Anais do 13º Encontro de Iniciação Científica e Pós-Graduação do ITA – XIII ENCITA / 2007
Instituto Tecnológico de Aeronáutica, São José dos Campos, SP, Brasil, Outubro, 01 a 04, 2007.

TRATAMENTO MULTIVARIADO DE DADOS POR ANÁLISE DE CORRESPONDÊNCIA E ANÁLISE DE AGRUPAMENTOS

Luciene Bianca Alves
ITA –Instituto Tecnológico de Aeronáutica
Praça Marechal Eduardo Gomes, 50 – Vila das Acácias
CEP 12228-900, São José dos Campos, SP
Bolsista Capes
luciene@ita.br

Mischel Carmen N. Belderrain
ITA –Instituto Tecnológico de Aeronáutica
Praça Marechal Eduardo Gomes, 50 – Vila das Acácias
CEP 12228-900, São José dos Campos, SP
carmen@ita.br

Rodrigo Arnaldo Scarpel
ITA –Instituto Tecnológico de Aeronáutica
Praça Marechal Eduardo Gomes, 50 – Vila das Acácias
CEP 12228-900, São José dos Campos, SP
rodrigo@ita.br

Resumo. *O objetivo deste trabalho é apresentar considerações de como a técnica de Análise de Agrupamentos pode auxiliar na interpretação dos resultados da Análise de Correspondência no tratamento multivariado de dados. Tal auxílio proporciona um melhor ajuste da associação das variáveis em análises que requerem soluções gráficas em maiores dimensões.*

Palavras chave: Análise Multivariada, Análise de Correspondência, Análise de Agrupamentos

1. Introdução

De acordo com Alves (2007), a demanda crescente de informações implica um melhor conhecimento de técnicas para a organização e interpretação de dados assim como para a interpretação dos resultados que podem ser gerados em cada tipo de aplicação. Nesse contexto, a Análise Multivariada dispõe de uma diversidade de técnicas que favorecem o entendimento de muitos fenômenos.

As técnicas existentes podem ser de diversas naturezas e devem ser utilizadas de acordo com o interesse da pesquisa. Dados qualitativos ou quantitativos podem restringir a aplicação de algumas delas. Assim, a escolha da técnica é justificada basicamente pelo que se pretende investigar através de um conjunto de dados.

Contudo, a especificação de uma determinada técnica não implica em desapossar a combinação de uma segunda técnica para atingir resultados. Uma análise conjunta, quando bem estruturada, pode revelar melhores respostas para uma análise.

As técnicas de Análise de Correspondência e Análise de Agrupamentos, consideradas técnicas de interdependência por analisar simultaneamente todas as variáveis, requerem a inclusão de diferentes tipos de dados para sua análise. Na técnica de Análise de Correspondência, tal inclusão é restrita a dados discretos (variáveis categóricas), enquanto na Análise de Agrupamentos é restrita à dados contínuos.

Este trabalho apresenta considerações sobre a utilização de ambas as técnicas ressaltando os seus atributos principais e diferentes aspectos quanto as suas aplicações convergindo para um mesmo objetivo. Logo, torna-se relevante a divulgação por pesquisas voltadas para esse fim, pois possíveis combinações tendem a otimizar a interpretação e, consequentemente, os resultados de uma determinada análise.

2. Técnicas de Análise Multivariada

A denominação "Análise Multivariada" corresponde a um conjunto de métodos e técnicas que analisam simultaneamente todas as variáveis na interpretação teórica do conjunto de dados. O primeiro passo para a utilização da análise multivariada é saber o que se pretende afirmar a respeito dos dados. A técnica e o método estatístico ideal para a aplicação devem ser escolhidos de acordo com o objetivo da pesquisa. Há diversas técnicas para a análise multivariada e cada uma delas se aplica a um objetivo de pesquisa específico.

Quanto à organização ou divisão das técnicas, alguns autores apresentam uma estrutura particular em seus trabalhos. Para Hair *et. al* (2005), tais técnicas são estabelecidas em dois grupos: a) Como técnicas de Dependência - Regressão Múltipla e Correlação múltipla, Análise Conjunta, Análise Discriminante Múltipla, Modelos Lineares de Probabilidade (*Logit* e *Probit*), Análise de Correlação Canônica, Análise Multivariada de Variância (MANOVA), Análise Multivariada de Covariância (MANCOVA), e Modelagem de Equações Estruturais; e b) Como técnicas de

Interdependência - Análise Fatorial, Análise de Agrupamentos, Escalonamento Multidimensional e Análise de Correspondência.

Como pode ser observado, as técnicas de Análise de Correspondência e Análise de Agrupamentos são classificadas como técnicas de interdependência, por se tratar de uma análise simultânea de todas as variáveis em estudo, na tentativa de se encontrar uma estrutura subjacente ao conjunto inteiro de variáveis. No entanto, cabe ressaltar que a primeira técnica se refere à inclusão de dados nominais, enquanto na Análise de Agrupamentos, tal inclusão menciona que os dados sejam mensurados em escala numérica contínuos.

2.1 Análise de Correspondência

A concepção geral da Análise de Correspondência (AC), entre outros aspectos, é que nela se permite a inclusão de variáveis categóricas, apropriadas para dados nominais. Recebe o nome de Análise de Correspondência pelo fato de estar interessada em estudar a correspondência entre as variáveis.

Conforme Carvalho e Struchiner (1992), este método permite a visualização gráfica das relações mais importantes de um grande conjunto de variáveis entre si (categóricas e contínuas categorizadas). A AC parte de uma matriz de dados representados por uma Tabela de Contingência e converge para um gráfico que exibe as linhas e as colunas da matriz como pontos de um espaço vetorial de dimensão menor que a original, de maneira a estabelecer relações entre linhas, colunas, e entre linhas e colunas, que possam ser interpretáveis (Greenacre e Hastie, 1984).

A AC Simples (AC de Tabelas de Contingência) em sua forma básica consiste na aplicação de tabelas de contingência de dupla entrada. Ao se tratar de tabelas de contingência com múltiplas entradas, a aplicação se refere a AC Múltipla (ACM) (AC de matrizes indicadoras e/ou matrizes de Burt). Segundo Greenacre (1987), a geometria da AC simples fornece as regras básicas para a sua interpretação. Todas as outras formas de AC são aplicações do mesmo algoritmo e outros tipos de matrizes de dados, com adaptações na sua interpretação.

De acordo com Faria (1993), na representação bidimensional, é possível verificar a projeção dos pontos originais sobre esse plano, mas não é possível averiguar quais deles estão mais próximos ou mais distantes. O ideal é ter o conhecimento da geometria de um conjunto de pontos em um espaço multidimensional. Desse modo, a análise pode ser realizada através de um gráfico aproximado, de menor dimensão, e assim, identificar onde esse gráfico é preciso e onde não o é. Em outras palavras, é construir um modelo e saber onde os seus dados se ajustam ou não, visando melhorar a qualidade de ajuste aos dados

A partir dos princípios geométricos da AC, é possível representar dentro do Espaço Euclidiano as distâncias entre os pontos linha e/ou coluna resultantes da associação entre as variáveis da tabela de contingência. Assim tem-se o gráfico denominado "Mapa de Correspondência" ou "Mapa Perceptual" que facilita a visualização das relações existentes entre as variáveis. (Lourenço, 1997).

Um dos pontos relevantes dessa técnica é que não há exigência de normalidade para a resposta estudada, e por conseqüência, os testes estatísticos aqui não são utilizados. Logo, a distribuição gráfica de seus resultados sugere a sua solução (Alves, 2007).

2.2 Análise de Agrupamentos

De acordo com Mingoti (2005), a Análise de Agrupamentos também é conhecida como Análise de Conglomerados ou Análise de Classificação ou *Cluster Analysis*. Seu objetivo é agrupar os elementos da amostra ou população em grupos. Os elementos de um mesmo grupo são homogêneos entre si, no que se refere às variáveis (características) que neles foram medidas. Por outro lado estes grupos já formados são heterogêneos entre eles em relação a estas mesmas características.

Segundo Hair *et al*.(2005), o objetivo principal da Análise de Agrupamentos é situar as observações homogêneas em grupos, a fim de definir uma estrutura para os dados. Para isto, são abordadas algumas questões básicas que devem ser consideradas durante a análise.

A primeira decisão na análise se refere à medida de similaridade que deve ser estabelecida. Ou seja, deve-se estabelecer a associação de dois objetos baseada nas variáveis da 'variável estatística de agrupamento'. Hair *et. al* (2005) define a 'variável estatística de agrupamento' como "o conjunto das variáveis que representam as características usadas para comparar objetos na análise de agrupamentos".

Para Mingoti (2005), é indispensável decidir *à priori,* a medida de similaridade que será utilizada para se proceder ao agrupamento de elementos. Para isto, existem medidas apropriadas para análise de variáveis qualitativas e quantitativas. As medidas apropriadas para variáveis quantitativas também são ditas 'de dissimilaridade'. Neste caso, quanto menores os seus valores, mais similares serão os elementos que estão sendo comparados. Algumas dessas medidas de similaridade são: Distância Euclidiana, Distância Generalizada (ou Ponderada) e Distância de Minkowsky.

A segunda decisão na análise se refere à formação do agrupamento do método hierárquico a ser empregado. É também um método aglomerativo, pelos agrupamentos serem formados pela combinação de outros já existentes, explica Hair *et. al*.. Existem vários métodos de agrupamentos hierárquicos e a maioria já se encontram disponíveis nos software estatísticos. À saber, Método de Ligação Simples, Método de Ligação Completa, Método de Ligação Média, Método do Centróide, Método de Ward, entre outros.

A última decisão na análise refere-se à escolha do número de agrupamentos. Deve-se haver um equilíbrio entre definir a estrutura mais básica com o nível de similaridade dentro dos agrupamentos porque quando o número de agrupamento diminui, a homogeneidade dentro dos grupos necessariamente diminui (HAIR, 2005).

Alguns dos critérios para a determinação do número ideal de agrupamentos citados por Mingoti (2005) são: Análise de Comportamento do Nível de Fusão (distância), Análise de Comportamento do Nível de Similaridade, Análise da Soma dos Quadrados entre Grupos, Correlação Semiparcial e Estatística Pseudo T^2.

Esta técnica dispõe do gráfico denominado Dendograma, em forma de árvore, no qual o nível de similaridade (ou dissimilaridade) é indicado na escala vertical. No eixo horizontal são relatados os elementos amostrais na ordem conveniente ao agrupamento (Mingoti, 2005).

Para maiores conhecimentos das medidas de similaridades, dos métodos de agrupamentos, e dos critérios de escolha do número de agrupamentos, consultar Mingoti (2005) e Hair *et. al.* (2005).

3. Tratamento de Dados por Análise de Correspondência e de Agrupamentos

A técnica de Análise de Agrupamentos pode ser utilizada para auxiliar na interpretação dos resultados da Análise de Correspondência. Esta recorrência se dá pela possibilidade de se obter a mesma disposição gráfica perceptual da AC no dendograma sucedido da Análise de Agrupamentos, o que permitirá uma melhor visualização das variáveis associadas.

Na maioria das vezes em que se pretende estudar a associação das varáveis utilizando-se da Análise de Correspondência, em especial, da AC Múltipla, é comum se deparar com grandes quantidades de variáveis que constituem de muitas categorias. Este fato pode resultar em uma baixa interpretação numa solução gráfica bidimensional, o que pode requerer uma análise em dimensões muito maiores. Logo, sabe-se que se torna impossível uma visualização gráfica multidimensional a partir da terceira dimensão.

Pelo exposto anteriormente propõe-se a interferência de uma segunda técnica de análise. Dado que no início do estudo o pesquisador esteja trabalhando com seus dados de entrada composto por variáveis categóricas, característica fundamental da AC, ela permite, além de todas as combinações gráficas dimensionais possíveis, a disposição das coordenadas das colunas das variáveis em cada dimensão. A partir de então, estes valores das coordenadas representam dados contínuos que retratam valores de distâncias.

Para um melhor entendimento, suponha que se tenha uma análise dos dados contendo os autovalores e percentuais de variação explicativa e acumulada de cada dimensão d_i (i: 1, ..., n). De acordo com Hair *et al.* (2005), o pesquisador deverá realizar uma observação comparativa e divergente sobre a variância adicional explicada referente à complexidade crescente na interpretação dos resultados. Assim, se a decisão da análise for sugerida por uma solução gráfica de dimensão d_k, tal que $3 < k \leq n$ para uma melhor visualização da associação das variáveis, torna-se necessário observar a localização das coordenadas das mesmas em cada dimensão d_i (i: 1, ..k) para se avaliar as distâncias entre elas. Sobre tais circunstâncias expostas, é sugerido o emprego da técnica de Análise de Agrupamentos, que possibilitará a disposição de tais variáveis sobre o dendograma, baseada nas distâncias das coordenadas nessas k dimensões. Ou seja, trata-se de uma técnica com recursos apropriados para esta aplicação, já que tem por finalidade agrupar objetos de uma amostra, de maneira que aqueles pertencentes a um mesmo grupo serão mantidos homogêneos entre si em relação às distâncias mais próximas e, por outro lado, heterogêneos em relação as mais distantes.

3.1 Análise Multidimensional dos Dados

A título de ilustração, será apresentado um exemplo de aplicação da análise de Correspondência Múltipla, com o mapa perceptual correspondente. O estudo se refere à associação de variáveis relativas à saúde física e mental de idosos de uma comunidade. Neste trabalho não serão apresentados os resultados obtidos, mas somente a disposição das variáveis no Mapa Perceptual Bidimensional. Trata-se de um estudo com cinco variáveis, constituídas de 13 categorias, sendo elas: gênero (feminino, masculino); idade (60-69 anos, 70-79 anos, 80+ anos); demência (sim, não); depressão (sim, não) e autonomia (sim, não).

Ao se aplicar a técnica de Análise de Correspondência sobre os dados, foi fornecida a seguinte composição gráfica, conforme exposta na Figura 1:

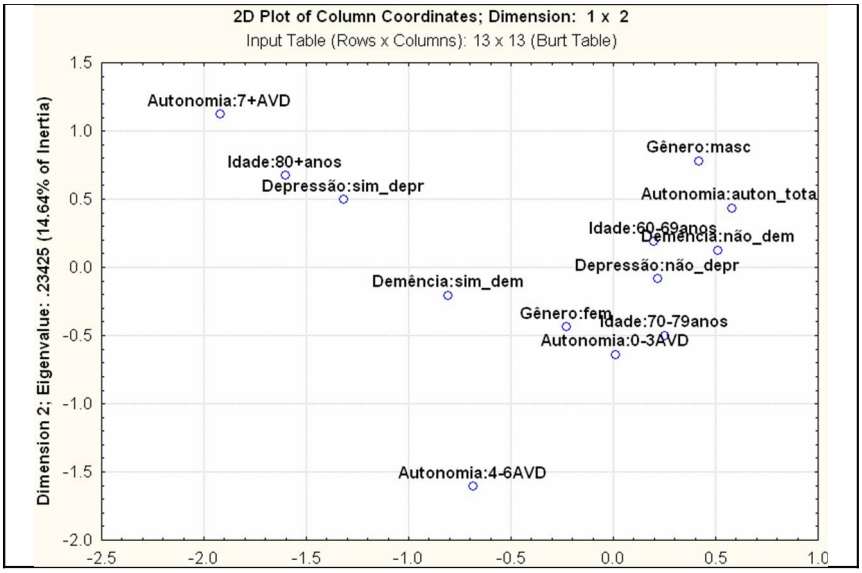

Figura 1 – Mapa Perceptual Bidimensional.

De acordo com a Figura 2, ao se analisar o percentual de variação explicada (percentual cumulativo) com a interpretabilidade (autovalor), foi decidido averiguar a associação das variáveis com 100% de explicação das variáveis, já que a interpretabilidade se manteria a um nível de 10%. Dessa forma, a análise gráfica se daria na oitava dimensão, o que não é possível obter.

Eigenvalues and Inertia for all Dimensions (base_dados.sta)
Input Table (Rows x Columns): 13 x 13 (Burt Table)
Total Inertia=1.6000

Number of Dims.	Singular Values	Eigen-Values	Perc. of Inertia	Cumulatv Percent	Chi Squares
1	0.595777	0.354950	22.18438	22.1844	470.6699
2	0.483993	0.234249	14.64058	36.8250	310.6185
3	0.468028	0.219051	13.69066	50.5156	290.4647
4	0.462491	0.213898	13.36865	63.8843	283.6329
5	0.415577	0.172704	10.79400	74.6783	229.0084
6	0.408053	0.166507	10.40669	85.0850	220.7911
7	0.367692	0.135197	8.44984	93.5348	179.2740
8	0.321626	0.103443	6.46521	100.0000	137.1678

Figura 2 – Autovalor e Percentual Cumulativo.

A partir de então, foram observadas as coordenadas das variáveis nas oito dimensões para se analisar as distâncias entre as variáveis. Foi utilizada a técnica de Análise de Agrupamentos para agrupar as variáveis conforme a distância apresentada entre as mesmas. Os dados referentes estão apresentados na Figura 3.

Column Coordinates and Contributions to Inertia (base_dados.sta)
Input Table (Rows x Columns): 13 x 13 (Burt Table)
Total Inertia=1.6000

Row Name	Row Number	Coordin. Dim.1	Coordin. Dim.2	Coordin. Dim.3	Coordin. Dim.4	Coordin. Dim.5	Coordin. Dim.6	Coordin. Dim.7	Coordin. Dim.8
Gênero:fem	1	-0.23265	-0.43517	0.03140	-0.17426	-0.406208	0.322578	0.10692	0.03079
Gênero:masc	2	0.41822	0.78226	-0.05644	0.31326	0.730207	-0.579872	-0.19221	-0.05535
Idade:60-69anos	3	0.19471	0.19166	-0.37311	0.70065	-0.417407	0.086532	0.02217	-0.19814
Idade:80+anos	4	-1.60539	0.67379	-0.36053	-1.27891	-0.555914	-0.951229	-0.68315	0.95679
Idade:70-79anos	5	0.24936	-0.49704	0.65428	-0.58435	0.782226	0.189147	0.19321	-0.03078
Demência:sim_dem	6	-0.80843	-0.20179	-0.31422	0.19174	0.217756	-0.396263	0.73982	-0.02933
Demência:não_dem	7	0.51088	0.12752	0.19857	-0.12117	-0.137610	0.250416	-0.46753	0.01853
Depressão:não_depr	8	0.21526	-0.08145	-0.12958	-0.20056	-0.080657	-0.162731	0.01612	-0.14189
Depressão:sim_depr	9	-1.31766	0.49855	0.79320	1.22767	0.493718	0.996108	-0.09868	0.86854
Autonomia:auton_total	10	0.57691	0.43695	-0.31642	-0.27099	-0.020702	0.282249	0.41382	0.30794
Autonomia:0-3AVD	11	0.00882	-0.63704	0.93215	0.54897	-0.326803	-0.765960	-0.19856	0.02786
Autonomia:4-6AVD	12	-0.68422	-1.60534	-1.93043	0.32522	1.112453	0.352568	-1.20082	-0.01972
Autonomia:7+AVD	13	-1.92290	1.12215	0.46947	-0.55436	0.041101	0.524743	-0.26198	-1.36878

Figura 3 – Coordenadas das Colunas das Variáveis.

Assim, a técnica de Análise de Agrupamentos forneceu o dendograma apresentado na Figura 4, com a devida disposição das variáveis, em relação as distâncias apresentadas anteriormente. Neste caso, foi definido o nível de similaridade pela métrica Distância Euclidiana, e o Método de Ward como o método hierárquico.

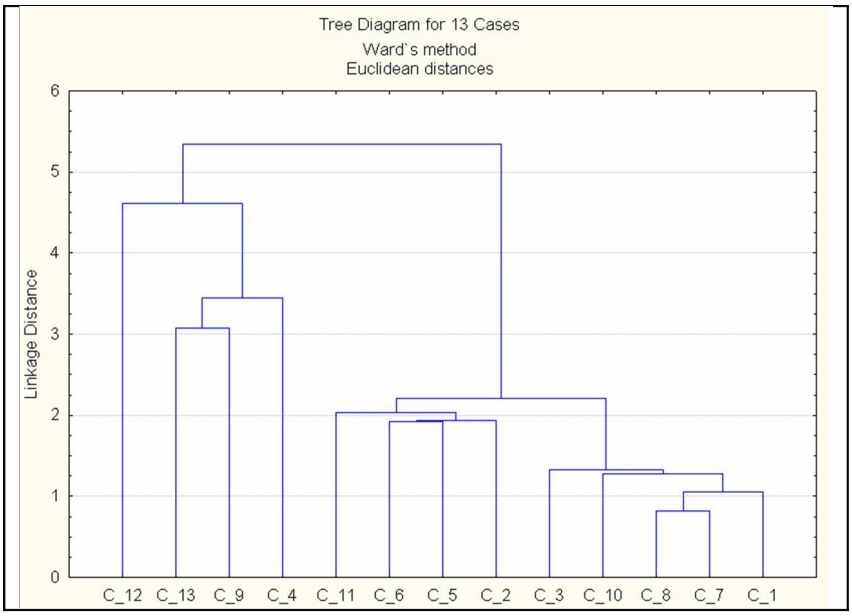

Figura 4 – Dendograma

Note que para cada nível de similaridade, é obtido um determinado número de agrupamento. Portanto, se faz necessário a escolha de um critério para determinar o número ideal de agrupamentos. Assim, dado este pela 'Análise de Comportamento do Nível de Fusão', foi verificado a classificação em cinco grupos, como dispostos na Figura 5. Observe que cada agrupamento foi designado por uma determinada cor e, para uma melhor leitura, o gráfico foi transposto de posição, mas mantendo a mesma configuração. A seta rosa indica o nível de distância indicado pelo critério escolhido.

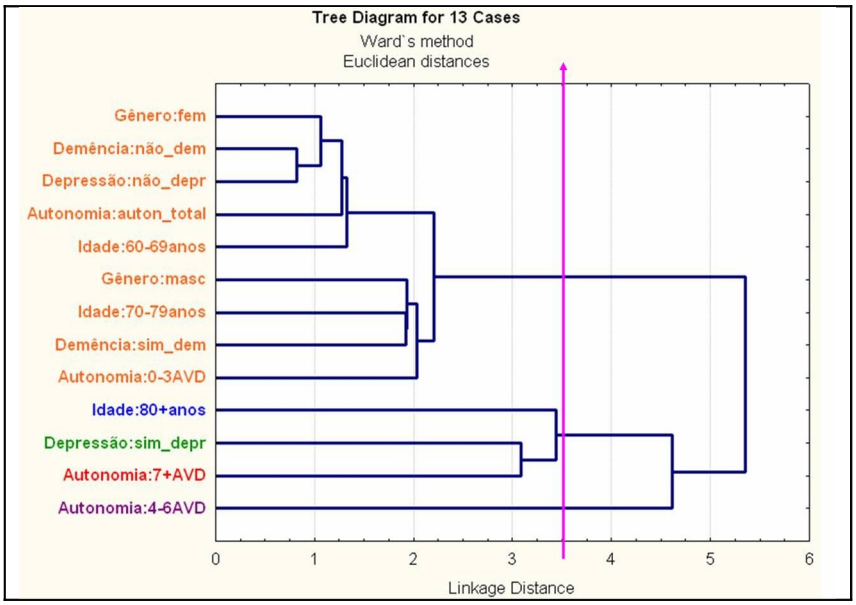

Figura 5 – Agrupamento das variáveis.

Portanto, ao se comparar a Figura 5 com a Figura 1, podem-se verificar os mesmos agrupamentos fornecidos pelas duas técnicas. A Figura 6 representa o agrupamento indicado pela Análise de Agrupamentos dentro do que fora indicado pela Análise de Correspondência. Observar como as demarcações são coincidentes.

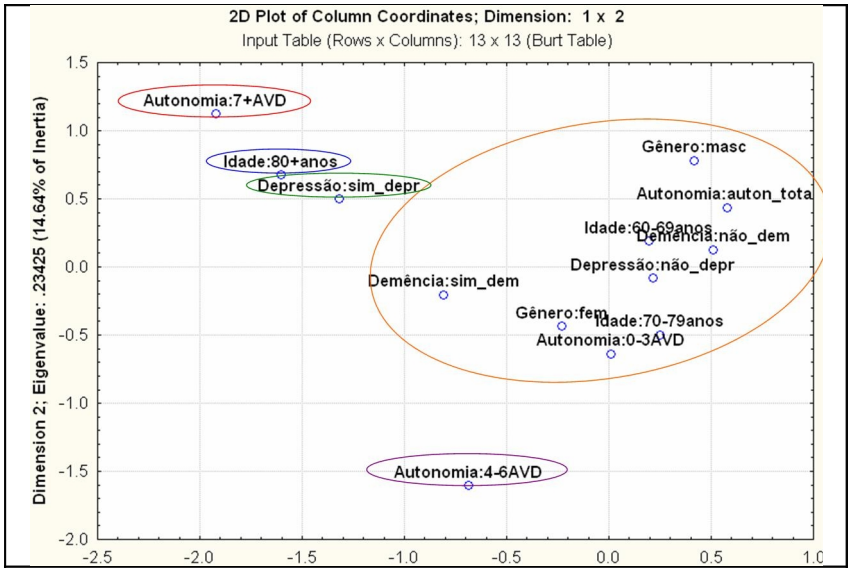

Figura 6 - Sobreposição gráfica da Análise de Correspondência com a Análise de Agrupamentos.

4. Conclusão

A técnica de Análise de Correspondência se revela uma técnica útil para disponibilizar a associação de variáveis categóricas através do seu dispositivo gráfico, gerado pelas relações existentes entre as mesmas. No entanto, a combinação de uma segunda técnica, a saber, da Análise de Agrupamentos, otimizou de forma considerável a interpretação de seus resultados.

Devido ao fato da ilustração exibida incluir pouca quantidade de categorias de varáveis, possibilita induzir o leitor a pensar que os grupos formados pela Análise de Agrupamentos já eram visíveis na disposição gráfica fornecida pela AC. Logo, esta escolha da quantidade foi primordial para que se pudessem visualizar as demarcações coincidentes das duas análises. E, por outro lado, uma ilustração com um grande número de variáveis, em que não fosse possível a mesma comparação, poderia gerar maiores suspeitas em relação ao resultado obtido, já que neste caso, não seria possível uma visualização gráfica bidimensional definida da AC.

Portanto, é interessante despertar o interesse por pesquisas que expõem a combinação de métodos, uma vez que podem resultar num aperfeiçoamento dos resultados de uma análise. Ainda que cada técnica possua suas particularidades e objetivos específicos de pesquisa, o ajuste de duas ou mais técnicas podem proceder a invenção de uma nova técnica.

5. Agradecimentos

Á CAPES pelo apoio financeiro.

6. Referências

Alves, L. B., 2007, Tratamento de Dados Multivariados por Análises de Correspondência e de Agrupamentos em dados de idosos de São José dos Campos. Dissertação de Mestrado - Instituto Tecnológico de Aeronáutica– Curso de Engenharia Aeronáutica e Mecânica, pp. 110.

Carvalho, M. S.; Strunchiner, C. J., 1992, Análise de Correspondência: uma aplicação do método à avaliação de serviços de vacinação. Caderno de Saúde Pública, Rio de Janeiro, Vol. 8 (3), pp. 287-301.

Faria, R. T., 1993, Tratamento de dados multivariados através de análise de correspondência em rochas carbonáticas.. Dissertação de Mestrado Campinas: Universidade Estadual de Campinas – UNICAMP. Instituto de Geociências. Área de Geologia de Petróleo, pp. 138.

Greenacre, M. J., 1987, Theory and applications of correspondence analysis. Journal of the American Statistical Association, Vol. 82, nº 398, pp. 437-477.

Greenacre, M.; Hastie, T., 1984, The Geometric Interpretation of Correspondence Analysis. Ed. Academic Press.

Hair, J. F. Jr., et al., 2005, Análise Multivariada de Dados. Ed Bookman, Porto Alegre, pp. 593.

Lourenço, E. B., 1997, Avaliação: contribuição da análise de correspondência para a avaliação docente, SP.

Mingoti, S. A., 2005, Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada. Belo Horizonte: Editora UFMG, pp. 297