7	79,19	10,11	1,75	2,54	0,64	1,21	2,48	0,17	0,48	0,06	0,15	1,10	0,02	0,00	0,05	0,05

Fonte - Elaborado com base em dados do IBGE, Censo Demográfico 2010

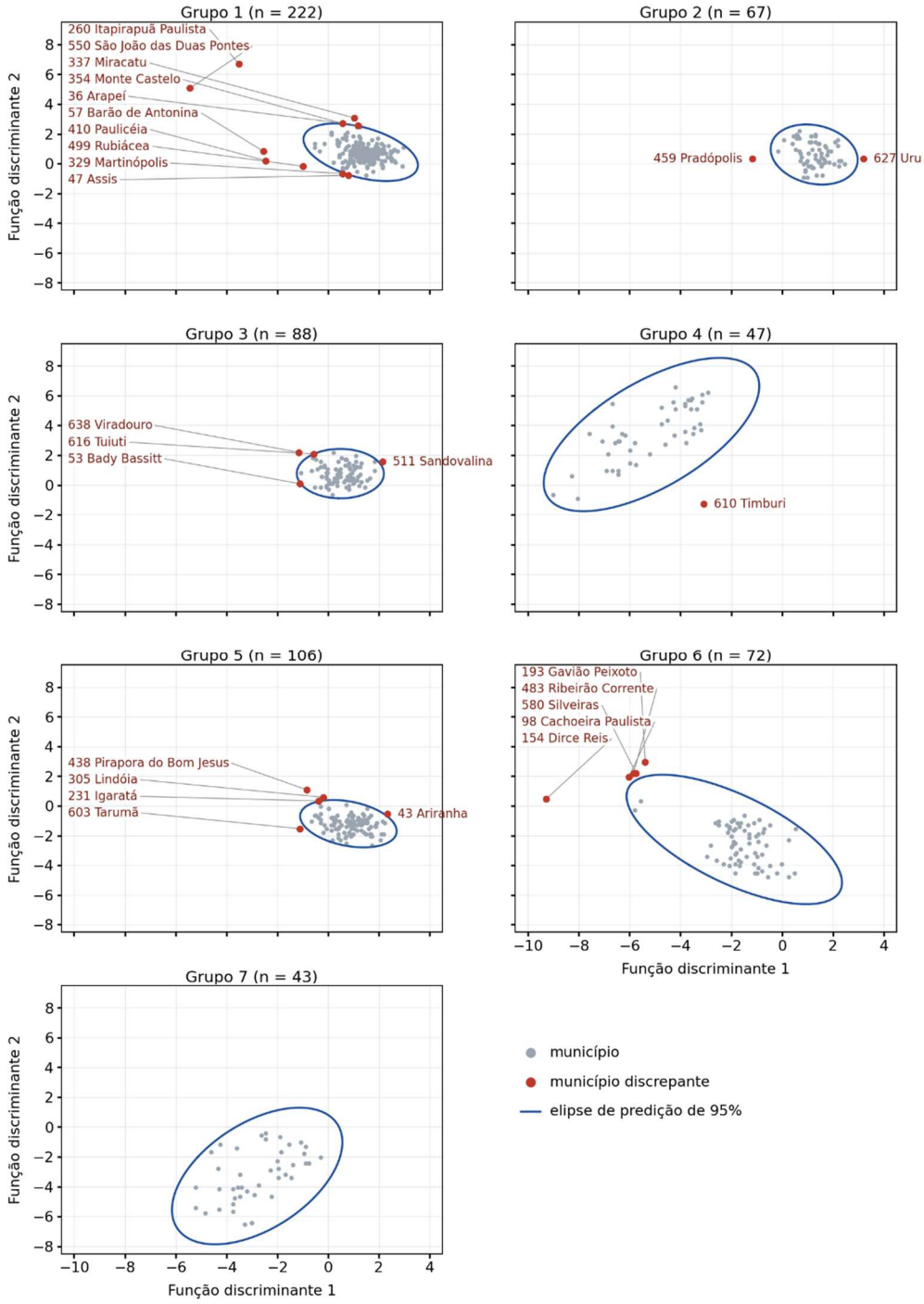
Observa-se também na Tabela 1 que o grupo com maior porcentagem de valores discrepantes foi o 5, com 4,7% de seus municípios considerados discrepantes.

A seguir, é apresentada a Figura 3 com os gráficos de dispersão entre a primeira e a segunda função discriminante para cada um dos diferentes grupos. Os casos situados fora da elipse de seu grupo correspondem a municípios considerados discrepantes.

É possível verificar também na Figura 3 que os casos considerados discrepantes são os dos municípios 36 (Arapeí), 47 (Assis), 57 (Barão de Antonina), 260 (Itapirapuã Paulista), 329 (Martinópolis), 337 (Miracatu), 354 (Monte Castelo), 410 (Paulicéia), 499 (Rubiácea) e 550 (São João das Duas Pontes) no grupo 1 (dez municípios); 459 (Pradópolis) e 627 (Uru) no grupo 2 (dois municípios); 53 (Bady Bassitt), 511 (Sandovalina), 616 (Tuiuti) e 638 (Viradouro) no grupo 3 (quatro municípios); 610 (Timburi) no grupo 4 (um município); 43 (Ariranha), 231 (Igaratá), 305 (Lindóia), 438 (Pirapora do Bom Jesus) e 603 (Tarumã) no grupo 5 (cinco municípios); 98 (Cachoeira Paulista), 154 (Dirce Reis), 193 (Gavião Peixoto), 483 (Ribeirão Corrente) e 580 (Silveiras) no grupo 6 (cinco municípios); e, finalmente, nenhum município do grupo 7 foi considerado atípico.

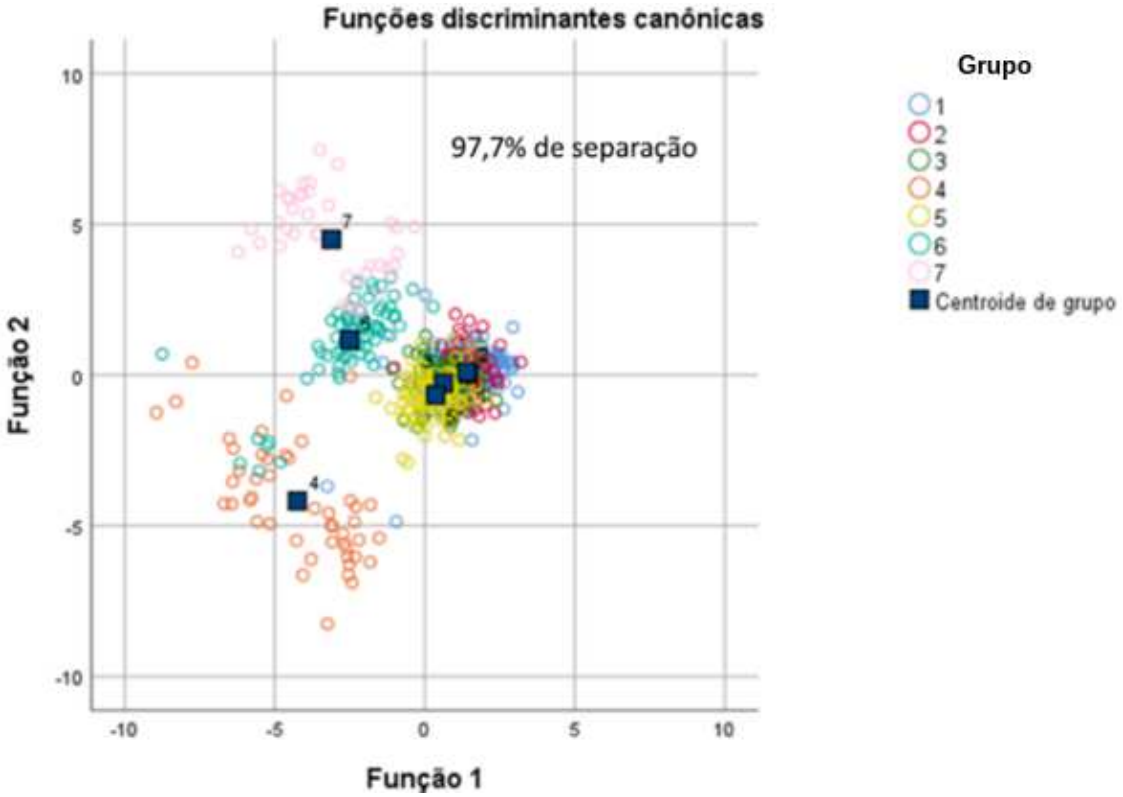
A Figura 4 apresenta o gráfico das funções discriminantes para todo o conjunto de dados. Nota-se que os grupos 1, 2, 3 e 5 se sobrepõem, que o grupo 6 se sobrepõe ao 7 e que o grupo 4 é o que melhor se separa dos demais.

Figura 3 - Gráficos das funções discriminantes, por grupo (dados transformados por logaritmo neperiano)



Fonte - Elaborado com base em dados do IBGE, Censo Demográfico 2010

Figura 4 - Gráfico das funções discriminantes para os grupos formados a partir dos dados transformados por logaritmo neperiano



Fonte - Elaborado com base em dados do IBGE, Censo Demográfico 2010

3.2 PARTE II: ESCORES DAS COMPONENTES PRINCIPAIS DOS DADOS COMPOSICIONAIS

Os escores das componentes principais também foram agrupados pelo método de Ward em sete grupos e, novamente, nenhum município ficou isolado no dendrograma. Na Tabela 2, fazendo-se análise similar à da Tabela 1, destacam-se os seguintes resultados:

- O maior grupo foi o 2, com 289 municípios, e o menor, o grupo 5, com apenas 20 municípios;
- Grupo 1: maiores proporções de DV, DI, DVO, DVF, DVI, DOI, DFI, DVOF, DVOI e DVFI e menor de SD;
- Grupo 3: menor proporção de DOI;
- Grupo 4: maior proporção de SD e menores de DV, DO, DF, DI, DVO, DVF, DOF, DOI, DVOF, DVOI, DOFI e DVOFI;
- Grupo 5: maiores proporções de DO, DF, DOF e DOFI e menor de DVI;
- Grupo 6: maior proporção de DOI e menor de DFI;
- Grupo 7: maior proporção de DVOFI e menor de DVFI.

Tabela 2 - Médias das proporções (%) das diferentes deficiências, por grupo formado a partir dos escores das componentes principais

Nº de discrepantes	%	n	Grupo	SD	DV	DO	DF	DI	DVO	DVF	DVI	DOF	DOI	DFI	DVOF	DVOI	DVFI	DOFI	DVOFI
3	3,0	101	1	73,30	13,09	1,81	2,77	0,73	1,63	3,50	0,22	0,53	0,07	0,28	1,61	0,06	0,20	0,05	0,15
12	4,2	289	2	79,07	10,89	1,68	2,23	0,61	1,16	2,22	0,15	0,39	0,06	0,19	1,02	0,03	0,15	0,04	0,11
5	3,8	132	3	77,42	10,68	1,97	2,82	0,62	1,36	2,72	0,19	0,52	0,04	0,21	1,16	0,03	0,16	0,04	0,07
2	4,5	44	4	83,38	8,40	1,49	2,10	0,49	0,96	1,61	0,06	0,32	0,04	0,16	0,81	0,02	0,09	0,03	0,04
-	-	20	5	75,41	10,96	2,32	3,40	0,64	1,40	3,02	0,01	0,71	0,06	0,26	1,41	0,04	0,14	0,06	0,15
-	-	24	6	78,99	10,15	1,91	2,52	0,56	1,36	2,49	0,10	0,51	0,07	0,01	1,09	0,03	0,11	0,05	0,07
-	-	35	7	78,74	10,39	1,67	2,56	0,64	1,20	2,60	0,06	0,49	0,06	0,18	1,16	0,03	0,00	0,05	0,17

Fonte - Elaborado com base em dados do IBGE, Censo Demográfico 2010

Observa-se ainda na Tabela 2 que o grupo com maior porcentagem de valores discrepantes foi o 4, com 4,5% dos municípios.

Na Figura 5, é possível observar os seguintes valores discrepantes: no grupo 1 (três municípios): 250 (Itaí), 185 (Florínia) e 234 (Ilha Comprida); no grupo 2 (doze municípios): 4 (Águas da Prata), 55 (Bálsamo), 105 (Cajati), 221 (Iaras), 238 (Indiaporã), 300 (Lavínia), 349 (Mongaguá), 359 (Motuca), 368 (Nova Aliança), 438 (Pirapora do Bom Jesus), 476 (Regente Feijó) e 637 (Vinhedo); no grupo 3 (cinco municípios): 205 (Guarani d'Oeste), 408 (Parisi), 475 (Redenção da Serra), 533 (Santa Salete) e 551 (São João de Iracema); no grupo 4 (dois municípios): 167 (Embaúba) e 601 (Taquarivaí), e, finalmente, nos grupos 5, 6 e 7 não foram detectados municípios com valores discrepantes.

A Tabela 3 relaciona os municípios considerados discrepantes com a transformação por logaritmo neperiano (Parte I) e com os escores das componentes principais (Parte II). Observa-se que o município de Pirapora do Bom Jesus foi considerado discrepante nas duas transformações.

Tabela 3 - Relação de municípios com proporções (%) consideradas discrepantes

Município	Código	Parte	Grupo	SD	DV	DO	DF	DI	DVO	DVF	DVI	DOF	DOI	DFI	DVOF	DVOI	DVFI	DOFI	DVOFI
Águas da Prata	4	II	2	79,32	8,64	2,66	3,15	0,45	1,19	2,25	0,12	0,53	0,08	0,20	1,15	0,00	0,16	0,04	0,04
Arapeí	36	I	1	77,46	10,84	2,22	3,82	0,62	0,99	2,34	0,25	0,49	0,12	0,00	0,49	0,00	0,12	0,12	0,12
Ariranha	43	I	5	73,91	14,32	1,22	1,91	1,04	0,99	3,59	0,41	0,29	0,00	0,12	1,28	0,17	0,46	0,12	0,17
Assis	47	I	1	79,80	9,97	1,73	2,47	0,48	1,08	2,13	0,15	0,55	0,01	0,13	1,30	0,03	0,07	0,05	0,04
Bady Bassitt	53	I	3	80,98	10,09	1,77	2,25	0,65	0,78	1,88	0,07	0,38	0,03	0,20	0,85	0,00	0,03	0,00	0,03
Bálsamo	55	II	2	83,93	6,12	1,53	2,88	0,29	0,82	2,41	0,00	0,47	0,00	0,35	0,82	0,18	0,12	0,00	0,06
Barão de Antonina	57	I	1	73,71	9,48	1,97	4,32	0,56	2,07	3,57	0,09	0,75	0,09	0,00	2,91	0,09	0,38	0,00	0,00
Cachoeira Paulista	98	I	6	78,07	10,55	1,99	2,49	0,82	1,24	2,77	0,00	0,53	0,04	0,14	1,21	0,04	0,04	0,07	0,00
Cajati	105	II	2	72,49	14,41	2,08	2,47	0,28	1,91	3,58	0,35	0,17	0,07	0,10	1,49	0,07	0,31	0,07	0,14
Dirce Reis	154	I	6	68,36	13,65	1,69	5,07	0,97	1,57	5,56	0,00	0,97	0,12	0,36	1,45	0,00	0,00	0,24	0,00
Embaúba	167	II	4	90,58	4,17	1,25	1,58	0,67	0,42	0,75	0,00	0,08	0,00	0,25	0,25	0,00	0,00	0,00	0,00
Florínia	185	II	1	63,30	18,82	2,21	2,63	0,95	1,89	5,47	0,53	0,74	0,00	0,32	2,52	0,11	0,32	0,00	0,21
Gavião Peixoto	193	I	6	86,37	5,01	2,14	2,06	0,29	0,66	1,92	0,00	0,22	0,22	0,44	0,22	0,07	0,15	0,22	0,00
Guarani d'Oeste	205	II	3	66,43	16,63	3,39	2,19	0,50	2,79	4,58	0,10	0,30	0,00	0,30	2,59	0,00	0,20	0,00	0,00
Iaras	221	II	2	83,36	8,64	1,17	1,56	0,54	0,98	2,29	0,15	0,15	0,10	0,24	0,49	0,05	0,10	0,05	0,15
Igaratá	231	I	5	71,14	17,37	1,38	1,56	0,30	2,40	3,35	0,06	0,36	0,00	0,18	1,50	0,00	0,24	0,00	0,18
Ilha Comprida	234	II	1	62,77	20,96	2,37	1,81	0,90	2,94	4,29	0,40	0,40	0,11	0,34	2,26	0,11	0,23	0,06	0,06
Indiaporã	238	II	2	76,20	8,56	2,04	3,14	0,86	1,49	3,61	0,08	1,02	0,08	0,24	1,81	0,00	0,31	0,08	0,47
Itaí	250	II	1	73,49	13,92	1,51	2,22	0,59	1,46	4,01	0,25	0,04	0,08	0,17	1,76	0,04	0,21	0,04	0,21
Itapirapuã Paulista	260	I	1	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Lavínia	300	II	2	85,46	6,42	0,97	2,39	0,51	0,80	1,59	0,11	0,45	0,06	0,17	0,85	0,00	0,11	0,06	0,06
Lindóia	305	I	5	75,89	13,48	1,38	2,44	0,60	1,33	2,99	0,05	0,55	0,00	0,05	0,87	0,05	0,09	0,09	0,14
Martinópolis	329	I	1	74,28	12,55	2,08	2,88	0,71	1,79	3,38	0,17	0,46	0,04	0,33	1,13	0,04	0,08	0,04	0,04
Miracatu	337	I	1	75,84	13,61	1,43	1,22	0,46	1,48	3,06	0,25	0,51	0,05	0,00	1,53	0,10	0,25	0,00	0,20
Mongaguá	349	II	2	71,77	16,32	1,66	2,40	0,40	1,89	3,06	0,15	0,38	0,04	0,17	1,28	0,06	0,15	0,09	0,17
Monte Castelo	354	I	1	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Motuca	359	II	2	74,54	17,34	0,82	1,09	0,34	1,37	2,32	0,27	0,41	0,00	0,14	0,96	0,00	0,20	0,07	0,14
Nova Aliança	368	II	2	84,94	6,35	1,49	2,15	0,97	0,97	0,97	0,15	0,41	0,05	0,31	0,92	0,00	0,20	0,05	0,05
Parisi	408	II	3	80,73	7,00	2,40	3,55	0,48	1,05	1,73	0,10	1,15	0,00	0,29	1,25	0,00	0,29	0,00	0,00
Paulicéia	410	I	1	79,75	12,33	1,47	1,61	0,76	1,00	1,85	0,05	0,24	0,00	0,00	0,76	0,05	0,00	0,05	0,09
Pirapora do Bom Jesus	438	I e II	5 e 2	83,00	9,88	1,01	1,75	0,47	0,67	1,61	0,03	0,40	0,00	0,10	0,77	0,07	0,13	0,00	0,10
Pradópolis	459	I	2	84,93	7,02	1,73	2,53	0,32	0,76	1,62	0,03	0,21	0,03	0,15	0,56	0,00	0,03	0,06	0,03
Redenção da Serra	475	II	3	68,19	14,78	2,85	2,94	1,04	2,25	4,15	0,26	1,12	0,00	0,43	1,73	0,00	0,26	0,00	0,00
Regente Feijó	476	II	2	79,82	7,77	1,66	3,94	0,68	0,90	2,77	0,05	0,60	0,03	0,08	1,44	0,00	0,11	0,08	0,08
Ribeirão Corrente	483	I	6	86,62	5,76	0,90	2,22	0,42	0,76	1,73	0,00	0,49	0,07	0,14	0,69	0,00	0,07	0,14	0,00
Rubiácea	499	I	1	80,51	9,16	1,49	2,13	0,75	1,38	2,77	0,21	0,53	0,11	0,00	0,64	0,11	0,11	0,11	0,00
Sandovalina	511	I	3	75,53	9,99	2,88	4,15	0,68	1,19	2,54	0,34	0,51	0,17	0,34	0,85	0,08	0,42	0,00	0,34
Santa Salete	533	II	3	75,48	8,58	2,32	5,31	0,82	1,09	3,27	0,27	1,09	0,14	0,27	1,23	0,00	0,14	0,00	0,00

Município	Código	Parte	Grupo	SD	DV	DO	DF	DI	DVO	DVF	DVI	DOF	DOI	DFI	DVOF	DVOI	DVFI	DOFI	DVOFI
São João das Duas Pontes	550	I	1	79,05	8,56	2,89	3,59	0,69	1,04	1,97	0,00	0,35	0,12	0,00	1,39	0,00	0,23	0,12	0,00
São João de Iracema	551	II	3	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Silveiras	580	I	6	85,64	5,31	1,42	2,68	0,89	1,16	1,37	0,00	0,37	0,05	0,21	0,68	0,11	0,05	0,05	0,00
Taquarivaí	601	II	4	82,69	10,84	1,34	1,22	0,52	0,93	1,57	0,17	0,00	0,06	0,06	0,58	0,00	0,00	0,00	0,00
Tarumã	603	I	5	75,73	13,09	1,68	2,69	0,47	1,09	3,12	0,08	0,43	0,00	0,27	1,25	0,00	0,08	0,00	0,04
Timburi	610	I	4	73,02	14,94	1,23	2,56	0,45	1,34	4,57	0,11	0,33	0,00	0,00	1,23	0,00	0,22	0,00	0,00
Tuiuti	616	I	3	78,38	12,33	1,33	1,85	0,66	1,75	1,80	0,05	0,57	0,14	0,09	0,85	0,00	0,14	0,00	0,05
Uru	627	I	2	66,40	17,75	1,58	2,38	1,11	2,69	2,54	1,27	0,63	0,16	0,32	2,22	0,00	0,63	0,16	0,16
Vinhedo	637	II	2	83,79	9,33	1,40	1,60	0,44	0,85	1,29	0,07	0,29	0,02	0,03	0,67	0,03	0,11	0,03	0,05
Viradouro	638	I	3	83,03	7,32	1,79	2,38	0,57	0,99	1,93	0,03	0,82	0,11	0,23	0,60	0,03	0,09	0,00	0,09

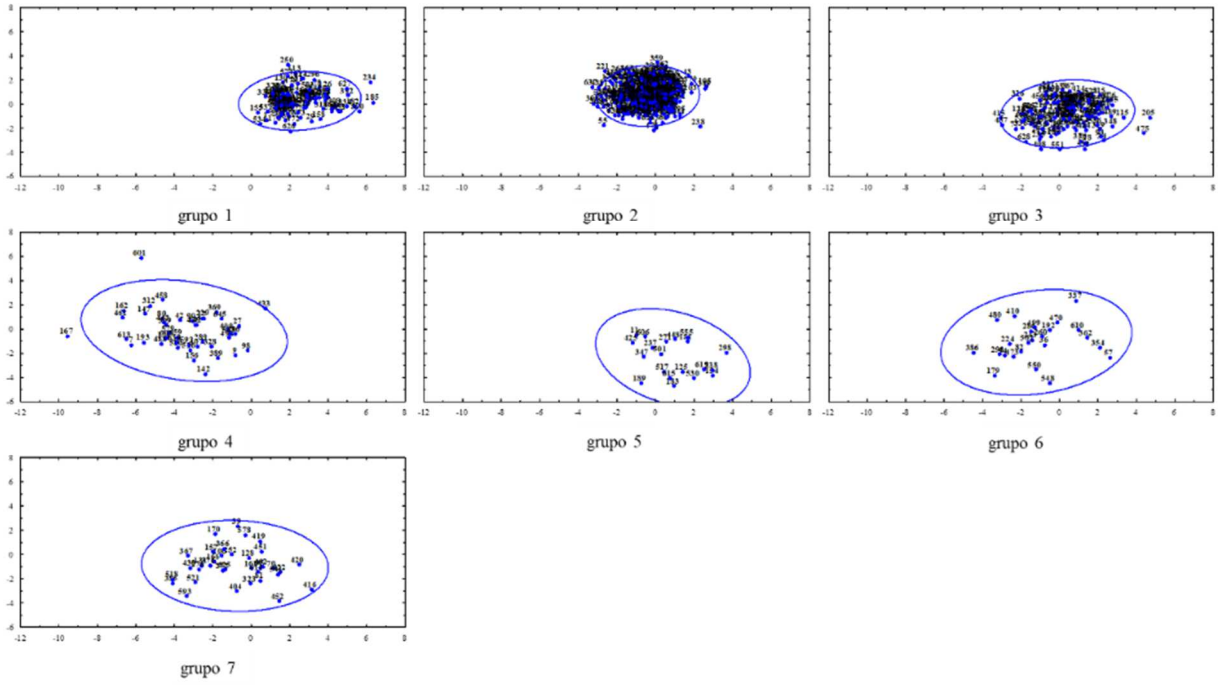
Nota - Pirapora do Bom Jesus foi considerada discrepante no grupo 5 da Parte I e no grupo 2 da Parte II. (...) Dado numérico não disponível: municípios identificados como discrepantes pela Figura 3 (Itapirapuã Paulista e Monte Castelo) e pela Figura 5 (São João de Iracema), cujas proporções não constam do material original

Fonte - Elaborado com base em dados do IBGE, Censo Demográfico 2010

Entre os municípios discrepantes, o que apresentou maior proporção de pessoas sem deficiência (SD) foi Embaúba (90,58%), e o de menor, Ilha Comprida (62,77%), cuja proporção de pessoas com deficiência (37,23%) é muito maior que a nacional (23,92%).

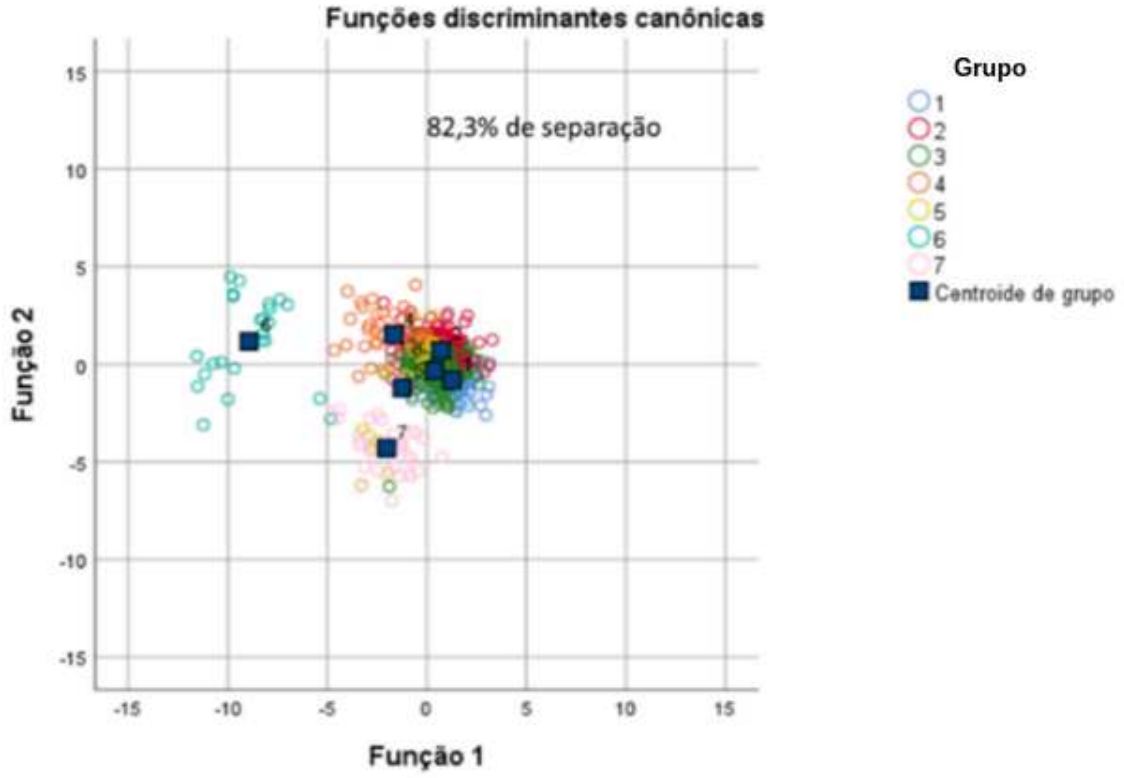
Figura 5 - Gráficos das funções discriminantes, por grupo (escores das componentes principais)

Os municípios discrepantes podem ser reunidos em três situações. Na primeira, a proporção de pessoas sem deficiência é muito inferior à média do grupo, com excesso de pessoas com deficiência em várias categorias: é o caso de Dirce Reis (SD de 68,36%, ante 78,77% no grupo 6, com DF de 5,07% e DVF de 5,56%, cerca do dobro das médias do grupo), Uru (SD de 66,40%; DVI de 1,27%, ante 0,18% no grupo), Guarani d'Oeste (SD de 66,43%; DO de 3,39% e DVOF de 2,59%), Redenção da Serra (SD de 68,19%), Florínia (SD de 63,30%) e Ilha Comprida (SD de 62,77%, com DV de 20,96%, ante 13,09% no grupo). Na segunda, ocorre o inverso: a proporção de pessoas sem deficiência é bem superior à média do grupo, como em Embaúba (90,58%, ante 83,38%), Ribeirão Corrente (86,62%), Gavião Peixoto (86,37%), Silveiras (85,64%) e Pradópolis (84,93%), com proporções muito baixas ou nulas de várias deficiências combinadas. Na terceira, a proporção de pessoas sem deficiência é próxima da média, mas a composição entre os tipos de deficiência é atípica, em geral pela ausência de pessoas em combinações pouco frequentes (DVI, DFI, DVFI) ou pelo excesso em uma delas, como em Arapeí (DF de 3,82% e nenhuma pessoa com DFI), São João das Duas Pontes (DO de 2,89% e DF de 3,59%) e Santa Salete (DF de 5,31%). Como essas combinações são raras, a presença ou a ausência de poucas pessoas pode pesar muito nas proporções de municípios com população pequena.



Fonte - Elaborado com base em dados do IBGE, Censo Demográfico 2010

Figura 6 - Gráfico das funções discriminantes para os grupos formados a partir dos escores das componentes principais



Fonte - Elaborado com base em dados do IBGE, Censo Demográfico 2010

Dos estudos comparativos entre os resultados obtidos utilizando transformações por logaritmo neperiano (Figuras 3 e 4 e Tabela 1) e utilizando escores das componentes principais (Figuras 5 e 6 e Tabela 2) foram obtidos os seguintes resultados:

- Nos dendrogramas, os dados transformados por logaritmo neperiano e os escores das componentes principais foram particionados em sete grupos, e não foram encontrados dados discrepantes. Acredita-se que isso se deva ao particionamento pelo método de Ward (grupos o mais homogêneos possível internamente e o mais heterogêneos possível entre si);
- Comparando as Figuras 3 e 5, a transformação por logaritmo neperiano (Figura 3) identificou o maior número de dados discrepantes, num total de 27, com ausência de valor discrepante apenas no grupo 7, enquanto na Figura 5 foram detectados 22 valores atípicos, com ausência de dado discrepante nos grupos 5, 6 e 7;
- Nas Figuras 4 e 6, nota-se boa discriminação em ambos os casos; porém, com a transformação por logaritmo neperiano, a separação foi de 97,7%, enquanto, com os escores das componentes principais, foi menor, de 82,3%. Como as funções discriminantes foram ajustadas aos próprios grupos formados pelo agrupamento dos mesmos dados, essas

porcentagens tendem a superestimar a capacidade de separação. Acredita-se que isso possa ter ocorrido devido a uma maior dispersão dos pontos dos dados composicionais ao serem transformados por logaritmo neperiano; e, finalmente,

- Comparando as Tabelas 1 e 2, a maior proporção de SD ocorreu no grupo 4 da Tabela 2 (83,38%), e a menor, no grupo 1, também da Tabela 2 (73,30%); por fim, o maior grupo foi o grupo 2 da Tabela 2, com 289 municípios.

Por deficiência, observaram-se os seguintes perfis:

- ✓SD (sem deficiência): a maior proporção ocorreu no grupo 4 com 83,4% e a menor foi no grupo 1 com 73,3%, ambos da Tabela 2;
- ✓ DV (deficiência visual): maior proporção foi no grupo 1 com 13,1%, enquanto a menor proporção foi no grupo 4 com 8,4%, ambos na Tabela 2;
- ✓ DO (deficiência auditiva): maior proporção foi no grupo 5 com 2,3%, enquanto a menor proporção foi no grupo 4 com 1,5%, ambos na Tabela 2;
- ✓ DF (deficiência física): maior proporção foi no grupo 5 com 3,4%, enquanto a menor proporção foi no grupo 4 com 2,1%, ambos na Tabela 2;
- ✓ DI (deficiência intelectual): maior proporção foi no grupo 1 com 0,7%, enquanto a menor proporção foi no grupo 4 com 0,5%, ambos na Tabela 2;
- ✓ DVO (deficiência visual e auditiva): maior proporção foi no grupo 1 com 1,6%, enquanto a menor proporção foi no grupo 4 com 1,0%, ambos na Tabela 2;
- ✓ DVF (deficiência visual e física): maior proporção foi no grupo 1 com 3,5%, enquanto a menor proporção foi no grupo 4 com 1,6%, ambos na Tabela 2;
- ✓ DVI (deficiência visual e intelectual): maior proporção foi no grupo 1 com 0,2%, enquanto a menor proporção foi no grupo 5 com 0,0%, ambos na Tabela 2;
- ✓ DOF (deficiência auditiva e física): maior proporção foi no grupo 5 com 0,7%, enquanto a menor proporção foi no grupo 4 com 0,3%, ambos na Tabela 2;
- ✓ DOI (deficiência auditiva e intelectual): maior proporção ocorreu nos grupos 2 e 3 com 0,1%, enquanto a menor proporção foi no grupo 5 com 0,0%, ambos na Tabela 1;
- ✓ DFI (deficiência física e intelectual): maior proporção foi no grupo 1 com 0,3%, enquanto a menor proporção foi no grupo 6 com 0,0%, ambos na Tabela 2;

- ✓ DVOF (deficiência visual, auditiva e física): maior proporção foi no grupo 1 com 1,6%, enquanto a menor proporção foi no grupo 4 com 0,8%, ambos na Tabela 2;
- ✓ DVOI (deficiência visual, auditiva e intelectual): maior proporção foi no grupo 1 da Tabela 2 com 0,1% e a menor foi no grupo 2 da Tabela 1 com 0,0%;
- ✓ DVFI (deficiência visual, física e intelectual): maior proporção foi no grupo 1 da Tabela 2 com 0,2%, enquanto a menor proporção foi no grupo 7 com 0,0% em ambas as tabelas;
- ✓ DOFI (deficiência auditiva, física e intelectual): maior proporção foi no grupo 2 com 0,1%, enquanto a menor proporção foi no grupo 3 com 0,0%, ambos na Tabela 1; e, finalmente,
- ✓ DVOFI (deficiência visual, auditiva, física e intelectual): maior proporção foi no grupo 7 da Tabela 2 com 0,2% e a menor foi no grupo 6 da Tabela 1 com 0,0%.

Da comparação entre as Tabelas 1 e 2 e a Figura 2, destaca-se que, com exceção dos grupos 1 e 5 da Tabela 2, todos os grupos têm proporção de pessoas SD maior que a do Brasil (76,08%).

Quanto às diferentes deficiências, 56 das 105 combinações de grupo e deficiência apresentaram proporções menores que as do Brasil, tanto na Tabela 1 quanto na Tabela 2.

Nota-se também que as diferenças dos resultados entre as Tabelas 1 e 2, neste caso, parecem pequenas.

4.CONCLUSÕES

- O grupo com maior quantidade de municípios foi o grupo 2 da Tabela 2, com 289 municípios, e o de menor número foi o grupo 5 da Tabela 2, com 20 municípios;
- A análise de agrupamentos, em ambos os casos, produziu grupos homogêneos internamente e heterogêneos entre si, o que dificultou a detecção de valores discrepantes pelo dendrograma;
- A transformação dos dados composicionais por logaritmo neperiano identificou mais municípios discrepantes (27) do que os escores das componentes principais (22), e o município de Pirapora do Bom Jesus foi considerado discrepante nas duas abordagens;
- Os resultados devem ser interpretados com algumas limitações. Esta versão não detalha o tratamento das proporções nulas antes da transformação logarítmica nem se foi usado o logaritmo de cada proporção ou a transformação log-razão centrada; não informa o número de componentes

principais retidas nem a matriz utilizada (covariância ou correlação); e não indica se as proporções municipais foram calculadas com os pesos amostrais do Censo. Além disso, como as funções discriminantes foram ajustadas aos próprios grupos formados, as taxas de separação tendem a ser otimistas, e as proporções de três municípios discrepantes não estão disponíveis na Tabela 3;

- Como propostas para trabalhos futuros, destacam-se:
  - ✓ Considerar outros estados e regiões e o país como um todo;
  - ✓ Trabalhar com escores de análise fatorial usando rotação varimax e extração por componentes principais;
  - ✓ Considerar outras variáveis com níveis composicionais como nível de instrução, tipo de trabalho, renda, raça e sexo, entre outras;
  - ✓ Considerar o conjunto de variáveis relacionadas a instrução, família, trabalho, renda, moradia e posse de bens; e, por fim,
  - ✓ Utilizar outras técnicas para detecção de valores discrepantes, como a análise de Procrustes.

5.REFERÊNCIAS BIBLIOGRÁFICAS

Aitchison, J. (1986). *The statistical analysis of compositional data*. London: Chapman and Hall. Reimpressão: Caldwell: The Blackburn Press, 2003.

Barbosa, J.J., Pereira, T.M. & Oliveira, F.L.P. (2018). Uma proposta para identificação de outliers multivariados. Ciência e Natura, 40(1). DOI: 10.5902/2179460X29535.

Barnett, V. & Lewis, T. (1994). *Outliers in statistical data* (3a ed.). Chichester: John Wiley & Sons.

Barroso, L.P. & Artes, R. (2003). Análise multivariada. In 48ª Reunião Anual da RBRAS e 10º SEAGRO. Lavras: UFLA, Departamento de Ciências Exatas.

Baxter, M.J. (1999). Detecting multivariate outliers in artefact compositional data. Archaeometry, 41(2), 321-338.

Brasil (1999). Decreto nº 3.298, de 20 de dezembro de 1999. Regulamenta a Lei nº 7.853, de 24 de outubro de 1989, dispõe sobre a Política Nacional para a Integração da Pessoa Portadora de Deficiência, consolida as normas de proteção, e dá outras providências. Diário Oficial da União, Brasília, DF, 21 dez. 1999.

Brasil (2015). Lei nº 13.146, de 6 de julho de 2015. Institui a Lei Brasileira de Inclusão da Pessoa com Deficiência (Estatuto da Pessoa com Deficiência). Diário Oficial da União, Brasília, DF, 7 jul. 2015.

Buccianti, A., Mateu-Figueras, G. & Pawlowsky-Glahn, V. (Eds.) (2006). *Compositional data analysis in the geosciences: from theory to practice*. London: Geological Society (Special Publications, 264).

Giroldo, F.R.S. (2008). *Alguns métodos robustos para detectar outliers multivariados*. (Dissertação de mestrado). Instituto de Matemática e Estatística, Universidade de São Paulo (IME-USP), São Paulo.

Hair Jr., J.F., Anderson, R.E., Tatham, R.L. & Black, W.C. (1996). *Multivariate data analysis*. Upper Saddle River: Prentice Hall.

IBGE – Instituto Brasileiro de Geografia e Estatística (2012). *Censo Demográfico 2010: características gerais da população, religião e pessoas com deficiência*. Rio de Janeiro: IBGE.

IBGE – Instituto Brasileiro de Geografia e Estatística (2018). *Nota técnica 01/2018: releitura dos dados de pessoas com deficiência no Censo Demográfico 2010 à luz das recomendações do Grupo de Washington*. Rio de Janeiro: IBGE.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. Philosophical Magazine, Series 6, 2(11), 559-572.

Ward Jr., J.H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58(301), 236-244.

AGRADECIMENTOS e COLABORAÇÕES

Agradecimentos ao IBGE pela disponibilização dos dados do Censo Demográfico 2010 e à Profa. Dra. Júlia Maria Pavan Soler pela sugestão e incentivo em trabalhar com o tema pessoa com deficiência.