

O envelhecimento populacional nos municípios do Sul/Sudoeste de Minas Gerais: análise de agrupamento

Larissa Gonçalves

Orientadora: Profa. Dra. Patrícia Ramos

Coorientador: Prof. Dr. Lincoln Frias

Programa de Pós-Graduação em Estatística Aplicada e Biometria
Universidade Federal de Alfenas

Sumário

- 1 Transição demográfica e envelhecimento
- 2 Dados e metodologia
- 3 Resultados parciais
- 4 Considerações finais
- 5 Referências bibliográficas

Transição demográfica

- Mudanças no regime demográfico
- Altera taxa de crescimento da população, estrutura etária e leva ao processo de envelhecimento populacional

Envelhecimento populacional

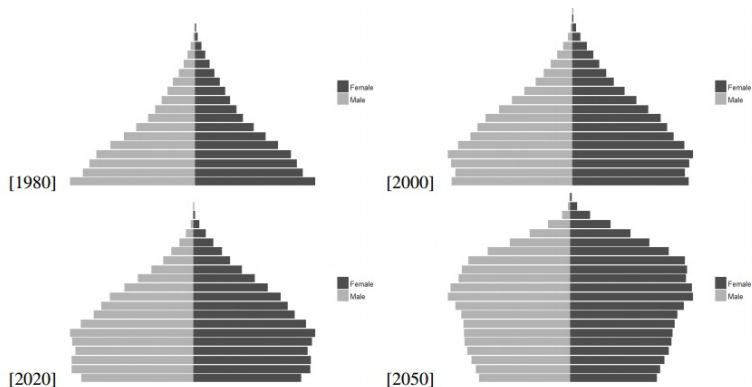


Figura: Pirâmides etárias absolutas do Brasil, 1980-2050.

Fonte: elaboração própria a partir de dados do *United States Census Bureau*, fonte disponível em: www.census.gov/population/international/data/idb

Objetivos

- Propor uma classificação dos municípios da mesorregião Sul/Sudoeste de Minas Gerais em relação ao processo de envelhecimento populacional.
- Realizar a análise de agrupamento utilizando as variáveis originais e, em seguida, com os escores dos componentes principais, com o intuito de comparar os resultados obtidos.
- Utilizar cinco métodos de agrupamento hierárquicos aglomerativos (ligação simples, ligação completa, distância média, centroide e Ward) e o método não hierárquico das k-médias.
- Analisar diferentes critérios para definição do número de grupos na partição final.

Dados

- Dados de 8 variáveis demográficas dos 146 municípios da mesorregião Sul/Sudoeste de Minas Gerais
- Censo demográfico 2010 do IBGE, consultado a partir do Atlas do Desenvolvimento Humano no Brasil 2013

Tabela 1: Descrição das variáveis demográficas

Sigla	Variável
espvida	esperança de vida ao nascer
tft	taxa de fecundidade total
mort1	mortalidade infantil
mort5	mortalidade até os 5 anos de idade
rd	razão de dependência
sobre40	probabilidade de sobrevivência até 40 anos
sobre60	probabilidade de sobrevivência até 60
t_env	taxa de envelhecimento

Metodologia

- Técnica multivariada de análise de agrupamento
- Análise exploratória: multicolinearidade entre as variáveis e normalidade
- Matriz de correlações entre as variáveis
- Procedimentos de Mardia e o teste Shapiro-Wilk para o caso multivariado
- Processo de agrupamento
- Variáveis originais

Metodologia

1) Seleção da medida de distância

- Distância de Mahalanobis e distância euclidiana

2) Padronização das variáveis

3) Escolha dos métodos de agrupamento

- Ligação simples, ligação completa, distância média, centroide e Ward. Em seguida, também foi utilizado o método não hierárquico das k-médias

4) Definição do número de grupos

- Pacote *NbClust*
- Trabalho de Milligan e Cooper (1985) e o trabalho de Tibshirani e Walther (2005)

Metodologia

Nove primeiros métodos considerados os melhores do artigo no trabalho de Milligan e Cooper (1985), sendo eles:

- pseudo F (CALINSKI; HARABASZ, 1974)
- $Je(2)/Je(1)$ (DUDA; HART et al., 1973)
- C-Index (HUBERT; LEVIN, 1976)
- gamma (BAKER; HUBERT, 1975)
- Beale (BEALE, 1969)
- estatística CCC (Cubic Clustering Criterion)
- ponto bisserial
- Gplus (ROHLF, 1974)
- Davies and Bouldin (DAVIES; BOULDIN, 1979)

Estatística gap (TIBSHIRANI; WALTHER, 2005)

Metodologia

- Escores dos componentes principais
- Representatividade dos k primeiros componentes conjuntamente com o método gráfico *scree plot*
- Resultados obtidos com os dados originais e os componentes principais serão comparados
- Mediana
- Coeficiente de variação

Análise descritiva das variáveis

Tabela 2 – Resumo estatístico das variáveis originais

	espvida	tft	mort1	mort5	rd	sobre40	sobre60	t_env
mínimo	73,03	1,33	10,35	12,11	37,68	92,33	79,54	5,46
1 quartil	74,44	1,79	13,40	15,63	43,26	93,23	81,67	8,49
mediana	75,56	1,95	14,45	16,86	44,85	93,92	83,32	9,39
média	75,46	1,95	14,69	17,09	45,24	93,85	83,15	9,45
3 quartil	76,28	2,08	16,18	18,86	47,27	94,35	84,36	10,32
máximo	78,15	2,70	18,50	21,55	53,20	95,99	87,58	14,85

Fonte: elaboração própria.

Análise descritiva das variáveis

Tabela 3 – Municípios com os melhores indicadores demográficos da mesorregião Sul/Sudoeste de Minas Gerais, 2010

município	espvida	município	tft	município	mort1	mort5
Passos	78,15	São Sebastião do Rio Verde	1,33	Passos	10,35	12,11
Itajubá	78,06	São João da Mata	1,39	Itajubá	10,50	12,12
Guaxupé	77,81	Espírito Santo do Dourado	1,41	Poços de Caldas	11,27	13,18
município	rd	município	sobre40	sobre60	município	t_env
Tocos do Moji	37,68	Itajubá	95,99	87,58	Córrego do Bom Jesus	14,85
São João da Mata	39,05	São Lourenço	95,62	86,63	Senador José Bento	13,65
Varginha	39,18	Passos	95,25	86,54	Pratápolis	12,97

Fonte: elaboração própria.

Análise descritiva das variáveis

Tabela 4 – Municípios com os piores indicadores demográficos da mesorregião Sul/Sudoeste de Minas Gerais, 2010

município	espvida	município	tft	município	mort1	mort5
Carmo da Cachoeira				Carmo da Cachoeira		
Divisa Nova	73,03	São Bento Abade	2,70	Divisa Nova	18,50	21,55
São Bento Abade				São Bento Abade		
São Tomé das Letras				São Tomé das Letras		
Bandeira do Sul				Bandeira do Sul		
Ibitiúra de Minas	73,14	Senador Amaral	2,55	Ibitiúra de Minas	18,4	21,34
Toledo				Toledo		
Natércia	73,28	Carmo da Cachoeira	2,53	Natércia	18,1	21,06
				Fortaleza de Minas		21,00
município	rd	município	sobre40	sobre60	município	t_env
		Carmo da Cachoeira				
São Tomás de Aquino	53,20	Divisa Nova	92,33	79,54	São Tomé das Letras	5,65
		São Bento Abade				
		São Tomé das Letras				
		Bandeira do Sul				
Divisa Nova	52,86	Ibitiúra de Minas	92,40	79,71	Carmo da Cachoeira	7,44
		Toledo				
Serrania	52,64	Natércia	92,49	79,92	São Bento Abade	5,90

Fonte: elaboração própria.

Análise descritiva das variáveis - normalidade multivariada

- Teste de Mardia rejeitou a hipótese de normalidade multivariada com um valor-p menor que 0,001
- O mesmo resultado foi obtido pelo teste Shapiro-Wilk multivariado, que também rejeitou a hipótese de normalidade dos dados com um valor-p menor que 0,001

Análise descritiva das variáveis - multicolinearidade

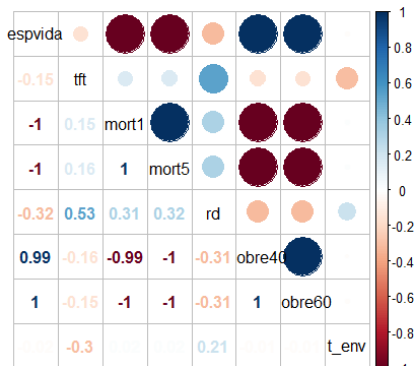


Figura: Correlações entre as variáveis

Agrupamentos com as variáveis originais

- Primeiro, as observações foram agrupadas de acordo com as 6 variáveis demográficas utilizando a distância de Mahalanobis e os 5 métodos hierárquicos aglomerativos.
- Foram obtidos os dendrogramas e os resultados da escolha do número final de grupos para cada caso, de acordo com os 10 critérios selecionados.

Agrupamentos com as variáveis originais

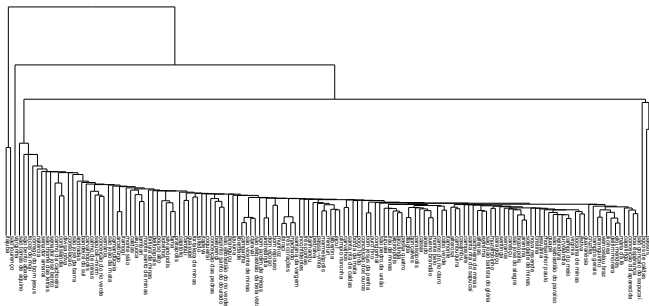


Figura: Dendrograma pelo método ligação simples e distância de Mahalanobis

Pelo dendrograma, um grupo é formado por Itajubá e São Lourenço e o outro pelos demais 144 municípios. Mesmo resultado obtido pela maioria dos critérios selecionados.

Agrupamentos com as variáveis originais

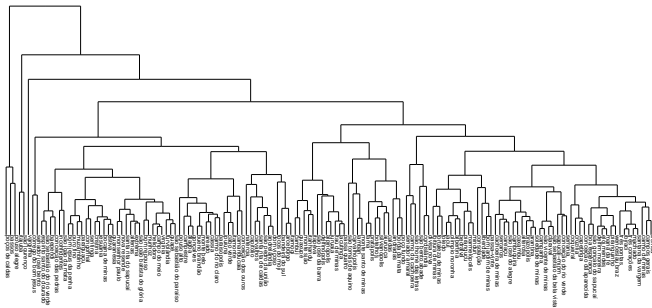


Figura: Dendrograma pelo método ligação completa e distância de Mahalanobis

Pelo dendrograma, um grupo é formado por três municípios, sendo eles Pouso Alegre, Passos e Poços de Caldas e o segundo contém os 143 municípios restantes. Mesmo resultado obtido pela maioria dos critérios selecionados.

Agupamentos com as variáveis originais

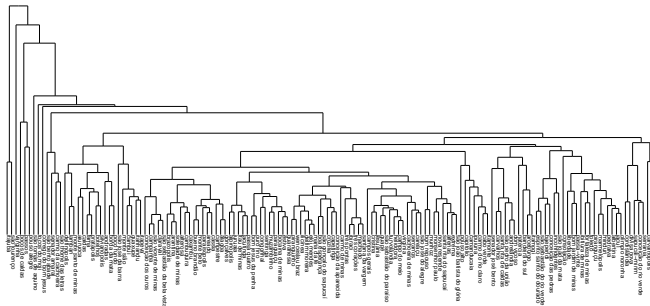


Figura: Dendrograma pelo método distância média e distância de Mahalanobis

Pelo dendrograma, há um grupo formado por Itajubá e São Lourenço, um contendo apenas a cidade de Varginha, outro formado por Pouso Alegre, Passos e Poços de Caldas e, por último, um grupo com as demais 140 observações. Pelos critérios selecionados o resultado é o mesmo do método ligação simples.

Agrupamentos com as variáveis originais

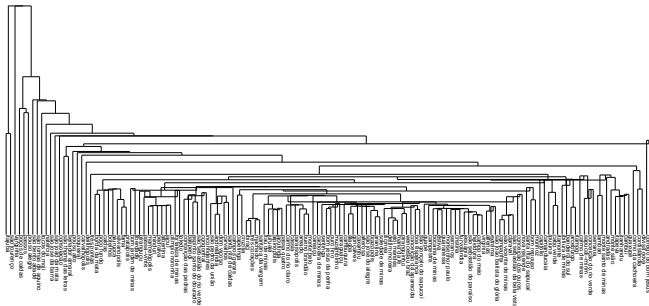


Figura: Dendrograma pelo método centroide e distância de Mahalanobis

Pelo dendrograma, Varginha se diferenciaria em relação às variáveis medidas em tal medida que, novamente, estaria sozinha em um grupo. O outro grupo seria formado por Itajubá e São Lourenço e um grande grupo pelos demais municípios. o resultado da análise dos critérios de número de grupos é o mesmo dos métodos de ligação simples e distância média.

Agrupamentos com as variáveis originais

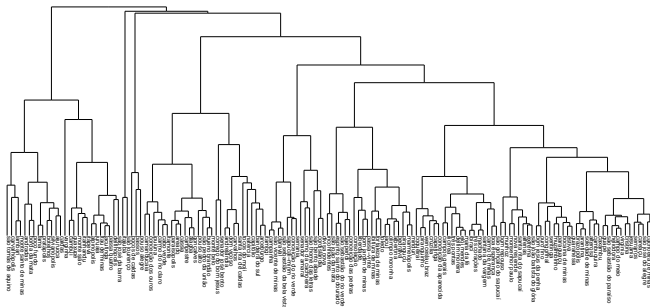


Figura: Dendrograma pelo método de *Ward* e distância de Mahalanobis

O dendrograma sugere um corte que produz 5 grupos finais. divisão dos municípios em cinco grupos (2, 3 26, 28 e 87 municípios). Essa divisão também sugere que Itajubá e São Lourenço formariam um único grupo, assim como Pouso Alegre, Passos e Poços de Caldas. Segundo os critérios, 26 municípios estariam alocados em um grupo e 120 em outro grupo.

Tabela 5: Número de observações nos grupos usando métodos hierárquicos aglomerativos e distância de Mahalanobis

Método	Grupo 1	Grupo 2
ligação simples	144	2
ligação completa	143	3
distância média	144	2
centroide	144	2
<i>Ward</i>	26	120

Fonte: elaboração própria

Agrupamentos com as variáveis originais

Tabela: Medianas das variáveis dos grupos obtidos por métodos hierárquicos aglomerativos e distância de Mahalanobis

Método	Grupo	espvida	tft	mort1	rd	sobre60	t_env
ligação simples	1	75,53	1,96	14,50	44,88	83,28	9,39
	2	77,67	1,67	11,00	42,92	87,06	9,50
ligação completa	1	75,53	1,96	14,50	44,92	83,28	9,47
	2	77,33	1,69	11,27	40,55	85,62	8,41
distância média	1	75,53	1,96	14,50	44,88	83,28	9,39
	2	77,67	1,67	11,00	42,92	87,06	9,50
centroide	1	75,53	1,96	14,50	44,88	83,28	9,39
	2	77,67	1,67	11,00	42,98	87,06	9,50
Ward	1	77,23	1,95	12,16	43,56	85,69	9,82
	2	75,24	1,96	14,95	44,99	82,86	9,38

Agrupamentos com as variáveis originais

Tabela: Coeficiente de variação (CV) das variáveis dos grupos obtidos pelos métodos hierárquicos aglomerativos e distância de Mahalanobis

Método	Grupo	espvida	tft	mort1	rd	sobre60	t_env
ligação simples	1	1,70	13,12	13,27	6,57	2,24	15,96
	2	0,71	7,62	6,43	4,56	0,84	12,13
ligação completa	1	1,69	13,13	13,14	6,46	2,27	15,85
	2	0,62	6,20	5,78	3,20	0,76	10,70
distância média	1	1,70	13,12	13,27	6,57	2,24	15,96
	2	0,71	7,62	6,43	4,56	0,84	12,13
centroide	1	1,70	13,12	13,27	6,57	2,24	15,96
	2	0,71	7,62	6,74	4,56	0,84	12,13
Ward	1	0,57	11,08	4,87	7,46	0,72	17,98
	2	1,51	13,61	11,79	6,34	2,03	15,33

- Cada método produziu um resultado
- Número de grupos na partição final
- Grupos muito desbalanceados
- Tentativa de confirmar essa representação dos dados
- Distância euclidiana com dados padronizados

Agrupamentos com as variáveis originais

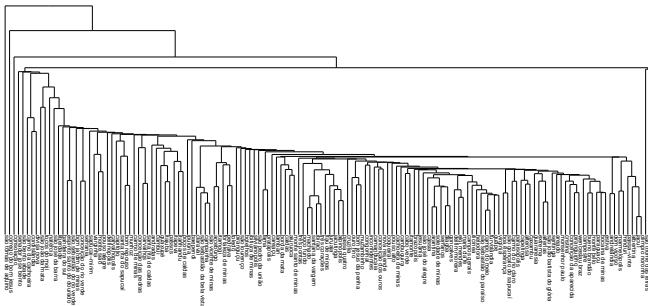


Figura: Dendrograma pelo método de ligação simples e distância de euclidiana

Dois grupos, um formado somente por São Tomás de Aquino e o outro pelos demais municípios. Três grupos um deles continuaria apenas com São Tomás de Aquino e um segundo formado por Córrego do Bom Jesus e, um terceiro grupo composto pelos outros 144 municípios. Os critérios sugerem dois grupos.

Agrupamentos com as variáveis originais

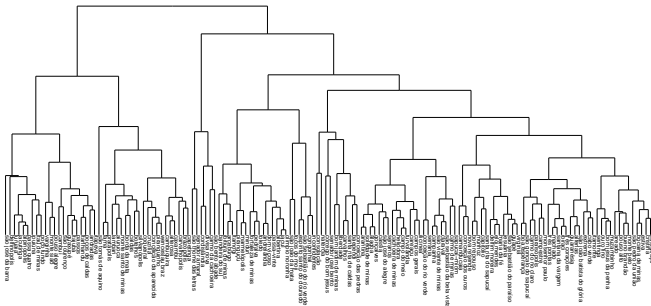


Figura: Dendrograma pelo método de ligação completa e distância de euclidiana

O dendrograma sugere um corte que produz 3 grupos finais. Os critérios de número de grupos sugerem 2 grupos com o ideal, dessa forma, em um grupo estariam 42 municípios e, no outro, 104 municípios.

Agrupamentos com as variáveis originais

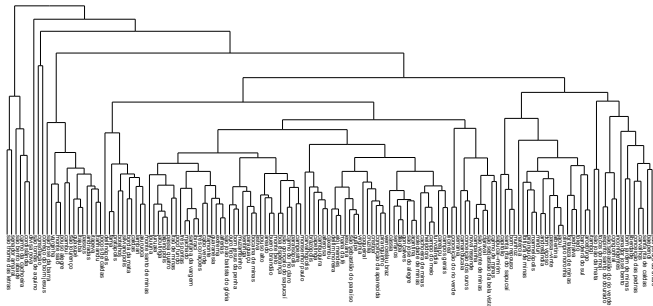


Figura: Dendrograma pelo método distância média e distância de euclidiana

Agrupamentos com as variáveis originais

- Corte no dendrograma sugerem quatro ou seis grupos para divisão das observações
- Pelos critérios de número de grupos, a maioria sugere um grupo com 140 municípios e outro formado por seis municípios (Carmo da Cachoeira, Cordislândia, Divisa nova, São Bento Abade, São Thomé das Letras e Senador Amaral).

Agrupamentos com as variáveis originais

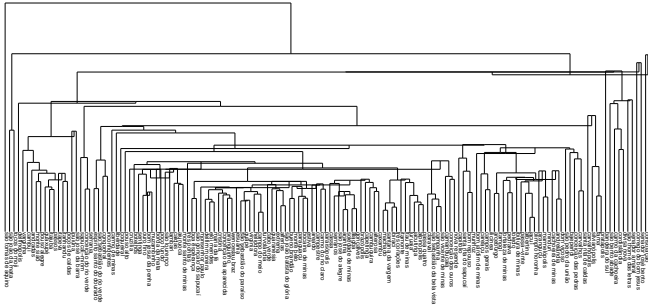


Figura: Dendrograma pelo método do centroide e distância de euclidiana

Tanto pela análise do dendrograma, quanto pelos critérios de número de grupos o resultado sugere que São Tomás de Aquino se diferencia dos demais municípios a ponto de estar separado de todos os outros em um grupo.

Agrupamentos com as variáveis originais

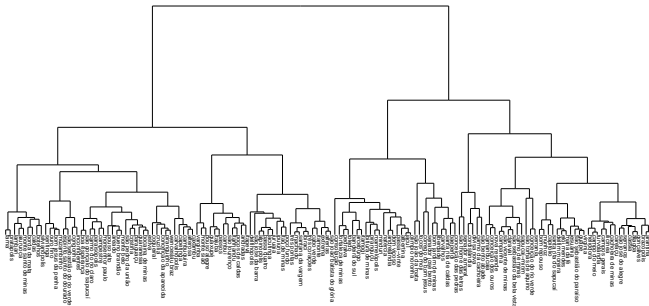


Figura: Dendrograma pelo método de *Ward* e distância de euclidiana

Fica evidente a separação das observações em dois grandes grupos, com aproximadamente o mesmo número de municípios. Com essa divisão, 77 municípios ficam inseridos em um grupo e os outros 69 municípios no outro grupo.

Agrupamentos com as variáveis originais

Tabela: Número de observações nos grupos usando métodos hierárquicos aglomerativos e distância euclidiana

Método	Grupo 1	Grupo 2
ligação simples	145	1
ligação completa	42	104
distância média	140	6
centroide	145	1
<i>Ward</i>	74	72

Fonte: elaboração própria

Agupamentos com as variáveis originais

Tabela: Medianas das variáveis dos grupos obtidos por métodos hierárquicos aglomerativos e distância euclidiana

Método	Grupo	espvida	tft	mort1	rd	sobre60	t_env
ligação simples	1	75,53	1,95	14,50	44,85	83,28	9,38
	2	77,46	2,43	11,80	53,20	86,00	10,70
ligação completa	1	76,96	1,92	12,45	44,45	85,32	9,76
	2	75,03	1,97	15,25	44,88	82,55	9,36
distância média	1	75,69	1,94	14,30	44,83	83,52	9,51
	2	73,03	2,49	18,50	49,66	79,54	6,67
centroide	1	75,53	1,95	14,50	44,85	83,28	9,38
	2	77,46	2,43	11,80	53,20	86,00	10,70
<i>Ward</i>	1	76,22	1,91	13,50	43,79	84,28	9,44
	2	74,41	1,98	16,20	46,36	81,64	9,36

Agupamentos com as variáveis originais

Tabela: Coeficiente de variação (CV) das variáveis dos grupos obtidos pelos métodos hierárquicos aglomerativos e distância euclidiana

Método	Grupo	espvida	tft	mort1	rd	sobre60	t_env
ligação simples	1	1,72	13,08	13,49	6,44	2,28	15,91
ligação completa	1	0,99	10,68	8,95	7,42	1,29	15,69
	2	1,36	13,91	10,28	6,20	1,83	15,91
distância média	1	1,65	12,10	13,05	6,31	2,18	14,79
	2	0,67	4,72	4,45	7,07	0,94	25,25
centroide	1	1,72	13,08	13,49	6,44	2,28	15,91
<i>Ward</i>	1	0,99	10,78	8,48	5,67	1,29	14,46
	2	1,23	14,39	8,99	6,46	1,67	17,32

Considerações finais

- Com o uso da distância de Mahalanobis houve uma tendência em alocar em um grupo menor os municípios com tendência a registrar melhores desempenhos nos indicadores considerados
- O uso da distância euclidiana apresentou tendência em separar em um grupo menor os municípios com piores desempenhos
- Não há um consenso sobre qual o melhor método de agrupamento
- Os próximos passos são a aplicação do método não hierárquico k-médias e os agrupamentos utilizando os escores dos componentes principais.

Referências bibliográficas

BARTHOLOMEW, D. J.; STEELE, F.; GALBRAITH, J.; MOUSTAKI, I. **Analysis of multi-variate social science data**. Boca Raton: CRC press, 2008.

CAMARANO, A. A. O. **Novo regime demográfico: uma nova relação entre população edesenvolvimento?**[S.I.]: Instituto de Pesquisa Econômica Aplicada (Ipea), 2014.

EVERITT, B. S.; LANDAU, S.; LEESE, M.; STAHL, D. **Cluster analysis**. 5. ed. UK: JohnWiley and Sons, 2011.

FERREIRA, D. F. **Estatística multivariada**. 2. ed. Lavras: Editora UFLA, 2011.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: umaabordagem aplicada**. Belo Horizonte: Editora UFMG, 2005.