Universidade Federal do Rio Grande do Sul
Instituto de Matemática
Cadernos de Matemática e Estatística
Série F: Trabalho de Divulgação

ANÁLISE DE CORRESPONDÊNCIA
APLICAÇÕES EM GENÉTICA

Sidia Maria Callegari-Jacques

Série F, no 2, DEZ/91
Porto Alegre, dezembro de 1991

ANÁLISE (FATORIAL) DE CORRESPONDÊNCIA: APLICAÇÕES EM GENÉTICA

Sidia Maria Callegari-Jacques

Departamento de Estatística
Instituto de Matemática
Universidade Federal do Rio Grande do Sul
Porto Alegre - RS - Brasil

1. INTRODUÇÃO

O método da análise de correspondência foi publicado pela primeira vez por *Hirschfeld* (1935), que propôs uma solução algébrica para a "correlação" entre linhas e colunas de uma tabela de contingência. Idéias similares, porém sob uma forma não matemática, foram sugeridas na mesma época por *Horst* (1935), que criou o termo "method of reciprocal averaging" para uso em psicometria. Posteriormente *Fisher* (1940) retomou o assunto, propondo uma análise discriminante para tabelas de contingência. Uma maior divulgação do método, no entanto, só ocorreu na década de 1970, devido em grande parte ao trabalho de um grupo de franceses liderados por *Jean-Paul Benzécri*. Este grupo cunhou a denominação de "análise fatorial de correspondências" (AFC).

Este texto vai apresentar um resumo da técnica, seus objetivos, a interpretação dos resultados e exemplos de aplicações em Genética. Descrições mais detalhadas da teoria, em vários níveis de complexidade, e outras aplicações podem ser encontradas em *Lebart & Fenelon* (1971), *Benzécri* (1973), *Hill* (1974), *Lebart et al* (1977), *Greenacre & Degos* (1977), *Nishisato* (1980), *Gifi* (1981), *Greenacre* (1984), *Lebart et al* (1984). Em português, *Verdinelli* (1980), *Souza* (1982) e *Souza* (1990).

A AFC é uma técnica exploratória, mais destinada a gerar hipóteses do que a testá-las. Foi delineada para dados categóricos, que costumam ser

organizados em tabelas de contingência. O método visa analisar a associação entre duas ou mais variáveis categóricas e permite visualizar mais facilmente a relação entre cada linha e cada coluna, o que é especialmente útil no caso das tabelas grandes.

Suponhamos uma tabela bidimensional (Tabela 1) na qual estão organizadas duas variáveis, por exemplo, população e tipo morfológico, cada uma dividida em categorias. As **categorias-linhas** representam as várias populações e as **categorias-colunas**, os tipos morfológicos. Cada casela contém o número de observações que correspondem a cada tipo morfológico, em cada população.

Pode-se representar as populações em um gráfico, de tantas dimensões quantas forem as categorias da variável "tipo morfológico" (Fig 1 e 2). Nestas representações, as freqüências utilizadas são as proporções em relação ao total da linha. A Fig. 1 representa as populações em um espaço bidimensional utilizando apenas os tipos x e y. Na Fig. 2 tem-se o espaço tridimensional, incluindo os três tipos morfológicos. Nota-se que as populações C e D, que aparecem muito próximas na Fig. 1, estão na realidade mais afastadas, o que pode ser percebido a partir dos dados da Tabela 1.

Este tipo de representação permite visualizar a proximidade entre as populações e também interpretar possíveis associações entre as populações estudadas e os vários tipos morfológicos. Assim, A e B associam-se ao tipo x, C e D a y. No entanto, se o número de categorias-colunas excede três, qualquer representação em um único gráfico é inviável. Uma das vantagens da análise de correspondência é justamente propiciar, para tabelas complexas, uma imagem gráfica a duas dimensões (se a estrutura dos dados permitir), conservando-se uma quantidade razoável da informação e perdendo-se um mínimo, sacrificado pela facilidade de interpretação. O objetivo e a técnica da AFC são muito semelhantes ao da análise de componentes principais (ACP). Deseja-se reduzir o espaço de representação das linhas (populações) para um subespaço de dimensões menores, sem perda substancial da informação (variabilidade). A diferença é que na AFC os dados não são quantitativos, mas qualitativos (categóricos).

TABELA 1. Frequências observadas (número de indivíduos) quanto a três tipos morfológicos (x, y e z) em quatro populações hipotéticas, de 100 indivíduos cada uma.

	TIPO MORFOLÓGICO			
POPULAÇÃO	x	y	z	Total
A	70	20	10	100
B	80	10	10	100
C	10	70	20	100
D	20	80	0	100
Total	180	180	40	400

TABELA 2. Dados da Tabela 1 transformados em proporções (p_{ij}) em relação ao total.

	TIPO MORFOLÓGICO			
POPULAÇÃO	x	y	z	Total ($p_{i.}$)
A	0,175	0,050	0,025	0,250
B	0,200	0,025	0,025	0,250
C	0,025	0,175	0,050	0,250
D	0,050	0,200	0	0,250
Total ($p_{.j}$)	0,450	0,450	0,100	1

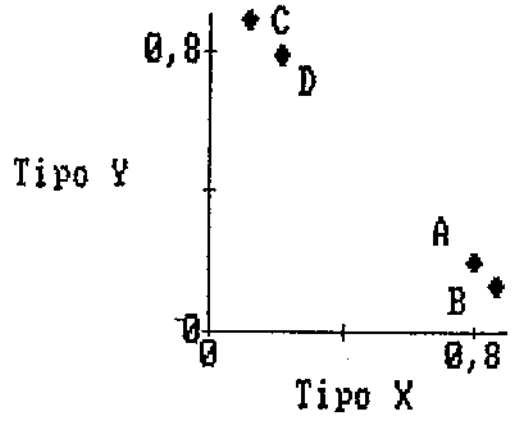


Fig.1. Representação gráfica dos pontos-linhas (populações) da Tabela 1, em um espaço bidimensional, usando as colunas x e y.

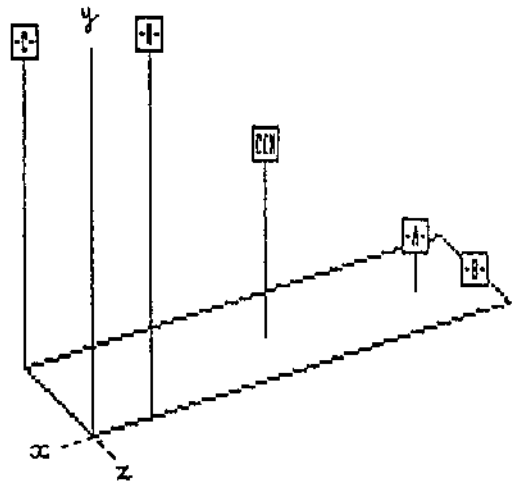


Fig.2. Representação gráfica dos pontos-linhas (populações) da Tabela 1, em um espaço tridimensional, usando as colunas x, y e z.

2. RESUMO DA METODOLOGIA DE CÁLCULO

Voltando à Tabela 1, divide-se inicialmente o número observado em cada cela ou casela (k_{ij}) pelo total da tabela (N), de forma a obter um total igual a 1. Com este procedimento, se obtêm as proproções observadas (p_{ij}) de cada cela em relação ao total estudado (Tabela 2). Assim, a proporção observada de indivíduos da população i a apresentar o tipo j é

$$p_{ij} = \frac{k_{ij}}{N}.$$

Dividindo os totais marginais também por N, têm-se as proporções marginais. Ao dividir as proporções observadas (p_{ij}) pela respectiva proporção marginal, que é o total destas proporções na linha ($p_{i.}$), tem-se o que se denomina o **perfil da linha** i (Tabela $3A$).

Os perfis-linhas podem ser usados como coordenadas para desenhar o diagrama de pontos que representa as populações em um espaço de m dimensões, sendo m o número de colunas. É natural que cada ponto tenha um peso proporcional ao número de indivíduos que representa. Isto é dado pela **massa**, que é a proporção marginal da linha i ($p_{i.}$; ver Tabela 2). A nuvem de pontos relativa às populações têm um **centro de gravidade** , ou **centróide**, que corresponde a uma população hipotética cujas coordenadas são as freqüências médias de cada tipo morfológico (isto é, as proporções marginais das colunas, $p_{.j}$). Na Fig. 2 está representada a nuvem das populações e o centróide (CEN). Suas coordenadas são: $x = 0,45$; $y = 0,45$ e $z = 0,10$. Neste exemplo, o centróide fica a igual distância de todos os pontos, porque todas as populações têm igual massa ($p_{1.} = p_{2.} = p_{3.} = p_{4.} = 0,25$).

A variabilidade dos pontos em relação ao centro de gravidade é dada pela **inércia**, que é a soma das distâncias, ao quadrado, de cada ponto em relação a este centro, ponderadas pela massa do ponto.

Uma representação gráfica pode ser feita também para as colunas (tipos morfológicos). A Fig. 3 representa a nuvem de pontos relativos aos três tipos morfológicos cujos perfis-colunas estão na Tabela $3B$. Estes tipos estão posicionados no espaço determinado pelas populações A, B e C somente, já que não se pode representar uma nuvem em um espaço tetradimensional. Nesta nuvem, o centro de gravidade representa o tipo morfológico "médio".

TABELAS 3A E 3B. Perfis-linhas ($p_{ij} / p_{i.}$) e perfis-colunas ($p_{ij} / p_{.j}$) dos dados das Tabelas 1 e 2.

A: PERFIS-LINHAS

	x	y	z
A	0,7	0,2	0,1
B	0,8	0,1	0,1
C	0,1	0,7	0,2
D	0,2	0,8	0

B: PERFIS-COLUNAS

	x	y	z
A	0,39	0,11	0,25
B	0,44	0,06	0,25
C	0,06	0,39	0,50
D	0,11	0,44	0

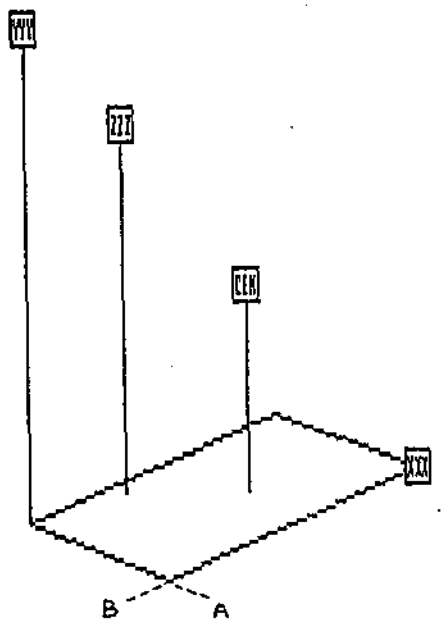


Fig.3. Representação gráfica dos pontos-colunas (tipos morfológicos) da Tabela 1, em um espaço tridimensional, determinado pelas populações A, B e C.

Como já foi mencionado, o procedimento matemático usado na AFC para reduzir o espaço de representação da informação é muito semelhante ao da ACP. Aqui, também se necessita uma matriz quadrada que informe sobre a similaridade entre linhas e entre colunas, respectivamente. A semelhança entre linhas (populações) será representada por uma **medida de distância**, dada pela diferença entre as proporções (condicionais) observadas em cada linha, isto é, a diferença entre os perfis das linhas duas a duas. Nota-se, no entanto, que se um dos tipos morfológicos (colunas) for, em geral, muito mais freqüente que os outros (massa maior), as diferenças encontradas quanto a este tipo terão importância desproporcionalmente grande na avaliação geral da diferença entre populações. A solução é introduzir, no cálculo da distância, uma ponderação pelo inverso da massa da coluna. Deste modo, o quadrado da distância será

$$d^2_{ii'} = \sum_j \frac{1}{p_{.j}} \left\{ \frac{p_{ij}}{p_{i.}} - \frac{p_{i'j}}{p_{i'.}} \right\}^2 .$$

Aqui, i e i' indicam duas linhas diferentes. O termo entre chaves é a diferença entre os perfis das linhas i e i'.

Esta distância é chamada "**distância tipo χ^2**" e foi escolhida porque apresenta a propriedade da "equivalência distribucional". Segundo esta propriedade, se duas colunas têm perfis idênticos, elas podem ser agregadas sem alterar as distâncias entre as linhas. O mesmo vale para reunião de linhas de perfis semelhantes: não se alterarão muito as distâncias entre as colunas. Esta propriedade é importante porque garante uma relativa estabilidade nos resultados face ao critério escolhido na construção das classes (*Lebart & Fenelon*, 1971: 231; *Lebart et al*, 1977).

A distância tipo χ^2, porém, não é uma soma de quadrados e então a abordagem que se vai utilizar, semelhante à da ACP, não é possível. Para poder realizar o mesmo tipo de análise, deve-se alterar a escala dos eixos de modo que as coordenadas do ponto i sofram a seguinte modificação:

$$\frac{p_{ij}}{p_{i.}} \longrightarrow \frac{p_{ij}}{p_{i.}\sqrt{p_{.j}}}$$

Pode-se demonstrar que esta mudança de escala não altera a distância entre dois pontos (*Lebart & Fenelon*, 1971: 233). Mas agora, as distâncias podem ser consideradas em relação ao centro de gravidade da nuvem (população

"média"), levando em consideração a massa de cada ponto. O conjunto destas distâncias é a inércia. A inércia total da nuvem pode ser calculada do seguinte modo:

$$I = \sum_i \sum_j \frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}} .$$

Ou seja, a informação tipo qui-quadrado dada pelas distâncias está sendo apresentada como inércia, que nada mais é do que χ^2 de associação da tabela de dados originais (k_{ij}) dividido pelo total geral N (*Moser*, 1989). Na verdade, a inércia é a distância χ^2 entre a lei das frequências observadas (p_{ij}) e a lei de independência ($p_{i.}p_{.j}$) (*Lebart & Fenelon*, 1971: 234).

Para os dados da Tabela 2, I é 0,4612. Multiplicando este valor pelo total estudado (400), tem-se 184,4, que é o valor do χ^2 de associação para os dados da Tabela 1.

A inércia, portanto, pode ser interpretada como:

1. Uma medida da variabilidade dos dados, isto é, da dispersão dos pontos no espaço, levando em consideração a massa dos pontos;

2. Uma medida do desvio das proporções observadas em relação às que seriam esperadas se houvesse independência entre linhas e colunas.

A AFC pode ser entendida como um método para a identificação de um subespaço ao longo do qual a inércia é máxima. A identificação deste subespaço é realizada como na ACP, buscando-se autovalores e os respectivos autovetores da matriz que representa a variabilidade entre as várias linhas, por um lado, e as várias colunas, por outro. Devido a simplificações possíveis especificamente na análise de correspondência (*Lebart & Fenelon*, 1971: 236), os cálculos se resumem a procurar os vetores próprios de uma matriz simétrica S, cujo termo geral é

$$s_{jk} = \sum_i \frac{p_{ij}\ p_{ik}}{p_{i.}\ \sqrt{p_{.j}\ p_{.k}}}$$

A inércia, portanto, pode ser decomposta em autovalores (lambda), que são os "momentos da inércia". O primeiro é trivial e não é utilizado na interpretação; seu valor é sempre um e ele garante que linhas e colunas possam ser representadas em um mesmo gráfico, colocando o centro das coordenadas

(origem) no centro de gravidade (*Hill*, 1974; *Murtagh & Heck*, 1987: 159; *Greenacre*, 1984: 92). Os demais decrescem de valor, como na ACP. Os autovetores correspondentes aos autovalores não triviais determinam os eixos principais da inércia. O primeiro eixo representa a maior fração da inércia, o segundo uma fração adicional, o terceiro outra fração adicional, porém menor, e assim sucessivamente. Os eixos são ortogonais, isto é, refletem dimensões independentes entre si.

É importante observar que na ACP, os componentes (eixos) refletem frações independentes da variância, enquanto na AFC os eixos representam frações independentes da inércia, onde a massa (importância relativa) de cada ponto é levada em consideração.

O mesmo tipo de abordagem apresentada para as linhas (gráfico, nuvem de pontos, centro de gravidade, inércia) pode ser utilizado para as colunas. Pode-se mostrar ainda que a decomposição da inércia das colunas é igual à das linhas (*Lebart & Fenelon*, 1971:238; *Greenacre & Degos*, 1977; *Lebart et al*, 1977:56), ou seja, os autovalores resultantes são iguais.

As novas coordenadas dos pontos-linhas podem ser organizadas em uma matriz A onde cada elemento $a_{i\alpha}$ é a coordenada do ponto i com respeito ao eixo principal α. Da mesma forma, as novas coordenadas dos pontos j são $b_{j\alpha}$, organizadas em uma matriz B (Tabelas $4A$ e $4B$). Os elementos das matrizes A e B são linearmente relacionados entre si, através dos vários autovalores (lambdas), do seguinte modo (*Greenacre & Degos*, 1977):

$$a_{i\alpha} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_j \left(\frac{p_{ij}}{p_{i.}}\right) b_{j\alpha} \; ; \quad b_{j\alpha} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_i \left(\frac{p_{ij}}{p_{.j}}\right) a_{i\alpha}.$$

Como resultados práticos da AFC, temos que:

1. A decomposição da inércia em autovalores fornece eixos ortogonais entre si, permitindo representar as categorias-linhas em um espaço dimensionalmente menor, sem perda substancial de informação.

2. O mesmo ocorre com relação às categorias-colunas.

3. Por causa de uma relação simétrica existente entre as novas coordenadas de linhas e colunas, pode-se plotá-las em um mesmo gráfico (Fig. 4), centrado no centro de gravidade, que é o mesmo para linhas e colunas (*Greenacre & Degos*, 1977).

TABELAS 4A E 4B. Novas coordenadas ($a_{i\alpha}$ e $b_{j\alpha}$) para os pontos-linhas (populações) e pontos-colunas (tipos morfológicos) obtidas após a análise de correspondências.

A: PONTOS-LINHAS

POP (i)	EIXO 1 (α=1)	EIXO 2 (α=2)
A	a_{11}	a_{12}
B	a_{21}	a_{22}
C	a_{31}	a_{32}
D	a_{41}	a_{42}

B: PONTOS-COLUNAS

TIPO (j)	EIXO 1 (α=1)	EIXO 2 (α=2)
x	b_{11}	b_{12}
y	b_{21}	b_{22}
z	b_{31}	b_{32}

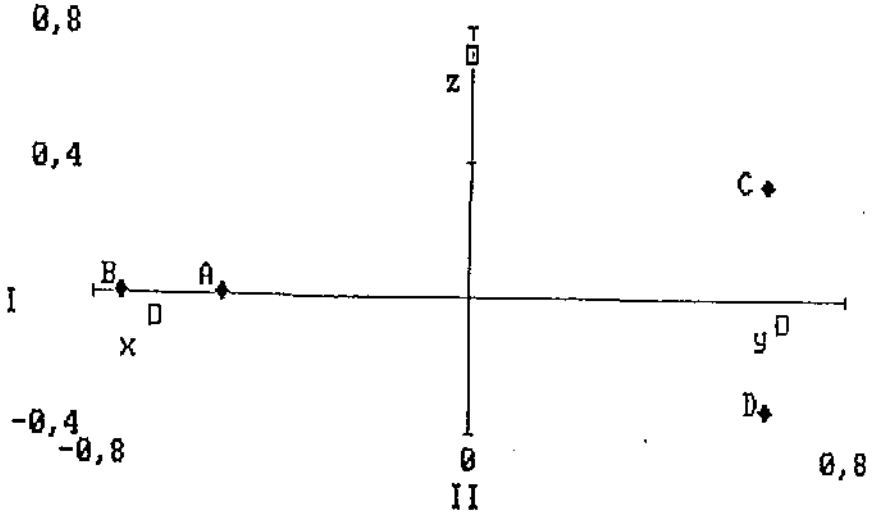


Fig.4. Representação gráfica simultânea de populações (linhas) e tipos morfológicos (colunas) da Tabela 1, através da técnica da análise de correspondência.

Através da Fig. 4, fica fácil verificar a proximidade das populações A e B, por um lado, e C e D por outro. Além disso, nota-se que o tipo x é mais freqüente nas duas primeiras populações, o y é mais freqüente em C e D e o tipo z não está associado preferentemente a nenhuma população .

Deste modo, a AFC permite a representação e a interpretação visual da posição relativa das nuvens de pontos-linhas e pontos-colunas em um único subespaço, que reflete as direções principais da dispersão nestas nuvens.

É imp ortante lembrar que a AFC não realiza testes de significância, de modo que antes de ser um substituto do teste do qui-quadrado para tabelas de contigência é, na verdade, um poderoso auxiliar na interpretação do seu resultado.

Por outro lado, a AFC permite analisar simultaneamente tabelas justapostas de dados, como no caso, por exemplo, de um questionário com 5 perguntas, cada uma admitindo várias respostas alternativas, apresentado à grupos de diferentes profissionais. O método da análise de correspondência múltipla é uma extensão da AFC para tabelas de contingência a duas dimensões e se caracteriza por apresentar regras simples de interpretação dos gráficos resultantes. *Lebart et al* (1984: cap.4) apresentam os detalhes da técnica. O exemplo de aplicação 5.2, a ser apresentado mais adiante, ilustra a aplicação da AFC múltipla a um estudo em antropometria.

3. INTERPRETAÇÃO DOS RESULTADOS DA AFC

3.1. QUANTO AOS GRÁFICOS:

(1) A origem do gráfico representa o centro de gravidade de linhas e colunas, isto é, ele representa a população "média" e o tipo "médio" para os dados da Tabela 1. É também o ponto onde o desvio em relação à hipótese de independência entre linhas e colunas é zero.

(2) Quanto mais afastado do centro está um ponto, mais ele se desvia da "média". É, também, maior a diferença entre as frequências observadas e as esperadas supondo independência, na casela que está representando.

(3) Se dois pontos-linhas (populações) estão próximos, são semelhantes quanto à variável representada nas colunas. O mesmo comentário vale para os pontos-colunas.

(4) Um ponto-população que está próximo à origem é muito semelhante à população "média".

(5) É legítimo interpretar a posição relativa de um ponto de uma nuvem em relação a **todos** os pontos da outra nuvem, isto é, interpretar a posição relativa de uma população em relação a todos os tipos morfológicos considerados juntos. Exceto em casos especiais, é perigoso interpretar diretamente a proximidade de dois pontos de nuvens diferentes (*Lebart et al*, 1984: 46).

3.2. QUANTO AOS EIXOS PRINCIPAIS:

(1) A **fração da inércia** representada por cada eixo é dada pelo autovalor deste eixo dividido pela soma dos autovalores.

(2) A **contribuição absoluta** de uma linha ao eixo 1 indica a porção da inércia explicada pelo eixo 1 que é atribuída a esta linha. Esta contribuição é calculada com base na semelhança entre as categorias (*Lebart et al.*, 1977: 60) e é apresentada como uma porcentagem, porque a soma das contribuições absolutas das linhas é 1. As contribuições absolutas das colunas também são avaliadas.

(3) A contribuição relativa mostra a parte da dispersão de uma categoria explicada por um eixo. Não é muito usada na interpretação da AFC.

A Tabela 5 apresenta os resultados da AFC aplicada ao exemplo das quatro populações e três tipos morfológicos. Ali estão a fração da inércia total representada por cada eixo e as contribuições absolutas de cada categoria. Nota-se que o primeiro eixo representa 88% da variabilidade encontrada. Não há muita diferença entre as populações quanto à contribuição (17-34) na formação do eixo 1, mas x (50) e y (50) são os que têm maior peso dentre os tipos morfológicos. Já z tem maior contribuição na formação do eixo 2. Sinais diferentes ao lado das contribuições indicam categorias que se opõem, quando se considera um eixo. A representação gráfica destes resultados já foi apresentada na Fig. 4.

4. COMPARAÇÃO ENTRE A ANÁLISE DE COMPONENTES PRINCIPAIS (ACP) E A ANÁLISE DE CORRESPONDÊNCIA (AFC)

(1) A ACP é utilizada em dados quantitativos; a AFC foi delineada para dados categóricos, embora seja utilizada também para dados quantitativos categorizados por faixas de valores.

'ABELA 5. Análise de correspondências nos dados da Tabela 1: roporção da inércia atribuída a cada eixo (entre parênteses) e ontribuições absolutas de cada categoria a cada eixo.

	EIXO 1 (88%)	EIXO 2 (12%)
POPULAÇÃO A	17 +	0
POPULAÇÃO B	34 +	0
POPULAÇÃO C	25 −	49 −
POPULAÇÃO D	25 −	51 +
TIPO x	50 +	5 +
TIPO y	50 −	5 +
TIPO z	0	90 −

(2) A ACP opera sobre a matriz de variância-covariância (ou de correlações); a AFC trabalha em cima de distâncias tipo qui-quadrado (métrica χ^2) entre perfis-linhas ou perfis-colunas (pontos).

(3) Na ACP, a dispersão dos pontos representa distâncias; a AFC leva em conta distâncias e massas. Na ACP, a inércia se reduz à variância e a massa é 1 para todas as linhas e colunas.

(4) Como na ACP, também na AFC podem ser obtidos os escores relativos à cada linha ou coluna. Olhando, por exemplo, do ponto de vista das linhas (populações), os escores do eixo 1 são os valores de cada população com respeito a uma nova variável, que é uma espécie de índice que reúne a informação quanto aos três tipos morfológicos em um único valor. Estes escores são muitas vezes utilizados para estudar, por exemplo, a distribuição geográfica dos tipos morfológicos (ver exemplo abaixo, quanto ao sistema Gm).

5. EXEMPLOS DE APLICAÇÃO DA ANÁLISE DE CORRESPONDÊNCIA EM GENÉTICA

5.1. EFEITO DE ESTERÓIDES NO DESENVOLVIMENTO ONTOGENÉTICO DE *DROSOPHILA MELANOGASTER*

O desenvolvimento desde o ovo fertilizado até o adulto, em *D. melanogaster*, é marcado, em diferentes estágios, pela ação do esteróide ecdisona. Um conhecido esteróide produzido por vertebrados é a vitamina D3. Existem plantas que apresentam o princípio ativo deste esteróide, porém na forma hidrossolúvel: *Solanum malacoxylon* é uma delas. Esta solanácea é comum no Mato Grosso (nome popular = "espichadeira") e no Rio Grande do Sul e sua ingestão por ovinos e bovinos causa calcinose zoótica, com grandes prejuízos à pecuária.

Visando conhecer os efeitos da ecdisona, da vitamina D3 e do esteróide encontrado no extrato aquoso de *S.malacoxylon* sobre o desenvolviento pupa-adulto de *D. melanogaster*, *Jung et al.* (1991) estudaram 2000 ovos de cada uma de três populações com velocidade de desenvolvimento diferentes: precoce, velocidade normal e desenvolvimento tardio. De cada população obtiveram-se quatro amostras que foram tratadas, respectivamente, com os três esteróides mencionados; uma, não tratada, foi usada como controle. Os

dados relativos à população de desenvolvimento tardio estão apresentados na Tabela 6. Foram encontrados: adultos viáveis (a) pupas com desenvolvimento interrompido (di) e histólise rápida da pupa, isto é, não diferenciação dos disco do imago (hr).

A Fig. 5 explica a associação identificada pelo χ^2 calculado nestes dados ($194,2; gl = 6; p < 0,001$). Nota-se que a histólise rápida é proporcionalmente mais freqüente quando o tratamento é com ecdisona (TE) e o desenvolvimento interrompido é mais comum com vitamina D3 (TV). A amostra tratada com esteróide de *Solanum* e a amostra controle (TS e TC) apresentam freqüência semelhante de adultos viáveis.

Doze condições experimentais foram avaliadas quanto ao efeito sobre a forma final observada nesta mosca. Elas resultaram da combinação de 3 populações (P: desenvolvimento precoce, C: normal, T: tardio) com 4 tratamentos (E: ecdisona, V: vitamina D3, S: extrato de *Solanum*, C: controle). O χ^2 obtido foi de $470,2$ ($gl = 22; p < 0,001$) indicando forte associação. Os dados foram então submetidos a uma AFC (Fig. 6), o que facilitou bastante a interpretação. Nota-se agora que o desenvolvimento interrompido ocorre com mais freqüência no tratamento com vitamina D3, especialmente se a população é de desenvolvimento precoce (PV). Por outro lado, a histólise rápida aparece associada fortemente à população de desenvolvimento precoce tratada com ecdisona (PE), e menos marcadamente com as populações precoce não tratada (PC) e tardia tratada com ecdisona (TE). As amostras tratadas com extrato de *Solanum* (TS, CS e PS) apresentam adultos viáveis com mais freqüência, especialmente se as populações originais são de desenvolvimento normal(CS).

5.2. UMA APLICAÇÃO PRÁTICA EM ANTROPOMETRIA

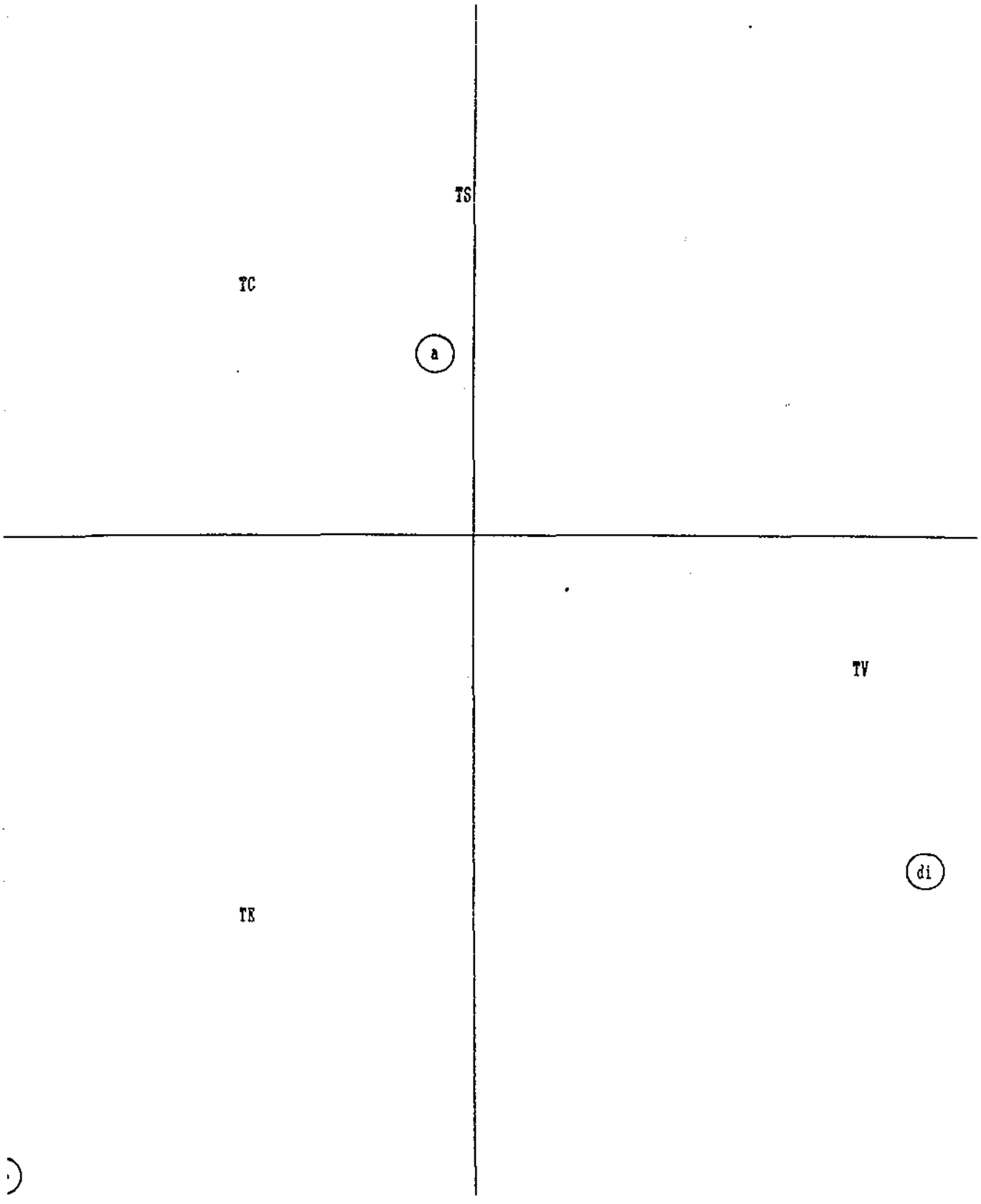
Em um estudo realizado por F. M. Salzano e colaboradores em indígenas Caingang de quatro aldeias do nosso Estado (*Salzano et al.*, 1980), foram analisados dados relativos a medidas antropométricas em 400 homens e mulheres adultos. Testes usuais de associação usando qui-quadrado resultaram em um valor estatisticamente significativo entre estatura (organizada em oito faixas de igual intervalo) e sexo ($235,4$; $gl = 7$; $p < 0,001$) e entre estatura e aldeia ($38,8; gl = 21; p < 0,05$).

A AFC permite reunir os dados relativos às três variáveis em uma única representação gráfica, considerando-se estatura como a variável dependente

TABELA 6. Número de adultos viáveis, pupas com desenvolvimento interrompido e formas com histólise rápida, obtidos após tratamento com três esteróides, em uma população de *Drosophila melanogaster* com desenvolvimento tardio.

	TRATAMENTO			
	Controle	Ecdisona	Vit. D3	*Solanum*
Adultos viáveis	165	176	189	338
Desenvol. interrompido	13	75	89	34
Histólise rápida	13	82	29	5

$\chi^2 = 194.187$; gl = 6 ; $p < 0.001$



ig.5. Desenvolvimento pupa-adulto em D. melanogaster e tratamentos que foram submetidos ovos de uma população de desenvolvimento ardio. TC: controle ; TS: extrato de S. malacoxylon; TV: vitamina 3; TE: ecdisona; a: adultos viáveis; di: pupa com desenvolvimento nterrompido; hr: histólise rápida da pupa. O eixo 1 (vertical) xplica 84% da inércia e o eixo 2, 16%.

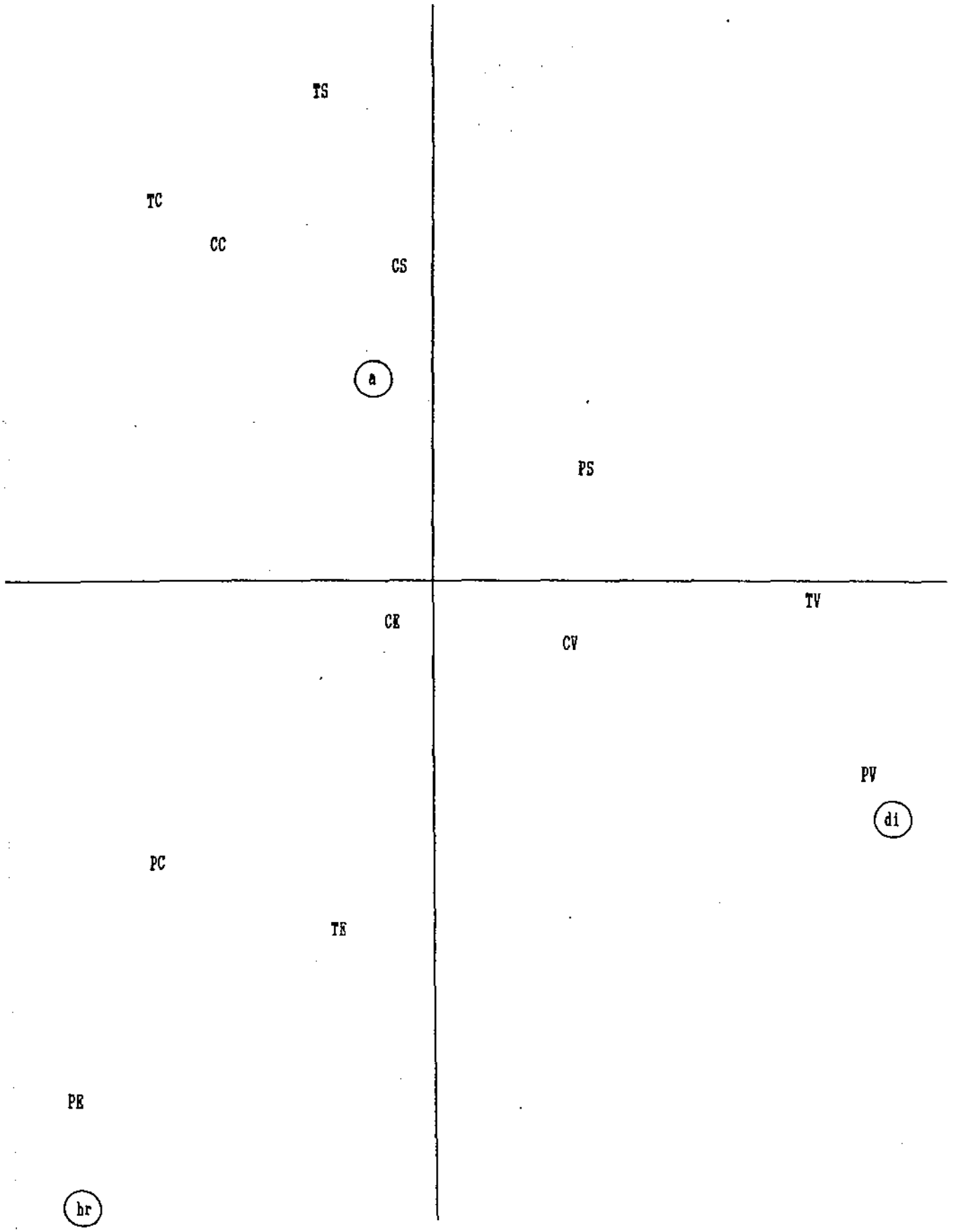


Fig.6. D. melanogaster: formas originadas de ovos submetidos a 12 tratamentos. T: pop. de desenvolvimento tardio; C: controle; P: pop. de desenvolvimento precoce; S: extrato de S. malacoxylon; V: vitamina D3; E: ecdisona; C: controle; a: adultos viáveis; di: pupa com desenvolvimento interrompido; hr: histólise rápida da pupa. O eixo 1 (vertical) representa 77% da inércia e o eixo 2, 23%.

e sexo e aldeias como variáveis independentes. Este tipo de análise é possível porque a AFC múltipla realiza a análise conjunta de tabelas justapostas de dados. A Tabela 7 apresenta os dados originais e a Fig. 7 mostra graficamente as associações encontradas, facilitando a interpretração dos resultados obtidos via qui-quadrado. Nota-se que no sexo feminino, são mais comuns as estaturas baixas ($ALT1$ a $ALT3$) e no sexo masculino, as pessoas altas são mais freqüentes ($ALT5$ a $ALT8$). Além disso, as estaturas mais elevadas são observadas na aldeia 3 ($ALD3$), ao passo que a aldeia 1 parece ser constituída de pessoas mais baixas. A estatura mais próxima da média é a $ALT4$.

É importante notar que é falsa a conclusão de que o sexo masculino é mais comum na aldeia 3, pois a tabela foi constituída de forma a considerar como variável dependente a estatura e qualquer análise da relação entre aldeia e sexo é ilegítima.

5.3. ASSOCIAÇÃO DO SISTEMA Gm COM LÍNGUA E GEOGRAFIA

A busca de associação entre freqüências gênicas e outros fatores, como a afiliação linguística, tem sido objeto de diversos estudos em populações indígenas americanas (ver, por exemplo, *Salzano & Callegari-Jacques*, 1989). Devido à grande variabilidade intra e inter-populacional, bem como ligações com genes relacionados com resistência a doenças e resposta imune, o sistema Gm de proteínas séricas (gamaglobulinas) tem sido objeto de inúmeras investigações. *Callegari-Jacques et al.* (1991) realizaram uma compilação exaustiva dos dados da literatura referentes a populações esquimós e indígenas das três Américas. Em uma das análises realizadas, foram selecionadas as 24 populações norte e centro-americanas que apresentaram um grau de miscigenação com não-indígenas inferior a 5%.

Segundo o critério de *Greenberg* (1987), estas 24 tribos reunem-se em 9 grupos lingüísticos: dois destes são classificados como Esquimós (Inuit, falado na costa do Oceano Ártico, desde o Alasca até a Groenlândia, e Yuit, com representantes que vivem na costa sudoeste do Alasca); um como Na-Dené (falado pelos Atabascos, do noroeste do Canadá e sudeste dos Estados Unidos da América (EUA); e seis como Ameríndios (Almosano e Keresiouano, falado em Labrador, centro-sul do Canadá e centro-leste dos EUA; Uto-Azteca, falado no oeste dos EUA e México; Penúcio Mexicano e Penúcio Zuni, com representantes no México; e Chibcha, falado na América Central, ao sul do Yucatán).

TABELA 7. Índios Caingang de quatro aldeias do Rio Grande do Sul: distribuição dos indivíduos segundo estatura (em cm), sexo e aldeia.

ESTATURA	SEXO		ALDEIA			
	MASC	FEM	ALD1	ALD2	ALD3	ALD4
139-143 (ALT1)	0	15	5	5	1	4
144-148 (ALT2)	4	59	16	29	8	10
149-153 (ALT3)	12	60	18	27	20	7
154-158 (ALT4)	56	29	17	25	32	11
159-163 (ALT5)	81	7	12	27	30	19
164-168 (ALT6)	60	1	12	19	26	4
169-173 (ALT7)	14	0	1	5	8	0
174 OU + (ALT8)	2	0	0	1	1	0

TABELA 8. Análise de correspondência entre o sistema Gm e línguas faladas por índios norte-americanos: proporção da inércia atribuída a cada eixo e contribuição absoluta de cada categoria à variabilidade representada pelos dois primeiros eixos.

CATEGORIA (Abreviatura)		EIXO 1 (64%)	EIXO 2 (31%)
Línguas:			
INUIT	(INUI)	17 +	7 -
YUIT	(YUIT)	12 +	18 -
ATABASCO	(ATHA)	10 +	13 -
ALMOSANO	(ALMO)	1 +	3 +
KERESIUANO	(KERE)	0	1 +
UTO-AZTECA	(UTOA)	0	36 +
PENÚCIO MEXICANO	(PENM)	5 -	3 -
PENÚCIO ZUNI	(PENZ)	0	1 +
CHIBCHA	(CHIB)	54 -	17 -
Haplótipos:			
*Gm*ag*	(ag)	1 +	14 +
*Gm*axg*	(axg)	62 -	28 -
*Gm*abOst*	(abs)	35 +	59 -
*Gm*fb*	(fb)	2 +	0
*Gm*ab*	(ab)	0	0

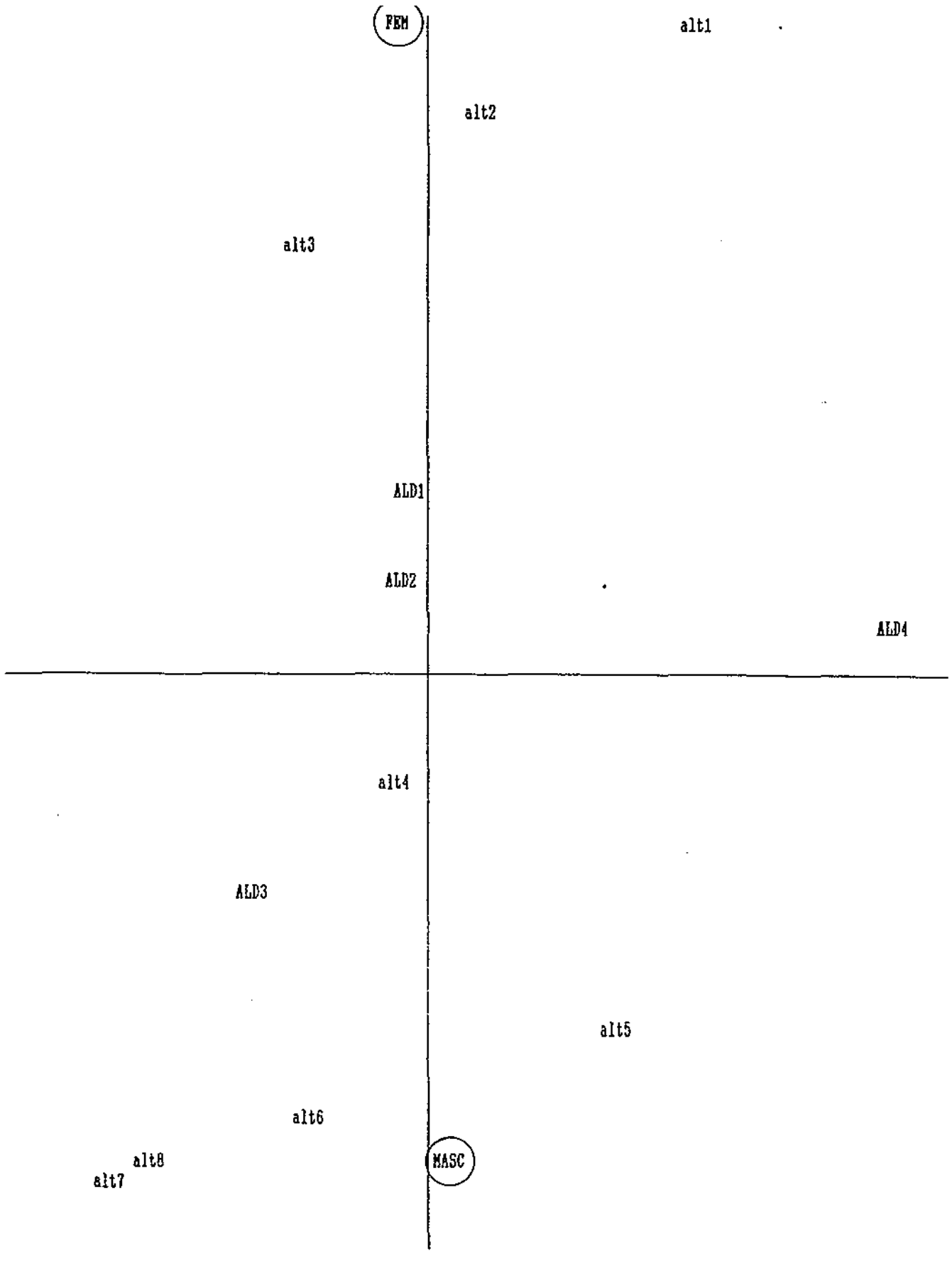


Fig.7. Análise da correspondência entre estatura, sexo e aldeia em índios Caingang do Rio Grande do Sul. O eixo principal 1 representa 93% da inércia e o eixo 2, 5%.

Os cinco haplótipos (combinações específicas de alelos em posição "cis" *Gm* analisados foram: $Gm \star ag$, $Gm \star axg$, $Gm \star ab0st$, $Gm \star fb$ e $Gm \star ab$. Na tabela de contingência utilizou-se, como freqüência observada, duas vezes o tamanho da amostra multiplicada pela freqüência do haplótipo.

A Tabela 8 mostra que os dois primeiros eixos representam 95% da variabilidade total (inércia), sendo 64% atribuído ao eixo 1. $Gm \star axg$ teve o maior peso (62%) na determinação do primeiro eixo, seguido de $Gm \star ab0st$ (35%). Os sinais associados a estes valores indicam que o eixo 1 reflete principalmente o antagonismo entre as frequências destes dois haplótipos: quando um é comum, o outro é raro.

Na Fig. 8 estão representadas as nuvens de pontos-linhas (grupos lingüísticos) e pontos-colunas (haplótipos *Gm*). Verifica-se que o eixo 1 separa $Gm \star axg$ e $Gm \star ab0st$ e que freqüências altas deste último são mais comuns nos esquimós de língua Yuit, seguidos dos Inuit e Atabascos. $Gm \star ab0st$ é mais raro em Ameríndios, especialmente nos Chibchas. Por outro lado, este grupo lingüístico apresenta freqüências altas de $Gm \star axg$, as quais decrescem até atingir os menores valores em Yuit.

$Gm \star ag$, por situar-se próximo do centro do gráfico, é o que se poderia chamar de haplótipo "padrão" e não é útil para discriminar entre as populações. Por outro lado, os indígenas que falam Penúcio Zuni e Uto Azteca são os mais próximos do que se poderia chamar de população "típica", aquela que apresentaria as freqüências médias para todos os haplótipos *Gm*. Os haplótipos $Gm \star fb$ e $Gm \star ab$, indicadores de miscigenação, aparecem associados aos esquimós Inuit e aos Chibchas.

Deste modo, a análise de correspondência mostra que, entre os indígenas da América do Norte, existe uma associação significativa do sistema Gm com grupo lingüístico e como estes distribuem-se por regiões distintas, também, indiretamente, com a localização geográfica das tribos estudadas.

A AFC foi também realizada no conjunto de 60 tribos das três Américas, para as quais havia dados disponíveis quanto os seis haplótipos do sistema *Gm*. Após a análise (resultados não apresentados aqui), as coordenadas de cada tribo relativas aos eixos 1 e 2 foram distribuídas ao longo do mapa do continente (Figs. 9 e 10). Observou-se um gradiente norte-sul bastante claro com respeito ao eixo principal 1. Quanto ao eixo 2, nota-se um segundo gradiente, menos nítido, na direção leste-oeste. Como no eixo 1 a contribuição maior foi a de $Gm \star axg$ (53%, dados não apresentados), seguida de $Gm \star ab0st$ (38%), se interpreta o gradiente como sendo devido principalmente à variação

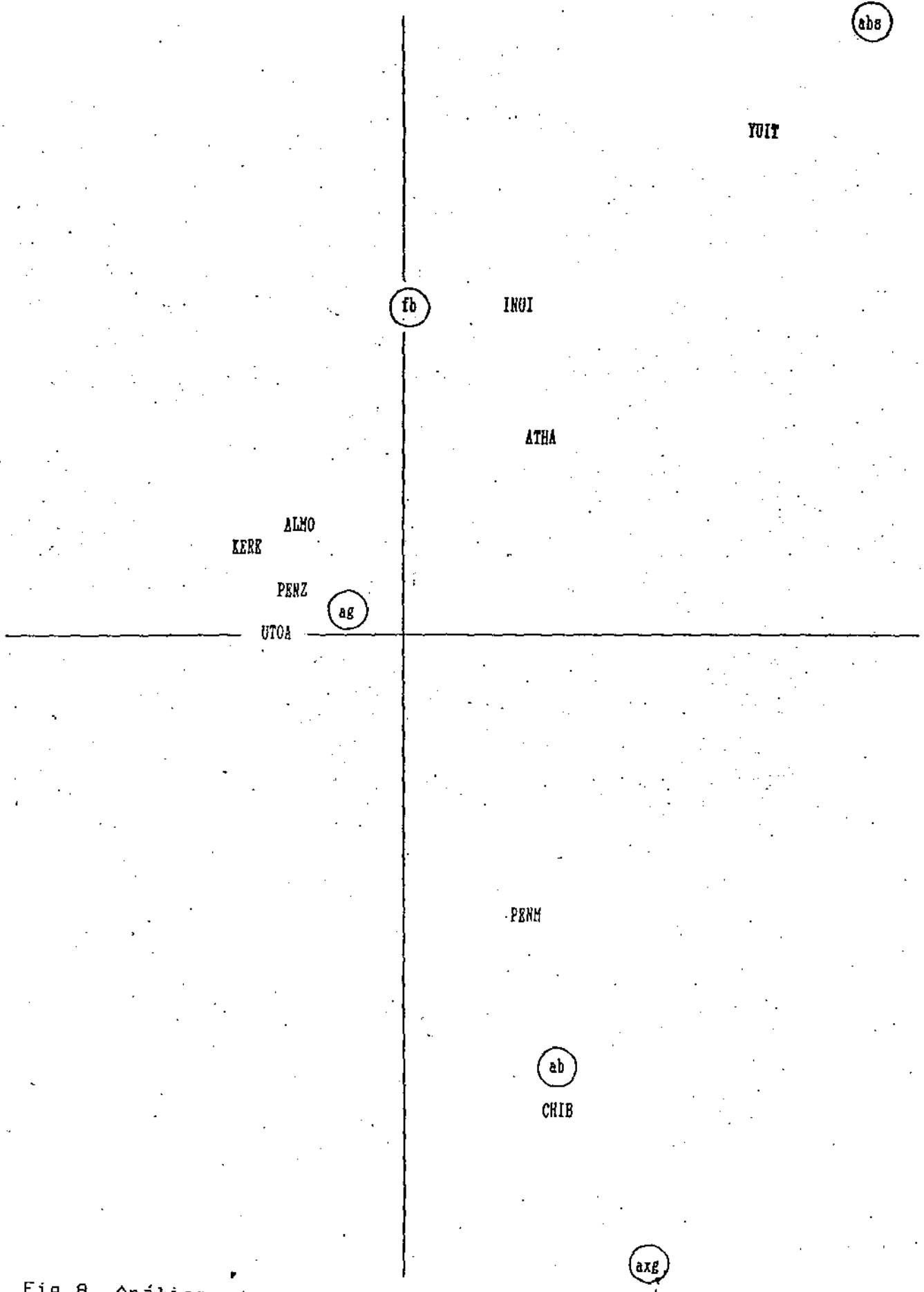


Fig.8. Análise da correspondência entre nove grupos de línguas faladas por indígenas norte e centro-americanos e cinco haplótipos do sistema Gm. O eixo 1 representa 64% da inércia e o eixo 2, 31%.

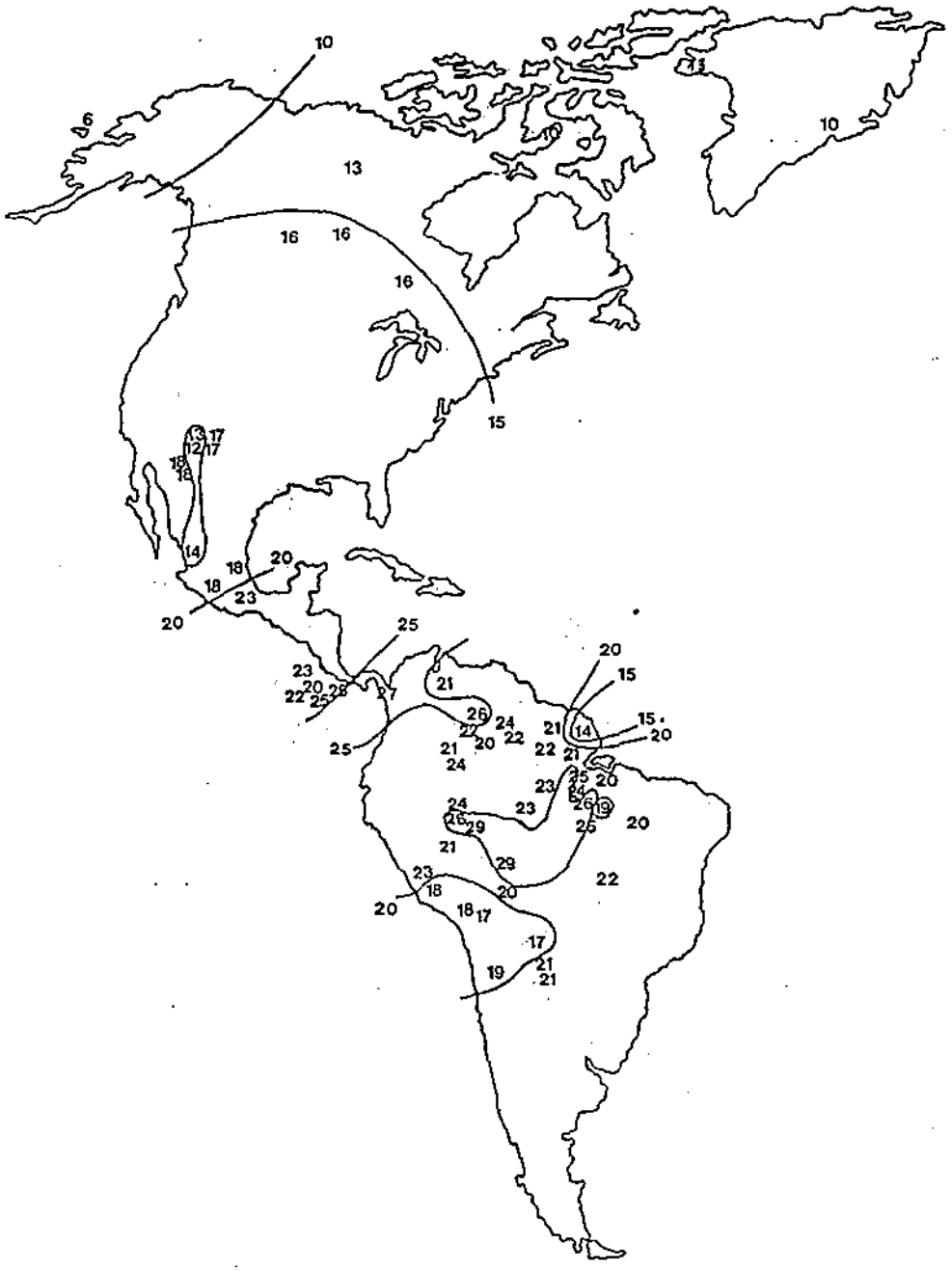


Fig.9. Análise de correspondências para 60 grupos indígenas americanos e haplótipos do sistema Gm: mapa das coordenadas relativas ao eixo 1 (57% da inércia).

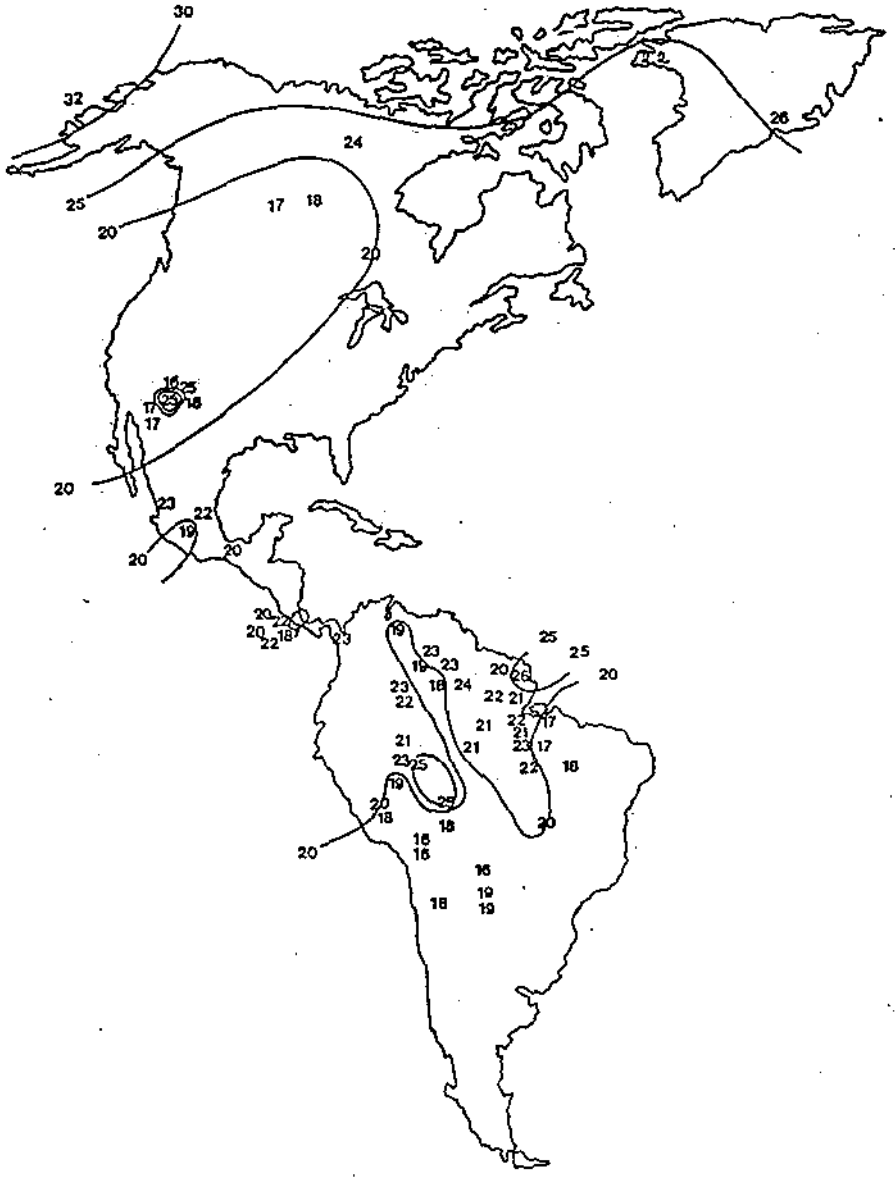


Fig.10. Análise de correspondências para 60 grupos indígenas americanos e haplótipos do sistema Gm: mapa das coordenadas relativas ao eixo 2 (32% da inércia).

norte-sul nas frequências de $Gm \star axg$, com influência menor do segundo haplótipo. No eixo 2, predominam $Gm \star ab0st$ (58%) e depois $Gm \star axg$ (27%). Assim, o padrão continua sendo ditado por estes dois haplótipos mas agora com predominância de $Gm \star ab0st$.

6. PROGRAMAS DE COMPUTADOR PARA ANÁLISE DE CORRESPONDÊNCIA

LE SPHINX - J. Moscarola & J. de Lagarde . 13 Chemim des Amarantes, 74600 Seynod, França.

NTSYS - J. Rohlf. 3 Heritage Lane, Setauket, 11733 New York, Estados Unidos da América.

Agradeço à Dra. Jandyra M. G. Fachel, do Departamento de Estatística, Instituto de Matemática da Universidade Federal do Rio Grande do Sul, a leitura cuidadosa deste texto, bem como o continuado incentivo para que eu o publicasse.

REFERÊNCIAS BIBLIOGRÁFICAS

1. BENZÉCRI, J. P. (1973) *Analyse des donées, tome II: Analyse des correspondences.* Dunod, Paris.

2. CALLEGARI-JACQUES,S.M.; SALZANO, F.M.; CONSTANS,J. & MAURIÈRES, P. (1991) Gm haplotype distribution in Amerindians - relationship with geography and language. (env. para publicação no *Yearbook of Physical Anthropology).*

3. FISHER,R.A. (1940) The precision of discriminant functions. *Ann Eugen. Lond.* **10**: 422-429.

4. GIFI, A. (1981) *Nonlinear multivariate analysis.* University of Leiden, Afdeling Datatheorie, Leiden, The Nederlands.

5. GREENACRE,M.J. (1984) *Theory and applications of correspondence analysis.* Academic Press, New York.

6. GREENACRE,M.J. & DEGOS,C. (1977) Correspondence analysis of HLA gene frequency data from 124 populations samples. *Am.J.Hum. Genet.* **29**: 60-75.

7. GREENBERG,J.H. (1987) *Language in the Americas.* Stanford University Press, Standford, Ca.

8. HIRSCHFELD,H.O. (1935) A connection between correlation and contingency. *Proc. Camb.Phil.Soc.* **31**: 520-524.

9. HILL,M.O. (1974) Correspondence analysis: a neglected multivariate method. *Appl.Statist.* **23**: 340-354.

10. HORST, P. (1935) Measuring complex attitudes. *J. Social Psychol.* 6: 369-374.

11. JUNG,I.B.C.; LACHINI,S. & OLIVEIRA,A.K. (1991) Alterações no tempo e no desenvolvimento de populações de *Drosophila melanogaster* selecionadas para velocidade de desenvolvimento tratadas com esteróides e análogo. *Rev.Bras.Genét.* **14**: 273 (resumo).

12. LEBART,L. & FENELON,J.P. (1971) *Statistique et informatique appliquées.* Dunod, Paris.

13. LEBART,L.; MORINEAU,A. & TABARD,N. (1977) *Techniques de la description statistique.* Dunod, Paris.

14. LEBART,L.; MORINEAU,A. & WARWICK, K.M. (1984) *Multivariate descriptive analysis: correspondence analysis and related techniques for large matrices.* Wiley-Interscience, New York.

15. MOSER,E.B. (1989) Exploring contingency tables with correspondence analysis. *Computer Applications in Bioscience*, 5: 183-189.

16. MURTAGH, F. & HECK, A. (1987) *Multivariate data analysis.* D. Reidel, Boston.

17. NISHIATO,S. (1980) *Analysis of categorical data: dual scalling and its applications.* University of Toronto Press, Toronto.

18. SALZANO,F.M. & CALLEGARI-JACQUES,S.M. (1989) *South American Indians: A case study in evolution.* Clarendon Press, Oxford.

19. SALZANO,F.M.; CALLEGARI-JACQUES,S.M.; FRANCO,M.H.L.P.; HUTZ,M.H.; WEIMER,T.A. & ROCHA,F.J. (1980) The Caingang revisted: blood genetics and anthropometry. *Am.J.Phys.Anthropol.* 53: 513-524.

20. SOUZA,A.M.R. de (1982) *Análise de correspondência.* Dissertação de Mestrado, Instituto de Matemática e Estatística, Universidade de São Paulo.

21. SOUZA,N.M. de (1990). *Análise de correspondência.* Monografia, Bacharelado em Estatística, Universidade Federal do Ro Grande do Sul.

22. VERDINELLI,M.A. (1980) *Análise inercial em Ecologia.* Tese de Doutoramento, Instituto Oceanográfico, Universidade de São Paulo, São Paulo.

Universidade Federal do Rio Grande Sul
Reitor: Professor Tuiskon Dick

Instituto de Matemática
Diretor: Professor Aron Taitelbaum
Núcleo de Atividades Extra Curriculares
Coordenador: Professor Jandyra M. G. Fachel
Secretária: Faraildes Beatriz da Silva

Os Cadernos de Matemática e Estatística publicam as seguintes séries:

Série A: Trabalho de Pesquisa
Série B: Trabalho de Apoio Didático
Série C: Colóquio de Matemática SBM/UFRGS
Série D: Trabalho de Graduação
Série E: Dissertações de Mestrado
Série F: Trabalho de Divulgação
Série G: Textos para Discussão

Toda correspondência com solicitação de números publicados e demais informações deverá ser enviada para:

NAEC - Núcleo de Atividades Extra Curriculares
Instituto de Matemática - UFRGS
Av. Bento Gonçalves, 9500
91.500 - Agronomia - POA/RS
Telefone: 36.98.22 ou 39.13.55 Ramal: 6176