REVISTA DE CIÊNCIAS AGRÁRIAS
Amazonian Journal
of Agricultural and Environmental Sciences
www.ajaes.ufra.edu.br

http://dx.doi.org/10.4322/rca.2014.014

ARTIGO ORIGINAL

Uso de análise multivariada no agrupamento de comunidades rurais

Use of multivariate analysis in the clustering of rural communities

Wilza da Silveira Pinto[1]*
Waldenei Travassos de Queiroz[1]
Manoel Malheiros Tourinho[1]
Paulo de Tarso Eremita da Silva[1]

[1]Universidade Federal Rural da Amazônia – UFRA, Instituto de Ciências Agrárias, Av. Presidente Tancredo Neves, 2501, Terra Firme, 66077-901, Belém, Pará, Brasil

Autor Correspondente:
*E-mail: wilza.pinto@ufra.edu.br

PALAVRAS-CHAVE
Análise de agrupamento
Faces de *chernoff*
Gráfico multidimensional

KEYWORDS
Cluster
Chernoff faces
Multidimensional graph

RESUMO: A identificação e a caracterização de agrupamentos de comunidades rurais em função de similaridades e diferenças são importantes para a proposição e a execução de políticas públicas focadas nas realidades locais, maximizando resultados e contribuindo para o aumento de sua eficácia, na medida em que são identificadas as comunidades com perfil semelhantes em função de sua estrutura social. O objetivo deste trabalho foi possibilitar, com o uso de técnicas de análise multivariada, a criação de grupos homogêneos de comunidades a serem assistidas por programas sociais e políticas públicas. As comunidades rurais com características semelhantes foram divididas em três grupos. O Grupo 1 congregou duas comunidades: Ribeira (C1) e Jacundaí (C2). O Grupo 2 abrangeu seis comunidades: São Manoel (C3), Santa Maria do Mirindeua (C4), Santo Cristo(C5), Centro Ouro (C7), São Sebastião (C8) e Santa Luzia do Tracuateua (C10). O Grupo 3, Santana do Baixo (C6) e Santa Maria do Tracuateua (C9). A partir da análise multivariada, constatatou-se que as comunidades apresentaram perfis diferenciados em relação às variáveis componentes da dimensão socioeconômica, confirmando os agrupamentos formados por algumas similaridades dentro dos grupos, assim como por heterogeneidade entre e dentro dos grupos.

ABSTRACT: *The identification and characterization of community clusters in terms of similarities and differences are important for the development and implementation of public policies focused on local realities, maximizing results and contributing to the increase of their effectiveness to the extent that communities of similar profile are identified according to their social structure. This paper aims at applying multivariate analysis to identify and create homogeneous groups of communities to be assisted by social programs and public policies. Specifically, the objective of this study was to typify, classify and characterize rural communities according to socioeconomic indicators and soil quality. Cluster analysis and graphical representation of Chernoff faces were used for this classification. Rural communities with similar characteristics were divided into three groups: Group 1: Ribeira (C1) and Jacundaí (C2); Group 2: São Manoel (C3), Santa Maria do Mirindeua (C4), Santo Cristo (C5), Centro Ouro (C7), São Sebastião (C8) and Santa Luzia do Tracuateua (C10); Group 3: Santana do Baixo (C6) and Santa Maria do Tracuateua (C9). The multivariate graph revealed that the communities present different profiles for the different variables in the socioeconomic dimension, confirming the clusters formed by some similarities within groups, as well as by heterogeneity between and within groups.*

Recebido: 05/08/2013
Aceito: 31/10/2013

1 Introdução

A identificação e a caracterização de agrupamentos de comunidades em função de similaridades e diferenças são importantes para a elaboração e a execução de ações públicas focadas nas realidades locais, maximizando resultados e contribuindo para o aumento de sua eficácia, na medida em que são sinalizadas as comunidades com perfil semelhante em função da estrutura social.

Não é possível se pensar em projetos ou programas de desenvolvimento local, sem considerar as realidades social, política e cultural das pessoas que ali vivem e produzem. Se assim não for, cada vez que se propõe um projeto que requeira esses aspectos, enfrentam-se problemas para sua execução e/ou manutenção. É necessário fixar prioridades às opções de solução que os projetos sugerem, objetivando implantar aqueles mais pertinentes às características e necessidades de cada comunidade a ser beneficiada. Para isto, são requeridos diagnósticos completos e confiáveis destas características e dos problemas e necessidades, de modo que a eficiência das políticas públicas possa ser incrementada significativamente se essas políticas tomarem em consideração as diferenças que, comumente, as comunidades apresentam entre si.

O uso da análise multivariada possibilita criar grupos homogêneos de comunidades a serem atendidas por programas e políticas públicas peculiares. Classificar e caracterizar comunidades rurais segundo indicadores socioeconômicos e de qualidade do solo podem ajudar no modo como atender cada comunidade ou cada grupo com projetos que sejam mais apropriados às características e necessidades, e aos entornos socioculturais e econômicos das mesmas.

Conforme Parsons (1969), os elementos estruturais de qualquer sistema social são: as unidades ou partes; os fatores de estruturação das relações; a interdependência dos atores e a coletividade, e o equilíbrio da organização. Estes elementos vão definir a estrutura social e a diferenciação entre sistemas sociais, dependendo de como esses elementos são desenvolvidos (FERRARI, 1983).

As estruturas sociais não são iguais na sociedade e variam segundo as oportunidades que são oferecidas à mobilidade social, que está relacionada com as variáveis nível de educação, ocupação, renda, papéis e *status* social dos atores na estrutura social (LOOMIS, 1960).

A análise de agrupamento (*Cluster analysis*), segundo Hair Junior et al. (2009), é um conjunto de técnicas estatísticas cujo objetivo é agrupar objetos segundo suas características, formando grupos ou agrupamentos homogêneos. Os objetos em cada agrupamento tendem a ser semelhantes entre si, porém diferentes dos demais objetos dos outros agrupamentos. Os grupos obtidos devem apresentar tanto uma homogeneidade interna (dentro de cada grupo), como uma grande heterogeneidade externa (entre grupos). Portanto, se a aglomeração for bem sucedida, quando representados em um gráfico, os objetos dentro dos agrupamentos estarão muito próximos e os agrupamentos distintos estarão afastados.

Para a aplicação da análise de agrupamento, inicialmente é necessário definir o problema de aglomeração e as variáveis a serem tratadas estatisticamente (MALHOTRA, 2001). Seleciona-se uma medida de distância dos conglomerados e se define o processo de aglomeração, que dependerá das variáveis em estudo e do problema em foco. Os agrupamentos resultantes devem ser interpretados em termos das variáveis usadas para constituí-los e de outras variáveis adicionais importantes. Finalmente, o pesquisador precisa avaliar a validade do processo de aglomeração.

Segundo Mingoti (2005), a técnica de Escalonamento Multidimensional – ou, simplesmente, EMD – é um procedimento matemático recomendável para representar graficamente um determinado número n de elementos ou objetos em um espaço de dimensão menor do que o original, usando como critério a distância entre os seus elementos. Os escalonamentos podem ser métricos e não métricos. Assim, as informações obtidas dos elementos amostrais são sintetizadas em q dimensões, tal que q seja menor ou igual ao número de objetos menos um, permitindo a construção de um gráfico de percepção que ofereça a visualização dos elementos amostrais com base na distância entre estes. Essas novas dimensões são construídas visando a preservar a similaridade ou as dissimilaridades entre os elementos amostrais entre si.

Complementando a análise de agrupamento, tem-se aplicado o método gráfico multivariado através do uso de disposições gráficas chamadas de técnicas iconográficas. Estas técnicas trabalham objetos geométricos com aparência paramétrica, que podem ser mapeados segundo atributos de uma base de dados (ESTIVALET; FREITAS, 2000). A ideia é mostrar as características essenciais de um domínio de dados por meio de ícones. Estas também são utilizadas para representações multidimensionais e podem ser compostas por atributos geométricos (forma, tamanho e orientação) e atributos de aparência (cor e textura), que podem ser associados aos itens de dados em análise.

Um dos primeiros trabalhos utilizando uma técnica baseada em ícones foi realizado por Chernoff (1973), mostrando que o ser humano tem sensibilidade a uma grande variedade de expressões faciais; o autor sugeriu, então, que ícones pudessem ser representados por faces, associando suas propriedades – tais como as formas da boca, cabelo e olhos – com atributo de dados. Desde que proposto por Chernoff (1973), o método de representação de dados multivariados graficamente por faces tem se tornado uma ferramenta de análise multivariada (FLURY; RIEDWYL, 1981). Este conjunto de técnicas tem como objetivo mapear os atributos em características particulares de ícones. Cada característica do ícone representa um atributo dos dados multidimensionais.

O objetivo deste trabalho foi possibilitar, com o uso de técnicas de análise multivariada, a criação de grupos homogêneos de comunidades a serem assistidas por programas sociais e políticas públicas.

2 Material e Métodos

O estudo foi desenvolvido na microrregião de Tomé-Açu, Município de Moju, Território Quilombola do Jambuaçu, nas seguintes comunidades: Ribeira, Jacundaí, São Manoel, Santa Maria do Mirindeua, Santo Cristo, Santana do Baixo, Bom Jesus do Centro Ouro, Santa Maria do Tracuateua, São Sebastião e Santa Luzia do Tracuateua.

A coleta de dados se deu com a aplicação de questionários nas comunidades objeto do estudo**,** do território quilombola do Jambuaçu. Os dados foram sistematizados em uma planilha do Excel, adotando-se uma nomenclatura para cada questão de acordo com a necessidade de organização dos dados para análises estatísticas univariada e multivariada.

Foram aplicados 298 questionários em uma população de 386 famílias, sendo observadas as seguintes variáveis respostas: carbono orgânico do solo (CO), produtividade da mandioca (PM), renda total (RT), tamanho da unidade de produção (TUP), área cultivada pela família (AC) e Área disponível em ha (AD). Essas variáveis foram analisadas pelas técnicas estatísticas multivariadas de agrupamento (*Cluster analisys*) e de escalonamento multidimensional, utilizando os programas computacionais MINITAB 15 e SPSS 19, respectivamente. Para obtenção dos gráficos dos perfis (*Faces de Chernoff*), foi acrescentada a variável relação território/ família (RTF) e usado o programa *Statistical Trial.*

Como o objetivo da análise de agrupamento é agrupar objetos semelhantes, é necessário obter uma medida da distância entre os mesmos. Os objetos com menor distância entre si são mais semelhantes; logo, são reunidos em um mesmo agrupamento. Já os mais distantes participam de agrupamentos distintos. Existem várias formas de medir a distância entre os objetos, porém a mais utilizada é a Distância Euclidiana, utilizada neste estudo. Esta medida de distância corresponde à soma dos quadrados das diferenças entre dois objetos para todas as variáveis.

Neste estudo, utilizou-se o processo de aglomeração hierárquico, que se caracteriza pelo estabelecimento de uma hierarquia ou estrutura em forma de árvore. Para formação dos agrupamentos, foi utilizado o Método de *Ward,* cujo objetivo é minimizar o quadrado da distância euclidiana às médias dos agrupamentos. Com o método, busca-se agrupar os agregados que apresentam a menor soma dos quadrados entre dois agrupamentos, calculada sobre todas as variáveis. Trata-se de um método que tende a proporcionar agregados com aproximadamente o mesmo número de observações (HAIR JUNIOR et al., 2009), sendo o mais utilizado em estudos de agrupamento na atualidade.

O método Escalonamento Multidimensional (EMD) foi aplicado aos dados com o objetivo de mapear as distâncias entre pontos, visando a obter uma representação gráfica espacial, o que permitiu identificar dimensões inerentes às avaliações feitas para determinados objetos. O EMD pode ser considerado como um procedimento de análise de agrupamento e foi utilizado neste trabalho como forma de confirmar o agrupamento anterior.

Para compor o estudo, levou-se em consideração a situação das comunidades em relação a três blocos de indicadores: social, econômico e qualidade do solo. No que se refere ao indicador social, foi selecionada a variável renda total; para o indicador econômico, foram utilizadas as variáveis produtividade da mandioca, tamanho da unidade de produção, área cultivada/ano/família e área disponível para cultivos sucessivos; finalmente, como indicador da qualidade do solo, selecionou-se a variável carbono orgânico da matéria orgânica do solo.

As variáveis foram padronizadas para converter cada escore original em um valor padronizado, em que a média é igual a zero e o desvio-padrão é igual a um. Essa transformação é necessária para eliminar a distorção introduzida pelas diferentes escalas das variáveis usadas na análise.

Como complemento da análise de agrupamento sobre os perfis multivariados foi então aplicado o método gráfico multivariado através do uso de disposições gráficas chamadas de técnicas iconográficas. Para analisar então os diferentes perfis das dez comunidades estudadas do Território Jambuaçu, foi selecionada a técnica *Faces de Chernoff* (CHERNOFF, 1973). A Tabela 1 apresenta as comunidades e os valores das variáveis respostas utilizadas nas análises multivariadas.

3 Resultados e Discussão

Aplicando-se o programa estatístico MINITAB 15 aos dados da Tabela 1**,** foi feita a análise de agrupamento de acordo com o Método de *Ward*, o qual possibilita a forma de visualização do resultado chamada de dendrograma (Figura 1). As linhas verticais representam os conglomerados unidos e as linhas horizontais, a distância euclidiana entre os mesmos.

Tabela 1. Comunidades e variáveis respostas utilizadas nas análises multivariadas.

Comunidades	CO	PM	RT	TUP	AC	AD	RTF
	(g kg^{-1})	(kg ha^{-1})	(R$)		(ha)		
Ribeira	13,35	13.200	14.517,44	7,05	3,02	2,91	21,02
Jacundaí	12,96	13.768	12.646,00	13,83	1,58	1,78	27,01
São Manoel	18,16	7.195	9.119,47	13,76	2,80	2,59	30,07
S. Maria do Mirindeua	14,29	7.330	10.500,00	18,71	1,79	4,10	22,9
Santo cristo	15,25	12.222	9.143,61	16,39	1,61	2,83	29,45
Santana do Baixo	17,47	11.000	11.321,60	36,50	1,00	4,10	36,07
Centro Ouro	14,53	8.100	10.704,92	9,50	1,21	1,98	26,61
São Sebastião	13,87	10.916	11.683,95	10,74	1,30	3,79	19,63
Santa Maria do Tracuateua	13,50	12.964	9.440,09	23,96	1,02	4,45	23,15
Santa Luzia do Tracuateua	14,58	7.200	10.871,14	11,36	0,96	3,68	10,7

CO= Carbono orgânico de solo seco; PM= Produtividade da mandioca; RT= Renda total; TUP= Tamanho da Unidade de Produção; AC= Área cultivada por ano pela família; AD= Área disponível; RTF=Relação território/família.

A partir deste gráfico, pode-se decidir sobre o número de agrupamentos.

Em conformidade com a Figura 1, verifica-se que as comunidades rurais foram divididas em três grupos com características semelhantes. O Grupo 1 foi formado por duas comunidades: Ribeira (C1) e Jacundaí (C2). O Grupo 2 agregou seis comunidades: São Manoel (C3), Santa Maria do Mirindeua (C4), Santa Luzia do Tracuateua (C10), São Sebastião (C8), Santo Cristo (C5) e Centro Ouro (C7). O grupo 3 reuniu duas comunidades: Santana do Baixo (C6) e Santa Maria do Tracuateua (C9).

Definindo-se 'Δ' como a matriz quadrada de ordem *n*, cujos elementos são as distâncias correspondentes (dissimilaridade) aos dados originais; seja D a matriz quadrada de ordem *n,* mas construída a partir dos dados correspondentes às distâncias estimadas dentro da nova dimensão resultante da aplicação do método de escalonamento multidimensional que, neste trabalho, resultou ser bidimensional (duas dimensões). A solução ideal para o escalonamento multidimensional corresponde à existência de uma correspondência máxima entre as distâncias contidas nessas matrizes.

Neste trabalho, essa correspondência foi muito eficiente, considerando-se que a técnica de escalonamento multidimensional confirmou os grupos formados na análise de agrupamento produzidos pelo método de Ward. No gráfico do escalonamento multidimensional (Figura 2), pode-se verificar a alocação das comunidades também formando grupos similares.

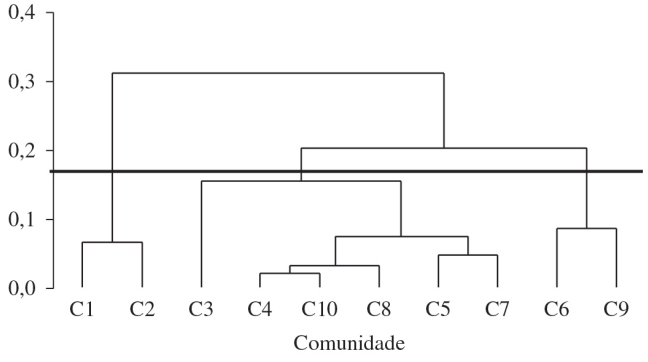

Figura 1. Dendrograma dos agrupamentos das comunidades segundo o método de *Ward*.

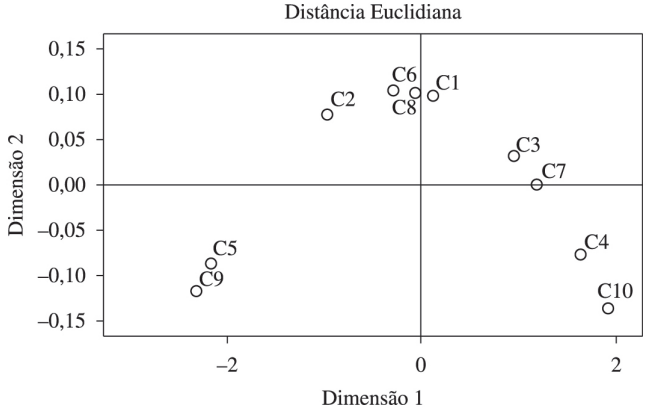

Figura 2. Representação gráfica do agrupamento considerando as duas dimensões resultantes obtidas no escalonamento multidimensional.

Para medir a adequação do ajuste, Kruskal (1964) propôs uma medida denominada Stress (*Standardized residual sum of squares*), para avaliar o quanto se aproximam as distâncias derivadas das distâncias correspondentes aos dados originais (medidas de dissimilaridade).

Para Hair Junior et al. (2009), o EMD – também conhecido como mapeamento perceptual – permite, ao pesquisador, determinar a imagem relativa percebida de um conjunto de objetos. O resultado da análise de escalonamento multidimensional, conforme pode ser verificado na Figura 2, confirmou completamente as informações obtidas pelo agrupamento feito pelo Método de *Ward*, ou seja, a formação dos três grupos com as mesmas composições.

O carbono orgânico do solo (CO) é a variável que controla a forma da face (Figura 3). Para valores maiores de carbono a face toma uma forma mais arredondada (São Manoel, 18,16 g kg^{-1}) e o inverso, uma forma elíptica (Jacundaí, 12,96 g kg^{-1}).

A produtividade da mandioca (PM) controla a altura da orelha. Neste caso, Jacundaí, que apresenta a maior média de produtividade (13.768,46 kg ha^{-1}), tem as orelhas numa posição superior, e a Comunidade de Santa Luzia do Tracuateua, que apresenta a menor média (7.200 kg ha^{-1})**,** tem as orelhas na parte inferior da face.

A renda total (RT) controla a altura da face. Conforme se verifica na Figura 3, quanto maior o valor da renda, mais longilínea é a face da comunidade Ribeira, com média de R$ 14.517,44 ano^{-1}. O menor valor é representado por São Manoel, com média de R$ 9.119,47 ano^{-1}, portanto uma face menos longilínea.

O tamanho da unidade de produção (TUP) controla a parte de cima da face. Valores superiores correspondem à face mais larga na parte superior (Santana do Baixo, com média de 36,50 ha), enquanto que o menor valor é representado por uma face com a parte superior mais estreita (Ribeira, com média de 7,05 ha).

A área cultivada/ano/família (AC) controla a parte de baixo da face (queixo). Para valores maiores, o queixo é mais largo (Ribeira, com a média de 3,02 ha), e para valores menores, mais estreito (Santa Luzia do Tracuateua, com a média de 0,96 ha).

O controle do tamanho do nariz é dado pela área disponível para cultivos sucessivos (AD). Para valores maiores, que é o caso de Santa Maria do Tracuateua, com média de 4,45 ha, maior é o nariz, e para o menor valor, o menor nariz; neste caso, Jacundaí, com a menor média de área disponível (1,78 ha).

O controle da altura da boca em relação ao nariz é dado pela variável razão território/família e, então, dependendo dos valores, a boca fica mais próxima ou mais afastada do nariz. Para valores maiores (Santana do Baixo, 36,07 ha/família), a boca se apresenta distante do nariz e, para valores mais baixos (Santa Luzia do Tracuateua, 10,70 ha/família), a boca fica mais próxima do nariz.

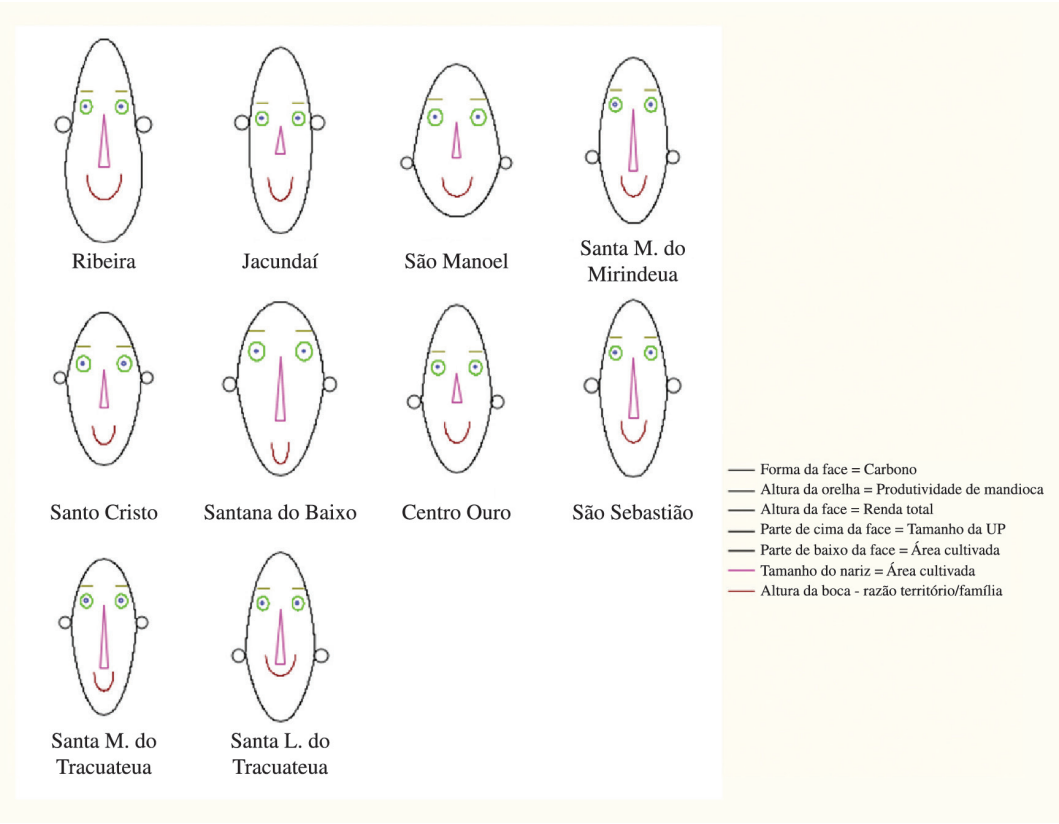

Figura 3. Configuração multivariada (*Faces de Chernoff*) do sistema social do Território Quilombola do Jambuaçu.

4 Conclusões

As comunidades apresentaram perfis diferenciados em relação às diferentes variáveis selecionadas na dimensão socioeconômica, confirmando os agrupamentos formados a partir da análise de agrupamento. Existem algumas similaridades dentro dos grupos, assim como heterogeneidade entre grupos e dentro dos grupos.

Os métodos aplicados neste estudo parecem promissores, pois produziram resultados que não são facilmente obtidos por análise estatística padrão e que possam ser visualizados de forma abrangente.

As dez comunidades estudadas, apesar de mostrarem alguma similaridade, apresentam estruturas sociais distintas. Isso pode induzir à aplicação de políticas e intervenções diferenciadas, que atendam às expectativas e aos anseios dos atores locais, e contribuam para o desenvolvimento comunitário.

Referências

CHERNOFF, H. The use of faces to represent points in K-Dimensional space graphically. *Journal of the American Statistical Association*, v. 68, n. 342, p. 361-367, 1973. http://dx.doi.org/10.1080/01621459.1973.10482434

ESTIVALET, L. F.; FREITAS, C. M. D. S. *O Uso de ícones na visualização de informações*. 2000. Dissertação (Mestrado em Computação)-Universidade Federal do Rio Grande do Sul, Porto Alegre, 2000.

FERRARI, A. T. *Fundamentos de Sociologia*. São Paulo: McGraw-Hill do Brasil, 1983.

FLURY, B.; RIEDWYL, H. Graphical Representation of Multivariate Data by Means of Asymmetrical Faces. *Journal of the American Statistical Association*, v. 76, n. 376, p. 757-765, dez. 1981. http://dx.doi.org/10.1080/01621459.1981.10477718

HAIR JUNIOR, J. F.; ANDRESON, R. E.; TATHAM, R. L.; BLACK, W. C. *Análise multivariada de dados*. São Paulo: Bookman, 2009.

KRUSKAL, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, v. 29, n. 1, p. 1-27,1964. http://dx.doi.org/10.1007/BF02289565

LOOMIS, C. *Social System*. D. Van Nostrand Co., 1960. PMCid:PMC405962.

MALHOTRA, N. K. *Pesquisa de marketing:* uma orientação aplicada. 3. ed. Porto alegre: Bookman, 2001.

MINGOTI, S. A. *Análise de dados através de métodos de estatística multivariada:* uma abordagem aplicada. Belo Horizonte: UFMG, 2005. PMid:16021257.

PARSONS, T. *Sociedades*: perspectivas evolutivas e comparativas. São Paulo: Pioneira, 1969.