ANÁLISE DO CONJUNTO DOS CANDIDATOS AO VESTIBULAR DA UESC NO ANO DE 2008 USANDO ANÁLISE DE CORRESPONDÊNCIA

Enio Galinkin JELIHOVSCHI[1]
Marcelo Inácio Ferreira Ferraz[1]

▪ RESUMO: O objetivo do presente trabalho foi utilizar a metodologia da análise de correspondência para analisar os questionários sócio-culturais respondidos pelos candidatos ao vestibular à Universidade Estadual de Santa Cruz do UESC no ano de 2008. Os gráficos bidimensionais resultantes mostram como se associam as variáveis, no caso as questões do questionário. Essas associações ligadas às componentes principais mostram quais são os fatores que levam à variabilidade ou inércia dos dados. Os resultados mostram que dois fatores polarizam esta inércia; o primeiro e mais importante que é o econômico, escola particular-pública, escolaridade alta-baixa dos pais; o segundo é o de competência do candidato.

▪ PALAVRAS-CHAVE: Análise de correspondência; questionário sócio-cultural.

1 Introdução

Analise de correspondência (CA, correspondence analysis) é uma técnica de análise exploratória de dados e redução de dimensionalidade. Foi desenvolvida para o uso principalmente em dados categóricos multivariados. A principio foi usada para entender as relações entre linhas e colunas de uma tabela de contingência depois que o teste qui-quadrado deu um resultado significativo. Com isto queremos dizer que a análise de correspondência é uma metodologia voltada para a análise de dados categóricos. Seu resultado é uma representação gráfica, simples e elegante, que leva a uma rápida interpretação e entendimento da estrutura por trás dos

[1]Universidade Estadual de Santa Cruz – UESC, Área de Estatística, Departamento de Ciências Exatas e Tecnológicas – DCET, CEP: 45662-900. Ilhéus, Bahia, Brasil. E-mail: *eniojelihovs@gmail.com / mfferraz@uesc.br*

dados. Em outras palavras, a análise de correspondência simplifica a complexidade de uma alta dimensionalidade, descrevendo toda a informação contida nos dados. A análise de correspondência teve seu início a mais de 60 anos atrás quando grandes estatísticos, inclusive Sir Ronald Fisher, começaram a trabalhar na interpretação da correlação entre linhas e colunas de uma tabela de contingência (2). A moderna aplicação gráfica, porém, surgiu na França impulsionada pelo matemático, linguista e analista de dados Jean Paul Benzécri seus colegas e estudantes (4). A característica mais importante da análise de correspondência é a sua consideração simultânea das categorias de duas variáveis categóricas, ou seja as linhas e colunas de uma tabela de contingência. Neste sentido sua representação gráfica é o que se chama de biplot. O desenvolvimento do biplot a partir dos trabalhos de Gabriel remete mais ou menos à mesma época do desenvolvimento de Benzécri (3).

Em resumo análise de correspondência é uma representação gráfica de baixa dimensão, em geral somente duas, no qual cada nível das variáveis usadas, ou cada nível dos fatores definidos nas linhas e colunas de uma tabela de contingência, aparecem como pontos e, dependendo do lugar onde aparecem no mapa, permite encontrar as informações buscadas. Além disso, tem uma exigência de entrada de dados muito flexível, basta uma matriz de números não negativos. Em geral para que a análise de correspondência tenha um resultado com real significância é importante que: (7)

- A matriz de dados seja grande tal que uma simples inspeção visual não revele a sua estrutura e a relação entre as variáveis.

- As variáveis são homogêneas no sentido de que faz sentido calcular distâncias estatísticas entre elas.

- A matriz de dados tem sua estrutura pouco entendida.

A forma de cálculo da CA pode ser entendida como um caso especial da Análise de Componentes Principais (PCA, Principal Components analysis) aplicado às linhas e colunas de uma tabela de contingência, ou seja uma matriz formada pela tabulação cruzada de duas variáveis. Em lugar da métrica euclidiana usual a métrica usada na CA é a distância quiquadrada. E, finalmente, enquanto que o conhecido teste qui-quadrado simplesmente mostra que existe uma relação entre linhas e colunas de uma matriz de dados, a CA mostra como é que estas variáveis são relacionadas. A CA também pode ser usada no caso de um número maior do que duas variáveis. Neste caso é chamada de análise de correspondência múltipla (MCA, multiple corespondence analysis).

O estudo realizado consistiu na análise de um questionário sócio-cultural, que vem inserido no Manual do Candidato, respondido pelos candidatos ao vestibular da Universidade Estadual de Santa Cruz - UESC - no ano de 2008. Neste ano foi introduzido o sistema de reserva de vagas (cotas) nesta universidade o qual preconiza metade das vagas para alunos de escolas públicas. Este questionário é constituido por informações escolares, culturais e econômicas, além de uma identificação social.

A MCA foi usada para realizar uma análise exploratória das questões (variáveis) escolares no sentido de entender as relações entre elas e também comparar a estrutura das respostas entre os candidatos cotistas e não-cotistas. O mesmo foi feito também para as informações econômicas. Também foram usadas o método de empilhamento de tabelas na CA para se tentar entender as relações entre as variáveis escolares e econômicas.

2 Metodologia

2.1 Conceitos e definições

A matriz original de dados é dada por $N(I, J)$, ou seja uma tabela de contingência.

Seja também:

- total das linhas: $n_{i+} = \sum_j n_{ij}$
- total das colunas: $n_{+j} = \sum_i n_{ij}$
- total geral: $n = \sum_j \sum_i n_{ij}$

2.1.1 Perfis (Profiles)

Os perfis de cada linha i é um vetor formado pelas frequências da linha i dividido pelo total desta linha.

$\frac{n_{ij}}{n_{i+}} \qquad (j = 1, 2, \ldots, J)$

Os perfis de cada coluna j é um vetor formado pelas frequências da coluna j dividido pelo total desta coluna.

$\frac{n_{ij}}{n_{+j}} \qquad (i = 1, 2, \ldots, I)$

Este conceitos de perfis é de uma importância fundamental em CA, pois estes vetores de frequência relativa, tendo sua soma igual a 1 ou 100%, tem características geoméricas muito especiais, e podem ser estudados como pontos no espaço dos perfis. Outro conceito importante é perfis médios, cujas definições são:

- - Perfil de linha médio $\frac{n_{+j}}{n} \qquad (j = 1, 2, \ldots, J)$
- - Perfil de coluna médio $\frac{n_{i+}}{n} \qquad (i = 1, 2, \ldots, I)$

2.1.2 Massa

Massa é outro conceito fundamental em CA.

- massa das linhas $r_i = \frac{n_{i+}}{n} \qquad (i = 1, 2, \ldots, I)$
- massa das colunas $c_j = \frac{n_{+j}}{n} \qquad (j = 1, 2, \ldots, J)$

Podemos ver a que a massa das linhas é o perfil de coluna médio e a massa das colunas é o perfil de linha médio.

2.1.3 Matriz de correspondência

Definimos como matriz de correspondência $\mathbf{P}$ à matriz original dividida pelo total geral n, ou seja $\mathbf{P} = \frac{1}{n}N(I,J)$, ou seja $p_{ij} = \frac{n_{ij}}{n}$. A matriz de correspondência mostra como uma unidade de massa é distribuída através das células da matriz original.

2.1.4 Distância

O resultado principal da CA é um mapa em duas ou três dimensões dos perfis de uma matriz de dados visto como pontos neste mapa, por esta razão é importante definir uma métrica a ser usada como distância entre os pontos. Neste caso usamos uma variante da métrica euclidiana chamada distância euclidiana ponderada (4).

A forma geral desta métrica é: $\sqrt{\sum_j \omega_j (x_j - y_j)^2}$. No nosso caso a métrica usada para medir a distância entre duas linhas ou duas colunas de perfis é chamada de *Distância* χ^2 *(Quiquadrada)*. A distância χ^2 entre duas linhas de perfis i e i' é definida como: $\sqrt{\sum_j (\frac{n_{ij}}{n_{i+}} - \frac{n_{i',j}}{n_{i',+}})^2 / c_j}$.

De forma similar definimos a distância χ^2 entre duas colunas de perfis.

2.1.5 Princípio da equivalência distribucional

Se duas linhas da matriz $N(I,J)$ são proporcionais, quer dizer tem exatamente os mesmos perfis, elas podem ser substituidas por uma linha cujas frequências são a soma das respectivas frequências, sem que isto afete a distância entre a colunas de $N(I,J)$.

Da mesma forma se duas colunas de $N(I,J)$ são proporcionais elas podem ser substituidas por uma coluna cujas frequências são a soma das respectivas frequências, sem que isto afete a distância entre as linhas de $N(I,J)$.

Isto quer dizer que se subdividimos uma categoria de uma variável (coluna por exemplo) em duas tais que suas frequências sejam proporcionais (uma o dobro da outra por exemplo) estas duas novas categorias não trazem nenhuma informação a mais em relação à outra variável (a das linhas).

Isto acontece porque a distância qui-quadrada satisfaz o princìpio da equivalência distribucional. A distância euclidiana, por outro lado, não satisfaz este princípio.

2.1.6 Inércia

Inércia de uma forma geral é a soma ponderada do quadrado das distâncias de um conjunto de pontos ao seu centro (ou centróide); na CA os pontos são os perfis, os pesos de ponderação são as massas dos perfis e as distâncias são as distâncias qui-quadradas.

A estatística qui-quadrada é uma medida global da diferença entre as frequências observadas numa tabela de contigência e as frequências esperadas

calculadas a partir da hipótese de independência (homogeneidade) dos perfis de linhas (ou de colunas).

A inércia (total) de uma tabela de contingência é a estatística qui-quadrada dividida pelo total n da tabela. Geometricamente, a inércia mede quão "longe"os perfis de linha (ou de colunas) estão situados do perfil de linha (coluna) médio. O perfil médio pode ser considerado como o representante da hipótese de independência (isto é igualdade) dos perfis.

$\frac{\chi^2}{n} = \sum_i \sum_j \frac{(p_{ij} - r_i c_j)^2}{r_i c_j}$

2.2 Redução de dimensionalidade

Perfis consistindo de m elementos estão situados exatamente em espaços de dimensão $m - 1$. Por esta razão, perfis com mais de quatro elementos, ou seja cuja matriz N tem ambos os números de linhas e colunas maior do que quatro, estão situados em espaços de dimensão maior do que três, o que impossibilita sua visualização direta.

Por este fato, o mais importante resultado da CA é um gráfico de duas (ou três) dimensões no qual todos os perfis aparecem como pontos. Isto é feito através de uma decomposição da inércia total, o que possibilita a identificação de um pequeno número de dimensões (duas ou três) no qual os perfis podem ser representados.

Se pudermos identificar um subespaço de duas ou três dimensões, que se situe perto de todos os pontos que representam os perfis, poderemos então projetar os perfis neste subespaço e observar nas posições projetadas como uma aproximação das suas verdadeiras posições no espaço de dimensão maior. O que se perde neste processo é quanto e em qual direção os perfis se projetam fora do subespaço mencionado acima. O que se ganha é uma é uma visão dos perfis que não seria possível de outra forma.

2.2.1 Critério para a redução de dimensionalidade

Em CA, nós estamos essencialmente buscando um espaço de duas ou três dimensões que seja o mais "perto"possível do conjunto de pontos que representam os perfis situados num espaço de dimensão superior. Seja S um candidato a este espaço. Para um perfil i qualquer, com massa m_i, podemos calcular a distância qui-quadrada entre este ponto e S a qual chamamos de $d_i(S)$. A proximidade entre este perfil e S será então $m_i[d_i(S)]^2$, ou seja a distância a o quadrado ponderada pela massa. A proximidade de todos os perfis a S será então a soma daquelas quantidades para todos os perfis: proximidade a $S = \sum_i m_i[d_i(S)]^2$

O objetivo da CA será então encontrar o subespaço S que minimiza este critério. A fidelidade ou precisão desta representação gráfica de S é medida pela quantidade chamada de *percentual de inércia* que é a razão da inércia dentro de S pela inércia total.

A forma mais elegante de tanto definir a teoria da CA como de calcular a solução da minimização acima é usando a metodologia matemática conhecida *decomposição em valor singular* SVD (singlar value decomposition). A SVD é um

dos resultados mais úteis na teoria da matrizes, e a metodologia mais importante usada nos métodos de redução de dimensionalidade em estatística como por exemplo análise fatorial ou análise de componentes principais (1).

A CA determina as principais componentes (ou eixo, ou fator) de inércia e para cada componente seu auto valor correspondente. A primeira componente principal é a reta cujo auto valor correspondente é o maior de todos. A segunda é, entre todas as retas perpendiculares à primeira, a que tem a maior inércia, e assim por diante. O subespaço que otimiza a minimização citada é aquele gerado pelas componentes principais. A inércia representada por uma componente principal é chamada de *inércia principal* cujo valor equivale ao auto valor desta componente.

2.2.2 Análise de linhas e colunas

A análise das linhas consiste em situar os perfis de linha no espaço total e depois projetá-las no espaço gerado pelas duas (ou três) primeiras componentes principais. Da mesma forma fazemos a análise das colunas. Estas duas análises estão intimamente conectadas e quando fazemos uma, de fato, também já fizemos a outra. Esta equivalência se deve ao fato de que ambos os perfis tem a mesma inércia total, a mesma dimensionalidade e a mesma decomposição em componentes principais.

A inércia total de uma tabela de contingência quantifica a quantidade de variação presente tanto nos perfis de linha como nos de coluna. Da mesma forma cada linha e cada coluna dão uma contribuição para a inércia total, chamadas respectivamente de inércia de linha e inércia de coluna. A inércia principal de cada linha ou coluna é a projeção deste ponto sobre a componente. Esta projeção representa sua contribuição à inércia da componente principal. Estas contribuições em valores relativos expressam melhor sua representação, que são as seguintes:

- A contribuição de uma linha ou coluna a uma componente, relativamente à inèrcia principal correspondente. Esta é contribuição relativa de uma linha ou coluna à composição desta componente, chamada de CTR, o que permite encontrar quais são os pontos que são os mais importantes na orientação do eixo da componente principal em questão e facilitam a interpretação da componente principal em questão.

- A contribuição de uma linha ou coluna a uma componente, relativamente à inercia do ponto correspondente, indicada como COR. Este valor permite encontrar diagnosticar quão bem representado no mapa está o ponto estudado e dependendo do resultado ele pode ser interpretado com confiança ou com mais cautela. Geometricamente estas quantidades são o quadrado do cosseno entre os pontos e os eixos das componentes principais, e também são interpretados como quadrado das correlações.

- A soma dos quadrados das correlações para um ponto numa solução de baixa dimensão dá uma medida da *qualidade* da representação do ponto neste espaço.

2.2.3 Pontos suplementares

Uma importante característica da CA é a possibilidade de introduzir pontos suplementares no gráfico (perfis de linha ou coluna). Os pontos suplementares são pontos que não contribuem para a orintação dos eixos das componentes principais mas que sua contribuições relativas são calculadas na CA.

Em síntese, pontos suplementares são aqueles que tem uma posição no espaço dos perfis, porém, não tem massa. Na prática são linhas ou colunas adicionais na tabela de contingência, que não exercem nenhuma influência nas componentes principais mas podem ser projetados no mapa de duas (três) dimensões. São usados para analisar novas linhas ou colunas agregadas à tabela sem que elas tenham influência na formação das componentes principais ou principalmente para suprimir a influência de outliers, que são pontos de pouca massa e altos valores de CTR e COR. Eles aparecem na periferia do mapa longe de todos os outros pontos (4).

2.2.4 Gráficos ou mapas

Os resultados da CA são gráficos ou mapas em duas (ou três) dimensões que representam as configurações da projeções dos pontos de perfis de linha e perfis de coluna. Em geral os gráficos são feitos usando como coordenadas as duas mais importantes componentes principais. É costume projetar tanto os perfis de linha quanto de colunas num só mapa apesar de que não se pode interpretar distâncias entre linhas e colunas, somente entre perfis de linha ou entre perfis de coluna. Se um ponto de linha está perto de um de coluna isto pode ser interpretado como uma certa relação entre uma categoria de uma variável (linha) e uma categoria da outra variável (coluna), mas não muito mais do que isto. Os pontos próximos da origem são aqueles que contribuem pouco para as inércias principais e são praticamente idênticos ao perfil médio. Os pontos de linha (coluna) situados próximos tem perfis similares. Proximidade de pontos em duas dimensões pode desaparecer quando se agrega uma terceira dimensão ao gráfico.

É importante salientar que são duas as formas pelas quais um ponto pode dar uma alta contribuição a uma inércia principal:

- Quando está localizado a uma grande distância da origem do eixo da componente principal.
- Quando tem uma grande massa, mesmo se a distância da origem for pequena.

Por esta razão é sempre necessário checar os resultados numéricos da CA, ou seja a massa, a CTR e a COR.

2.3 Análise de correspondência múltipla

A análise de correspondência múltipla (MCA, multiple corespondence analysis) é a CA aplicada à análise de três ou mais variáveis de uma só vez, quando estamos interessados na associação entre estas variáveis, ou seja, em quão forte e de qual maneira estas variáveis estão inter-relacionadas. Vale lembrar que a tabela de

contingência é o resultado de uma tabulação cruzada de duas variáveis, cada uma com suas categorias, na qual uma variável é representada pelas linhas da tabela e a outra pelas colunas. No caso da MCA, as variáveis são apresentadas como colunas e os objetos nos quais as variáveis são medidas na linhas. As variáveis são codificadas com variáveis simuladas (dummy variables) 0 ou 1 numa matriz indicadora $\mathbf{Z}$ que tem tantas colunas como o total de categorias de todas as variáveis e tantas linhas quanto o número total de objetos (ou casos). Os valores em cada linha são $0s$ com exceção dos $1s$ que indicam as categorias para cada variável correspondente à resposta de cada caso, portanto todas as linhas tem a mesma massa, pois todas tem a mesma quantidade de $1s$.

Uma outra alternativa de codificação destes dados é a *matriz Burt*, chamada $\mathbf{B}$ que é relacionada a $\mathbf{Z}$ pela fórmula $\mathbf{B} = \mathbf{Z}^T\mathbf{Z}$ que é uma matriz simétrica das categorias formada por todas as tabelas de contingências de todas as variáveis duas a duas inclusive os bolcos diagonais que são as tabulações de uma variável com ela mesma.

MCA é simplesmente a aplicação do algoritmo CA sobre $\mathbf{Z}$ ou sobre $\mathbf{B}$. As duas são quase equivalentes (4).

3 Resultados

No vestibular da UESC de 2008 participaram 13662 candidatos que responderam ao questionário sócio-cultural citado acima. As respostas a este questionário constituem uma tabela que pode ser facilmente codificada no formato dummy 0 ou 1, ou seja, cada pergunta é uma variável e as respostas são suas categorias. A principio foram realizadas três MCA análises, uma para cada tipo de informação ; escolares, econômicas e culturais.

O software usado é o pacote ca, (5), que está no ambiente computacional estatístico R, (6).

3.1 Informações escolares

É importante salientar que, devido à dificuldade em se analisar um grande número de variáveis e suas categorias o que obscurece demasiadamente o gráfico, devemos sempre selecionar um subconjunto informativo e que, ao nosso ver, contém praticamente toda a informação buscada. Por esta razão, das 17 questões contidas nas informações escolares foram selecionadas 8. Além disto, das questões selecionadas foram mais uma vez selecionadas as respostas (ou categorias) que tiveram uma massa mínima, maior do que 10. A massa de uma categoria está relacionada com sua frequência total, ou seja, quando esta frequência é baixa e categoria pouco influencia na formação das componentes principais e portanto na explicação da variação.

As questões e as respostas selecionadas estão na Tabela 1

O resultado da análise de correspondência múltipla estão na Tabela 2

Tabela 1: Questões e respostas selecionadas, informações escolares. Vestibular UESC, 2008

Questão	Categoria	Código	%
E1-Ano em que concluiu o Ensino Médio.	2007	E1A	30
	2006	E1B	21
	2005	E1C	13
	1991 a 2001	E1G	14
E2-Período em que cursou o Ensino Médio.	Todo diurno	E2A	70
	Maior parte diurno	E2B	14
	Todo noturno	E2C	11
E3-tipo de estabelecimento em que cursou o Ensino Médio.	Todo na escola pública	E3C	60
	Todo na escola particular	E3D	30
E4-Modalidade de Ensino Médio que concluiu.	Científico ou Formação Geral	E4A	83
E5-Freqüentou algum curso pré-vestibular?	Não	E5A	46
	Sim, particular	E5B	34
	Sim, programas populares	E5C	20
E6-Quantos vestibulares já prestou na UESC.	Nenhum	E6A	51
	Um	E6B	24
	Dois	E6C	14
E7-Grau de escolaridade do pai.	Ensino fundamental incompleto	E7B	30
	Ensino médio completo	E7E	29
	Superior completo	E7G	13
E8-Grau de escolaridade da mãe.	Ensino fundamental incompleto	E8B	23
	Ensino médio completo	E8E	34
	Superior incompleto	E8F	9
	Superior completo	E8G	16

- mass é proporcional à quantidade de candidatos que responderam ao item A,B... da questão do número marcado; exemplo E1A resposta A à questão E1
- inr é a inércia da pergunta
- qlt é a qualidade da representação bidimensional
- k é a coordenada de cada categoria na componente principal k.
- cor é a proporção da inércia explicada pela componente principal k. cor1 + cor2 = qlt
- ctr é a contribuição da pergunta à inércia da CP.

As quantidades na tabela estão multiplicadas por 1000 (isto é, as coordenadas e as massas, enquanto que para cor e ctr significa que estão representados em permills ($^{0}/_{00}$)

Tabela 2: Resultado da MCA, Informações escolares. Vestibular UESC, 2008

	name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
1	E1A	46	993	133	-357	402	81	432	591	323
2	E1B	32	380	19	-22	7	0	-158	373	30
3	E1C	19	845	22	111	95	3	-311	750	69
7	E1G	18	880	66	562	786	79	-194	94	25
9	E2A	104	964	11	-100	882	14	-31	82	4
10	E2B	19	896	5	163	874	7	26	22	0
11	E2C	14	924	55	635	910	76	79	14	3
15	E3C	83	986	77	314	963	112	48	23	7
16	E3D	46	967	79	-419	923	111	-91	43	14
17	E4A	122	751	4	-51	718	4	-11	33	1
21	E5A	65	856	23	-34	29	1	181	826	80
22	E5B	51	968	44	-117	145	10	-280	823	149
23	E5C	28	979	59	473	967	87	52	12	3
24	E6A	74	993	33	-72	103	5	210	889	122
25	E6B	35	716	20	68	72	2	-203	644	53
26	E6C	20	752	26	153	160	6	-294	592	64
48	E7B	43	931	74	411	899	101	78	32	10
51	E7E	42	647	21	-177	559	18	-70	88	8
53	E7G	19	811	51	-479	777	60	-101	34	7
55	E8B	33	941	92	527	906	126	103	35	13
58	E8E	50	526	17	-120	389	10	-71	137	9
59	E8F	13	969	12	-314	948	18	-47	21	1
60	E8G	23	843	54	-456	820	67	-76	23	5

As componentes principais foram formadas pelas seguintes combinações:

PC1 (k=1) E1G+E2C+E3C+E5C+E7B+E8B-E3D-E7G-E8F-E8G

PC2 (k=2) E1A+E5A+E6A-E1B-E1C-E5B-E6B-E6C

As duas primeiras componentes principais explicam 90% da inércia total.

Analisando a Figura 1 podemos ver os seguintes resultados:

O cículo inferior congrega os pontos E6C, E1C, E6B, que estão pertos de E1B e E5B. Isto significa que quem conclui o ensino médio dois anos antes do vestibular está associado a quem prestou um ou dois vestibulares,e estão também razoavelmente associados a quem concluiu o ensino médio um ano antes do vestibular e a quem frequentou um curso vestibular particular.

O círculo mais à direita congrega os pontos E5C, E2C, E8B, E7B e o ponto E3C um pouco mais afastado à esquerda. Isto significa que quem frequentou cursinho pré-vestibular em programas populares, tem o pai e a mãe com ensino fundamental incompleto e fez o ensino médio todo noturno e todo na escola pública.

O círculo do meio congrega os pontos E8E, E7E, E2A e E4A. Isto significa que quem tem a mãe e o pai com ensino médio completo, fez o ensino médio todo diurno estão

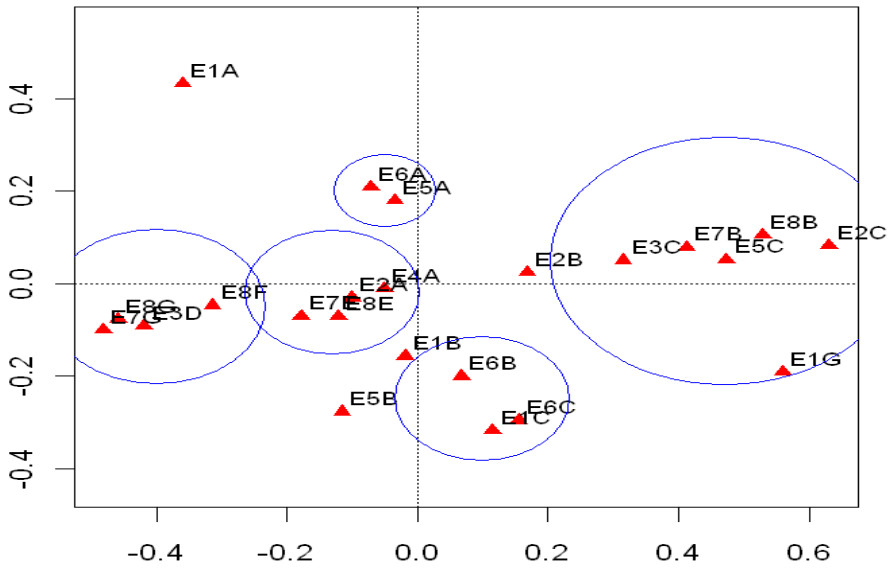


Figura 1: Gráfico de MCA, informações escolares. Vestibular UESC, 2008.

fortemente relacionados e concluiu o ensino médio no científico estão fortemente relacionados.

O círculo mais à esquerda congrega os pontos E3D, E7G, E8G e E8F. Isto significa que quem cursou todo o ensino médio em escola particular, o pai tem curso superior completo e a mãe com curso superior completo e incompleto estão fortemente relacionados.

O pequeno círculo superior relaciona os pontos E5A e E6A, ou seja quem terminou o colegial no ano do vestibular e não fez nenhum cursinho pré-vestibular.

O ponto E1A -concluiu o ensino médio no ano do vestibular está isolado no quarto quadrante e o ponto E1G -concluiu o ensino médio entre 1992 a 2002 está isolado no segundo quadrante quase simétrico ao E1A.

A primeira componente principal (PC1) opõe principalmente o grupo do círculo mais à direita com o mais à esquerda ou seja, opõe quem frequentou cursinho pré-vestibular em programas populares, concluiu o ensino médio entre 1992 e 2002, tem o pai e a mãe com ensino fundamental incompleto e fez o ensino médio todo noturno e todo na escola pública, com quem cursou todo o ensino médio em escola particular, cujo o pai tem curso superior completo e a mãe com curso superior incompleto e completo.

A segunda componente principal (PC2) opõe E6A, E5A e E1A a E1B, E1C, E5B, q18B, q18C, ou seja opõe quem não prestou nenhum vestibular anterior na

UESC, não frequentou nenhum curso pré-vestibular e terminou o ensino médio num colégio particular, e por isto, não precisou fazer o cursinho, com quem frequentou um curso vestibular particular, concluiu o ensino médio um ou dois anos antes do vestibular, prestou um ou dois vestibulares anteriormente.

A PC1 explica a variação de um ponto de vista sócio-econômico, candidatos de classe média (A e B) versus candidatos das classes mais baixas economicamente classes C e D. Isto mostra que mesmo nas variáveis escolares a principal variação se deve a fatores econômicos. A PC2 explica a variação do ponto de vista da competência acadêmica dos candidatos.

3.2 Informações econômicas

O critério de seleção das questões e respectivas respostas foi o mesmo da análise anterior.

As questões e as respostas selecionadas estão na Tabela 3

Tabela 3: Questões e respostas selecionadas, informações econômicas. Vestibular UESC, 2008

Questão	Categoria	Código	%
F1-Ocupação do pai.	Profissional liberal	F1B	12
	Postos médios de Supervisão	F1C	16
	Ocupação de Apoio e Pequenos Negócios	F1D	21
	Aposentado	F1E	14
	Outra	F1J	22
F2-Ocupação da mãe.	Postos médios de Supervisão	F2C	15
	Ocupação de Apoio e Pequenos Negócios	F2D	14
	Dona de Casa	F2G	30
	Outra	F2J	15
F3-Com quantas pessoas você reside?	2 pessoas	F3B	22
	3 pessoas	F3C	27
	4 a 6 pessoas	F3D	44
F4-Renda mensal de toda a família.	De um a dois salários mínimos	F4C	22
	De dois a três salários mínimos	F4D	23
	De três a cinco salários mínimos	F4E	26
	De cinco a dez salários mínimos	F4F	16
F5-Sua participação na renda familiar?	Não trabalho, recebo ajuda financeira da família	F5A	66
	Trabalho e contribuo parcialmente para o sustento da família	F5D	17
F6-Quantas pessoas vivem dessa renda?	Duas	F6B	11
	Três	F6C	21
	Quatro	F6D	31
	Cinco	F6E	19
F7-Pretende trabalhar enquanto faz o curso superior?	Sim, apenas em estágios para treinamento	F7B	35
	Sim, desde o primeiro ano, em tempo parcial	F7D	35
	Sim, desde o primeiro ano, em tempo integral	F7E	19
F8-Imóvel onde sua família reside.	Próprio	F8A	80
	Alugado	F8B	15
F9-Qual o meio de transporte que você mais utiliza?	Ônibus	F9A	71
	Carro próprio ou da família	F9D	17
F10-Se aprovado(a) neste Vestibular, qual será a sua mais provável situação de moradia?	Com a própria família	F10A	61
	Com parentes	F10B	10
	Pensão, pensionato ou república	F10D	21

As componentes principais foram formadas pelas seguintes combinações:

PC1 (k=1) F3B+F4C+F5D+F6B+F7E-F1B-F1C-F2C-F4F-F5A-F7B-F9D-F10D

PC2 (k=2) F3B+F6B+F6C-F3D-F6E

As duas primeiras componentes principais explicam 78% da inércia total.

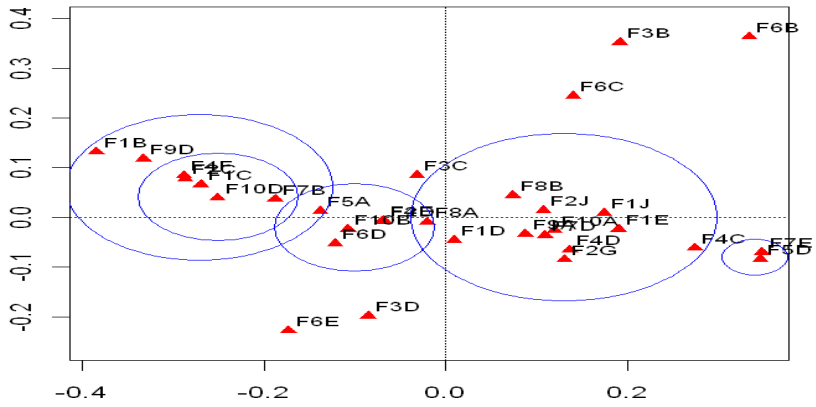


Figura 2: Gráfico de MCA, informações econômicas. Vestibular UESC, 2008.

-.5cm Analisando a Figura 2 podemos ver os seguintes resultados:

O círculo menor mais à esquerda congrega os pontos F4F, F2C, F1C, e F10D um pouco acima à esquerda está o ponto F9D e abaixo à direita o ponto F7B acima, um pouco mis afastado à esquerda o ponto F1B. Isto significa que quem tem renda familiar de cinco a 10 salários, tem a mãe e o pai em postos médios de supervisão e vai morar em pensão, ou seja, não são da região da UESC, estão bem relacionados. Um pouco menos estão os que utilizam carro próprio ou da familia e vão trabalhar somente em estágios de treinamento.

O círculo do meio congrega os pontos F4E, F2D, F10B, F6D e F5A. Isto significa que estão fortemente relacionados quem tem a renda mensal familiar de três a a cinco salários, a mãe com a ocupação de apoio a pequenos negócios, vai viver com parentes, tem quatro pessoas em casa vivendo da renda familiar e o estudante não trabalha e recebe ajuda da familia.

O outro círculo central à esquerda do anterior congrega os pontos F8B, F9A, F7D, F2J, F10A, F2G, F4D, F1J, F1E. Isto significa que estão bem relacionados aqueles cuja família vive em casa alugada, utiliza ônibus, vai trabalhar desde o primeiro ano em tempo parcial, a mãe trabalha em outras ocupações ou é dona de casa e o pai também se ocupa em outras ou é aposentado, vai morar com a família ou seja, é da região da UESC, a família tem renda mensal de dois a três salários.

O círculo mais à direita congrega os pontos F7E, F5D e um pouco mais afastado à esquerda F4C. Isto significa que estão fortemente relacionados quem vai trabalhar

desde o primeiro ano, em tempo integral, trabalha e contribui parcialmente para o sustento da família e, menos relacionado quem tem a renda familiar mais de um até dois salários mínimos.

Os pontos F6E, F3D, F1D, F8A, F3C, F6C, F3B, F6B seguem o sentido da CP2 fortemente negativa e CP1 negativa passando pelo ponto zero até as duas CP´s fortemente positivas. Isto significa cinco pessoas vivendo da renda familiar, reside com 4 a 6 pessoas, o pai tem ocupação de apoio e pequenos negócios, a família vive em casa própria, reside com 3 pessoas, 3 pessoas vivem da renda familiar, reside com 2 pessoas, 2 pessoas vivendo da renda familiar.

A primeira componente principal opõe principalmente o grupo do círculo mais à esquerda com o mais à direita ou seja, opõe quem tem renda familiar de cinco a 10 salários, quem tem a mãe e o pai em postos médios de supervisão, quem vai morar em pensão ou seja não são da região da UESC, quem utiliza carro próprio ou da familia e quem vai trabalhar somente em estágios de treinamento, com quem vai trabalhar desde o primeiro ano em tempo integral, quem trabalha e contribue parcialmente para o sustento da família, quem tem a renda familiar mais de um até dois salários mínimos e quem tem 2 pessoas vivendo da renda familiar. Ou seja a PC1 opõe renda média alta com baixa renda.

A segunda componente principal opõe principalmente quem tem famíla grande vivendo da renda com quem tem famíla pequena vivendo da renda. Provavelmente também opõe renda média e baixa renda familiar.

3.3 Informações culturais

As questões e as respostas selecionadas estão na Tabela 4

Tabela 4: Questões e respostas selecionadas, informações culturais. Vestibular UESC, 2008

Questão	Categoria	Código	%
C1-Excetuando os escolares, quantos livros, em média, você lê por ano?	1 a 2	C1B	24
	3 a 5	C1C	41
	6 a 10	C1D	22
	11 a 20	C1E	7
C2-Qual o meio que você mais utiliza para se manter informado(a)?	Televisão	C2A	47
	Jornal	C2C	10
	Revistas	C2D	10
	Internet	C2F	31
C3-Além dos estudos, com quais atividades você mais ocupa seu tempo?	Leitura	C3A	21
	Cinema, vídeo	C3C	9
	Esportes	C3D	18
	Religião	C3E	12
	Música	C3F	12
	Outros	C3I	21
C4-Acesso à Internet.	Não	C4A	6
	Sim, em casa	C4B	50
	Sim, apenas em outros locais	C4C	45

As componentes principais foram formadas pelas seguintes combinações:

PC1 (k=1) C2A+C2C+C3A+C4A+C4C-C2F-C3D-C3B

PC2 (k=2) C1B+C2A-C1D-C1E-C2D-C3A

As duas primeiras componentes principais explicam 98% da inércia total.

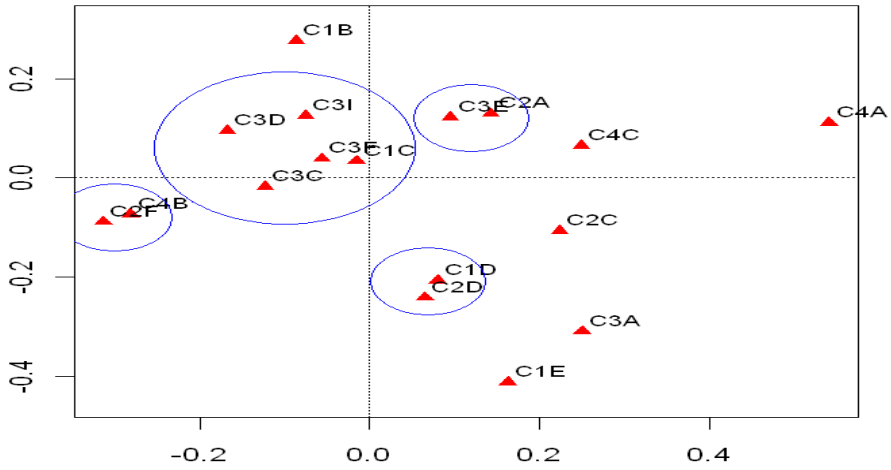


Figura 3: Gráfico de MCA, informações culturais, para os candidatos ao vestibular da UESC/2008.

Analisando a Figura 3 podemos ver os seguintes resultados:

O círculo menor mais à esquerda congrega os pontos C4B e C2F. Isto significa que quem utiliza a internet para se manter informado e tem a internet em casa, estão relacionados.

O círculo do meio congrega os pontos C3D, C3C, C3I, C3F, C1C. Isto significa que estão fortemente relacionados quem ocupa o tempo com música,lê de 3 a 5 livros por ano, ocupa o tempo com esportes, cinema e outros.

O círculo inferior à direita congrega os pontos C1D e C2D. Isto significa que estão fortemente relacionados quem lê de 6 a 10 livros por ano e usa revistas para se manter informado ou seja pessoas acostumados a ler.

O círculo superior à direita congrega os pontos C3E e C2A. Isto significa que estão fortemente relacionados quem utiliza a televisão para se manter informado e ocupa o tempo com outros meios. Provavelmente passa o tempo vendo televisão também.

O sentido que mostra o caminho dos pontos C1E, C3A, C2C, C4C, C1B seguem o sentido da CP2 de negativa a positivo, ou seja quem lê de 11 a 20 livros por ano; ocupa o tempo com leitura; lê jornal (todos CP2 negativa); acessa a internet em outros locais;lê de 1 a 2 livros por ano. exatamente o caminho do mais ao menos intelectualizado.

A relação linear C2A+C2C+C3A+C4A+C4C-C2F-C3D-C4B que forma a CP1 mostra que ela opõe quem se mantem informado pela televisão, pelo jornal, ocupa o tempo com leitura, não tem acesso à internet e tem acesso apenas em outros locais, com quem se mantem informado pela internet, ocupa o tempo com esportes, tem acesso à internet em casa. Basicamente ela opõe classe média com classe mais pobre e o intelectualizado com o mais esportivo.

A relação linear C1B+C2A-C1D-C1E-C2D-C3A que forma a CP2 mostra que ela opõe quem lê de 1 a 2 livros por ano e se mantem informado pela televisão, com quem lê de 6 a 10 e de 11 a 20 livros por ano, se mantém informado por revistas e ocupa o tempo com leitura. Bàsicamente a CP2 opõe o grupo menos intelectualizado com o mais intectualizado.

3.4 Informações escolares e culturais

A Figura 4 é um gráfico simétrico,de CA, no qual tanto os perfis das linhas como das colunas estão mostrados nas coordenadas principais.

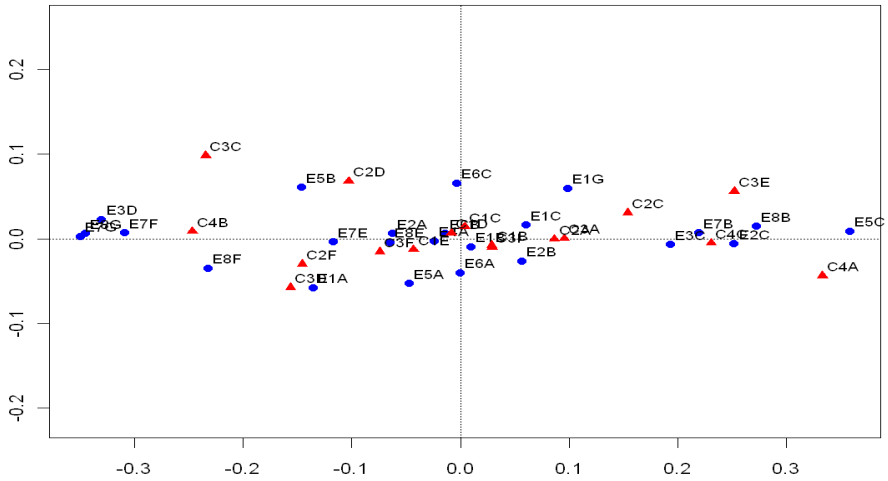


Figura 4: Gráfico de CA, informações escolares-culturais, tabela empilhada. Vestibular UESC, 2008.

Esta tabela de contigência analisada é o que se chama de tabela empilhada.

Primeiramente fazemos a tabulação cruzada de cada pergunta das informações escolares com cada pergunta das informaçõe culturais, de modo que as informções escolares sejam as linhas e as culturais as colunas. Depois empilhamos, para cada pergunta cultural, todas as tabelas seguindo a ordem das informações escolares, uma em cima da outra, e depois juntamos lado a lado para cada pergunta cultural. Neste caso, teoricamente, o total de cada tabela empilhada seria o de todos os candidatos

inscritos, 13662. Se assim fosse, a inércia total seria exatamente a média aritmética das inércias de cada tabela, porém, como há muitos casos de NA´s, os totais na verdade são um pouco diferentes uns dos outros o que aumenta o valor da inércia total em relação à média citada.

No gráfico foram apresentadas somente as mesmas categorias das usadas nas MCA's acima.

A primeira componente principal (principal component PC1) explica 80% da inercia e a segunda PC2 explica 6%. Isto mostra porque os pontos estão praticamente todos concentrados ao redor e ao longo da PC1. No sentido da PC1, as linhas (informações escolares, em preto) tem mais à esquerda os pontos E3D, E8G e E7G (Todo na escola particular,Superior completo pai e mãe).

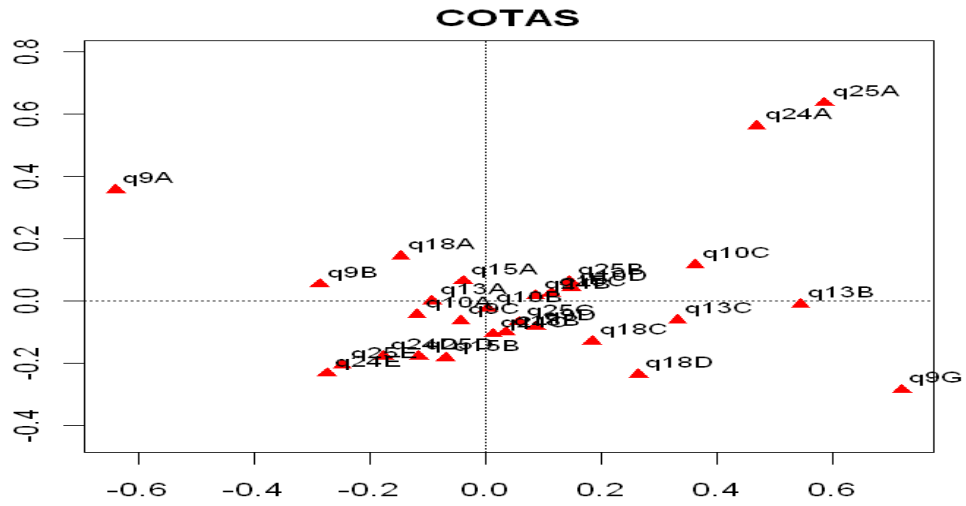
Depois vêm os pontos E5B, E7E, E1A (Sim, (cursinho pre´-vestibular)particular, Ensino médio completo(pai), concluiu ensino médio 2008). Do lado direito da origem, E3C, E2C, E8B, E5C [Todo na escola pública(ensino médio), Todo noturno, Ensino fundamental incompleto(mãe), Sim, programas populares(cursinho pré-vestibular)].

O caminho das informações escolares denota outra vez, que as diferenças principais são a de ordem econômica que se polariza na forma de estudo dos pais , e escolas particulares e públicas e os candidatos mais pobres tendo de cursar as escolas públicas noturnas pois tem de trabalhar. Apesar de que a maioria dos candidatos que cursaram ensino médio em escola pública o fizeram no turno todo diurno E2A, esta resposta não ficou polarizada pois todos os estudantes que fizeram o ensino médio particular também o fizeram todo diurno. a resposta E1A está perto da origem no lado esquerdo. Os candidatos que cursaram no todo noturno, todos estudaram em escolas públicas o que põe q10C junto com E3C.

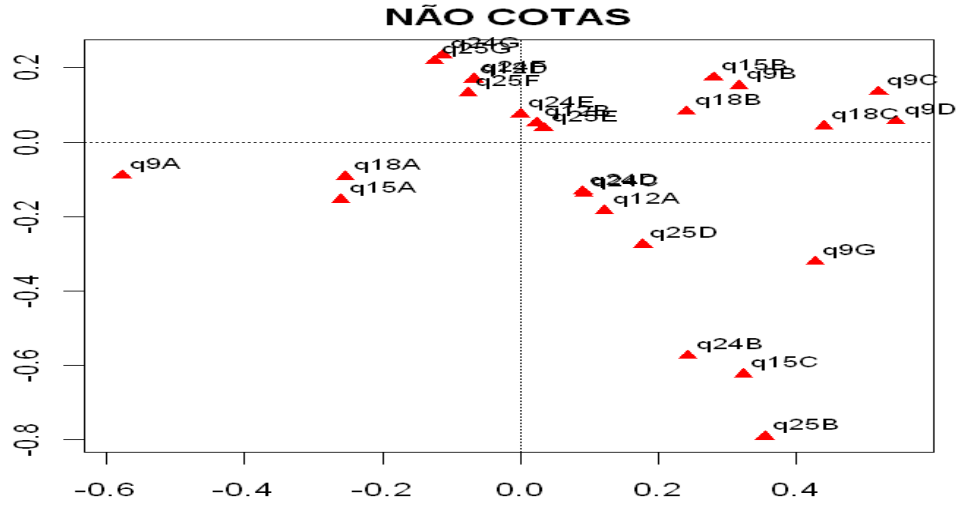
A polarização principal é escola particular - escola pública.

As colunas (informações sócio-culturais, em vermelho) tem mais à esquerda os pontos C4B, C3C, depois C3D, C2F, C2D (Sim, em casa(internet), Cinema-vídeo (atividades), esportes(atividades), internet(informação), revistas(informação)). Do lado direito da origem estão C2C, C4C, C3E, C4A (Jornal(informações), Sim-apenas em outros locais(internet), Religião(atividades), Não(internet)) Da mesma forma a polarização é a econômica de um lado candidatos de classe média (internet em casa, etc.) e os de menor poder aquisitivo (internet em outros lugares ou não tem acesso, etc.)

O importante desta tabela, é que ela mostra que as atividades dos candidatos de classe média estão associadas justamente àqueles que estudaram em escola particular,os pais tem alta escolarização, enquanto que os candidatos das classes menos favorecidas que estudaram em escola pública, pais com baixa escolaridade estão associadas a outras atividades como religião, não tem internet em casa ou somente em lanhouses.Podemos notar que as associações entre estas duas informações estão, no fundo, ligadas ao nível econômico e classe social.



(a) Candidatos cotistas



(b) candidatos não cotistas

Figura 5: Informações escolares; cotistas, não cotistas.

3.5 MCA, comparação cotistas e não-cotistas informações escolares

As duas primeiras componentes principais explicam 86% da inércia dos não cotistas e 76% dos cotistas. Pelo fato dos candidatos estarem diferenciados entre cotistas (todos que estudaram o ensino fundamental e médio em escola pública) e os não cotistas (todos que estudaram o ensino fundamental e médio em escola particular), é de se esperar que o feito econômico seja menos acentuado.

De fato, ao longo da CP1 dos não cotistas está uma oposição entre os candidatos que concluiram o ensino médio no ano do vestibular e os que concluiram nos anos anteriores. Também mostra uma oposição entre os que prestaram o vestibular pela primeira vez e não fizeram curso pré-vestibular e os que prestaram pela segunda ou mais vezes e fizeram curso pré-vestibular. Isto mostra que o fator que descreve a CP1 é a diferença entre os candidatos recém formados do ensino médio e talvez os que estão, em média, mais bem preparados e os outros. Na CP2 a oposição se mostra entre pais com alto nível de instrução versus os de nível mais baixo, o que neste caso aparece a diferença econômica.

Entre os cotistas, podemos notar que os pontos q9A de um lado e q9G do outro são pontos extremos, ou seja, muitos candidatos que terminaram o ensino médio há mais de 8 anos voltaram a tentar o vestibular. Neste caso também CP1 está formada pela oposição entre os candidatos mais bem preparados, que neste caso tem os pais com um maior nível de instrução, com os menos preparados, que neste caso tem os pais com o nívelde instrução mais baixo.

Conclusões

Neste primeiro estudo sobre os candidatos ao vestibular da UESC de 2008, analisamos as respostas do questionário sócio-cultural ao qual todos os candidatos respondem e conseguimos avaliar quais são os fatores principais que formam a estrutura de associação das variáveis e atuam na variabilidade dos dados que gera esta estrutura.

O principal fator é a diferença econômica entre as duas classes sociais, às quais os candidatos pertencem. O segundo fator é a diferença entre os candidatos que estão prestando o vestibular pela primeira vez e os outros que já prestaram vestibular anteriormente. Já quando os candidatos são divididos entre cotistas e não cotistas a diferença econômica se apresenta mais diluida.

A visualização gráfica da CA mostrou ser uma ferramenta poderosa para este tipo de análise exploratória. Numa segunda etapa, depois que obtivermos dados sobre a situação acadêmica dos candidatos cotistas que passaram no vestibular, poderemos usar as principais variáveis definidoras dos fatores mencionados acima para efetuarmos uma análise inferencial e usá-las para encontrar probabilidades de sucesso dos candidatos cotistas durante seu período na universidade.

JELIHOVSCHI E. G.; FERRAZ M. I. F. Analysis of the set of candidates to the State University of Santa Cruz (UESC) entrance exam in 2008, using correspondence analysis. Rev. Bras. Biom., São Paulo, v.28, n.4, p.117-136, 2011.

▪ *ABSTRACT: The aim of this study is the use of the methodology of correspondence analysis (CA) to analyze the responses to the sociocultural questionnaire answered by the candidates of the entrance exam to the State University of Santa Cruz (UESC) in 2008. The resulting two-dimensional graphs show how the variables (questions), are associated. These associations linked to the principal components show which are the factors that lead to inertia or variability of the data. The results show that two factors polarize this inertia, the first and most important is the economic factor, (private school-public school, parents high-low level of schooling), and the second is the competence of the candidates (first exam-earlier ones).*

▪ *KEYWORDS: Correspondence analysis; sociocultural questionnaire.*

Referências

1 Eckart, C.; Young, G. The approximation of one matrix by another of lower rank. Psychometrika, v.3, n.1, p.211–218, 1936.

2 Fisher, R.A. The use of multiple measurements in taxonomic problems. The Annals of Eugenics, n.7, p.179-188, 1936.

3 Murtagh, F. Correspondence Analysis and Data Coding with Java and R. Londres: Chapman Hall/CRC, 2005. 230p.

4 Greenacre, M . Correspondence Analisys in Practice. 2.ed. Boca Raton, FL: Chapman Hall/CRC, 2007. 280p.

5 Greenacre, M; Nenadic, O. ca: Simple, Multiple and Joint Correspondence Analysis. R package version 0.33. http://CRAN.R-project.org/package=ca. 2010.

6 R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing,Vienna: http://www.R-project.org, 2010

7 http://www.unesco.org/webworld/idams/advguide/Chapt6$_5$.*htm*

Recebido em 26.6.2010.

Aprovado após revisão em 04.3.2011.