**Análise Fatorial Múltipla para
Tabelas de Contingência**

Valéria Troncoso Baltar

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO GRAU
DE
MESTRE EM ESTATÍSTICA

Área de Concentração: **Estatística**

Orientadora: Profa. Dra. Lúcia Pereira Barroso

São Paulo, 21 de fevereiro de 2005

Análise Fatorial Múltipla para Tabelas de Contingência

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por Valéria Troncoso Baltar e aprovada pela comissão julgadora.

São Paulo, janeiro de 2005

Banca examinadora

- **Profa. Dra. Lúcia Pereira Barroso (orientadora) – IME-USP**
- **Prof. Dr. Rinaldo Artes – IME-USP**
- **Profa. Dra. Édina Shisue Miazaki - UnB**

Agradecimentos

Eu gostaria de agradecer muita gente, amigos, parentes, colegas, a todos que passaram pela minha vida durante este processo e que o tornaram mais especial. É claro que em primeiro lugar agradeço meus pais e irmãos que me agüentaram estes anos, e que acima de tudo acreditaram que ia dar certo. Agradeço também aos meus tios, tias, primos, primas e avós. Acho que no fundo todo mundo teve alguma influência no meu trabalho.

Um parágrafo eu gostaria de dedicar aos meus amigos especiais do Dante (Neide, Ro, Nei, Ari, Bete, Lu...) e, sobretudo, ao pessoal do LEE. Ao Julio, meu chefinho, por me incentivar e me proporcionar tantas oportunidades. À Camila, ao Valdir, a Helen e ao Clóvis por estarem sempre presentes, por segurarem as pontas na minha ausência, por emprestarem o divã para eu chorar as pitangas, por me ajudarem nos momentos cegos de programação, por me fazerem companhia no cafezinho, por se preocuparem comigo, enfim por tudo. Obrigada. Agradeço também ao Ari que estava sempre atento achando que a qualquer momento eu ia adoecer e cuidou direitinho para que isso não acontecesse.

Um agradecimento também especial para Lúcia, orientadora, amiga e companheira de casa, cerveja e cava.

Agradeço ao Dr. Eduardo Sousa por autorizar minha ida a Barcelona e a todo o pessoal do departamento pessoal que me ajudou com a parte burocrática. Agradeço a Fundação Adib Jatene pelo suporte financeiro e também por proporcionar condições de eu trabalhar na minha própria dissertação durante meu horário de trabalho.

É claro que gostaria de agradecer as minhas amigas queridas e companheiras de IME, a Rê, a Dri e a Diana e dizer que adorei conhecê-las e que se tornaram grandes amigas. A Rê, nem se fala, é praticamente uma irmã, sempre presente. Agradeço a Liginha, a Carol Ribas, a Mari Baltar, a Maricleidas, a Crisão, ao Bá, ao Rui, a Vivian, ao Tetel, a Carlota, a Márcia, ao Eddie, ao Klaudiuzs, ao Pedro, ao Fred, ao Marcos, a Gianne, ao Alcimar e a todos os meus amigos que estavam sempre por aqui, de um jeito, ou de outro.

Agradeço ao apoio dos amigos em Barcelona: a Cris Serra, a Marga, a Núria, à Iéri e a Salsa! Viva a Salsa, salvadora das tensões! Ao Hector Rene e a Monica Becuè por nos receberem tão bem na UPC.

Agraïments a la meva amiga Cris López pels bons moments que vam passar juntes a Barcelona. Amb ella vaig compartir llargues passejades descobrint alhora que tindria per sempre una bona amiga en una bonica ciutat.

I would like to thank Michael Greenacre that have received Lúcia and me with open heart, and have supported us at UPF. I have to thank you, Mike, chiefly for the beautiful days in Greixer, and in Barcelona, for the office (you were so kind in lending it to me every time I have needed), the coffee and the beer, the lunch and for everything. This time with you in Barcelona was really wonderful experience of life.

Agradeço a todo o pessoal do Geopi que me forneceu os dados e também oportunidades de trocar informações nas reuniões na UNICAMP.

E é claro, agradeço o também a UPF por me receber em Barcelona e a Alβan pelo apoio financeiro.

Apoio

Fundação Adib Jatene, Instituto Dante Pazzanese de Cardiologia

Programa **Alβan**, Programa de bolsas de alto nível da União Européia para América Latina, nº de identificação E03M26648BR

Resumo

Atualmente muitos estudos de diferentes iniciativas governamentais e acadêmicas buscam identificar um Sistema Nacional de Inovação em Saúde. Esses estudos tentam caracterizar os setores de atividade econômica envolvidos em Saúde. Nesta dissertação a caracterização baseia-se nos registros da versão 4.1 do Diretório dos Grupos de Pesquisa no Brasil (Gpesq-4), no qual os mesmos registram Saúde como área do conhecimento científico ou setor de atividade econômica em quaisquer de suas linhas de pesquisa. No Gpesq-4, os grupos podem registrar até três áreas do conhecimento e até três setores econômicos. O fato de estas variáveis apresentarem multiplicidade de respostas e, além disso, em uma das variáveis termos a presença de ordem de importância, tornou a análise destes dados um desafio e uma forte motivação para estudar as alternativas disponíveis para este tipo de problema. O objetivo foi buscar alternativas de análise para esta situação e o resultado são quatro formas diferentes para traçar este retrato do sistema, relacionando as áreas e os setores. A primeira solução é uma Análise de Correspondência Simples (ACS) para uma tabela de múltipla resposta com todos os cruzamentos de área e setor, onde o total da tabela supera o tamanho da população. A segunda solução também é uma ACS, porém em uma tabela de contingência na qual temos as combinações observadas entre áreas e entre setores. Uma terceira solução é o uso da Análise de Correspondência Múltipla (ACM) para indicadores de cada área e de cada setor e por último, Análise Fatorial Múltipla para Tabelas de Contingência (AFMTC), na qual levamos em conta a ordem de importância que os respondentes agregaram aos setores econômicos. Nesta dissertação descrevemos essas três técnicas, com maior ênfase para AFMTC; desenvolvemos o programa em R para sua aplicação e apresentamos as análises dos dados do Gpesq-4.

Índice

1. Introdução e Objetivos.....	1
1.1 Introdução.....	1
1.2 Objetivos.....	2
2. Metodologia.....	4
2.1. Análise de Correspondência Simples.....	4
2.2. Análise de Correspondência Múltipla.....	18
2.4. Análise Fatorial Múltipla para Tabelas de Contingência.....	33
3. Aplicação	
Retrato das relações entre Área do Conhecimento Científico e Setor de	
Atividade Econômica: Sistema Nacional de Inovação em	
Saúde.....	50
3.1. Análise de Correspondência Simples da Tabela de Pares de Área-Setor.....	51
3.2. Análise de Correspondência Simples das Combinações de Áreas e Combinações de Setores.....	61
3.3. Análise de Correspondência Múltipla Considerando Todas as Categorias de Áreas e de Setores como Indicadores.....	70
3.4. Análise Fatorial Múltipla Incluindo a Importância Agregada aos Setores...	80
3.5 Conclusões da Aplicação.....	90
4. Conclusões.....	92
5. Anexo.....	93
Anexo A: Rótulos para áreas e setores usadas no Capítulo 3.....	93
Anexo B: Tabela para AFMTC.....	94
Anexo C: Programa em R para análise da Seção 2.1.....	97
Anexo D: Programa em R para análise da Seção 2.2.....	102
Anexo E: Programa em R para análise da Seção 2.3.....	106
6. Referências Bibliográficas.....	114

1. Introdução e Objetivos

1.1 Introdução

A idéia que motivou esta dissertação surgiu do grande interesse em estudar técnicas multivariadas voltadas para dados categorizados. No dia a dia da pesquisa, muitos estudos, principalmente os que visam tipificar uma amostra ou relacionar características, necessitam da utilização de técnicas multivariadas para se ter uma compreensão das relações como um todo.

Este trabalho surgiu especificamente devido à dificuldade de se analisar as relações entre duas variáveis dos registros do Diretório Brasileiro de Grupos de Pesquisa (versão 4.1)¹ coletado em 2000. Este diretório é um projeto desenvolvido desde 1992 pelo CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico. A base de dados contém informações sobre os grupos de pesquisa em atividade no Brasil. Cinco censos foram realizados: 1993, 1995, 1997, 2000 e 2002. Atualmente os dados são atualizados diariamente e podem ser consultados no *site* <http://lattes.cnpq.br/diretorio/>. Cada grupo de pesquisa desta base pode ter mais de uma linha de pesquisa e cada uma delas é um respondente deste censo.

Estes dados apresentam uma grande quantidade de informação sobre quem está fazendo pesquisa, onde e que tipo de pesquisa. Consideramos duas variáveis categorizadas, “setor de atividade econômica” e “área do conhecimento científico” cujos rótulos estão no Anexo A. Cada respondente pode escolher até três opções diferentes para cada uma destas variáveis e para setor foi agregada a ordem de importância. Selecionamos desta base todos os grupos de pesquisa que indicaram Saúde, seja como área de conhecimento científico, seja como setor de atividade econômica, em pelo menos uma de suas linhas, totalizando 12.814 registros.

¹ Dados do Subprojeto “Inserção do CNPq em Sistemas de Inovação - Priorização de Atividades a Partir do Estudo da Área de Saúde” que, por sua vez, integra o Projeto “Capacitação em Inovação e Prospecção Tecnológica: Aspectos Conceituais e Aplicações”, realizado pelo Grupo de Estudos sobre Organização da Pesquisa e da Inovação (GEOPI) do Departamento de Política Científica e Tecnológica da Universidade Estadual de Campinas (DPCT-UNICAMP), com subsídios do Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (Processo institucional n. 680106/99-5). Sob a coordenação dos professores Sergio R. R. Queiroz e Sergio L. M. Salles-Filho, este projeto amparou vários estudos no período de 2000 a 2002.

Analisar as relações entre estas duas variáveis foi o desafio e a motivação para este estudo, pois além de apresentar multiplicidade na escolha das respostas, uma das variáveis apresenta a informação sobre a ordem de importância.

A análise de correspondência é uma técnica usual para visualizar dados categorizados e para analisar associação entre variáveis qualitativas. Neste caso, como os respondentes podiam escolher mais do que uma categoria de resposta para cada variável, a análise de correspondência pode ser aplicada de diversas maneiras. Apresentamos quatro formas diferentes de olhar estes dados, usando análise de correspondência simples (ACS) de duas maneiras, análise de correspondência múltipla (ACM) e análise fatorial múltipla para tabelas de contingência (AFMTC).

Uma vez que ACS e ACM já são técnicas bem conhecidas, a maior contribuição deste trabalho é o detalhamento da AFMTC, pouco encontrado na literatura, e o desenvolvimento de um programa em R para sua aplicação, dado que não existe um aplicativo computacional que o tenha implementado.

Assim mesmo, introduzimos ACS e ACM de uma forma bem geral e didática, consciente de que existem trabalhos muito mais detalhados e mais desenvolvidos (Greenacre, 1984, 1988, 1994; Lebart e Morineau, 1984; Escofier e Pagès, 1992; Van der Heijden, 1987).

Esta dissertação está estruturada em quatro capítulos. O primeiro deles, introduz e descreve os objetivos do estudo, o segundo descreve cada uma das técnicas estatísticas empregadas com a ajuda de um exemplo simples e didático; este está dividido em 3 seções, uma para cada técnica. O Capítulo 3 é a aplicação para a base de dados dos registros do Diretório dos Grupos de Pesquisa no Brasil (Gpesq-4) usando as técnicas descritas no capítulo anterior. E por último, o Capítulo 4 traz um apanhado geral das conclusões deste trabalho.

1.2 Objetivos

O propósito deste estudo consiste primeiramente em estudar diversas técnicas de análise multivariada aplicadas a dados categorizados, buscando idéias para solucionar o problema introduzido na seção anterior. O principal objetivo é um aprofundamento teórico nos estudos de análise de correspondência, e a busca dentro desta técnica para uma adaptação à

dificuldade de se trabalhar com este tipo de dado. Além disso, este projeto também tem como objetivo traçar um retrato das relações entre “áreas do conhecimento científico” e “setores de atividade econômica” para caracterizar o sistema nacional de inovação em saúde.

2. Metodologia

Neste capítulo são descritos três métodos de análise multivariada que permitem analisar relações entre variáveis categorizadas. Os métodos multivariados são adequados ao estudo de fenômenos complexos, em geral descritos por um número grande de variáveis. O primeiro método, ACS, é utilizado quando se dispõe de duas variáveis categorizadas, o segundo método, ACM, quando temos várias variáveis categorizadas e, por último, AFMTC quando o interesse é estudar a relação entre duas variáveis ao longo do tempo, ou ao longo de uma terceira característica. Desde um ponto de vista prático estes métodos permitem a extração bastante rápida da informação de um conjunto de dados. Neste capítulo, os métodos são apresentados enfatizando os aspectos geométricos de tal forma a obter uma visão espacial das variáveis (linhas/colunas). A aplicação das técnicas é ilustrada no estudo da relação entre rendimentos e escolaridade da população ocupada na Região Metropolitana de São Paulo e, em alguns casos, com a inclusão da variável sexo. Os dados obtidos para estes exemplos foram extraídos da Pesquisa Nacional por Amostra de Domicílios (PNAD) referente ao ano de 2002. Escolhemos este exemplo mais simples para facilitar o entendimento das técnicas e no Capítulo 3 apresentamos a análise do problema descrito no Capítulo 1.

2.1 Análise de Correspondência Simples

A Análise de Correspondência Simples (ACS) surgiu do interesse em estudar tabelas de contingência, ou tabelas cruzadas, que são construídas a partir do cruzamento de duas variáveis categorizadas. A ACS descreve e representa graficamente a relação entre as duas variáveis. Esta técnica produz não só as representações em dimensões menores do que a dimensão original da tabela, mas também os coeficientes apropriados que servem de guia na interpretação correta e completa dos resultados da análise. “Esta técnica de análise tem quatro objetivos principais, o primeiro deles é analisar toda a informação contida na tabela, o segundo é representar graficamente a estrutura da tabela, o terceiro é fornecer estatísticas de controle complementar, e o quarto é analisar a estrutura da tabela respeitando o fato desta resumir uma relação simétrica entre os caracteres observados” (Crivisqui, 1993). O autor

refere-se tanto à “toda informação”, quanto a “estrutura da tabela” como sendo as relações entre as variáveis e refere-se ao termo “simétrica” no sentido de que, tanto na linha como na coluna, as variáveis são do mesmo tipo (categorizadas), e as estatísticas de controle ajudam a entender quanto do total da informação está sendo representada pelo novo espaço euclidiano adotado.

Para auxiliar o entendimento desta técnica, usaremos um exemplo simples, porém real, que consiste em analisar a relação entre as variáveis “Rendimento” e “Anos de estudo” referidas à população ocupada e com rendimento positivo da Região Metropolitana de São Paulo (RMSP), em 2002. A base de dados utilizada é da Pesquisa Nacional por Amostra de Domicílios de 2002 (PNAD). Este exemplo foi desenvolvido passo a passo no software R, que pode ser obtido gratuitamente no *site* <http://www.r-project.org/> e os programas que foram usados estão no Anexo C.

Para avaliar esta relação entre escolaridade e rendimento na ocupação principal, realizou-se, preliminarmente, uma adequação das remunerações das pessoas ocupadas em função das horas trabalhadas na semana². Esta adequação visa eliminar o problema de se comparar rendimentos mensais de populações que trabalham um número de horas diferente (no geral mulheres trabalham em média menos horas semanais do que os homens). Essa nova variável foi agrupada em cinco classes de rendimentos (em salários mínimos)³: ‘menos de 1’, ‘de 1 a 3’, ‘de 3 a 5’, ‘de 5 a 10’ e ‘mais de 10’ salários mínimo (SM). A variável escolaridade também foi agrupada de acordo com os anos de estudo nas seguintes quatro classes: ‘menos de 8 anos’ (ensino fundamental incompleto), ‘de 8 a 10 anos’ (ensino fundamental completo ou médio incompleto), ‘de 11 a 14 anos’ (ensino médio completo ou superior incompleto) e ‘15 ou mais anos’ (superior completo). A tabela de contingência, que descreve a distribuição conjunta das variáveis anos de estudo e rendimento, usada neste exemplo, está apresentada na Tabela 2.1.1.

² A formulação desta variável leva em consideração o valor do rendimento mensal no trabalho principal (v_1) e o número de horas trabalhadas na semana neste trabalho (v_2). O valor do salário mínimo em 2002 era de R\$ 200,00, e a jornada completa de trabalho de 44 horas semanais. Assim o cálculo é feito da seguinte maneira: $\text{rendimento} = v_1 / [v_2(4,28)(1,06)]$, onde consideramos 30 dias mensais divididos pelos 7 dias semanais, concluindo que temos 4,28 semanas num mês. O fator 1,06 é o fator que pondera o Salário-hora semanal multiplicado pelas 4,28 semanas $((200/44)4,28)$.

³ A categorização da variável rendimento é exatamente ‘menos de 1 SM’, ‘de 1 a 2,99 SM’, ‘de 3 a 4,99 SM’, ‘de 5 a 9,99 SM’ e ‘mais de 10 SM’. São usadas no decorrer do texto os rótulos simplificados ‘menos de 1’, ‘de 1 a 3’, ‘de 3 a 5’, ‘de 5 a 10’ e ‘mais de 10’, respectivamente.

Tabela 2.1.1: Tabela de contingência da distribuição conjunta das variáveis “anos de estudo” e “rendimento”, para a população ocupada na RMSP em 2002 (matriz N).

Anos de Estudo	Rendimentos em Salários Mínimos					Total
	Menos de 1	De 1 a 3	De 3 a 5	De 5 a 10	Mais de 10	
Menos de 8 anos	277.733	1.441.023	380.847	137.202	36.587	2.273.392
De 8 a 10 anos	120.565	843.966	301.011	157.988	39.916	1.463.446
De 11 a 14 anos	91.466	1.166.632	598.701	538.024	199.572	2.594.395
15 ou mais anos	8.316	88.974	162.147	385.833	571.281	1.216.551
Total	498.080	3.540.595	1.442.706	1.219.047	847.356	7.547.784

Fonte: IBGE-PNAD, 2002

Na Análise de Correspondências Simples (ACS) a distribuição conjunta das variáveis “anos de estudo” e “rendimento”, descrita na Tabela 2.1.1, pode ser representada por uma matriz N de dimensão I (número de linhas) x J (número de colunas) formada por n observações (pessoas ocupadas). Usualmente no lugar das frequências absolutas costumam-se usar tabelas com as frequências relativas. Então, se N é a matriz de dimensão $I \times J$ com total geral n , consideramos $P = (1/n)N$ a matriz de correspondência, com total geral 1. A linha dos totais que fornece a distribuição da variável rendimento e a coluna dos totais que fornece a distribuição da variável anos de estudo correspondem às distribuições marginais (somatórios por colunas e por linhas respectivamente). Os vetores de marginais linha (l) e coluna (c) expressam as somas por coluna e linha respectivamente da matriz P . Observe que, no nosso exemplo, $I = 4$ e $J = 5$, com $n = 7.547.784$.

Tabela 2.1.2: Distribuição conjunta das variáveis ‘anos de estudo’ e ‘rendimentos’ na RMSP em 2002 (matriz P), vetores l e c .

Anos de Estudo	Rendimentos em Salários Mínimos					Vetor l
	Menos de 1	De 1 a 3	De 3 a 5	De 5 a 10	Mais de 10	
Menos de 8 anos	0,0368	0,1909	0,0505	0,0182	0,0048	0,3012
De 8 a 10 anos	0,0160	0,1118	0,0399	0,0209	0,0053	0,1939
De 11 a 14 anos	0,0121	0,1546	0,0793	0,0713	0,0264	0,3437
15 ou mais anos	0,0011	0,0118	0,0215	0,0511	0,0757	0,1612
Vetor c	0,0660	0,4691	0,1911	0,1615	0,1123	1

A Tabela 2.1.2 descreve a distribuição conjunta das variáveis (a matriz P) com as frequências relativas que nos indica que há uma concentração de rendimentos na faixa de ‘1 a

3 SM', enquanto, com relação à escolaridade a distribuição é mais uniforme, apresentando proporções mais elevadas nas faixas de 'menos de 8' e 'de 11 a 14 anos' de estudo.

Consideremos as matrizes diagonais, D_l e D_c , com l e c nas suas diagonais, respectivamente. A representação matricial dos perfis linha e coluna são respectivamente:

$$L = D_l^{-1}P = \begin{bmatrix} \tilde{l}_1^T \\ \cdot \\ \cdot \\ \cdot \\ \tilde{l}_J^T \end{bmatrix} \quad \text{e} \quad C = D_c^{-1}P = \begin{bmatrix} \tilde{c}_1^T \\ \cdot \\ \cdot \\ \cdot \\ \tilde{c}_J^T \end{bmatrix}.$$

As linhas de L e de C são os valores das linhas e colunas da matriz N divididos pelas respectivas somas por linha ou coluna.

No exemplo, a partir da matriz N (Tabela 2.1.1), calculamos as matrizes L e C , expostas nas Tabelas 2.1.3 e 2.1.4 respectivamente.

Tabela 2.1.3: Distribuição das frequências relativas em relação aos totais de cada linha (matriz de perfis linha L).

Anos de Estudo	Rendimentos em Salário Mínimo					Total
	Menos de 1	De 1 a 3	De 3 a 5	De 5 a 10	Mais de 10	
Menos de 8 anos	0,1222	0,6339	0,1675	0,0604	0,0160	1,00
De 8 a 10 anos	0,0823	0,5767	0,2057	0,1080	0,0273	1,00
De 11 a 14 anos	0,0353	0,4497	0,2308	0,2073	0,0769	1,00
15 ou mais anos	0,0068	0,0731	0,1333	0,3172	0,4696	1,00
Vetor c	0,0660	0,4691	0,1911	0,1615	0,1123	1,00

Notamos que a matriz L (Tabela 2.1.3) de perfis linha já nos indica que existe uma relação entre as duas variáveis: à medida que os anos de estudos vão aumentando, as frequências relativas de rendimento tendem a se concentrar nas faixas mais elevadas.

Da mesma forma que é possível observar a relação entre rendimento e escolaridade na matriz L , na matriz C (Tabela 2.1.4) podemos também verificar que à medida que se consideram faixas de maior rendimento a escolaridade também é superior.

Tabela 2.1.4: Distribuição das frequências relativas em relação aos totais de cada coluna (matriz de perfis coluna C).

Anos de Estudo	Rendimentos em Salário Mínimo					Vetor l
	Menos de 1	De 1 a 3	De 3 a 5	De 5 a 10	Mais de 10	
Menos de 8 anos	0,5576	0,4070	0,2640	0,1125	0,0432	0,3012
De 8 a 10 anos	0,2421	0,2384	0,2086	0,1296	0,0471	0,1939
De 11 a 14 anos	0,1836	0,3295	0,4150	0,4414	0,2355	0,3437
15 ou mais anos	0,0167	0,0251	0,1124	0,3165	0,6742	0,1612
Total	1,00	1,00	1,00	1,00	1,00	1,00

A ACS permite avaliar o grau de associação entre as variáveis comparando os perfis linha e os perfis coluna, bem como utilizar a métrica do χ^2 para testar estatisticamente a independência⁴ (homogeneidade) entre as variáveis. Para analisar a tabela de contingência podemos raciocinar em termos de perfis comparando as categorias (proximidades) de uma mesma variável. Na busca de uma representação geométrica definem-se dois espaços de representação: a nuvem dos perfis-linha e a nuvem dos perfis-coluna.

Os I pontos de perfis linha $\tilde{l}_1, \dots, \tilde{l}_I$, conformam uma nuvem de pontos (nuvem dos perfis-linha) no espaço J -dimensional, formado pelas J colunas; e os J pontos de perfis coluna, $\tilde{c}_1, \dots, \tilde{c}_J$, conformam uma nuvem de pontos no espaço I -dimensional (nuvem perfis-coluna), constituído pelas I linhas da tabela de contingência. Na nuvem perfis-linha definimos como massas os I elementos de l e na nuvem perfis-coluna os J elementos de c . A métrica em ambos casos é dada pela distância euclidiana ponderada, que é a distância χ^2 . Os centróides (centros de gravidade) das nuvens linha e coluna são respectivamente:

$$c = L'l \quad \text{e} \quad l = C'c.$$

A variabilidade total nestes espaços, linhas e colunas, é medida pela inércia (dispersão) total, que é a soma ponderada das distâncias euclidianas ao quadrado, de cada ponto com relação ao seu centróide.

A inércia total (in) dos pontos linha é dada por $in(I) = \text{traço}[D_l(L - 1c^T)D_c^{-1}(L - 1c^T)^T]$ e a inércia total dos pontos coluna é dada por $in(J) = \text{traço}[D_c(C - 1l^T)D_l^{-1}(C - 1l^T)^T]$, onde 1 é o vetor de 1's de dimensão adequada.

⁴ Teste χ^2 sob hipótese nula de independência (Bussab e Morettin, 2002; Pereira, 1999).

A inércia total é a mesma em ambas nuvens, e é igual ao *Qui-Quadrado* médio da matriz N , ou seja, é igual à estatística χ^2 dividida pelo total geral n . Isto é,

$$in(I) = in(J) = \text{traço}[D_l^{-1}(P - lc^T)D_c^{-1}(P - lc^T)^T] = \sum_i \sum_j \frac{(p_{ij} - l_i c_j)^2}{l_i c_j} = \frac{\chi^2}{n},$$

onde p_{ij} é o elemento da i -ésima linha e j -ésima coluna da matriz P , l_i é o i -ésimo elemento do vetor l e c_j é o j -ésimo elemento do vetor c .

Podemos verificar se existe relação entre duas variáveis qualitativas quando em alguma casela da tabela P há um valor de p_{ij} diferente do produto $c_j l_i$. Se p_{ij} for superior ao produto $c_j l_i$, as categorias referentes a i e j têm associação positiva: dizemos que i e j se atraem. E se ocorre o contrário, i e j se associam negativamente e dizemos que há uma repulsa entre essas categorias.

Podemos avaliar a semelhança entre duas linhas ou entre duas colunas usando a distância χ^2 entre elas. A definição desta distância é análoga para linhas e colunas:

Distância χ^2 entre duas linhas:

$$d_{\chi^2}(\text{perfil linha } i, \text{perfil linha } k) = \sum_j \frac{1}{c_j} \left(\frac{p_{ij}}{l_i} - \frac{p_{kj}}{l_k} \right)^2.$$

Distância χ^2 entre duas colunas:

$$d_{\chi^2}(\text{perfil coluna } j, \text{perfil coluna } q) = \sum_i \frac{1}{l_i} \left(\frac{p_{ij}}{c_j} - \frac{p_{iq}}{c_q} \right)^2.$$

A distância entre duas linhas depende somente da diferença entre seus respectivos perfis, sendo uma soma ponderada dos quadrados destas diferenças. A ponderação $1/c_j$ equilibra a influência das colunas sobre a distância entre as linhas, agregando um peso maior a categorias mais raras. Isto ocorre de forma análoga para as colunas.

As Tabelas 2.1.5 e 2.1.6 fornecem as distâncias entre perfis linha e entre perfis coluna bem como suas respectivas distâncias aos perfis médios permitindo avaliar a relação entre as variáveis escolaridade e rendimento.

Podemos observar que a distância entre as linhas vai aumentando à medida que a escolaridade é maior, por exemplo, entre ‘menos de 8 anos’ e ‘de 8 a 10 anos’ a distância χ^2

vale 0,0537 e entre ‘de 8 a 10 anos’ e ‘de 11 a 14 anos’ vale 0,1545 (Tabela 2.1.5). O mesmo ocorre entre as colunas (Tabela 2.1.6). Com relação aos perfis gerais temos que as categorias de escolaridade mais elevada e a renda mais elevada são as mais distantes da média.

Tabela 2.1.5: Distâncias entre perfis linha e entre os perfis linha e o perfil médio linha

	Menos de 8 anos	De 8 a 10 anos	De 11 a 14 anos	15 ou mais anos	Perfil médio Linha (<i>l</i>)
Menos de 8 anos	0	0,0537	0,3745	3,1182	0,2544
De 8 a 10 anos	0,0537	0	0,1545	2,6681	0,1120
De 11 a 14 anos	0,3745	0,1545	0	1,8122	0,0475
15 ou mais anos	3,1182	2,6681	1,8122	0	1,6921

Tabela 2.1.6: Distâncias entre perfis coluna e entre perfis coluna e o perfil médio coluna

	Menos de 1	De 1 a 3	De 3 a 5	De 5 a 10	Mais de 10
Menos de 1	0	0,1377	0,5045	1,4737	3,7646
De 1 a 3	0,1377	0	0,1410	0,9120	3,2676
De 3 a 5	0,5045	0,1410	0	0,3689	2,3483
De 5 a 10	1,4737	0,9120	0,3689	0	0,9681
Mais de 10	3,7646	3,2676	2,3483	0,9681	0
Perfil médio coluna (<i>c</i>)	0,4343	0,1628	0,0353	0,3169	1,9991

A distância χ^2 tem a propriedade de equivalência distribucional: se dois perfis linha são idênticos, estas duas linhas podem ser substituídas por apenas uma (na matriz original N seus valores são somados) sem afetar a geometria do perfil coluna, ou seja, as distâncias entre os perfis coluna não se alteram. Na prática, encontrar duas linhas com perfis idênticos é muito raro, porém em um caso no qual elas sejam quase iguais podemos juntá-las sem que a ACS apresente grandes diferenças. Esta propriedade pode ser muito útil para ajudar na redução do número de categorias de uma determinada variável.

Se a dimensão original da tabela é K , a idéia principal da ACS é reduzir a dimensionalidade para subespaços K^* - dimensionais de nuvens linhas e colunas ($K^* < K$), de maneira a minimizar as distâncias entre os pontos que representam as categorias no espaço K -dimensional e os pontos projetados no espaço de dimensão menor, K^* . Estes sub-espacos são definidos pelos K^* vetores singulares (Mardia *et al.*, 1979) de $P-lc^T$, nas métricas $D_c^{-1/2}$ e $D_l^{-1/2}$, correspondentes aos maiores valores singulares μ_{K^*} , ou seja, os eixos que são gerados

pela Decomposição do Valor Singular (DVS) de $D_l^{-1/2}(P - lc^T)D_c^{-1/2}$ que são responsáveis por maximizar a inércia, garantindo que sejam ortogonais entre si.

Os eixos principais são obtidos pela DVS da matriz P centrada e padronizada, denotada por M :

$$M = D_l^{-1/2}(P - lc^T)D_c^{-1/2} = UD_\mu V^T, \text{ onde } U^T U = V^T V = I,$$

D_μ é a matriz diagonal dos valores singulares (μ_k) de $M^T M$, ordenados de forma decrescente, $\mu_1 \geq \dots \geq \mu_K > 0$, U é a matriz cujas colunas são os autovetores de MM^T e V é a matriz cujas colunas são os autovetores de $M^T M$. Então podemos calcular as coordenadas padrão, X e Y :

$$X = D_l^{-1/2}U \quad \text{e} \quad Y = D_c^{-1/2}V.$$

As coordenadas principais dos perfis linha e coluna para a representação simétrica estão relacionadas às coordenadas padrão X e Y por um simples re-escalonamento. Estas coordenadas são úteis pois linhas e colunas representadas simetricamente em um mesmo mapa facilitam a interpretações das relações.

Coordenadas principais do perfil linha F :

$$F = XD_\mu$$

Coordenadas principais do perfil coluna G :

$$G = YD_\mu.$$

As coordenadas com relação a um subespaço, ótimo, K^* -dimensional, referem-se às K^* primeiras colunas de F e G . As fórmulas de transição são a relação entre F e G .

Transição de linha (F) para coluna (G):

$$GD_\mu = D_c^{-1}P^T F$$

Transição de coluna (G) para linha (F):

$$FD_\mu = D_l^{-1}P^T G.$$

A porcentagem da inércia explicada por um eixo é dada pela razão entre o autovalor associado ao eixo e a soma de todos os autovalores: $\left(\lambda_k / \sum_k \lambda_k\right) 100$.

No exemplo temos a seguinte seqüência de tabelas, ou matrizes, para o cálculo das coordenadas:

Tabela 2.1.7: Matriz D_l

0,3012	0	0	0
0	0,1939	0	0
0	0	0,3437	0
0	0	0	0,1612

Tabela 2.1.8: Matriz D_c

0,0660	0	0	0	0
0	0,4691	0	0	0
0	0	0,1911	0	0
0	0	0	0,1615	0
0	0	0	0	0,1123

Tabela 2.1.9: Matriz para DVS, $M = D_l^{-1/2} (P - lc^T) D_c^{-1/2}$

0,1200	0,1320	-0,0296	-0,1381	-0,1575
0,0281	0,0692	0,0146	-0,0587	-0,1117
-0,0701	-0,0166	0,0531	0,0669	-0,0618
-0,0924	-0,2321	-0,0531	0,1555	0,4281

Tabela 2.1.10: Valores Singulares de $M^T M$, Autovalores de M , porcentagem da inércia explicada por eixo, e porcentagem da inércia explicada acumulada por eixo

Eixo (k)	μ_k	$\lambda_k = \mu_k^2$	%	% Acumulado
1	0,5986	0,3584	92,52	92,52
2	0,1698	0,0288	7,43	99,95
3	0,0133	0,0002	0,05	100,00
4	0,0000	0,0000	0,0	100,00
Inércia total (λ)		0,3874	100	

Tabela 2.1.11: Matriz de autovetores U

-0,4305	0,5937	-0,4012	0,5488
-0,2454	-0,0013	0,8637	0,4403
-0,0075	-0,7516	-0,3022	0,5863
0,8686	0,2875	0,0425	0,4015

Tabela 2.1.12: Matriz de autovetores V

-0,2311	0,5734	-0,4960	0,3292
-0,4599	0,1418	0,1446	0,6074
-0,0624	-0,4289	0,4671	0,5189
0,3482	-0,5156	-0,6641	0,3844
0,7810	0,4487	0,2717	0,3251

Tabela 2.1.13: Coordenadas padrão dos perfis linha (Matriz X)

-0,7845	1,0818	-0,7311	1,0000
-0,5572	-0,0030	1,9614	1,0000
-0,0127	-1,2820	-0,5154	1,0000
2,1634	0,7160	0,1059	1,0000

Tabela 2.1.14: Coordenadas padrão dos perfis coluna (Matriz Y)

-0,8996	2,2322	-1,9310	1,2814
-0,6714	0,2070	0,2111	0,8868
-0,1428	-0,9811	1,0684	1,1868
0,8663	-1,2828	-1,6524	0,9565
2,3310	1,3390	0,8110	0,9704

Tabela 2.1.15: Coordenadas principais dos perfis linha (Matriz F)

-0,4696	0,1837	-0,0098	0,0000
-0,3336	-0,0005	0,0262	0,0000
-0,0076	-0,2177	-0,0069	0,0000
1,2951	0,1216	0,0014	0,0000

Tabela 2.1.16: Coordenadas principais dos perfis coluna (Matriz G)

-0,5385	0,3790	-0,0258	0,0000
-0,4019	0,0351	0,0028	0,0000
-0,0855	-0,1666	0,0143	0,0000
0,5186	-0,2178	-0,0220	0,0000
1,3954	0,2274	0,0108	0,0000

As coordenadas F e G no plano, segundo a inércia, já comentada, devem representar os dados sem muita perda de variabilidade, pois representam 99,9% da inércia. As distâncias entre os pontos podem ser calculadas com a distância euclidiana ao quadrado por meio das coordenadas destas matrizes e os resultados não deverão ser muito diferentes dos resultados das Tabelas 2.1.5 e 2.1.6, mesmo quando não usamos todas as coordenadas. Por exemplo, a

distância entre a origem, que representa o perfil médio, e o ponto da categoria ‘menos de 8 anos’ de estudos pode ser obtida pela seguinte conta:

$$d^2 = (-0,4696 - 0)^2 + (0,1837 - 0)^2 = 0,2543.$$

O valor da distância calculada pelos dados originais foi de 0,2544 (tabela 2.1.5), quase igual à distância calculada sob as coordenadas no primeiro plano, que foi de 0,2543. O mesmo ocorre para todas as demais categorias.

Para auxiliar na interpretação dos eixos, calculamos a decomposição da inércia por eixo (Tabela 2.1.10) e por categoria de variável nos eixos (Tabela 2.1.17). Devido à ortogonalidade entre os eixos, a soma das inércias de uma nuvem de pontos sobre cada eixo é igual a inércia total da nuvem. Cada eixo representa uma parte da relação entre as duas variáveis. O valor da inércia em um eixo significa a importância do eixo para explicar a relação entre as variáveis e pode ser vista como um percentual da inércia total.

A situação mais extrema de dependência é uma matriz simétrica e diagonal na qual uma categoria de uma variável se associa a apenas uma categoria da outra. Neste caso, cada eixo da ACS teria inércia igual a 1. Desta forma, quanto mais próxima de 1 a inércia de um eixo estiver, mais próxima desta situação limite de dependência estará a matriz de dados. A soma da decomposição dessa inércia em cada eixo, por cada uma das variáveis, é igual à inércia por eixo, da Tabela 2.1.10.

Outra informação importante é a contribuição que cada categoria de cada uma das variáveis tem para a inércia de um eixo. Esta contribuição é calculada em percentuais; calcula-se primeiramente a inércia de cada categoria da variável linha, que é dada através do produto $l l F^2$ e também a inércia para cada categoria de coluna, que é $c^T l G^2$. Observe que cada coluna destas matrizes F e G representa um eixo. A soma destes valores de inércia (nas linhas) para cada eixo é exatamente a inércia por eixo da Tabela 2.1.10, então a contribuição para a inércia é a proporção destas inércias por categorias de cada eixo com relação à inércia total do eixo, ou seja, cada caséla de $l l F^2$ e de $c^T l G^2$ é dividida pela soma por linha, que equivale aos λ_k 's, multiplicada por 100.

Observe, na Tabela 2.1.17, que a categoria que mais contribui para a inércia do primeiro eixo é a categoria ‘15 ou mais anos’ da variável de escolaridade, com 75,43% da

inércia do eixo e para rendimento é a categoria ‘mais de 10’ SM com 61% da inércia, o que já era previsto, pois justamente estas categorias são as mais distantes da média.

Tabela 2.1.17: Percentual de inércia por categoria de variável para os três primeiros eixos.

		Eixo 1	Eixo 2	Eixo 3
Escolaridade	Menos de 8 anos	18,54	35,25	16,10
	De 8 a 10 anos	6,02	0,00	74,59
	De 11 a 14 anos	0,01	56,49	9,13
	15 ou mais anos	75,43	8,26	0,18
	Total	100	100	100
Rendimento	Menos de 1	5,34	32,88	24,60
	De 1 a 3	21,15	2,01	2,09
	De 3 a 5	0,39	18,40	21,82
	De 5 a 10	12,12	26,58	44,10
	Mais de 10	61,00	20,13	7,39
	Total	100	100	100

No segundo eixo as categorias ‘de 11 a 14 anos’ e ‘menos de 8 anos’ têm uma alta contribuição para a inércia. Com relação à variável de rendimento, ‘menos de 1’ SM, ‘de 5 a 10’ SM, ‘mais de 10’ SM e ‘de 3 a 5’ SM apresentam boa contribuição para a inércia. Ou seja, a inércia do segundo eixo se deve principalmente às coordenadas destas categorias.

Somente a categoria de ‘8 a 10 anos’ de estudo não apresentou boa contribuição para a inércia no primeiro plano e somente no terceiro eixo ela contribui satisfatoriamente (74,59%).

A Figura 2.1.1 representa as categorias de “anos de estudo” e de “rendimento” de acordo com as coordenadas de F e G para os dois primeiros eixos (duas primeiras colunas), simultaneamente.

Uma coluna j atrai mais uma linha i quanto maior for valor de p_{ij} , e desta forma temos também que as categorias projetadas longe do centro do gráfico chamam a atenção por se distanciarem da média (Escofier e Pagès, 1992).

Figura 2.1.1- Mapa simétrico de ACS no primeiro plano para as variáveis escolaridade e renda

Essa representação em um mesmo gráfico é natural do ponto de vista de que tanto linhas quanto colunas têm a mesma natureza, são categorias. Mas vale ressaltar que neste gráfico simétrico, que representa linhas e colunas através das fórmulas de transição, as comparações entre categorias de uma mesma variável são feitas por meio de distâncias, enquanto a comparação de uma categoria de uma variável com uma de outra variável deve ser feita em termos gerais, ou seja, uma categoria de uma variável deve ser comparada com o conjunto de categorias da outra variável, sem levar em consideração as distâncias entre categorias de variáveis diferentes propriamente ditas.

Na ACS podemos optar por interpretar os mapas conhecidos como mapas assimétricos, que usam as coordenadas padrão de uma variável e as coordenadas principais da outra variável. Nesses mapas, podemos interpretar as distâncias entre categorias de variáveis diferentes.

A interpretação do mapa, Figura 2.1.1, é muito simples. O primeiro comentário é a respeito da forma parabólica da trajetória que as categorias formam; essa forma é muito comum quando a variável analisada é ordinal, como é o caso. Esta forma parabólica é conhecida como efeito Guttman (Escofier e Pagès, 1992; Wolf and Gabler, 1998), e quando há uma dependência perfeita entre as duas variáveis a parábola é também perfeita, simétrica. Nota-se que há uma trajetória crescente do lado esquerdo para o lado direito do mapa, e

podemos observar também que há uma proximidade maior entre os pontos de uma mesma variável da parte esquerda do mapa. A linha tracejada e as categorias em *itálico* representam as projeções das categorias de escolaridade e a linha contínua, as categorias de rendimentos. O aumento dos anos de estudos destes trabalhadores é acompanhado por um aumento de rendimentos, e vice-versa.

O mapa da Figura 2.1.1 indica que há uma similaridade maior entre os anos de estudos ‘menos de 8 anos’ e ‘de 8 a 10 anos’. Note que, estas duas categorias são as mais próximas dentre todas as categorias de escolaridade, podendo ser observada a menor distância na Tabela 2.1.5. Isto indica que existe um perfil de renda mais parecido entre elas e, além disso, o fato destas categorias se encontrarem mais próximas às duas menores faixas de rendimento ‘menos de 1’ e ‘de 1 a 3’ SM indica que estas categorias de escolaridade apresentam uma frequência mais elevada, do que o geral, de rendimentos mais baixos. A categoria ‘menos de 8 anos’ ainda tem uma proximidade maior com ‘menos de 1’ SM do que a categorias seguintes de anos de estudos. A categoria ‘de 11 a 14 anos’ de estudo está muito próxima do centro, indicando que é um perfil parecido com o perfil médio, e a categoria ‘de 3 a 5’ SM também está próxima desta faixa de escolaridade e do centro. Estas proximidades com o centro indicam um distanciamento pequeno com relação ao perfil médio geral. Observa-se que a faixa de escolaridade ‘15 ou mais anos’ de estudo é a que mais se distancia do centro, e também de todas as outras categorias de escolaridade. Este perfil é diferente dos demais por apresentar uma proporção mais elevada do que as demais categorias de “anos de estudos” com “rendimento” ‘mais de 10’ SM, fora isso também apresenta uma atração com rendimentos ‘de 5 a 10’ SM.

Elementos suplementares, que são linhas ou colunas que não contribuem para o cálculo dos eixos, podem ser incluídos na análise. Uma linha suplementar pode ser projetada no mapa com o uso da fórmula de transição de linha para coluna e as colunas suplementares podem ser projetadas através da fórmula de transição de colunas para linhas. Esta forma de projeção é interessante para se ter uma idéia da localização de pontos para os quais julgamos importante usar um peso nulo, como por exemplo, para o caso de não resposta para uma das variáveis.

2.2 Análise de Correspondência Múltipla

Na Seção 2.1 discutimos o caso mais simples e fundamental da Análise de Correspondência, no qual a matriz de dados contém a informação sobre a relação entre duas variáveis qualitativas. Nesta seção estudamos as relações conjuntas entre mais do que duas variáveis.

A Análise de Correspondência Múltipla (ACM), assim como a ACS, tem como objetivo a redução das dimensões dos dados. O pesquisador tem interesse pelas análises simultâneas das relações entre as diversas variáveis e pela comparação de todas as unidades de observação por meio de todas as categorias de respostas observadas. A ACM permite analisar uma amostra de n indivíduos descritos por J variáveis qualitativas. Uma aplicação muito comum é a análise das respostas de um questionário, onde cada pergunta constitui uma variável, e as respostas são as categorias nas quais o indivíduo se classifica, supondo que ele se classifique em apenas uma categoria por variável.

Essa metodologia explora os dados evitando que se faça, o que é muito comum, muitas tabelas de contingência. Muitas vezes busca-se resumir a informação selecionando aquelas que são julgadas interessantes segundo o grau de significância estatística do teste χ^2 sobre todas elas, interpretando erroneamente o p-valor do teste em termos de “força da relação” entre duas variáveis.

A base de dados, representada pela Tabela de Dados Brutos (TDB), Tabela 2.2.1, apresenta os indivíduos nas linhas e as variáveis nas colunas. Na interseção da linha i (indivíduo) com a coluna j (variável) tem-se o valor x_{ij} que representa uma codificação de categoria de resposta. Essa codificação não apresenta propriedades numéricas, então para representar estes dados monta-se uma Tabela Disjunta Completa (TDC). Nesta tabela as linhas representam os indivíduos e as colunas representam as categorias das variáveis. Na interseção do indivíduo i com a coluna q (categoria q) encontramos o valor zero ou o valor um: se igual a um esta categoria foi opção deste indivíduo, caso contrário não. Chamamos esta matriz, Z , de indicadora.

Tabela 2.2.1: Tabela de Dados Brutos (TDB), onde cada coluna apresenta um código de variável.

	Variável 1	Variável 2	...	Variável j	...	Variável J
1	x_{11}	x_{12}		x_{1j}		x_{1J}
2	x_{21}	x_{22}		x_{2j}		x_{2J}
...						
i	x_{i1}	x_{i2}		x_{ij}		x_{iJ}
...						
n	x_{n1}	x_{n2}		x_{nj}		x_{nJ}

A Tabela 2.2.2 representa a matriz indicadora: $Z = [Z_1 Z_2 \dots Z_j \dots Z_J]$.

Tabela 2.2.2: Tabela Disjunta Completa (TDC) ou matriz indicadora (Z), onde cada coluna representa um indicador de categoria.

	Variável I						Variável j				Variável J			
	cat. 1	cat. 2	(...)	cat. q_1	(... q ...)	cat. $q_{j-1}+1$	(...)	cat. q_j	(...)	cat. $q_{J-1}+1$	(...)	cat. Q	Marginal l	
1	0	0		1		0		1		0		1	J	
...														
i					z_{iq}								J	
...														
n													J	
Marginal c	$n_{1\bullet}$	$n_{2\bullet}$		$n_{q_1\bullet}$		$n_{(q_{j-1}+1)\bullet}$		$n_{q_j\bullet}$		$n_{(q_{J-1}+1)\bullet}$		$q_{Q\bullet}$	nJ	

$$z_{iq} = \begin{cases} 0 & \text{se a categoria } q \text{ não estiver presente no indivíduo } i \\ 1 & \text{se a categoria } q \text{ estiver presente no indivíduo } i, \end{cases}$$

onde, q indica a categoria ($q = 1, \dots, Q$), e $Q_j = q_j - q_{j-1}$ indica o número de categorias na variável j .

As Q colunas da matriz Z são variáveis indicadoras das Q categorias das J variáveis observadas, de modo que a soma nas colunas dessa tabela é uma marginal constante e igual a J . A marginal coluna (soma por linhas) de Z corresponde à distribuição de frequências das categorias observadas nos n indivíduos, e a soma dessas marginais evidentemente é igual à nJ .

A TDC (Z) indica o mesmo que a TDB, mas com dimensão muito maior, porém apresenta propriedades numéricas. “A origem da terminologia ‘tabela disjunta completa’ é

a seguinte: o conjunto de valores z_{iq} de um mesmo indivíduo, para as categorias de uma mesma variável é formada pelo valor 1 uma vez (completa) e só uma vez (disjunta)” (Escofier e Pagès, 1992). Esta tabela é então uma hiper-tabela e na maior parte das vezes é impossível manusear uma tabela tão grande, mas é possível construir uma tabela que contém o conjunto de tabelas de contingência entre as variáveis tomadas duas a duas. Esta tabela é chamada de tabela ou matriz de Burt (Benzécri, *et al.* 1976; Lebart *et al.*, 1987; Greenacre, 1984; Crivisqui, 1993) e não é exatamente uma tabela de contingência, mas sim uma justaposição de tabelas de contingência, onde cada indivíduo aparece $J^2 n$ vezes.

Tabela 2.2.3: Tabela ou matriz de Burt (TB), matriz $Z^T Z$

		Variável 1			(…)	Variável j			(…)	Variável J					marginal
		1	(…)	q_1		$q_{j-1}+1$	(…)	q_j		$q_{j-1}+1$	(…)	q'	(…)	Q	
I	1	$n_{1\bullet}$	…	0		.	.	.		.	.	$n_{(1)q'}$	.	.	$(n_{1\bullet})J$
	(…)	.		.						.		.			…
	q_1	0	…	$n_{q_1\bullet}$		.	.	.		.	.	.	.	.	$(n_{q_1\bullet})J$
	(…)				…				…						
j	$q_{j-1}+1$	.				$n_{(q_{j-1}+1)\bullet}$		0		.		.			$(n_{(q_{j-1}+1)\bullet})J$
	(…)	.								.		.			…
	q	.	…	…		0	…	0		.	.	$n_{qq'}$	.	n_{qQ}	$(n_{q\bullet})J$
	(…)	.								.		.			…
	q_j	.				0		$n_{q_j\bullet}$		.		.			$(n_{q_j\bullet})J$
	(…)				…				…						
J	$q_{j-1}+1$	.								$n_{(q_{j-1}+1)\bullet}$	.	.		0	$(n_{(q_{j-1}+1)\bullet})J$
	(…)	.								.	.	.		.	…
	q'	.	…	…		…	…	…		0	…	$n_{q'\bullet}$	…	0	$(n_{q'\bullet})J$
	(…)	.								.		.		.	…
	Q	.								0		.		$n_{Q\bullet}$	$(n_{Q\bullet})J$
marginal		$(n_{1\bullet})J$	…	$(n_{q_1\bullet})J$	…	$(n_{(q_{j-1}+1)\bullet})J$	…	$(n_{q_j\bullet})J$	…	$(n_{(q_{j-1}+1)\bullet})J$	…	$(n_{q'\bullet})J$	…	$(n_{Q\bullet})J$	$J^2 n$

Cada bloco da diagonal da matriz de Burt (Tabela 2.2.3) é o cruzamento de uma variável com ela mesma, dentro do qual os elementos fora da diagonal são nulos. Os elementos que estão na diagonal destes blocos são as frequências simples de cada categoria das variáveis. Analogamente a uma matriz de covariâncias os blocos fora da diagonal mostram as relações entre os pares de variáveis, mas apresentam menos informações que a hiper-tabela, TDC, pois não apresentam a informação por indivíduo.

A grande vantagem da matriz de Burt é a propriedade de escalonamento ótimo (a matriz de Burt é simétrica e quadrada); as coordenadas das categorias no primeiro eixo provêm de uma escala ótima, no sentido que o escore obtido destes valores para as categorias contém a máxima variância. Assim como no caso simples, o interesse é encontrar uma quantificação para valores categorizados de cada variável que tenha máxima discriminação entre os respondentes.

Para estudar as categorias temos duas considerações a serem feitas: ou consideramos uma categoria como uma variável indicadora definida sobre o conjunto dos indivíduos, isto é, uma coluna de Z ; ou consideramos como um grupo de indivíduos dos quais se conhece a distribuição sobre o conjunto de categorias, seja uma linha ou coluna da tabela de Burt. Ou seja, trabalhando com a matriz Z ou com a $Z^T Z$. Para cada um destes casos a noção de semelhança é diferente. No primeiro caso (Z) a semelhança entre duas categorias deve se basear em sua associação mútua: duas categorias se parecem mais, quanto maior é a sua presença ou ausência simultânea em um grande número de indivíduos. As outras modalidades não interferem. O segundo caso, tabela de Burt, é análogo ao uso de tabelas de freqüências, uma linha da tabela de Burt caracteriza a associação da categoria com as categorias de todas as variáveis: duas categorias são semelhantes quanto maior ou menor for sua associação com categorias comuns.

O exemplo que usamos nesta seção para a matriz indicadora com mais de duas variáveis é a análise das mesmas variáveis “anos de estudo” e “rendimento” da seção anterior incluindo a informação sobre o “sexo” do trabalhador ocupado. A variável “sexo” pode assumir duas categorias: “homem” ou “mulher”.

A matriz indicadora deste exemplo é $Z = [Z_1 \ Z_2 \ Z_3]$ onde, Z_1 representa as categorias de “sexo”, Z_2 representa as categorias de “anos de estudo” e Z_3 de “rendimento”. Esta matriz indicadora tem $n = 7.547.784$ indivíduos, $J = 3$ variáveis e $Q = 2+4+5 = 11$ categorias de variáveis. Devido ao fato desta matriz Z ser muito grande, optamos por apresentar diretamente a tabela de Burt e não a tabela indicadora (Tabela 2.2.4).

Intuitivamente poderíamos pensar em analisar a matriz $Z_1^T [Z_2 \ Z_3]$, mas estaríamos ignorando as relações entre Z_2 e Z_3 . Imaginando que usualmente, principalmente analisando respostas de um questionário, o número de variáveis é bem mais elevado e a relação entre todas as questões é importante, esta análise não seria suficiente.

Consideramos a matriz indicadora geral $Z = [Z_1 \dots Z_J]$, com n linhas e Q colunas onde a j -ésima matriz, Z_j , tem $Q_j - Q_{j-1}$ categorias correspondentes à variável j . Ao longo de Z , temos Jn valores 1, n em cada sub-matriz Z_j . Cada linha de Z_j soma 1 e de Z soma J , enquanto que as colunas somam $1^T Z$, que é a soma do número de vezes que cada categoria aparece em Z . A partir daqui usaremos q quando nos referirmos a uma coluna de Z e q_j quando nos referirmos a uma coluna de Z_j .

O vetor de massas das linhas de Z e Z_j são respectivamente:

$$l^Z = \left(\frac{1}{nJ} \right) \quad \text{e} \quad l_j^Z = \left(\frac{1}{n} \right).$$

O vetor de massas das colunas de Z e de Z_j são definidos respectivamente por:

$$c^Z = \left(\frac{1}{Jn} Z^T 1 \right) \quad \text{e} \quad c_j^Z = \left(\frac{1}{Jn} Z_j^T 1 \right).$$

A soma das massas das colunas de Z_j é $1/J$, para $j = 1, \dots, J$. Cada variável recebe a mesma massa que é distribuída pelas categorias da variável de acordo com as frequências de cada uma delas.

O centróide das colunas de Z_j é a origem de todos os perfis coluna, cada sub-nuvem de categorias está centrada na origem.

A inércia total dos perfis coluna de Z é:

$$in(J) = \frac{Q}{J} - 1$$

e a inércia de perfil coluna de Z_j é:

$$in(Q_j) = \frac{Q_j - 1}{J},$$

onde Q_j é o número de categorias da variável j .

A contribuição para a inércia de uma variável discreta aumenta linearmente com o número de categorias.

A inércia de uma categoria particular q é dada por:

$$in(q) = \frac{1}{J} - c_q^Z,$$

e a contribuição para a inércia de uma categoria aumenta à medida que a resposta a esta categoria decresce, com limite em $1/J$ (Greenacre, 1984).

O número de dimensões não triviais com inércia positiva é de no máximo $Q-J$ (soma do número de categorias menos um de cada variável). Os perfis linha são médias ponderadas com mesmo peso dos perfis coluna, re-escalados pela inversa da raiz da inércia principal do eixo, da mesma maneira que na ACS.

A matriz de correspondência para análise de Z , é: $P^Z = (1/nJ)Z$.

Para simplificar a análise, usaremos o exemplo da PNAD para análise da matriz de Burt, que apresenta os mesmos resultados da análise da matriz indicadora, porém sem representar a informação para cada indivíduo. A matriz de Burt deste exemplo é dada por:

$$Z^T Z = \begin{bmatrix} Z_1^T Z_1 & Z_1^T Z_2 & Z_1^T Z_3 \\ Z_2^T Z_1 & Z_2^T Z_2 & Z_2^T Z_3 \\ Z_3^T Z_1 & Z_3^T Z_2 & Z_3^T Z_3 \end{bmatrix}.$$

Cada sub-matriz $Z_j^T Z_{j'}$, onde $j \neq j'$ é uma tabela de contingência que relaciona a variável j com a variável j' para os n indivíduos.

Como a matriz de Burt é simétrica, positiva e semi-definida os resultados das coordenadas de linhas e colunas da análise de correspondências são idênticos. Greenacre (1984) demonstra que as coordenadas de linhas/colunas da análise de $Z^T Z$ são idênticas aos das colunas da análise de Z . A única diferença está no valor da inércia e conseqüentemente na escala da coordenada: $\lambda^B = (\lambda^Z)^2$, onde λ^Z é a inércia da matriz Z e λ^B é a inércia da matriz de Burt.

O valor da inércia total da matriz Z do exemplo vale 2,667 (11/3-1), que é o resultado direto da soma dos valores singulares da decomposição espectral com base na matriz de Burt.

A ACM pode ser feita tanto na matriz de Burt ($Z^T Z$) como na matriz Z , mas como a ACM não é o objetivo principal deste trabalho, usaremos somente a análise da matriz de Burt. Um melhor detalhamento da análise de Z pode ser encontrado em Greenacre (1984, 1993).

Tabela 2.2.4: Tabela de Burt para as variáveis “sexo”, “anos de estudo” e “rendimento”.

	Homem	Mulher	Menos de 8	De 8 a 10	De 11 a 14	15 ou mais	Menos de 1	De 1 a 3	De 3 a 5	De 5 a 10	Mais de 10	Marginal
Homem	4397087	0	1448516	907181	1449345	592045	229501	1950732	923815	738403	554636	13191261
Mulher	0	3150697	824876	556265	1145050	624506	268579	1589863	518891	480644	292720	9452091
Menos de 8	1448516	824876	2273392	0	0	0	277733	1441023	380847	137202	36587	6820176
De 8 a 10	907181	556265	0	1463446	0	0	120565	843966	301011	157988	39916	4390338
De 11 a 14	1449345	1145050	0	0	2594395	0	91466	1166632	598701	538024	199572	7783185
15 ou mais	592045	624506	0	0	0	1216551	8316	88974	162147	385833	571281	3649653
Menos de 1	229501	268579	277733	120565	91466	8316	498080	0	0	0	0	1494240
De 1 a 3	1950732	1589863	1441023	843966	1166632	88974	0	3540595	0	0	0	10621785
De 3 a 5	923815	518891	380847	301011	598701	162147	0	0	1442706	0	0	4328118
De 5 a 10	738403	480644	137202	157988	538024	385833	0	0	0	1219047	0	3657141
Mais de 10	554636	292720	36587	39916	199572	571281	0	0	0	0	847356	2542068
Marginal	13191261	9452091	6820176	4390338	7783185	3649653	1494240	10621785	4328118	3657141	2542068	67930056

Fonte: IBGE-PNAD, 2002.

Se q e q' são categorias de uma mesma variável $n_{qq'}$ será zero, exceto para o caso em que $q=q'$, que será igual a $n_{q\cdot}$.

Já que as coordenadas das linhas e das colunas da análise da matriz de Burt são idênticas, usaremos somente uma notação para as coordenadas padrão, Γ^B , e para as coordenadas principais, G^B .

Tabela 2.2.5: Tabela de freqüências relativas, com relação ao total geral J^2n (matriz P)

	Homem	Mulher	Menos de 8	De 8 a 10	De 11 a 14	15 ou mais	Menos de 1	De 1 a 3	De 3 a 5	De 5 a 10	Mais de 10	I^B
Homem	0,0647	0,0000	0,0213	0,0134	0,0213	0,0087	0,0034	0,0287	0,0136	0,0109	0,0082	0,1942
Mulher	0,0000	0,0464	0,0121	0,0082	0,0169	0,0092	0,0040	0,0234	0,0076	0,0071	0,0043	0,1391
Menos de 8	0,0213	0,0121	0,0335	0,0000	0,0000	0,0000	0,0041	0,0212	0,0056	0,0020	0,0005	0,1004
De 8 a 10	0,0134	0,0082	0,0000	0,0215	0,0000	0,0000	0,0018	0,0124	0,0044	0,0023	0,0006	0,0646
De 11 a 14	0,0213	0,0169	0,0000	0,0000	0,0382	0,0000	0,0013	0,0172	0,0088	0,0079	0,0029	0,1146
15 ou mais	0,0087	0,0092	0,0000	0,0000	0,0000	0,0179	0,0001	0,0013	0,0024	0,0057	0,0084	0,0537
Menos de 1	0,0034	0,0040	0,0041	0,0018	0,0013	0,0001	0,0073	0,0000	0,0000	0,0000	0,0000	0,0220
De 1 a 3	0,0287	0,0234	0,0212	0,0124	0,0172	0,0013	0,0000	0,0521	0,0000	0,0000	0,0000	0,1564
De 3 a 5	0,0136	0,0076	0,0056	0,0044	0,0088	0,0024	0,0000	0,0000	0,0212	0,0000	0,0000	0,0637
De 5 a 10	0,0109	0,0071	0,0020	0,0023	0,0079	0,0057	0,0000	0,0000	0,0000	0,0179	0,0000	0,0538
Mais de 10	0,0082	0,0043	0,0005	0,0006	0,0029	0,0084	0,0000	0,0000	0,0000	0,0000	0,0125	0,0374
c^B	0,1942	0,1391	0,1004	0,0646	0,1146	0,0537	0,0220	0,1564	0,0637	0,0538	0,0374	1,0000

A seqüência de cálculos nesta análise é a mesma da seção anterior, o que muda é a matriz N , que agora é a matriz de Burt, Tabela 2.2.4. A soma de cada linha de cada sub-

matriz $Z_j^T Z_{j'}$, para todos os j e j' é igual ao vetor nJc_j^Z e para a matriz $Z^T Z$ é $nJ^2 c^Z$, então a matriz de perfis linha (L^B) de $Z^T Z$ é:

$$L^B = \left(\frac{1}{J} \right) \begin{bmatrix} I & L_{12} & \dots & L_{1J} \\ L_{21} & I & \dots & L_{2J} \\ \dots & \dots & \dots & \dots \\ L_{J1} & \dots & \dots & I \end{bmatrix}, \text{ onde } L_{jj'} \text{ é a matriz de perfis linha da tabela de}$$

contingência $Z_j^T Z_{j'}$.

Só há uma transição, das coordenadas de colunas para elas mesmas, na coordenada padrão, que é:

$\Gamma^B = L^B \Gamma^B (D_\lambda^B)^{-1/2}$, onde a matriz D_λ^B é a matriz diagonal com os valores da inércia da matriz de Burt.

Γ^B pode ser dividida em J grupos: Γ_j^B ($j=1, \dots, J$).

$$\Gamma_j^B = \left(\frac{1}{J} \right) \left(\Gamma_j^B + \sum_{j \neq j'} L_{jj'} \Gamma_{j'}^B \right) (D_\lambda^B)^{-1/2}$$

Tabela 2.2.6: Tabela de perfis linha, da Tabela de Burt, chamada de matriz L .

	Homem	Mulher	Menos de 8	De 8 a 10	De 11 a 14	15 ou mais	Menos de 1	De 1 a 3	De 3 a 5	De 5 a 10	Mais de 10	Total
Homem	1,0000	0,0000	0,3294	0,2063	0,3296	0,1346	0,0522	0,4436	0,2101	0,1679	0,1261	3
Mulher	0,0000	1,0000	0,2618	0,1766	0,3634	0,1982	0,0852	0,5046	0,1647	0,1526	0,0929	3
Menos de 8	0,6372	0,3628	1,0000	0,0000	0,0000	0,0000	0,1222	0,6339	0,1675	0,0604	0,0161	3
De 8 a 10	0,6199	0,3801	0,0000	1,0000	0,0000	0,0000	0,0824	0,5767	0,2057	0,1080	0,0273	3
De 11 a 14	0,5586	0,4414	0,0000	0,0000	1,0000	0,0000	0,0353	0,4497	0,2308	0,2074	0,0769	3
15 ou mais	0,4867	0,5133	0,0000	0,0000	0,0000	1,0000	0,0068	0,0731	0,1333	0,3172	0,4696	3
Menos de 1	0,4608	0,5392	0,5576	0,2421	0,1836	0,0167	1,0000	0,0000	0,0000	0,0000	0,0000	3
De 1 a 3	0,5510	0,4490	0,4070	0,2384	0,3295	0,0251	0,0000	1,0000	0,0000	0,0000	0,0000	3
De 3 a 5	0,6403	0,3597	0,2640	0,2086	0,4150	0,1124	0,0000	0,0000	1,0000	0,0000	0,0000	3
De 5 a 10	0,6057	0,3943	0,1125	0,1296	0,4413	0,3165	0,0000	0,0000	0,0000	1,0000	0,0000	3
Mais de 10	0,6545	0,3455	0,0432	0,0471	0,2355	0,6742	0,0000	0,0000	0,0000	0,0000	1,0000	3
c^B	0,1942	0,1391	0,1004	0,0646	0,1146	0,0537	0,0220	0,1564	0,0637	0,0538	0,0374	1

Apesar de termos usado a matriz de Burt para os cálculos, podemos interpretar segundo os dois pontos de vista, pois sabemos que o resultado é proporcional para as duas análises.

Coletando os termos em Γ_j^B e lembrando que $\Gamma^Z = \Gamma^B$ e $D_\lambda^Z = (D_\lambda^B)^{1/2}$ temos as seguintes coordenadas padrão para as categorias da variável j :

$$\Gamma_j^Z (JD_\lambda^Z - I) = \left(\sum_{j \neq j'} L_{jj'} \Gamma_{j'}^Z \right), \text{ onde } I \text{ é a matriz identidade.}$$

Tabela 2.2.7: Distribuição da inércia ao longo dos eixos.

Eixo (k)	$\lambda_k^B = (\lambda_k^Z)^2$	$\mu = \lambda_k^Z$	%	% Acumulado
1	0,2842	0,5331	19,99	19,99
2	0,1524	0,3903	14,64	34,63
3	0,1289	0,3590	13,46	48,09
4	0,1118	0,3344	12,54	60,63
5	0,1100	0,3317	12,44	73,07
6	0,1039	0,3224	12,09	85,16
7	0,0733	0,2707	10,15	95,31
8	0,0157	0,1251	4,69	100,00
Inércia total	0,9802	2,6667	100,00	

As coordenadas padrão Γ^B e as principais G^B são calculadas exatamente da mesma forma que na ACS, usando a nova matriz P . No cálculo de G^B há uma ponderação pela raiz do valor singular por eixo:

$$\Gamma^B = D_l^{-1/2} U \quad \text{e} \quad G^B = D_\mu^{1/2} \Gamma^B = D_{\lambda^Z}^{1/2} \Gamma^B = D_{\lambda^B}^{1/4} \Gamma^B,$$

onde D_μ é a matriz diagonal com os valores singulares da decomposição espectral da matriz de Burt, U é a matriz dos autovetores e D_l é matriz diagonal l^B (ou c^B)

Tabela 2.2.8: Coordenadas padrão dos perfis coluna (Matriz Γ^B)

Variável	Categorias	Eixo 1	Eixo 2	Eixo 3	Eixo 4
Sexo	H	-0,0459	0,1460	-1,1254	0,2140
	M	0,0641	-0,2038	1,5706	-0,2987
Escolaridade	menos 8 anos	-0,9649	-1,2917	-0,2626	0,8745
	8 a 10 anos	-0,6863	0,0388	-0,7192	-2,0374
	11 a 14 anos	-0,0126	1,5258	0,6278	0,4417
	15 ou mais anos	2,6555	-0,8867	0,0171	-0,1251
Rendimento	menos de 1 SM	-1,0827	-2,8540	1,6046	3,4422
	de 1 a 3 SM	-0,8182	-0,2752	0,1842	-1,1171
	de 3 a 5 SM	-0,1853	1,2913	-1,4653	1,4375
	de 5 a 10 SM	1,0595	1,5232	1,2653	0,4940
	mais de 10 SM	2,8462	-1,5623	-1,0385	-0,5139

A Tabela 2.2.8 apresenta cada coordenada padrão para os primeiros quatro eixos. A Tabela 2.2.9 apresenta as coordenadas principais para os mesmos eixos. O gráfico pode ser feito com qualquer uma dessas coordenadas, a escolha implica em uma mudança de escalas.

Tabela 2.2.9: Coordenadas principais dos perfis coluna (Matriz G^B)

Variável	Categorias	Eixo 1	Eixo 2	Eixo 3	Eixo 4
Sexo	H	-0,034	0,091	-0,674	0,124
	M	0,047	-0,127	0,941	-0,173
Escolaridade	menos 8 anos	-0,704	-0,807	-0,157	0,506
	8 a 10 anos	-0,501	0,024	-0,431	-1,178
	11 a 14 anos	-0,009	0,953	0,376	0,255
	15 ou mais anos	1,939	-0,554	0,010	-0,072
Rendimento	menos de 1 SM	-0,790	-1,783	0,961	1,990
	de 1 a 3 SM	-0,597	-0,172	0,110	-0,646
	de 3 a 5 SM	-0,135	0,807	-0,878	0,831
	de 5 a 10 SM	0,774	0,952	0,758	0,286
	mais de 10 SM	2,078	-0,976	-0,622	-0,297

Cada sub-tabela (bloco) fora da diagonal aparece duplicada nesta tabela de Burt, que é simétrica e quadrada. O nosso interesse se concentra na parte triangular fora da diagonal desta matriz, onde estão as relações entre todas as variáveis.

Estas sub-tabelas diagonais acabam elevando o valor da inércia total, já que uma matriz diagonal representa associação perfeita. Este fato faz com que a inércia total seja alta, enquanto que o percentual de inércia por eixos principais seja baixo. Isto nos dá uma impressão inadequada de que os mapas representam muito pouco da variabilidade dos dados.

Depois de obtidas as coordenadas de ACM, as coordenadas dos eixos principais podem ser re-escaladas para que a solução do mapa represente melhor as associações fora do bloco diagonal (somente a escala muda). Esta solução se aproxima mais da Análise de Correspondência Conjunta (Greenacre, 1994), que é a análise da tabela triangular fora da diagonal, sem grandes dificuldades computacionais.

Antes de re-escalar os eixos é preciso avaliar a inércia média da tabela de Burt fora da diagonal. Isto pode ser calculado direto na tabela de Burt ou a partir da inércia total da mesma.

Se todos os eixos tivessem a mesma contribuição (se tivéssemos uma matriz onde cada indivíduo respondesse somente uma das possíveis categorias nas J variáveis e cada categoria tivesse apenas uma resposta), os J eixos na ACM de Z teriam inércia de $1/J$ cada (Greenacre, 1984) e então, as inércias por eixo consideradas importantes seriam aquelas que superam este valor.

Greenacre (1990, 1993, 1994) propôs uma correção para cada inércia dos eixos para apenas estas inércias $In(Z)$ que superam $1/J$. Este cálculo baseia-se na inércia da matriz de Burt:

$$\left(\frac{J}{J-1}\right)^2 \left(\sqrt{\lambda_k^B} - \frac{1}{J}\right)^2,$$

onde J é o número de variáveis e λ_k^B é a inércia de Burt do eixo k ($k=1,...,Q-J$). Este resultado é devido ao fato da inércia total ser J^2 vezes a inércia de Burt das matrizes fora da diagonal e a soma das inércias dos blocos diagonais $Q-J$ (Q é o número total de categorias).

O ideal seria trabalhar somente com o bloco triangular fora da diagonal, porém computacionalmente ainda é muito complicado e o resultado desta transformação acima é quase o mesmo de se trabalhar com a matriz triangular contendo o bloco diagonal (difícilmente há grandes diferenças, (Greenacre, 1994)).

No exemplo temos $J = 3$, e os valores que serão usados para re-escalar os eixos estão descritos na Tabela 2.2.10.

Tabela 2.2.10: Coeficientes para re-escalamento dos eixos da ACM.

Eixo (k)	$In(B)$ $((\lambda_k^Z)^2)$	$In(Z)$ (λ_k^Z)	Inércia Corrigida (λ_k^*)	% Inércia	Coeficiente de correção
1	0,2842	0,5331	0,0898	91,08	0,2996
2	0,1524	0,3903	0,0073	7,40	0,0855
3	0,1289	0,3590	0,0015	1,52	0,0385
4	0,1118	0,3344	0,0000	100,00	0,0016
5	0,1100	0,3317			
6	0,1039	0,3224			
7	0,0733	0,2707			
8	0,0157	0,1251			
Inércia total	0,9801	2,6667	0,0986		

A Tabela 2.2.11 apresenta as coordenadas principais corrigidas pelo coeficiente apresentado para cada eixo na Tabela 2.2.10. Esta correção é feita pela simples multiplicação das coordenadas principais dadas na Tabela 2.2.9 pelos respectivos coeficientes de correção.

Tabela 2.2.11: Coordenadas principais corrigidas ($\sqrt{\lambda_k^*} G_B$)

Variável	Rótulo	Eixo 1	Eixo 2	Eixo 3	Eixo 4
Sexo	H	-0,010	0,008	-0,026	0,000
	M	0,014	-0,011	0,036	0,000
Escolaridade	menos 8 anos	-0,211	-0,069	-0,006	0,001
	8 a 10 anos	-0,150	0,002	-0,017	-0,002
	11 a 14 anos	-0,003	0,082	0,014	0,000
	15 ou mais anos	0,581	-0,047	0,000	0,000
Rendimento	menos de 1 SM	-0,237	-0,152	0,037	0,003
	de 1 a 3 SM	-0,179	-0,015	0,004	-0,001
	de 3 a 5 SM	-0,041	0,069	-0,034	0,001
	de 5 a 10 SM	0,232	0,081	0,029	0,000
	mais de 10 SM	0,623	-0,083	-0,024	0,000

As Figuras 2.2.1 e 2.2.2 são os mapas de correspondência resultantes da ACM da tabela de Burt. Caso tivéssemos usado a tabela disjunta completa poderíamos fazer a interpretação destes mapas com relação aos indivíduos; as coordenadas dos mesmos podem ser obtidas pela fórmula de transição. Esta análise nos permitiria realizar uma tipologia dos indivíduos, baseada nas semelhanças e diferenças dos mesmos. Assim, quanto maior o número de categorias em comum, mais próximo um indivíduo está do outro. Como não é o objetivo desta dissertação enfatizar o uso da ACM para matriz Z , não realizamos a análise dos indivíduos.

As correções das coordenadas usadas na Figura 2.2.2 mostram que a representação, tanto no primeiro quanto no segundo eixo, sofrem um condensamento; isto porque foi excluída parte da inércia “ilusória” causada pelos blocos diagonais da matriz de Burt (observe as diferenças de escalas no gráfico).

Figura 2.2.1: Mapa da ACM sem a correção dos eixos

Figura 2.2.2: Mapa da ACM com correção de eixos

Com relação às variáveis, podemos resumir um conjunto de variáveis a um pequeno número de variáveis numéricas, desde que estas estejam ligadas a um mesmo assunto, usando para tanto os eixos encontrados. Por exemplo, o primeiro eixo pode ser considerado como uma nova variável quantitativa que resume toda a informação com relação principalmente às categorias de ‘15 ou mais anos’ de estudos (com coordenada 0,581), ‘de 5 a 10’ SM (com peso 0,232) e ‘mais de 10’ SM (com peso 0,623). Observamos que as relações encontradas na ACS se mantêm e que as categorias de homens e de mulheres se encontram no centro do mapa, aparentando que não há relação entre ambas categorias. Entretanto, se observarmos, na Tabela 2.2.12 a coluna que se refere à contribuição para o terceiro eixo, notamos que as relações existentes entre ambos sexos estão nessa dimensão.

Tabela 2.2.12: Contribuição de cada categoria para a inércia do eixo.

Categorias	Eixo 1	Eixo 2	Eixo 3	Eixo 4
Homem	0,04	0,41	24,59	0,89
Mulher	0,06	0,58	34,32	1,24
Menos de 8	9,35	16,75	0,69	7,68
De 8 a 10	3,04	0,01	3,34	26,83
De 11 a 14	0,00	26,67	4,52	2,24
15 ou mais	37,89	4,22	0,00	0,08
Menos de 1	2,58	17,92	5,66	26,06
De 1 a 3	10,47	1,18	0,53	19,51
De 3 a 5	0,22	10,62	13,68	13,17
De 5 a 10	6,04	12,49	8,62	1,31
Mais de 10	30,32	9,13	4,04	0,99
Total	100	100	100	100

A contribuição de cada categoria para a inércia do eixo (Tabela 2.2.12), permite definir o “grau de generalidade” de cada eixo. Se um eixo está formado por um grande número de categorias, este apresenta um alto grau de generalidade (o eixo resume muita informação). Já quando este grau é baixo, esta informação é muito específica. Se poucas categorias foram observadas em poucos indivíduos, estas categorias terão uma forte contribuição para a inércia geral, conseqüentemente os primeiros eixos serão formados por estas categorias raras. Este problema deve ser levado em consideração pelo pesquisador, que deve tentar manter as categorias com um número razoável de indivíduos e um número relativamente homogêneo de categorias por variável.

Observamos que a variável sexo contribui muito pouco para inércia dos dois primeiros eixos e sua contribuição para o terceiro eixo se eleva bastante, sendo a variável de maior contribuição para inércia do eixo 3. Este fato explica porque no plano aparentemente homens e mulheres não se mostram com perfis diferentes. As relações entre rendimentos e escolaridade para homens e mulheres é diferente, porém de difícil visualização no primeiro plano. As Figuras 2.2.3 e 2.2.4 mostram que a categoria homem apresenta profundidade (valor negativo para eixo 3) enquanto a mulher está com maior altura (acima do primeiro plano).

Figura 2.2.3: Mapa da ACM com correção de eixos para eixos 1 e 3

Figura 2.2.4: Mapa da ACM com correção de eixos para eixos 2 e 3

Observamos nas Figuras 2.2.3 e 2.2.4 que as categorias de rendimentos ‘Menos de 1’ e ‘De 1 a 3’ e também a ‘De 5 a 10’ apresentam valores positivos para o eixo 3, assim como a categoria do sexo feminino. Já os rendimentos ‘De 3 a 5’ e ‘Mais de 10’ são negativos, assim como o sexo masculino. Com relação à escolaridade, observe que na Figura 2.2.3 da esquerda para direita há um aumento gradual dos anos de estudos e com relação à média as mulheres apresentam uma tendência leve para um número mais elevado de anos de estudos.

Os elementos suplementares também podem ser incluídos na ACM, são considerados como perfis de peso nulo, e as fórmulas de transição são usadas para que eles possam ser incluídos nos mapas.

2.3 Análise Fatorial Múltipla para Tabelas de Contingência

A Análise Fatorial Múltipla para Tabelas de Contingência (AFMTC), introduzida por Bécue & Pagès (2004), é uma proposta para estudar um conjunto de tabelas de contingência de dupla entrada, ao longo de uma terceira variável, com a restrição de que as tabelas apresentem pelo menos uma variável em comum. Neste texto fixamos a variável descrita nas linhas da tabela como a variável comum, permitindo que as colunas acomodem variáveis diferentes. O tipo de tabela que esta técnica analisa está descrito pela Tabela 2.3.1, onde n_{ijk} é o número de indivíduos na linha i , coluna j da subtabela ou bloco k ($i = 1, 2, \dots, I; k = 1, 2, \dots, K; j = 1, 2, \dots, J_k$).

Tabela 2.3.1: Justaposição de tabelas de contingências em K blocos.

	Bloco I						Bloco k				Bloco K			
Categorias	cat. 1	cat. 2	(...)	cat. J_1	(...)	(...)	cat. j	(...)	(...)	cat. 1	(...)	cat. J_K	Marginal	
1	n_{111}						n_{1jk}					n_{1JK}	$n_{1..}$	
...														
i	n_{i11}						n_{ijk}					n_{iJ_KK}	$n_{i..}$	
...														
I	n_{I11}						n_{Ijk}					n_{IJ_KK}	$n_{I..}$	
Marginal	$n_{.11}$						$n_{.jk}$					$n_{.J_KK}$	n	

Todos os cálculos efetuados nesta análise são baseados nas proporções relativas com relação ao total geral. Desta forma, $p_{ijk} = n_{ijk} / n$ e $\sum_{ijk} p_{ijk} = 1$. O perfil linha da tabela geral P é $p_{i..} = \sum_{jk} p_{ijk}$ e o perfil coluna é $p_{.jk} = \sum_i p_{ijk}$. A marginal linha do bloco k é

$$\frac{P_{i\bullet k}}{P_{\bullet\bullet k}} = \frac{\sum_j P_{ijk}}{P_{\bullet\bullet k}} \text{ enquanto a marginal coluna é } \frac{P_{\bullet jk}}{P_{\bullet\bullet k}} = \frac{\sum_i P_{ijk}}{P_{\bullet\bullet k}}, \text{ onde o total por bloco é}$$

$$P_{\bullet\bullet k} = \sum_{ij} P_{ijk}.$$

Para um melhor entendimento desta técnica, primeiramente faremos uma introdução à Análise Fatorial Múltipla (AFM). A AFM é uma técnica muito usada para tratar dados nos quais temos um conjunto de linhas, usualmente indivíduos, descrito por vários blocos de variáveis (Escofier e Pagès, 1992, Pagès, 1997). Dentro de cada bloco de variáveis é necessário que o tipo das variáveis seja o mesmo, porém de um bloco para outro podemos ter tipos diferentes. Esta técnica analisa uma tabela similar à Tabela 2.3.1, em que cada coluna de um mesmo bloco representa uma variável, e não necessariamente categorizada; no caso da AFMTC temos apenas variáveis categorizadas.

Este tipo de análise visa representar todos os blocos de variáveis conjuntamente, numa análise usualmente chamada de análise global. Se a análise for feita separadamente, uma para cada bloco, nada nos garante que não haja problemas de rotação, o que não permite uma análise direta.

Se todos os blocos de variáveis fossem tratados de uma só vez poderíamos ter um grande problema de apenas um dos blocos de variáveis predominar na inércia dos primeiros eixos. Para evitar este problema a AFM balanceia estes blocos de acordo com a inércia intra-blocos (inércia de cada sub-bloco).

A influência de um bloco é resultante da estrutura das variáveis que pertencem a esse bloco, ou seja, a inércia do bloco. Por exemplo, se um bloco apresenta uma inércia mais elevada que os demais, esse bloco apresentará uma alta influência no primeiro eixo da análise global. Para evitar este problema a AFM normaliza as tabelas com relação à inércia do primeiro eixo de cada tabela separadamente, ou seja, pondera cada variável do bloco k por $1/\lambda_1^k$, onde λ_1^k é o maior autovalor da DVS do bloco k . Desta forma, a ponderação transformará a maior inércia de eixo de cada bloco, em 1. Para encontrar o λ_1^k de cada bloco leva-se em consideração o tipo de variável considerada dentro de cada bloco; caso sejam quantitativas, usa-se a ACP e caso qualitativas, usa-se a AC.

Esta forma de ponderação não balanceia a inércia total de um bloco, assim um bloco com alta dimensionalidade terá uma alta influência na análise global (contribuirá para

vários eixos). Porém, este bloco não apresenta razão para contribuir particularmente para o primeiro eixo. Da mesma forma, um bloco unidimensional contribuirá muito para somente um eixo, que poderá, ou não, ser o primeiro da análise global. Vale ressaltar que a ponderação é igual dentro de cada bloco, não alterando a estrutura das relações.

Chamamos de nuvem parcial de indivíduos aquela que representa a nuvem dos indivíduos relacionados às variáveis do bloco k (N_I^k).

O objetivo principal da AFM se resume na seguinte pergunta: Existe uma estrutura comum nestas nuvens N_I^k ? As distâncias entre os pontos-linha comuns (homólogos) são grandes? Duas linhas próximas com relação a uma sub-tabela também são próximas com relação à outra sub-tabela?

Para avaliar esta questão é necessário sobrepor estas análises dos blocos, em um mesmo mapa. Para tanto encontra-se as coordenadas de cada nuvem N_I^k , depois juntam-se todas as nuvens de maneira a deixar os pontos homólogos o mais próximos possível um do outro. Esta técnica se aproxima à Análise *Procrustes*, que minimiza distâncias entre pontos homólogos (Gower, 1984), mas projeta N_I^k nos eixos da análise global como “pontos suplementares”. De fato, estes não são suplementares porque os dados contribuem para os eixos, já que também foram usados para o cálculo dos eixos da análise global.

Voltando à AFMTC, queremos comparar as estruturas de várias tabelas de contingência que apresentam as mesmas linhas, porém não necessariamente mesmas colunas, geralmente comparando populações diferentes (Pagès e Bécue, in press). Entende-se como estrutura de uma tabela de contingência, as relações entre suas linhas e colunas e assim como na ACS, é visualizada por meio de representação gráfica. Vários métodos foram desenvolvidos para analisar as relações entre duas variáveis qualitativas ao longo de uma terceira variável, mas a grande contribuição da AFMTC é que as colunas das tabelas não precisam ser iguais nos diferentes blocos.

Adotando o ponto de vista de análise de correspondência de tabelas justapostas e imaginando que as tabelas podem apresentar marginais bem diferentes umas das outras (não proporcionais) a ACS de cada tabela separadamente não compara a estrutura interna entre elas, pois seus centróides são diferentes. Uma generalização de ACS foi proposta por

Benzécri (1983) e por Escofier e Drouot (1983) e é chamada de Análise de Correspondência Interna (ACI) que resolve este problema de centrar as sub-tabelas em suas marginais.

Para compreender esta ACI é necessário levar em consideração a relação entre ACS e Análise de Componentes Principais (ACP). Os eixos principais da ACS podem ser obtidos através da ACP de uma tabela Y na qual cada elemento desta tabela é o resíduo com relação ao modelo de independência (homogeneidade), ponderado pela inversa do produto cruzado das marginas linha e coluna.

$$y_{ij} = \frac{p_{ij} - p_{i\bullet} p_{\bullet j}}{p_{i\bullet} p_{\bullet j}}.$$

A ACP usa a métrica D_I como peso linha e D_J como peso coluna (Escofier e Pagès, 1998). Essas matrizes D_I e D_J são diagonais e D_I apresenta $p_{i\bullet}$ na sua diagonal enquanto que D_J apresenta $p_{\bullet j}$.

Uma generalização da ACS para qualquer modelo M , e para qualquer métrica P (peso, termo geral p_i) e Q (termo geral q_j) foi proposta por Escofier (1983, 1984). O termo geral m_{ij} de M é o valor esperado de p_{ij} sob o modelo considerado, e realizar uma ACS de uma tabela de contingência com relação ao modelo M nas métricas P e Q é equivalente a realizar ACP nas métricas P e Q da matriz com termo geral:

$$\frac{p_{ij} - m_{ij}}{p_i q_j}.$$

A ACI estende esta idéia para blocos de tabelas de contingência de tal forma que a dispersão das estruturas comuns para todas as tabelas juntas seja representada pelos eixos globais, responsáveis pela máxima inércia da tabela global. E assim como em ACS, a estrutura de cada tabela de contingência corresponde à discrepância com relação ao modelo de independência (homogeneidade) e também da mesma forma como em ACS as métricas e pesos são induzidos pelas marginais linha e coluna das tabelas.

O modelo de independência intra-tabela considera as marginais de cada sub-tabela de tal forma que:

$$m_{ijk} = \left(\frac{p_{i\bullet k}}{p_{\bullet\bullet k}} \right) \left(\frac{p_{\bullet j k}}{p_{\bullet\bullet k}} \right).$$

O vetor de pesos linha (métrica do espaço coluna) é a marginal linha da tabela geral $p_{i..}$, que é a média ponderada, pelos totais de cada bloco, das marginais linha do bloco. A escolha por este peso linha permite que sejam introduzidos pontos suplementares, já que para todos os blocos o peso é o mesmo.

Os pesos coluna (métrica do espaço linha) devem ser escolhidos de tal forma que cada bloco de coluna possa ser comparado na análise global, levando em consideração que uma tabela de maior dimensão sempre terá maior influência sobre o eixo.

Em tabelas de contingência duas situações podem aumentar a inércia:

- Estrutura forte da tabela do bloco k , ou seja, relação forte entre as duas variáveis elevando os autovalores naquele bloco.
- Total geral elevado no bloco k , por exemplo, uma amostra maior que as demais, elevando assim o percentual total da tabela k : $p_{..k}$.

Para balancear essas influências de cada tabela, usamos a idéia de AFM, de padronizar a maior inércia por eixo de cada análise separada, para 1, por isso dividimos o peso da coluna do bloco k por λ_1^k (maior autovalor da ACI do bloco k). Resumindo, usaremos a idéia de ACI, que centra as tabelas e também a idéia de AFM para ponderar as inércias por blocos, chegando a uma configuração geral que chamamos de AFMTC.

Então usamos as métricas D_I , que é uma matriz diagonal com os valores $p_{i..}$, porque precisamos de um peso comum para as linhas e a métrica D_J .

Os componentes da métrica no espaço linha são D_{J_k} com termo geral $p_{.jk}/p_{..k}$ dividido por λ_1^k , que apresentam várias propriedades (Escofier e Pagès, 1998). No geral, a estrutura intra-tabela não é modificada e, exceto em casos muito especiais, o primeiro eixo da análise global não é gerado por apenas uma tabela.

A AFMTC pode ser realizada através de uma DVS da tabela Z na qual o termo geral é o resíduo padronizado com relação ao modelo de independência calculado para cada

bloco nas métricas D_I e D_{J_k} , onde $Z_k = \frac{p_{ijk} - \frac{p_{i..k} p_{.jk}}{p_{..k}}}{\sqrt{p_{i..k} p_{.jk}}}$, e a matriz para DVS é $D_I^{1/2} Z D_K^{1/2}$, onde $Z = [Z_1 \dots Z_k \dots Z_K]$ e $D_K = \text{diag}\{\frac{1}{\lambda_1^1} D_{J_1}, \dots, \frac{1}{\lambda_1^K} D_{J_K}\}$.

Note que Z (da ACI) e Y (ACS) são parecidas, mas a ponderação é diferente. Em ambos os casos medem-se as distâncias com relação ao modelo de independência, porém com pesos diferentes. A necessidade de um peso comum para a mesma linha torna a comparação entre os Z_k possível. Se os perfis linha das tabelas de cada bloco são parecidos, a análise de Y e Z são parecidas, mas se os perfis diferem, a análise de Z é mais apropriada para a comparação das Z_k .

A Tabela 2.3.2 é a justaposição de duas tabelas de contingências que são usadas como exemplo nesta seção. A primeira tabela relaciona “Anos de estudo” com “Rendimento” para os homens ocupados da Região Metropolitana de São Paulo, a segunda relaciona as mesmas variáveis para as mulheres. Será que esta relação ocorre de uma mesma maneira para homens e mulheres?

Tabela 2.3.2: Tabela de Contingência N de “Escolaridade” *versus* “Rendimento” por “Sexo”, 2002

Escolaridade (em anos)	Homem						Mulher						Total Geral
	Rendimentos em Salário Mínimo Hora						Rendimentos em Salário Mínimo Hora						
	Menos de 1	De 1 a 3	De 3 a 5	De 5 a 10	Mais de 10	Total	Menos de 1	De 1 a 3	De 3 a 5	De 5 a 10	Mais de 10	Total	
Menos de 8	130551	867278	301013	113918	35756	1448516	147182	573745	79834	23284	831	824876	2273392
De 8 a 10	55710	485598	211206	122236	32431	907181	64855	358368	89805	35752	7485	556265	1463446
De 11 a 14	38252	560435	347570	345929	157159	1449345	53214	606197	251131	192095	42413	1145050	2594395
15 ou mais	4988	37421	64026	156320	329290	592045	3328	51553	98121	229513	241991	624506	1216551
Total	229501	1950732	923815	738403	554636	4397087	268579	1589863	518891	480644	292720	3150697	7547784

Fonte: IBGE-PNAD, 2002.

Tabela 2.3.3: Tabela de frequências relativas, matriz P

Escolaridade (em anos)	Homem						Mulher						$P_{i..}$
	Rendimentos em Salário Mínimo Hora						Rendimentos em Salário Mínimo Hora						
	Menos de 1	De 1 a 3	De 3 a 5	De 5 a 10	Mais de 10	$P_{i..1}$	Menos de 1	De 1 a 3	De 3 a 5	De 5 a 10	Mais de 10	$P_{i..2}$	
Menos de 8	0,0173	0,1149	0,0399	0,0151	0,0047	0,1919	0,0195	0,0760	0,0106	0,0031	0,0001	0,1093	0,3012
De 8 a 10	0,0074	0,0643	0,0280	0,0162	0,0043	0,1202	0,0086	0,0475	0,0119	0,0047	0,0010	0,0737	0,1939
De 11 a 14	0,0051	0,0743	0,0460	0,0458	0,0208	0,1920	0,0071	0,0803	0,0333	0,0255	0,0056	0,1517	0,3437
15 ou mais	0,0007	0,0050	0,0085	0,0207	0,0436	0,0784	0,0004	0,0068	0,0130	0,0304	0,0321	0,0827	0,1612
$P_{..jk}$	0.0304	0.2585	0.1224	0.0978	0.0735	$P_{..1}$ 0.5826	0.0356	0.2106	0.0687	0.0637	0.0388	$P_{..2}$ 0.4174	1.0000

A Tabela 2.3.3 é a matriz de frequências relativas, que é usada como base para os cálculos da matriz Z . As marginais desta matriz são os pesos, linha e coluna, usados para o cálculo da ACI.

A Tabela 2.3.4 é a matriz Z adaptada para que a comparação possa ser feita em um mesmo gráfico. O primeiro passo é encontrar os autovalores das ACI para cada uma destas tabelas Z_k e para tanto precisamos das matrizes de métricas D_{J_k} , para cada bloco k , que apresenta $\frac{p_{\cdot jk}}{p_{\cdot \cdot k}}$ na diagonal. A matriz de pesos D_I é a mesma para as duas ACI, que é diagonal com $p_{i \cdot \cdot}$.

Para calcular os respectivos λ_1^k , basta fazer a DVS das matrizes $D^{1/2} Z_k D_{J_k}^{1/2}$ para cada bloco k . O autovalor de cada ACI é o correspondente maior valor singular ao quadrado. Estes autovalores são usados para ponderar a inércia entre blocos.

Tabela 2.3.4: Tabela adaptada para AFMTC (Z).

	Homem					Mulher				
	Menos de 1	De 1 até 3	De 3 até 5	De 5 até 10	Mais de 10	Menos de 1	De 1 até 3	De 3 até 5	De 5 até 10	Mais de 10
Menos de 8	0,7949	0,3824	-0,0119	-0,5815	-0,8797	0,9502	0,3289	-0,3584	-0,7084	-0,8598
De 8 a 10	0,1879	0,2198	0,1151	-0,2103	-0,7625	0,3348	0,2520	-0,0180	-0,5269	-0,7787
De 11 a 14	-0,4740	-0,1231	0,1356	0,4040	-0,1346	-0,4809	0,0520	0,3507	0,1054	-0,6358
15 ou mais	-0,7005	-0,7164	-0,4054	0,4781	2,8481	-1,1529	-1,0286	-0,0565	1,7328	3,8993

O passo seguinte é alterar a matriz de métricas usando estes autovalores parciais. A matriz diagonal D_K tem em sua diagonal os valores $\frac{p_{\cdot jk}}{p_{\cdot \cdot k} \lambda_1^k}$. A matriz de pesos é a mesma

D_I e a AFMTC é obtida pela DVS da matriz $D_I^{1/2} Z D_K^{1/2}$. As matrizes D_I , D_{J_1} , D_{J_2} e D_K do exemplo são dadas nas Tabelas 2.3.5, 2.3.6, 2.3.7 e 2.3.8, respectivamente.

Tabela 2.3.5: Matriz D_I

0,3012	0	0	0
0	0,1939	0	0
0	0	0,3437	0
0	0	0	0,1612

Tabela 2.3.6: Matriz D_{J_1}

0,0522	0	0	0	0
0	0,4436	0	0	0
0	0	0,2101	0	0
0	0	0	0,1679	0
0	0	0	0	0,1261

Tabela 2.3.7: Matriz D_{J_2}

0,0852	0	0	0	0
0	0,5046	0	0	0
0	0	0,1647	0	0
0	0	0	0,1526	0
0	0	0	0	0,0929

Tabela 2.3.8: Matriz D_K

0,1731	0	0	0	0	0	0	0	0	0
0	1,4714	0	0	0	0	0	0	0	0
0	0	0,6968	0	0	0	0	0	0	0
0	0	0	0,5570	0	0	0	0	0	0
0	0	0	0	0,4183	0	0	0	0	0
0	0	0	0	0	0,1678	0	0	0	0
0	0	0	0	0	0	0,9933	0	0	0
0	0	0	0	0	0	0	0,3242	0	0
0	0	0	0	0	0	0	0	0,3003	0
0	0	0	0	0	0	0	0	0	0,1829

Tabela 2.3.9: Inércia (autovalor) por eixo das ACI's para a cada tabela e da AFMTC da tabela global

Z_1 : Homens				Z_2 : Mulheres			Z : Análise Global		
Eixo	Autovalor	%	% Acumulado	Autovalor	%	% Acumulado	Autovalor	%	% Acumulado
1	0,3015	90,8	90,8	0,5080	92,3	92,3	1,9859	90,9	90,9
2	0,0304	9,2	100,0	0,0421	7,6	99,9	0,1970	9,0	99,9
3	0,0001	0,0	100,0	0,0005	0,1	100,0	0,0021	0,1	100,0
4	0,0000	0,0	100,0	0,0000	0,0	100,0	0,0000	0,0	100,0
Total	0,3320	100		0,5506	100		2,185	100	

A Tabela 2.3.9 destaca os resultados com relação à inércia para cada uma das ACI separados por sexo, e também o resultado para a análise global. O número de eixos é igual ao número de linhas da tabela (que é menor do que número de colunas), mas observe que um eixo pode ser excluído da análise, pois não contribui para inércia, como esperado. Observamos que a variabilidade entre as mulheres é maior do que entre os homens (a soma

dos autovalores é maior) e, além disso, as mulheres têm uma maior contribuição dessa inércia no primeiro eixo. Com relação à análise global, temos que o primeiro eixo tem uma grande contribuição para a inércia, com mais de 90%, assim como nas análises parciais.

Depois de encontrados os valores singulares e as matrizes U e V da DVS da análise global, o cálculo das coordenadas das linhas da tabela é feito de forma similar ao descrito nas Seções 2.1 e 2.2. A matriz D_K pode ser decomposta em duas matrizes diagonais, D_{1/λ_1^k}

com os valores $\frac{1}{\lambda_1^k}$ e D_{J_k} com os valores $\frac{P_{\bullet jk}}{P_{\bullet\bullet k}}$. Então temos, $D_K = D_{1/\lambda_1^k} D_{J_k}$ e a matriz

para DVS global é $D_I^{1/2} Z D_{J_k}^{1/2} D_{1/\lambda_1^k}^{1/2}$. As matrizes usadas nos cálculos das coordenadas são as matrizes de pesos D_I e D_{1/λ_1^k} . As coordenadas padrão X e Y são:

$$X = D_I^{-1/2} U \quad \text{e} \quad Y = D_{1/\lambda_1^k}^{-1/2} V.$$

As coordenadas principais do perfil linha é $F = XD_\mu$ e as coordenadas principais do perfil coluna $G = YD_\mu$, onde D_μ é matriz diagonal com valores singulares da DVS global. As Tabelas de 2.3.10 à 2.3.18 exibem, passo a passo, as matrizes relativas à análise do exemplo “Anos de estudos” *versus* “Rendimento”, por “sexo”.

Tabela 2.3.10: Matriz $D_I^{-1/2}$

1,8221	0	0	0
0	2,2710	0	0
0	0	1,7057	0
0	0	0	2,4908

Tabela 2.3.11: Matriz de autovetores U

-0,4260	0,5869	-0,4158	0,5488
-0,2464	0,0208	0,8631	0,4403
-0,0122	-0,7585	-0,2843	0,5863
0,8704	0,2825	0,0369	0,4015

Tabela 2.3.12: Matriz $D_{1/\lambda_i^2}^{-1/2}$

0,5491	0	0	0	0	0	0	0	0	0	0
0	0,5491	0	0	0	0	0	0	0	0	0
0	0	0,5491	0	0	0	0	0	0	0	0
0	0	0	0,5491	0	0	0	0	0	0	0
0	0	0	0	0,5491	0	0	0	0	0	0
0	0	0	0	0	0	0,7127	0	0	0	0
0	0	0	0	0	0	0	0,7127	0	0	0
0	0	0	0	0	0	0	0	0,7127	0	0
0	0	0	0	0	0	0	0	0	0,7127	0
0	0	0	0	0	0	0	0	0	0	0,7127

Tabela 2.3.13: Matriz de autovetores V

-0,1322	0,3647	-0,3760	-0,0953
-0,3122	0,2696	0,1645	-0,6922
-0,0902	-0,2051	0,3256	-0,4500
0,1710	-0,5290	-0,1223	-0,4032
0,5896	0,1349	-0,3455	-0,3257
-0,1912	0,3619	-0,2372	-0,0961
-0,3282	-0,0708	-0,0696	0,1574
0,0256	-0,3566	0,1949	0,0565
0,3218	-0,1029	-0,3658	-0,0157
0,5015	0,4248	0,5986	0,0569

Tabela 2.3.14: Matriz D_μ

1,4092	0	0	0
0	0,4439	0	0
0	0	0,0457	0
0	0	0	0

Tabela 2.3.15: Coordenadas padrão dos perfis linha (matriz X)

-0,7762	1,0694	-0,7576	1,0000
-0,5597	0,0472	1,9601	1,0000
-0,0208	-1,2937	-0,4850	1,0000
2,1681	0,7037	0,0919	1,0000

Tabela 2.3.16: Coordenadas principais dos perfis linha (matriz F)

-1,0939	0,4747	-0,0346	0,0000
-0,7887	0,0209	0,0896	0,0000
-0,0293	-0,5743	-0,0222	0,0000
3,0553	0,3124	0,0042	0,0000

Tabela 2.3.17: Coordenadas padrão dos perfis coluna (matriz Y)

-0,0726	0,2003	-0,2065	-0,0521
-0,1714	0,1480	0,0903	-0,3546
-0,0495	-0,1126	0,1788	-0,2771
0,0939	-0,2905	-0,0672	-0,2208
0,3238	0,0741	-0,1897	-0,1722
-0,1363	0,2580	-0,1691	-0,1094
-0,2339	-0,0504	-0,0496	0,1203
0,0183	-0,2541	0,1389	0,0449
0,2294	-0,0733	-0,2607	-0,0174
0,3575	0,3028	0,4266	0,0373

Tabela 2.3.18: Coordenadas principais dos perfis coluna (matriz G)

-0,1023	0,0889	-0,0094	0,0000
-0,2416	0,0657	0,0041	0,0000
-0,0698	-0,0500	0,0082	0,0000
0,1323	-0,1289	-0,0031	0,0000
0,4563	0,0329	-0,0087	0,0000
-0,1921	0,1145	-0,0077	0,0000
-0,3296	-0,0224	-0,0023	0,0000
0,0258	-0,1128	0,0064	0,0000
0,3232	-0,0325	-0,0119	0,0000
0,5037	0,1344	0,0195	0,0000

Resumindo, primeiramente adaptamos a ACI a partir da idéia de ACS, encontramos as inércias para os primeiros eixos da ACI de cada sub tabela k , e usamos estes valores para ponderar o peso coluna que será usado na AFMTC. Isto feito, a AFMTC está baseada na DVS da matriz centrada e ponderada $D_I^{1/2} Z D_K^{1/2}$. A seqüência da Tabela 2.3.19 resume todas estas informações, comparando as matrizes para DVS em cada situação.

A fórmula de transição de coluna para linha é obtida a partir da manipulação da matriz usada da DVS e das coordenadas principais F e G . Temos que,

$$\begin{cases} D_I^{1/2} Z D_{J_k}^{1/2} D_{1/\lambda_1^k}^{1/2} = U D_\mu V^T \\ F = D_I^{-1/2} U D_\mu \\ G = D_{1/\lambda_1^k}^{-1/2} V D_\mu \\ U^T U = V^T V = I. \end{cases}$$

A manipulação destas informações nos permite concluir que $F_i^s = \frac{1}{\lambda_1^k} \frac{1}{\mu_s} \sqrt{\frac{p_{\bullet jk}}{p_{\bullet\bullet k}}} Z_{ijk} G_{is}$, onde o índice s indica o eixo da coordenada.

Tabela 2.3.19: Resumo esquemático das etapas da AFMTC

	Termo Geral: Y	Peso Linha: D_l	Peso Coluna: D_c	Matriz para DVS
ACS	$\frac{p_{ijk} - \frac{p_{i\bullet k}}{p_{\bullet\bullet k}} p_{\bullet jk}}{\frac{p_{i\bullet k}}{p_{\bullet\bullet k}} p_{\bullet jk}}$	$\frac{p_{i\bullet k}}{p_{\bullet\bullet k}}$	$\frac{p_{\bullet jk}}{p_{\bullet\bullet k}}$	$D_l^{1/2} Y D_c^{1/2}$
ACI	$\frac{p_{ijk} - \frac{p_{i\bullet k}}{p_{\bullet\bullet k}} p_{\bullet jk}}{p_{i\bullet\bullet} p_{\bullet jk}}$	$p_{i\bullet\bullet}$	$\frac{p_{\bullet jk}}{p_{\bullet\bullet k}}$	$D_l^{1/2} Y D_c^{1/2}$
	Termo Geral: Z	Peso Linha: D_l	Peso Coluna: D_K	Matriz para DVS
AFMTC	$\frac{p_{ijk} - \frac{p_{i\bullet k}}{p_{\bullet\bullet k}} p_{\bullet jk}}{p_{i\bullet\bullet} p_{\bullet jk}}$	$p_{i\bullet\bullet}$	$\frac{p_{\bullet jk}}{p_{\bullet\bullet k}} \frac{1}{\lambda_1^k}$	$D_l^{1/2} Z D_K^{1/2}$

Com intuito de comparar as linhas de cada bloco pode-se calcular as coordenadas parciais dos perfis linha de cada sub-tabela. Para tanto usamos a seguinte fórmula de transição adaptada: $F_i^s = \frac{1}{\lambda_1^k} \frac{1}{\mu_s} K \sqrt{\frac{p_{\bullet jk}}{p_{\bullet\bullet k}}} Z_{ijk} G_{is}$. A média das coordenadas dos perfis linha parciais é igual ao perfil global, também chamado de perfil médio.

Da mesma forma como em ACS e ACM podemos calcular a contribuição de cada categoria para a inércia. A contribuição das categorias de linha e coluna são

$$\text{respectivamente: } Cr_i^s = \frac{p_{i..} F_{is}^2}{\sum_i p_{i..} F_{is}^2} \text{ e } Cr_j^s = \frac{\frac{1}{\lambda_1^k} G_{js}^2}{\sum_j \frac{1}{\lambda_1^k} G_{js}^2}.$$

A Tabela 2.3.20 destaca as coordenadas principais calculadas para cada linha de cada bloco, a distância de cada categoria até o centro, e também a contribuição de cada categoria para a inércia do eixo. A distância da categoria até a origem é a distância euclidiana com relação ao vetor zero: $d_{x^2}(\text{perfil linha } i, \text{origem}) = \sum_s F_{is}^2$.

Tabela 2.3.20: Coordenadas e contribuição de cada linha em cada bloco para os eixos.

Bloco		Distância até origem	Coordenadas			Contribuição		
			eixo 1	eixo 2	eixo 3	eixo 1	eixo 2	eixo 3
< 8 anos	Média	1,4231	-1,0939	-0,4747	0,0346	18,1483	34,4464	17,2857
	Homem	3,3466	-1,1946	-0,8011	-0,3967	11,9123	12,7750	5,2367
	Mulher	2,3460	-0,9932	-0,1483	0,4660	11,9123	12,7750	5,2367
8 a 10 anos	Média	0,6305	-0,7887	-0,0209	-0,0896	6,0734	0,0432	74,4957
	Homem	1,4171	-0,8397	-0,1944	-0,4706	1,9671	2,3224	2,6303
	Mulher	1,1050	-0,7377	0,1525	0,2914	1,9671	2,3224	2,6303
11 a 14 anos	Média	0,3311	-0,0293	0,5743	0,0222	0,0148	57,5283	8,0842
	Homem	0,6900	0,1254	0,6133	-0,1593	32,0676	0,2088	1,0579
	Mulher	0,6345	-0,1839	0,5352	0,2037	32,0676	0,2088	1,0579
> 14 anos	Média	9,4326	3,0553	-0,3124	-0,0042	75,7636	7,9822	0,1363
	Homem	17,9012	2,9750	0,4228	1,6472	4,0529	34,6937	41,0752
	Mulher	19,8291	3,1356	-1,0476	-1,6556	4,0530	34,6937	41,0752

Tabela 2.3.21: Quadrado dos cossenos para eixos de cada linha.

	$\cos^2 \theta$		
	eixo 1	eixo 2	eixo 3
< 8 anos	0,8408	0,1583	0,0008
8 a 10 anos	0,9866	0,0007	0,0127
11 a 14 anos	0,0026	0,9959	0,0015
> 14 anos	0,9897	0,0103	0,0000

A Tabela 2.3.21 apresenta os cossenos quadrados para cada linha e eixo da AFMTC, este valor é $\cos^2 \theta(i,s) = \frac{F_{is}^2}{\sum_s F_{is}^2}$. Observe que a interpretação é diferente da contribuição da categoria para a inércia do eixo, sendo interpretada como a contribuição do eixo para a inércia da categoria i .

Tabela 2.3.22: Inércias para cada eixo separadas por bloco e para tabela global

Eixo	Z ₁ : Homens			Z ₂ : Mulheres			Z: Análise Global		
	Autovalor	%	% Acumulado	Autovalor	%	% Acumulado	Autovalor	%	% Acumulado
1	0,9929	90,16	90,16	0,9930	91,63	91,63	1,9859	90,9	90,90
2	0,1075	9,77	99,92	0,0895	8,26	99,89	0,1970	9,00	99,90
3	0,0009	0,08	100,00	0,0012	0,11	100,00	0,0021	0,10	100,00
4	0,0000	0,00	100,00	0,0000	0,00	100,00	0,000	0,00	100,00

A Tabela 2.3.22 resume as inércias para cada eixo separadas por blocos da AFMTC. Os cálculos destas inércias podem ser obtidos a partir da Tabela 2.3.23. A inércia de um eixo para um dado bloco é a soma dos quadrados das coordenadas principais deste eixo neste bloco dividido pela inércia do primeiro eixo de cada bloco, ou seja, a inércia eixo-bloco vale: $In_s^k = \frac{1}{\lambda_1^k} \sum_j G_{jk}^2$ para cada eixo s . Este valor será sempre menor ou igual a 1, e quando 1 significa que o primeiro eixo coincide com o primeiro eixo a ACI do bloco. Os valores destes autovalores para as duas sub-tabelas são muito próximos de 1, indicando que o primeiro eixo da análise global esta relacionada aos dois primeiros eixos, de cada bloco, das análises parciais.

A Tabela 2.3.23 destaca as coordenadas principais calculadas para cada coluna de cada bloco, a distância de cada categoria até o centro, e também a contribuição de cada categoria para a inércia do eixo. Os rendimentos mais elevados são os que mais contribuem para a inércia, justamente porque apresentam as menores frequências observadas, distanciando-se da média.

Tabela 2.3.23: Coordenadas e contribuição de cada coluna em cada bloco para os eixos

Bloco	Rendimento	Distância até origem	Coordenada			Contribuição		
			eixo 1	eixo 2	eixo 3	eixo 1	eixo 2	eixo 3
Homem	Menos de 1 S.M. (H)	0,0185	-0,1023	-0,0889	0,0094	1,7467	13,3004	14,1410
	1 a 3 S.M. (H)	0,0627	-0,2416	-0,0657	-0,0041	9,7465	7,2693	2,7044
	3 a 5 S.M. (H)	0,0074	-0,0698	0,0500	-0,0082	0,8141	4,2075	10,5990
	5 a 10 S.M. (H)	0,0342	0,1323	0,1289	0,0031	2,9251	27,9861	1,4968
	Mais de 10 S.M. (H)	0,2093	0,4563	-0,0329	0,0087	34,7645	1,8195	11,9385
Mulher	Menos de 1 S.M. (M)	0,0501	-0,1921	-0,1145	0,0077	3,6576	13,0995	5,6255
	1 a 3 S.M. (M)	0,1092	-0,3296	0,0224	0,0023	10,7709	0,5010	0,4842
	3 a 5 S.M. (M)	0,0134	0,0258	0,1128	-0,0064	0,0658	12,7135	3,8004
	5 a 10 S.M. (M)	0,1057	0,3232	0,0325	0,0119	10,3565	1,0580	13,3822
	Mais de 10 S.M. (M)	0,2722	0,5037	-0,1344	-0,0195	25,1523	18,0453	35,8280

Tabela 2.3.24: Cossenos quadrados e correlação entre fator e a variável ajustada.

Bloco	Rendimento	$\cos^2 \theta$			Correlação entre fator e variável ajustada		
		eixo 1	eixo 2	eixo 3	eixo 1	eixo 2	eixo 3
Homem	Menos de 1 S.M. (H)	0,5669	0,4283	0,0048	-0,7529	-0,6544	0,0695
	1 a 3 S.M. (H)	0,9308	0,0689	0,0003	-0,9648	-0,2625	-0,0165
	3 a 5 S.M. (H)	0,6551	0,3359	0,0090	-0,8094	0,5796	-0,0948
	5 a 10 S.M. (H)	0,5129	0,4869	0,0003	0,7161	0,6977	0,0166
	Mais de 10 S.M. (H)	0,9945	0,0052	0,0004	0,9972	-0,0719	0,0190
Mulher	Menos de 1 S.M. (M)	0,7369	0,2619	0,0012	-0,8585	-0,5117	0,0345
	1 a 3 S.M. (M)	0,9954	0,0046	0,0000	-0,9977	0,0678	0,0069
	3 a 5 S.M. (M)	0,0494	0,9476	0,0030	0,2223	0,9734	-0,0548
	5 a 10 S.M. (M)	0,9886	0,0100	0,0013	0,9943	0,1001	0,0367
	Mais de 10 S.M. (M)	0,9322	0,0664	0,0014	0,9655	-0,2576	-0,0374

A tabela 2.3.24 destaca a contribuição do eixo para a inércia da categoria de coluna ($\cos^2 \theta$) e também a correlação entre o eixo e a variável ajustada que é obtida por $\sqrt{\cos^2 \theta}$ mantendo o sinal da coordenada. Estas correlações são usadas como coordenadas para o gráfico simétrico de representação, pois tornam menores as diferenças de escala entre linhas e colunas mantendo as relações encontradas entre ambas (Figura 2.3.1).

Figura 2.3.1 – Mapa de AFMTC no primeiro plano com as projeções de linhas e colunas simultaneamente.

Na Figura 2.3.1 as linhas azuis referem-se às informações de “anos de estudos” e, as linhas vermelhas representam as variáveis de “rendimento”. Nas linhas contínuas temos homens e nas tracejadas, mulheres. Observamos que à medida que os níveis de estudos vão aumentando há um aumento dos rendimentos, tanto para homens como para mulheres. É evidente também que, as linhas que ligam as categorias de escolaridade das mulheres são mais curtas que as dos homens, até a categoria de 11 a 14 anos de estudos, mostrando que existe uma maior homogeneidade de rendimentos entre estas mulheres até esta categoria de idade, ou seja, mulheres com níveis de escolaridade diferentes não apresentam grandes diferenças com relação aos perfis de rendimento. Entre os homens estas distâncias são maiores. Acompanhando os pontos que representam as categorias de escolaridade e rendimentos das mulheres, observamos que, mulheres com ‘menos de 8 anos’ de estudo estão entre os intervalos de ‘menos de 1’ SM e ‘de 1 a 3’ SM. As mulheres ‘de 8 a 10 anos’ de estudos estão entre os intervalos ‘de 1 a 3’ SM e ‘de 3 a 5’ SM, porém mais próximas de rendimento ‘de 1 e 3’ SM. As mulheres ‘de 11 a 14 anos’ de estudo estão entre os intervalos ‘de 1 a 3’ SM e ‘de 3 a 5’ SM, aproximando-se mais de rendimentos ‘de 3 e 5’ SM.

Com relação aos homens, os que apresentam ‘menos de 8 anos’ de estudos estão mais próximos de rendimentos menores do que 1 SM, os ‘de 8 a 10 anos’ de estudos se

localizam entre 'de 1 e 3' SM e 'de 3 a 5' SM, aproximando-se mais de rendimentos 'de 1 e 3' SM. Os homens 'de 11 a 14 anos' de estudos estão entre 'de 3 a 5' SM e 'de 5 a 10' SM, mais próximos do último.

Ressaltamos que as mulheres com rendimentos próximos ou superiores a 5 SM são aquelas que apresentaram mais do que 14 anos de estudo, e os homens com este rendimentos apresentaram níveis de escolaridade inferiores. Também observamos que há uma proximidade entre as rendas 'de 5 a 10' SM e 'mais de 10' SM entre as mulheres, e entre os homens esta distância é maior. Isto significa que o perfil educacional das mulheres nestas categorias de rendimentos são mais semelhantes e, além disso, são elevados os seus níveis de escolaridade. Já para os homens nestes intervalos de rendimentos há uma proximidade com duas categorias de escolaridade, mais de 14 anos de estudos e entre 11 e 14 anos de estudos.

Concluimos então que há uma diferença entre o perfil educacional e rendimentos de homens e mulheres trabalhadores da Região Metropolitana de São Paulo: para um mesmo rendimento as mulheres apresentam escolaridade mais elevada que os homens. E para as escolaridades mais elevadas, as mulheres apresentam principalmente duas faixas de rendimento, enquanto que para os homens a maior faixa de escolaridade está relacionada à maior faixa de rendimento. A AFMTC mostrou que as relações entre as categorias destas duas variáveis ocorrem de forma diferente entre homens e entre mulheres. As mulheres apresentam uma dificuldade maior para alcançar salários mais altos e os homens atingem melhores salários mesmo que apresentem menor grau de escolaridade do que as mulheres.

3. Aplicação

Retrato das relações entre Área do Conhecimento Científico e Setor de Atividade Econômica: Sistema Nacional de Inovação em Saúde

O objetivo neste capítulo é apresentar a aplicação das técnicas descritas no Capítulo 2 para o exemplo que motivou este estudo. A aplicação das três formas de análise descritas no Capítulo 2 servem para caracterizar as relações entre as duas variáveis descritas anteriormente, “Área do Conhecimento Científico” e “Setor de Atividade Econômica” que são questões do Gpesq4.

As duas variáveis apresentam multiplicidade de respostas, cada uma delas contendo até três respostas para um mesmo respondente. Além disso, a variável “Setor de Atividade Econômica” apresenta as respostas em ordem de importância.

O fato das variáveis apresentarem esta peculiaridade de multiplicidade e também ordem de importância para uma delas, nos fez optar por quatro formas de análise. A primeira consiste em contar cada par de área-setor e construir uma tabela de frequências na qual foi aplicada ACS, ressaltando que esta não é uma tabela de contingência pois para cada variável, o respondente foi contado mais de uma vez. Na segunda análise tabulamos todas as combinações que ocorreram entre áreas e também entre setores, algumas vezes agrupando combinações mais raras, configurando uma tabela de contingência na qual as categorias de respostas são diferentes das originais, e aplicamos novamente ACS. A terceira análise consiste em criar indicadores para cada área e para cada setor, para a construção da hiper-tabela de Burt usada na ACM, observando que cada variável passa a ser um indicador de presença ou ausência daquela categoria para cada respondente. Por último, usamos as áreas definidas na segunda análise (combinações de áreas) para construção de uma hiper-tabela onde cada bloco de coluna é formado por “Setores de Atividade Econômica” em ordem de importância na qual, para estas três tabelas justapostas, usamos a técnica de AFMTC.

Neste capítulo, além das técnicas descritas no Capítulo 2, usamos também análise de agrupamento hierárquico, com o intuito de ajudar a tipificar os setores, segundo as áreas, ou

mesmo ajudar a interpretar as distâncias visualizadas nos mapas produzidos pelas diferentes análises (Seções 3.1, 3.2, e 3.3).

Este tipo de agrupamento é uma fusão sucessiva dos objetos, na qual se juntam os grupos mais próximos, passo a passo, até que se tenha apenas um grupo (Bussab *et al.*, 1990). Depois se decide onde o agrupamento deve ser interrompido, optando-se por algum número de grupos que discrimine bem os dados. O critério de agrupamento usado neste estudo foi o método de Ward, que maximiza a variância entre os grupos e minimiza a variância dentro dos grupos. Para saber mais sobre esta técnica, ver Lebart *et al.* (1984, 1987).

Depois de concluído o agrupamento de setores, nas Seções 3.1, 3.2 e 3.3, foi usada a análise de resíduos para caracterizar os grupos de setores, que consiste em calcular o resíduo, ou seja, a diferença entre o observado e o esperado (sob a hipótese de independência), padronizar este resíduo e avaliar quais diferenças são relevantes para a composição da inércia (Pereira, 1999 e Lebart *et al.*, 1984, 1987).

Os resultados são apresentados em quatro sub-seções, cada uma contendo uma análise diferente. Todos os rótulos das tabelas e gráficos encontram-se no Anexo A.

3.1. Análise de Correspondência Simples da Tabela de Pares de Área-Setor

A ACS da tabela de pares de área-setor consiste em contar quantas vezes cada par ocorreu; as frequências dos pares área-setor podem ser vistas na Tabela 3.1.1.

Por exemplo, a frequência 311 significa que este número de indivíduos marcou a área de 'Ciências Agrárias' (A) e setor 'Produção Animal' (Pr.Anim). Devido ao fato dos respondentes terem a opção de marcar até três respostas para cada variável, a contagem total desta tabela supera o tamanho populacional, totalizando 24.837 ocorrências.

Frequências relativas (perfis linha de setores de atividade econômica) podem ser vistas na Tabela 3.1.2. Alguns setores têm mais do que metade dos registros em apenas uma área: para 'Produção Animal' 66,3% são da área de 'Ciências Agrárias', para os setores de 'Indústria Química' e 'Recursos Minerais' 59,9% e 54,3% são da área de 'Ciências Exatas e da Terra', para o setor 'Energia', 'Irrigação e Drenagem' e 'Indústria Metal-Mecânica' 55,8%, 58,3% e 63,4% são da área de 'Engenharia'.

Tabela 3.1.1: Frequências entre “área do conhecimento científico” e “setor de atividade econômica”, com possíveis pares de resposta

Setor	Área									Total
	A	B	S	Ex	H	So	En	L	nr	
Pr. Anim	311	108	20	5		1			24	469
Bio	238	1.002	401	145	4	2	71		212	2.075
I.Qui	19	23	19	175			46		10	292
I.C.Civ			3	6	1	3	6		1	20
Inform	13	18	68	166	21	11	59	3	51	410
Seg. Pub		5	13	5	18	7	3		15	66
Cult		9	84	4	181	18	3	15	68	382
Edu	31	264	1.033	129	563	83	54	31	400	2.588
I.Elet	1		3	43	1	3	51		4	106
Ene	2	7	4	52		1	86		2	154
C.Amb	87	494	180	289	62	21	109	1	147	1.390
Sau	486	2.428	4.661	425	649	139	349	24	1.535	10.696
Habit			3	2	10	12	11		7	45
Irriga. e Drenag	2			3			7			12
Trab	2	6	122	4	128	77	19	1	86	445
I.M-Mec	5		4	38	3	1	92		2	145
Met. e Clim	2	1	3	10	3		2		2	23
Rec. Minerais		1		19		5	6		4	35
N.Materiais	11	12	100	155	1	1	87		30	397
Norm. e Instrum	10	4	24	38	9	3	34		11	133
Nutr	256	111	322	25	13	3	23		86	839
I.Farm	27	280	201	212	3	1	50		92	866
Mec. Prec			3	6			8		2	19
Qual.Produt	105	24	77	28	18	21	35	1	43	352
D.Rur	72	19	18	21	36	27	14		24	231
Saneam	23	59	41	25	9	10	73		16	256
Rec. Mar	8	54		35	5	2	15		10	129
P. Espacial	1			11	1		11		1	25
Inf e Gest C&T	5	4	30	6	10	24	16		21	116
Telec		1	1	25	7	3	26		4	67
I.Têxtil	1	2		7		2	4			16
Transp			2	5	2	6	14			29
D.Urb	6	14	22	19	58	64	17	1	32	233
P.Veg	101	48	14	30	1		6		18	218
O.Set	114	337	431	140	195	96	75	20	150	1.558
Total	1.939	5.335	7.907	2.308	2.012	647	1.482	97	3.110	24.837

Fonte: Gpesq4

As Tabelas 3.1.1 e 3.1.2 apresentam 35 linhas e 9 colunas, dificultando bastante a visualização das possíveis relações entre as variáveis. Outra tabela poderia ser sugerida, por

exemplo, a de perfil-coluna, com as mesmas dimensões das duas anteriores. Para auxiliar na visualização das possíveis relações existentes nestas tabelas, sugerimos o uso da ACS na Tabela 3.1.1, por mais que a mesma não seja uma tabela de contingência, a ACS pode ser perfeitamente aplicada. A técnica de ACS foi aplicada diretamente na Tabela 3.1.1, da forma como foi explicado na Seção 2.1.

O mapa de ACS está ilustrado na Figura 3.1.1, onde a inércia total é 0,705 e o primeiro plano contém 63,7% dela (Tabela 3.1.3). A terceira dimensão contribui com 18,1% da inércia. A maior parte das categorias de setores apresenta boa qualidade de representação (contribuição para inércia) neste primeiro plano; oito delas têm menos de 50% da inércia explicada nestas duas dimensões. A qualidade de representação das áreas também é boa, mas 'Ciências Biológicas', 'Ciências Sociais Aplicadas' e 'Linguística, Letras e Artes' necessitam mais do que duas dimensões para explicar 50% da inércia.

Na Figura 3.1.1, os pontos azuis representam os setores e os triângulos vermelhos representam as áreas. Neste mapa podemos observar a proximidade entre o setor 'Saúde' e a área de 'Ciências da Saúde' e 'Ciências Biológicas', como esperado. Sua posição no mapa está justamente entre estas duas áreas, devido à maneira como os dados foram coletados. O setor 'Saúde' é o maior setor (10.696 registros) e nas Tabelas 3.1.1 e 3.1.2 podemos verificar que a maior ocorrência é na área de 'Ciências da Saúde', com 4.661 ocorrências (43,6%). Checando o percentual de 'Ciências da Saúde' neste setor, comparado com o percentual geral de 'Ciências da Saúde' (31,8%), podemos ver que aqui o percentual é muito mais elevado. No caso da área de 'Ciências Biológicas' o percentual do setor 'Saúde' é 22,7% (2.428 em 10.696) um pouco mais elevado do que o percentual em todos os setores, 21,5% (5.335 em 24.837).

No mapa da Figura 3.1.1 podemos identificar outras relações importantes, como por exemplo, a proximidade entre os setores de 'Produção Animal' e 'Produção Vegetal' direcionando-se para a área de 'Ciências Agrárias'. Já os setores de 'Nutrição', 'Biotecnologia', 'Qualidade e Produtividade' e 'Desenvolvimento Rural' estão na direção da área de 'Ciências Biológicas'.

Tabela 3.1.2: Perfil linha dos setores de atividade econômica nos pares de área-setor

Sector	Área									Total
	A	B	S	Ex	H	So	En	L	nr	
Pr. Anim	66,3	23,0	4,3	1,1		0,2			5,1	100
Bio	11,5	48,3	19,3	7,0	0,2	0,1	3,4		10,2	100
I.Qui	6,5	7,9	6,5	59,9			15,8		3,4	100
I.C.Civ			15,0	30,0	5,0	15,0	30,0		5,0	100
Inform	3,2	4,4	16,6	40,5	5,1	2,7	14,4	0,7	12,4	100
Seg. Pub		7,6	19,7	7,6	27,3	10,6	4,5		22,7	100
Cult		2,4	22,0	1,0	47,4	4,7	0,8	3,9	17,8	100
Edu	1,2	10,2	39,9	5,0	21,8	3,2	2,1	1,2	15,5	100
I.Elet	0,9		2,8	40,6	0,9	2,8	48,1		3,8	100
Ene	1,3	4,5	2,6	33,8		0,6	55,8		1,3	100
C.Amb	6,3	35,5	12,9	20,8	4,5	1,5	7,8	0,1	10,6	100
Sau	4,5	22,7	43,6	4,0	6,1	1,3	3,3	0,2	14,4	100
Habit			6,7	4,4	22,2	26,7	24,4		15,6	100
Irriga. e Drenag	16,7			25,0			58,3			100
Trab	0,4	1,3	27,4	0,9	28,8	17,3	4,3	0,2	19,3	100
I.M-Mec	3,4		2,8	26,2	2,1	0,7	63,4		1,4	100
Met. e Clim	8,7	4,3	13,0	43,5	13,0		8,7		8,7	100
Rec. Minerais		2,9		54,3		14,3	17,1		11,4	100
N.Materiais	2,8	3,0	25,2	39,0	0,3	0,3	21,9		7,6	100
Norm. e Instrum	7,5	3,0	18,0	28,6	6,8	2,3	25,6		8,3	100
Nutr	30,5	13,2	38,4	3,0	1,5	0,4	2,7		10,3	100
I.Farm	3,1	32,3	23,2	24,5	0,3	0,1	5,8		10,6	100
Mec. Prec			15,8	31,6			42,1		10,5	100
Qual.Produt	29,8	6,8	21,9	8,0	5,1	6,0	9,9	0,3	12,2	100
D.Rur	31,2	8,2	7,8	9,1	15,6	11,7	6,1		10,4	100
Saneam	9,0	23,0	16,0	9,8	3,5	3,9	28,5		6,3	100
Rec. Mar	6,2	41,9		27,1	3,9	1,6	11,6		7,8	100
P. Espacial	4,0			44,0	4,0		44,0		4,0	100
Inf e Gest C&T	4,3	3,4	25,9	5,2	8,6	20,7	13,8		18,1	100
Telec		1,5	1,5	37,3	10,4	4,5	38,8		6,0	100
I.Têxtil	6,3	12,5		43,8		12,5	25,0			100
Transp			6,9	17,2	6,9	20,7	48,3			100
D.Urb	2,6	6,0	9,4	8,2	24,9	27,5	7,3	0,4	13,7	100
P.Veg	46,3	22,0	6,4	13,8	0,5		2,8		8,3	100
O.Set	7,3	21,6	27,7	9,0	12,5	6,2	4,8	1,3	9,6	100
Total	7,8	21,5	31,8	9,3	8,1	2,6	6,0	0,4	12,5	100

Tabela 3.1.3: Distribuição da inércia ao longo dos eixos da ACS

Eixo (k)	Autovalor (λ_k)	%	% Acumulado
1	0,2461	34,91	34,91
2	0,2029	28,78	63,69
3	0,1276	18,10	81,78
4	0,0540	7,65	89,44
5	0,0439	6,23	95,67
6	0,0272	3,86	99,53
7	0,0025	0,36	99,88
8	0,0008	0,12	100,00
Total	0,7050	100	

O setor 'Educação' se direciona para a área de 'Ciências da Saúde' e também das áreas de 'Ciências Humanas' e 'Linguística, Letras e Artes'. Os setores de indústrias, 'Informática', 'Telecomunicações', 'Saneamento', etc, estão relacionados às áreas de 'Ciências Exatas e da Terra' e 'Engenharias'.

Com intuito de complementar a análise da Figura 3.1.1, e auxiliar numa tipificação das relações encontradas, aplicamos análise de agrupamento hierárquico pelo método de Ward aos setores (Lebart *et al.*, 1984) e desta forma podemos resumir a configuração dos pontos obtidos no mapa. O dendrograma relativo a esta análise está representado na Figura 3.1.2, e decidimos particioná-lo em cinco grupos que nos pareceram homogêneos.

Na Figura 3.1.2 o primeiro grupo (de baixo para cima) é formado pelos setores de 'Ciências Ambientais' que se juntou a 'Recursos do Mar', depois à 'Indústria Farmacêutica', 'Saneamento' e por último à 'Biotecnologia'. Este grupo está localizado na parte central do mapa da Figura 3.1.1, no segundo quadrante. O segundo grupo é formado somente pelo setor da 'Saúde'. O terceiro reúne outros oito setores e se localiza na parte inferior do mapa, quase que totalmente no quarto quadrante. O quarto grupo é formado pelos setores ligados à terra e por último, o quinto grupo, é formado pelos setores ligados à indústria.

Figura 3.1.1: Primeiro plano da ACS nos pares área-setor

A Tabela 3.1.4 e a Figura 3.1.3 mostram as comparações entre os grupos de acordo com cada área do conhecimento científico. Calculamos o resíduo padronizado na Tabela 3.1.4 para identificar quais áreas são as responsáveis pelas diferenças entre os grupos (Lebart et al., 1984). O sinal negativo indica que a frequência observada é menor do que a frequência esperada de acordo com os perfis gerais. O sinal positivo é o oposto. A Tabela 3.1.5 resume os setores em cada grupo e as áreas relacionadas com resíduo padronizado maior do que 2, que são aquelas que julgamos caracterizar tais grupos.

A Figura 3.1.3 destaca, para cada grupo, a proporção de registros nas áreas correspondentes a resíduos ajustados superiores a 2.

Na Figura 3.1.4 os grupos são representados por números de 1 a 5. O Grupo 1 tem cinco setores e representa 19,0% da população. Este grupo tem a maior proporção das áreas ‘Ciências Biológicas’, ‘Ciências Exatas e da Terra’ e ‘Engenharias’. Note que na população geral temos 21,5% de ‘Ciências Biológicas’ enquanto que neste grupo 40,1%. O resíduo

entre estas proporções (sob hipótese de independência) é de 34,5 desvios-padrão. A população geral tem 9,3% de ‘Ciências Exatas e da Terra’ e 6,0% de ‘Engenharias’, enquanto que neste grupo temos 15,0% e 6,7%, respectivamente. Notamos que este grupo está relacionado à biologia e meio ambiente.

O Grupo 2, com um setor, ‘Saúde’, tem a maior proporção da área de ‘Ciências da Saúde’ e tem também proporções significativas de ‘não-resposta’ para área, e para ‘Ciências Biológicas’. A população total tem 31,8%, 12,5% e 21,5% para ‘Ciências da Saúde’, ‘não-resposta’ e ‘Ciências Biológicas’, enquanto que neste grupo temos 43,6%, 14,4% e 22,7%, respectivamente. Este grupo está relacionado à saúde e biologia.

O Grupo 3, com oito setores, tem a maior proporção da área de ‘Ciências Humanas’. As áreas relacionadas são ‘Ciências Humanas’, ‘Ciências Sociais Aplicadas’, ‘Linguística, Letras e Artes’ e ‘não-resposta’ com, respectivamente, 8,1%, 2,6%, 0,4% e 12,5% para população total e 21,4%, 7,0%, 1,3% e 14,3% para o grupo. Podemos observar que este grupo está relacionado a humanas.

O Grupo 4, que inclui cinco setores, tem a maior proporção de ‘Ciências Agrárias’ (40,1% no grupo contra 7,8% na população). Está relacionado à agricultura e alimentação.

Tabela 3.1.4: Distribuição das áreas nos grupos para pares área-setor, segundo perfil-linha e resíduo padronizado

Grupo		Área do Conhecimento Científico									Total
		A	B	S	Ex	H	So	En	L	nr	
1	Perfil linha	8,1	40,1	17,5	15,0	1,8	0,8	6,7	0,0	10,1	100
	Resíduo padronizado	0,9	34,5	-23,6	14,9	-17,7	-8,8	2,5	-4,5	-5,5	
2	Perfil linha	4,5	22,7	43,6	4,0	6,1	1,3	3,3	0,2	14,4	100
	Resíduo padronizado	-16,7	4,1	34,5	-25,1	-10,2	-11,2	-15,6	-3,7	7,6	
3	Perfil linha	2,9	11,8	32,0	5,7	21,4	7,0	3,6	1,3	14,3	100
	Resíduo padronizado	-15,2	-19,7	0,3	-10,4	40,7	23,1	-8,2	11,5	4,6	
4	Perfil linha	40,1	14,7	21,4	5,2	3,2	2,5	3,7	0,0	9,2	100
	Resíduo padronizado	57,7	-7,9	-10,8	-6,8	-8,6	-0,4	-4,6	-2,6	-4,8	
5	Perfil linha	3,6	3,7	12,4	40,3	2,6	2,1	28,6	0,2	6,6	100
	Resíduo padronizado	-7,1	-19,6	-18,8	48,2	-9,1	-1,5	43,2	-1,7	-8,1	
Total Perfil linha		7,8	21,5	31,8	9,3	8,1	2,6	6,0	0,4	12,5	100

Tabela 3.1.5: Caracterização dos grupos para pares área-setor

Grupo	Setor	Área
1 (19,0% n=4.716)	C.Amb	B
	Bio	Ex
	I.Farm	En
2 (43,1%, n=10.696)	Sau	S
		nr
		B
3 (21,9%, n=5.433)	Edu	H
	O. Set	So
	Seg.Pub	L
	Trab	nr
4 (8,5%, n=2.109)	Nutr	D. Rur
	P. Veg	Pr.Anim
	Qual.Produt	A
5 (7,6%, n=1.883)	N.Materiais	I.Qui
	Norm. e Instrum	I.C.Civ
	Telec	Met
	Inform	Min
	Mec. Prec	Ene
	P.Espacial	Irr
	I. Textil	Mec
	I.Elet	Tra
		Ex
		En

Finalmente, o Grupo 5 é composto por dezesseis setores e tem a maior proporção de ‘Ciências Exatas e da Terra’ e ‘Engenharias’. As proporções na população são 9,3% e 6,0% respectivamente, contra 40,3% e 28,6% no grupo. Este grupo está relacionado a tecnologia.

Na análise de resíduos da Tabela 3.1.4, conseqüentemente na Figura 3.1.3 também, podemos observar que as áreas características de cada grupo são justamente aquelas que mais se aproximam dos setores destes grupos na Figura 3.1.4 (representadas por triângulos). E quanto mais próximas, ou mais ao centro dos setores do grupo a área estiver, maior é a força desta relação e com isso maior é o seu resíduo padronizado.

Figura 3.1.2: Dendrograma do agrupamento para pares área-setor

Figura 3.1.3: Distribuição das áreas nos grupos para pares área-setor

Figura 3.1.4: Primeiro Plano de ACS nos pares área-setor separados por grupos

3.2. Análise de Correspondência Simples das Combinações de Áreas e Combinações de Setores

Esta análise inicia-se na busca das combinações entre as áreas, assim como das combinações entre os setores. Note que cada linha de pesquisa marcou pelo menos um setor de atividade econômica, porém não necessariamente área do conhecimento científico. Isto significa que a área pode não apresentar resposta.

A tabela de contingência para esta análise é a Tabela 3.2.1 e os perfis-linha correspondentes estão na Tabela 3.2.2. Alguns rótulos novos estão descritos no Anexo A.

Esta tabela apresenta o tamanho populacional (12.814) e podemos ver quando uma categoria aparece sozinha ou associada a outra(s). Esta é uma reflexão mais detalhada do que os dados contêm.

A maioria dos pesquisadores trabalha exclusivamente na área de 'Ciências da Saúde' - 4.280 em 12.814 linhas de pesquisa - isto representa 33,4% das linhas de pesquisa. A segunda categoria de área mais freqüente é 'Ciências Biológicas' (19,2% do total). É interessante notar que 1.863 respondentes não declararam a área de conhecimento científico.

Aplicamos ACS na Tabela 3.2.1 e os resultados estão representados no mapa da Figura 3.2.1. A inércia do primeiro plano concentra 69,9% da inércia total, que foi de 0,7007 (Tabela 3.2.3).

Tabela 3.2.1: Tabela de contingência para “área do conhecimento científico” e “setor de atividade econômica”, nas combinações de áreas e combinações de setores

Setor	Área																Total
	S	B	Ex	A	H	En	So	L	SB	SH	ExEn	AB	S2	S3	O	nr	
Sau	2.489	892	34	90	126	42	24	1	172	29		12	21	7	12	816	4.767
Bio	35	121	11	21		2			5			3		1	1	28	228
C.Amb	8	98	34	3	1	3			5		2	1	1		8	26	190
Edu	41	12	13		73		11	1		8					4	26	189
Nutr	33	8	1	32	1	2			5	1			1	2	11	20	117
O. Set	57	83	52	25	31	7	21	8		2	2	2	3	3	8	33	337
Sal+Edu	474	57	5	1	120	3	5	4	16	34		2	3	1	10	169	904
Sal+Bio	76	298	6	8		8			74		2	5	3	3	5	83	571
Sau+Out	102	80	9	23	24	7	11	2	12	4				5	1	30	310
Sau+Nut	123	16	2	26	1				8			2	9	1		25	213
Sau+Edu+Out	104	13			38	13	4	1		1	2		1	1	1	26	205
Sal+Ind.Farm+Bio	21	81	15	5		8			17				2	3	9	17	178
Sal+Ind.Farm	30	51	14			22			10						7	37	171
Sal+C.Amb	21	60	22	3	1	1			10		2	1	5	1	4	28	159
Sal+Trab	47	2	1		38	2	6	1		3			1	1	4	24	130
Sau+Cult+Edu	33	3			44		1	3		12			1		3	19	119
Sal+Bio+Out	32	50	4	4	1		1		1				1		2	5	101
Sal+Bio+Edu	16	52	2	1	1	1			10	1		1		1		15	101
Sal+Edu+Nut	47	6		9	3				2	1			1			11	80
Sal+Edu+C.Amb	12	15	3	7	7	2			3	4				3		14	70
Sal+Bio+Pr.Anim	1	14		42								9	1	1		2	70
Sal+Bio+C.Amb	6	39	5	5					8						1	3	67
S2	98	33	30	45	23	56	7	1	5	5	8	5	27	7	8	51	409
S3	269	147	116	132	54	93	30	1	27	9	25	4	47	22	33	160	1.169
Os1	35	22	132	78	31	75	27	2	2	1	7	4	2		15	40	473
Os2	34	118	154	106	65	62	21	8	3	4	15	20	8	2	17	78	715
Os3	36	86	157	110	73	88	43	1	4	3	22	5	12	3	51	77	771
Total	4.280	2.457	822	776	756	497	212	34	399	122	87	76	150	68	215	1.863	12.814

Cinco categorias de combinações de áreas não apresentam boa qualidade de representação no primeiro plano. São elas:

- i) ‘Linguística, Letras e Artes’;
- ii) ‘Ciências da Saúde mais Ciências Humanas’;
- ii) ‘Ciências Agrárias mais Ciências Humanas’;
- iv) ‘Ciências da Saúde mais uma área diferente de Ciências Biológicas e Ciências Humanas’;
- v) ‘Ciências da Saúde mais duas outras áreas’.

Tabela 3.2.2: Perfil-linha para “setores de atividade econômica” nas combinações de áreas e combinações de setores

Setor	Área															Total
	S	B	Ex	A	H	En	So	L	SB	SH	ExEn	AB	S2	S3	O	nr
Sal	52,2	18,7	0,7	1,9	2,6	0,9	0,5	0,0	3,6	0,6	-	0,3	0,4	0,1	0,3	17,1
Bio	15,4	53,1	4,8	9,2	-	0,9	-	-	2,2	-	-	1,3	-	0,4	0,4	12,3
C.Amb	4,2	51,6	17,9	1,6	0,5	1,6	-	-	2,6	-	1,1	0,5	0,5	-	4,2	13,7
Edu	21,7	6,3	6,9	-	38,6	-	5,8	0,5	-	4,2	-	-	-	-	2,1	13,8
Nutr	28,2	6,8	0,9	27,4	0,9	1,7	-	-	4,3	0,9	-	-	0,9	1,7	9,4	17,1
O. Set	16,9	24,6	15,4	7,4	9,2	2,1	6,2	2,4	-	0,6	0,6	0,6	0,9	0,9	2,4	9,8
Sal+Edu	52,4	6,3	0,6	0,1	13,3	0,3	0,6	0,4	1,8	3,8	-	0,2	0,3	0,1	1,1	18,7
Sal+Bio	13,3	52,2	1,1	1,4	-	1,4	-	-	13,0	-	0,4	0,9	0,5	0,5	0,9	14,5
Sal+Out	32,9	25,8	2,9	7,4	7,7	2,3	3,5	0,6	3,9	1,3	-	-	-	1,6	0,3	9,7
Sal+Nut	57,7	7,5	0,9	12,2	0,5	-	-	-	3,8	-	-	0,9	4,2	0,5	-	11,7
Sal+Edu+Out	50,7	6,3	-	-	18,5	6,3	2,0	0,5	-	0,5	1,0	-	0,5	0,5	0,5	12,7
Sal+Ind.Farm+Bio	11,8	45,5	8,4	2,8	-	4,5	-	-	9,6	-	-	-	1,1	1,7	5,1	9,6
Sal+Ind.Farm	17,5	29,8	8,2	-	-	12,9	-	-	5,8	-	-	-	-	-	4,1	21,6
Sal+C.Amb	13,2	37,7	13,8	1,9	0,6	0,6	-	-	6,3	-	1,3	0,6	3,1	0,6	2,5	17,6
Sal+Trab	36,2	1,5	0,8	-	29,2	1,5	4,6	0,8	-	2,3	-	-	0,8	0,8	3,1	18,5
Sal+Cult+Edu	27,7	2,5	-	-	37,0	-	0,8	2,5	-	10,1	-	-	0,8	-	2,5	16,0
Sal+Bio+Out	31,7	49,5	4,0	4,0	1,0	-	1,0	-	1,0	-	-	-	1,0	-	2,0	5,0
Sal+Bio+Edu	15,8	51,5	2,0	1,0	1,0	1,0	-	-	9,9	1,0	-	1,0	-	1,0	-	14,9
Sal+Edu+Nut	58,8	7,5	-	11,3	3,8	-	-	-	2,5	1,3	-	-	1,3	-	-	13,8
Sal+Edu+C.Amb	17,1	21,4	4,3	10,0	10,0	2,9	-	-	4,3	5,7	-	-	-	4,3	-	20,0
Sal+Bio+Pr.Anim	1,4	20,0	-	60,0	-	-	-	-	-	-	-	12,9	1,4	1,4	-	2,9
Sal+Bio+C.Amb	9,0	58,2	7,5	7,5	-	-	-	-	11,9	-	-	-	-	-	1,5	4,5
S2	24,0	8,1	7,3	11,0	5,6	13,7	1,7	0,2	1,2	1,2	2,0	1,2	6,6	1,7	2,0	12,5
S3	23,0	12,6	9,9	11,3	4,6	8,0	2,6	0,1	2,3	0,8	2,1	0,3	4,0	1,9	2,8	13,7
Os1	7,4	4,7	27,9	16,5	6,6	15,9	5,7	0,4	0,4	0,2	1,5	0,8	0,4	-	3,2	8,5
Os2	4,8	16,5	21,5	14,8	9,1	8,7	2,9	1,1	0,4	0,6	2,1	2,8	1,1	0,3	2,4	10,9
Os3	4,7	11,2	20,4	14,3	9,5	11,4	5,6	0,1	0,5	0,4	2,9	0,6	1,6	0,4	6,6	10,0
Total	33,4	19,2	6,4	6,1	5,9	3,9	1,7	0,3	3,1	1,0	0,7	0,6	1,2	0,5	1,7	14,5

Das 31 categorias de setores, dez não apresentaram boa qualidade neste plano:

- i) ‘Educação’;
- ii) ‘Nutrição’;
- iii) ‘Saúde mais Outro Setor’;
- iv) ‘Saúde mais Nutrição’;
- v) ‘Saúde mais Indústria Farmacêutica’;
- vi) ‘Saúde mais Educação mais Nutrição’;
- vii) ‘Saúde mais Educação mais Ciências Ambientais’;
- viii) ‘Saúde mais Biotecnologia mais Produção Animal’;
- ix) ‘Saúde mais outro setor diferente daqueles que aparecem agrupados na Tabela 3.2.1’; e

x) 'Outro Setor'.

Tabela 3.2.3: Distribuição da inércia ao longo dos eixos da ACS das combinações de áreas e combinações de setores

Eixo (k)	Autovalor (λ_k)	%	% Acumulada
1	0,3278	46,78	46,78
2	0,1617	23,08	69,87
3	0,0797	11,37	81,24
4	0,0470	6,71	87,95
5	0,0287	4,10	92,05
6	0,0129	1,85	93,90
7	0,0114	1,63	95,53
8	0,0091	1,30	96,82
9	0,0066	0,95	97,77
10	0,0053	0,75	98,52
11	0,0036	0,51	99,03
12	0,0029	0,42	99,45
13	0,0017	0,24	99,69
14	0,0012	0,18	99,87
15	0,0009	0,13	100,00
Total	0,7007	100	

Na Figura 3.2.1, os pontos azuis representam os setores e os triângulos vermelhos, as áreas. Podemos observar que o setor 'Saúde' e a área de 'Ciências da Saúde' mais uma vez estão muito próximos e quando verificamos 'Saúde', 'Biotecnologia mais outro setor diferente' ou 'Saúde mais Biotecnologia' encontramos a área de 'Ciências Biológicas' e a área 'Ciências da Saúde mais Ciências Biológicas' próximas também, exceto para o setor 'Saúde mais Biotecnologia mais Produção Animal' que está mais próximo da área de 'Ciências Agrárias mais Ciências Biológicas', 'Ciências Exatas e da Terra' e 'Ciências Agrárias'. Notamos também que há alguns setores próximos das áreas de 'Ciências Humanas' e 'Ciências da Saúde mais Ciências Humanas'; estes setores se localizam na parte superior da Figura 3.2.1 e referem-se ao setor 'Saúde mais Cultura' ou 'Saúde mais Educação' e também ao setor 'Educação' isoladamente.

Para ajudar a tipificar os setores, aplicamos análise de agrupamento hierárquico e o dendrograma está representado na Figura 3.2.2. Podemos encontrar uma boa separação para três grupos. Os grupos são representados pelos números de 1 a 3 na Figura 3.2.3. A Tabela 3.2.4 e a Figura 3.2.4 mostram os percentuais de áreas para cada

grupo e a Tabela 3.2.5 mostra a caracterização destes grupos, com relação ao resíduo padronizado da Tabela 3.2.4 (maior do que 2).

Os setores isolados de 'Nutrição e Alimentação' e 'Outros Setores' foram alocados no Grupo 1, 'Biotecnologia' e 'Ciências Ambientais' no Grupo 2 e 'Educação' e 'Saúde' no Grupo 3. Em geral, combinações de setores envolvendo Saúde mais um ou mais dois setores e envolvendo Biotecnologia, encontram-se no Grupo 2, se os setores estão relacionados a humanas no Grupo 3 e quando temos uma mistura, no Grupo 1.

O Grupo 1, com 31,7% da população, tem como área mais significativa a área de 'Ciências Exatas e da Terra' (15,8% no grupo contra 6,4% na população). Seus setores estão ligados a áreas misturadas.

O Grupo 2 representa 13,8% da população e as áreas mais significativas estão ligadas a Saúde e Biológicas. 'Ciências Biológicas' representam 48,1% deste grupo, contra 19,2% na população total, enquanto que 'Ciências da Saúde mais Ciências Biológicas' 7,9% no grupo, contra 3,1% na população.

O Grupo 3 é o maior grupo, com 54,5% (6.987) dos respondentes. As áreas relacionadas a este grupo são 'Saúde' isolada, 'Ciências da Saúde mais Ciências Humanas', 'Ciências Humanas' isolada e 'não-resposta' para área, que representam 49,7%, 1,4%, 6,8% e 16,6% respectivamente para o grupo, contra 33,4%, 1,0%, 5,9% e 14,5% para a população. Este grupo está ligado a humanas e educação, indicando que a maioria das linhas de pesquisa estão nas universidades.

Figura 3.2.1: Primeiro Plano da ACS nas combinações de áreas e combinações de setores

Figura 3.2.2: Dendrograma do agrupamento para combinação de área e para combinação de setores

Figura 3.2.3: Primeiro plano de ACS para combinação de áreas e combinação de setores por grupos

Tabela 3.2.4: Distribuição de combinação de áreas por grupo

	Grupo 1		Grupo 2		Grupo 3		Total
	Perfil linha	Resíduo Aj.	Perfil linha	Resíduo Aj.	Perfil linha	Resíduo Aj.	
S	13,90%	-31,9	13,90%	-18,7	49,70%	42,8	33,4%
B	12,60%	12,9	48,10%	33,3	15,70%	-11	19,2%
Ex	15,80%	29,6	6,40%	0	1,00%	-27,6	6,4%
A	14,00%	25,8	2,80%	-6,1	2,20%	-19,9	6,1%
H	6,80%	3,1	0,20%	-10,9	6,80%	4,7	5,9%
Em	9,40%	22,2	2,50%	-3,1	1,00%	-18,6	3,9%
So	3,70%	12,2	0,10%	-5,7	0,90%	-7,5	1,7%
L	0,50%	3,8	0,00%	-2,3	0,20%	-1,9	0,3%
SB	1,10%	-8,8	7,90%	12,5	3,00%	-0,5	3,1%
SH	0,60%	-2,7	0,10%	-4,2	1,40%	5,4	1,0%
ExEn	1,90%	11,9	0,30%	-1,9	0,00%	-9,8	0,7%
AB	1,20%	6,2	0,60%	0,2	0,20%	-5,9	0,6%
S2	2,50%	9,4	0,70%	-2,1	0,50%	-7,4	1,2%
S3	1,00%	4,8	0,50%	-0,1	0,30%	-4,4	0,5%
O	3,50%	11,1	2,10%	1,5	0,50%	-11,4	1,7%
Nr	11,40%	-7	13,70%	-1,1	16,60%	7,3	14,5%
Total	100%		100%		100%		100%

Tabela 3.2.5: Caracterização dos grupos para combinação de áreas e combinação de setores

Grupo	Setor	Área
1 (31,7% n=4.061)	S2	O. Set
	S3	O1
	Sau+Bio+Pr.Anim	O2
	Nutr	O3
		Ex
		A
2 (13,8%, n=1.766)	Bio	En
	C.Amb	So
	Sau+Bio	ExEn
	Sau+C.Amb	O
	Sau+Ind.Farm	S2
		AB
3 (54,5%, n=6.987)		S3
		L
		H
		B
		SB
	Edu	Sau+Bio+Edu
	Sau	Sau+Bio+C.Amb
	Sau+Edu	Sau+Bio+Out
	Sau+Trab	Sau+Bio+Ind.Farm
	Sau+Nut	Sau+O
		Sau+Cult+Edu
		Sau+Edu+C.Amb
		Sau+Edu+Nut
		Sau+Edu+O

Figura 3.2.4- Distribuição da combinação de áreas nos grupos

A Figura 3.2.4 ilustra a mesma caracterização dos grupos exposta pela Tabela 3.2.5. Note-se que o Grupo 2 é o que apresenta um menor número de áreas características.

3.3. Análise de Correspondência Múltipla Considerando Todas as Categorias de Áreas e de Setores como Indicadores

Uma maneira de se considerar todas as categorias de resposta para as variáveis de múltipla resposta área e setor, é verificar a presença ou ausência de cada categoria. No máximo cada respondente pode apresentar três respostas para área e três para setor, no mínimo uma resposta para setor e nenhuma para área (pois há casos de não-resposta para área). Para tanto, criamos indicadores para cada categoria de resposta. Neste caso não estamos buscando relações entre duas variáveis categorizadas e sim entre muitas categorias de resposta, sem diferenciar tratar-se de área ou setor. Ou seja, nesta análise estamos considerando cada indicador como uma variável e que estes indicadores são homogêneos no sentido de que são todos binários.

Aplicamos ACM para estes 43 indicadores, e o mapa nos indica a associação entre todos eles. Temos um total de 12.814 linhas de pesquisa.

As Figuras 3.3.1 e 3.3.2 resumem a proporção de presença de cada um desses indicadores, em ordem decrescente, sendo que a primeira destaca áreas e a segunda, setores. A área ‘Ciências Biológicas’ e o setor ‘Educação’ são os indicadores com maior frequência depois de área e setor ‘Saúde’.

Figura 3.3.1: Proporção de presença para os indicadores de área

A Tabela 3.3.1 mostra as inércias para cada eixo e as inércias corrigidas para os eixos principais, da maneira como foi descrito na Seção 3.2. A Figura 3.3.3 mostra a porcentagem explicada da inércia corrigida por eixo.

Considerando-se a inércia ajustada para os eixos mais importantes, notamos que a proporção de inércia supera os 65% a partir do quinto eixo, e como nesta análise temos 43 indicadores, a redução de dimensionalidade para cinco eixos é bastante satisfatória.

Figura 3.3.2: Proporção de presença para os indicadores de setor

Tabela 3.3.1: Distribuição da inércia ao longo dos eixos para ACM

Eixo (k)	Inércia da Tabela de Burt (λ_k^B)	Inércia Corrigida (λ_k^*)	%	% Acumulado
1	0,003283	0,001215	30,2	30,2
2	0,002304	0,000642	16,0	46,2
3	0,001927	0,000447	11,1	57,3
4	0,001399	0,000210	5,2	62,5
5	0,001149	0,000119	3,0	65,5
6	0,000980	0,000068	1,7	67,2
7	0,000930	0,000055	1,4	68,5
8	0,000876	0,000042	1,0	69,6
9	0,000745	0,000017	0,4	70,0
10	0,000692	0,000010	0,2	70,3
11	0,000676	0,000008	0,2	70,5
12	0,000645	0,000005	0,1	70,6
13	0,000610	0,000002	0,1	70,6
14	0,000590	0,000001	0,0	70,7
15	0,000581	0,000001	0,0	70,7
16	0,000566	0,000000	0,0	70,7
17	0,000548	0,000000	0,0	70,7
18	0,000534			
19	0,000524			
20	0,000515			
21	0,000506			
22	0,000484			
23	0,000475			
24	0,000467			
25	0,000428			
26	0,000420			
27	0,000416			
28	0,000392			
29	0,000353			
30	0,000346			
31	0,000331			
32	0,000313			
33	0,000296			
34	0,000282			
35	0,000269			
36	0,000246			
37	0,000240			
38	0,000207			
39	0,000190			
40	0,000161			
41	0,000114			
42	0,000112			
43	0,000053			
Total	0,027181			

Figura 3.3.3: Distribuição percentual da inércia corrigida

Considerando que, se todos os indicadores tivessem a mesma contribuição, esta seria de 2,33% em cada um dos eixos (100% dividido por 43 eixos) e as contribuições de variáveis que superam este valor em algum dos eixos são, aqui, vistas como contribuições importantes.

Tabela 3.3.2: Percentual de contribuição para inércia de cada um dos cinco primeiros eixos, quando esta superou 2,33%, em ordem decrescente

Eixo	% Contribuição	Eixo	% Contribuição
1º Eixo		Ex	3,0
Sau	17,4	I.M-Mec	2,8
Ex	12,7	Cult	2,6
S	10,7	Nutr	2,5
En	9,9	Qual.Produt	2,4
I.M-Mec	5,8		
I.Qui	5,1	4º Eixo	
C.Amb	4,6	B	16,9
Ene	4,6	A	12,1
N.Materiais	3,4	C.Amb	11,4
I.Elet	3,3	Nutr	8,1
		Rec.Mar *	6,2
2º Eixo		S	5,5
H*	13,2	Pr.Anim	5,2
Edu*	11,2	I.M-Mec	4,8
Bio*	10,4	Qual.Produt	3,9
B*	10,0	Bio	3,0
A*	8,0	En	2,7
Cult*	7,1		
Pr.Anim*	6,2	5º Eixo	
So*	5,9	Ex	9,7
Trab*	5,4	Saneam *	8,9
D.Urb*	4,1	I.Qui	7,6
		D.Urb	5,8
3º Eixo		Ind.Farm	5,5
A	17,0	En	5,4
Pr.Anim	10,6	Edu	5,0
D.Rur *	6,7	Inform *	4,0
S	5,3	P.Espacial *	4,0
Sau	5,1	Transp *	4,0
So	5,0	Habit *	3,9
H	4,8	I.M-Mec	3,6
D.Urb	4,8	Cult	3,5
N.Materiais	3,6	I.C.Civ *	3,5
En	3,5	H	2,9
P.Veg *	3,4	C.Amb	2,8

* indicador ainda não apareceu em nenhum eixo anterior.

A Tabela 3.3.2 apresenta os indicadores e suas respectivas contribuições para a inércia dos cinco primeiro eixos, desde que a mesma supere os 2,33%. Estas contribuições foram apresentadas em ordem decrescente de importância.

O primeiro eixo concentra 30,2% da inércia e nele os indicadores de setores: 'Saúde', 'Indústria Metal-Mecânica', 'Indústria Química', 'Ciências Ambientais', 'Energia', 'Novos Materiais', e 'Indústria Eletro-Eletrônica'; mais os indicadores de

áreas: 'Ciências Exatas e da Terra', 'Ciências da Saúde', e 'Engenharias', apresentaram contribuições importantes.

O segundo eixo, com 16% da inércia, teve como variáveis “contribuidoras” para a inércia, os indicadores de setores: 'Educação', 'Biotecnologia', 'Cultura', 'Produção Animal', 'Trabalho', 'Desenvolvimento Urbano' e também as áreas: 'Ciências Humanas', 'Ciências Biológicas', 'Ciências Agrárias', e 'Ciências Sociais Aplicadas'. Notamos que até então nenhum destes indicadores apareceu como indicadores importantes para contribuição da inércia do primeiro eixo.

No terceiro eixo, com 11,1% da inércia, os setores e áreas se repetem muito e os setores de 'Desenvolvimento Rural' e 'Produção Vegetal' são dois com contribuição importante que ainda não tinham sido citados.

No quarto eixo, que explica 5,2% da inércia, aparece pela primeira vez o setor 'Recursos do Mar e Águas Internacionais'.

No quinto eixo, com 3,0% da inércia, o destaque se dá para os setores: 'Saneamento', 'Informática', 'Pesquisa Espacial', 'Transporte', 'Habitação' e 'Indústria da Construção Civil'.

Na Figura 3.3.4 percebemos a proximidade entre as áreas 'Ciências Sociais Aplicadas', 'Ciências Humanas' e 'Linguística, Letras e Artes', e observamos que 'Ciências Sociais Aplicadas' tem sinal negativo no eixo 1, oposto à 'Saúde'. Comparando os pontos azuis (setores) e rosas (áreas) e observando estes mesmos indicadores de áreas, vimos que os setores de 'Segurança Pública e Criminal', 'Cultura', 'Trabalho' e 'Educação' se aproximam deste grupo de áreas.

Na parte superior do mapa encontramos a área de 'Ciências Agrárias' juntamente com os setores de 'Produção Vegetal' e 'Produção Animal'.

A área 'Ciências Biológicas' se aproxima dos setores 'Biotecnologia' e 'Nutrição e Alimentação'. A área de 'Ciências da Saúde' tem o setor 'Saúde' muito próximo e se opõe aos demais indicadores.

Figura 3.3.4: Mapa de ACM no primeiro plano

Uma alternativa para resumir estas proximidades e afastamentos é o uso de análise de agrupamento. Neste trabalho optamos pelo método hierárquico de Ward aplicado em todos os eixos corrigidos e a Tabela 3.3.3 resume os resultados. Vale ressaltar que o agrupamento foi feito nas áreas e setores de acordo com suas coordenadas, neste caso teremos grupos que misturam áreas e setores (consideramos como categoria de setor ou área apenas a presença). O agrupamento feito desta maneira nos ajuda a interpretar o mapa no plano, devendo-se ter consciência de que a variabilidade total é pouco representada no plano, pois o agrupamento é feito com base em todos os eixos.

A Tabela 3.3.3 ilustra a formação de cada grupo. O primeiro grupo é formado por 11 indicadores, incluindo o setor 'Saúde' e a área de 'Ciências da Saúde'; este grupo se parece aos Grupos 1 e 2 somados, da análise na Seção 3.1 (pares área-setor). O segundo grupo é parecido com o Grupo 4, o terceiro é parecido com o terceiro da mesma análise e o quarto grupo se parece ao quinto da análise da Seção 3.1.

Tabela 3.3.3: Descrição dos grupos da ACM

Grupo 1	
Ind.Farm	Nutr
Bio	B
A	Qual.Produt
Sau	O.Set
Pr.Anim	S
P.Veg	
Grupo 2	
N.Materiais	Ex
Norm e Instrum	D.Rur
Saneam	Em
Inform	Rec.Mar
C.Amb	I.Qui
Grupo 3	
Seg.Pub	So
L	Trab
H	Cult
Inf e Gest C&T	D.Urb
Edu	Habit
Grupo 4	
I.M-Mec	Mec Pre
I.Elet	Ene
Rec.Minerais	Irrigac e Drenag
Telec	I.Textil
Met e Clim	I.C.Civ
Transp	P.Espacial

O mapa na Figura 3.3.5 ilustra a posição de cada indicador de acordo com o grupo a que ele pertence.

Podemos resumir a caracterização dos grupos da seguinte maneira:

Grupo 1: Indicadores ligados a setores de ‘Saúde’ e ‘Biotecnologia’ e preocupados com áreas de ‘Ciências Biológicas’ e ‘Saúde’.

Grupo 2: Indicadores de setores ligados a terra e áreas ligadas a Ciências Exatas.

Grupo 3: Indicadores de setores ligados a ‘Educação’, ‘Cultura’ e outros.

Grupo 4: Indicadores ligados a indústria e tecnologia.

Nesta análise as relações se dão entre setores-setores, setores-áreas e áreas-áreas e os grupos encontrados são muito parecidos com os grupos das duas análises anteriores.

Figura 3.3.5: Mapa de ACM no primeiro plano, destacando indicadores por grupo

No método pares área-setor (Seção 3.1) temos a formação de um grande grupo que é o setor 'Saúde' que se caracteriza pelas áreas das 'Ciências da Saúde' e 'Ciências Biológicas'. Nesta análise temos um grupo parecido, porém também representado pelo setor 'Biotecnologia' (Grupo 1). Na análise combinações de áreas e combinações de setores (Seção 3.2) temos o Grupo 2, de setores de 'Saúde' ligados a setores principalmente de Biotecnologia e outros e mais o setor de 'Biotecnologia' isolado. Este grupo foi caracterizado pelas áreas de 'Ciências Biológicas' e 'Ciências da Saúde mais Ciências Biológicas'.

Na Seção 3.1 o Grupo 3 caracteriza-se por apresentar os setores ligados a 'Cultura', 'Educação' e 'Trabalho'; o perfil tem relação com as áreas de 'Ciências Humanas', 'Linguística, Letras e Artes' e 'Ciências Sociais Aplicadas', muito parecido com o Grupo 3 da Seção 3.2, porém com perfil diferente, por este incluir área de 'Ciências da Saúde'. Estes também são parecidos com o Grupo 1 da presente análise, que inclui as áreas 'Ciências Biológicas' e 'Ciências da Saúde'.

3.4. Análise Fatorial Múltipla Incluindo a Importância Agregada aos Setores

Esta análise consiste em montar uma tabela com 16 linhas de combinações de áreas e 104 setores nas colunas, formando três blocos-colunas relativos à importância dos setores (veja tabela do Anexo B). Esta análise é a única, das quatro apresentadas, que leva em conta a ordem na escolha de setores. No entanto, a correlação entre as ordens não é considerada.

A ACI, realizada através de ACP na matriz transformada de acordo com a fórmula descrita na Seção 2.3, para cada tabela, primeiramente nos indica uma maior variabilidade para as respostas de primeira importância com relação aos setores, apesar desta tabela apresentar uma categoria a menos de setor. A inércia do primeiro eixo é maior para os setores de primeira ordem também, e já no primeiro plano os setores de segunda importância têm maior proporção de inércia acumulada (Tabela 3.4.1).

Tabela 3.4.1: Inércias e proporções de inércia total, no primeiro eixo e no primeiro plano

	Inércia Total	Inércia no 1º eixo	Inércia no 1º plano
ACI 1	1,0566	0,3489 (33,02%)	0,5801 (54,90%)
ACI 2	0,9214	0,2789 (30,27%)	0,5164 (56,04%)
ACI 3	0,7998	0,2214 (27,68%)	0,4081 (51,02%)

Figura 3.4.1: Mapa de ACI no primeiro plano, destacando áreas, para setores de importância 1

Com relação à tabela de importância 1, vimos que as áreas que mais se relacionam são ‘Ciências Exatas mais Engenharias’, ‘Ciências Exatas e da Terra’ e ‘Engenharias’, que estão no extremo do primeiro quadrante. Temos também ‘Ciências Agrárias’ próxima de ‘Ciências Agrárias mais Ciências Biológicas’. ‘Ciências da Saúde mais Ciências Humanas’ está entre as áreas ‘Ciências Sociais Aplicadas’, ‘Linguística, Letras e Artes’ e ‘Ciências Humanas’ e ‘Ciências da Saúde’ (veja segundo quadrante). No terceiro quadrante notamos a proximidade entre ‘Ciências Biológicas’ e ‘Ciências da Saúde mais Ciências Biológicas’; e Ciências da Saúde com mais uma ou duas áreas, diferente das citadas, estão próximas entre si e também de ‘Outras’ áreas.

Figura 3.4.2: Mapa de ACI no primeiro plano, destacando setores, para setores de importância 1

Com relação aos setores (Figura 3.4.2) observamos que ‘Educação’, ‘Cultura’ e ‘Trabalho’ são setores que mais se aproximaram das áreas ‘Ciências Humanas’, ‘Ciências Sociais Aplicadas’ e ‘Linguística, Letras e Artes’ e ‘Ciências da Saúde mais Ciências Humanas’. O setor ‘Saúde’ se aproxima da área ‘Ciências da Saúde’, extremo do segundo quadrante. Na parte inferior da Figura 3.4.2 encontramos os setores ‘Produção Animal’, ‘Biotecnologia’, ‘Nutrição e Alimentação’ e ‘Produção Vegetal’ que são os mais próximos das áreas ‘Ciências Agrárias’ e ‘Ciências Agrárias mais Ciências Biológicas’. Os setores de indústria se direcionam para as áreas ‘Ciências Exatas e da Terra’ e ‘Engenharias’.

Figura 3.4.3: Mapa de ACI no primeiro plano, destacando áreas, para setores de importância 2

A Figura 3.4.3 representa a análise da tabela de importância 2. As áreas que mais se relacionam são: ‘Ciências Exatas e da Terra’ e ‘Engenharias’; a que mais se aproxima destas é ‘Ciências Exatas mais Engenharias’, no extremo do terceiro quadrante. Temos também ‘Ciências Agrárias’ próxima de ‘Ciências Agrárias mais Ciências Biológicas’. ‘Ciências da Saúde mais Ciências Humanas’ está muito próxima a ‘Ciências Humanas’ e ‘Linguística, Letras e Artes’. Já ‘Ciências da Saúde’ isolada está próxima da ‘não-resposta’ (veja o quarto quadrante). No primeiro quadrante notamos a proximidade entre ‘Ciências Biológicas’ e ‘Ciências da Saúde mais Ciências Biológicas’ e também ‘Ciências da Saúde mais outras duas áreas’. Note que as relações que apareceram nesta análise são muito parecidas com as do mapa da tabela de importância 1.

Figura 3.4.4: Mapa de ACI no primeiro plano, destacando setores, para setores de importância 2

Na Figura 3.4.4 podemos observar um destaque para o setor de ‘Produção Animal’ que é o que mais se direciona no sentido de ‘Ciências Agrárias’ da Figura 3.4.3, seguido por ‘Produção Vegetal’, ‘Nutrição e Alimentação’ e ‘Biotecnologia’, similar à relação que observamos na tabela de importância 1.

Os setores de indústria também caminham na direção das ‘Engenharias’ e ‘Ciências Exatas e da Terra’ e os setores ‘Educação’, ‘Cultura’ e ‘Trabalho’ se direcionam no sentido das áreas de ‘Ciências Humanas’, ‘Ciências da Saúde mais Ciências Humanas’ e ‘Linguística, Letras e Artes’.

Figura 3.4.5: Mapa de ACI no primeiro plano, destacando áreas, para setores de importância 3

Na Figura 3.4.5, importância 3, temos que as áreas ‘Ciências Exatas e da Terra’ e ‘Engenharias’ são as que mais se aproximam de ‘Ciências Exatas e da Terra mais Engenharias’, em sentido extremo do quarto quadrante, porém esta relação não é tão forte quanto nas duas tabelas de importância anteriores.

‘Ciências Agrárias mais Ciências Biológicas’ está um pouco mais afastada de ‘Ciências Agrárias’ isolada, aproximando-se mais de ‘Ciências Biológicas’, ‘Ciências da Saúde mais Ciências Biológicas’, ‘Ciências da Saúde mais duas outras áreas’ e ‘Ciências da Saúde mais uma área’ (primeiro e segundo quadrantes).

‘Ciências Sociais Aplicadas’, ‘Ciências da Saúde mais Ciências Humanas’ e ‘Ciências Humanas’ isolada estão próximas e localizam-se no terceiro quadrante.

Figura 3.4.6: Mapa de ACI no primeiro plano, destacando setores, para setores de importância 3

Na Figura 3.4.6 observamos que o setor 'Produção Animal' é atraído pela área de 'Ciências Agrárias'. 'Trabalho' e 'Cultura' se direcionam para as áreas de 'Ciências Humanas' e 'Ciências Sociais Aplicadas', e 'Energia', 'Transporte' e 'Telecomunicações' se direcionam para 'Ciências Exatas e da Terra mais Engenharias'.

Nota-se que as relações são mais fracas, nesta tabela de importância três, talvez porque esta tabela tenha menos respostas e conseqüentemente menor inércia.

Por mais que a primeira tabela apresente maior inércia, as relações entre as áreas de conhecimento parecem ocorrer de uma maneira muito parecida nos três casos.

Tabela 3.4.2: Inércias e percentuais de inércia para cada eixo da AFMTC

Eixo	Inércia	%	% Acumulado
1	2,8376	28,5	28,5
2	2,3240	23,4	51,9
3	1,3364	13,4	65,4
4	0,8386	8,4	73,8
5	0,6849	6,9	80,7
6	0,5657	5,7	86,4
7	0,4392	4,4	90,8
8	0,2375	2,4	93,2
9	0,1847	1,9	95,1
10	0,1405	1,4	96,5
11	0,1071	1,1	97,5
12	0,0909	0,9	98,5
13	0,0755	0,8	99,2
14	0,0466	0,5	99,7
15	0,0307	0,3	100

A análise global apresenta 65,4% da inércia até o terceiro eixo fatorial e considerando-se que a matriz apresenta 16 linhas por 104 colunas, esta redução de dimensionalidade está muito boa.

A correlação entre as linhas das tabelas se mostrou bem alta, o que pode ser um indicativo de uma estrutura bem parecida entre elas (veja Tabela 3.4.3). O cálculo destas correlações é obtido facilmente, selecionando-se as primeiras colunas de F_1 , F_2 e F_3 e também de F e calculando-se as correlações de *Pearson* entre elas. Notamos que as maiores correlações ocorrem entre as primeiras coordenadas linhas parciais entre as tabelas de importância 1 e 2, seguida pela correlação entre 2 e 3 e por último entre 1 e 3. Com relação às primeiras coordenadas linha totais, esta apresentou maior correlação com as primeiras coordenadas linha da tabelas de importância 2.

Tabela 3.4.3: Correlação entre os primeiros eixos-linha entre as tabelas de importâncias 1, 2 e 3 e geral

	Importância 1	Importância 2	Importância 3	Todos
Importância 1	1,0000			
Importância 2	0,9261	1,0000		
Importância 3	0,8388	0,9135	1,0000	
Todos	0,9545	0,9823	0,9539	1,0000

‘Ciências Agrárias’ e ‘Ciências Agrárias mais Ciências Biológicas’ no geral se aproximam, porém quando analisados separadamente por importância de setor, notamos que esta relação é mais forte especialmente nas importâncias 1 e 2. Na prioridade três notamos a maior distância entre os pontos e uma proximidade maior de ‘Ciências Agrárias mais Ciências Biológicas’ com ‘Ciências da Saúde mais duas outras áreas’ e com ‘Ciências Biológicas’ isolada.

Figura 3.4.8: Primeiro plano da AFMTC para os setores

‘Ciências Exatas e da Terra mais Engenharias’ estão muito próximas de ‘Ciências Exatas’ na importância um; na importância dois isto também ocorre, já na três esta distância aumenta.

Observe na Figura 3.4.8 (setores) que há uma proximidade entre os setores ‘Educação’, ‘Cultura’, ‘Trabalho’ e ‘Desenvolvimento Urbano’ em todas as tabelas de importância, porém para a primeira delas não há uma proximidade destes setores com ‘Habitação’ e já na tabela de importância 3, sim. Observe que na tabela de primeira

importância há uma relação entre o setor 'Saúde' com 'Educação', 'Trabalho' e 'Cultura'. Na tabela de segunda importância temos que, o setor 'Saúde' se relaciona mais com 'Desenvolvimento Rural' e 'Outros Setores'. E para a tabela de importância três, há uma proximidade com os setores de 'Biotecnologia', 'Produção Vegetal' e 'Recursos do Mar'.

Observamos também que quando se trata da importância 1 há uma relação mais forte entre 'Produção Animal', 'Biotecnologia' e 'Nutrição e Alimentação', já na importância 2 a relação mais forte que ocorre com o setor de 'Produção Animal' é com 'Produção Vegetal'.

3.5. Conclusões da Aplicação

Com relação aos resultados do Capítulo 3, as quatro formas de analisar os dados neste projeto apresentaram resultados muito parecidos, os das Seções 3.1 e 3.3 são os mais similares; os resultados das Seções 3.2 e 3.4 mostraram um pouco mais de detalhe, com relações que não foram identificadas nas anteriores.

A análise realizada na Seção 3.3, ACM, é mais trabalhosa do que a da Seção 3.1, porém não agrega muita informação; observe que a matriz usada na Seção 3.1 é uma parte da matriz de Burt usada na Seção 3.3.

Os resultados apresentados nas Seções 3.2 e 3.4 foram capazes de destacar relações entre áreas e entre setores, que nas demais análises ficaram obscuras e, na última, foi possível identificar relações ligadas à importância atribuída aos setores. Vale ressaltar que o propósito da AFMTC é a comparação entre tabelas de contingência, então a resposta que a análise da Seção 3.4 nos fornece é com relação às diferentes estruturas de relação que podem ser encontradas de uma tabela para outra, ou seja, há diferenças entre as relações de áreas e setores quando olhamos a ordem de importância do setor? Duas linhas que são próximas em uma tabela se mantêm próximas em outra?

As Seções 3.2 e 3.4 são complementares, já que a primeira não identifica a ordem de importância, mas sim as relações entre áreas e entre setores de todas as combinações encontradas.

Realizar a análise da Seção 3.2 pode ser impraticável se o número de categorias for muito elevado ou se as respostas apresentarem muita multiplicidade.

Concluimos então, que os resultados do presente estudo, podem ajudar nas tomadas de decisão do administrador de Ciência e Tecnologia, podendo-se ter uma preocupação maior com os grupos de pesquisa que colaboram mais com a melhoria da qualidade de vida dos serviços de saúde, por exemplo. Se possível, o melhor é realizar as análises das seções 3.2 e 3.4.

4. Conclusões

A contribuição desta dissertação é importante devido ao fato destas técnicas de “análise fatorial” conjunta com análise de correspondência serem muito pouco usadas no Brasil e, mesmo na Europa, ainda serem muito recentes e estarem ainda em desenvolvimento. No Brasil muitas teses desenvolvem as técnicas de ACS e ACM e na tentativa de comparação de tabelas não há nada que use a ponderação entre as tabelas, como a AFMTC faz. Acreditamos que o fato de disponibilizarmos um documento em português com um exemplo didático e, além disso, com os programas desenvolvidos em R é muito importante principalmente porque os pacotes estatísticos usuais ainda não apresentam AFMTC implementada.

5. Anexo

Anexo A: Rótulos para áreas e setores usadas no Capítulo 3

Setor		Área	
O.Set	Outros Setores	A	Ciências Agrárias
Bio	Biotecnologia	B	Ciências Biológicas
C.Amb	Ciências Ambientais	S	Ciências da Saúde
Cult	Cultura	Ex	Ciências Exatas e da Terra
D.Rur	Desenvolvimento Rural	H	Ciências Humanas
D.Urb	Desenvolvimento Urbano	So	Ciências Sociais Aplicadas
Edu	Educação	En	Engenharias
Ene	Energia	L	Linguística, Letras e Artes
Habit	Habitação	nr	Sem resposta
I.C.Civ	Indústria da Construção Civi	O	Outras Áreas
I.Elet	Indústria Eletro-Eletrônica	S2	Ciências da Saúde mais uma outra área diferente Ciências Biológicas e Ciências Humanas
I.Farm	Indústria Farmaceutica		
I.M-Mec	Indústria Metal-Mecânica		
I.Qui	Indústria Química	S3	Ciências da Saúde mais duas outras áreas
I.Textil	Indústria Textil		
Inf e Gest C&T	Informação e Gestão C&T	Rec.Minerais	Recursos Minerais
Inform	Informática	Saneam	Saneamento
Irriga e Drenag	Irrigação e Drenagem	Sau	Saúde
Mec Prec	Mecânica de Precisão	Seg.Pub	Segurança Pública e Criminal
Met e Clim	Meteorologia e Climatologia	Telec	Telecomunicações
Norm e Instrum	Normalização e Instrumentação	Trab	Trabalho
N.Materiais	Novos Materiais	Transp	Transporte
Nutr	Nutrição e Alimentação	S2	Saúde mais um outro setor
P.Espacial	Pesquisa Espacial	S3	Saúde mais dois outros setores
Pr.Anim	Produção Animal	Os1	Um único outro setor, sem incluir Saúde
P.Veg	Produção Vegetal	Os2	Combinação de dois outros setores sem incluir Saúde
Qual.Produt	Qualidade e Produtividade		
Rec.Mar	Recursos do Mar e Águas Internacionais	Os3	Combinação de três outros setores sem incluir Saúde

Anexo B: Tabela para AFMTC

Área	Setor - Importância 1																												Total						
	Bio	C. Amb	Cult	D. Rur	D. Urb	Edu	Ene	Habit	I.C. Civ	I. Elet	Ind. Farm	I.M- Mec	I. Qui	Inf e Gest C&T	Infor- m	Irrigac e Drenag	Mec Prec	Met e Clim	Norm e Instrum	N. Mate- riais	Nutr	P.Espa- cial	Pr. Anim	P. Veg	Qual. Produ	Rec. Mar	Rec. Mine- rais	Sane- am		Sau	Seg. Pub	Telec	Trab	Transp	O. Set
S	104	28	18	3	3	343	1	-	1	-	47	-	4	4	7	-	-	-	-	12	123	-	3	3	1	-	-	3	3.494	2	-	12	-	64	4.280
B	539	277	-	2	3	64	-	-	-	-	42	-	2	-	1	-	-	1	1	-	26	-	21	8	4	5	-	5	1.339	1	-	-	-	116	2.457
Ex	63	164	1	3	6	49	16	-	-	13	76	11	68	3	74	-	3	2	11	93	3	3	1	4	10	1	1	2	84	-	2	1	-	54	822
A	143	31	-	24	-	3	-	-	-	-	4	-	8	-	9	1	-	-	2	1	119	-	135	49	32	3	-	5	174	-	-	-	-	33	776
H	1	21	76	11	13	264	-	1	-	-	-	-	-	-	-	-	-	1	5	1	2	-	-	-	2	2	-	2	277	2	1	27	-	47	756
En	27	38	1	5	8	10	44	-	2	22	7	55	8	5	9	1	3	-	8	37	6	2	-	1	13	7	-	29	111	1	10	6	2	19	497
So	1	7	7	5	29	42	-	-	1	-	-	-	-	9	2	-	-	-	1	1	-	-	-	-	2	-	3	-	51	1	-	22	-	28	212
L	-	-	8	-	-	13	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-	11	34
SB	68	19	-	-	-	8	-	-	-	-	13	-	1	-	-	-	-	-	-	-	14	-	-	-	1	-	-	-	275	-	-	-	-	-	399
SH	-	3	6	1	-	41	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-	-	-	1	64	1	-	-	-	3	122
ExEn	3	19	-	-	-	2	9	-	-	3	1	5	7	-	8	-	-	-	1	10	-	2	-	-	-	1	-	1	10	-	3	-	-	2	87
AB	23	4	-	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	18	2	2	-	-	2	19	-	-	-	-	2	76
S2	7	14	-	1	-	8	-	1	-	-	10	1	1	3	3	-	-	-	1	5	9	-	4	-	-	-	-	3	76	-	-	-	-	3	150
S3	6	6	-	3	1	5	1	-	-	-	2	-	-	1	2	-	-	1	-	-	3	-	2	2	-	-	-	-	30	-	-	-	-	3	68
O	27	37	5	4	6	30	1	-	-	2	4	-	1	2	6	1	-	-	1	1	14	-	-	5	2	1	-	2	39	2	1	10	-	11	215
nr	140	90	32	8	7	206	1	1	-	1	47	1	2	4	17	-	1	-	1	9	45	1	12	12	13	2	2	2	1.127	4	-	33	-	42	1.863
Total	1.152	758	154	73	76	1.088	73	3	4	41	253	73	102	31	138	3	7	5	32	170	367	8	196	86	82	22	6	57	7.172	14	17	111	2	438	12.814

Anexo B: Tabela para AFMTC (continuação)

Área	Setor - Importância 2																														Total					
	Bio	C. Amb	Cult	D. Rur	D. Urb	Edu	Ene	Habit	I.C. Civ	I. Elet	Ind. Farm	I.M- Mec	I. Qui	I Textil	Inf e Gest C&T	Infor- m	Irrigac e Drenag	Mec Prec	Met e Clim	Norm e Instrum	N. Mate- riais	Nutr	P.Espa- cial	Pr. Anim	P. Veg	Qual. Produ	Rec. Mar	Rec. Mine- rais	Sane- am	Sau		Seg. Pub	Telec	Trab	Transp	O. Set
S	113	43	20	4	5	474	1	-	2	-	50	-	3	-	6	20	-	1	1	9	43	114	-	2	1	35	-	-	17	426	4	-	57	-	131	1.582
B	212	97	2	7	7	96	2	-	-	-	90	-	4	1	2	5	-	-	-	-	3	36	-	31	21	10	30	1	25	440	2	-	2	-	95	1.221
Ex	19	33	1	11	4	31	13	1	2	17	65	9	62	4	1	32	-	-	2	15	28	6	2	1	5	9	17	12	3	99	3	7	-	1	30	545
A	34	23	-	21	1	7	1	-	-	-	13	2	6	-	2	1	-	-	1	4	6	67	-	104	29	35	1	-	5	126	-	-	1	-	37	527
H	1	9	51	8	15	140	-	2	-	-	-	2	-	-	3	10	-	-	-	3	-	4	-	-	-	6	1	-	1	142	6	2	38	1	48	493
En	10	19	1	1	3	8	16	6	2	14	30	21	11	2	8	13	-	2	-	13	17	2	1	-	1	13	3	2	24	97	1	3	2	5	15	366
So	-	1	-	10	10	17	1	6	1	1	-	1	-	2	6	2	-	-	-	1	-	-	-	-	-	5	-	1	4	21	2	-	13	-	24	129
L	-	-	2	-	1	6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	9	-	-	-	-	4	22	
SB	51	9	-	-	-	22	-	-	-	-	22	-	1	-	-	-	-	-	-	1	2	6	-	1	1	1	-	-	6	72	1	-	-	-	14	210
SH	-	-	8	1	1	32	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	1	-	-	-	-	-	-	-	33	-	-	1	-	3	81
ExEn	4	2	-	-	2	-	4	-	-	3	1	4	5	-	-	7	2	1	-	4	5	-	1	-	-	-	2	1	2	18	-	3	-	-	5	76
AB	11	5	-	-	-	2	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	7	-	11	1	-	-	-	2	12	-	-	-	-	2	54
S2	5	5	1	-	1	6	-	-	-	-	5	1	2	-	2	22	-	1	-	-	5	12	-	2	3	4	-	-	1	36	1	-	4	1	2	122
S3	4	1	-	1	3	4	-	-	-	-	4	-	-	-	-	3	-	-	1	-	-	3	-	2	1	2	-	-	2	18	-	-	1	-	5	55
O	8	9	8	9	5	12	-	1	1	-	26	1	4	-	-	8	-	-	-	-	-	2	-	2	1	4	5	-	3	34	-	1	10	-	2	156
nr	56	46	24	10	12	147	1	3	-	2	32	1	5	-	13	21	-	1	1	7	15	32	-	10	3	10	6	-	7	325	5	1	30	-	48	874
Total	528	302	118	83	70	1.004	39	19	8	37	339	42	103	9	43	145	2	6	6	57	124	292	4	166	67	134	65	17	102	1.908	25	17	159	8	465	6.513

Anexo B: Tabela para AFMTC (continuação)

Área	Setor - Importância 3																														Total					
	Bio	C. Amb	Cult	D. Rur	D. Urb	Edu	Ene	Habit	I.C. Civ	I. Elet	Ind. Farm	I.M- Mec	I. Quil	I. Textil	Inf e Gest C&T	Infor- m	Irrigac e Drenag	Mec Prec	Met e Clim	Norm e Instrum	N. Mate- riais	Nutr	P.Espa- cial	Pr. Anim	P. Veg	Qual. Produ	Rec. Mar	Rec. Mine- rais	Sane- am	Sau		Seg. Pub	Telec	Trab	Transp	O. Set
S	30	25	20	4	5	73	-	1	-	1	36	2	2	-	10	7	-	1	-	11	29	25	-	-	-	27	-	-	5	81	3	1	38	1	186	624
B	45	34	4	4	3	44	4	-	-	-	62	-	6	1	2	5	-	-	-	1	4	7	-	17	9	4	13	-	13	130	1	1	1	-	93	508
Ex	7	15	-	1	4	19	2	1	4	4	19	7	20	3	-	9	-	2	2	5	13	4	1	1	8	5	7	4	6	85	1	6	2	2	34	303
A	14	7	-	15	2	15	-	-	-	-	8	2	3	1	-	-	-	-	1	4	4	21	1	29	11	34	1	-	5	101	-	-	1	-	35	315
H	-	6	16	6	14	42	-	4	-	-	1	1	-	-	-	3	-	-	-	-	-	2	-	-	-	2	-	-	2	62	5	2	32	1	58	259
En	6	15	1	2	2	24	6	4	2	5	4	4	7	2	3	8	3	1	-	7	5	-	4	-	-	4	1	1	8	50	-	3	3	4	18	207
So	-	3	2	3	6	3	-	4	-	-	-	-	-	-	1	3	-	-	-	-	-	-	-	-	-	3	-	-	1	17	-	1	15	4	19	85
L	-	-	1	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3	-	-	1	-	1	7	
SB	6	12	-	-	1	4	-	-	-	-	8	-	1	-	-	-	-	-	-	-	1	5	-	-	1	-	-	-	2	28	-	-	2	-	1	72
SH	1	4	10	-	-	2	-	1	-	-	-	-	-	-	2	-	-	-	-	1	-	-	-	-	-	1	-	-	-	6	-	-	3	-	3	34
ExEn	5	1	-	-	-	1	6	-	-	1	-	1	3	-	-	-	-	-	2	-	1	1	1	-	-	-	-	1	5	11	-	3	-	1	5	49
AB	2	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	1	-	-	1	-	-	10	-	-	-	-	2	19	
S2	5	4	1	-	1	10	1	-	-	2	3	-	2	-	1	3	-	-	-	-	3	2	-	2	1	4	-	-	1	11	1	-	2	-	7	67
S3	1	6	-	-	1	1	-	-	-	-	1	-	1	-	1	-	-	-	-	1	-	1	-	2	1	1	-	-	-	9	-	-	2	-	6	35
O	3	3	1	2	4	3	-	1	-	1	5	-	4	-	2	3	-	-	-	-	1	3	1	-	4	3	3	1	3	27	2	1	6	2	15	104
nr	16	11	12	6	13	47	-	3	1	1	13	-	3	-	4	13	-	-	1	3	6	9	-	2	3	20	2	2	7	83	6	3	23	-	60	373
Total	141	146	68	43	56	290	19	19	7	15	160	17	52	7	26	54	3	4	6	33	67	82	8	54	38	108	28	9	58	714	19	21	131	15	543	3.061

Anexo C: Programa em R para análise da Seção 2.1.

```
# PROGRAMA PARA EFETUAR A ANÁLISE DE CORRESPONDÊNCIA SIMPLES #
# NO R – PASSO A PASSO #
# A TABELA DE ENTRADA DE DADOS DEVE SER UM ARQUIVO.TXT #
# SEPARADO POR TABULAÇÃO #

# LEITURA DA BASE DE DADOS – ARQUIVO.TXT SEPARADO POR TAB

N <- read.table("C:\\base.txt",sep="\t")

# CRIAR A MATRIZ P, COM RELAÇÃO AO TOTAL GERAL
n=sum(N)
P<-N/n
# Criar o vetor de perfil linha médio
lin<-array(0:0, dim=c(nrow(N),1))
for (i in 1:nrow(N)) {
  sum=0;
  for (j in 1:ncol(N)) {
    sum = sum + P[i,j];
  }
  lin[i,1] = sum
}
# Criar o vetor de perfil coluna médio
col<-array(0:0,dim=c(1,ncol(N)))
for (j in 1:ncol(N)) {
  sum=0;
  for (i in 1:nrow(N)) {
    sum = sum + P[i,j];
  }
  col[1,j] = sum
}

# Criar a matriz de perfis coluna
L <-array (0:0,dim=c(nrow(N),ncol(N)))
for (j in 1:ncol(N)) {
  for (i in 1:nrow(N)) {
    L[i,j] = P[i,j]/lin[i,1]
  }
}

# Criar a matriz de perfis linha
C <-array (0:0,dim=c(nrow(N),ncol(N)))
for (i in 1:nrow(N)) {
  for (j in 1:ncol(N)) {
    C[i,j] = P[i,j]/col[1,j]
```



```

    }
}

# Criar Matriz Diagonal (1 sobre Raiz de perfil linha médio)
Dlin<-array(0:0, dim=c(nrow(N),nrow(N)))
for (i in 1:nrow(N)) {
  Dlin[i,i]=1/sqrt(lin[i,1])
}

# Criar Matriz Diagonal (1 sobre de raiz de perfil coluna médio)
Dcol<-array(0:0, dim=c(ncol(N),ncol(N)))
for (j in 1:ncol(N)) {
  Dcol[j,j]=1/sqrt(col[1,j])
}

# Calcular a matriz sob hipótese de independência
M<-array(0:0, dim=c(nrow(N),ncol(N)))
for (i in 1:nrow(N)) {
  for (j in 1:ncol(N)) {
    M[i,j] = col[1,j]*lin[i,1]
  }
}

#Calcular matriz diagonal (1/Dlin)#
Dlin1<-array(0:0, dim=c(nrow(N),nrow(N)))
for (i in 1:nrow(N)) {
  Dlin1[i,i]=1/(lin[i,1])
}

#Calculando 1/Dcol#
Dcol1<-array(0:0, dim=c(ncol(N),ncol(N)))
for (j in 1:ncol(N)) {
  Dcol1[j,j]=1/(col[1,j])
}

# Calcular a inércia total
Dlin<-as.matrix(Dlin)
P<-as.matrix(P)
M<-as.matrix(M)
Dcol<-as.matrix(Dcol)
Dlin1<-as.matrix(Dlin1)
Dcol1<-as.matrix(Dcol1)
inercia<-sum(diag((Dlin1) %*% (P-M) %*% (Dcol1) %*% t(P-M)))

```



```

# Calcular SVD da matriz centrada
Dlin<-as.matrix(Dlin)
Dcol<-as.matrix(Dcol)

MSVD <- Dlin %*% (P-M) %*% Dcol

d<-svd(MSVD)$d
U<-svd(MSVD)$u
V<-svd(MSVD)$v

# Calcular Matrizes X e Y de coordenadas padrão
X<-Dlin %*% U
Y<-Dcol %*% V

# Calcular Matrizes F e G de coordenadas principais
D <-diag(d)

F<-X %*% D
G <- Y %*% D

# Calcular inércia explicada por categoria linha e eixo
somaln<-array(0:0, dim=c(1,nrow(N)))
for (j in 1:nrow(N)) {
  sum=0;
  for (i in 1:nrow(N)) {
    sum = sum + lin[i,1]*F[i,j]^2;
  }
  somalin[1,j] = sum
}
lexlin <-array (0:0,dim=c(nrow(N),nrow(N)))
for (i in 1:nrow(N)) {
  for (j in 1:nrow(N)) {
    lexlin[i,j] = 100*( lin[i,1]*F[i,j]^2/somalin[1,j])
  }
}

# Inércia explicada por categoria eixo-coluna
somacol<-array(0:0, dim=c(1,ncol(N)))
for (j in 1:nrow(N)) {
  sum=0;
  for (i in 1:ncol(N)) {
    sum = sum + col[1,i]*G[i,j]^2;
  }
  somacol[1,j] = sum
}

```



```

lexcol <-array (0:0,dim=c(ncol(N),nrow(N)))
for (i in 1:ncol(N)) {
  for (j in 1:nrow(N)) {
    lexcol[i,j] = 100*( col[1,i]*G[i,j]^2/somacol[1,j])
  }
}

```

Distâncias entre as linhas.

```

Distlin <-array (0:0,dim=c(nrow(N),nrow(N)))
for (i in 1:nrow(N)) {
  for (k in 1:nrow(N)){
    soma=0;
    for (j in 1:ncol(N)) {
      soma = soma +(L[i,j]-L[k,j])^2/col[j]
    }
    Distlin[i,k] = soma
  }
}

```

Distância entre a linha e o centro (média).

```

Distlin1 <-array (0:0,dim=c(nrow(N),1))
for (i in 1:nrow(N)) {
  soma=0;
  for (j in 1:ncol(N)) {
    soma = soma +(L[i,j]-col[j])^2/col[j]
  }
  Distlin1[i] = soma
}

```

Distâncias entre as colunas.

```

Distcol <-array (0:0,dim=c(ncol(N),ncol(N)))
for (j in 1:ncol(N)) {
  for (q in 1:ncol(N)){
    soma=0;
    for (i in 1:nrow(N)) {
      soma = soma +(C[i,j]-C[i,q])^2/lin[i]
    }
    Distcol[j,q] = soma
  }
}

```

Distâncias entre as colunas e o centro (média).

```

Distcol1 <-array (0:0,dim=c(1,ncol(N)))
for (j in 1:ncol(N)) {
  soma=0;
  for (i in 1:nrow(N)) {

```



```

    soma = soma +(C[i,j]-lin[i])^2/lin[i]
  }
  Distcol1[j] = soma
}

# RESULTADOS
# Matriz de dados, de proporção geral e de perfis Linha e Coluna
N
P
L
C
# Distancias entre linhas e entre colunas
Distlin
Distcol
# inércia total
inércia
# Inércia por eixo
d^2
# Coordenadas Padrão (linha e coluna)
X
Y
# Coordenadas principais (linha e coluna)
F
G
# % de inércia por linha-eixo e coluna-eixo
lexlin
lexcol

# MAPA DE CORRESPONDÊNCIA SIMÉTRICO
a<-min(F,G)
b<-max(F,G)

plot(c(a-0.2,b+0.2), c(a-0.2,b+0.2),pch= 2,xlab="Eixo 1",ylab="Eixo 2",type="n", lwd=2)

# colocar todos os labels de linhas entre "aspas"
text(F[,1],F[,2],labels= c("menos 8 anos","8 a 10 anos","11 a 14 anos","15 ou +
anos"),col="dark green", font=8, pos="3")
# colocar todos os labels de colunas entre "aspas"
text(G[,1],G[,2],labels= c("menos de 1","de 1 a 3","de 3 a 5","de 5 a 10","mais de
10"),col="blue", font=9, pos="3")
abline(0,0,0,0, lty=2, col="black",lwd=1)
lines(F[,1],F[,2],lty=3,col="dark green")
lines(G[,1],G[,2],lty=5,col="blue")

```



```
# Como rodar diretamente uma ACS:
```

```
library(mva)
```

```
library(MASS)
```

```
library(ctest)
```

```
corresp(N)
```

```
mapa = corresp(N,nf=2)
```

```
plot(mapa)
```

```
# corresp multipla: mca
```

Anexo D: Programa em R para análise da Seção 2.2.

```
# PROGRAMA PARA EFETUAR A ANÁLISE DE CORRESPONDÊNCIA #
```

```
# MÚLTIPLA NO R – PASSO A PASSO #
```

```
# A TABELA DE ENTRADA DE DADOS DEVE SER UM ARQUIVO.TXT #
```

```
# TEM QUE DEFINIR O VALOR Q QUE É O Nº DE VARIÁVEIS #
```

```
# SEPARADO POR TABULAÇÃO, COM A MATRIZ DE BURT #
```

```
# LEITURA DA BASE DE DADOS – ARQUIVO.TXT SEPARADO POR TAB #
```

```
# Inserir número de variáveis
```

```
Q<-3
```

```
# Inserir número de categorias
```

```
J<-nrow(N)
```

```
# Ler arquivo.txt separado por tabulação
```

```
N <- read.table("C:\\MBurt1.txt",sep="\t")
```

```
# Criar a matriz P, com relação ao total geral
```

```
n=sum(N)
```

```
P<-N/n
```

```
#Criar o vetor de perfil linha médio (igual perfil coluna)
```

```
lin<-array(0:0, dim=c(nrow(N),1))
```

```
for (i in 1:nrow(N)) {
```

```
  sum=0;
```

```
  for (j in 1:nrow(N)) {
```

```
    sum = sum + P[i,j];
```

```
  }
```

```
  lin[i,1] = sum
```

```
}
```

```
# Criar a matriz de perfis linha (igual ao das coluna)
```

```
L <-array (0:0,dim=c(11,11))
```



```

for (j in 1:11) {
  for (i in 1:11) {
    L[i,j] = N[i,j]/N[i,i]
  }
}

# Criar matriz diagonal com 1 sobre a raiz de perfil linha geral
Dlin<-array(0:0, dim=c(nrow(N),ncol(N)))
for (i in 1:nrow(N)) {
  Dlin[i,i]=1/sqrt(lin[i,1])
}

# Criar a matriz sobre a hipótese de independência
M<-array(0:0, dim=c(nrow(N),ncol(N)))
for (i in 1:nrow(N)) {
  for (j in 1:ncol(N)) {
    M[i,j] = lin[j,1]*lin[i,1]
  }
}

# Calcular inércia total
inercia<-J/Q - 1

# Calcular SVD da matriz centrada
P<-as.matrix(P)
MSVD <- Dlin %*% (P-M) %*% t(Dlin)

d<-svd(MSVD)$d
U<-svd(MSVD)$u
V<-svd(MSVD)$v

# Calcular as coordenadas padrão, matriz X
X<-Dlin %*% U

# Calcular as coordenadas principais, matriz F
D <-diag(d)
F<-X %*% D

# Inércia explicada por categoria eixo-linha/coluna
somalin<-array(0:0, dim=c(1,ncol(N)))
for (j in 1:ncol(N)) {
  sum=0;
  for (i in 1:nrow(N)) {
    sum = sum + lin[i,1]*F[i,j]^2;
  }
}

```



```

    somalin[1,j] = sum
  }

  lexlin <-array (0:0,dim=c(nrow(N),ncol(N)))
  for (i in 1:nrow(N)) {
    for (j in 1:ncol(N)) {
      lexlin[i,j] = 100*(lin[i,1]*F[i,j]^2/somalin[1,j])
    }
  }

  # Correção das Inércias #
  cont=0;
  parada=1/Q;
  for (i in 1:(nrow(N)-Q)) {
    if (d[i] > parada) {cont= cont+1} else {cont=cont}}

  In<-array(0:0, dim=c(cont,1))
  for (i in 1:cont) {
    In[i]= ((Q/(Q-1))^2)*(d[i]-(1/Q))^2
  }
  # Correção dos eixos F #
  F1<-array (0:0,dim=c(nrow(N),cont))
  for (j in 1:cont) {
    for (i in 1:nrow(N)) {
      F1[i,j]= F[i,j]*sqrt(In[j])
    }
  }

  # RESULTADOS
  # Matriz de dados, proporção e de perfil linha
  N
  P
  L
  # inércia total
  inercia
  # Inércia por eixo
  t(t(d^2))
  #Inércia corrigida
  In
  # Coordenadas padrão (linha/coluna)
  X
  # Coordenadas principais (linha/coluna)
  F
  # Coordenadas principais corrigidas
  F1

```



```
# % de inércia por linha-eixo e coluna-eixo
Iexlin
```

```
# MAPA DE CORRESPONDÊNCIA SIMÉTRICO
```

```
plot(c(min(F1[,1])-0.05,max(F1[,1])+0.05), c(min(F1[,2])-0.05,max(F1[,2])+0.05),pch=
2,xlab="Eixo 1",ylab="Eixo 2",type="n", lwd=2)
abline(0,0,0,0, lty=2, col="black",lwd=1)
```

```
# para separar as categorias de variáveis diferentes: fazer passo a passo – rótulos de cada
categoria separadamente e entre “aspas” #
```

```
text(F1[1,1],F1[1,2],labels= c("H"),col="blue", font=9, pos="3")
text(F1[2,1],F1[2,2],labels= c("M"),col="blue", font=9, pos="3")
```

```
text(F1[3,1],F1[3,2],labels= c("menos 8 anos"),col="dark green", font=9, pos="3")
text(F1[4,1],F1[4,2],labels= c("8 a 10 anos"),col="dark green", font=9, pos="3")
text(F1[5,1],F1[5,2],labels= c("11 a 14 anos"),col="dark green", font=9, pos="3")
text(F1[6,1],F1[6,2],labels= c("15 ou + anos"),col="dark green", font=9, pos="3")
```

```
text(F1[7,1],F1[7,2],labels= c("menos de 1"),col="dark red", font=9, pos="3")
text(F1[8,1],F1[8,2],labels= c("de 1 a 3"),col="dark red", font=9, pos="3")
text(F1[9,1],F1[9,2],labels= c("de 3 a 5"),col="dark red", font=9, pos="3")
text(F1[10,1],F1[10,2],labels= c("de 5 a 10"),col="dark red", font=9, pos="3")
text(F1[11,1],F1[11,2],labels= c("mais de 10"),col="dark red", font=9, pos="3")
```

```
# Ligar as categorias de uma mesma variável com uma mesma cor
```

```
lines(c(F1[1,1],F1[2,1]),c(F1[1,2],F1[2,2]),lty=5,col="blue")
lines(c(F1[3,1],F1[4,1],F1[5,1],F1[6,1]),c(F1[3,2],F1[4,2],F1[5,2],F1[6,2]),lty=5,col= "dark
green")
lines(c(F1[7,1],F1[8,1],F1[9,1],F1[10,1],F1[11,1]),c(F1[7,2],F1[8,2],F1[9,2],F1[10,2],F1[11,2
]),lty=5,col="dark red")
```


Anexo E: Programa em R para análise da Seção 2.3.

```
# PROGRAMA PARA EFETUAR A AFMTC PARA JUSTAPOSIÇÃO DE K      #
# TABELAS, ALINHADAS POR LINHA                                #
# INSTRUÇÕES: ESTE PROGRAMA DEVE SER RODADO EM 2 PASSOS      #
# SEGUIR AS INSTRUÇÕES DE CADA PASSO                          #

# PASSO 1: LER A MATRIZ N COM AS FREQUÊNCIAS DA TABELA N (TXT) #
# INFORMAR OS VALORES DE K (nº de sub-tabelas), J (vetor contendo o número de #
# colunas em cada sub-tabela Nk), IN_TAB (vetor com as posições iniciais das colunas de#
# cada sub-tabela), FIN_TAB (vetor com posições finais de cada sub-tabela)      #

      # PROGRAMA ADAPTADO PARA O CASO DE DUAS TABELAS #

# Entrada da Matriz de Dados com frequências de tabelas justapostas por linha
N <- read.table("C:\\N.txt",sep="\t")

# N° de tabelas (K) e n°s de linhas e colunas de N#
K=2
J<- array(c(5,5),dim=c(1,K))
# Inserir posições de inicio e fim de cada sub-tabela
in_tab<-array(c(1,J[1]+1))
fin_tab<-array(c(J[1],ncol(N))

#Calculando Matriz P, com frequências relativas ao total geral n... (P(ijk))#
SN=sum(N)
P<-N/SN

# Perfil Geral Linha -P(i.) #
lin<-array(0:0, dim=c(nrow(N),1))
for (i in 1:nrow(N)) {
  sum=0;
  for (j in 1:ncol(N)) {
    sum = sum + P[i,j];
  }
  lin[i,1] = sum
}

# Perfil Geral Coluna - P(.jk) #
col<-array(0:0, dim=c(1,ncol(N)))
for (j in 1:ncol(N)) {
  sum=0;
  for (i in 1:nrow(N)) {
    sum = sum + P[i,j];
  }
}
```



```

col[1,j] = sum
}
# Soma Marginal de cada Tabela k - P(i.k) #
lin_k<-array(0:0, dim=c(nrow(N),K))
for (k in 1:K) {
  for (i in 1:nrow(N)) {
    sum=0;
    for (j in in_tab[k]:fin_tab[k]) {
      sum = sum + P[i,j];
    }
    lin_k[i,k] = sum;
  }
}

# Calculando percentual somado em cada sub-tabela -P(..k) #
SP<-array(0:0, dim=c(1,K))
for (k in 1:K) {
  sum=0;
  for (i in 1:nrow(N)) {
    for (j in in_tab[k]:fin_tab[k]) {
      sum = sum + P[i,j];
    }
  }
  SP[1,k] = sum;
}

# Calculando Matriz Pijk/P.jk #
X <-array(0:0, dim=c(nrow(N),ncol(N)))
for (i in 1:nrow(N)) {
  for (j in 1:ncol(N)) {
    X[i,j] = P[i,j]/col[1,j];
  }
}

# Calculando Matriz Pi.k/P..k (P..k = SP1 e SP2) #
Y <-array(0:0, dim=c(nrow(N),ncol(N)))
for (k in 1:K) {
  for (i in 1:nrow(N)) {
    for (j in in_tab[k]:fin_tab[k]) {
      Y[i,j] = lin_k[i,k]/SP[k];
    }
  }
}

# Matriz WP é [P(ijk)/P(.jk) - (P(i.k)/P(..k)]/P(i..) #
WP <-array(0:0, dim=c(nrow(N),ncol(N)))

```



```

for (i in 1:nrow(N)) {
  for (j in 1:ncol(N)) {
    WP[i,j] =(X[i,j] - Y[i,j])/lin[i,1]
  }
}

# PASSO 2: SEPARAR A MATRIZ WP EM K MATRIZES PARA EFETUAR K ACI #
# APLICAR DVS PARA CADA UMA DELAS SEPARADAMENTE E GUARDAR #
# OS AUTOVALORES EM K VETORES (dk) #

# Separando as tabelas #
Z1 <-array(0:0, dim=c(nrow(N),J[1]))
for (i in 1:nrow(N)) {
  for (j in in_tab[1]:fin_tab[1]) {
    Z1[i,j] = WP[i,j];
  }
}
Z2 <-array(0:0, dim=c(nrow(N),J[2]))
for (i in 1:nrow(N)) {
  for (j in in_tab[2]:fin_tab[2]) {
    Z2[i,j-fin_tab[1]] = WP[i,j];
  }
}

# seguir criando outras matrizes, se K> 2#

#Matrizes M para cada DVS - Matriz diag (P(.jk)/P(..k))#
M1 <-array(0:0, dim=c(J[1], J[1]))
for (j in 1: J[1]) {
  M1[j,j]=col[1,j]/SP[1]
}

M2<-array(0:0, dim=c(J[2], J[2]))
for (j in in_tab[2]:fin_tab[2]) {
  M2[j-fin_tab[1],j-fin_tab[1]]=col[1,j]/SP[2]
}

# seguir criando outras matrizes, se K> 2#

#Matriz D para CP – Matriz Diag(P(i..))#
D <- as.vector(lin)
D <- diag(D)

```



```

# Calculando Autovalores das ACI #
MSVD1 <- D^0.5 %*% Z1 %*% M1^0.5
d1<-svd(MSVD1)$d

MSVD2 <- D^0.5 %*% Z2 %*% M2^0.5
d2<-svd(MSVD2)$d
# seguir com as DVS se K>2 #

# Usando os autovalores das ACI para ponderar  $p*jk/d1^2$  #
colf <- array(0:0, dim=c(1,ncol(N)))
for (k in 1:K) {
  for (j in in_tab[k]:fin_tab[k]) {
    colf[1,j]=col[1,j]/SP[k];
  }
}

lamb_a <- array(0:0, dim=c(ncol(N),ncol(N)))
for (j in 1:J[1]) {
  lamb_a[j,j]= 1/d1[1]^2
}
for (j in in_tab[2]:fin_tab[2]) {
  lamb_a[j,j]= 1/d2[1]^2
}
# seguir cálculos caso K>2 #

# Matriz M para AFMTC #
Mf<-as.vector(colf)
Mf<-diag(Mf)

# Decomposição do Valor Singular Final #
MSVDf <- D^0.5 %*% WP %*% Mf^0.5 %*% lamb_a^0.5
df<-svd(MSVDf)$d
Uf<-svd(MSVDf)$u
Vf<-svd(MSVDf)$v

Df <-diag(df)

Dlinf <- array(0:0, dim=c(nrow(N),nrow(N)))
for (i in 1:nrow(N)) {
  Dlinf[i,i]= 1/sqrt(lin[i,1])
}

Lamb_f <- array(0:0, dim=c(ncol(N),ncol(N)))
for (j in 1:J[1]) {
  lamb_f[j,j]= d1[1]
}

```



```

for (j in in_tab[2]:fin_tab[2]) {
  Correc12 [j,j]=d2[1]
}
# continuar caso K>2 #

# Calcular Coordenadas padrão
X<-Dlinf %*% Uf
F<-X %*% Df

#Calcular coordenadas principais
Y<-lamb_f %*% Vf
G <- Y %*% Df

# Calcular as coordenadas parciais para linhas de tabelas diferentes #
Mb<- array(0:0, dim=c(ncol(N),ncol(N)))
for (j in 1:J[1]) {
  Mb[j,j]=sqrt(colf[1,j])
}
for (j in in_tab[2]:fin_tab[2]) {
  Mb[j,j]=sqrt(colf[1,j])
}

Da <-array(0:0, dim=c(nrow(N),nrow(N)))
for (i in 1:nrow(N)) {
  Da[i,i] = 1/df[i]
}

# CALCULANDO AS COORDENADAS PRINCIPAIS PARCIAIS

Fa <- t(Mb %*% lamb_a %*% t(WP)) %*% t(Da %*% t(G))

# Para Tabela 1#
Zero<-array(0:0, dim=c(nrow(N),J[1]))
Z1a<-cbind(Z1,Zero)
Fa1 <- t(Mb %*% lamb_a %*% t(Z1a)) %*% t(Da %*% t(G))

# Para Tabela 2#
Z2a<-cbind(Zero,Z2)
Fa2 <- t(Mb %*% lamb_a %*% t(Z2a)) %*% t(Da %*% t(G))

Fa1<-2*Fa1
Fa2<-2*Fa2

plot(c(-2,3),c(-2,3), pch= 2,xlab="Eixo 1",ylab="Eixo 2",type='n', cex = 1.5)
points(F[,1],F[,2],lty=2,col="red",pch=c("1","2","3","4"))
lines(F[,1],F[,2],lty=4,col="red")

```



```

points(G[,1],G[,2],lty=2,col="blue",pch=c("a","b","c","d","e","A","B","C","D","E"))
lines(G[,1],G[,2],lty=2,col="blue")
abline(0,0,0,0, lty=2, col="black")

```

```

# Inércia explicada por categoria eixo-linha
min<- min(nrow(N),ncol(N))
somalin<-array(0:0, dim=c(1,min))
for (j in 1:min) {
  sum=0;
  for (i in 1:nrow(N)) {
    sum = sum + lin[i,1]*F[i,j]^2;
  }
  somalin[1,j] = sum
}
lexlin <-array (0:0,dim=c(4,4))
for (i in 1:nrow(N)) {
  for (j in 1:min) {
    lexlin[i,j] = 100*( lin[i,1]*F[i,j]^2/somalin[1,j])
  }
}

```

```

# Inércia explicada por categoria eixo-coluna#
somacol<-array(0:0, dim=c(1,ncol(N)))
for (j in 1:min) {
  sum=0;
  for (i in 1:ncol(N)) {
    sum = sum + (1/Correc12[i,i]^2)*G[i,j]^2 ;
  }
  somacol[1,j] = sum
}

```

```

lexcol <-array (0:0,dim=c(ncol(N),4))
for (i in 1:ncol(N)) {
  for (j in 1:4) {
    lexcol[i,j] = 100*((1/Correc12[i,i]^2)*G[i,j]^2/somacol[1,j])
  }
}

```



```

# Inércia explicada por categoria eixo-linha – Para tabelas Parciais de linhas #
somalin1<-array(0:0, dim=c(1,min))
for (j in 1:min) {
  sum=0;
  for (i in 1:nrow) {
    sum = sum + lin[i,1]*(F[i,j]-Fa1[i,j])^2;
  }
  somalin1[1,j] = sum
}

Iexlin1 <-array (0:0,dim=c(nrow(N),min))
for (i in 1:nrow(N)) {
  for (j in 1:min) {
    Iexlin1[i,j] = 50*( lin[i,1]*(F[i,j]-Fa1[i,j])^2/somalin1[1,j])
  }
}

# Parcial 2#
somalin2<-array(0:0, dim=c(1,4))
for (j in 1:4) {
  sum=0;
  for (i in 1:4) {
    sum = sum + lin[i,1]*(F[i,j]-Fa2[i,j])^2;
  }
  somalin2[1,j] = sum
}

Iexlin2 <-array (0:0,dim=c(4,4))
for (i in 1:4) {
  for (j in 1:4) {
    Iexlin2[i,j] = 50*( lin[i,1]*(F[i,j]-Fa2[i,j])^2/somalin2[1,j])
  }
}

# Calculando o cos^2 para linhas #
Tcos_f<-array(0:0, dim=c(nrow(N),1))
for (i in 1:nrow(N)) {
  sum=0;
  for (j in 1:min) {
    sum = sum + lin[i,1]*F[i,j]^2;
  }
  Tcos_f[i,1] = sum;
}
cos_f<-array(0:0, dim=c(nrow(N),min))
for (i in 1:nrow(N)) {
  for (j in 1:min) {
    cos_f[i,j] = lin[i,1]*F[i,j]^2/Tcos_f[i,1];
  }
}

```



```

}
}

# Cosseno quadrado de coordenadas colunas OK#
Tcos_g<-array(0:0, dim=c(ncol(N),1))
for (i in 1:ncol(N)) {
  sum=0;
  for (j in 1:min) {
    sum = sum + col12b[1,i]*G[i,j]^2;
  }
  Tcos_g[i,1] = sum;
}
cos_g<-array(0:0, dim=c(ncol(N),min))
for (i in 1:ncol(N)) {
  for (j in 1:min) {
    cos_g[i,j] = col12b[1,i]*G[i,j]^2/Tcos_g[i,1];
  }
}

# Raiz quadrada dos cossenos quadrados de colunas #
Unit_G<-sign(G)*sqrt(cos_g)

# Calculando as inércias por eixo -bloco #
d<-cbind(d1^2,d2^2)
in_bloc <- array(0:0, dim=c(nrow(N),K))
for (k in 1:K) {
  for (j in 1:nrow(N)) {
    sum=0;
    for (i in in_tab[k]:fin_tab[k]) {
      sum=sum + G[i,j]^2;
    }
    in_bloc[j,k]=sum/d[1,k];
  }
}

```


6. Referências Bibliográficas

Bécue-Bertaut M, Pagès J. A principal axes method for comparing contingency tables: MFACT. *Computacional Statistics & Data Analysis*. 2004; 45: 481-503.

Bussab WO, Miazaki ES, Andrade DF. *Introdução à análise de agrupamentos*. São Paulo: 9º Simpósio Nacional de Probabilidade e Estatística, 1990.

Bussab W, Morettin P. *Estatística básica*. 5.ed. São Paulo: Saraiva, 2002.

Benzécri J-P, Bressolier C, Thomas Y. Deux analysis de données granulométriques en géomorphologie. *Cahier de l'Analyse des Données* 1, 145-160. 1976.

Benzécri, J-P. Analyse de l'inertie intraclasse par l'anlyse d'un tableau de contingente, *Les Cahiers de l'Analyse des Donnes*, 1983; 8(3), 351-358.

Crivisqui EM. *Analisis factorial de correspondencias: un instrumento de investigación en ciencias sociales*. Edición del Laboratorio de informática social: Universidad Católica de Asunción. 1993

Escofier B, Drout D. Analyse des differences entre plusieurs tableaux de fréquence (analysis of the differences among several frequency tables). *Les Cahiers de l'Analyse des Données* 8 (4), 491-499. 1983.

Escofier B. Analyse factorielle em referência à um modèle: application à l'analyse d'um tableau d'échanges (principal axés method in reference to a model: application to the analysis of na exchanges table). *Rev. Statist. Appl.* 32 (4), 25-36. 1984.

Escofier B, Pagès J. *Analisis factoriales simples y múltiples*. Bilbao: Dunod. 1992.

Escofier B, Pagès J. *Multiple and simple principal axes methods; objectives, methods and interpretation*. Paris: Dunod. 1998.

Gower J. *Procrustes analysis*. *Handbook of applicable mathematics*. Wiley, Chichester, pp 397-405. Lloyd, E.H (Ed.) 1984, vol 6.

Greenacre JM. *Theory and application of correspondence analysis*. London: Academic Press. 1984.

Greenacre MJ. Correspondence analysis of multivariate categorical data by weighted least-square. *Biometrika* 1988; 75 (3): 457-67.

Greenacre MJ. Some limitations of multiple correspondence analysis. *Computational statistics quarterly* 3: 249-256. 1990. Heidelberg: Physica Verlag.

Greenacre MJ. Multivariate generalizations of correspondence analysis. *Multivariate analysis: future directions* 2, 1993. 327-40.

Greenacre MJ. *Correspondence analysis in practice*. London: Academic Press, 1994.

Lebart L, Morineau A, Warwick KM. *Multivariate descriptive statistical analysis: correspondence analysis related techniques for large matrices*. New York: Wiley, 1984.

Lebart L, Morineau A, Tabard N. *Techniques de la description statistique*. Paris: Dunod. 1987.

Mardia KV, Kent JT e Bibby JM. *Multivariate analysis*. New York: Academic Press, 1979.

Pagès J. *Multiple factor analysis*. IV International Meeting of Multivariate Data Analysis. Paris: New Spad, 1997.

Pagès J, Bécue-Bertaut M. *Multiple factor analysis for contingency tables*. 2004-2005 (in press).

Pereira JCR. *Análise de dados qualitativos: estratégias metodológicas para ciências da saúde, Humanas e Sociais*. São Paulo: Edusp, 1999.

Van der Heijden PGM. *Correspondence analysis of longitudinal categorical data*. Leiden: DSWO Press, 1987. (M&T series, vol 8).

Wolff KE, Gabler S. Comparison of visualization in formal concept analysis and correspondence analysis. *Visualization of categorical data*. 1998; 85-105.

Disponível em URL <http://www.r-project.org/>

Disponível em URL <http://lattes.cnpq.br/diretorio/>