Naresh K. Malhotra
Georgia Institute of Technology

PESQUISA DE MARKETING
UMA ORIENTAÇÃO APLICADA

4ª EDIÇÃO

M249p Malhotra, Naresh
Pesquisa de marketing : uma orientação aplicada / Naresh Malhotra ; tradução Laura Bocco. – 4. ed. – Porto Alegre : Bookman, 2006.
720 p. ; 28 cm.

ISBN 85-363-0650-5

1. Marketing – Pesquisa. I. Título.

CDU 658.8.012.12

Catalogação na publicação: Júlia Angst Coelho – CRB Provisório 05/05

Tradução:
Laura Bocco

Consultoria, supervisão e revisão técnica desta edição:
Tânia Modesto Veludo-de-Oliveira
Mestre em Administração pela FEA/USP

Bookman
2006

CAPÍTULO 20

Análise de *Clusters*

"Todos nós acreditamos que qualquer população é composta de segmentos distintos. Se trabalharmos com as variáveis adequadas, a análise de *clusters* nos ajudará a ver se existem grupos que são mais semelhantes entre si do que a membros de outros grupos".

Tom Myers, consultor sênior, Burke Customer Satisfaction Associates, Burke, Inc..

OBJETIVOS

Após a leitura deste capítulo, o aluno deverá estar apto a:

1. Descrever o conceito básico e o objetivo da análise de *clusters* e sua importância na pesquisa de marketing.
2. Discutir as estatísticas associadas à análise de *clusters*.
3. Explicar o procedimento de análise de *clusters*, inclusive a formulação do problema, a seleção de uma medida de distância, a escolha de um procedimento de aglomeração, a decisão sobre o número de *clusters*, a interpretação dos *clusters* e o traçado do seu perfil.
4. Descrever a finalidade e os métodos de avaliação da qualidade dos resultados dos *clusters* e de sua confiabilidade e validade.
5. Descrever as aplicações dos *clusters* não-hierárquicos e dos *clusters* de variáveis.

Aspectos gerais

Tal como a análise fatorial (Capítulo 19), a análise de *clusters* (ou de conglomerados) estuda todo um conjunto de relações interdependentes. A análise de *clusters* não faz distinção entre variáveis dependentes e independentes. Ao contrário, examina relações de interdependência entre todo o conjunto de variáveis. O objetivo principal da análise de *clusters* é classificar objetos em grupos relativamente homogêneos com base no conjunto de variáveis considerado. Os objetos em um grupo são relativamente semelhantes em termos dessas variáveis e diferentes de objetos de outros grupos. Quando utilizada dessa maneira, a análise de *clusters* é o inverso da análise de fatores pelo fato de reduzir o número de objetos, e não o número de variáveis, concentrando-os em um número muito menor de *clusters*.

Este capítulo descreve o conceito básico de análise de *clusters*. As etapas de uma análise de *clusters* são abordadas e ilustradas no contexto da aglomeração hierárquica utilizando um programa de computador. Apresenta-se a seguir uma aplicação da aglomeração não-hierárquica, seguida por um estudo da aglomeração de variáveis.

PESQUISA ATIVA

PROJETO DA LOJA DE DEPARTAMENTOS

***Análise de* cluster**

No projeto da loja de departamentos, os entrevistados foram agrupados com base na importância atribuída por eles mesmos a cada fator relativo aos critérios de escolha utilizados na seleção de uma loja de departamentos. Os resultados indicaram que os entrevistados poderiam ser agrupados em quatro segmentos. Testaram-se estatisticamente as diferenças entre os segmentos. Assim, cada segmento contém entrevistados relativamente homogêneos em relação aos seus critérios de escolha. Estimou-se então o modelo de escolha da loja separadamente para cada segmento. Esse procedimento teve como resultado modelos de escolhas de loja que melhor representavam o procedimento subjacente de escolha pelos entrevistados nos segmentos específicos.

PESQUISA REAL

Sorveterias para regiões "quentes"

A Häagen-Dazs Shoppe Co. (*www.haagen-dazs.com*), com mais de 300 sorveterias espalhadas pelos Estados Unidos, tinha interesse em expandir sua base de clientes. O objetivo era identificar segmentos de consumidores potenciais que viessem ampliar o volume das vendas. Para esse fim, utilizou a geodemografia – um método de agrupar consumidores com base em suas características geográficas, demográficas e de estilos de vida. Fez-se uma pesquisa inicial para estabelecer perfis demográficos e psicográficos dos usuários da Häagen-Dasz Shoppe, incluindo ainda freqüência de compras, hora em que os clientes apareciam, dia da semana e outras variáveis relacionadas com o uso do produto. Anotaram-se também os endereços e os códigos de endereçamento postal dos entrevistados. A seguir, os respondentes foram distribuídos por 40 *clusters* geodemográficos com base em um procedimento de aglomeração elaborado pela Claritas. Para cada *cluster* geodemográfico, comparou-se o perfil dos clientes de Häagen-Dazs com o perfil do *cluster* a fim de determinar o grau de penetração. De posse dessa informação, a Häagen-Dazs pôde também identificar vários grupos potenciais

A Häagen-Dazs aumentou sua penetração identificando *clusters* geodemográficos que ofereciam potencial para vendas crescentes de sorvete.

de consumidores. Além de expandir a base de clientes da empresa, a propaganda do produto foi organizada de forma a atingir novos clientes com base nesse princípio. Em agosto de 2001, a Häagen-Dazs lançou uma campanha satírica que mostrava um grupo de auto-ajuda conversando sobre o prazer proporcionado pelo programa de alimentação Häagen-Dazs. O *slogan* dessa campanha foi "*Pleasure is the path to joy*" ("O prazer é o caminho para a felicidade"). Essa tirada de humor sobre os grupos de auto-ajuda do fim da década de 1990 centrou-se em alcançar o grupo de pessoas por volta dos vinte anos de idade da década de 2000.[1] ■

O exemplo da Häagen-Dazs ilustra o uso da aglomeração para chegar a segmentos homogêneos, com o propósito de formular estratégias específicas de mercado. No exemplo da loja de departamentos, utilizou-se a aglomeração para agrupar entrevistados para análise multivariada subseqüente.

CONCEITO BÁSICO

A análise de *clusters* é uma técnica usada para classificar objetos ou casos em grupos relativamente homogêneos chamados de *clusters* (ou conglomerados). Os objetos em cada *cluster* tendem a ser semelhantes entre si, mas diferentes de objetos em outros *clusters*. Essa análise é também chamada de *análise de classificação**, ou *taxonomia numérica*.[2] Dedicaremos nossa atenção a procedimentos de aglomeração que destinam cada objeto a um único *cluster*.[3] A Figura 20.1 mostra uma situação ideal de *cluster*, em que os *clusters* se apresentam distintamente separados segundo duas variáveis: consciência de qualidade (variável 1) e atitude em relação aos preços (variável 2). Observe que cada consumidor se enquadra em um *cluster* e que não há sobreposição de áreas. Por outro lado, a Figura 20.2 ilustra uma situação de *cluster* mais encontrada na prática. Na Figura 20.2, as fronteiras de alguns *clusters* não são nítidas e a classificação dos consumidores nem sempre é óbvia, já que muitos deles podem ser enquadrados em um ou outro *cluster*.

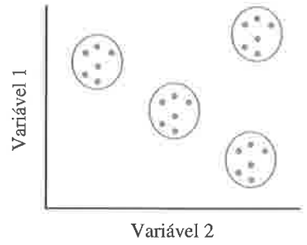

Figura 20.1 Uma situação ideal de *cluster*.

Tanto a análise de *clusters* quanto a análise discriminante se referem à classificação. Entretanto, a análise discriminante exige o conhecimento prévio da composição do grupo ou *cluster* para cada objeto ou caso incluídos para então definir uma regra de classificação. Em contrapartida, na análise de *clusters* não há qualquer informação *a priori* sobre a composição do grupo ou *cluster* para qualquer um de seus objetos. Os grupos ou *clusters* são sugeridos pelos dados, e não definidos *a priori*.[4]

[1] Emma Reynolds, "Is Haagen-Dazs Shrewd to Drop Its Sexy Image?" Marketing (September 6, 2001): 17; Liz Stuart, "Haagen-Dazs Aims to Scoop a Larger Share," *Marketing Week*, 19 (46/2) (February 21, 1997): 26; Dwight J. Shelton, "Birds of a Geodemographic Feather Flock Together," *Marketing News* (August 28, 1987): 13.

* N. de R.T.: Também é conhecida por Análise *Q*, contrastando com a Análise Fatorial (Análise *R*).

[2] Para uma aplicação recente da análise de *clusters*, ver Wendy W. Moe e Peter S. Fader, "Modeling Hedonic Portfolio Products: A Joint Segmentation Analysis of Music Compact Disc Sales," *Journal of Marketing Research*, 38 (3) (August 2001): 376-388; e George Arimond, "A Clustering Method for Categorical Data in Tourism Market Segmentation Research," *Journal of Travel Research,* 39 (4) (May 2001): 391-397.

[3] Também estão disponíveis métodos de aglomeração sobrepostos que permitem que um objeto seja agrupado em mais de um *cluster*. Ver Anil Chaturvedi, J. Douglass Carroll, Paul E. Green e John A. Rotondo, "A Feature-Based Approach to Market Segmentation via Overlapping K-Centroids Clustering," *Journal a/Marketing Research,* 34 (August 1997): 370-377.

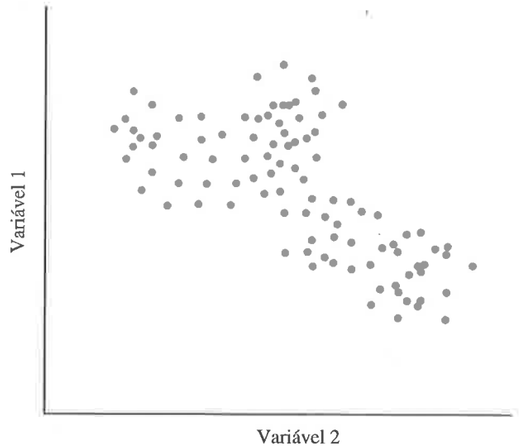

Figura 20.2 Uma situação prática de *cluster*.

[4] Podem-se encontrar excelentes discussões sobre os inúmeros aspectos da análise de *clusters* em Brian S. Everitt, Sabine Landau e Morven Leese, *Cluster Analysis,* 4th ed. (Oxford, UK: Oxford University Press, 2001); e H. Charles Romsburg, *Cluster Analysis for Researchers* (Melbourne: Krieger Publishing Company, 1990).

[5] Jafar Ali, "Micro-Market Segmentation Using a Neural Network Model Approach," *Journal of International Consumer Marketing* (2001): 7; Vicki Douglas, "Questionnaires Too Long? Try Variable Clustering," *Marketing News,* 29 (5) (February 27, 1995): 38; Girish Punj e David Stewart, "Cluster Analysis in Marketing Research: Review and Suggestions for Application," *Journal of Marketing Research,* 20 (May 1983): 134-148.

A análise de *clusters* tem sido utilizada em pesquisa de marketing para vários propósitos, incluindo:[5]

- *Segmentação do mercado.* Por exemplo, os consumidores podem ser agrupados com base nas vantagens que esperam da compra de um produto. Cada *cluster* consistiria em consumidores relativamente homogêneos quanto às vantagens que procuram.[6] Essa abordagem é chamada de *segmentação por vantagem ou benefícios*.

PESQUISA REAL

Os turistas exigentes, os estudiosos e os escapistas

Em um estudo sobre os padrões de decisão entre turistas em férias de várias nacionalidades, 260 entrevistados forneceram informações sobre seis tipos de orientação psicográfica: psicológica, educacional, social, de descanso, fisiológica e estética. Aplicou-se a análise de *clusters* para agrupar os entrevistados em segmentos psicográficos. Os resultados sugeriram que há três segmentos significativos baseados nesses estilos de vida. O primeiro segmento (53%) consistiu em indivíduos situados no alto de quase todas as escalas de estilo de vida. Esse grupo foi designado "os exigentes". O segundo grupo (20%), designado "os estudiosos", situou-se no alto da escala de nível da educação. O último grupo (26%) ficou no topo da escala do descanso e no início da escala social; foi chamado de "os escapistas". Formularam-se estratégias específicas de mercado para atrair elementos em cada segmento. Um estudo realizado em 2000 examinou a imagem que a Tailândia tinha para 510 turistas como destino de

[6] Para observar o uso da análise de *clusters* na segmentação, ver George Arimond, "A Clustering Method for Categorical Data in Tourism Market Segmentation Research," *Journal of Travel Research,* 39 (4) (May 2001): 391-397; William D. Neal, "Advances in Market Segmentation," *Marketing Research* (Spring 2001): 14-18; e Mark Peterson e Naresh K. Malhotra, "A Global View of Quality of Life: Segmentation Analysis of 165 Countries," *International Marketing Review,* 17 (I)(2000): 56-73.

A análise de *clusters* revelou distintos segmentos de marketing dentre os turistas de férias.

viagem internacional. O estudo avaliou o impacto da imagem do destino a partir da probabilidade de o viajante retornar à Tailândia. A amostra para esse estudo foi composta de viajantes internacionais que visitaram a Tailândia e que estavam partindo do Aeroporto Internacional de Bangkok. O estudo envolveu o uso de uma abordagem de amostra de três estágios que incluiu amostragem estratificada proporcional, de *cluster* e amostragem aleatória sistemática. A amostragem de *clusters* foi usada para selecionar aleatoriamente vôos de partida do Aeroporto Internacional de Bangkok. Os resultados do estudo revelaram que a Tailândia tem uma imagem negativa de problemas ambientais e sociais. Ao mesmo tempo, entretanto, tem uma imagem positiva como um destino de viagem seguro, associada com paisagens de grande beleza natural, cultura, gastronomia e hotéis, além de bons locais para compras. Acima de tudo, a maioria dos turistas, especialmente os "escapistas", indicaram que retornariam à Tailândia para passar as férias. Assim, a Tailândia deveria fazer um esforço especial para alcançar o segmento "escapista", já que o país atrairia a maioria desses turistas.[7] ■

- *Compreensão do comportamento do comprador*. A análise de *clusters* pode ser utilizada para identificar grupos homogêneos de compradores. A seguir, o comportamento de compra de cada grupo pode ser examinado separadamente, como no projeto da loja de departamentos, no qual os entrevistados foram agrupados com base na importância por eles atribuída a cada fator nos critérios utilizados para a seleção de uma loja de departamentos. A análise de *clusters* também foi usada para identificar os tipos de estratégia que os compradores de automóveis utilizam para obter informações externas.
- *Identificação das oportunidades para um novo produto*. Ao aglomerar marcas e produtos, é possível determinar conjuntos concorrentes dentro do mercado. As marcas no mesmo *cluster* concorrem mais acirradamente entre elas do que com marcas de outros *clusters*. Uma firma pode analisar suas ofertas atuais comparando-as com as de seus concorrentes, a fim de identificar oportunidades potenciais para um novo produto.
- *Seleção de mercados de teste*. Ao formar *clusters* homogêneos de cidades, é possível selecionar cidades comparáveis para testar várias estratégias de marketing.
- *Redução de dados*. A análise de *clusters* pode servir como instrumento de redução de dados para criar *clusters* ou subgrupos de dados mais fáceis de manejar do que observações individuais. A seguir, são feitas análises multivariadas subseqüentes com base nos *clusters*, e não mais nas observações individuais. Por exemplo, para descrever diferenças de comportamento dos consumidores em relação ao uso de determinado produto, os consumidores podem ser aglomerados primeiramente em grupos. Examinam-se então as diferenças entre os grupos com auxílio da análise discriminante múltipla.

[7] Tom J. Brown, Hailin Qu e Bongkosh Ngamsom Rittichainuwat, "Thailand's International Travel Image: Mostly Favorable," *Cornell Hotel and Restaurant Administration Quarterly*, 42 (2) (April 2001): 85-95; Chul-Min Mo, Mark E. Havitz e Dennis R. Howard, "Segmenting Travel Markets with the International Tourism Role (ITR) Scale," *Journal of Travel Research*, 33(1) (Summer 1994): 24-31; George P. Moschis e Daniel C. Bello, "Decision-Making Patterns Among International Vacationers: A Cross-Cultural Perspective," *Psychology & Marketing* (Spring 1987): 75-89.

ESTATÍSTICAS ASSOCIADAS À ANÁLISE DE *CLUSTERS*

Antes de discutir as estatísticas associadas à análise de *clusters*, vale mencionar que a maioria dos métodos de aglomeração consiste em procedimentos relativamente simples que não têm o apoio de um raciocínio estatístico rigoroso. Ao contrário, a maioria dos métodos de aglomeração é heurística, baseada em algoritmos. Assim, a análise de *clusters* contrasta fortemente com a análise da variância, de regressão, a análise discriminante e a análise fatorial, que se baseiam em um rigoroso raciocínio estatístico. Embora muitos métodos de aglomeração tenham importantes propriedades estatísticas, não se pode deixar de reconhecer a simplicidade de fundamentos de tais métodos.[8] As estatísticas e conceitos a seguir estão ligados à análise de *clusters*.

Esquema de aglomeração. Um esquema de aglomeração fornece informações sobre os objetos ou casos a serem combinados em cada estágio de um procedimento hierárquico de aglomeração.

***Centróide de* clusters.** São os valores médios das variáveis para todos os casos ou objetos em um *cluster* particular.

***Centros de* clusters.** São os pontos de partida iniciais em um *cluster* não-hierárquico. Os *clusters* são construídos em torno desses centros ou *sementes*.

***Associação a um* cluster.** Indica o *cluster* ao qual pertence cada objeto ou caso.

Dendrograma. Também chamado de *gráfico em árvore*, é um dispositivo gráfico para apresentar os resultados de aglomeração. As linhas verticais representam *clusters* unidos. A posição da reta na escala indica as distâncias às quais os *clusters* foram unidos. O dendrograma é lido da direita para a esquerda. A Figura 20.8 é um dendrograma.

***Distâncias entre centros de* clusters.** Indicam o grau de separação dos pares individuais de *clusters*. Os *clusters* que se apresentam muito separados são distintos e, por conseguinte, desejáveis.

Diagrama em sincelos (*icicle*). Um diagrama em sincelos (pingentes de gelo) é uma apresentação gráfica de resultados de *clusters*, e é chamado assim porque se assemelha a uma fileira de pingentes de gelo no beiral de um telhado. As colunas correspondem aos objetos que estão sendo aglomerados e as linhas correspondem ao número de *clusters*. Um diagrama em sincelos é lido de baixo para cima. A Figura 20.7 é um diagrama em sincelos.

Matriz de coeficientes de semelhança/distância. Uma matriz de coeficientes de semelhança/distância é o triângulo inferior de uma matriz que contém distâncias pareadas entre objetos ou casos.

[8] Brian S. Everitt, Sabine Landau e Morven Leese, *Cluster Analysis,* 4th ed. (Oxford, UK: Oxford University Press, 2001).

COMO FAZER ANÁLISE DE *CLUSTERS*

As etapas de uma análise de *clusters* estão descritas na Figura 20.3. O primeiro passo consiste em formular o problema definindo as variáveis sobre as quais se baseará a aglomeração. Escolhe-se, a seguir, uma medida apropriada de distância. A medida de distância determina quão semelhantes ou diferentes são os objetos que estão sendo aglomerados. Foram elaborados vários procedimentos de aglomeração, e o pesquisador deve escolher aquele que se afigura apropriado ao problema em questão. A decisão sobre o número de *clusters* exige julgamento por parte do pesquisador. Os *clusters* derivados devem ser interpretados em termos das variáveis usadas para constituí-los e perfilados em termos de variáveis adicionais importantes. Finalmente, o pesquisador precisa avaliar a validade do procedimento de aglomeração.

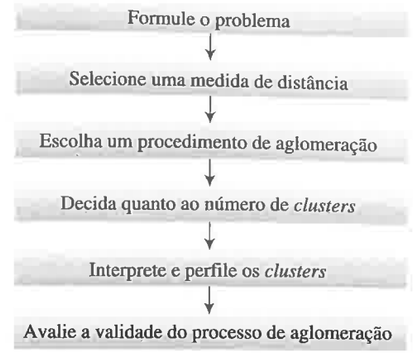

Figura 20.3 Como fazer uma análise de *cluster*.

Formular o problema

Talvez a parte mais importante da formulação de um problema de *cluster* seja a escolha das variáveis sobre as quais se baseará o procedimento de aglomeração. A simples inclusão de uma ou duas variáveis irrelevantes pode distorcer uma solução que, não fosse por isso, se revelaria útil. Basicamente, o conjunto de variáveis escolhidas deve descrever a semelhança entre objetos em termos relevantes para o problema de pesquisa de marketing. As variáveis devem ser escolhidas com base em pesquisas passadas, na teoria, ou em função das hipóteses que estão sendo testadas. Na pesquisa exploratória, o pesquisador deve exercer julgamento e aplicar a intuição.

A título de ilustração, consideremos um *cluster* de consumidores baseado em atitudes em relação a compras. Com base em pesquisa anterior, identificaram-se seis variáveis referentes à atitude. Solicitou-se aos consumidores que expressassem seu grau de concordância com as seguintes afirmações (com base em uma escala de sete pontos: 1 = discordo, 7 = concordo):

V_1: Fazer compras é divertido.
V_2: As compras afetam meu orçamento.
V_3: Combino compras com refeições fora de casa.
V_4: Procuro a melhor oferta quando compro.
V_5: Não me preocupo com compras.
V_6: Podemos economizar bastante comparando preços.

A Tabela 20.1 exibe os dados obtidos de uma amostra pré-teste de 20 entrevistados. Observe que, na prática, a aglomeração se faz em amostras muito maiores – 100 ou mais. Para ilustrar o procedimento de aglomeração, utilizou-se uma amostra de tamanho pequeno.

TABELA 20.1

Dados sobre atitudes para aglomeração

Caso Nº	V_1	V_2	V_3	V_4	V_5	V_6
1	6	4	7	3	2	3
2	2	3	1	4	5	4
3	7	2	6	4	1	3
4	4	6	4	5	3	6
5	1	3	2	2	6	4
6	6	4	6	3	3	4
7	5	3	6	3	3	4
8	7	3	7	4	1	4
9	2	4	3	3	6	3
10	3	5	3	6	4	6
11	1	3	2	3	5	3
12	5	4	5	4	2	4
13	2	2	1	5	4	4
14	4	6	4	6	4	7
15	6	5	4	2	1	4
16	3	5	4	6	4	7
17	4	4	7	2	2	5
18	3	7	2	6	4	3
19	4	6	3	7	2	7
20	2	3	2	4	7	2

Selecionar uma medida de distância ou de semelhança

Como o objetivo da aglomeração é agrupar objetos semelhantes, torna-se necessária alguma medida para avaliar quão semelhantes ou diferentes são os objetos. A abordagem mais comum consiste em avaliar a semelhança em termos de distância entre pares de objetos. Os objetos com menor distância entre si são mais semelhantes que objetos com maior distância. Há várias maneiras de calcular a distância entre dois objetos.[9]

A medida de semelhança mais comumente utilizada é a ***distância euclidiana*** ou o seu quadrado. A distância euclidiana é a raiz quadrada da soma dos quadrados das diferenças dos valores para cada variável. Outras medidas de distância também estão disponíveis. A *distância de Manhattan* (ou *city block*) entre dois objetos é a soma dos valores absolutos das diferenças para cada variável. A *distância de Chebychev* entre dois objetos é o valor absoluto da maior diferença de valores para qualquer variável. Para nosso exemplo, vamos utilizar o quadrado da distância euclidiana.

distância euclidiana
Raiz quadrada da soma dos quadrados das diferenças de valores para cada variável.

No caso de as variáveis serem medidas em unidades muito diferentes, a solução por aglomerado será influenciada pelas unidades de medida. Em um estudo sobre compras em um supermercado, as variáveis de atitude podem ser medidas em uma escala Likert de nove pontos; preferência, em termos da freqüência de visitas mensais e quantia gasta em compras; e lealdade à marca, em termos de percentagem de despesas com compras de alimentos feitas no supermercado favorito. Nesses casos, antes de aglomerar os entrevistados, devemos padronizar os dados reescalonando as variáveis de modo a terem média zero e desvio-padrão um. Embora a padronização possa remover a influência da unidade de medida, pode também reduzir as diferenças entre grupos em variáveis que melhor discriminem grupos ou *clusters*. É conveniente também eliminar os casos com valores atípicos (*outliers*).[10]

A utilização de diferentes medidas de distância pode levar a diferentes resultados de aglomeração. Assim, é conveniente utilizar medidas diferentes e comparar os resultados. Escolhida uma medida de distância ou de semelhança, passamos a escolher um procedimento de aglomeração.

[9] Para uma discussão detalhada sobre as diferentes medidas de similaridade e as fórmulas para calculá-las, ver Eric T. Bradlow, "Subscale Distance and Item Clustering Effects in Self-Administered Surveys: A New Metric," *Journal of Marketing Research* (May 2001): 254-261; Victor Chepoi e Feeder Dragan, "Computing a Median Point of a Simple Rectilinear Polygon," *Information Processing Letters,* 49 (6) (March 22, 1994): 281-285; e H. Charles Romsburg, *Cluster Analysis for Researchers* (Melbourne: Krieger Publishing Company, 1990).

[10] Para uma discussão mais detalhada das questões envolvidas na padronização, ver H. Charles Romsburg, *Cluster Analysis for Researchers* (Melbourne: Krieger Publishing Company, 1990).

Escolher um procedimento de aglomeração

A Figura 20.4 é uma classificação de procedimentos de aglomeração. Estes podem ser hierárquicos ou não-hierárquicos. A ***aglomeração hierárquica*** se caracteriza pelo estabelecimento de uma hierarquia, ou estrutura em forma de árvore. Os métodos hierárquicos podem ser aglomerativos ou divisivos. O ***cluster aglomerativo*** começa com cada objeto em um *cluster* separado. Os *clusters* são formados agrupando-se os objetos em *clusters* cada vez maiores. O procedimento continua até que todos os objetos sejam membros de um único *cluster*. O ***cluster divisivo*** começa com todos os objetos agrupados em um único *cluster*, que é então partido, ou dividido, até que cada objeto esteja em um *cluster* separado.

aglomeração hierárquica
Procedimento de aglomeração caracterizado pelo desenvolvimento de uma hierarquia ou estrutura em forma de árvore.

cluster *aglomerativo*
Procedimento de aglomeração hierárquica em que cada objeto tem início em um *cluster* separado. Formam-se os *clusters* agrupando-se os objetos em *clusters* cada vez maiores.

cluster *divisivo*
Procedimento de aglomeração hierárquica em que todos os objetos partem de um *cluster* gigante. Formam-se os *clusters* dividindo-se esse *cluster* gigante em *clusters* cada vez menores.

Os métodos aglomerativos são de uso comum em pesquisa de marketing. Consistem em métodos de encadeamento (*linkage methods*), métodos de erros de soma de quadrados, ou métodos de variância, e métodos centróides. Os ***métodos de encadeamento*** compreendem o encadeamento único, o encadeamento completo e o encadeamento médio. O método de ***encadeamento único*** se baseia na distância mínima (regra do vizinho mais próximo). Os dois primeiros objetos aglomerados são os que apresentam a menor distância entre eles. Identifica-se a menor distância mais próxima aglomerando o terceiro objeto com os dois primeiros, ou formando-se um novo *cluster* de dois objetos. Em cada estágio, a distância entre dois *clusters* é a distância entre seus dois pontos mais próximos (ver Figura 20.5). Dois *clusters* podem incorporar-se em qualquer estágio por meio do encadeamento mais curto entre eles. Continua-se o procedimento até que todos os objetos estejam em um único *cluster*. O método do encadeamento único só funciona se todos os *clusters* estiverem bem-definidos. O método do ***encadeamento completo*** é similar ao encadeamento único, mas se baseia na distância *máxima*, ou método do vizinho *mais afastado*. No encadeamento completo, a distância entre dois *clusters* é calculada como a distância entre seus dois pontos mais afastados. O método do ***encadeamento médio*** funciona de maneira análoga, mas, nesse método, a distância entre dois aglomerados se define como a média das distâncias entre todos os pares de objetos, onde cada membro de um par provém de cada um dos *clusters* (Figura 20.5). Como se pode constatar, o método do encadeamento médio utiliza informações sobre todos os pares de distâncias, e não apenas as distâncias mínima ou máximas. Por essa razão, costuma ser preferido aos métodos de encadeamento único e de encadeamento completo.

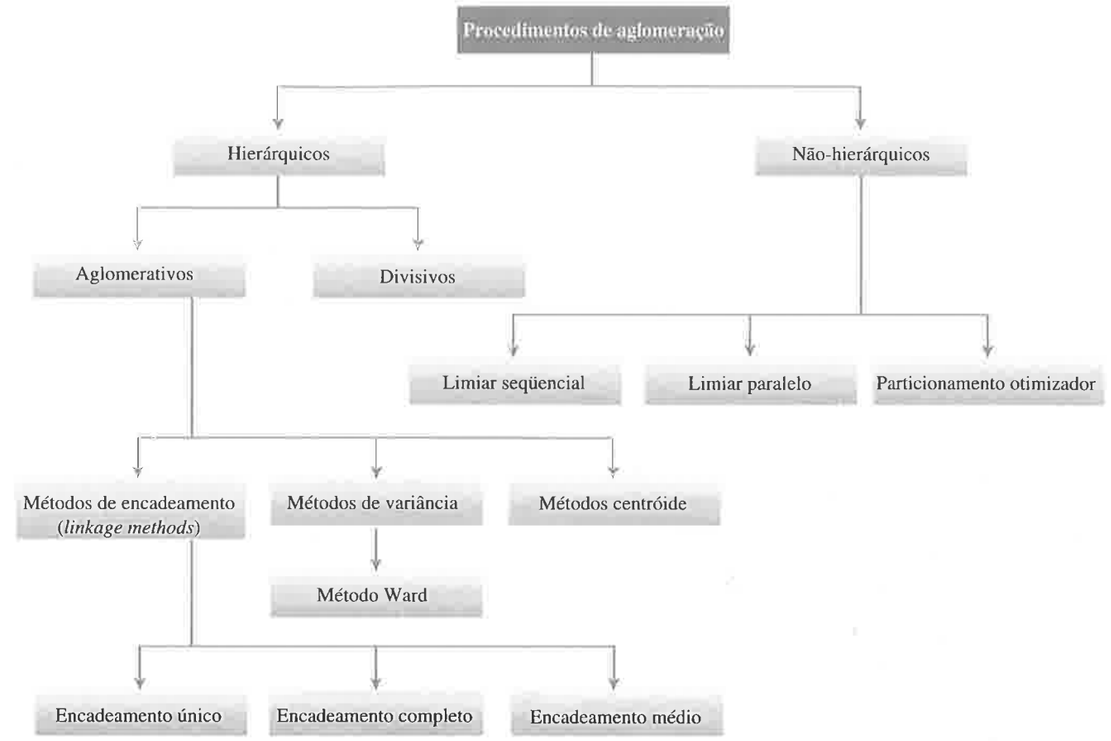

Figura 20.4 Uma classificação dos procedimentos de aglomeração.

***métodos de encadeamento* (linkage methods)**
Métodos aglomerativos de *cluster* hierárquico que agrupam os objetos com base no cálculo da distância entre eles.

encadeamento único
Método de encadeamento baseado na distância mínima ou regra do vizinho mais próximo.

encadeamento completo
Método de encadeamento baseado na distância máxima ou regra do vizinho mais afastado.

encadeamento médio
Método de encadeamento baseado na distância média entre todos os pares de objetos, onde cada membro de um par é extraído de cada um dos *clusters*.

método da variância
Método aglomerativo de *cluster* hierárquico, em que os *clusters* são gerados de modo a minimizar a variância dentro do *cluster*.

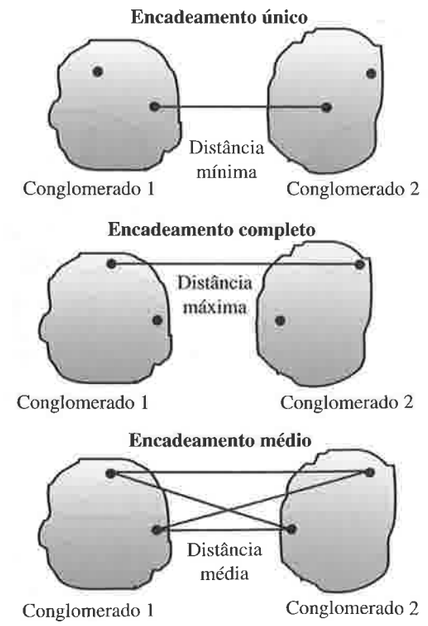

Figura 20.5 Métodos de aglomeração por encadeamento.

Os ***métodos de variância*** procuram gerar *clusters* para minimizar a variância dentro dos *clusters*. O ***método Ward*** é um método de variância bastante utilizado. Para cada *cluster* calculam-se as médias de todas as variáveis. Calcula-se então, para cada objeto, o quadrado da distância euclidiana às médias do *cluster* (Figura 20.6). Somam-se essas distâncias para todos os objetos. Em cada estágio, combinam-se os dois *clusters* que apresentarem menor aumento na soma global de quadrados dentro dos *clusters*. No ***método centróide***, a distância entre dois *clusters* é a distância entre seus centróides (médias para todas as variáveis), conforme a Figura 20.6. Cada vez que se agrupam objetos, calcula-se um novo centróide. Dos métodos hierárquicos, os métodos do encadeamento médio e o Ward têm-se revelado superiores aos outros métodos.[11]

O segundo tipo de procedimento de aglomeração, os métodos ***não-hierárquicos de aglomeração***, costumam ser chamados de aglomeração de *k* médias. Esses métodos compreendem o limiar seqüencial, o limiar paralelo e o particionamento otimizador. No ***método do limiar seqüencial***, escolhe-se um centro de aglomeração e todos os objetos a menos de um valor pré-determinado a contar do centro são agrupados juntos. Escolhe-se então um novo centro de aglomeração, ou semente, repetindo-se o procedimento para os pontos não-aglomerados. Um objeto aglomerado a uma semente não é mais levado em consideração para aglomerações subseqüentes. O ***método do limiar paralelo*** opera de modo semelhante, com a diferença que se escolhem simultaneamente vários centros de aglomeração e os objetos dentro do limiar são agrupados com o centro mais próximo. O ***método de particionamento otimizador*** difere dos dois anteriores porque os objetos podem ser posteriormente reatribuídos a *clusters* a fim de otimizar um critério global, tal como a distância média dentro de *clusters* para um dado número de *clusters*.

método Ward
Método de variância em que se deve minimizar o quadrado da distância euclidiana às médias dos aglomerados.

método centróide
Método de variância de aglomeração hierárquica em que a distância entre dois aglomerados é a distância entre seus centróides (médias para todas as variáveis).

cluster não-hierárquico
Procedimento que inicialmente determina ou assume um centro de *cluster* e em seguida agrupa todos os objetos que estão a menos de um valor pré-especificado do centro.

método do limiar seqüencial
Método de aglomeração não-hierárquica em que se escolhe um centro de *cluster* e se agrupam todos os objetos que estão a menos de um valor especificado em relação ao centro.

método do limiar paralelo
Método de aglomeração não-hierárquica que especifica de uma só vez vários centros de *clusters*. Todos os objetos que estão a menos de um valor pré-determinado do centro são agrupados juntos.

método do particionamento otimizador
Método de aglomeração não-hierárquica que permite reatribuir posteriormente objetos a *clusters* de modo a otimizar um critério global.

Duas grandes desvantagens dos procedimentos não-hierárquicos são que o número de *clusters* deve ser preestabelecido e que a escolha dos centros de *clusters* é arbitrária. Além disso, os resultados da aglomeração podem depender de como se escolhem os centros. Muitos programas não-hierárquicos escolhem os primeiros *k* (*k* = número de *clusters*) casos sem valores faltantes como centros iniciais dos *clusters*. Assim, os resultados da aglomeração podem depender da ordem das observações nos dados. Mas a aglomeração não-hierárquica é mais rápida que os métodos hierárquicos e tem seu mérito quando o número de objetos ou observações é grande. Sugeriu-se que os métodos hierárquicos e os não-hierárquicos sejam usados de forma conjugada. Primeiro, obtém-se uma solução inicial por meio de um procedimento hierárquico, como o encadeamento médio ou o método Ward. Os números de *clusters* e de centróides são então usados como entradas para o método de particionamento otimizador.[12]

[11] Brian Everitt, Sabine Landau e Morven Leese, *Cluster Analysis,* 4th ed. (Oxford, UK: Oxford University Press, 2001); e G. Milligan, "An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms," *Psychometrika,* 45 (September 1980): 325-342.

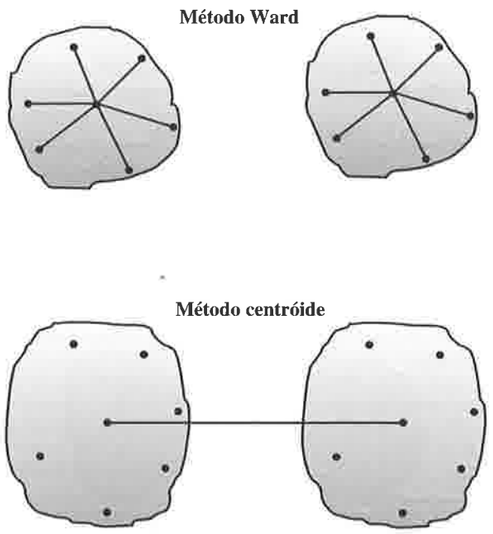

Figura 20.6 Outros métodos de *cluster* aglomerativo.

[12] Brian Everitt, Sabine Landau e Morven Leese, *Cluster Analysis,* 4th ed. (Oxford, UK: Oxford University Press, 2001).

A escolha de um método de aglomeração e a escolha de uma medida de distância estão inter-relacionadas. Por exemplo, devem-se usar os quadrados das distâncias euclidianas com os métodos Ward e centróides. Vários procedimentos não-hierárquicos também utilizam quadrados de distâncias euclidianas.

Utilizaremos o método Ward para ilustrar a aglomeração hierárquica. A Tabela 20.2 dá o resultado obtido aglomerando-se os dados da Tabela 20.1. O esquema de aglomeração contém informações úteis, mostrando o número de casos ou *clusters* combinados em cada estágio. A primeira linha representa o estágio 1, com 19 *clusters*. Nesse estágio, são combinados os entrevistados 14 e 16, conforme mostram as colunas "*Clusters* combinados". O quadrado da distância euclidiana entre esses dois entrevistados é dado na coluna intitulada "Coeficiente". A coluna "Estágio em que o *cluster* aparece primeiramente" indica o estágio em que primeiro se forma um *cluster*. A título de ilustração, uma entrada de 1 no estágio 6 indica que o entrevistado 14 foi agrupado inicialmente no estágio 1. A última coluna, "Próximo estágio", indica o estágio em que outro caso (entrevistado) ou *cluster* é combinado com este. Como o número na primeira linha da última coluna é 6, vemos que, no estágio 6, o entrevistado 10 é combinado com o 14 e o 16, formando um *cluster* único. Da mesma for-

TABELA 20.2

Resultados da aglomeração hierárquica

RESUMO DO PROCESSAMENTO DE CASOS [a, b]

Válidos		Casos faltantes		Total	
N	Percentagem	N	Percentagem	N	Percentagem
20	100,0	0	0,0	20	100,0

[a] Quadrado da distância euclidiana usado
[b] Método Ward

MÉTODO WARD
ESQUEMA DE AGLOMERAÇÃO

	Clusters combinados			Estágio em que o *cluster* aparece primeiramente		
Estágio	*Cluster* 1	*Cluster* 2	Coeficiente	*Cluster* 1	*Cluster* 2	Próximo estágio
1	14	16	1,000	0	0	6
2	6	7	2,000	0	0	7
3	2	13	3,500	0	0	15
4	5	11	5,000	0	0	11
5	3	8	6,500	0	0	16
6	10	14	8,167	0	1	9
7	6	12	10,500	2	0	10
8	9	20	13,000	0	0	11
9	4	10	15,583	0	6	12
10	1	6	18,500	0	7	13
11	5	9	23,000	4	8	15
12	4	19	27,750	9	0	17
13	1	17	33,100	10	0	14
14	1	15	41,333	13	0	16
15	2	5	51,833	3	11	18
16	1	3	64,500	14	5	19
17	4	18	79,667	12	0	18
18	2	4	172,667	15	17	19
19	1	2	328,600	16	18	0

ASSOCIAÇÃO A UM *CLUSTER*

Caso	4 *clusters*	3 *clusters*	2 *clusters*
1	1	1	1
2	2	2	2
3	1	1	1
4	3	3	2
5	2	2	2
6	1	1	1
7	1	1	1
8	1	1	1

(*continua*)

TABELA 20.2

Resultados da aglomeração hierárquica

9	2	2	2
10	3	3	2
11	2	2	2
12	1	1	1
13	2	2	2
14	3	3	2
15	1	1	1
16	3	3	2
17	1	1	1
18	4	3	2
19	3	3	2
20	2	2	2

ma, a segunda linha representa o estágio 2 com 18 *clusters*. No estágio 2, estão agrupados os entrevistados 6 e 7.

Outra parte importante do resultado aparece no gráfico em sincelos (*icicles*) da Figura 20.7. As colunas correspondem aos objetos que estão sendo aglomerados, neste caso, os entrevistados rotulados de 1 a 20. As linhas correspondem aos números de *clusters*. Essa figura deve ser lida de baixo para cima. Inicialmente, todos os casos são considerados como *clusters* individuais. Como há 20 entrevistados, há 20 *clusters* iniciais. Em um primeiro passo, combinam-se os dois objetos mais próximos, o que resulta em 19 *clusters*. A última linha da Figura 20.7 mostra esses 19 *clusters*. Os dois casos – entrevistados 14 e 16 – que foram combinados neste estágio não têm espaço vago (em branco) separando-os. A linha número 18 corresponde ao próximo estágio, com 18 *clusters*. A essa altura, agrupam-se os entrevistados 2 e 13; há 18 *clusters*; 16 deles consistem em entrevistados individuais e dois contêm dois entrevistados cada um. Cada passo subseqüente conduz à formação de um novo *cluster*, de uma de três maneiras: (1) agrupam-se dois casos individuais, (2) agrupa-se um caso a um *cluster* já existente ou (3) agrupam-se dois *clusters*.

Outro dispositivo gráfico útil para exibir resultados de *clusters* é o dendrograma (ver Figura 20.8). O dendrograma é lido da esquerda para a direita. As linhas verticais representam *clusters* que são unidos. A posição da reta na escala indica as distâncias às quais os *clusters* foram unidos. Como muitas das distâncias nos primeiros estágios têm magnitudes semelhantes, é difícil dizer em que seqüência se formaram alguns dos *clusters* iniciais. Entretanto, é claro que, nos dois últimos estágios, as distâncias em que os *clusters* estão sendo combinados são grandes. Essa informação é útil para decidir sobre o número de *clusters*.

CASO / NÚMERO DE *CLUSTERS*	18	19	16	14	10	4	20	9	11	5	13	2	8	3	15	17	12	7	6	1
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
3	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
4	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
5	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
6	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
7	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
8	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
9	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
10	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
11	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
12	X		X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X
13	X		X	X	X	X		X		X	X	X	X	X	X	X	X	X	X	X
14	X		X	X	X	X		X		X	X	X	X	X	X	X		X	X	X
15	X		X	X	X			X		X	X	X	X	X	X	X		X	X	X
16	X		X	X	X			X		X	X	X		X	X	X		X	X	X
17	X		X	X	X			X		X		X		X	X	X		X	X	X
18	X		X	X	X			X		X		X		X	X	X		X	X	X
19	X		X	X	X			X		X		X		X	X	X		X	X	X

Figura 20.7 Gráfico em sincelos verticais (*icicle*) utilizando o método Ward.

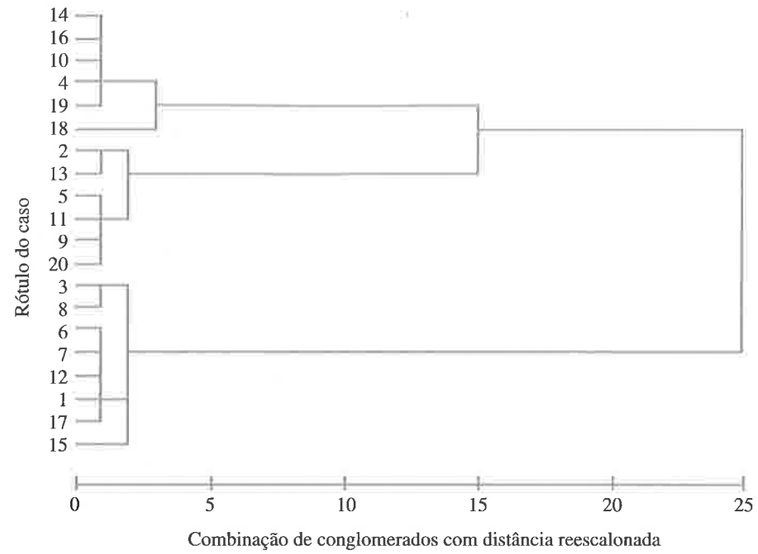

Figura 20.8 Dendrograma utilizando o método Ward.

É possível também obter informações quanto à associação dos casos aos *clusters*, desde que se especifique o número do *cluster*. Embora essa informação possa ser obtida a partir de um gráfico em sincelos (*icicles*), uma visualização na forma de tabela também pode ser útil. A Tabela 20.2 dá a associação dos *clusters* para os diversos casos, dependendo de a solução final conter dois, três ou quatro *clusters*; esse tipo de informação ajuda a decidir quanto ao número de *clusters*.

Decidir sobre o número de *clusters*

Um problema relevante na análise de *clusters* é a decisão quanto ao seu número. Não há regras difíceis nem fáceis; dispomos, todavia, de algumas diretrizes:

1. Considerações teóricas, conceituais ou práticas podem sugerir um certo número de *clusters*. Por exemplo, se a finalidade da aglomeração for identificar segmentos de mercado, a administração pode precisar de um número específico de *clusters*.
2. Na aglomeração hierárquica, as distâncias em que os *clusters* são combinados podem ser utilizadas como critérios. Essa informação pode ser obtida do esquema de aglomeração ou do dendrograma. Em nosso caso, vemos pelo esquema de aglomeração da Tabela 20.2 que o valor na coluna "Coeficiente" mais do que dobra repentinamente entre os estágios 17 (3 *clusters*) e 18 (2 *clusters*). Da mesma forma, nos dois últimos estágios do dendrograma da Figura 20.8, os *clusters* estão sendo combinados a grandes distâncias. Portanto, parece adequada uma solução de três *clusters*.
3. No *cluster* não-hierárquico, a razão da variância total dentro do grupo para a variância entre grupos pode ser grafada em função do número de *clusters*. O ponto em que ocorre um "cotovelo", ou uma virada brusca, indica um número apropriado de *clusters*. Em geral não compensa aumentar o número de *clusters* além desse ponto.
4. Os tamanhos relativos dos *clusters* devem ser significativos. Na Tabela 20.2, mediante uma contagem simples de freqüência nos *clusters*, vemos que uma solução de três *clusters* resulta em *clusters* com oito, seis e seis elementos. Entretanto, passando a uma solução com quatro *clusters*, os tamanhos respectivos são oito, seis, cinco e um. Como não tem sentido lidarmos com um *cluster* de apenas um caso, nessa situação é preferível uma solução com três *clusters*.

Interpretar e perfilar os *clusters*

Interpretar e traçar o perfil dos *clusters* envolve o exame dos respectivos centróides. Os centróides representam os valores médios dos objetos contidos no *cluster* em cada uma das variáveis. Os centróides permitem descrever cada *cluster* atribuindo-lhe um nome ou rótulo. Se o programa de aglomeração não imprimir essa informação, ela pode ser obtida através da análise discriminante. A Tabela 20.3 dá os centróides ou valores médios para cada *cluster* em nosso exemplo. O *cluster* 1 tem valores relativamente altos nas variáveis V_1 (fazer compras é divertido) e V_3 (combino compras com refeições fora de casa). Tem também um valor baixo em V_5 (não me preocupo com compras). Logo, o *cluster* 1 pode ser rotulado como "compradores envolvidos e que gostam de diversão". Esse *cluster* consiste nos casos 1, 3, 6, 7, 8, 12, 15 e 17. O *cluster* 2 é precisamente o oposto, com valores baixos em V_1 e V_3 e um valor elevado em V_5; esse *cluster* pode ser rotulado como "compradores apáticos". Os elementos do *cluster* 2 são

TABELA 20.3

Centróides de *clusters*

	Médias de variáveis					
***Cluster* nº**	V_1	V_2	V_3	V_4	V_5	V_6
1	5,750	3,625	6,000	3,125	1,750	3,875
2	1,667	3,000	1,833	3,500	5,500	3,333
3	3,500	5,833	3,333	6,000	3,500	6,000

os casos 2, 5, 9, 11, 13 e 20. O *cluster* 3 tem valores altos em V_2 (as compras afetam meu orçamento), V_4 (procuro a melhor oferta quando compro) e V_6 (podemos economizar comparando preços). Assim, esse *cluster* pode ser rotulado de "compradores econômicos". O *cluster* 3 compreende os casos 4, 10, 14, 16, 18 e 19.

Em geral, é conveniente traçar o perfil dos *clusters* em termos de variáveis que não foram utilizadas na aglomeração e que podem incluir variáveis demográficas, psicográficas, de consumo do produto, de utilização dos meios de comunicação ou outras. Por exemplo, os *clusters* podem ter sido estabelecidos com base em benefícios procurados. Pode-se fazer um desenho posterior do perfil em termos de variáveis demográficas e psicográficas para enfocar esforços de marketing para cada *cluster*. As variáveis que se diferenciam significativamente entre *clusters* podem ser identificadas via análise discriminante e análise da variância de um fator.

Avaliar a confiabilidade e a validade

Dados os diversos julgamentos que a análise de *clusters* acarreta, nenhuma solução de aglomeração deve ser aceita sem alguma avaliação de sua confiabilidade e validade. Os procedimentos formais para avaliar a confiabilidade e a validade de soluções de *clusters* são complexos e nem sempre totalmente defensáveis.[13] Por isso, vamos omiti-los aqui. Não obstante, os procedimentos a seguir permitem uma verificação adequada da qualidade dos resultados da aglomeração.

1. Fazer análise de *clusters* sobre os mesmos dados utilizando diferentes medidas de distância. Comparar os resultados em relação às medidas para verificar a estabilidade da solução.
2. Utilizar métodos diferentes de aglomeração e comparar os resultados.
3. Separar aleatoriamente os dados em duas metades. Fazer a aglomeração separadamente sobre cada metade. Comparar os centróides dos *clusters* nas duas subamostras.
4. Eliminar variáveis aleatoriamente. Fazer a aglomeração com base no conjunto reduzido de variáveis. Comparar os resultados com os obtidos por aglomeração baseada em todo o conjunto de variáveis.
5. Em um *cluster* não-hierárquico, a solução pode depender da ordem dos casos no conjunto de dados. Faça repetições múltiplas utilizando diferentes ordens de casos até que a solução se estabilize.

Damos mais uma ilustração do *cluster* hierárquico com um estudo de diferenças entre estratégias de marketing de firmas estadunidenses, japonesas e inglesas.

PESQUISA REAL

É um mundo pequeno

Para um estudo de concorrentes estadunidenses, japoneses e ingleses, obtiveram-se dados de entrevistas pessoais detalhadas com dirigentes e tomadores de decisão em marketing para grupos definidos de produtos em 90 empresas. Para controlar as diferenças de mercado, a metodologia se baseou na comparação de 30 empresas inglesas com suas principais concorrentes estadunidenses e japonesas no mercado britânico. O estudo envolveu 30 trios de empresas, cada um composto de uma firma inglesa, uma estadunidense e uma japonesa que competiam diretamente umas com as outras.

A maioria dos dados sobre as características do desempenho, estratégia e organização das empresas foi coletada em escalas de diferencial semântico de cinco pontos. O primeiro estágio da análise envolveu a análise fatorial de variáveis que descreviam as estratégias e atividades de marketing das empresas. Empregaram-se escores fatoriais para identificar grupos de empresas similares utilizando a rotina de aglomeração hierárquica de Ward (método Ward). Chegou-se a uma solução de seis *clusters*.

Interpretou-se a composição dos seis *clusters* em termos do desempenho original, estratégia e variáveis organizacionais.

[13] Para uma discussão formal sobre confiabilidade, validade e teste de significância na análise de *clusters*, ver Michael J. Brusco, J. Dennis Cradit e Stephanie Stahl, "A Simulated Annealing Heuristic for a Bicriterion Partitioning Problem in Market Segmentation," *Journal of Marketing Research,* 39 (1) (February 2002): 99-109; Hui-Min Chen, "Using Clustering Techniques to Detect Usage Patterns in a Web-Based Information System," *Journal of the American Society for Information Science and Technology,* 52 (11) (September 2001): 888; S. Dibbs e P. Stern, "Questioning the Reliability of Market Segmentation Techniques," *Omega*, 23 (6) (December 1995): 625-636; G. Ray Funkhouser, "A Note on the Reliability of Certain Clustering Algorithms," *Journal of Marketing Research,* 30 (February 1983): 99-102; T. D. Klastorin, "Assessing Cluster Analysis Results," *Journal of Marketing Research,* 20 (February 1983): 92-98; e S. J. Arnold, "A Test for Clusters," *Journal of Marketing Research*, 16 (November 1979): 545-551.

CLUSTERS ESTRATÉGICOS

Cluster	*I*	*II*	*III*	*IV*	*V*	*VI*
Nome	*Inovadores*	*Negociantes de qualidade*	*Negociantes de preço*	*Negociantes de produto*	*Negociantes experimentados*	*Vendedores agressivos*
Tamanho	22	11	14	13	13	17
Bem sucedido (%)	55	100	36	38	77	41
Nacionalidade (%):						
japonesa	59	46	22	31	15	18
estadunidense	18	36	14	31	54	53
inglesa	23	18	64	38	31	29

Todos os *clusters* contiveram algumas empresas bem-sucedidas, embora uns significativamente mais que outros. Os *clusters* apoiaram a hipótese de que as empresas bem-sucedidas eram semelhantes, independentemente da nacionalidade, porque em todos os *clusters* se encontraram empresas estadunidenses, inglesas e japonesas. Constatou-se, entretanto, uma preponderância de empresas japonesas nos *clusters* mais bem-sucedidos e uma predominância de empresas inglesas nos dois *clusters* com menor sucesso. Aparentemente, as empresas japonesas não desenvolvem estratégias exclusivas; ao contrário, a maioria delas utiliza estratégias que se revelam eficazes no mercado inglês.

Os resultados indicam que há estratégias genéricas que caracterizam empresas bem-sucedidas independentemente do seu ramo. É possível identificar três estratégias bem-sucedidas. A primeira é a dos negociantes de qualidade. Essas empresas têm força em marketing e em pesquisa e desenvolvimento; elas concentram seus recursos técnicos mais na busca de alta qualidade do que em simples inovações. Essas empresas se caracterizam como organizações empreendedoras, pelo planejamento a longo prazo e por um senso bem comunicado de missão. A segunda estratégia genérica é a dos inovadores, que se revelam mais fracos em pesquisa e desenvolvimento avançados, mas são empreendedores e motivados por uma busca a inovação. O último grupo bem-sucedido é o dos negociantes experimentados, que são essencialmente orientados para o lucro e têm acentuadas habilidades de marketing. Os três parecem consistir em negócios essencialmente orientados para o mercado. Os Estados unidos, em 2000/2001, investiram em cerca de 50% de todos os projetos britânicos, e o Japão também continua sendo um investidor fundamental, especialmente nos mercados de automóveis e de eletrônicos.[14] ■

[14] William Pedder, "Annual Report," *Invest UK Web Site* (March 31, 2001)(*www.invest.uk.com/investing/annual.cfm?d=ar_cereport&action=pdisp*); John Saunders, Rosalind H. Forrester, "Capturing Learning and Applying Knowledge: An Investigation of the Use of Innovation Teams in Japanese and American Automotive Firms," *Journal of Business Research,* 47 (1) (January 2000): 35; Peter Doyle, John Saunders e Veronica Wong, "International Marketing Strategies and Organizations: A Study of U.S., Japanese, and British Competitors," in Paul Bloom, Russ Winer, Harold H. Kassarjian. Dehra L. Scammon, Bart Weitz, Robert E. Spekman, Vijay Mahajan e Michael Levy, Eds., *Enhancing Knowledge Development in Marketing,* Series No. 55 (Chicago: American Marketing Association, 1989): 100-104

APLICAÇÕES DE AGLOMERAÇÃO NÃO-HIERÁRQUICA

Ilustramos o procedimento não-hierárquico utilizando os dados da Tabela 20.1 e um método de particionamento otimizador. Com base nos resultados da aglomeração hierárquica, foi predeterminada uma solução de três *clusters*. A Tabela 20.4 apresenta os resultados. Os centros iniciais dos *cluster* são os valores dos três casos selecionados aleatoriamente. Em alguns programas, são selecionados os primeiros três casos. Os centros de classificação dos *clusters* são centros provisórios usados para a atribuição de casos. Cada caso é atribuído ao centro de classificação do *cluster* mais próximo. Os centros de classificação são atualizados até que se atinjam os critérios de finalização. Os centros finais de aglomeração representam as médias das variáveis para os casos nos *clusters* finais. No SPSS Windows, eles são arredondados até o inteiro mais próximo.

A Tabela 20.4 também apresenta a associação a um *cluster* e a distância entre cada caso e seu centro de classificação. Observe que as inclusões em *clusters* dadas na Tabela 20.2 (aglomeração hierárquica) e na Tabela 20.4 (aglomeração não-hierárquica) são idênticas. (O *cluster* 1 da Tabela 20.2 é rotulado como *cluster* 3 na Tabela 20.4, e o *cluster* 3 da Tabela 20.2 é rotulado como *cluster* 1 na Tabela 20.4). As distâncias entre os centros finais de *clusters* indicam que os pares de *clusters* estão bem separados. Apresenta-se o teste *F* univariado para cada variável de aglomeração. Esses testes *F* são apenas descritivos. Os casos ou objetos são sistematicamente atribuídos a *clusters* a fim de maximizar diferenças nas variáveis de aglomeração, de modo que as probabilidades resultantes não devem ser interpretadas como um teste da hipótese nula por não haver diferença entre *clusters*.

O exemplo seguinte, de escolha de um hospital, ilustra melhor a aglomeração não-hierárquica.

PESQUISA REAL

Segmentação com precisão cirúrgica

Para identificar segmentos de preferência por hospitais, utilizou-se a análise de *cluster* ao classificar entrevistados que escolheram hospitais para internação. A aglomeração baseou-se nas razões apresentadas pelos entrevistados para escolher determinado hospital. Compararam-se os perfis demográficos

TABELA 20.4

Resultados da aglomeração não-hierárquica

CENTROS INICIAIS DOS *CLUSTERS*

	Cluster 1	2	3
V_1	4	2	7
V_2	6	3	2
V_3	3	2	6
V_4	7	4	4
V_5	2	7	1
V_6	7	2	3

HISTÓRICO DE ITERAÇÃO[a]

	Alteração nos centros de *clusters*		
Iteração	**1**	**2**	**3**
1	2,154	2,102	2,550
2	0,000	0,000	0,000

[a] Convergência alcançada devido a pequena ou nenhuma alteração de distância. A distância máxima pela qual qualquer centro se alterou é 0,000. A distância mínima entre centros iniciais é 7,746.

ASSOCIAÇÃO A UM *CLUSTER*

Nº do caso	*Cluster*	Distância
1	3	1,414
2	2	1,323
3	3	2,550
4	1	1,404
5	2	1,848
6	3	1,225
7	3	1,500
8	3	2,121
9	2	1,756
10	1	1,143
11	2	1,041
12	3	1,581
13	2	2,598
14	1	1,404
15	3	2,828
16	1	1,624
17	3	2,598
18	1	3,555
19	1	2,154
20	2	2,102

CENTROS FINAIS DOS *CLUSTERS*

	Cluster 1	2	3
V_1	4	2	6
V_2	6	3	4
V_3	3	2	6
V_4	6	4	3
V_5	4	6	2
V_6	6	3	4

DISTÂNCIAS ENTRE CENTROS FINAIS DE *CLUSTERS*

Cluster	1	2	3
1		5,568	5,698
2	5,568		6,928
3	5,698	6,928	

ANOVA

	Cluster		Erro			
	Quadrado da média	gl	Quadrado da média	gl	F	sig.
V_1	29,108	2	0,608	17	47,888	0,000
V_2	13,546	2	0,630	17	21,505	0,000
V_3	31,392	2	0,833	17	37,670	0,000
V_4	15,713	2	0,728	17	21,585	0,000
V_5	22,537	2	0,816	17	27,614	0,000
V_6	12,171	2	1,071	17	11,363	0,000

Os testes *F* devem ser usados apenas para fins descritivos porque os *clusters* foram escolhidos para maximizar as diferenças entre os casos em diferentes *clusters*. Os níveis de significância observados não são corrigidos para isso e, portanto, não podem ser interpretados como testes da hipótese de que as médias dos *clusters* são iguais.

NÚMERO DE CASOS EM CADA *CLUSTER*

Cluster 1	6.000
2	6.000
3	8.000
Válidos	20,000
Faltantes	0,000

(*continua*)

Utilizou-se a análise de *cluster* para classificar os clientes que selecionam hospitais para internamento como antiquados, afluentes, conscientes do valor, ou que acham que têm direito a tudo.

dos entrevistados agrupados para saber se os segmentos poderiam ser s eficientemente identificados.

Utilizou-se o Quick Cluster (SPSS), um método de aglomeração de variância mínima, para agrupar os entrevistados com base em suas respostas aos itens da preferência por um hospital. Minimizaram-se os quadrados das distâncias euclidianas entre todas as variáveis do *cluster*. Como indivíduos distintos têm percepções diferentes de escalas de importância, as classificações atribuídas por cada indivíduo foram padronizadas antes da aglomeração. Os resultados indicaram que a melhor classificação dos entrevistados seria em quatro *clusters*. Aplicou-se duas vezes o procedimento de validação cruzada para análise de *clusters* sobre as duas metades da amostra total.

Como era de se esperar, os quatro grupos diferiram substancialmente por suas distribuições e respostas médias às razões para suas preferências hospitalares. Os nomes atribuídos aos quatro grupos refletiram as características demográficas e as razões pelas preferências hospitalares: Antiquados, Afluentes, Conscientes do Valor e Profissionais "Querem-tudo".[15] ■

[15] Alfred Lin, Leslie A. Lenert, Mark A. Hlatky, Kathryn M. McDonald, et al., "Clustering and the Design of Preference-Assessment Surveys in Healthcare," *Health Services Research,* 34 (5) (December 1999): 1033-1045; Edward J. Holohean, Jr., Steven M. Banks e Blair A. Maddy, "System Impact and Methodological Issues in the Development of an Empirical Typology of Psychiatric Hospital Residents," *Journal of Mental Health Administration,* 22 (2) (Spring 1995): 177-188; e Arch G. Woodside, Robert L. Nielsen, Fred Walters e Gale D. Muller, "Preference Segmentation of Health Care Services: The Old-Fashioneds, Value Conscious, Affluents, and Professional Want-It-Alls," *Journal of Health Care Marketing* (June 1988): 14-24.

AGLOMERAÇÃO DE VARIÁVEIS

Por vezes, utiliza-se a análise de *clusters* para aglomerar variáveis com a finalidade de identificar grupos homogêneos. Nessa instância, as unidades usadas para análise são as variáveis, calculando-se as medidas de distância para todos os pares de variáveis. Por exemplo, o coeficiente de correlação – seja em valor absoluto ou com sinal – pode ser usado como medida de semelhança (o oposto da distância) entre variáveis.

A aglomeração hierárquica de variáveis pode auxiliar na identificação de variáveis isoladas, ou variáveis que dão uma contribuição única para os dados. O *cluster* também pode ser usado para reduzir o número de variáveis. Associada a cada *cluster*, existe uma combinação linear das variáveis no aglomerado, chamada de *componente do cluster*. Um grande conjunto de variáveis pode, com freqüência, ser substituído pelo conjunto de componentes do *cluster* sem grande perda de informação. Todavia, um determinado número de componentes de *cluster* em geral não explica tanto da variância quanto o mesmo número de componentes principais. Por que, então, utilizar a aglomeração de variáveis? As componentes de um *cluster* em geral são mais fáceis de interpretar do que as componentes principais, mesmo que estas estejam rotacionadas.[16] Ilustramos a aglomeração de variáveis com um exemplo da pesquisa de propaganda.

PESQUISA REAL

Nada mais que sentimentos

Em 2002, a Polaroid enfrentou-se repentinamente com uma concorrência crescente de câmeras digitais, portanto os executivos de marketing decidiram centrar-se no potencial emocional de sua linha principal de câmeras. Uma campanha de US$ 15 milhões veiculou o novo *slogan*, "Clique, instantaneamente" implicando que a câmera Polaroid pode alterar um sentimento "aqui e agora". A campanha de propaganda foi projetada para evocar sentimentos emotivos em consumidores quando vissem os novos comerciais da Polaroid.

A Polaroid baseou essa campanha em um estudo realizado para sensações provocadas pela propaganda. Um total de 655 sensações foi reduzido a um conjunto de 180 que os entrevistados consideraram mais prováveis de serem estimuladas pela propaganda. Esse grupo foi aglomerado com base em julgamentos de similaridade entre sensações, resultando em *clusters* de 31 sensações divididos em 16 *clusters* positivos e 15 negativos.[17]

Sensações positivas	*Sensações negativas*
1. Lúdica/infantil	1. Temerosa
2. Amistosa	2. Má/doente
3. Humorística	3. Confusa
4. Encantada	4. Indiferente
5. Interessada	5. Entediada
6. Forte/confiante	6. Triste
7. Calorosa/terna	7. Ansiosa
8. Relaxada	8. Incapaz/tímida
9. Enérgica/impulsiva	9. Feia/estúpida
10. Ansiosa/excitada	10. Piedosa/decepcionada
11. Contemplativa	11. Louca
12. Orgulhosa	12. Desagradável
13. Persuadida/expectante	13. Desgostosa
14. Vigorosa/desafiadora	14. Irritada
15. Admirada	15. Mal-humorada/frustrada
16. Definida/informada	

Assim, 655 reações à propaganda foram reduzidas a um núcleo de 31. Os anunciantes têm agora um conjunto manipulável de sensações para entender e medir as reações emocionais à propaganda. Quando medidas, essas sensações podem proporcionar informações sobre a capacidade de um anúncio de persuadir os consumidores-alvo, como no caso da câmera Polaroid. ■

A análise de *clusters*, especialmente a aglomeração de objetos, também costuma ser utilizada em pesquisa de marketing internacional (como no próximo exemplo) e pode ser aplicada também na pesquisa de avaliações éticas (como no exemplo posterior).

PESQUISA REAL

Percepção de equivalência de um produto – antes uma raridade, hoje uma realidade

Como os consumidores em diferentes países percebem as marcas em diferentes categorias de produtos? Surpreendentemente, a resposta é que a taxa de paridade de percepção de um produto é bastante alta. A paridade observada de um produto significa que os consumidores consideram todas/quase todas as marcas em uma categoria de produto como semelhantes entre si. Um estudo recente feito pelo BBDO Worldwide mostra que dois terços dos consumidores pesquisados em 28 países consideram equivalentes as marcas em 13 categorias de produtos desde companhias aéreas até cartões de crédito e café. A paridade observada foi, em média, de 63% para todas as categorias em todos os países. Os japoneses têm a mais alta percepção de paridade em todas as categorias de produtos, com 99%, e os colombianos têm a mais baixa percepção, com 28%. Encarados por categoria de produtos, os cartões de crédito detêm a mais alta percepção de paridade – 76%, e os cigarros, a mais baixa – 52%.

O BBDO aglomerou os países com base em percepções de paridade de produto, chegando a *clusters* que exibiam níveis e padrões de percepção de paridade semelhantes. O maior valor de percepção de paridade provém da região da Ásia/Pacífico (83%), que inclui Austrália, Japão, Malásia, Coréia do Sul e França. Não é de surpreender que a França figure nessa lista porque, para a maioria dos produtos, os franceses usam em larga escala uma propaganda eminentemente visual, orientada para as sensações. O *cluster* seguinte consiste em mercados influenciados pelos Estados Unidos (65%), que incluem Argentina, Canadá, Hong Kong, Kuwait, México, Cingapura e os próprios EUA. O terceiro *cluster*, em grande parte países europeus (60%), compreende Áustria, Bélgica, Dinamarca, Itália, Holanda, África do Sul, Espanha, Reino Unido e Alemanha.

O que tudo isso significa é que, para diferenciar um (a) produto/marca, a propaganda não pode se centrar apenas no desempenho do produto; deve também relacionar o produto com a vida da pessoa de uma forma importante. Além disso, exige-se um esforço de marketing muito maior na região da Ásia/Pacífico e na França para diferenciar as marcas da concorrência e estabelecer uma imagem única. Um fator relevante nessa paridade crescente é, sem dúvida, a emergência do mercado global. Um estudo realizado em 2001 explorou as questões subjacentes ao conteúdo informativo factual da propaganda nas condições de paridade de produto e tipo de produto. Os dados para essa pesquisa foram obtidos da análise de conteúdo de mais de 17.000 propagandas em jornais e 9.800 comerciais televisivos. A análise mostrou que as propagandas de produtos com baixa paridade contêm mais informações factuais que suas contrapartes. Quando ambas as condições foram vistas juntas, a paridade influenciou o conteúdo informativo factual, mas não em mesmo grau que o tipo de produto. O estudo revelou que, acima de tudo, quando se trata de incluir informações factuais nas propagandas, os anunciantes reagem mais ao tipo de produto que à paridade de produto.[18] ■

PESQUISA REAL

Aglomerando profissionais de marketing com base em avaliações éticas

Pode-se aplicar a análise de *clusters* para explicar diferenças em percepções éticas, utilizando-se uma grande escala multiitens e multidimensional elaborada para avaliar quão éticas são diferentes situações. Uma dessas escalas foi criada por Reidenbach e Robin. Essa escala tem 29 itens, compondo cinco dimensões que medem como um entrevistado julga determinada ação. Por exemplo, um entrevistado lê sobre um pesquisador de marketing que revelou dados reservados de um de seus clientes a um segundo cliente. O entrevistado deve então completar a escala ética de 29 itens, indicando, por exemplo, se essa ação é:

Justa	:_:_:_:_:_:_:_:	Injusta
Tradicionalmente aceitável	:_:_:_:_:_:_:_:	Inaceitável
Envolve violação de um contrato verbal	:_:_:_:_:_:_:_:	Não envolve violação de um contrato verbal

Essa escala poderia ser aplicada a uma amostra de profissionais de marketing. Aglomerando entrevistados com base nesses 29 itens, é preciso investigar duas questões importantes. Em primeiro lugar, como diferem os *clusters* em relação às cinco dimensões éticas: neste caso, Justiça, Relativismo, Egoísmo, Utilitarismo e Deontologia. Em segundo lugar, que tipos de firma compõem cada *cluster*? Os *clusters* podem ser descritos em termos de classificação por indústria (NAICS – North American Industry Classification System), tamanho da firma e lucratividade da firma. As respostas a essas duas questões devem permitir visualizar que tipos de firma utilizam quais dimensões para avaliar situações éticas. Por exemplo, as grandes e as pequenas firmas se enquadram em *clusters* diferentes? Situações questionáveis são mais aceitáveis por firmas mais lucrativas do que por firmas menos lucrativas? Um estudo empírico realizado em 2001 comparou as percepções taiwanesas e estadunidenses sobre a ética corporativa. Usou-se um questionário auto-aplicável que consistia em cinco medidas. Uma das medidas, valores morais individuais, foi avaliada usando a escala de Reidenbach e Robin. Os resultados mostraram que em ambas as culturas as percepções individuais de ética corporativa parecem determinar o comprometimento organizacional mais que os valores morais individuais.[19] ■

APLICAÇÕES NA INTERNET E EM COMPUTADORES

No SPSS, o programa principal para aglomeração de objetos ou casos é CLUSTER. Podem-se calcular diferentes medidas de distância, estando disponíveis todos os procedimentos de aglomeração hierárquica discutidos aqui. Para a aglomeração não-hierárquica, pode-se utilizar o programa QUICK CLUSTER. Esse programa é particularmente útil para aglomerar um grande número de casos. Todas as opções por padrão resultarão em um *cluster* de *k* médias. Para aglomerar variáveis, as medidas de distância devem ser calculadas por meio das variáveis utilizando-se o programa PROXIMITIES. Essa matriz de proximidade pode ser lida em CLUSTER para obtermos um agrupamento das variáveis.

No SAS, pode-se utilizar o programa CLUSTER para o *cluster* hierárquico de casos ou objetos. Estão disponíveis todos os procedimentos de aglomeração discutidos aqui, assim como alguns outros adicionais. Pode-se obter o *cluster* não-hierárquico de casos ou objetos recorrendo-se a FASTCLUS. Para a aglomeração de variáveis, utiliza-se o programa VARCLUS. Os dendrogramas não são calculados automaticamente, mas podem ser obtidos com a aplicação do programa TREE.

No MINITAB, a análise de *cluster* pode ser acessada com a função de observação Multivariate>Cluster. Dispõe-se também de *Cluster* de Variáveis e Aglomerado de K médias. A análise de *clusters* não pode ser acessada no Excel.

[16] Brian Everitt, Sabine Landau e Morven Leese, *Cluster Analysis,* 4th ed. (Oxford, UK: Oxford University Press. 2001); e Vicki Douglas. "Questionnaire Too Long? Try Variable Clustering," *Marketing News.* 29 (5) (February 27, 1995): 38.

[17] Aaron Baar, "Polaroid Ads Play Up Emotion," *Adweek,* 42 (15) (April 9, 2001): 2; Thorolf Helgesen, "The Power of Advertising Myths and Realities," *Marketing & Research Today*, 24 (2) (May 1996): 63-71; David A. Aaker, Douglas M. Stayman e Richard Ve/ina, "Identifying Feelings Elicited by Advertising," *Psychology & Marketing* (Spring 1988): 1-16.

[18] Gergory M. Pickett, "The Impact of Product Type and Parity on the Informational Content of Advertising," *Journal of Marketing Theory ami Practice,* 9 (3) (Summer 2001): 32-43; Fred Zandpour e Katrin R. Harich, "Think and Feel Country Clusters: A New Approach to International Advertising Standardization." *International Journal of Advertising*, 15 (4) (1996): 325-344; e Nancy Giges, "World's Product Parity Perception High," *Advertising Age* (June 20, 1988).

[19] John P. Fraedrich, Neil C. Hemdon, Jr. e Quey-Jen Yeh, "An Investigation of Moral Values and the Ethical Content of the Corporate Culture," *Journal of Business Ethics*, 30 (1) (March 2001): 73-85; Ishmael P. Akaah, "Organizational Culture and Ethical Research Behavior," *Journal of the Academy of Marketing Science,* 21 (1) (Winter 1993): 59-63; e R. E. Reidenhach e D. P. Robin, "Some Initial Steps Toward Improving the Measurement of Ethical Evaluations of Marketing Activities," *Journal of Business Ethics*, 1 (1988): 871-879.

SPSS Windows

Para selecionar esse procedimento usando o SPSS para Windows, clique em:

Analyze>Classify>Hierarchical Cluster...

Analyze>Classify>K-Means Cluster...

Para aglomerar as variáveis, use o *Cluster* Hierárquico e, a seguir, selecione a opção variáveis dentro de *cluster*.

FOCO NA BURKE

Quando a Burke apresenta a um cliente os resultados de uma análise de *cluster*, há três questões importantes a serem respondidas:

1. O que os *clusters* informam sobre o mercado?
2. Que variáveis comandam a aglomeração?
3. Quão diferentes são os *clusters*?

O que os clusters informam sobre o mercado?

Para responder a essa pergunta, a Burke utiliza tipicamente dados de entrevistados que não foram incluídos no procedimento de aglomeração. Por exemplo, é possível encontrar quatro *clusters* de entrevistados baseados em suas avaliações sobre os benefícios do produto. Também coletamos dados sobre o comportamento passado e as intenções de compra para formularmos um novo conceito. Se os *clusters* não apresentarem quaisquer diferenças administrativamente significativas sobre essas medidas de comportamento e de intenções, fica difícil justificar perante a administração a utilidade dos *clusters*. Como a finalidade do *cluster* é criar grupos que sejam o mais semelhantes possível, não há qualquer garantia de que eles apresentem diferenças quando submetidos a um valor externo. Não podemos simplesmente aceitar que os *clusters* sejam realmente diferentes pelo simples fato de terem sido criados.

Que variáveis comandam a aglomeração?

Se utilizarmos diversas variáveis para criar o *cluster*, devemos ter muito cuidado para não criar um sistema implícito de ponderação. Por exemplo, em um projeto para uma indústria de automóveis, foram propostas 20 questões sobre as características desejadas em um carro novo para uso na aglomeração de entrevistados. Ficou imediatamente claro que sete das questões se relacionavam direta ou indiretamente com a economia, oito delas se referiam à imagem, três diziam respeito a preço/valor e duas refletiam velocidade/aceleração. Naturalmente, uma análise fatorial ajudou a formular essas observações. Podia-se prever que as questões que causariam maior impacto no *cluster* levariam a administração a concluir que "os *clusters* de entrevistados parecem refletir fortemente a economia e a imagem". Seria de se admirar que isso não ocorresse. Quando calculamos a distância euclidiana para examinar diferenças entre entrevistados, a soma de quadrados inclui 15 de economia e imagem, com apenas 5 de preço/valor ou velocidade/aceleração. Se todos eles estiverem em escalas comparáveis, as duas últimas categorias terão pouca chance de superar a força das duas primeiras. Seria mais razoável reduzir o número de questões de modo a nos aproximarmos o máximo possível ao mesmo número refletindo grupos de questões fortemente correlacionados. Se não fizermos isso, o número de questões formuladas sobre um tópico pode eventualmente ter maior influência sobre nossos resultados do que o próprio tópico.

Quão diferentes são os clusters?

No primeiro tópico acima, discutiu-se a idéia de determinar se os *clusters* são úteis em razão de variáveis externas. Devemos também examinar se os *clusters* são efetivamente "diferentes" com base nas variáveis usadas para criá-los. Podemos colocar nossos dados em um algoritmo de aglomeração e, se determinarmos que ele deve parar em dois *clusters*, teremos dois *clusters* em razão da natureza do procedimento, e não necessariamente em conseqüência da lógica ou da estrutura que existe na população. Devemos atentar para os *clusters* e ver se as diferenças acusam magnitude e estabilidade que suscitem confiança.

1. É improvável que os *clusters* sejam diferentes em todas as questões que utilizamos como entrada para o procedimento de aglomeração. Embora os procedimentos estatísticos não sejam realmente válidos quando aplicados a *clusters* criados sistematicamente, eles proporcionam uma percepção de como se formam os *clusters*. A ANOVA de um fator nos informa se as questões individuais são diferentes ao longo dos *clusters* com base em uma estatística que seria apropriada para uma amostra probabilística (naturalmente, não é isto que temos... mas trata-se de um "índice" útil). A aplicação da análise discriminante tem ainda maior apelo, pois mostra quais das questões constituiriam discriminadores potenciais entre grupos que levam em conta a colinearidade entre esses previsores.
2. A significância gerencial é um problema diferente. Suponhamos que nossos *clusters* pareçam diferentes quanto a esses índices estatísticos (ANOVA e análise discriminante). Isso não significa que as diferenças sejam suficientemente grandes para serem consideradas úteis pela administração. Por exemplo, uma questão para a qual os clusters foram "diferentes" foi a classificação da economia e, no entanto, 90% das classificações ficaram entre 6 e 9 em uma escala de 10 pontos. Seria necessário ter mais evidências para podermos considerar esse ponto como uma diferença significativa para a administração. As classificações se apresentam diferentes em grau de "positividade", mas não diferentes pelo fato de umas serem altas e outras baixas. Trata-se de um problema difícil para o qual não há resposta pronta. Devemos dar-nos por satisfeitos pelo fato de podermos interpretar essas diferenças numéricas em termos de decisões administrativas significativas.

RESUMO

Utiliza-se a análise de *clusters* para classificar objetos ou casos e, eventualmente, variáveis em grupos relativamente homogêneos. Os grupos ou *clusters* são sugeridos pelos dados, e não definidos *a priori*.

As variáveis sobre as quais se faz o *cluster* devem ser selecionadas com base em pesquisa passada, na teoria, nas hipóteses que estão sendo testadas ou no julgamento do pesquisador. Deve-se escolher uma medida adequada de distância ou de semelhança. A medida mais comumente usada é a distância euclidiana ou o seu quadrado.

Os procedimentos de aglomeração podem ser hierárquicos ou não-hierárquicos. O *cluster* hierárquico se caracteriza pelo desenvolvimento de uma hierarquia ou de uma estrutura em forma de árvore. Os métodos hierárquicos podem ser aglomerativos ou divisivos. Os métodos aglomerativos dividem-se em métodos de encadeamento (*linkage methods*), métodos de variância e métodos centróide. Os métodos de encadeamento compreendem o encadeamento único, o encadeamento completo e o encadeamento médio. Um método de variância bastante usado é o método Ward. Os métodos não-hierárquicos são chamados freqüentemente de aglomeração de *k* médias. Esses métodos podem ser classificados como limiar seqüencial, limiar paralelo e particionamento otimizador. Os métodos hierárquicos e não-hierárquicos podem ser utilizados em seqüência. A escolha de um procedimento de aglomeração e a escolha de uma medida de distância estão inter-relacionadas.

O número de *clusters* pode basear-se em considerações teóricas, conceituais ou práticas. Na aglomeração hierárquica, as distâncias às quais se combinam os aglomerados é um critério importante. Os tamanhos relativos dos *clusters* devem ser significativos. Os *clusters* devem ser interpretados em termos de centróides de aglomeração. Em geral convém perfilar os *clusters* em termos de variáveis que não tenham sido usadas para formar o *cluster*. A confiabilidade e a validade das soluções de *clusters* podem ser avaliadas de diferentes maneiras.

TERMOS E CONCEITOS FUNDAMENTAIS

esquema de aglomeração, 574
centróide de *clusters*, 574
centros de *clusters*, 574
associação a um *cluster*, 574
dendrograma, 574
distâncias entre centros de *clusters*, 574
diagrama em sincelos (*icicle*), 574
matriz de coeficientes de semelhança/distância, 574
distância euclidiana, 576
aglomeração hierárquica, 576
cluster aglomerativo, 576
cluster divisivo, 576
métodos de encadeamento (*linkage methods*), 577
encadeamento único, 577
encadeamento completo, 577
encadeamento médio, 577
método da variância, 577
método Ward, 578
método centróide, 578
cluster não-hierárquico, 578
método do limiar seqüencial, 578
método do limiar paralelo, 578
método do particionamento otimizador, 578

EXERCÍCIOS

Perguntas

1. Discuta a semelhança e a diferença entre análise de *cluster* e análise discriminante.
2. Cite algumas aplicações da análise de *cluster* em marketing.
3. Defina sucintamente os seguintes termos: dendrograma, gráfico em sincelos (*icicle*), esquema de aglomeração e associação a um *cluster*.
4. Qual é a medida de semelhança mais comumente usada em análise de *clusters*?
5. Apresente uma classificação dos procedimentos de aglomeração.
6. Por que se prefere em geral o método do encadeamento médio aos métodos de encadeamento único e encadeamento completo?
7. Quais são as duas principais desvantagens dos procedimentos de aglomeração não-hierárquicos?
8. Quais são as diretrizes existentes para decidir quanto ao número de *clusters*?
9. O que está envolvido na interpretação de *clusters*?
10. Cite algumas variáveis adicionais usadas para o traçar o perfil de *clusters*.
11. Descreva alguns procedimentos para avaliar a qualidade de soluções de *clusters*.
12. Como se utiliza a análise de *clusters* para agrupar variáveis?

Problemas

1. As afirmações a seguir são verdadeiras ou falsas?
   a. Os métodos hierárquicos e não-hierárquicos de aglomeração sempre dão resultados diferentes.
   b. Devemos sempre padronizar os dados antes de fazermos análise de *cluster*.
   c. Pequenas distâncias entre coeficientes no esquema de aglomeração implicam que estamos mesclando casos diferentes.
   d. Não importa a medida de distância utilizada; as soluções de *clusters* são sempre as mesmas.
   e. É aconselhável analisar o mesmo conjunto de dados utilizando diferentes procedimentos de aglomeração.

EXERCÍCIOS PARA INTERNET E COMPUTADORES

1. Analise os dados da Tabela 20.1 utilizando os seguintes métodos: (a) encadeamento único; (b) encadeamento completo; e (c) método centróide. Utilize SPSS, SAS, ou MINITAB. Compare seus resultados com os apresentados na Tabela 20.2.
2. Realize a análise seguinte sobre os dados relativos à Nike apresentados nos exercícios para Internet e computadores 1 do Capítulo 15. Considere apenas as seguintes variáveis: consciência, atitude, preferência, intenção e lealdade à Nike.
   a. Aglomere os respondentes com base nas variáveis identificadas usando a aglomeração hierárquica. Use o método Ward e o quadrado das distâncias euclidianas. Quantos clusters você recomenda e por quê?
   b. Aglomere os respondentes com base nas variáveis identificadas usando a aglomeração de *k* médias e o número de *clusters* identificados em (a). Compare os resultados com aqueles obtidos em (a).
3. Realize a seguinte análise sobre os dados sobre estilo de vida fora de casa apresentados nos exercícios para Internet e computadores 2 do Capítulo 15. Considere apenas as seguintes variáveis: a importância conferida a desfrutar da natureza, relacionar-se com o clima, viver em harmonia com o meio ambiente, exercitar-se regularmente e encontrar-se com outras pessoas (V_2 a V_6).
   a. Aglomere os respondentes com base nas variáveis identificadas usando a aglomeração hierárquica. Use o método Ward e o quadrado das distâncias euclidianas. Quantos *clusters* você recomenda e por quê?
   b. Aglomere os respondentes com base nas variáveis identificadas usando os seguintes métodos hierárquicos: (a) encadeamento único (vizinho mais próximo), (b) encadeamento completo (vizinho mais afastado) e (c) método centróide.
   c. Aglomere os respondentes com base nas variáveis identificadas usando a aglomeração de *k* médias e o número de *clusters* identificados em (a). Compare os resultados com aqueles obtidos em (a).
4. Realize a seguinte análise dos dados sobre calçados esportivos apresentados nos exercícios para Internet e computadores 3 no Capítulo 17. Considere apenas as seguintes variáveis: avaliações dos tênis quanto a conforto (V_2), estilo (V_3) e durabilidade (V_4).
   a. Aglomere os respondentes com base nas variáveis identificadas usando a aglomeração hierárquica. Use o método Ward e o quadrado das distâncias euclidianas. Quantos *clusters* você recomenda e por quê?
   b. Aglomere os respondentes com base nas variáveis identificadas usando a aglomeração de *k* médias e o número de *clusters* identificados em (a). Compare os resultados com aqueles obtidos em (a).
5. Analise os dados coletados no exercício Trabalho de Campo para aglomerar os entrevistados usando os métodos hierárquicos e não-hierárquicos. Use um dos pacotes de *software* discutidos neste capítulo.
6. Analise os dados coletados no exercício Trabalho de Campo para aglomerar as 15 variáveis que medem a atitude dos consumidores quanto a companhias aéreas e vôos. Use um dos programas descritos neste capítulo.

ATIVIDADES

Trabalho de campo

1. Como um consultor de pesquisa de marketing de uma grande companhia aérea, você deve determinar as atitudes dos consumidores em relação a companhias aéreas e aos vôos. Construa uma escala de 15 itens para esse fim. Em um grupo de cinco alunos, obtenha dados para essa escala e para características demográficas padrão de 50 chefes de família, homens ou mulheres, em sua comunidade. Cada aluno deve realizar 10 entrevistas. Esses dados serão utilizados para aglomerar os entrevistados e para aglomerar as 15 variáveis que medem as atitudes dos consumidores quanto a companhias aéreas e vôos.

Discussão em grupo

2. Em um pequeno grupo, discuta o papel da análise de *clusters* ao analisar dados de pesquisa de marketing. Destaque as maneiras em que a análise de *clusters* pode ser usada junto com outros procedimentos de análise de dados.