TRATAMENTO GRÁFICO E MATEMÁTICO-ESTATÍSTICO DE DADOS MULTIVARIADOS: ANÁLISES DE AGRUPAMENTOS A PARTIR DAS TÉCNICAS MATRIZ ORDENÁVEL, COMPONENTES PRINCIPAIS, WARD E K-MÉDIAS

Márcia M. Duarte dos Santos(*), Miguel C. Sanchez (**) & Sueli Ap. Mingoti (***)

Abstract

In this paper some methods commonly used in multivariate cluster analysis are discussed and compared by using a specific data set. The main objective is to show that the graphical method is efficient and similar to the Ward and K-Means methods which are based upon mathematical and statistical theories. Thesimilarity found between the graphical and the statistical methods suggests that although it is more subjective the graphical method is a valid technique that could be used to determine the clusters of a sample or a population.,

INTRODUÇÃO

No processo de tratamento matemático-estatístico de um conjunto de dados, diagramas, matrizes, tabelas, quadros e mapas são comumente empregados na fase de organização dos dados. A partir dessas mensagens, medidas e testes são efetuados para a consecução dos objetivos de análise. Nesse processo, recorre-se novamente às mensagens citadas para se comunicar, desta feita, os resultados da análise e os de sua interpretação.

Os dados apresentados em diagramas, matrizes, tabelas, quadros e mapas, em razão de estarem organizados nessas mensagens, podem revelar, em algumas situações, informações referentes aos objetivos de análise, apreendidas pelo pesquisador sem o auxílio de uma análise matemática e estatística. Citam-se, como exemplo, a correlação de duas variáveis visualmente apresentada num diagrama de dispersão, a classificação de dados de uma variável representada num diagrama triangular, a distribuição espacial ou a dinâmica espaço-temporal de um fato, fenômeno ou evento transcrito cartograficamente, dentre outros.

Verifica-se, de outras vezes, que as informações de interesse do pesquisador só podem ser patenteadas sem o auxílio de medidas e testes estatístico-matemáticos, após transcrições dos dados em mensagens diferentes das que serviram inicialmente para sua organização ou após reconstruções sucessivas dessas mensagens. Nesse caso, assim como no anterior, pode-se falar em tratamento gráfico de dados, o qual compreende princípios e regras que definem técnicas e procedimentos para sua aplicação.

Algumas das técnicas de tratamento gráfico são particularmente indicadas para a análise de dados multivariados e têm sido muito utilizadas. Observa-se que a atração por essas técnicas deve-se, de um lado, à relativa simplicidade teórica do método gráfico de tratamento de dados, e, de outro lado, à afinidade que muitos usuários têm em relação à representação gráfica. Nota-se, também, que a tendência de emprego dessas técnicas é de crescimento, considerando-se as facilidades para sua operacionalização com a disseminação de aplicativos computacionais gráficos.

Entre as técnicas gráficas de tratamento de dados multivariados, cita-se a Matriz Ordenável como uma das mais utilizadas, tendo em vista a análise de conglomerados. Essa técnica é descrita por Santos & Sanches (1996) num trabalho que objetiva, a partir de um exemplo de aplicação, esclarecer a natureza do método gráfico e discutir aspectos relacionados ao interesse e à propriedade de utilizá-lo, num contexto científico de tratamento de dados.

Os resultados do trabalho de Santos & Sanches ensejaram uma oportunidade para ampliar as discussões a propósito do método gráfico de tratamento de dados. Nesse sentido, os dados tratados pelos autores citados foram submetidos à análises matemático-estatísticas de agrupamento, muito difundidas e testadas, visando à discussão desses resultados, assim como o obtido por intermédio do método gráfico. Desse modo, pretende-se alcançar implicações a propósito da eficácia do tratamento gráfico de dados para se fundamentar sua escolha, tendo em vista um elenco de técnicas passíveis de serem utilizadas, mas de outra natureza.

OS DADOS ESTUDADOS E AS TÉCNICAS MATEMÁTICO-ESTATÍSTICAS EMPREGADAS

Os dados multivariados estudados referem-se ao desempenho dos estudantes que prestaram o Vestibular de 1994 da UFMG. Esse desempenho, de acordo com os dados apresentados no Quadro 1, é expresso através

(*) (Depto. de Geografia, IGC, UFMG)
(**) (Depto. de Estatística, ICEX, UFMG)
(***) (Depto. de Planejamento, IGCE, UNESP)

CURSO, NA PRIMEIRA ETAPA.

Nº	CURSOS	PROVAS							
		Port.	Mat.	Geog.	Hist.	L.Est.	Fís.	Quím.	Biol.
1	Administração D.	100	92	80	78	98	78	52	74
2	Administração N.	96	96	74	66	92	78	58	64
3	Arquitetura	100	95	85	67,5	97,5	92,5	53,7	76,2
4	Belas Artes (L/B)	88,3	23,3	56,6	35	70	11,6	8,8	31,6
5	Biblioteconomia	91,4	20	40	32,8	45,7	4,2	4,2	14,2
6	C. da Computação	100	100	92,8	87,1	100	100	68,5	94,2
7	C.Biológ. D.(L/B)	100	76,2	88,7	53,7	93,7	41,2	50	95
8	C. Biológ. N.(L)	95	52,5	75	55	62,5	27,5	35	77,5
9	C. Contábeis	98,7	65	58,7	57,5	86,2	50	18,7	30
10	C. Econômicas	98,7	73,7	72,5	53,7	93,7	66,2	32,5	61,2
11	C. Sociais (L/B)	98,4	27,6	61,5	70,7	76,9	21,5	16,9	35,3
12	Comunicação Social	100	91,6	100	91,6	100	70	55	95
13	Direito	100	86,6	91,3	85,6	98	77,3	49,3	81,3
14	Ed. Física (L/B)	99	23	31	22	74	28,3	19	40
15	Enfermagem (L/B)	97,5	42,5	61,2	26,2	86,2	8,7	16,2	70
16	Eng. Civil	97,5	89,5	64	47	92,5	85,5	50,5	57,5
17	Eng. Elétrica	95	100	81,2	66,2	93,7	100	70	77,5
18	Eng. Mecânica	98,7	98,7	78,7	52,5	95	98,7	66,2	81,2
19	Eng. Metalúrgica	96	82	88	68	84	76	52	64
20	Eng. de Minas	95	75	75	32,5	62,5	45	25	42,5
21	Eng. Química	98	92	64	60	92	84	70	78
22	Estatística	93,3	83,3	63,3	46,6	96,6	23,3	23,3	53,3
23	Farmácia	100	51,6	52,5	37,5	90	60,8	56,6	81,6
24	Filosofia (L/B)	95	25	77,5	62,5	72,5	15	12,5	40
25	Física D. (B)	100	86,6	90	33,5	86,6	86,6	56,6	53,3
26	Física N. (L)	96,6	76,6	66,6	43,3	60	80	16,6	43,3
27	Fisioterapia	100	70	75	42,5	92,5	62,5	35	85
28	Geografia D.(L/B)	77,5	25	70	62,5	45	5	10	30
29	Geografia N.(L)	90	16,6	90	63,3	86,6	16,6	6,6	26,6
30	Geologia	63,3	60	76,6	46,6	83,3	33,3	26,6	33,3
31	História D. (L/B)	100	50	82,5	82,5	80	25	15	42,5
32	História N. (L/B)	95	30	97,5	82,5	77,5	25	12,5	40
33	Letras (L)	96,2	20,8	55,8	40,8	90,4	4,5	8,7	30,8
34	Matemática D.(L)	96,6	66,6	26,6	16,6	66,6	46,6	16,6	36,6
35	Matemática D.(B)	95	80	65	35	60	60	20	30
36	Matemática N.(L)	96,6	86,6	56,6	40	73,3	90	26,6	33,3
37	Medicina	100	99,3	99,6	90	100	95,3	96,5	100
38	Med. Veterinária	99,1	76,6	90	65,8	94,1	58,3	54,1	94,1
39	Música (L/B)	71,8	21,8	31,2	37,5	46,8	12,5	6,2	12,5
40	Odontologia	100	100	95	74,1	100	84,1	74,1	99,1
41	Pedagogia D.(L)	95	20	58,3	38,3	70	6,6	5	31,6
42	Pedagogia N.(L/B)	95	15	58,3	30	55	3,3	5	33,3
43	Psicologia (L/B)	100	61,6	83,3	63,3	94,1	24,1	26,6	77,5
44	Química D.(L/B)	95	62,5	87,5	35	55	32,5	52,5	52,5
45	Química N.(L)	96,6	60	80	56,6	60	46,6	66,6	56,6
46	Terap. Ocupacional	97,5	55	62,5	37,5	77,5	17,5	17,5	67,5
	UFMG	94,3	63,1	72	53,7	80,6	49,1	35,2	57,1

Quadro 1: Percentagem de aprovados no Vestibular de 1994 da UFMG que obtiveram metade ou mais de pontos, por prova e por curso, na primeira etapa.

Chart 1: Percentage of aproved applicant for the 94 selection in UFMG that had more than half of the points, by test and course into the first stage.

da percentagem de aprovados que obtiveram a metade ou mais pontos em cada uma das oito provas da primeira etapa do concurso vestibular, segundo os quarenta e seis cursos oferecidos pela Universidade. O conjunto de dados estudados compreende, então, oito variáveis quantitativas e uma população de 46 cursos.

Tendo em vista o objetivo que norteou o tratamento desses dados, estes foram submetidos à aplicação de métodos matemáticos de análise de agrupamentos que mais se utilizam de critérios estatísticos, ou seja, os métodos de Ward e K-Médias (Johnson & Wickern, 1992). A escolha desses métodos é justificada, também, pelo fato de corresponderem a procedimentos que operam segundo diferentes esquemas de formação de conglomerados. Esses esquemas classificados como aglomerativos hierárquicos e não hierárquicos são representados, respectivamente, pelos métodos de Ward e K-Médias.

As análises de agrupamentos de Ward e K-Médias foram escolhidas, ainda, por serem consideradas as mais apropriadas para o estudo de variáveis quantitativas, sob o ponto de vista estatístico. Nessas análises, os percentuais de aprovados registrados para as oito provas são utilizados simultaneamente na comparação dos cursos feita matematicamente. A cada passo do processo de análise, os cursos cujos alunos apresentam desempenhos similares vão sendo agrupados, de acordo com uma medida de distância, ou seja, empregando-se um critério matemático de similaridade (Johnson & Wickern, 1992). No caso, dos dados tratados neste artigo, a medida empregada é a distância Euclidiana simples (Basilevsky, 1983). Brevemente, o método de análise de conglomerados (Anderberg, 1973) pode ser descritoda seguinte forma: Seja $X = (X_1 , X_2 , ..., X_p)$, o conjunto contendo as p variáveis medidas em cada um dos n elementos amostrais, aqui representados por: $(E_1 , E_2 ,..., E_n)$. Com base no conjunto X os n elementos amostrais são agrupados em grupos g_i de modo que :

(i) se $E_l \; e \; E_k \in g_i \Rightarrow E_l \; e \; E_k$ são semelhantes ;

(ii) se $E_l \; e \; E_k \notin g_i \Rightarrow E_l \; e \; E_k$ são distintos.

No método hierárquico de Ward, a cada passo dois conglomerados se juntam formando um novo conglomerado, sendo que conglomerados agrupados em um determinado passo qualquer, permanencem juntos até o estágio final do procedimento. O método não-hierárquico das K-Médias, por sua vez, permite que conglomerados que se uniram num determinado passo do procedimento venham a se separar ao longo da execução do algoritmo de aplicação do método. Deste modo, o método das K-Médias proporciona uma maior flexibilidade ao pesquisador. Ambos, entretanto, estão baseados no princípio estatístico da análise de variância.

O objetivo dessas análises de conglomerados norteou, também, a aplicação de um procedimento que associou uma análise de componentes principais a uma de agrupamento, o método de Ward. A escolha desse procedimento propiciou o estudo dos dados através de uma variável alternativa que sintetiza para cada curso, a informação conjunta dos percentuais de estudantes nas oito provas. Essa variável, chamada de componente principal, corresponde ao desempenho global dos estudantes de cada curso. De uma naneira suscinta, a técnica de Análise de Componentes Principais (Dillon & Goldstein, 1984) pode ser descrita da seguinte forma: Seja $X = (X_1 , X_2 , ..., X_p)$ o vetor aleatório contendo as p variáveis de interesse observadas em n elementos amostrais. As componentes principais denotadas por Y_j, j = 1,2,..., p , são definidas como:

$$Y_j = \sum_{k=1}^{p} c_{jk} \; X_k \qquad c_j$$

Onde os coeficientes do vetor $c_j = (c_{j1} , c_{j2} , ..., c_{jp})$ devem ser escolhidos de forma a satisfazer as seguintes condições :

(i) Var (Y_j) é máxima ;

(ii) Cov (Y_j , Y_l) = 0 , qualquer l $\neq$ j ;

(iii) $c_j' \; c_j = 1$

onde, Var(.) denota variância e Cov(.,.) denota covariância.

A solução do modelo descrito envolve na realidade o conhecimento dos Autovalores e Autovetores da matriz de covariância (ou de correlação) do vetor aleatório X. Pode ser mostrado que o vetor c_j que satisfaz as condições (i)-(iii) é simplesmente o Autovetor normalizado (ou seja de comprimento igual a 1) correspondente ao Autovalor λ_j , da matriz de covariâncias (ou correlação) do vetor X , j = 1, 2, ..., p.

Neste caso, as seguintes afirmações são válidas :

(a) A variância da j-ésima componente principal é igual a λ_j ;

b) $V.T.(X) = \sum_{i=1}^{p} Var(X_i) = \sum_{j=1}^{p} Var(Y_j) = V.T.(Y)$

onde V.T(.) denota variância total, e $Y = (Y_1 , Y_2 , ..., Y_p)$ é o vetor contendo as p-componentes principais. Pode ser mostrado que :

(c) $V.T.(X) = \sum_{j=1}^{p} \lambda_j$;

(d) Corr (Y_j , Y_l) = 0 , qualquer $j \neq l$, onde Corr(.) denota o coeficiente de correlação linear de Pearson.

Deste modo, temos que a estrutura de variância e covariância das p variáveis aleatórias do vetor aleatório X é explicada por um conjunto de p combinações lineares destas variáveis, combinações estas que são não correlacionadas e portanto, mais fáceis de serem entendidas conjuntamente. Na prática a matriz de covariâncias (ou correlação) do vetor X é estimada pela matriz de covariâncias amostral (ou correlação amostral), da qual os Autovalores e Autovetores normalizados correspondentes são obtidos para a construção das p componentes principais.

Para os dados analisados neste trabalho, a matriz de correlação estudada foi aquela referente as correlações das porcentagens de estudantes das 8 provas consideradas no vestibular. A primeira componente principal aqui denotada por Y, representou o desempenho global dos candidatos dos 46 cursos da UFMG nas várias provas. O valor desta componente principal, Y, para cada curso é obtido pela combinação linear das percentagens de estudantes que alcançaram notas acima ou igual a metade dos pontos em cada uma das oito provas da primeira etapa do vestibular. Essas percentagens são padronizadas pela média e pelo desvio padrão que, por sua vez, são calculados, para cada uma da oito provas, usando-se as respectivas percentagens observadas nos quarenta e seis cursos da UFMG. Esse procedimento é demonstrado a seguir:

$$Y = 0{,}252293(NP^*) + 0{,}388441(NM^*) + 0.326485(NG^*) + 0{,}298169(NH^*) + 0{,}359239(NLE^*) + 0{,}375818(NF^*) + 0{,}401620(NQ^*) + 0{,}397963(NB^*)$$

onde , NP*, NM*... NB* representam respectivamente, as percentagens de estudantes em Português, Matemática, ..., Biologia, padronizadas pela média e desvio padrão.

As variáveis são padronizadas da seguinte forma: considerando-se a prova de Português (NP) como exemplo, NP* = [(NP - média (NP)) / desvio padrão (NP)], onde a média de NP e o desvio padrão de NP são calculados usando-se as 46 percentagens de Português observadas.

Conhecidos assim os valores da componente principal Y para cada um dos cursos da UFMG, procedeu-se ao agrupamento dos mesmos recorrendo-se ao método Ward. Nota-se que o valor da componente principal definida para cada um dos cursos da UFMG indica, se for positivo, que a percentagem de aprovados no curso tem, globalmente, a tendência de se situar acima da percentagem média da UFMG em cada prova. Ao contrário, se o valor da componente for negativo mostra a tendência dos aprovados de se posicionarem abaixo da percentagem média geral de cada prova. No Quadro 2, são apresentados os valores calculados da componente Y, ordenados decrescentemente, onde se destaca os valores positivos dos negativos. Na aplicação dos métodos de agrupamento, a decisão sobre o número final de grupos ou conglomerados a serem formados, ou seja, sobre o momento de se interromper o processo de agrupamento, foi definida pelos resultados da análise efetuada através da Matriz Ordenável. Nesse estudo, essa decisão foi tomada ao se obter seis grupos, considerando-se que, de um lado, a qualidade da partição decresceria bastante se fosse observado um nível de fusão menor; e, de outro lado, que esse número proporcionaria a classificação da população, segundo as categorias mais comumente utilizadas na consideração de desempenho estudantil, quais sejam, ótimo, muito bom, bom, razoável, fraco e muito fraco.

CURSOS		*VALOR DA COMPONENTE*
Nº	*Denominação*	*PRINCIPAL Y*
37	*Medicina*	*4,39846*
6	*Computação*	*3,73163*
40	*Odontologia*	*3,55751*
12	*Comunicação Social*	*3,24979*
13	*Direito*	*2,65819*
17	*Engenharia Elétrica*	*2,64607*
3	*Arquitetura*	*2,54005*
18	*Eng.Mecânica*	*2,50953*
1	*Administração Diurno*	*2,34722*
38	*Veterinária*	*2,14045*
21	*Engenharia Química*	*2,01821*
2	*Administração Noturno*	*1,77881*
7	*C. Biológicas Diurno*	*1,69247*
19	*Engenharia Metalúrgica*	*1,57182*
25	*Física Diurno*	*1,36536*
16	*Engenharia Civil*	*1,13785*
27	*Fisioterapia*	*0,99923*
10	*Ciências Econômicas*	*0,77698*
43	*Psicologia*	*0,67687*
23	*Farmácia*	*0,49950*
45	*Química Noturno*	*0,22875*
31	História diurno	-0,25608
22	Estatística	-0,27687
36	Matemática Noturno	-0,33282
8	C.Biológicas Noturno	-0,40906
44	Química Diurno	-0,56791
32	História Noturno	-0,56836
9	Ciências Contábeis	-0,6023
26	Física Noturno	-0,65411
20	Engenharia de Minas	-0,97386
46	Terapia Ocupacional	-1,02710
35	Matemática Diurno	-1,21132
15	Enfermagem	-1,28924
11	Ciências Sociais	-1,37014
24	Filosofia	-1,53215
29	Geografia Noturno	-1,56914
30	Geologia	-1,83493
33	Letras	-2,21950
34	Matemática Diurno	-2,27848
14	Educação Física	-2,58463
41	Pedagogia Diurno	-2,73374
4	Belas Artes	-2,88004
28	Geografia Diurno	-3,19722
42	Pedagogia Noturno	-3,26960
5	Biblioteconomia	-4,12709
39	Música	-4,74115

Quadro 2: Valor da Componente Principal Y, segundo os Cursos da UFMG.

Chart 2: Main Component Y for each course of UFMG.

A propósito do procedimento de análise empreendido através da técnica Matriz Ordenável, observa-se que a formação de conglomerados é norteada pelo mesmo princípio que fundamenta as operações de agrupamento dos métodos matemático-estatísticos. Esse princípio diz respeito à constituição de grupos os mais heterogêneos possível, mas cujos elementos sejam

similares. No entanto, para a consecução desse propósito, no âmbito de um sistema gráfico de análise, os dados a serem estudados são transcritos numa imagem e, posteriormente, os elementos dessa imagem são rearranjados, a partir da percepção visual das semelhanças e das diferenças entre eles.

A imagem, a partir da qual se inicia a análise, tal como a mostrada na Figura 1, é construída observando-se regras que determinam o número de diagramas que a compõem. Essas regras definem, também, o tipo, a escala, assim como, a disposição dos diagramas no plano de representação, sempre tendo em vista a modalidade de análise empreendida. (Santos & Sanches, 1996).

A análise, por sua vez, guiada pela percepção visual, consiste na comparação do padrão gráfico apresentado pelos elementos da imagem, ou seja, de suas colunas e barras que representam, respectivamente, as variáveis e a população. A partir desse procedimento, os elementos da imagem são permutados, procurando-se justapor os semelhantes até se chegar a uma ordenação que se considera ideal.

A imagem ideal dos dados, num contexto de uma operação de classificação a partir de eixos ortogonais, corresponde à melhor configuração da forma que se delineia no plano de representação, a saber, a de uma diagonal ou a de uma paralela a uma das coordenadas do plano. No caso dos dados estudados, a imagem ideal tende à realização de uma diagonal, considerando os padrões de tonalidades, claras e escuras, definidos no plano de representação, como se mostra na Figura 2.

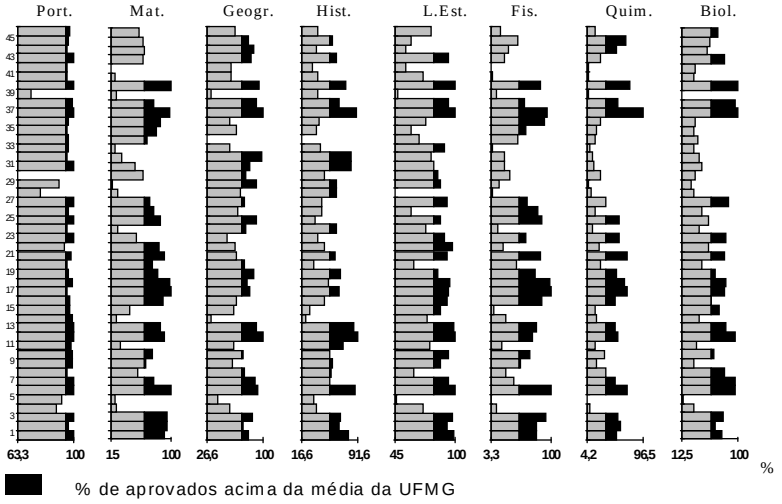

Figura 1: Matriz de Tratamento não manipulada.
Transição Gráfica dos dados apresentados no quadro 1.

Figure 1: Working Matrix without manipulation.
Graphic Transition of date showed in Chart 1.

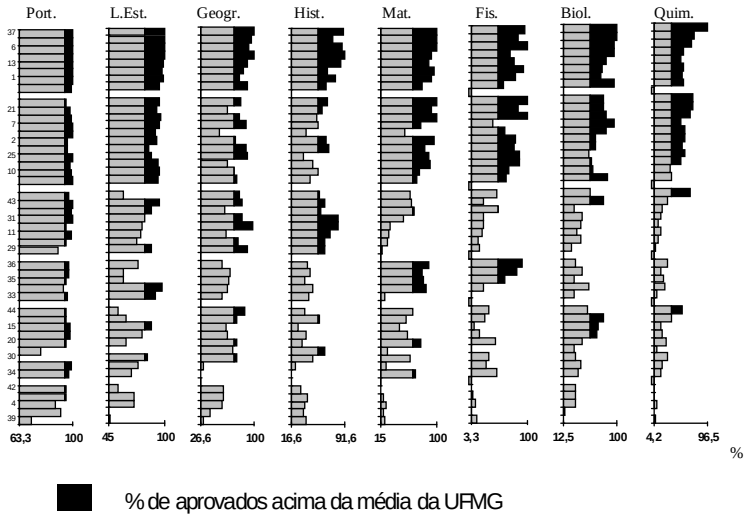

Figura 2: Matriz de Tratamento Manipulada.

Figure 2: Manipulated Working Matrix.

Observa-se que a imagem ideal define, até certo ponto, a qualidade do processo de classificação. Se a forma que os dados permitem criar encontra-se bem delineada no plano de representação, em função das permutações, é provável que se obtenha conglomerados cujos elementos sejam os mais semelhantes possível. É importante ressaltar, entretanto, o fato de que o grau de similaridade interna dos grupos relaciona-se, também, ou com o número de conglomerados que se considera desejável atingir, tendo em vista objetivos específicos, ou com o número de conglomerados a partir do qual as diferenças entre os grupos tornam-se menores, o que conseqüentemente acresce a heterogeneidade interna dos grupos.

Nota-se que o processo de análise de agrupamento empreendido através da Matriz Ordenável se assemelha à princípio aos das técnicas aglomerativas não hierárquicas; pois, parte-se do princípio de que se tem *n* conglomerados, onde *n* corresponde ao tamanho da população, e de que, em qualquer estágio do processo aglomerativo, os elementos unidos anteriormente podem ser separados para compor um outro grupo. A diferença reside no fato de que o método das K-Médias tem a ele associado uma medida matemático-estatística para comparar a similaridade dos vários cursos, enquanto que na Matriz Ordenável a comparação é feita segundo a percepção visual do pesquisador.

OS RESULTADOS ALCANÇADOS ATRAVÉS DAS ANÁLISES DE AGRUPAMENTOS: DISCUSSÕES

Os grupos definidos através das análises matemático-estatísticas e gráfica são apresentados na Figura 3, cujas linhas foram ordenadas pelos resultados alcançados com a aplicação da Matriz Ordenável. Esses grupos estão representados graficamente e são nomeados de acordo com a interpretação dada ao desempenho dos estudantes, nas diversas provas da primeira etapa do concurso Vestibular da UFMG de 1994.

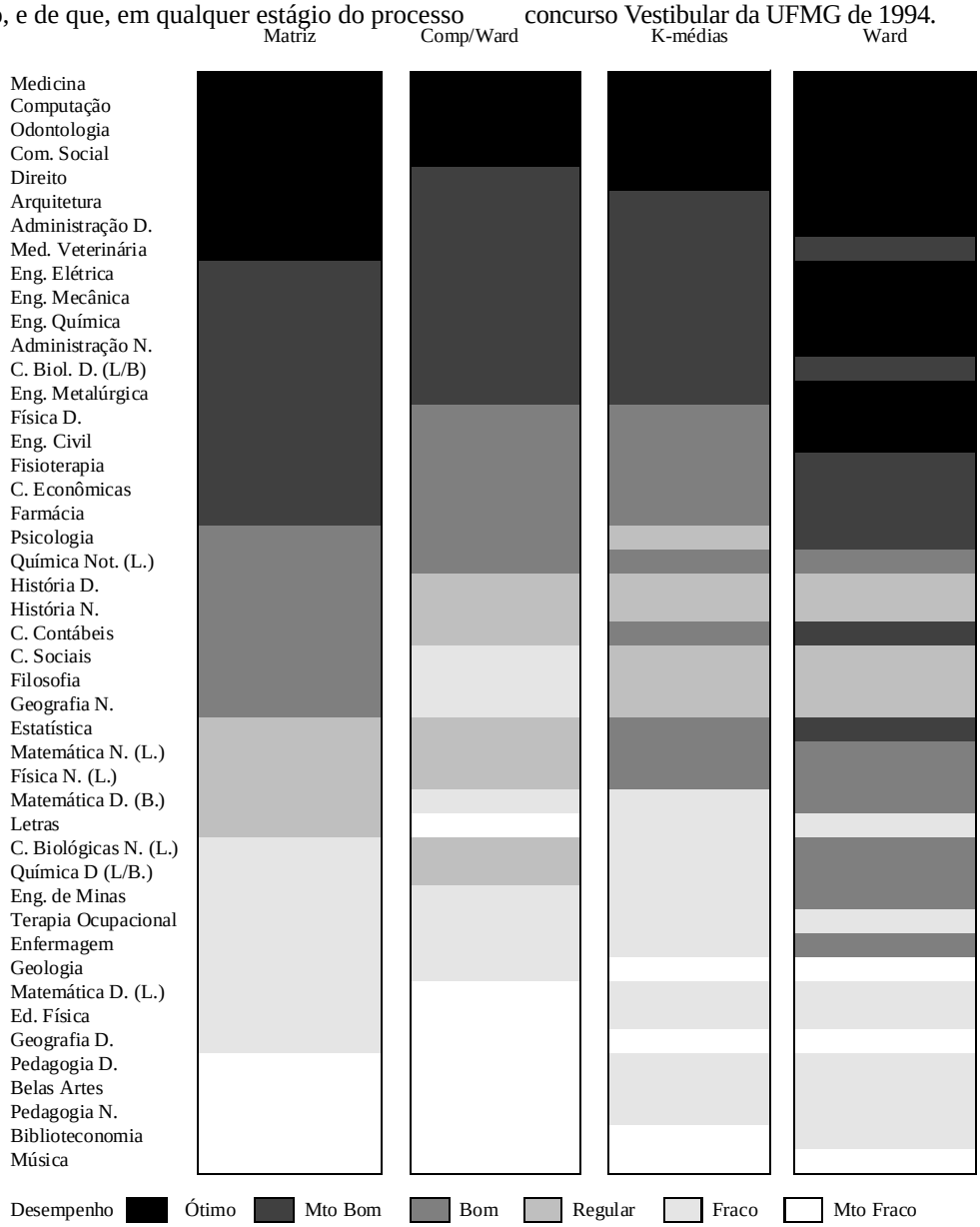

Figura 3: Agrupamento dos Cursos da UFMG, Segundo o Desempenho dos Aprovados no Vestibular de 1994, nas oito Provas da Primeira Etapa do Concurso, a partir de Análises de Conglomerados Diferentes.

Figure 3: Results for different group analyses by course in UFMG in the 94 selection.

De acordo com os resultados apresentados na Figura 1, verifica-se que as classificações não concordam inteiramente. As diferenças observadas dizem respeito ao número total de cursos compreendidos nos grupos e, conseqüentemente, à composição desses grupos. Essas diferenças foram quantificadas e os dados que as expressam são mostrados no Quadro 3. A concordância entre as classificações foram analisadas estatísticamente através do coeficiente de concordância de Kappa (Woolson, 1987). Os resultados indicaram que a um nível de significância de 5% há concordância entre as classificações obtidas pelo método da Matriz Ordenável e respectivamente pelos métodos de Ward (K= 0,3079; Z = 6,60) , K-Médias (K=0,3652; Z = 7,74) e Componentes Principais (K=0,4253; Z = 8,11), sendo esta última a concordância mais significativa.

Entretanto, a partir dos dados do Quadro 3, destaca-se, com relação aos resultados alcançados através do método gráfico, que:

- a maior concordância é observada nos resultados obtidos através da análise de conglomerados que associa a técnica de componente principal ao método de Ward; ou seja, em média, 55,27%, da composição dos grupos obtidos através dessa análise está de acordo com a composição da classificação do método gráfico;

- a menor concordância, 30,56%, é assinalada aos resultados obtidos através do método de Ward;

- as maiores semelhanças de resultados, considerando-se cada uma das três classificações matemático-estatísticas, são observadas nos Grupos I e II, que compreendem os cursos, nos quais os vestibulandos apresentam desempenhos classificados como ótimo e muito bom;

- as maiores diferenças de resultados, observando-se também as três classificações matemático-estatísticas, referem-se à composição dos grupos intermediários das classificações - os Grupos II e III, caracterizados por desempenhos classificados como bom e regular. Nota-se que, quer na classificação obtida através do método de K-médias, quer na do método de Ward, a composição do Grupo III discorda totalmente da composição definida pelo método gráfico.

Esses resultados, de modo geral, eram esperados e são atribuídos às formas específicas dos vários métodos de realizar os agrupamentos. Quanto a essas formas, podem ser considerados dois aspectos que as caracterizam para se discutir seus efeitos nos resultados alcançados, quais sejam: de um lado, a observação ou não da propriedade de hierarquia; e, de outro lado, a consideração das mensurações obtidas, relativas às oito variáveis estudadas, segundo cada elemento da população, para se definir o grau de similaridade desses elementos, ou então, a consideração de uma variável que sintetiza as mensurações referentes às oito variáveis originais.

Verifica-se, então, que, o resultado mais discordante com o obtido através do método gráfico é assinalado ao método matemático-estatístico que observa a propriedade da hierarquia, o de Ward, ao contrário do método gráfico. Mas, por outro lado, constata-se que o resultado mais concordante com o do método gráfico não foi assinalado ao método matemático-estatístico não hierárquico, o K-Médias, mas sim ao procedimento que associa a uma técnica de análise de agrupamento hierárquica a uma de Componentes Principais.

Desse modo, considera-se que a observação ou não da propriedade da hierarquia que caracteriza os esquemas de classificação dos métodos comparados influencia apenas secundariamente os resultados alcançados. O aspecto dos métodos que parece influenciar significativamente os resultados diz respeito ao modo de se considerar as mensurações realizadas para os elementos da população. Quanto a esse aspecto, a Matriz Ordenável e o procedimento que associa uma análise de Componentes Principais ao Método de agrupamento de Ward operam de modo muito semelhante; daí a maior concordância dos resultados obtidos através dessas técnicas.

RESULTADOS DO MÉTODO GRÁFICO	RESULTADOS DOS MÉTODOS MATEMÁTICO-ESTATÍSTICOS								
	Componentes Principais e Ward			K-Médias			Ward		
Número de Cursos Compreendidos nos Grupos Definidos pela Matriz Ordenável	Cursos Total	Cursos Concordantes		Cursos Total	Cursos Concordantes		Cursos Total	Cursos Concordantes	
	Abs.	Abs.	% Total	Abs.	Abs.	% Total	Abs.	Abs.	% Tota
Grupo I 8	4	4	100,0	5	5	100,0	14	7	50,0
Grupo II 11	10	6	60,0	9	6	66,66	8	4	50.0
Grupo III 8	7	2	28,57	10	2	20,0	8	1	12,5
Grupo IV 5	8	3	37,5	6	0	0	5	0	0
Grupo V 9	8	4	50,0	12	7	58,33	8	3	37,5
Grupo VI 5	9	5	55,56	4	2	50,0	3	1	33,33
Média	-	-	55,27	-	-	49,17	-	-	30,56

Quadro 3: Concordância dos Resultados das Análises de Conglomerados realizadas através de Métodos Matemático-Estatísticos e Gráfico.

Chart 3: Comparision of group Analyses Results using math-statistics and graphic methods.

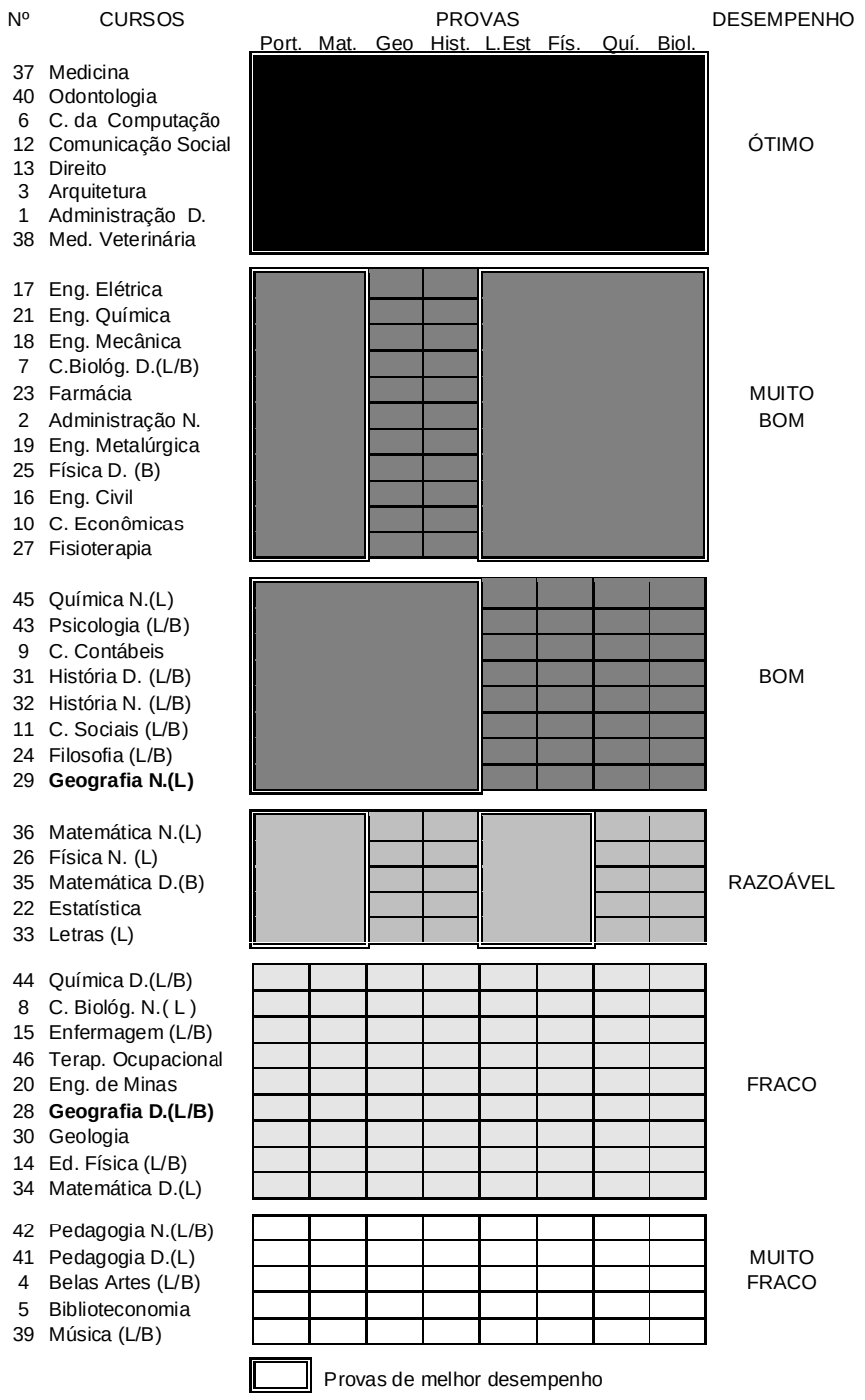

Figura 4: Matriz de Comunicação de Informações Úteis.
Figure 4: Comunicacion of useful information Matrix.

No procedimento matemático-estatístico, em questão, cria-se uma nova variável para sintetizar o desempenho dos estudantes nas várias provas e para proceder, por seu intermédio, a formação dos agrupamentos, conforme se descreveu anteriormente. Nesse sentido, o procedimento considera o desempenho global dos estudantes nas oito provas, ao invés de observar os resultados particulares de cada uma dessas provas.

Por sua vez, no procedimento gráfico, os desempenhos dos vestibulandos para cada uma das oito provas são representados em diagramas distintos, segundo os cursos que compreendem a população. Porém, os cursos são agrupados, considerando-se o desempenho geral dos vestibulandos, uma vez que são distinguidos: de um lado, pela observação do comprimento predominante dos segmentos de barras, correspondentes aos desempenhos referentes a cada uma das provas; e, de outro lado, observando-se o padrão

de tonalidades, claro e escuro, que pode ser percebido no conjunto das barras referentes aos cursos.

Desse modo, ao se empreender essa análise procura-se aproximar os cursos que apresentam as mesmas tendências, quanto ao montante do percentual de aprovados com 50 ou mais pontos nas oito provas do concurso Vestibular e à situação desses aprovados em relação a média da UFMG, nas provas consideradas.

Nota-se que, no âmbito da análise de agrupamento empreendida através do método gráfico, o modo de se considerar a situação dos aprovados de cada um dos cursos estudados, em relação à média da UFMG, é diferente das formas seguidas pelos métodos matemático-estatísticos. Esse aspecto do método pode explicar, ainda, as diferenças dos resultados que, de modo geral, foram alcançados, bem como, os resultados que foram particularizados, quanto ao grau de concordância relativa à composição dos seis grupos obtidos.

Verificou-se que, apesar da maior ou menor concordância das classificações obtidas através dos quatro métodos utilizados, os resultados encontrados são consistentes. A consistência dos resultados pode ser exemplificada pelo fato das maiores semelhanças entre as classificações terem sido verificadas sempre para os mesmos grupos, a saber, para os grupos que apresentam os melhores desempenhos, Grupos I e II, e, de uma forma apenas um pouco menos expressiva, para os de pior desempenho, Grupos V e VI. Por outro lado, a consistência dos resultados é exemplificada, também, pelas discordâncias dos resultados, em relação à composição dos Grupos III e IV.

Nesse sentido, há concordância de resultados quando se trata de cursos, cujos vestibulandos apresentam, de modo geral, desempenhos homogêneos em relação a média da UFMG, para todas as provas estudadas, seja apresentando resultados sempre acima da média, seja apresentando resultados sempre abaixo da média. As discordâncias ocorrem, então, quando os resultados são heterogêneos, ou seja, variáveis em função da prova considerada, ora situando-se acima, ora abaixo da percentagem média da UFMG.

O fato dos vestibulandos apresentarem desempenhos homogêneos ou heterogêneos nas oito provas do concurso, segundo os cursos estudados, é apreendido, é claro, por todas as técnicas matemático-estatísticas empregadas nesse estudo. Essas diferenças causam efeito na classificação dos cursos, na medida em que influenciam a medida de similaridade considerada para agrupá-los. Porém, essas diferenças são consideradas explicitamente apenas quando se emprega o método gráfico.

Na Matriz Ordenável, os desempenhos homogêneos ou heterogêneos dos vestibulandos são percebidos na imagem que serve de base para a análise, como pode ser visto na Figura 2, apresentada anteriormente. Essas diferenças, expressas através dos padrões formados pelos comprimentos dos segmentos de barra e, sobretudo, por suas tonalidades, permitem identificar tendências, quais sejam, de desempenhos destacados, quer no conjunto das provas, quer em algumas delas, ou de desempenhos sem destaque, para todos os elementos da população estudada.

Por sua vez, essas tendências são consideradas no procedimento de agrupamento dos cursos, como pode ser visto através dos resultados da análise mostrados na Figura 3, que também foi apresentada anteriormente, ou como pode ser observado na Figura 4, de modo mais destacado.

CONSIDERAÇÕES FINAIS

Os resultados deste trabalho mostram que o método gráfico Matriz Ordenável, apropriado para a sintetização da informação de variáveis e para o agrupamento de dados, pode ser eficazmente aplicável naquelas situações onde os métodos estatísticos correspondentes são utilizados. A grande concordância observada entre este método e os de Componente Principal, Ward e K-Médias sugere que apesar da grande subjetividade envolvida na aplicação do método da Matriz Ordenável, este apresentou uma solução próxima daquelas obtidas por métodos que têm fundamentação matemática. Entretanto, uma avaliação mais segura a propósito da eficácia do método gráfico exige novos testes com outros conjunto de dados.

Em termos práticos, os métodos estatísticos levam certa vantagem em relação ao método gráfico, apesar dos avanços concernentes ao desenvolvimento dos aplicativos computacionais disponíveis para o emprego desse método. Os métodos estatísticos podem ser aplicados rapidamente a qualquer conjunto de dados, tendo em vista a grande disponibilidade de *softwares* estatísticos apropriados para essas análises. Desse modo, esses métodos representam boas alternativas ao da Matriz Ordenável.

BIBLIOGRAFIA

ANDERBERG, M. R. Cluster Analysis for Applications. New York : Academic Press, 1973.

BASILEVSKY, A. Applied Matrix Algebra in the Statistical Sciences. New York: North-Holland, 1983.

DILLON, W. R. & GOLDSTEIN, M. Multivariate Analysis Methods and Applications. New York : John Wiley & Sons, 1984.

JOHNSON, R. A & WICKERN, D. W. Applied Multivariate Statistical Analysis. 3ª ed. New Jersey: Prentice Hall, 1992.

SANTOS, M. M. D. dos & SANCHEZ, M. C. O Tratamento Gráfico de um Conjunto de Dados - Estudo da Técnica Matriz Ordenável Quantitativa. *Geografia*, Rio Claro, vol. 21, nº 1, abril 1996, 77-101 (no prelo).

WOOLSON, R. Statistical Methods for the Analysis of Biomedical Data. New York : John Wiley & Sons, 1987.