



APRESENTAÇÃO DO GUIA DE AULA

Esse guia foi desenvolvido para o mini-curso de Análise Multivariada de Dados com apoio de Softwares Estatísticos a ser apresentado durante a 63ª edição do SBPC em Goiania-GO. O suporte a esse material provém de dois softwares em versão Demo e Trial: Sphinx® e Systat®, disponíveis para download através dos sites de seus distribuidores.

O objetivo desse material é avaliar suporte para o participante que contará com outro recurso didático que são os arquivos-exemplos disponíveis no cd do curso. Com relação à bibliografia, foram empregadas duas obras fundamentais para o pesquisador e disponíveis em Português: Hair *et al* (2009) e Fávero *et al* (2010), sendo a primeira obra clássica internacional sobre Análise Multivariada de Dados e a segunda uma excelente publicação brasileira com apelo mais prático, mas igualmente rica na essência da teoria estatística.

Esse material foi desenvolvido pelo Prof. Solon Bevilacqua, do departamento de Administração da UFG campus Catalão-GO. O e-mail para contato é solbev@gmail.com e telefone 64-3441.5623.

O QUE É ANÁLISE MULTIVARIADA?

Definir Análise Multivariada não é tarefa simples e, de acordo com Hair *et al* (2009) todos os métodos estatísticos que simultaneamente analisam múltiplas medidas sobre cada indivíduo ou objeto sob investigação pode ser considerada Análise Multivariada. “Qualquer análise simultânea de mais de duas variáveis de certo modo pode ser considerada Análise Multivariada” (p. 26).

Escalas de Medida

As variáveis utilizadas em estudos que envolverão Análise Multivariada objetivam tratar fenômenos de dispersão e posição para os dados. O primeiro fenômeno possibilita combinar análises com desvio padrão, média, bem como correlação. Já o segundo possibilita análises como histogramas de classes. Dessa forma, já no projeto do estudo, se torna evidente prever o tipo de questão a ser construída em função da análise futura.

De acordo com Hair *et al* (2009, p. 27) “analisar dados envolve a partição, a identificação e a medida de variação em um conjunto de variáveis, seja entre elas ou entre uma variável dependente e uma ou mais variáveis independentes”. Basicamente há dois tipos de dados, os considerados não métricos (qualitativos) e os métricos (quantitativos). Os primeiros são aqueles classificados como atributos, características ou propriedades categóricas que identificam ou descrevem um objeto, são aquelas classificações como um estado. As variáveis métricas prevêm quantidade ou magnitude.

As variáveis mais comuns não métricas são aquelas tidas como Ordinais, Categóricas e Intervalares com zero arbitrário. As métricas são variáveis com zero absoluto ou as denominadas *Razão*, também chamadas de *Abertas Numéricas*.

Escalas como ‘Likert’, conforme exemplos, a seguir são considerados não métricas:

1 []	2 []	3 []	4 []	5 []	6 []
-------	-------	-------	-------	-------	-------

[] Muito Insatisfeito	[] Insatisfeito	[] Satisfeito	[] Muito Satisfeito
------------------------	------------------	----------------	----------------------



REGRESSÃO MÚLTIPLA

A Regressão Múltipla representa uma das técnicas multivariadas mais comuns entre os pesquisadores e sua difusão se deve principalmente à amplitude de seu espectro e a facilidade de adoção. Seu emprego está condicionado ao envolvimento de uma variável dependente ou de resposta e 'n' variáveis independentes ou preditoras.

A problemática da técnica envolve prever o valor da variável dependente através de valores desconhecidos das variáveis independentes, como por exemplo, o lucro relacionado a várias operações de pesquisa e desenvolvimento; propaganda e publicidade.

A relação entre as variáveis é medida pelo coeficiente de correlação (r) que varia de -1 a 1, indicando correlação positiva ou negativa. Quanto mais próxima de zero, menos significativa será. Valores negativos indicam que se uma variável se torna maior, a outra fica menor, sendo, portanto inversamente proporcional.

Como se mensura a confiança no modelo?

Comumente se utiliza a soma dos quadrados dos erros (SSE), mas sua divisão pela Soma total dos Quadrados é tida como mais usual e essa será aquela abordada nesse guia. A origem dessa divisão apresenta o Coeficiente de Correlação (R^2). Quando a equação da Regressão contém mais de uma variável preditora o valor do R^2 contempla todos os efeitos combinados de toda a variável estatística na previsão. Se tivermos, por exemplo, um valor de 0,85, temos que 85% da variação na variável dependente é explicada pelas variáveis independentes. No que se refere ao tamanho da amostra e conforme Hair *et al* (2009) em amostras pequenas, caracterizadas por menos de 20 observações, são apropriadas análise apenas por Regressão Simples.

É importante ressaltar que dificilmente teremos R^2 centrado exatamente em 1 em função de que, de acordo do Fávero *et al* (2010) é praticamente impossível termos todos os pontos em cima da reta de tendência. Se supostamente tivéssemos os pontos com correlação perfeita, não haveria resíduos, portanto um modelo com ausência de *outliers*.

O autor reforça que sempre haverá ressalvas na interpretação do R^2 visto que essa medida captura apenas a importância das variáveis contidas no modelo, desconsiderando variáveis periféricas. Assim ocorre com o exemplo citado pelos autores, que envolve a análise de um R^2 em 0,30 para uma cesta de ações. A explicação parece pequena, mas estamos desconsiderando a existência de inúmeros outros fatores periféricos.

ANÁLISE FATORIAL [AF]

A AF é uma técnica de redução de dados que objetiva promover análises que reportem a estrutura das inter-relações entre um grande número de variáveis, definindo um conjunto de dimensões latentes comuns (fatores), conforme afirma Hair *et al* (2009). Essa técnica difere das demais técnicas de dependência em função de se considerar as variáveis originais do modelo como dependentes, todas serão inter-correlacionadas, sem a pré existência de uma variável de resposta. Adicionalmente em função de seu caráter de redução de dados a técnica também pode ser considerada como exploratória.

Com relação ao procedimento de execução da AF é recomendado que não se operacionalize a técnica sem que haja 50 casos, sendo recomendado pelo autor mais de 100 observações ou ainda outra proporção de 20 casos para cada variável.

Como se mensura a confiança no modelo?

O relatório de saída de softwares como o Systat® já apresentam no corpo das análises o valor encontrado para o RSQ, Correlação Quadrática. O coeficiente alternativo de correlação ao quadrado



(RSQ), informa a proporção de variabilidade dos dados do modelo, sempre oscilando entre zero e 1, onde o primeiro valor representa que o modelo é ruim e o segundo o contrário.

Em Hair *et al* (2009, p. 98) temos mais duas opções para a confiabilidade das análises com a AF que poderão também ser mais bem estudadas: o teste Bartlett de Esfericidade e a MAS (medida de adequação da amostra). Em Fávero *et al* (2010, p. 254) os dois testes também são descritos de forma prática.

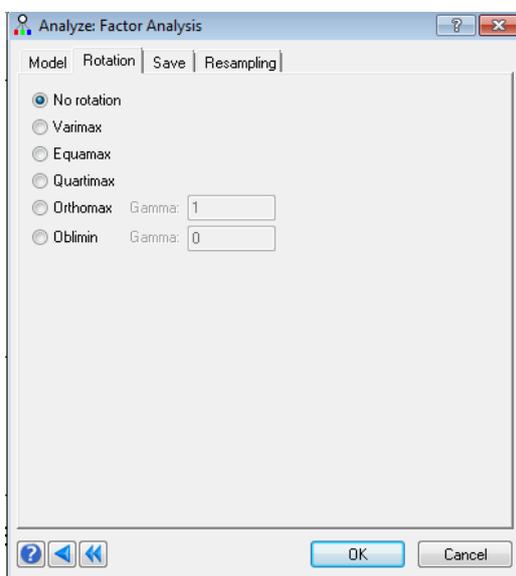
Quantos fatores devo extrair?

Basicamente temos três procedimentos a adotar, sendo o mais tradicional o critério a priori, baseado em estudos preliminares e secundários. O pesquisador dominando quantos fatores existem, já informa ao software de análise quantos fatores ele deseja calcular.

O segundo critério é o da raiz latente, igualmente muito utilizada e, nesse caso vamos considerar apenas os autovalores ou raízes latentes maiores do que 1;

O teste *scree* normalmente é apresentado diretamente por softwares como o Systat® e sua análise é intuitiva no sentido de que apenas observando a tendência dos dados é possível determinar o número correto de fatores a extrair. Quando o gráfico inicia sua ‘horizontalização’ é o momento o corte para os fatores.

Verificada a consistência do modelo e o número de fatores a extrair roda-se a operação no software específico obtendo a carga fatorial que são as correlações de cada variável sob exame. Conforme Hair *et al* (2009, p. 103) “essas cargas indicam o grau de correspondência entre a variável e o fator, sendo que as maiores cargas nomeiam os fatores. Comumente o pesquisador não adota algum tipo de rotação, mas conforme o autor, a rotação é indicada pois simplifica a estrutura fatorial, e normalmente é difícil determinar se fatores não rotacionados serão significativos. A seguir é apresentada a tela do Systat® para as rotações possíveis e costumeiramente empregadas.



O objetivo da rotação é redistribuir a variância dos primeiros fatores para os últimos para atingir um padrão fatorial mais simples. O pesquisador pode ir retirando variáveis com baixa carga fatorial a fim de ajustar o modelo e aumentar sua eficácia. Maiores detalhes podem ser obtidos em Hair *et al* (2009, p. 103).

Figura 1: Análise Fatorial
Fonte: Systat®



O modelo de AF para melhor significância estatística deve obedecer ao critério do tamanho da amostra. Dessa forma um projeto com amostra **maior** aceita cargas fatoriais **menores** nas análises. Hair *et al* (2009, p. 107) propõe um modelo para identificação de cargas fatoriais significantes com base no tamanho da amostra.

Tabela 2: Uso gerencial da Análise Conjunta

Carga Fatorial	Tamanho necessário da amostra para significância
0,30	350
0,35	250
0,40	200
0,45	150
0,50	120
0,55	100
0,60	85
0,65	70
0,70	60
0,75	50

Fonte: Hair *et al et al* (2009, p. 107)

A Análise Fatorial é considerada excelente como técnica de redução e exploração de dados, entretanto suas múltiplas interpretações quando dos procedimentos acarreta inúmeras divergências. As duas principais limitações se referem à forma que os fatores são extraídos e no tocante à rotação dos eixos fatoriais. Ambos os procedimentos deixam margem para análises variáveis o que impacta no resultado final da AF.

Integração com outras técnicas

Sendo o objetivo da técnica de AF reduzir a quantidade de dados transformando o modelo em um conjunto menor de variáveis (fatores), não correlacionados. Dessa forma podemos avançar para empregar outras técnicas de forma sequencial. Segundo Fávero *et al* (2010) os escores fatoriais podem ser utilizados, caso não haja perda significativa de variância explicada pelos fatores gerados” (p. 260). As técnicas a serem utilizadas em integração poderiam ser a Regressão Múltipla, Análise de Discriminante ou ainda Regressão Logística.

ANÁLISE CONJUNTA DE ATRIBUTOS

A Análise Conjunta de Atributos é uma técnica sofisticada e emergente de análise e tem registrado expressiva aceitação e utilização. Segundo Aaker, Kumar e Day (2004) já na década de 1980 mais de 400 análises conjuntas para fins comerciais eram realizadas anualmente. Conforme os autores as aplicações para a técnica que recebem destaque seriam:

- Em análises em que os produtos ou serviços possuem um número de atributos, cada qual com dois ou mais níveis;
- Onde não existam ainda combinações possíveis entre os níveis de atributos;
- Onde possa existir expansão dos níveis dos atributos para além daqueles disponíveis na ocasião;
- Onde a tendência geral seja provável e já conhecida sobre os atributos

A Análise Conjunta é adequada para algumas formas de segmentação. Sua aplicação principal recai diretamente sobre a medição de preferências do comprador para os níveis de atributo de produto e, conforme Green e Krieger (1991) nos benefícios que o comprador possa obter a partir dos atributos desse consumo.



Tabela 2: Uso gerencial da Análise Conjunta

Segmentação de Mercado	Comportamento do Consumidor
Isolar grupos de clientes	Formação de Conceito de Produto
Identificar oportunidades	Definição de contribuição de cada Atributo
Identificar combinações de objetos	Definição de Preferências

Fonte: Adaptado de Hair *et al et al* (2009, p. 327)

Nesse sentido a CA apresenta ao entrevistado um conjunto de atributos independentes, geralmente sob forma hipotética ou uma alternativa real e, em seguida, exige avaliação ou preferência da alternativa. Dessa forma, a CA é superior a outros métodos, pois permite uma visão mais realista para os *trade-offs*.

Como a CA revela dados de utilidade a partir de amostra selecionada, segundo Motta (1987, p. 22) é possível estudar as discrepâncias de utilidade ao longo dessa amostra, “procurando identificar grupos distintos de consumidores, sendo útil para análise de segmentação”. Poder-se-ia submeter as utilidades obtidas em cada nível a um programa de aglutinação dos entrevistados, a partir das respostas fornecidas individualmente.

Adicionalmente, Bittencourt e Fachel (1999), afirmam que através de combinações que representam conjuntos de escolhas reais, a técnica proporciona a identificação de configurações ótimas de produtos ou serviços, que esclarecem potencial de novas oportunidades de mercado ou potencial de consumo. Portanto o objetivo da técnica é determinar as contribuições das variáveis preditoras e de seus respectivos níveis, para construir um modelo consistente de preferências de consumo.

A Análise Conjunta pode também ser útil quando houver interesse em avaliar as percepções dos consumidores em relação a diversos estímulos pertencentes a um mesmo propósito. Conforme Motta (1987) as utilidades presentes em cada estímulo confere ganhos frente ao método tradicional decompositivo, afinal o consumidor se posiciona frente a uma cesta de opções e não para utilidades isoladas. Reagindo a estímulos apresentados pela combinação de níveis de atributos, o consumidor dá margem para que se avance até sua estrutura mental de compensação, provendo informações valiosas sobre seu sistema de valores e de avaliação de preferências.

Com relação a Análise de Conglomerado, Hair *et al* (2009) aborda a técnica sugerindo seu emprego para a definição da relação de interdependência entre o conjunto de variáveis. Segundo o autor, as aplicações mais clássicas envolvem a segmentação de mercado, compreensão do comportamento do consumidor e identificação das oportunidades de um novo produto. Conforme Fávero *et al* (2009) a variável estatística de agrupamento se define como o conjunto de atributos das observações que servirão de base para a determinação da similaridade entre elas. Essa base gerada a partir de atributos torna-se mais robusta se proveniente de uma técnica que privilegie o emprego desses dados de entrada, como é o caso da Análise Conjunta.

Considerando a metodologia original de Green e Krieger (1991) e as limitações encontradas pelos autores, tem-se que a ferramenta é tida como um recurso de planejamento e de análise de sensibilidade para explorar alternativas de produtos e estratégias de preços. A técnica não fora testada para atributos intangíveis como conforto, e muita pesquisa adicional será necessária antes que o seu potencial seja comprovado.

Operacionalização da AC

O início de um trabalho envolvendo a AC prevê a escolha de atributos (tangíveis ou intangíveis) e a consequência definição de seus níveis. Softwares com o Systat® reportam em seu relatório de saída um modelo para as utilidades parciais, que indicam a utilidade de cada atributo e seus níveis. No exemplo a seguir consideramos que o problema de pesquisa seja identificar como é formada a cesta de consumo de determinado produto considerando dois atributos: Conforto e Preço.



Escolhemos o atributo *Conforto* em seus níveis 'alto', 'médio' e 'baixo' e *Preço*, '23.000,00' e '24.000,00'. Temos uma relação de níveis em 3*2 que irá possibilitar a criação de 6 *trade off's*. A terceira coluna será destinada para a nota a ser dada para cada possibilidade de escolha.

Quantidade excessiva de perfis para resposta origina complicação para o respondente, pois o mesmo terá que se posicionar para um número grande de possibilidades de escolha. Uma forma intuitiva de identificar a confiabilidade do modelo é a observação do diagrama de Shepard para o ajuste dos dados, conforme na figura a seguir.

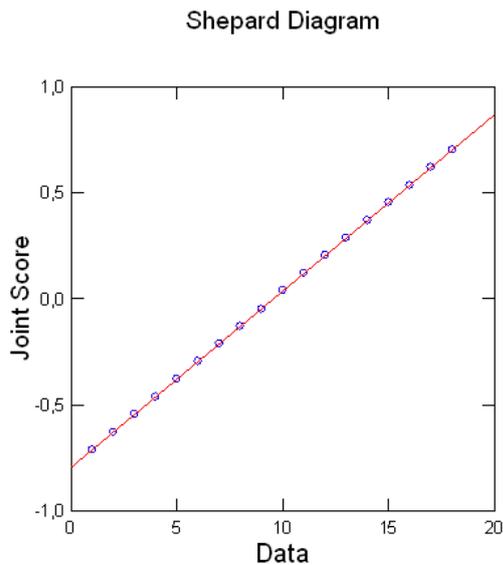


Figura 2: Diagrama de Shepard
 Fonte: Systat®

Temos que a soma (em módulo) das utilidades parciais para *Conforto* gera o valor de 1,43 e para *Preço* registra 0,70. A utilidade total do modelo é 2,13, reportando que a utilidade total para *Conforto* é 66,94% da utilidade total e 32,06% para *Preço*. Fundamentalmente os consumidores adquirem esse produto em função do Conforto que proporciona.

Parameter Estimates (Part Worth's)

CONFORTO(1)	CONFORTO(2)	CONFORTO(3)	PRECO(1)	PRECO(2)
0,688	-0,371	-0,371	0,380	-0,326

Figura 3: Utilidades Parciais
 Fonte: Systat®

ESCALONAMENTO MULTIDIMENSIONAL

A técnica de análise multivariada Escalonamento Multidimensional [EMD] tem como finalidade no dizer de Fávero *et al* (2010) mapear distâncias entre pontos em uma representação gráfica espacial e, portanto, ajuda o pesquisador na identificação de dimensões-chave inerentes a avaliações feitas por respondentes quanto a determinados objetos (p. 307).

A técnica permite tratar dados métricos e não métricos a partir de medidas de dissimilaridades onde valores maiores representam distância maior e o mesmo ocorre para os valores menores, tornando a técnica conforme Fávero (2010), bastante intuitiva.



	Estímulo A	Estímulo B	Estímulo C
Estímulo A	-		
Estímulo B	9	-	
Estímulo C	8	4	-

Quadro 1: Matriz de dissimilaridade

Fonte: Adaptado de Fávero *et al* (2010, p. 311)

O processo para o cálculo do relatório é tido como simples e sua execução exige que o pesquisador insira as respostas para a matriz e determine o número de dimensões a analisar, conforme a seguir.

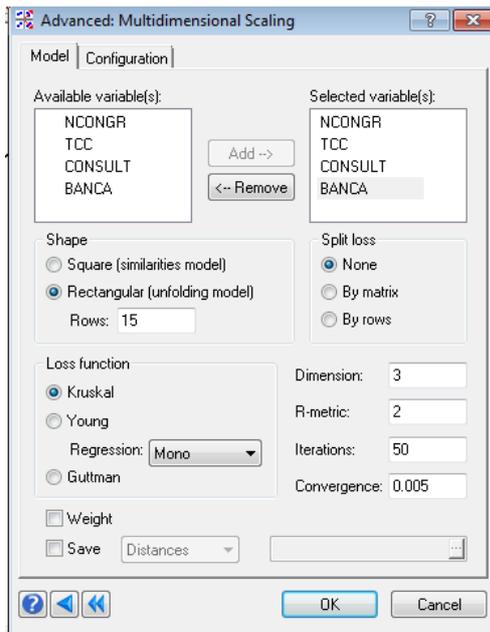


Figura 3: Utilidades Parciais

Fonte: Systat®

A confiabilidade do modelo é apresentada a partir de dois elementos: o nível de Stress e o MSQ. Adicionalmente está a disposição do pesquisador o diagrama de Shepard para visualização da disposição dos dados.

Como se mensura a confiança no modelo?

Stress

Kruskal (1964) propôs uma medida de adequação de ajuste para avaliar o quanto se aproximam as distâncias derivadas dos dados de dissimilaridade daqueles fornecidos pelos respondentes. Quanto maior o nível de Stress, pior o ajuste dos dados, considerando que seu valor mínimo será zero. O autor sugere que o Stress seja tratado da seguinte forma:



Tabela 2: Uso gerencial da Análise Conjunta

Stress	Adequação do Ajuste
20%	Pobre
10%	Razoável
5%	Bom
2,5%	Excelente
0%	Perfeito

Fonte: Adaptado de Hair *et al et al* (2009, p. 327)

RSQ (Correlação Quadrática)

O coeficiente alternativo de correlação ao quadrado (RSQ), informa a proporção de variabilidade dos dados do modelo, sempre oscilando entre zero e 1, onde o primeiro valor representa que o modelo é ruim e o segundo o contrário.

No caso a seguir, do arquivo 'Professores.sav' de Fávero *et al* (2010) é possível construir uma tabela como a seguir, onde é apresentado o Stress para cada dimensão, bem como o RSQ. Ambos indicadores de adequação para o estudo, sendo que o primeiro está próximo da perfeição.

Iteration History

Iteration	STRESS
0	0,166
1	0,126
2	0,102
3	0,089
4	0,081
5	0,075
6	0,071
0	0,087
1	0,074
2	0,068
3	0,064
4	0,061

Stress of Final Configuration : 0,061
Proportion of Variance (RSQ) : 0,932

Figura 4: Stress
Fonte: Systat®

ANÁLISE DE AGRUPAMENTOS

A técnica de Análise de Agrupamentos ou *Cluster Analysis* se refere a um procedimento de interdependência para os dados a partir da idéia de similaridade, conforme Fávero *et al* (2010). Sua aplicação pode ser observada em áreas distintas, desde marketing até ciências biológicas, afinal seu objetivo é segmentar dados em grupos homogêneos e heterogêneos entre si. Para fins didáticos a história dessa técnica bem como visões de outros autores pode ser aprofundada em Fávero *et al* (2010, p. 197).

Com relação a Análise de Conglomerado, Malhotra (2001) aborda a técnica sugerindo seu emprego para a definição da relação de interdependência entre o conjunto de variáveis. Segundo o autor, as aplicações mais clássicas envolvem a segmentação de mercado, compreensão do comportamento do consumidor e identificação das oportunidades de um novo produto. Conforme Fávero *et al* (2010) a variável estatística de agrupamento se define como o conjunto de atributos das observações que servirão de base para a determinação da similaridade entre elas. Essa base gerada a partir de atributos torna-se mais robusta se proveniente de uma técnica que privilegie o emprego desses



dados de entrada, como é o caso da Análise Conjunta. Para Brochado e Martins (2008) os métodos hierárquicos permitem recuperar a heterogeneidade existente nos dados e a obtenção simultânea de estimativas ao nível do indivíduo.

Alguns problemas associados ao emprego isolada da Análise de Conglomerados são conhecidos e detalhados em Aldenderfer e Blashfield (1984) e Everitt (1980). Destaca-se a questão da validade dos grupamentos, se efetivamente são genuínos, conferindo imprecisão para o estudo mercadológico. Utilizar a Análise Fatorial previamente não resolveria essa fragilidade considerando que, segundo Aldenderfer e Blashfield (1984), poderia ocorrer uma perda de dados relevantes acarretando falha de informação para a Análise de Conglomerados. Considerando a metodologia original de Green e Krieger (1991) e as limitações encontradas pelos autores, tem-se que a ferramenta é tida como um recurso de planejamento e de análise de sensibilidade para explorar alternativas de produtos e estratégias de preços. A técnica não fora testada para atributos intangíveis, e muita pesquisa adicional será necessária antes que o seu potencial seja comprovado.

Malhotra (2001) aborda essa técnica, sugerindo seu emprego para a definição da relação de interdependência entre o conjunto de variáveis. Segundo o autor, as aplicações mais clássicas envolvem a segmentação de mercado, compreensão do comportamento do consumidor e Identificação das oportunidades de um novo produto, sendo o último, objetivo de aplicação dessa técnica nesse estudo.

Com relação a questão de confiabilidade, duas providencias podem ser tomadas, de fácil emprego, a primeira seria separar o grupo de dados em duas metades, e comparar os centróides dos clusters nas duas subamostras e a segunda seria ir realizando repetições múltiplas até a estabilização da solução (Aeker, 2001).

Operacionalização da Análise de Agrupamentos

A Análise de Agrupamentos obedece a uma lógica simples de operação, obedecendo ao seguinte fluxo de procedimento, conforme Hair *et al* (2009, p. 389):

Estágio 1. Definição do objetivo da Análise de Agrupamentos:

Explorar ou formar classificação de objetos baseados na experiência;
Simplificação de dados;
Identificação de relação.

Nesse estágio se seleciona as variáveis de agrupamento.

Estágio 2. Definição do projeto de pesquisa;

Seleção de *outliers*;

Definição da medida de similaridade: correlacional ou de distância. A primeira raramente é utilizada, afinal conforme Hair *et al*, (2009, p. 393) “porque a ênfase na maioria das aplicações da análise de agrupamentos é sobre as magnitudes dos objetos, e não sobre os padrões de valores.”

Definição da medida de distância: Conforme o autor “essa medida se concentra na magnitude dos valores e representam casos similares que estão próximos, mas podem ter padrões muito diferentes dos encontrados quando se usam medidas correlacionais” (p.393). Distâncias menores representam maior similaridade. O Software Systat® disponibiliza 15 medidas de distância conforme a figura 5 e nesse caso nos deteremos apenas na euclidiana por ser a mais difundida.



Figura 4: Stress
Fonte: Systat®

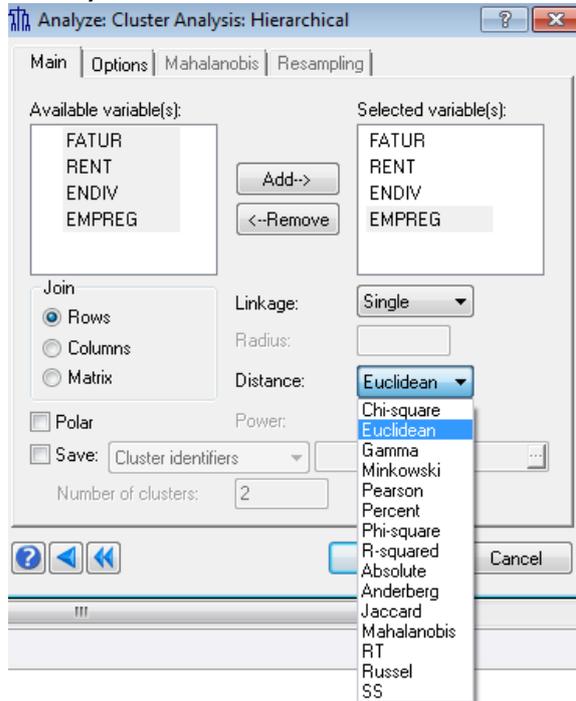


Figura 5: Medidas de Distância
Fonte: Systat®

Estágio 3: Suposições em análise de agrupamentos:

Nesse estágio a preocupação principal recai na representatividade da amostra e na multicolinearidade. A primeira preocupação é tida como normal em qualquer pesquisa quantitativa que haja interesse em fazer inferências para a população. A segunda pode ser facilmente resolvida com o empregada da distância de Mahalanobis (D^2) que segundo Hair *et al* (2009, p395) executa um processo de padronização dos dados e “estabelece uma escala em termos de desvio padrão, bem como soma as variância-covariância acumulada dentro dos grupos”.

Estágio 4: Determinação de agrupamentos e avaliação do ajuste geral

Nesse estágio o pesquisador irá decidir entre a escolha do método hierárquico e o não hierárquico. Analisando agrupamentos pelo método Hierárquico. O detalhamento dos dois métodos são aprofundados em Hair *et al* (2009, p. 398-404) e em Fávero *et al* (2010, p. 204-224). Ambos os autores questionam sobre qual método deve ser escolhido e nas duas publicações encontramos a resposta centrada no problema de pesquisa.

Qual o melhor método?

O método hierárquico não requer que o pesquisador conheça a priori o número de cluster para iniciar a partição. Em contrapartida os objetos após o agrupamento permanecerão no mesmo cluster. No método não hierárquico existe a necessidade de se conhecer o número sugerido de agrupamentos e necessita de uma semente inicial e o número indicativo de cluster a serem formados. O método *K-means* o mais popular dos métodos não hierárquicos é bastante inadequado quando a semente inicial aleatória é utilizada sem a aplicação prévia de um método hierárquico. Nesse guia os procedimentos serão feitos com o método hierárquico.



Os autores reforçam que o método de escolha para o número de agrupamentos depende de uma decisão *ad hoc*, “em conceituação empírica com qualquer conceituação de relações teóricas que possam sugerir um número natural de agrupamentos” (Hair *et al*, p. 404).

Estágio 5: Interpretação dos agrupamentos

Nessa fase o pesquisador irá nomear os grupos a partir das características dos mesmos, tais como questões demográficas e psicográficas. Trata-se de um trabalho de rotulamento com interpretações relevantes a partir das particularidades exclusivas.

Estágio 6. Validação e perfil dos agrupamentos

Em Hair *et al* (2009) tem-se que para validar a operação e os agrupamentos é interessante buscar recursos a outra técnica de análise multivariada, a Análise de Discriminante, onde se compara perfis de escores médios dos agrupamentos. A variável dependente categórica é a identificação prévia dos grupos e as variáveis independentes são as características demográficas, psicográficas, entre outras.

Os seis estágios em um exemplo prático:

Conforme a figura 5, optamos pela Análise de Cluster do tipo hierárquico e Distância Euclidiana. O arquivo a ser adotado é o “Siderur_clusterteste.sav”. Escolhemos as 4 variáveis disponíveis e rodamos a operação no Systat®.

Clusters Joining		at Distance	No. of Members
Case 9	Case 8	1.184,242	2
Case 18	Case 17	2.048,779	2
Case 22	Case 21	2.569,770	2
Case 25	Case 24	3.913,055	2
Case 18	Case 16	5.301,662	3
Case 25	Case 23	6.200,537	3
Case 18	Case 15	6.783,541	4
Case 18	Case 19	7.214,163	5
Case 22	Case 20	7.300,045	3
Case 5	Case 4	8.082,063	2
Case 18	Case 22	10.871,339	8
Case 14	Case 13	12.698,789	2
Case 14	Case 18	13.507,416	10
Case 9	Case 10	14.557,486	3
Case 7	Case 6	17.612,418	2

Figura 7: Análise Individual dos Agrupamentos

Fonte: Systat®

Em análise visual do gráfico de dendograma já é possível perceber que há formação de 3 grupos distintos. A melhor forma para comprovar visualmente a presença desses grupos é traçar uma linha vertical que corte os agrupamentos finais, conforme a figura 6. Os dados agrupados seguem conforme a figura 7.

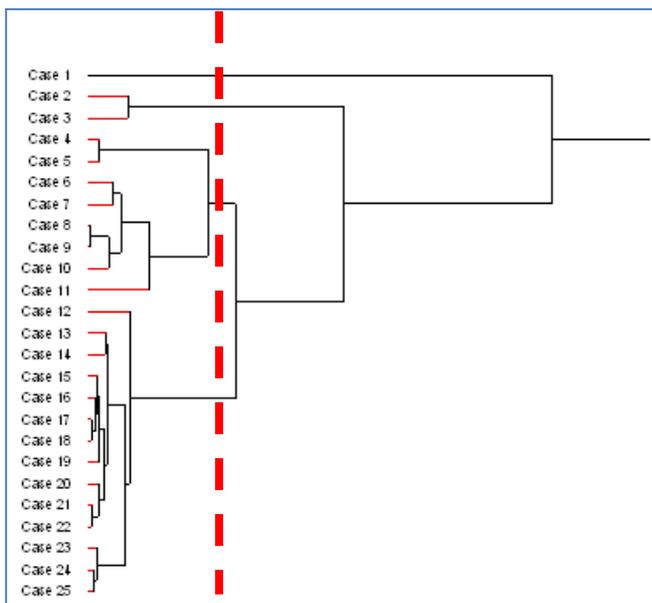


Figura 6: Análise dos Agrupamentos
Fonte: Systat®

REFERÊNCIAS BIBLIOGRÁFICAS

- AAKER, D.A., KUMAR, V.D., & DAY, G.S. *Pesquisa de Marketing*. São Paulo: Atlas, 2001
- ALDENDERFER, M. S.; BLASHFIELD, R. K. *Cluster analysis*. Beverly Hills, CA: Sage, 1984.
- BITTENCOURT, H. R., e J. M. FACHEL. Análise Conjunta de Atributos (Conjoint Analysis), Porto Alegre, n. 1, p. 1-46, 1999. *Documentos de Estudo do PPGA*, 1999: 1-46.
- BROCHADO, A; MARTINS, F. Aspectos Metodológicos da Segmentação de Mercado: base de segmentação e métodos de classificação. *Revista Brasileira de Gestão de Negócios*. São Paulo, V. 10, n 27. 132-149-Abril-Junho, 2008.
- EVERITT, B. *Cluster analysis*. New York, NY: Haltesd, 1980.
- FÁVERO, L. P; BELFIORE, P; DA SILVA, F. L; CHAN, B. L. *Modelagem multivariada para tomada de decisões*. São Paulo: Campus, 2009. p. 544
- GREEN, P. E.; KRIEGER, A. M. Segmenting markets with conjoint analysis. *Journal of Marketing*, v.55, n. 4, p.20-31, 1991
- HAIR, J; ANDERSON, R. E.; TATHAM, R. L.; BLACK W. C. *Análise Multivariada de Dados*. 5. ed. Porto Alegre: Bookman, 2005.
- KRUSKAL, J. B. *Nonmetric Multidimensional Scaling: A Numerical Method*. *Psychometrika* 2: 115-129, 1964
- MALHOTRA, N. *Pesquisa de Marketing*. Porto Alegre: Bookman, 2001.
- MOTTA, P. C. *Análise Conjunta*. Rio de Janeiro: PUC - RJ, Mímeo, 1997.