Análise dos componentes principais e análise fatorial na pesquisa geográfica: alguns problemas e questões*

R. J. Johnston
Universidade de Sheffield

RESUMO

São discutidos, considerando-se as estruturas geográficas, quatro problemas relativos ao uso da análise dos componentes principais, baseados nos coeficientes de correlação. A conclusão geral é de que o método não se aplica a muitos conjuntos de dados geográficos.

1 — INTRODUÇÃO

As técnicas de análise dos componentes principais e de análise fatorial tornaram-se muito comuns para a pesquisa geográfica. Infelizmente, a natureza de muitos dos conjuntos de dados usados numa pesquisa deste tipo apresenta um número de problemas que poderiam surgir na interpretação dos resultados analíticos. O objetivo do presente trabalho é o de esclarecer quatro destes pro-

* Tradução de Angela Maria Rocha Lima Diego, do CEDIT. Transcrito de *The South African Geographical Journal*, Vol. 59, n.º 1, april, 1977, com autorização dos editores.

blemas, fazendo-se uma referência particular à pesquisa na Geografia Humana.

O objetivo principal das duas técnicas é a redução do número de variáveis "significantes" numa matriz de dados pela retirada das redundâncias linearmente relacionadas. Com efeito, as técnicas são descrições complexas das relações lineares, embora sejam usadas freqüentemente como testes de hipóteses, e existam métodos (por exemplo, análise fatorial de múltiplo grupo; Timms, 1971) que fornecem provas mais exatas. O componente principal e a análise fatorial diferem no tratamento que dão ao "problema da comunalidade". Na análise dos componentes principais são analisadas todas as variâncias no conjunto de dados, enquanto que na análise fatorial apenas a variância comum estimada — sendo analisada a porção da variância em cada variável que está relacionada em outras variáveis — está sujeita a um exame minucioso (para uma descrição completa de diferenças entre os métodos, ver Rummel, 1970). Estas diferenças são irrelevantes para o objetivo do presente trabalho, e a discussão entrará no quadro da análise dos componentes principais.

A análise dos componentes principais baseia-se na análise mais geral de *eigenfunctions*. Embora os *eigenvaluer* e *eigenvectors* possam ser extraídos de qualquer matriz simétrica quadrada, a maior parte das análises de componentes principais na Geografia tem operado em matrizes de coeficientes de correlação Produto Momento de Pearson (Johnston, 1973a). Assim, a partir de uma matriz dados D compreendendo n colunas que representam as diferentes variáveis, e m linhas relativas às unidades de observação, forma-se uma matriz R de correlação $n \times n$, a qual é *input* para obtenção de *eigenfunctions* da análise de componentes principais. A derivação da matriz R é, assim, crucial para a série total de operações.

2 — O PROBLEMA DOS DADOS

Um aspecto comum a muitas análises de componentes principais em Geografia — tal como, a amplamente conhecida ecologia fatorial — é que os dados estão em forma de proporção ou percentagem. Tais dados podem influenciar seriamente a natureza da matriz R de correlação, e, portanto, da solução dos componentes principais, especialmente quando todas as categorias de um mesmo conjunto de dados (por exemplo, todos os grupos de idade ou todas as categorias de renda), estão incluídas.

A figura 1 apresenta um exemplo bastante simplificado de como isto funciona. Temos seis variáveis em nossa análise, das quais duas são as seguintes:

X_1 = percentagem da população masculina da cidade;

X_2 = percentagem da população feminina da cidade.

A definição das variáveis X_3 a X_6 é irrelevante, exceto para se observar que elas não envolvem categoria de um mesmo fenômeno. As variáveis X_1 e X_2 juntas formam um conjunto de números fechados, de forma que:

$$X_1 + X_2 = 100 \qquad (1)$$

dispomos de dados para cinco cidades, para as quais os valores de X_1 são:

CIDADES	VALORES X_1
1	45
2	50
3	53
4	60
5	70

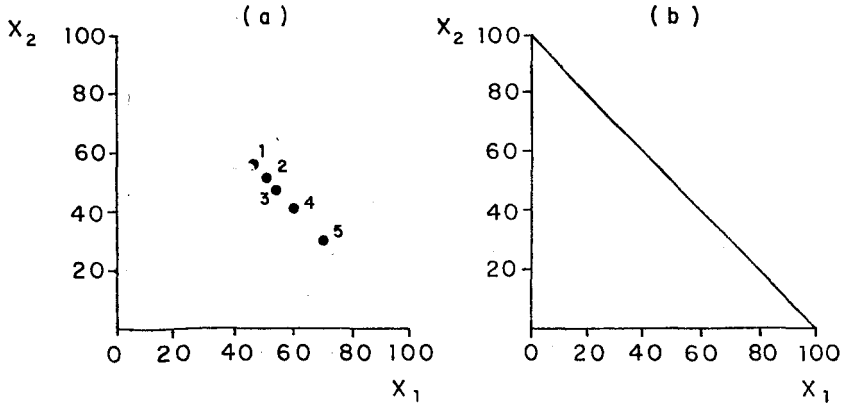

Fig. 1 — *Correlação entre X_1 (percentagem de homens) e X_2 (percentagem de mulheres).*

Em função da igualdade definida na equação (1), os valores de X_2 estão agora fixados. Representando graficamente, poderemos observar, na figura 1(a), que — e não é de se surpreender — todos caem numa linha reta, e a correlação entre X_1 e X_2 é $r_{12} = -1{,}00$.

O padrão descrito na figura 1(a) pode ser generalizado para mostrar as restrições na correlação. Isto é feito na figura 1(b). Se os dados analisados satisfizerem a igualdade da equação (1), é na linha diagonal que todas as observações cairão.

A conseqüência de se introduzir as duas variáveis X_1 e X_2 numa análise de componentes principais pode agora ser demonstrada. Para nossas seis variáveis a matriz de correlação é como se vê na tabela 1, e a matriz de correlação é reproduzida num espaço vetorial bidimensional na figura 2(a) (para representação de matrizes de correlação ver Rummel, 1967, 1970). Usando o método do centróide, a posição da primeira componente principal foi calculada, dando os *loadings* apresentados na tabela 2. O diagrama e os *loadings* indicam a inclusão feita de ambos X_1 e X_2; uma vez que as duas variáveis são altamente e inversamente correlacionadas, *estamos de fato medindo a mesma variável duas vezes*. Isto é uma função dos nossos dados e não algum aspecto mais importante ou inter-relação estrutural. Por serem X_1 e X_2 duplas medidas de X_1, a posição dos componentes

TABELA 1

Matriz de correlação hipotética

VARIÁVEIS	X_1	X_2	X_3	X_4	X_5	X_6
X_1	1,00	—1,00	0,98	—0,50	0,34	—0,50
X_2	—1,00	1,00	—0,98	0,50	—0,34	0,50
X_3	0,98	—0,98	1,00	—0,64	0,17	—0,34
X_4	—0,50	0,50	—0,64	1,00	0,64	—0,50
X_5	0,34	—0,34	0,17	0,64	1,00	—0,98
X_6	—0,50	0,50	0,34	—0,50	—0,98	1,00

(antes e depois da rotação) é influenciada na direção de X_1; dá-se duas vezes o peso das variáveis $X_3 \ldots . X_6$. Isto pode ser visto na figura 2(b), que é a matriz de correlação excluindo X_2. O componente está mais afastado de X_1 do que estava na figura 2(a), conforme indicado pelos *loadings* na tabela 2.

A inclusão de duas variáveis que reúnem as exigências de igualdade da equação (1) é extremamente rara, porque está claro que se calcularmos a percentagem masculina e a percentagem feminina, a mesma coisa estará sendo mensurada duas vezes. Entretanto, são freqüentemente empregados conjuntos de números fechados que envolvem mais que duas categorias, e eles também introduzem distorções consideráveis para uma análise de componentes principais.

Vejamos o exemplo do conjunto de dados hipotéticos da tabela 3, onde há três variáveis formando a igualdade:

$$X_1 + X_2 + X_3 = 100 \qquad (2)$$

Estamos interessados na correlação entre X_1 e X_2, r_{12}. Para a divisão eleitoral A, $X_1 = 50$ e $X_2 = 25$. O valor máximo de X_1 é 75, sendo $X_3 = 25$, e isto limitaria $X_2 = 0$; similarmente, se $X_2 = 70$, sendo $X_3 = 25$, e então $X_1 = 5$. Assim, uma vez que o valor de X_1 foi dado na figura 1(b), o valor de X_2, foi fixado, então na figura 3(a), uma vez que o percentual de Nacionalista é conhecido, a percentagem do *United Party* é fixada, dado que a percentagem do *Progressive*

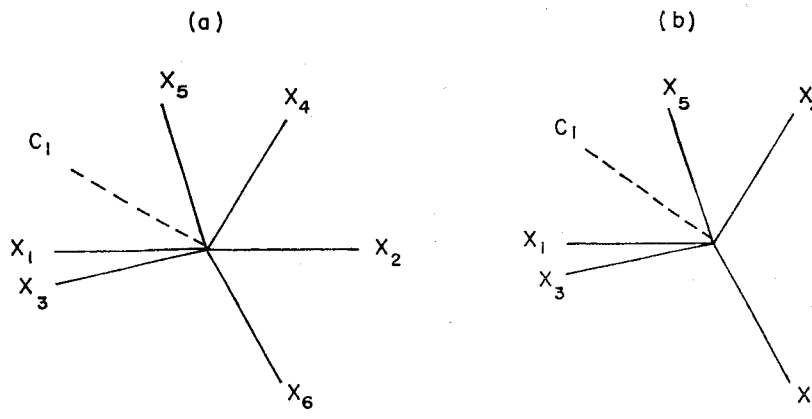

Fig. 2 — *Representação geométrica de duas matrizes de correlação — uma incluindo e outra excluindo a variável X_3 — e as aproximações dos primeiros componentes principais destas matrizes.*

TABELA 2

Loadings relativos à primeira componente principal

	X_1	X_2	X_3	X_4	X_5	X_6
Com todas as seis variáveis incluídas.....	0,88	—0,88	0,84	—0,77	0,71	—0,82
Excluindo a variável X_2.....	0,81		0,76	0,80	0,76	—0,86

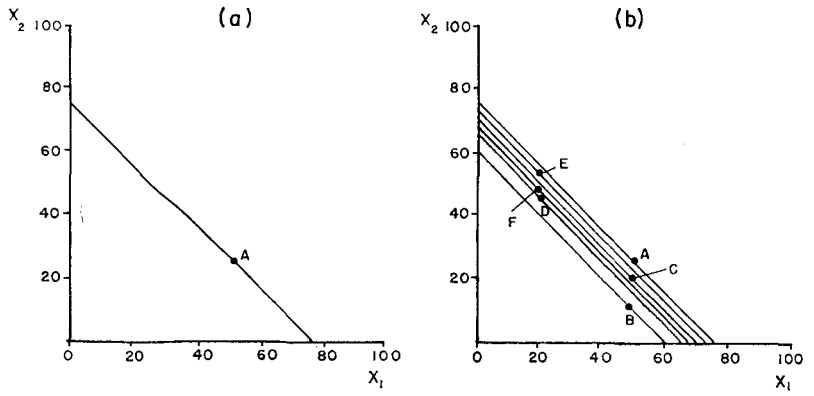

Fig. 3 — *Correlação entre X_1 (percentagem Nacionalista) e X_2 (percentagem do United Party) com X_3 (percentagem do Progressive-Reform) sendo dada.*

Party é também fixada. A única diferença entre as figuras 3(a) e 1(b) é que enquanto no último caso a linha diagonal ao longo da qual todos os valores devem ficar passa através dos pontos $X_1 = 100$, $X_2 = 0$ e $X_1 = 0$, $X_2 = 100$, no primeiro ela passa através de $X_1 = 75$, $X_2 = 0$ e $X_1 = 0$, $X_2 = 75$, sendo, o valor máximo de X_1 e X_2, fixado pelo valor de X_3.

Se agora representarmos graficamente as linhas diagonais para todas as seis observações (Figura 3(b)), veremos que a distribuição dos valores para X_1 e X_2 se restringe a um lado relativamente estreito do diagrama bidimensional. Conseqüentemente, não é de se estranhar que X_1 e X_2 sejam alta e inversamente correlacionados ($r_{12} = -0{,}95$). Na verdade, as restrições do "lado" do gráfico no qual os valores de X_1 e X_2 podem ser colocados tornam uma correlação negativa relativamente alta quase uma certeza, mesmo com uma distribuição aleatória de seus valores.

O que aconteceria se houvesse uma ampla escala de valores para X_3? Dados hipotéticos deste tipo são mostrados na tabela 4 e na figura 4. A correlação de X_1 e X_2 é claramente positiva ($r_{12} = +0{,}30$). Um estudo do diagrama sugere que a primeira situação é quase que certamente impossível; seria possível se conseguir uma correlação negativa entre X_1 e X_2, mas pela lei das probabilidades o sistema é tão limitado que seria uma ocorrência rara.

TABELA 3

Conjunto de dados hipotéticos I

DIVISÃO ELEITORAL	*NACIONALIST* (X_1)	*UNITED PARTY* (X_2)	*PROGRESSIVE-REFORM* (X_3)
A.....	50	25	25
B.....	50	10	40
C.....	50	20	30
D.....	20	45	35
E.....	20	52	28
F.....	20	48	32

TABELA 4

Conjunto de dados hipotéticos II

DIVISÃO ELEITORAL	PERCENTAGEM DOS VOTOS COMPUTADOS		
	NACIONALIST (X_1)	*UNITED PARTY* (X_2)	*PROGRESSIVE-REFORM* (X_3)
A.....	10	10	80
B.....	25	15	60
C.....	30	30	40
D.....	20	50	30
E.....	20	60	20
F.....	45	45	10

A conclusão a que queremos chegar é que com um conjunto de números fechados preenchendo as exigências de igualdade da equação (2), é provável que, correlacionando X_1 com X_2:

(1) se a amplitude de valores de X_3 é relativamente pequena, é provável que r_{12} seja negativo;

(2) se a amplitude de valores de X_3 é relativamente grande, pode ocorrer r_{12} positivo, embora seja mais provável que a correlação seja negativa;

(3) é provável que estas correlações sejam "significantemente" diferentes de zero.

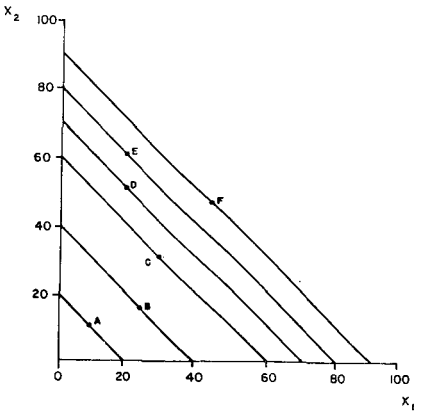

Fig. 4 — *Correlação entre X_1 (percentagem Nacionalista) e X_2 (percentagem do United Party) dada uma larga escala de valores para X_3 (percentagem do Progressive-Reform).*

A realidade destas conclusões pode ser demonstrada por uma pequena simulação, usando-se os dados das tabelas 3 e 4. Nestas tabelas, os valores de X_3 são tomados como fixos, e os valores de X_1 são simulados através do uso de tabelas de números aleatórios. Assim, os valores de X_2 são também fixados. As correlações de X_1 com X_2 são mostradas por dez simulações, cada uma com:

(1) uma pequena amplitude de valores para X_3 (Tabela 3);

(2) uma ampla amplitude de valores para X_3 (Tabela 4); e,

(3) nenhuma restrição — de forma que ambos X_1 e X_2 são simulados por números aleatórios.

Os dados e correlações resultantes são mostrados na tabela 5, e as distribuições de freqüência para os coeficientes de correlação estão na figura 5. É claro que uma pequena amplitude de valores para X_3 produzirá uma alta correlação negativa entre X_1 e X_2 — mesmo se as distribuições de valores para estas duas variáveis são produzidas probabilisticamente, como na figura 5(a). Com uma escala de valores mais ampla para X_3, são possíveis significativas correlações positivas, figura 5(b), mas são ainda mais prováveis correlações negativas (isto é porque a área total do gráfico bidimensional disponível é, de fato, um triângulo isósceles). Estas duas distribuições, simuladas sob a situação de restrição de igualdade (2), estão em contraste marcante com as distribuições de correlações entre núme-

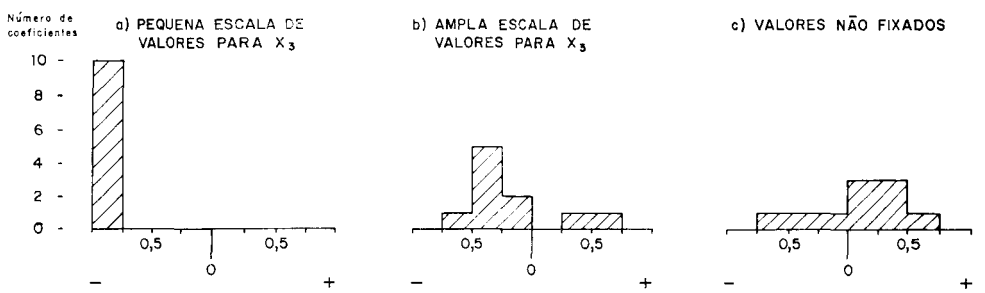

Fig. 5 — *Distribuições de freqüência dos coeficientes de correlação na tabela 5.*

TABELA 5

Correlação simulada de conjuntos numéricos fechados

	DIVISÃO ELEITORAL						
	A	*B*	*C*	*D*	*E*	*F*	r_{12}
Pequena amplitude de valores para X_3							
Valor fixado de X_3	25	40	30	35	28	32	
	25(50)	31(29)	61(9)	9(56)	26(46)	54(14)	— 0,96
	3(72)	57(3)	1(69)	33(32)	23(49)	17(51)	— 0,99
	14(61)	9(51)	25(45)	25(40)	10(12)	49(19)	— 0,94
	38(37)	25(35)	0(70)	48(17)	63(9)	59(9)	— 0,97
Valores simulados de $X_1(X_2)$	11(64)	11(49)	28(42)	50(15)	37(35)	43(25)	— 0,95
	43(32)	55(5)	35(36)	3(62)	65(7)	50(18)	— 0,97
	49(32)	16(44)	53(17)	49(16)	42(30)	51(17)	— 0,94
	36(39)	26(34)	25(45)	17(48)	50(22)	18(50)	— 0,91
	7(68)	26(34)	11(59)	4(61)	12(60)	59(9)	— 0,97
	43(32)	12(48)	7(63)	31(34)	44(28)	33(35)	— 0,94
Ampla amplitude de valores para X_3							
Valor fixado de X_3	80	60	40	30	20	10	
	10(10)	4(36)	38(22)	59(11)	74(6)	78(12)	— 0,64
	16(4)	34(6)	22(38)	4(66)	49(31)	79(11)	— 0,48
	17(3)	19(21)	56(4)	60(10)	55(25)	88(2)	— 0,28
	8(12)	23(17)	46(14)	46(24)	57(23)	53(37)	+ 0,66
Valores simulados de X_1 (X_2)	1(19)	4(36)	27(33)	27(43)	8(72)	4(86)	— 0,20
	1(19)	37(3)	9(51)	5(65)	47(33)	12(78)	— 0,45
	4(16)	26(14)	19(41)	56(14)	64(16)	43(47)	— 0,11
	1(19)	36(4)	46(14)	1(69)	55(25)	30(60)	— 0,44
	3(17)	6(34)	15(45)	37(33)	50(30)	41(49)	+ 0,35
	9(11)	9(31)	16(44)	8(62)	65(15)	16(74)	— 0,40
Nenhum valor fixado							
	89(22)	5(66)	41(27)	38(50)	98(44)	34(87)	— 0,57
	4(0)	39(30)	0(47)	31(82)	37(24)	36(85)	+ 0,37
	98(54)	64(59)	86(13)	91(43)	38(39)	52(25)	+ 0,11
	41(71)	7(51)	33(84)	38(63)	31(26)	52(18)	— 0,20
	28(73)	86(59)	61(87)	10(23)	86(90)	32(71)	+ 0,60
Valores simulados de X_1 (X_2)	65(21)	83(73)	43(89)	59(98)	29(97)	37(59)	— 0,33
	65(55)	8(59)	29(93)	63(80)	23(22)	15(37)	+ 0,35
	97(83)	93(91)	0(37)	18(63)	89(0)	80(73)	+ 0,23
	4(16)	58(55)	56(89)	66(50)	27(77)	18(49)	+ 0,52
	86(59)	43(84)	62(65)	1(59)	78(90)	54(45)	+ 0,20

ros aleatórios, escolhidos sem restrições. Podemos concluir então que, *ceteris paribus* é mais provável que o uso de um conjunto de números fechados produza uma correlação diferente de zero, do que se não houvesse restrições nos valores das variáveis consideradas.

Com respeito ao uso dos conjuntos de números fechados nas análises de componentes principais, as implicações da discussão acima são que elas podem influenciar severamente a estrutura do componente resultante. Vemos na figura 2 que o uso de duas variáveis medindo a mesma coisa influencia a posição de componentes. Com conjuntos de números fechados compreendendo três variáveis, ao invés de duas, vimos que, mesmo com uma distribuição aleatória dos valores, as restrições do sistema são tais que são prováveis correlações que não sejam zero, negativo ou positivo, de acordo com a escala de valores para a terceira variável. Estas quase que com certeza não seriam extremamente altas como no caso da figura 2, mas poderiam ser suficientemente significativas para que, especialmente se um número de correlações como estas estivesse na matriz R, a natureza da estrutura do componente fosse de fato pré-determinada pelo conjunto de dados.

Os conjuntos de números fechados com três variáveis têm sido usados nos estudos geográficos. Os mais comuns são conjuntos com cerca de cinco a dez categorias, entretanto, como em estudos que envolvem todos os grupos etários numa população, todas as filiações religiosas, ou todos os grupos raciais. Estes estão sujeitos às mesmas restrições, conforme ilustrado nas figuras 3 e 4; se considerarmos a correlação r_{12}, a igualdade é:

$$X_1 + X_2 + (X_3 + X_4 + \ldots X_n) = 100, \quad (3)$$

onde n é o número de categorias no conjunto de números fechados. Com cinco variáveis, existem ao todo dez pares de intercorrelações (X_1X_2, X_1X_3, etc.). Em cada caso é criada uma situação de conjuntos de números fechados, porque, como no caso do par X_1X_3, uma vez que $(X_2 + X_4 + X_5)$ e X_1 são fixados, então X_3 não tem graus de liberdade. Usando-se um conjunto de números fechados que compreende qualquer número de categorias, portanto, é provável manter-se numa matriz R de correlação os valores de r_{xy} que influenciarão a estrutura do componente resultante. *Isto não significa que nenhum dos valores de r_{xy} que não seja zero seja real e, portanto, substancialmente interpretável: isto não implica, entretanto, que pelo menos parte do valor de r_{xy} seja o produto de distribuições dos valores das outras variáveis no conjunto. Desde que não podemos separar esta parte, não podemos interpretar nem as correlações nem os loadings.*

A idéia desta crítica pareceria ser a de que conjuntos de números fechados não deveriam ser usados nas análises de componentes. Isto criaria problemas para a pesquisa geográfica, na qual todas as categorias de um conjunto — toda a renda ou grupos ocupacionais, por exemplo, — deveriam ser incluídas num estudo para conseguir a descrição exigida. É possível se evitar os problemas aqui discutidos, trocando-se o denominador das equações de percentagem. Com um conjunto de três variáveis, poderemos ter, então,

$$V_1 = X_1/(X_1 + X_2 + X_3) \times 100 \quad (4)$$

$$V_2 = X_2/(X_2 + X_3) \times 100 \quad (5)$$

e a correlação de V_1 e V_2 evitaria as restrições impostas para a correlação de X_1 com X_2. Com mais do que três categorias no conjunto,

entretanto, o número de denominadores separados é grande e a consistência de seus resultados é incerta. Um método alternativo, embora ainda pouco explorado, seria usar um algoritmo de escala multidimensional não métrica numa matriz de índices de desigualdade (Duncan & Duncan, 1955) para todos os pares de variáveis no conjunto de números fechados, como foi feito por Klaff (1973). As escalas restauradas de análises de vários conjuntos poderiam, então, se desejado, ser introduzida numa análise de componentes, se a escala fosse em uma matriz para todos os pares de observações em vez de para todos os pares de variáveis.

3 — INTENSIDADE E SEGREGAÇÃO

Muitas análises de componentes principais e análise fatorial na Geografia Humana são realizadas para descrever o grau de padronização de um conjunto de variáveis para um número de áreas. As técnicas são usadas por causa da redundância nas variáveis, e da necessidade de isolar as "dimensões básicas" e descrever seus padrões espaciais. Assim, no agora bem conhecido "campo" da ecologia fatorial, o objetivo é descrever a posição de cada unidade de observação — usualmente uma área residencial mais ou menos relevante como uma "vizinhança" ou "comunidade" — num *continuum*, por exemplo, de *status* sócio-econômicos. E desde que o objetivo da maior parte da ecologia fatorial não é fornecer um grande número de descrições únicas, mas desenvolver uma teoria de *ecologia urbana comparativa* (Berry, 1971), os pesquisadores desejam contrastar a natureza de vários padrões residenciais. Tais contrastes podem ser interurbanos, como nos estudos da Nova Zelândia (Timms, 1970; Johnston, 1973b), ou podem ser intra-urbanos, comparando, talvez, a padronização residencial nas várias áreas de grupo de Cape Town e Durban, Johannesburg, Pretoria e Port Elizabeth.

Mas as análises dos componentes não podem fornecer toda a informação necessária para ecologias urbanas comparativas. Elas podem descrever a covariância entre grupos de variáveis e, através da derivação de *scores*, podem situar qualquer área num *continuum* derivado das variáveis. Mas, como são usualmente conduzidas, elas não podem ser usadas para descrever a intensidade de segregação residencial, ou seja, até que ponto dois grupos raciais vivem separados espacialmente. A razão para isto repousa na natureza dos coeficientes de correlação usados na matriz básica R que é *input* na análise de componentes. A fórmula para um coeficiente de correlação *produto-momento* é:

$$\frac{\Sigma\ (x_i - \bar{x})\ (y_i - \bar{y})\ 1/N}{\delta_x\ \delta_y}, \quad (6)$$

onde x_i é o valor por variável x na observação i; $\bar{x}$ é a média de todos os valores de x, y_i; y refere-se à variável y, e δ_x, δ_y são os desvios-padrão das variáveis x e y respectivamente. Disto, vemos que os coeficientes de correlação são baseados em números adimensional, ou desvios Z, nos quais cada valor de x é reescrito em termos da média e desvio-padrão desta ordenação, de forma que:

$$Z_{xi} = (x_i - \bar{x})/\delta_x \quad (7)$$

Este processo de padronização torna cada conjunto de números comparável, e assegura que os coeficientes de correlação caiam na escala $\pm$ 1,0, como também garante que a "intensidade" ou tamanho

dos valores originais seja insignificante para a análise.

Em termos de ecologia fatorial, e particularmente ecologia fatorial comparativa, a importância desta afirmação é que apenas a padronização relativa de cidades, e não a intensidade desta padronização, pode ser descrita. A quantidade de segregação é eliminada uma vez que as variáveis originais e os *scores* são escritos na forma de desvio *Z*. Por exemplo, temos duas cidades, cada uma dividida em seis áreas, para as quais medimos a percentagem da população que é formada por membros da Igreja Protestante Holandesa. Os dados originais são:

CIDADES	ÁREAS					
	A	B	C	D	E	F
I	60	54	48	52	46	40
II	82	63	44	56	37	18

Para cada cidade a percentagem em média de aderentes em cada área é a mesma — $X_I = X_{II} = 50$ — mas o desvio-padrão difere — $\delta_I = 6{,}325$ e $\delta_{II} = 20{,}000$. Entretanto, estas últimas diferenças são suprimidas numa análise de correlação usando as variáveis, uma vez que em termos de desvio *Z*, os valores são:

CIDADES	ÁREAS					
	A	B	C	D	E	F
I	+1,58	+0,63	−0,32	+0,32	−0,63	−1,58
II	+1,60	+0,65	−0,30	+0,30	−0,65	−1,60

Em ecologias fatoriais separadas de duas cidades, a maior segregação espacial dos membros da Igreja Protestante Holandesa na cidade II poderia passar desapercebida.

Portanto, as análises de componentes descrevem padrões relativos sem dimensão de separação espacial, e não a intensidade absoluta da separação. Esta descrição relativa pode ser tudo o que se deseja. Mas se, como geógrafos, estamos interessados na intensidade, estaremos, então, anulando nossos fins, se usarmos as análises de componentes como tem sido feito atualmente na maioria dos estudos. Há alternativas para este problema na metodologia geral, como Berry (1961) demonstrou num antigo ensaio, mas que poucos seguiram; isto envolve evitar o processo de padronização, ou, alternativamente, de reescala, depois dos *eigenfunctions* terem sido extraídos. Sugeri outras alternativas em outra parte deste trabalho, e os argumentos usados não serão aqui repetidos (Johnston, 1973a, 1976). Nenhum deles é totalmente satisfatório, e será preciso fazer uma investigação futura do problema; uma escala multidimensional oferece possibilidades, embora seja necessária uma reescala para preservar as intensidades de separação espacial mostrada pelos índices de desigualdade.

4 — "HETEROSCEDASCITY" E MÉDIAS RESIDUAIS

Duas das exigências do modelo linear geral são a igualdade de médias e de variâncias na distribuição condicional de resíduos, e a *homoscedascity*. Se uma das duas não for alcançada, o resultado é um coeficiente de regressão distorcido (Poole & O'Farrell, 1971; Mather & Openshaw, 1974). Ambas são freqüentemente violadas nas análises de componentes principais de matrizes de correlação.

A primeira violação refere-se ao problema do conjunto de números fechados discutido acima. Conforme a figura 1 (b) e figura 3 (a), a dispersão possível para qualquer ponto está ao longo de uma única linha diagonal; a dispersão máxima para um conjunto de pontos

está num triângulo isósceles limitado pela diagonal mostrada na figura 1(b), partindo-se do princípio de que a escala de valores para $(X_3 + X_4 + \ldots X_n)$ é 0 a 100. Se há uma distribuição uniforme de valores de X_1 e X_2 sobre a "escala de restrição", conforme mostrado na figura 4, não é provável, então, que as exigências de médias iguais sejam seriamente violadas, mas a concentração da maior parte dos pontos numa parte da escala poderia levar a pequenas observações influenciando consideravelmente o valor de r_{12}, e, assim, a natureza da estrutura do componente. A figura 6 ilustra isto por um conjunto de dados hipotéticos; os valores para duas observações — *A* e *B* — estão seriamente deslocadas com a tendência identificada pela maioria dos pontos, e "influenciará" a correlação. Isto não se tornará um problema sério, mas alguns "valores estranhos" poderiam influenciar substancialmente a correlação e as estruturas do componente.

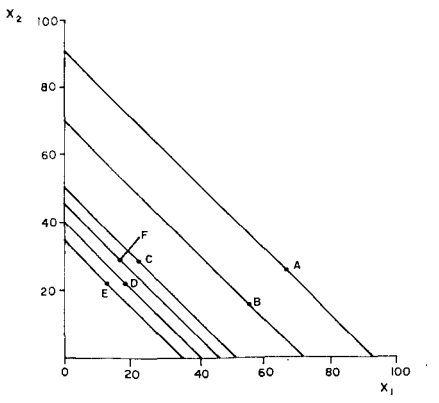

Fig. 6 — *Um exemplo hipotético das médias desiguais e problemas de variância numa correlação que envolve um conjunto de números fechados.*

O segundo tipo de violação é possivelmente muito mais sério, uma vez que tem implicações importantes (Johnston, 1971); isto porque o tipo de inter-relação entre duas variáveis não está descrito adequadamente por um coeficiente de correlação. Vejamos o caso das duas variáveis seguintes:

X_1 — percentagem da população recebendo mais de R10 000 por ano; e

X_2 — percentagem da população nascida no sul da Europa.

A "teoria" residencial "clássica" locação/alocação nos diz que é provável que imigrantes, como os vindos do sul da Europa, estão concentrados em áreas residenciais de baixa renda, e então poderíamos antecipar um alto valor negativo para r_{21}. Mas o que aconteceria se houvesse relativamente poucos imigrantes sul europeus em nossa cidade, e estes estivessem concentrados em poucas das áreas residenciais de baixa renda? A distribuição de pontos seria, então, como o da figura 7. Algumas das áreas de baixa renda — à esquerda do eixo X_1 — têm altas percentagens de sul europeus, mas outras não; todas as áreas de alta renda — à direita do eixo X_1 — têm baixas percentagens de sul europeus. Assim, o que temos é um conjunto de relações lógicas, onde se lê:

se *A* (alta renda), então *B* (poucos sul europeus) *e*

se *C* (muitos sul europeus), então *D* (baixa renda) *mas não*

se *D*, então *C*.

Um conjunto deste tipo não pode ser propriamente descrito por um coeficiente de correlação, mesmo depois da transformação de X_1 e X_2 para tentar alcançar linearidade. Ajustando uma linha de regressão à distribuição de pontos na figura 7, produz-se uma correlação de apenas $r_{21} = -0{,}6$. Se isto fizesse parte de uma matriz de correlação submetida a uma análise de componentes principais, poderia ser que X_1 e X_2 surgissem com *loadings* altos em compo-

nentes diferentes (ver Johnston, 1973c). A interpretação substantiva disto seria que duas variáveis são independentes em seus padrões espaciais; isto ocorreria apenas por causa do uso de um sistema linear que é irrelevante para a descrição desta inter-relação.

Tais relações podem ser comuns na Geografia Humana; deve-se perguntar se estudo de análise das "teorias" de padrões residenciais e de estrutura funcional de lugares centrais, por exemplo, revelaria muitas destas estruturas lógicas. Se assim for, então nosso uso de análise de componentes principais estará obscurecendo o esforço para descrevê-las.

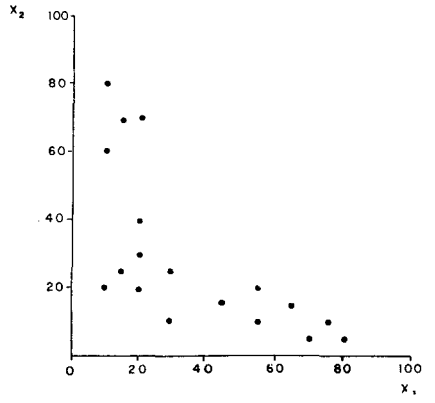

Fig. 7 — *O problema da homoscedascity de dependência não-linear ilustrado pela relação entre percentagem de recebedores de altos salários (X_1) e percentagem nascida no Sul da Europa (X_2).*

5 — INTERPRETAÇÃO DE COMPONENTES E SCORES DAS COMPONENTES

O primeiro objetivo de muitas análises de componentes principais conduzidas por geógrafos humanos tem sido o de substituir um número de variáveis por uma nova variável composta e localizar as observações no *continuum* padronizado deste composto. Assim, muitas ecologias fatoriais resultam em mapas, por exemplo, da variável do *status* sócio-econômico numa cidade, cujos mapas podem ser um fim em si próprios ou podem ser usados como base para futuras investigações.

Combinando-se as observações aos componentes obtem-se os *scores* do componente, derivados pelo produto da matriz de *loadings* L, pela matriz D de dados (observar que D é reescrito primeiro na forma de desvio Z e L é padronizado pelos *eigenvalues* relevantes, de forma que cada vetor de *scores* tem o mesmo desvio-padrão). Estes *scores* são interpretados como os "mapas" das variáveis compositas, mas freqüentemente são encontrados vários problemas na interpretação.

O primeiro destes problemas foi notado por Joshi (1972), que observou que um mapa de *scores* num componente identificado como *status* sócio-econômico não corresponde a seu conhecimento do padrão espacial da cidade estudada. A razão para isto está na identificação do componente por *loadings* pouco significativos, uma falha comum que foi apontada por Palm e Caruso (1972). Tomemos uma análise que envolve doze variáveis, cujas *loadings* num componente particular são:

ESPECIFICAÇÃO	VARIÁVEIS					
	X_1	X_2	X_3	X_4	X_5	X_6
Loadings.....	0,90	0,85	0,80	0,50	0,40	0,39

ESPECIFICAÇÃO	VARIÁVEIS					
	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
Loadings.....	0,38	0,42	0,47	0,39	0,10	0,20

De acordo com a prática comum, este componente provavelmente seria interpretado em termos das variáveis X_1, X_2 e X_3 — como representando o *status* sócio-econômico — mas sete das outras variáveis, que podem ter pouco a ver com o conceito do *status* sócio-

econômico, também têm *loadings* bastante substanciais no componente. Na derivação dos *scores*, através do produto matricial *DL*, estas sete podem vir a ser tão importantes quanto as primeiras três em determinar o *score* para qualquer área — i.e., se ela tiver altos valores em pelo menos alguns dos $X_4 \ldots X_{10}$ e apenas valores moderados nos $X_1 \ldots X_3$. O vetor de *score* está, assim, longe de ser um composto de variáveis X_1 a X_3 apenas, e ou a interpretação do componente ou a derivação do "mapa composto" deverá estar atenta a isto. À medida que os computadores se tornam maiores, e mais e mais variáveis são alimentadas nas análises de componentes, este problema pode se tornar bastante sério.

Um segundo problema refere-se também à interpretação. Suponhamos que uma análise de componentes produz um padrão de *loadings* em três de suas variáveis, como a seguinte:

PADRÃO DE *LOADINGS*	COMPONENTES	
	I	II
X_1 Percentagem masculina ocupada	0,90	0,10
X_2 Percentagem masculina com graduação	0,90	0,05
X_3 Renda média masculina	0,70	0,70

O primeiro componente poderia ser interpretado como *status* sócio-econômico geral, e o segundo indicando que alguma porção da variação na renda não está relacionada a *status* ocupacionais e educacionais. O padrão de *scores* no componente *I* poderia, então, ser interpretado como o mapa de *status* sócio-econômico geral, e no componente *II* como um mapa residual de renda. Mas a última interpretação poderia ser incorreta. No produto matricial *DL* não há divisão da variável de renda na porção que está relacionada à ocupação/educação, e na que não está. Os *scores* no componente *II* forneceriam um mapa de variações de renda, mas *não* das variações residuais de renda de uma regressão de renda em relação à ocupação e educação.

Finalmente, há um problema baseado na "super-interpretação" de um componente. Isto pode ser ilustrado por um exemplo (ver Johnston, 1973c), cujas *loadings* são:

PADRÃO DE *LOADINGS*	COMPONENTES	
	I	II
X_1 Percentagem nascida na Grécia	0,70	0,70
X_2 Percentagem nascida na Itália	0,70	—0,70
X_3 Percentagem nascida na Inglaterra	—0,80	0,00

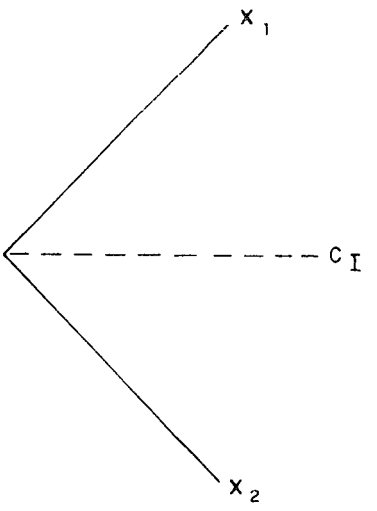

Fig. 8 — *Duas variáveis ortogonais (X_2 e X_1) com loadings positivos de + 0,7 no mesmo componente (C_1).*

Comumente, o primeiro componente poderia ser interpretado como uma dimensão do *status* do migrante; seus *scores* positivos poderiam ser interpretados como áreas residenciais "sul européias", e seus *scores* negativos como o "gueto inglês". Mas, com referência aos *loadings* no componente *II*, as variáveis X_1 e X_2 podem estar muito pouco relacionadas (real-

mente, dois *loadings* de 0,7 poderiam ser baseados numa correlação de $r_{12} = 0{,}0$, conforme indicado pela figura 8, onde ambas X_1 e X_2 têm estes *loadings,* mas são ortogonais entre si). Os *scores* no componente *I* são, então, provavelmente, um composto sem expressão, tirando uma média dos padrões gregos e italianos e não indicando segregação.

Estes problemas de interpretação de *score* de componentes são causados pelo fato da análise dos componentes geográficos raramente alcançar uma estrutura simples, com cada variável tendo um *loading* de *c.* $\pm$ 1,0 e todas as outras de 0,0. Por causa disto, os padrões de *score* de componente podem ter pouco valor.

Apenas certas variáveis poderiam ser mapeadas, conforme sugere Joshi, 1972. Pode ser que outros métodos provem ser melhores. Os geógrafos têm que decidir se querem ordenar observações ao longo de um *continuum,* ou se a classificação de áreas é suficiente, caso este em que, especialmente considerando o problema do conjunto de números fechados, uma técnica desenvolvida por Semple pode se revelar mais adequada (Semple & Scorrar, 1975).

6 — CONCLUSÃO

O fato de a literatura de Geografia nas duas últimas décadas ser desordenada em exemplos do uso, mau uso e abuso de várias técnicas estatísticas, indica um processo em estudo. Quanto às análises de componentes e análises fatoriais, indicam os problemas específicos de muitos conjuntos de dados geográficos, alguns dos quais foram aqui mencionados (para outros, ver Clark, Davies & Johnston, 1974). Apesar da orientação inconstante na Geografia Humana, que vai do empirismo funcional à explicação estrutural (Johnston, 1977), há ainda uma necessidade de descrição correta dos padrões espaciais. As análises de componentes principais e análises fatoriais em muito ajudaram a alcançar tal descrição. O presente trabalho apenas esboçou alguns problemas do uso destes métodos, embora quase sempre os tenha tratado em profundidade. Algumas alternativas foram sugeridas, muitas delas envolvendo um retorno a técnicas mais simples, menos sofisticadas. É claro, no entanto, que será necessária uma maior reflexão e maior experiência nos problemas de descrever sucintamente os padrões complexos.

7 — BIBLIOGRAFIA

BERRY, B. J. L., 1961: Basic patterns of economic development. In Ginsburg. N. (ed.) *Atlas of Economic Development.* University of Chicago Press. Chicago, 110-119.

BERRY, B. J. L. (ed.), 1971: Comparative Factorial Ecology. *Econ. Geogr.,* 47, 3.

CLARK, D., DAVIES, W. K. D. & JOHNSTON, R. J., 1974: The application of factor analysis in human geography. *The Statistician,* 23, 259-281.

DUNCAN, O. D., DUNCAN, B., 1955: Occupational stratification and residential differentation. *Amer. J. Sociol.,* 50, 493-503.

JOHNSTON, R. J., 1971: Some limitations of social area analysis and factorial ecology. *Econ. Geogr.,* 47, 314-323.

JOHNSTON, R. J., 1973a: Possible extensions to the factorial ecology method: a note. *Envir. Plann.* A, 5, 719-734.

JOHNSTON, R. J., 1973b: Residential differentation in major New Zealand urban areas: a comparative factorial ecology. In B. D. Clark and M. B. Gleave (eds.). *Social Patterns in Cities.* Institute of British Geographers. Special Publication 5, 143-168.

JOHNSTON, R. J., 1973c: Social area change in Melbourne 1961-1966: a sample exploration. *Austr. Geogr. Studs.,* 11, 79-98.

JOHNSTON, R. J., 1976: *The World Trade System: Some Inquiries into its Spatial Structure.* G. Bell & Sons, Ltd., London.

JOHNSTON, R. J., 1977: The internal structure of the city. *Progress in Human Geography,* 1.

JOSHI, T. R., 1972: Towards computing factor scores. In W. P. Adams and F. Helleiner (eds.). *International Geography,* 2 University of Toronto Press. Toronto, 906-908.

KLAFF, V. Z., 1973: Ethnic segregation in urban Israel. *Demography,* 10, 161-184.

MATHER, P. M. & OPENSHAW, S., 1974: Multivariate methods and geographic data *The Statistician,* 23, 283-308.

NEWTON, P. W. & JOHNSTON, R. J., 1976: Residential area characteristics and residential area homogeneity: further thoughts on extensions to the factorial ecology method. *Envir. Plann.* A, 543-552.

PALM, R. & CARUSO, D. J., 1972: Labelling in factorial ecology. *Ann. Assoc. Amer. Geogr.,* 62, 122-133.

PAOOLE, M. A. & O'FARRELL, P. N., 1971: The assumptions of the linear regression model. *Trans. Inst. Brit. Geogr.,* 52, 145-158.

RUMMEL, R. J., 1967: Understanding factor analysis. *J. Conflict Resolution,* 40, 440-480.

RUMMEL, R. J., 1970: *Applied Factor Analysis,* Northwestern University Press. Evanston.

SEMPLE, R. K. & SCORRAR, D. A., 1975: Canadian international trade. *Can. Geogr.* 19, 135-148.

TIMMS, D. W. G., 1970: Modernisation and the factorial ecology of the Cook Islands, Brisbane and Auckland. *Austr. N. Z. J. Sociol.,* 6, 139-149.

TIMMS, D. W. G., 1971: *The Urban Mosaic,* Cambridge University Press, Cambridge.