Análise de Trajetória

RODOLPHO A. SIMAS

INTRODUÇÃO

Este artigo procura apresentar, de maneira operacional, o método denominado *Análise de Trajetória*. Deve-se ao geneticista Sewall Wright os seus fundamentos. Desde a apresentação dos primeiros trabalhos em 1918, este método vem sendo bastante debatido até os dias de hoje por outros pesquisadores, o que muito contribuiu para o seu aperfeiçoamento. Atualmente a análise de trajetória é aplicada nos mais diversos ramos do conhecimento humano. Dentro do contexto geográfico este método se apresenta como um instrumento poderoso para uma abordagem realista de sistemas multivariados, onde se admite a existência de causação. No entanto, como afirmou o próprio Wright, a análise de trajetória *"não está de forma alguma restrita às relações que possam ser descritas como sendo de causa e efeito. Pode ser aplicado a sistemas lineares puramente matemáticos e funde-se com o método de regressão múltipla"*.

ANÁLISE DE TRAJETÓRIA

A análise de trajetória é um método de decomposição e interpretação das relações lineares, aditivas e unidirecionais em um grupo de variáveis que se admite serem mensuráveis em uma escala de intervalo, embora algumas delas, na realidade, não possam ser medidas ou possam ser

puramente hipotéticas — por exemplo, os *fatores* na análise fatorial. Supõe-se ainda que:

1) exista uma ORDEM CAUSAL (FRACA) entre as variáveis;
2) as relações entre estas variáveis sejam de CAUSAS FECHADAS.

1 — PRINCÍPIOS DA ANÁLISE DE TRAJETÓRIA

1.1 — Definição de Efeito Causal

Propomos a seguinte definição *operacional* com uma aproximação inicial da idéia de causação:

X_1 É CAUSA DE X_0 SE E SÓ SE X_0 POSSA SER MUDADA PELA MANIPULAÇÃO DE X_I E SOMENTE X_1.

Notamos, primeiramente, que a noção de causação implica em predição, mas de um tipo particular, pois aquelas puramente estatísticas ou matemáticas, que não impliquem em noção de geração de mudança estão excluídas. Em segundo, para compreender o que se entende por *somente* na definição deve-se compreender a noção de hierarquia causal e a noção de controle. Por enquanto, devemos observar que a manipulação de X_1 sozinha não implica que todas as outras causas de X_0 estejam controladas ou permaneçam constantes. Se mudarmos ou manipularmos X_I sozinha provocaremos alterações em muitas outras variáveis que são afetadas por X_1. Modificações induzidas por X_1 podem afetar X_0. Estas mudanças induzidas em outras variáveis não podem estar controladas ou constantes enquanto examinamos o efeito de X_1 em X_0.

A definição anterior de causa sugere o critério de causação e o meio de medir o efeito causal. Primeiro, para estabelecer conclusivamente que X_I é causa de X_0 deve-se executar um experimento *ideal* no qual todas as outras variáveis relevantes são mantidas constantes enquanto manipula-se a variável causal. Segundo, devem existir mudanças em paralelo na variável X_0. Usaremos tal evidência como critério final de que X_1 é causa de X_0.

Em um experimento ideal a relação entre mudanças manipuladas em X_1 (que serão indicadas por x_I) e as mudanças que decorrem em X_0 (que serão indicadas por x_0) deve ser uma função linear da forma

$$x_0 = c_{01}\, x_1$$

onde c_{01} é uma constante que representa as mudanças em x_0 para uma unidade de mudança em x_1. Note que as letras minúsculas são usadas para indicar que não estamos falando sobre a relação entre X_I e X_0 em seus estados naturais. Observe-se que a afirmativa precedente é verdadeira sob o postulado do determinismo causal. O coeficiente c_{01} assim medido será denominado COEFICIENTE LINEAR DO EFEITO CAUSAL ou simplesmente COEFICIENTE DO EFEITO.

Dada uma regressão de Y em X, por exemplo

$$Y = a + bX$$

o coeficiente b da regressão não pode ser interpretado como coeficiente do efeito. Mede simplesmente a diferença esperada entre dois grupos que são diferentes em X por uma unidade. No entanto, se são encontradas as suposições de ordem causal fraca e fechamento causal, então

os dois coeficientes (b_{yx} e c_{yx}) são equivalentes. Se os coeficientes de uma regressão são interpretados como coeficientes do efeito pela admissão explícita daquelas suposições, em geral por meio de um diagrama de trajetória, então está se efetuando uma interpretação analítica das trajetórias, ou mais simplesmente uma ANÁLISE DE TRAJETÓRIA.

1.2 — Ordem causal

A primeira suposição geral exigida em uma análise de trajetória é a de uma ordenação causal fraca nas variáveis. Dado um par de variáveis X_i e X_j, supõe-se ou sabe-se que X_i pode (ou não) afetar X_j, mas X_j NÃO PODE AFETAR X_i, diz-se, então, que há uma ORDEM CAUSAL FRACA $X_i \geqslant X_j$.[1] Embora a ordem causal não seja sempre inequívoca, a suposição de ordem fraca é sustentável em uma ampla variedade de situações de investigação.

1.3 — Fechamento causal

A segunda suposição geral exige o FECHAMENTO CAUSAL para a análise de trajetória. Dada uma covariação bivariada entre, digamos X e Y, e uma ordenação causal conhecida, digamos $X \geqslant Y$, a covariação observada entre X e Y pode ser devido à:

1) dependência causal de Y em X, unicamente;
2) sua dependência causal em algumas variáveis externas, ou
3) combinação de (1) e (2).

Um exemplo simplificado das possíveis estruturas causais subjacentes a uma covariação é apresentado na figura 1.

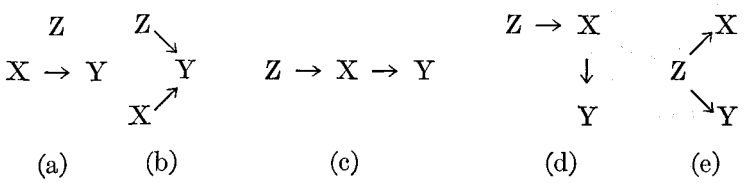


Figura 1 — TIPOS DE ESTRUTURAS CAUSAIS

Em (a), Z não está conectada com X ou Y; em (b), Z é uma causa de Y mas não de X; em (c), Z é uma causa de X e Y, mas o efeito de Z em Y está totalmente contido em X ou por intermédio de X. Nos diagramas (a), (b) e (c), os subsistemas bivariados em X e Y são todos, EM RELAÇÃO A SUA COVARIAÇÃO, DE CAUSAÇÃO FECHADA À INFLUÊNCIA EXTERNA. Em (d), a covariação entre X e Y é parcialmente devido à dependência causal de Y em X e em parte devido a sua repartição direta de uma causa comum, no caso Z. Em (e), a covariação entre X e Y é totalmente devido a sua dependência comum e direta de uma causa Z externa.

---

1 Usaremos o símbolo $\geq$ para designar uma ordem causal fraca, mas não deve ser interpretado como significando *"maior do que ou igual a"*.

2 — SISTEMAS RECURSIVOS

A noção de equações recursivas é fundamental na análise de trajetória. A idéia básica é que as variáveis podem ser arranjadas hierarquicamente em termos de suas prioridades causais de tal forma que se torne possível desprezar-se variáveis que sejam claramente dependentes de um dado subconjunto de variáveis. Por exemplo, suponha que estejamos considerando quatro variáveis X_1, X_2, X_3 e X_4. Se admitirmos existir uma ordenação causal fraca $X_1 \geqslant X_2 \geqslant X_3 \geqslant X_4$, neste caso X_4 não afeta X_1, X_2 e X_3 então independente da influência que qualquer uma destas variáveis possa ter em X_4, é justificável ignorar-se X_4 quando considerarmos as interrelações entre as três primeiras variáveis. Da mesma forma X_3 não influencia X_1 e X_2, e é justificável ignorar-se X_3 no estudo de suas interrelações. Finalmente, como X_2 não afeta X_1, então podemos escrever o seguinte sistema de equações estruturais

$$
\begin{aligned}
&(X_1 \text{ exógena}) \\
X_2 &= b_{21}\, X_1 + e_u \\
X_3 &= b_{31}\, X_1 + b_{32}\, X_2 + e_v \qquad \text{(I)} \\
X_4 &= b_{41}\, X_1 + b_{42}\, X_2 + b_{43}\, X_3 + e_w
\end{aligned}
$$

Aqui nos livramos dos termos constantes pela suposição de que cada variável esteja sendo medida em termos dos desvios de sua média, e *e* representa o termo de ERRO que acumula todos os efeitos de todas as fontes de variação omitidas. É claro que esta idéia básica pode ser estendida a qualquer número de variáveis, até onde desejarmos supor tal ordenação causal.

Pode-se mostrar que para as equações do sistema podem ser obtidas estimativas não tendenciosas dos coeficientes das regressões, supondo-se que o termo de erro em cada equação seja não correlacionado com as variáveis predeterminadas daquela equação e em todas as equações anteriores. A variável estritamente exógena X_1 é predeterminada em cada equação. Além disso, X_2 é uma variável predeterminada na equação — X_3, enquanto que X_3 é uma variável predeterminada na equação — X_4. Assim e_u é não correlacionado com X_1; e_v é correlacionado com X_1 e X_2; e e_w é não correlacionado com X_1, X_2 e X_3. Em conseqüência disso os erros são mutuamente não correlacionados.

O que significam tais suposições em termos do comportamento das variáveis externas não contidas explicitamente no modelo? Se se supõe que as variáveis externas tenham um efeito direto em SOMENTE UMA das variáveis explícitas, então podem ser encontradas as suposições acima. Observe que uma variável implícita poderia ter um efeito INDIRETO em alguma variável através de uma das X_i restantes sem violar as suposições. Mas se um fator implícito afeta DIRETAMENTE duas ou mais variáveis explícitas, então este fator normalmente será correlacionado com uma das variáveis independentes em sua equação, e não serão encontradas as suposições. Se este for o caso, as estimativas de mínimos quadrados serão tendenciosas. Para superar este problema tal variável deve ser incluída explicitamente no sistema. Finalmente, em algum ponto o investigador parará e fará a suposição simples de que todos os fatores implícitos restantes operam em uma única variável explícita.

3 — DIAGRAMA DE TRAJETÓRIA

Embora não seja intrínseca a análise de trajetória, a representação diagramática do sistema de equações recursivas é de grande auxílio na reflexão sobre suas propriedades. Este sistema consiste de q FATORES ou CAUSAS PRIMÁRIAS e p EFEITOS RESULTANTES, supõe-se que estas p + q variáveis estejam mutuamente associadas por uma rede de trajetórias causais. O diagrama representa esta rede pelo dispositivo devido a Wright, que descreve as trajetórias causais por meio de uma seta de uma ponta, que une a causa (cauda) ao efeito correspondente (ponta). Neste diagrama toda variável incluída, medida ou hipotética, ou é representada (por setas) como COMPLETAMENTE determinada por outras ou como um fator primário.

O necessário fechamento formal do diagrama exige a introdução de um símbolo para o conjunto de fatores residuais desconhecidos que afetam cada variável (que não represente um dos fatores primários), a menos que se possa supor com segurança que haja determinação completa pelos fatores conhecidos.

A característica dos fatores primários como *causas primárias* ou *variáveis não-resposta* é representada pela restrição de que UMA SETA DE UMA PONTA NUNCA PODE APONTAR PARA UM FATOR PRIMÁRIO. Não existe nenhuma outra restrição quanto à posição da seta, mesmo duas setas apontando em direções opostas são permitidas e pode ser dada uma interpretação, desde que uma das variáveis não seja um fator primário, mas tal tipo de trajetória não será tratado aqui.

As representações para as trajetórias causais apresentadas acima são ilustradas na estrutura abaixo:

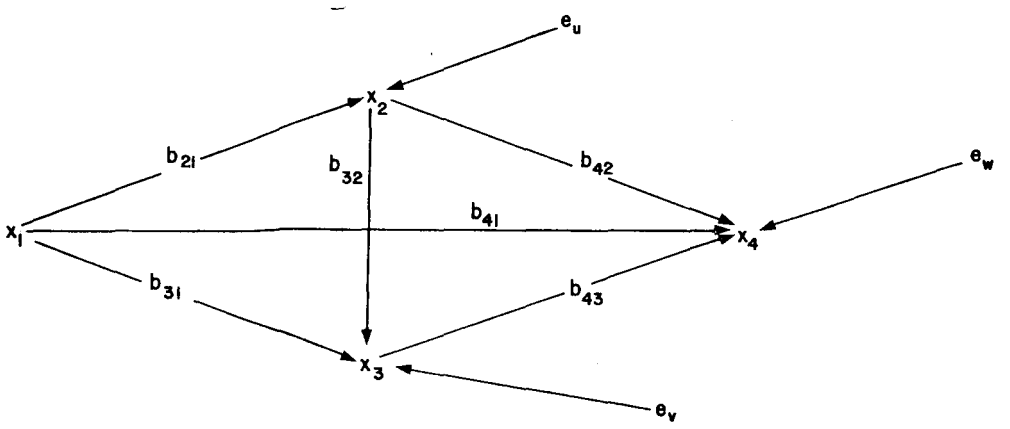


FIGURA 2

O diagrama de trajetória pode ser interpretado em termos das equações estruturais. INTERPRETA-SE A VARIÁVEL NA PONTA DE UMA OU MAIS SETAS, COMO SENDO UMA FUNÇÃO DAQUELAS VARIÁVEIS NAS CAUDAS DAS SETAS QUE A ELA CONVERGEM. Assim, o diagrama da figura 2 é a representação do sistema de equações estruturais (I).

Trataremos aqui somente das funções lineares, as relações não lineares podem algumas vezes ser sistematicamente transformadas, por todo o diagrama, em relações lineares. Onde forem pequenos, dentro do in-

tervalo de variação real, os desvios da linearidade, pode-se obter resultados aproximados sem transformação. Além disso, a regressão linear sempre pode ser interpretada como a melhor aproximação linear para a relação, quando esta última for não linear.

4 — O COEFICIENTE DE TRAJETÓRIA

Consideremos o seguinte diagrama *

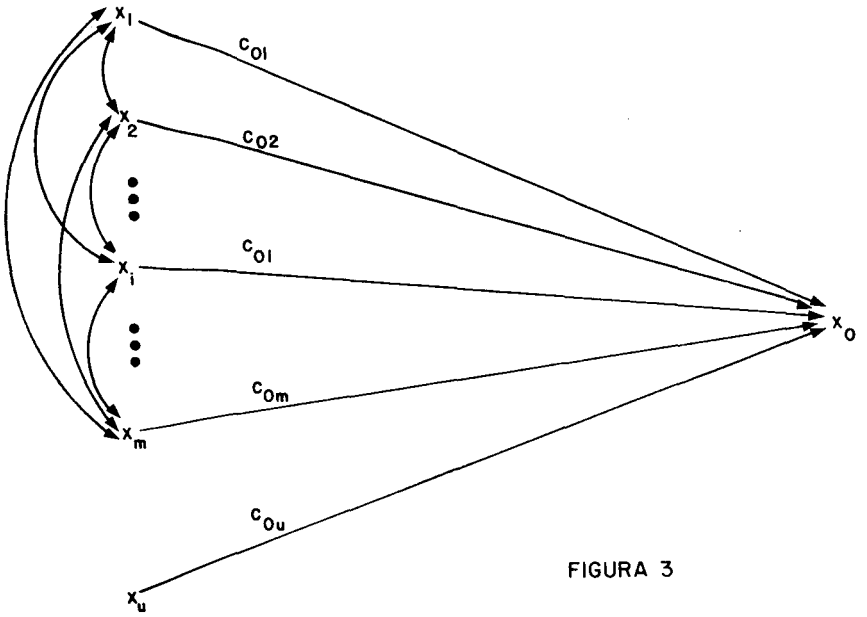


FIGURA 3

onde a variável X_0 é representada como completamente determinada pelos fatores imediatos X_1, X_2, ..., X_m, X_u, que são representados como correlacionados, com exceção da variável residual X_u. Admite-se serem lineares todas as relações, então

$$X_0 = c_0 + c_{01} X_1 + c_{02} X_2 + \ldots + c_{0i} X_i + \ldots + c_{0m} X_m + c_{0u} X_u$$

Os coeficientes c_{0i} são do tipo dos coeficientes parciais de regressão, mas estão em um sistema que envolve X_u (a menos que se saiba ser completa a determinação pelos outros fatores imediatos). Assim, outras variáveis hipotéticas não mensuráveis estão envolvidas.

O COEFICIENTE DA REGRESSÃO DE TRAJETÓRIA C_{0i} mede a contribuição em termos absolutos que se supõe que X_i faça DIRETAMENTE em X_0, do ponto de vista representado no diagrama. Como os coeficientes de variáveis não mensuráveis só podem ser tratados na forma padronizada, a homogeneidade exige que todos assim sejam tratados.

* Para maior compreensão sobre este tipo de trajetória curva, veja a seção 6.

Agora, padronizando-se as variáveis do sistema, i. é, fazendo-se

$$x_i = \frac{X_i - \overline{X}_i}{\sigma_i} \qquad i = 0, 1, 2, \ldots, m$$

obtém-se o COEFICIENTE DE TRAJETÓRIA

$$p_{0i} = \frac{\sigma_i}{\sigma_0} c_{0i} \qquad i = 1, \ldots, m$$

observe que tal coeficiente é adimensional, portanto as variáveis mensuráveis e não mensuráveis podem ser tratadas na mesma base. O coeficiente de trajetória mede o efeito direto de uma variável sobre a outra, i. é, modificando-se de um desvio-padrão a variável x_i, a variável x_0 sofre uma mudança esperada em p_{0i}. Estes coeficientes são determinados pela regressão múltipla,

$$x_0 = p_{0i}\, x_1 + p_{02}\, x_2 + \ldots + p_{0m}\, x_m$$

sendo o coeficiente de trajetória residual obtida pela expressão

$$p_{0u} = \sqrt{1 - R^2}$$

onde R^2 é o quadrado do coeficiente de correlação múltipla.

5 — TEOREMA FUNDAMENTAL DA ANÁLISE DE TRAJETÓRIA

O princípio que se segue das equações do sistema recursivo (I) é que a correlação entre um par qualquer de variáveis pode ser escrita em termos de trajetórias a partir das variáveis antecedentes.

Podemos escrever o coeficiente de correlação de Pearson entre X_i e X_j, por

$$r_{ij} = \frac{1}{n} \sum_{k=1}^{n} x_{ik}\, x_{jk}$$

onde n representa o número de observações.

A partir desta fórmula chega-se a EXPRESSÃO GERAL DO TEOREMA BÁSICO da análise de trajetória

$$r_{ij} = \sum_{q} p_{iq}\, r_{jq} \qquad (1)$$

onde i e j indicam duas variáveis na forma padronizada de um sistema recursivo e o índice q percorre todas as variáveis das quais as trajetórias levam diretamente a X_i. Assim, pela aplicação sucessiva da fórmula em r_{jq} a correlação r_{ij} pode ser expandida até onde for possível a decomposição de r_{jq}.

5.1 — Regra de Decomposição da Correlação

A decomposição da correlação r_{ij}, obtida pela aplicação sucessiva da expressão geral do teorema básico, pode ser lida diretamente do diagrama seguindo a regra devida a Wright:

"Leia para trás a partir da variável i, daí para frente até a variável j, formando o produto de todas as trajetórias ao longo do percurso, depois some os produtos correspondentes a todos percursos possíveis. A mesma variável não pode ser interceptada mais de uma vez em um único percurso. Em nenhum momento é permitido voltar para trás após ter-se começado a ir para frente. A correlação bidirecional é usada

para seguir para frente ou para trás, mas se houver mais de uma correlação bidirecional no diagrama, somente uma poderá ser usada em um único percurso."

5.2 — Decomposição das Covariações na Estrutura de Trajetória do Modelo Geral

À medida que mais variáveis são incorporadas ao modelo, diminui a proporção das relações que são decompostas puramente em termos das suposições básicas e aumenta a proporção das relações que são testadas parcialmente em relação ao fechamento causal e são examinadas em relação aos processos causais. Façamos, portanto, a decomposição das covariações bivariadas da estrutura causal correspondente a um modelo geral onde consideramos as variáveis como padronizadas, utilizando para isso o Teorema Fundamental.

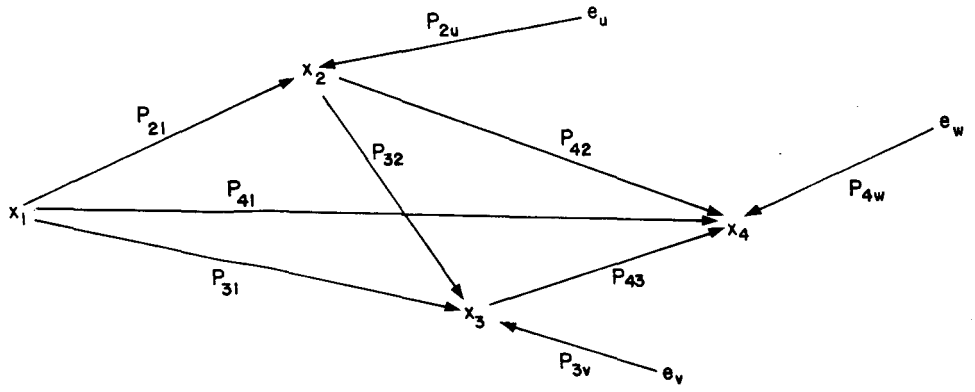


FIGURA 4

As equações do modelo são

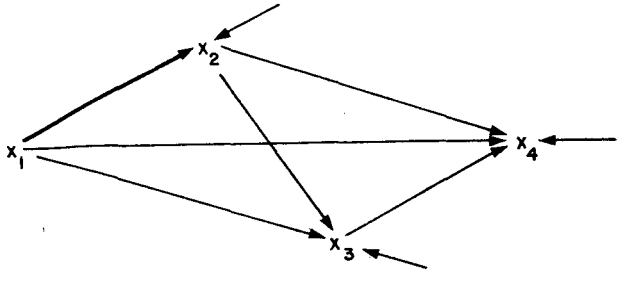
(x_1 exógena)

$$x_2 = p_{21}\, x_1 + p_{2u}\, e_u$$

$$x_3 = p_{32}\, x_2 + p_{31}\, x_1 + p_{3v}\, e_v$$

$$x_4 = p_{43}\, x_3 + p_{42}\, x_2 + p_{41}\, x_1 + p_{4w}\, e_w$$

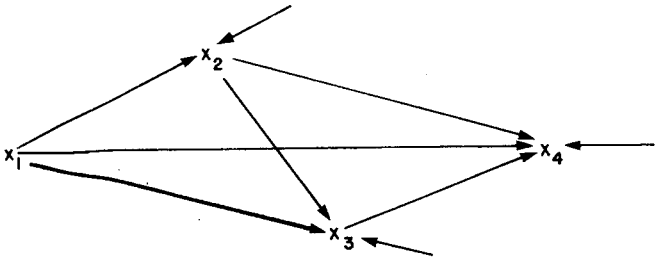
1) Vemos que a correlação total entre x_1 e x_2 é gerada pelo efeito direto, p_{21}



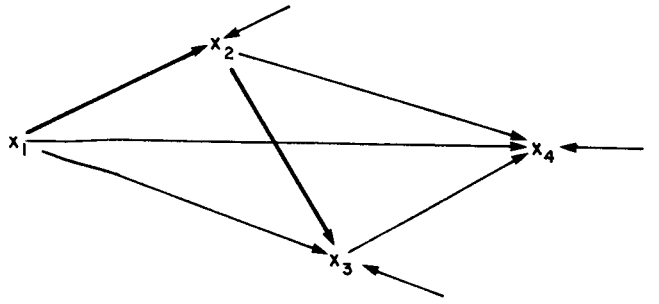
então,

$$p_{21} = p_{21}$$

2) A correlação entre x_1 e x_3 é gerada por duas trajetórias distintas, de tal forma que p_{31} é igual ao efeito direto, p_{31}



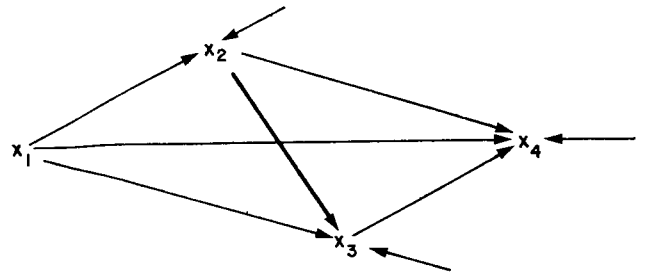
MAIS o efeito indireto, $p_{32}\ p_{21}$



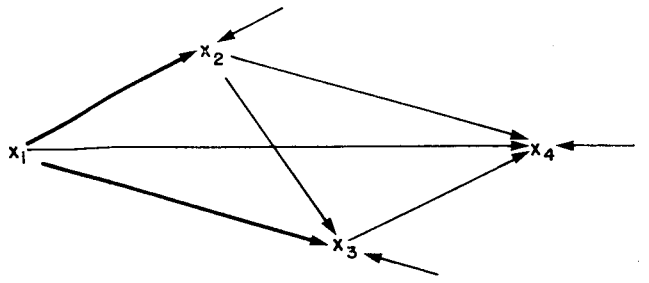
portanto,

$$p_{31} = p_{31} + p_{32}\ p_{21}$$

3) A situação é diferente com relação a x_2 e x_3, aqui temos a correlação total (p_{32}) gerada pela soma do efeito direto, p_{32}



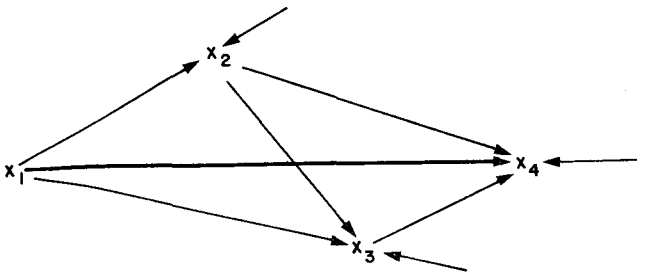
MAIS a correlação devido a uma causa comum (x_1), $p_{31}\ p_{21}$



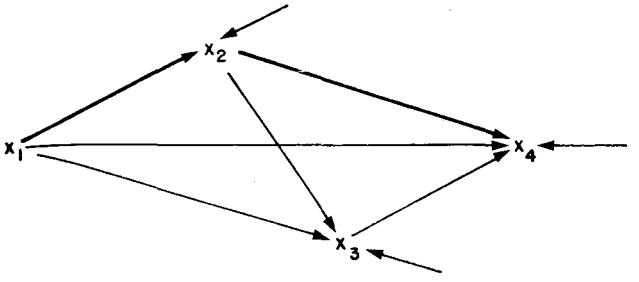
portanto,

$$p_{32} = p_{32} + p_{31}\ p_{21}$$

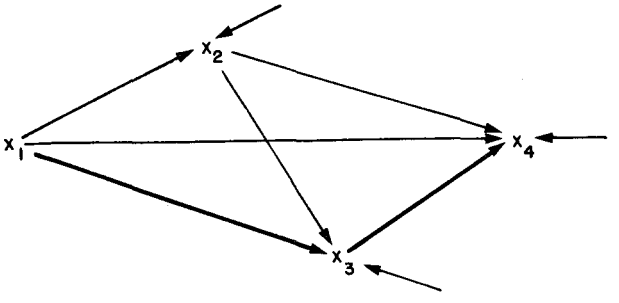
4) A correlação entre x_1 e x_4 é gerada por quatro conexões causais distintas; p_{41} é igual ao efeito direto, p_{41}



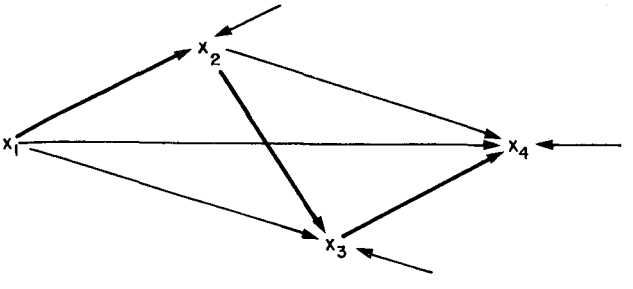
MAIS o efeito indireto via x_2, $p_{42}\ p_{21}$



MAIS o efeito indireto via x_3, $p_{43}\ p_{31}$



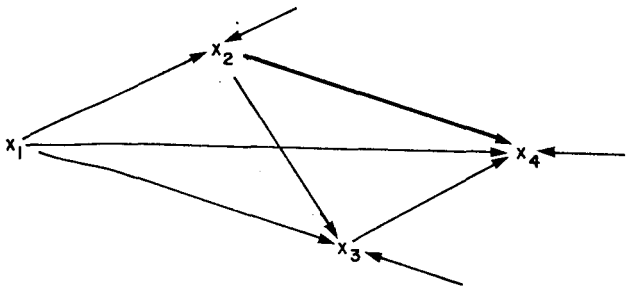
MAIS o efeito indireto via x_3 e x_2, $p_{43}\ p_{32}\ p_{21}$



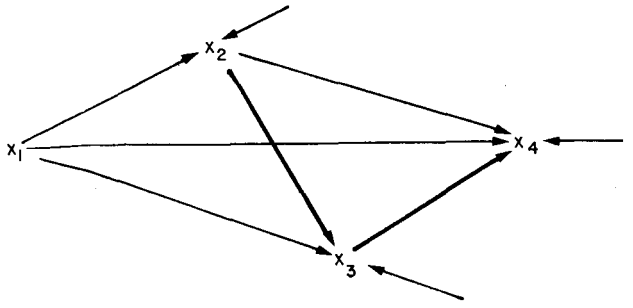
assim,

$$p_{41} = p_{41} + p_{42}\ p_{21} + p_{43}\ p_{31} + p_{43}\ p_{32}\ p_{21}$$

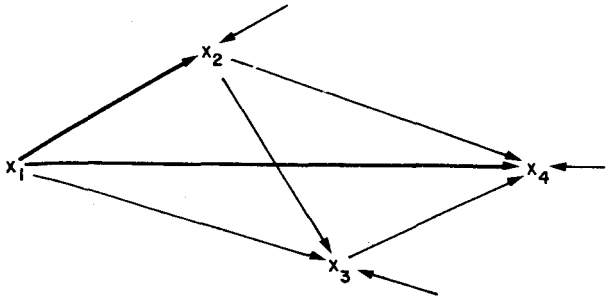
5) Na geração da correlação entre x_2 e x_4 participam um efeito indireto e uma correlação devido a causas comuns; p_{42} é igual ao efeito direto, p_{42}



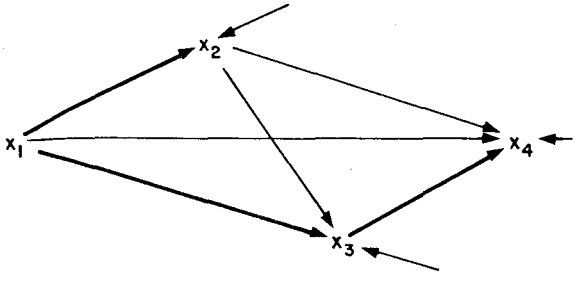
MAIS o efeito indireto via x_3, $p_{43}\ p_{32}$



MAIS a correlação devido a x_1 operando como uma causa comum, diretamente, $p_{41}\ p_{21}$



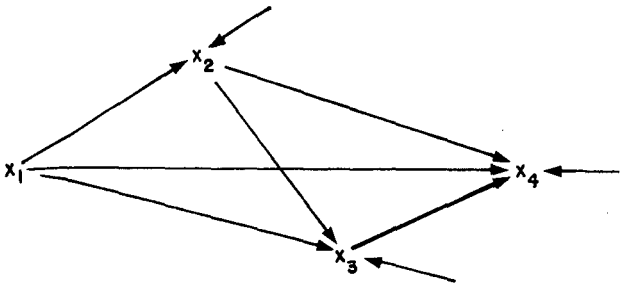
e indiretamente (via x_3), $p_{43\ 31}\ p_{21}$



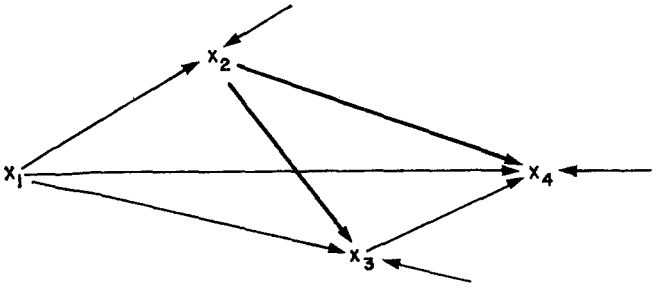
então,

$$P_{42} = p_{42} + p_{43}\ p_{32} + p_{41}\ p_{21} + p_{43}\ p_{31}\ p_{21}$$

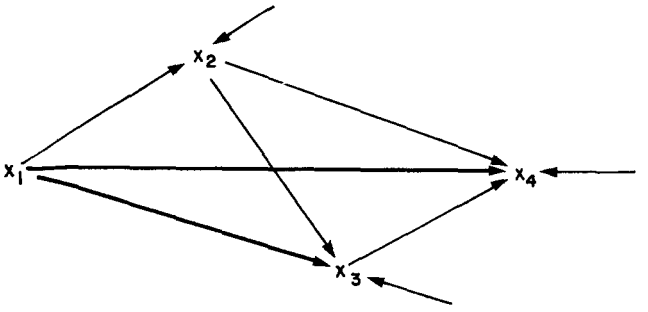
6) Não existem no modelo efeitos indiretos produzindo correlação entre x_3 e x_4; mas existem duas causas comuns. Então, p_{43} é igual ao efeito direto, p_{43}



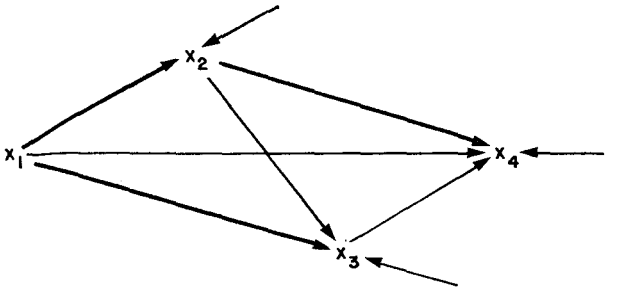
MAIS a correlação devido a causas comuns, atuando diretamente, $p_{42}\ p_{32}$



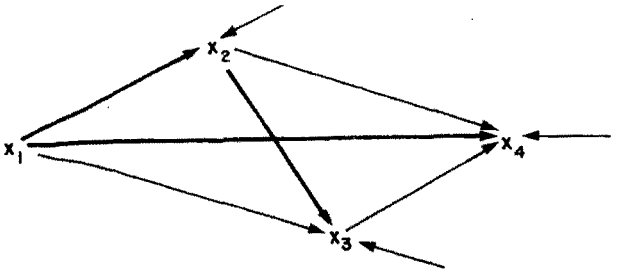
e $p_{41}\ p_{31}$



ou indiretamente, $p_{42}\ p_{31}\ p_{21}$



e $p_{41}\ p_{32}\ p_{21}$



portanto,

$$p_{43} = p_{43} + p_{42}\, p_{32} + p_{41}\, p_{31} + p_{42}\, p_{32}\, p_{21} + p_{41}\, p_{32}\, p_{21}$$

Tais resultados podem ser colocados na tabela que se segue.

TABELA 1

Decomposição da covariação bivariada

RELAÇÃO BIVARIADA X_i X_j	COVARIAÇÃO TOTAL ρ_{ij}	CAUSA				EFEITO TOTAL c_{ij} = $= p_{ij} + I$	NÃO CAUSAL $\rho_{ij} - c_{ij} =$ CD + CI
		Direta p_{ij}	Indireta (I)	Comum			
				Direta (CD)	Indireta (CI)		
X_2 X_1	ρ_{21}	ρ_{21}	—	—	—	ρ_{21}	—
X_3 X_1	ρ_{31}	ρ_{31}	$p_{32}\, p_{21}$	—	—	p_{31}	—
X_3 X_2	ρ_{32}	p_{32}	—	$p_{31}\, p_{21}$	—	p_{32}	$\rho_{32} - p_{32}$
X_4 X_1	ρ_{41}	p_{41}	$p_{42}\, p_{21} + p_{43}\, p_{31} + p_{43}\, p_{32}\, p_{21}$	—	—	ρ_{41}	
X_4 X_2	ρ_{42}	p_{42}	$p_{43}\, p_{32}$	$p_{41}\, p_{21}$	$p_{43}\, p_{31}\, p_{21}$	$p_{42} + p_{43}\, \rho_{32}$	$p_{42} - (p_{42} + p_{43}\, p_{32})$
X_4 X_3	ρ_{43}	p_{43}	—	$p_{42}\, p_{32} + p_{41}\, p_{31}$	$p_{42}\, p_{31}\, p_{21} + p_{41}\, p_{32}\, p_{21}$	p_{43}	$p_{43} - p_{43}$

A coluna EFEITO TOTAL dá o coeficiente do efeito causal total, quer dizer, mudando-se em 1 σ_j variável X_j então a variável X_i mudará em c_{ij}.

Uma inspeção de cada linha da tabela 1 indica que a relação (X_2 X_1) é a única para a qual a análise de trajetória não fornece nenhuma informação além da contida na correlação bivariada e nas suposições iniciais do modelo geral. Observe-se que a soma das causas comuns dá a parte não causal ou espúria da covariação, e a última coluna dá uma forma alternativa para o cálculo desta parte da correlação.

5.3 — Decomposição da Variação da Variável Dependente

Um caso importante da (1) é considerado quando (i = j) e é conhecido como FÓRMULA PARA A DETERMINAÇÃO COMPLETA DE X_i.

$$r_{ii} = 1 = \sum_{q} p_{iq}^2 + 2 \sum_{q} \sum_{q'} p_{iq}\, r_{qq'}\, p_{iq'}$$

onde q e q' (q' > q) incluem todas as variáveis envolvidas, sejam ou não residuais. Por exemplo, a aplicação desta fórmula ao diagrama da figura 2 resultaria em

$$1 = \sum_{j=1}^{m} p_{0j}^2 + 2 \sum_{j=1}^{m} \sum_{k=j+1}^{m} p_{0j}\, p_{0k}\, r_{jk} + p_{0u}^2$$

agora, multiplicando-se esta expressão pela variância de X_0, σ_0^2, temos

$$\sigma_0^2 = \sigma_0^2 \left[\sum_{j=1}^{m} p_{0j}^2 \right] + \sigma_0^2 \left[2 \sum_{j=1}^{m} \sum_{k=j+1}^{m} p_{0j}\, p_{0k}\, r_{jk} \right] + \sigma_0^2 \left[p_{0u}^2 \right] \qquad (2)$$

Assim, vemos que a fórmula para a determinação completa de uma variável dependente (X_0) decompõe sua variação em três partes

$$\text{VARIAÇÃO TOTAL DE } X_0 = \left[\begin{array}{l}\text{PROPORÇÃO DETERMINADA} \\ \text{DIRETAMENTE PELOS FATO-} \\ \text{RES IMEDIATOS}\end{array}\right] +$$

$$+ \left[\begin{array}{l}\text{PROPORÇÃO DEVIDO ÀS INTERCORRELAÇÕES ENTRE} \\ \text{AS VARIÁVEIS INDEPENDENTES}\end{array}\right] +$$

$$+ \text{ [PROPORÇÃO DETERMINADA PELO RESÍDUO } X_u\text{]}$$

5.4 — Decomposição da Variação das Variáveis dependentes do Modelo Geral

A tabela 2 ilustra o desmembramento da variação total das variáveis dependentes pela aplicação da expressão (2).

TABELA 2

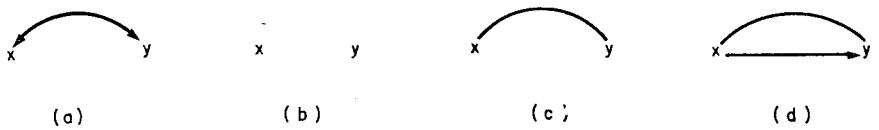
Decomposição da variação da variável dependente

VARIÁVEL DEPENDENTE	VARIAÇÃO		
	Determinada Diretamente pelos Fatores Imediatos	Determinada pelas Intercorrelações entre as Variáveis Independentes	Determinada pelo Resíduo
X_2	p_{21}^2	—	p_{2u}^2
X_3	$p_{31}^2 + p_{32}^2$	$2\, p_{31}\, p_{32}\, p_{12}$	p_{3u}^2
X_4	$p_{41}^2 + p_{42}^2 + p_{43}^2$	$2 [p_{41}\, p_{42}\, \rho_{12} + p_{41}\, p_{43}\, \rho_{13} + p_{42}\, p_{43}\, \rho_{23}]$	p_{4w}^2

6 — INTERPRETAÇÃO DOS RESULTADOS SOB SUPOSIÇÕES ESPECÍFICAS E AMBÍGUAS

Até agora temos considerado somente modelos gerais de trajetória em que todas as relações bivariadas são supostas estarem sujeitas a uma ordem causal fraca e o sistema como um todo tem um fechamento causal. Contudo, inúmeras suposições diferentes podem ser incorporadas em um dado modelo de trajetória. A natureza de cada suposição tem implicações importantes na estimação dos coeficientes de trajetória e na identificação dos coeficientes de efeito.

A análise de trajetória não é um procedimento para a demonstração da existência de causação. Trata-se de um método de extração das conseqüências de um conjunto de suposições causais que o pesquisador deseja impor ao sistema de relações. Como iremos demonstrar, a incorporação de suposições ambíguas ao modelo leva a ambigüidades na interpretação dos resultados. Para a discussão das implicações de se acrescentar outras suposições ao modelo geral, apresentamos nas figuras (a), (b), (c) e (d) várias convenções para representar revelações bivariadas.



(a) (b) (c) (d)

A curva de duas pontas em (a) representa uma correlação não analisada. Neste caso, a trajetória entre X e Y permanece ambígua, a covariação pode ser causal ou espúria e a direção da causação pode ser de X para Y ou vice-versa. A ausência de seta ou curva em (b) significa a não covariação entre X e Y.

A curva simples em (c) representa uma covariação não causal ou completamente espúria entre X e Y. Finalmente, a curva e a seta em (d) representam uma relação bivariada que é em parte causal e em parte espúria.

Consideremos em primeiro lugar a suposição do tipo (a). Na figura 4, ao invés de se admitir uma ordem causal (fraca) entre X_1 e X_2, suponhamos que X_1 e X_2 sejam exógenas, i.é, a relação entre estas variáveis é considerada como dada e que se admite como desconhecida a verdadeira relação causal. Um modelo geral de análise de trajetória como tal suposição é em geral discrito como o da figura 5.

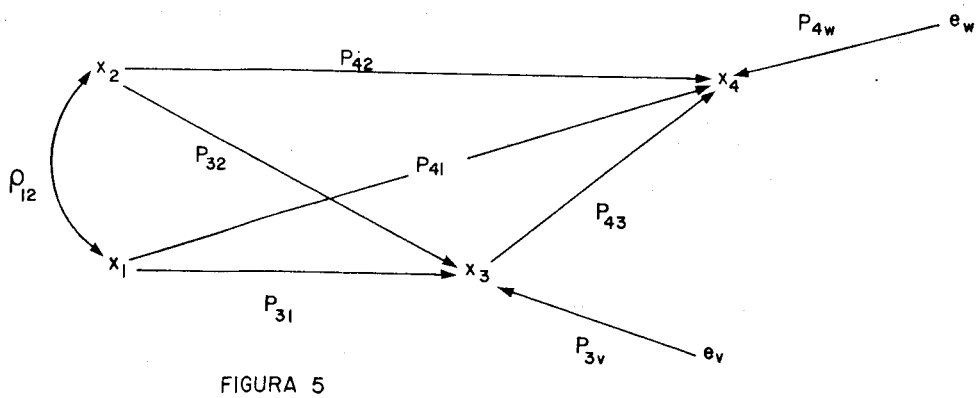


FIGURA 5

As equações do modelo são

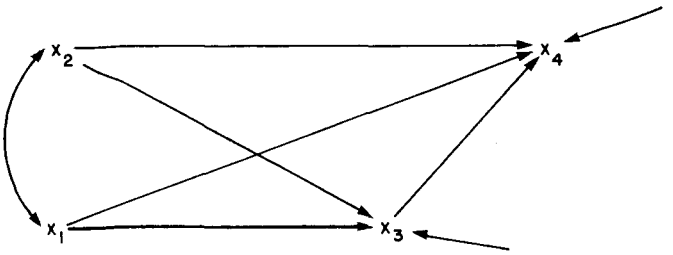
(x_I e x_2 exógenas)

$$x_3 = p_{32}\, x_2 + p_{31}\, x_1 + p_{3v}\, e_v$$

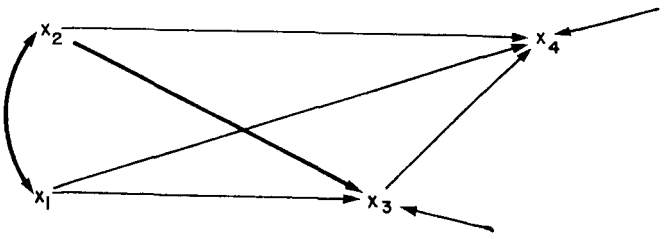
$$x_4 = p_{43}\, x_3 + p_{42}\, x_2 + p_{4I}\, x_1 + p_{4w}\, e_w$$

A decomposição das correlações entre as variáveis do modelo representado pela figura 5 é feita a seguir:

1) Considere a correlação entre x_1 e x_3. A correlação ρ_{31} é gerada pelo efeito direto, p_{31}



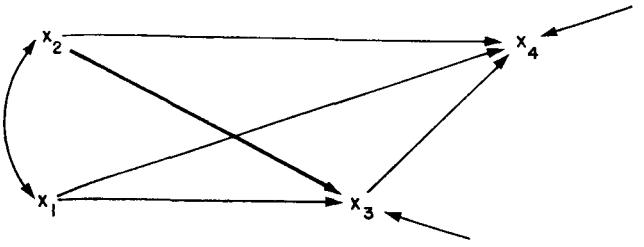
MAIS o termo $p_{32}\ p_{12}$, que consiste no produto do efeito direto de x_2 em x_3 e a correlação entre x_1 e x_2. Este produto representa uma contribuição para ρ_{31} devido ao fato de que OUTRA CAUSA de x_3 (no caso x_2) está correlacionada (da ordem de p_{12}) com a causa que estamos examinando no momento (ou seja, x_1)



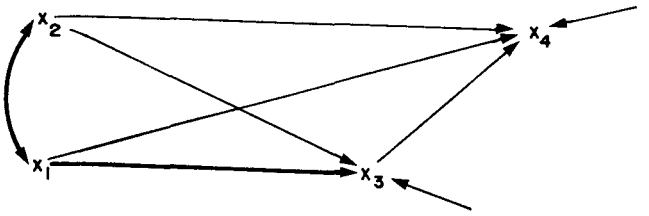
portanto,

$$\rho_{31} = p_{31} + p_{32}\ \rho_{12}$$

2) Da mesma maneira, particionamos ρ_{32} nas componentes: efeito direto p_{32}



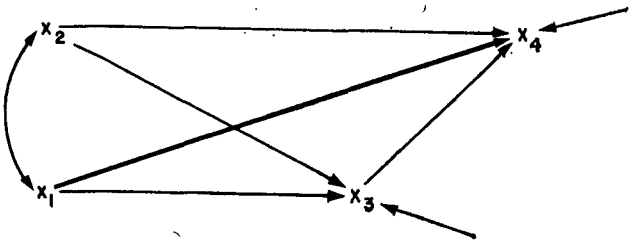
MAIS a correlação devido à correlação com outra causa (x_2), $p_{31}\ \rho_{12}$



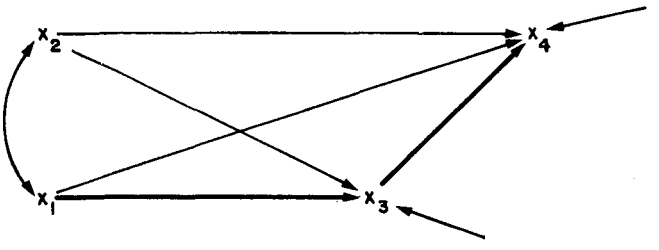
portanto,

$$\rho_{32} = \rho_{32} + p_{31}\ \rho_{12}$$

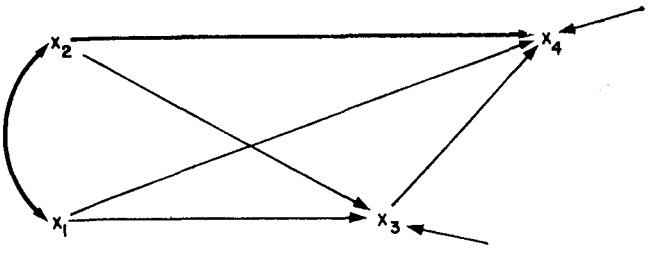
3) Para ρ_{41}, temos o efeito direto, p_{41}



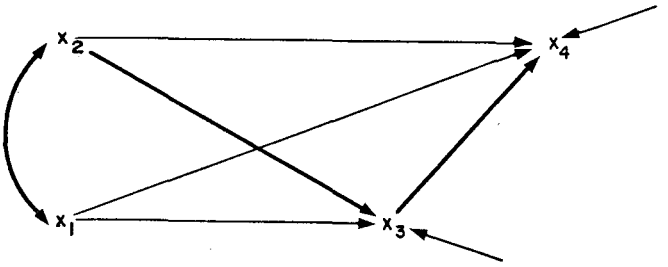
MAIS o efeito indireto, $p_{43}\ p_{31}$



MAIS a correlação devido à correlação de x_1 com outra causa (x_2), ambas operando diretamente, $p_{42}\ \rho_{12}$



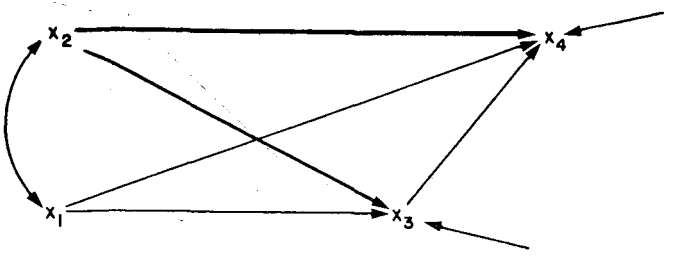
e indiretamente, $p_{43}\ p_{32}\ \rho_{12}$



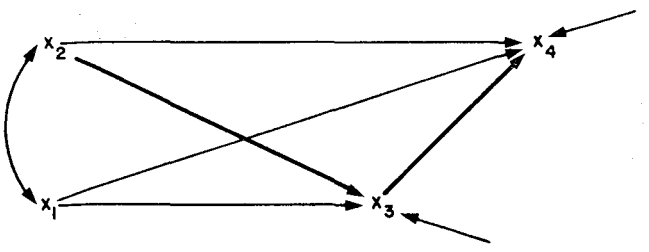
portanto,

$$\rho_{41} = p_{41} + p_{43}\ p_{31} + p_{42}\ \rho_{12} + p_{43}\ p_{32}\ \rho_{12}$$

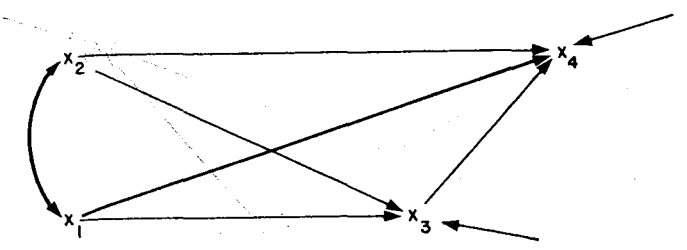
4) Decompomos ρ_{42} no efeito direto, p_{42}



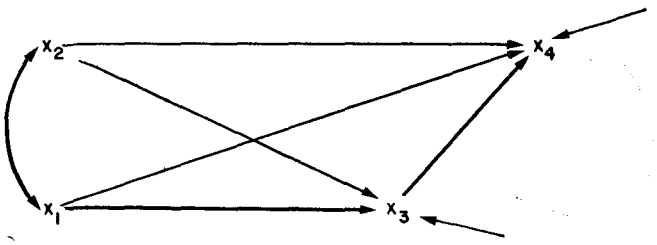
MAIS o efeito indireto, $p_{43}\ p_{32}$



MAIS a correlação devido à correlação de x_2 com outra causa (x_1), ambas operando diretamente, $p_{41}\ \rho_{12}$



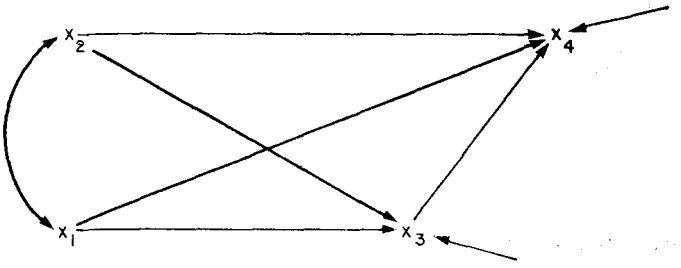
e indiretamente, $p_{43}\ p_{31}\ \rho_{12}$



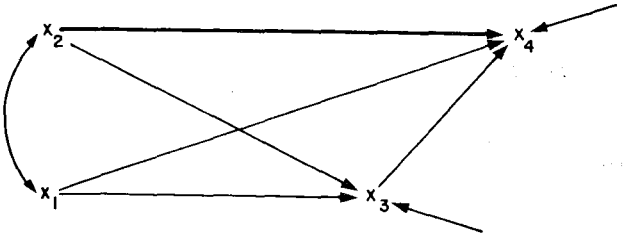
portanto,

$$\rho_{42} = p_{42} + p_{43}\ p_{32} + p_{41}\ \rho_{12} + p_{43}\ p_{31}\ \rho_{12}$$

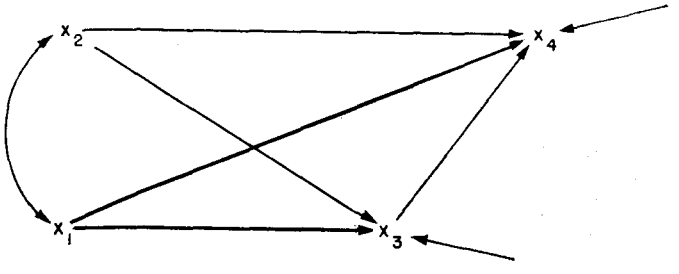
5) Como na versão anterior do modelo, ρ_{43} não envolve nenhum efeito indireto, mas a correlação gerada pelas causas comuns (x_1 e x_2) envolvem os efeitos diretos destas causas e a correlação devido ao fato de estarem correlacionadas. Daí, ρ_{43} é igual ao efeito direto, p_{43}



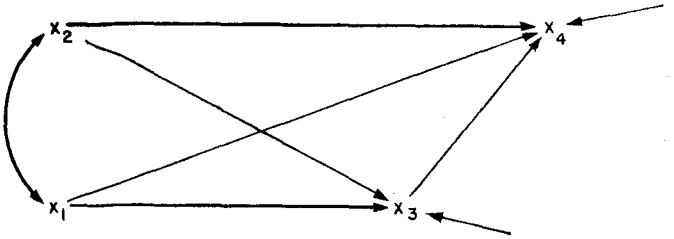
MAIS a correlação devido a x_2 como uma causa comum, $p_{42}\ p_{32}$



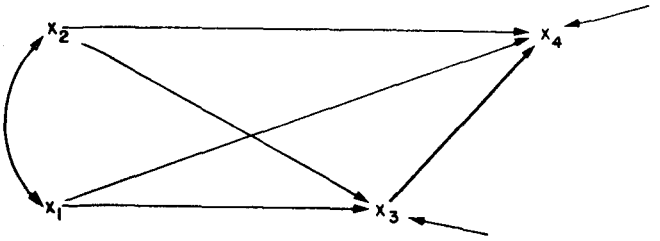
e x_1 como uma causa comum, $p_{41}\ p_{31}$



MAIS a correlação devido à correlação de x_1 com outra causa comum (x_2), $p_{42}\ p_{31}\ \rho_{12}$



e a correlação devido à correlação de x_2 com outra causa comum (x_1), $p_{41}\ p_{32}\ \rho_{12}$



portanto,

$$\rho_{43} = p_{43} + p_{42}\ p_{32} + p_{41}\ p_{31} + p_{42}\ p_{31}\ \rho_{12} + p_{41}\ p_{32}\ \rho_{12}$$

Estes resultados podem ser sintetizados na tabela 3.

TABELA 3

Decomposição da covariação bivariada

RELAÇÃO BIVARIADA $x_i\ x_j$	CORRELAÇÃO TOTAL = ρ_{ij}	EFEITO DIRETO EM x_i	CORRELAÇÃO DEVIDO A CAUSAS COMUNS E/OU CORRELACIONADAS
$x_3\ x_1$	ρ_{31}	p_{31}	$p_{32}\,\rho_{12}$
$x_3\ x_2$	ρ_{32}	p_{32}	$p_{31}\,\rho_{12}$
$x_4\ x_1$	ρ_{41}	p_{41}	$p_{43}\,p_{31} + p_{42}\,\rho_{12} + p_{43}\,p_{32}\,\rho_{12}$
$x_4\ x_2$	ρ_{42}	p_{42}	$p_{43}\,p_{32} + p_{41}\,\rho_{12} + p_{43}\,p_{31}\,\rho_{12}$
$x_4\ x_3$	ρ_{43}	p_{43}	$p_{42}\,p_{32} + p_{41}\,\rho_{31} + p_{42}\,p_{31}\,\rho_{12} + p_{41}\,p_{32}\,\rho_{12}$

Uma propriedade indesejável neste modelo é a covariação *ambígua* entre as variáveis exógenas, o que nos impede de extrair os efeitos destas duas variáveis. A nossa teoria é incapaz de nos dizer se x_1 causa x_2, x_2 causa x_1, mutuamente se influenciam, ambas são efeitos de uma ou mais causas comuns ou correlacionadas, ou se alguma combinação destas situações se verificam. Neste caso, não podemos saber com certeza

se uma mudança iniciada em, digamos, x_1 terá efeitos indiretos via x_2 ou não, já que não sabemos se x_2 depende de x_1. Segue-se que não podemos, com este modelo, estimar o coeficiente de efeito de x_1 em, digamos, x_4. Pode haver ou não uma conexão causal indireta de x_1 através de x_2 para x_4. E como não sabemos nada sobre este fato, então não sabemos incluir tal efeito indireto em nossa estimativa do coeficiente do efeito total.

Consideremos agora a trajetória do Tipo (c), em que a teoria existente sugere que a covariação entre as variáveis exógenas é não causal. Tal suposição pode ser incorporada a estrutura de trajetória, e o modelo pode ser representado como o da figura 6.

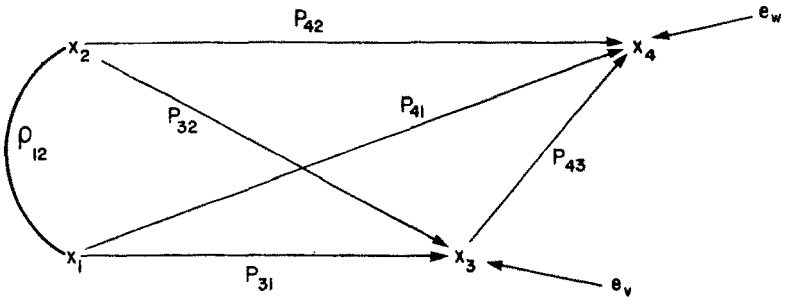


FIGURA 6

A estimação dos coeficientes segue o procedimento normal. A diferença entre este modelo e o apresentado na figura 5 é que o efeito causal de x_2 ou x_1 em x_4 está claramente definido na figura 6. Neste caso, somos capazes de dizer que a mudança de uma unidade em x_2 leva uma mudança de $p_{42} + p_{43}\, p_{32}$ em x_4, devido à suposição de que mudanças em x_1 não causarão qualquer mudança em x_2 e vice-versa. Neste modelo, ao se incorporar às suposições de ordem causal fraca e fechamento causal, a suposição de que ρ_{12} é não causal, todas as relações causais no modelo terão interpretações causais.

Outro exemplo de introdução de suposições poderosas para obter-se uma interpretação mais completa, que de outra forma seria ambígua, é ilustrado na figura 7.

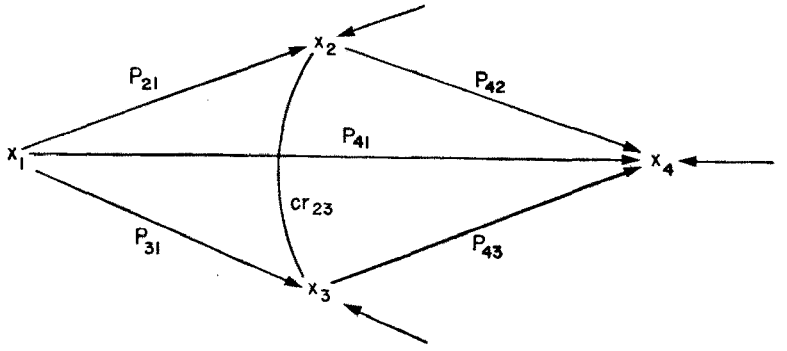


FIGURA 7

Supõe-se nesta figura que toda a covariação entre x_2 e x_3 é não causal, parte de sua covariação é devida a sua dependência comum em variáveis que estão fora do modelo e não correlacionadas com x_1. Dada esta suposição, estimam-se p_{21} e p_{31} pelas equações de regressão simples $x_2 = p_{21}\, x_1$ e $x_3 = p_{31}\, x_1$. O COEFICIENTE DE COVARIAÇÃO RESIDUAL da curva é dado por $cr_{23} = \rho_{23} - p_{21}\, p_{31}$. Os coeficientes restantes são estimados pela equação de regressão $x_4 = p_{41}\, x_1 + p_{41}\, x_2 + p_{43}\, x_3$. Assim como o modelo de trajetória da figura 6, a suposição de nenhuma trajetória causal entre x_2 e x_3 na figura 7 permite uma identificação completa do efeito causal total das variáveis de ordem inferior naquelas de ordem superior. Se a relação entre x_2 e x_3 fosse tratada como uma correlação não analisada, então não somente permaneceriam ambíguos os efeitos indiretos de todas as variáveis, como seria impraticável o método usual de estimação.

7 — O MODELO RESTRITO

Até agora só consideramos modelos nos quais todos os caminhos diretos indicados pela ordenação causal estão, de fato, presentes no modelo. Trataremos agora dos procedimentos utilizados em modelos recursivos que não apresentam uma ou mais trajetórias.

Suponha que o modelo, ainda recursivo, especifique claramente que um ou mais coeficientes de trajetória sejam nulos, como exemplo seja o seguinte diagrama.

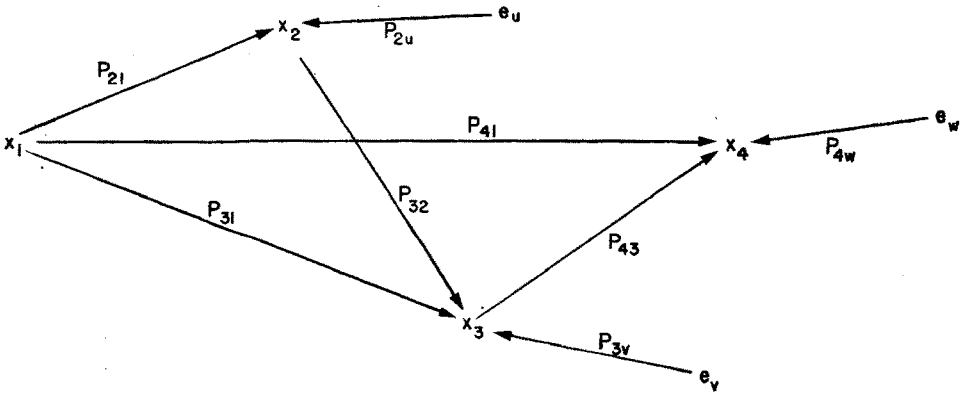


FIGURA 8

As equações do modelo são

(x_1 exógena)

$$x_2 = p_{21}\, x_1 + p_{2u}\, e_u$$

$$x_3 = p_{32}\, x_2 + p_{31}\, x_1 + p_{3v}\, e_v$$

$$x_4 = p_{43}\, x_3 + p_{41}\, x_1 + p_{4w}\, e_w$$

onde acrescentamos a restrição de que $p_{42} = 0$. Continuamos a supor que as variáveis (inclusive o termo de erro) estejam na forma padrão. Em cada equação do modelo o termo de erro é não correlacionado com as variáveis predeterminadas. Além disso, como consideramos ser o

modelo completamente recursivo, então o termo de erro em cada equação é não correlacionado com as variáveis predeterminadas em todas as equações precedentes. (Nesta seção, permitiremos que a correlação nula entre dois termos de erro seja uma conseqüência da correlação nula entre os termos de erro e as variáveis predeterminadas anteriormente), a força desta especificação tem uma importância especial com referência ao presente exemplo, pois $\rho_{2w} = 0$, embora x_2 não apareça (explicitamente) na equação $-x_4$.

Portanto, temos as seguintes suposições nos termos de erro: $\rho_{1u} = \rho_{1v} = \rho_{2w} = \rho_{1w} = \rho_{2v} = \rho_{3w} = 0$. E como conseqüência destas suposições, achamos que também serão verdadeiras as suposições $\rho_{uv} = \rho_{wu} = \rho_{vw} = 0$. E neste caso as estimativas pelo método dos mínimos quadrados serão não tendenciosas.

As equações normais para a equação $-x_2$ e a equação $-x_3$ são as mesmas de antes, e as regressões de mínimos quadrados nestas equações fornecem as estimativas para p_{21}, p_{2u}, p_{32}, p_{31} e p_{3v}.

Agora, as equações normais para a equação $-x_4$, obtêm-se multiplicando-se esta equação por cada variável predeterminada e calculando-se a expectância dos termos, achamos então

$$\rho_{41} = p_{41} + p_{43}\ \rho_{31}$$

$$\rho_{42} = p_{41}\ \rho_{21} + p_{43}\ \rho_{32}$$

$$\rho_{43} = p_{41}\ \rho_{31} + p_{43}$$

Supondo-se que os p's sejam conhecidos, temos três equações em dois coeficientes de trajetória desconhecidos, p_{41} e p_{43}. Em termos matemáticos, a solução para os p's é superdeterminada. Na linguagem dos modelos de equações estruturais, a equação $-x_4$ é SUPERIDENTIFICADA. No caso de uma equação no modelo ser super-identificada, podemos deduzir que se verificam uma ou mais RESTRIÇÕES SUPERIDENTIFICANTES se o modelo for verdadeiro. No exemplo existem três soluções distintas para cada coeficiente, se o modelo se verifica, então os valores obtidos em todas as três devem ser iguais.

Agora, a restrição superidentificante deve se verificar em qualquer POPULAÇÃO na qual se aplica o modelo. Mas se tivermos somente valores AMOSTRAIS das correlações, não podemos esperar que o modelo se verifique exatamente, nem podemos esperar que as três soluções para cada coeficiente de trajetória sejam exatamente iguais. Neste caso, para estimar os coeficientes de trajetória devemos escolher uma das soluções. É evidente que a solução preferida deve ser aquela obtida a partir das equações

$$\rho_{41} = p_{41} + p_{43}\ \rho_{31}$$

$$\rho_{43} = p_{41}\ \rho_{31} + p_{43}$$

observe que as estimativas de p_{41} e p_{43} obtidas a partir deste sistema são exatamente os coeficientes da regressão de x_4 em x_1 e x_3. Então a regra geral é:

"Em um modelo completamente recursivo (onde a correlação entre cada par de termo de erro é nula) estime-se os coeficientes em cada equação pela regressão de mínimos quadrados da variável dependente nas variáveis predeterminadas incluídas na equação."

A base para esta regra é a prova de que a variância amostral de um $\hat{p}$ estimado pelos mínimos quadrados é menor do que a variância de qualquer outra estimativa não tendenciosa do mesmo coeficiente, mesmo se tal estimativa parece usar mais informação quando combina as correlações que envolvem as variáveis predeterminadas.

7.1 — Teste para a Restrição Superidentificante

No exemplo precedente discutimos a estimação na suposição de que o modelo e, em particular, a restrição superidentificante no modelo eram consideradas serem verdadeiras. Mas o investigador pode não se sentir seguro nesta especificação. Na realidade, o investigador pode estar procedendo a um estudo precisamente para testar este aspecto de sua teoria que diz que um certo coeficiente deve ser nulo.

Em nosso exemplo o problema na especificação do modelo é se $p_{42} =$ ou $p_{42} \neq 0$. Em outras palavras, devemos decidir entre as duas especificações da equação $-x_4$

$$x_4 = p_{43}\, x_3 + p_{41}\, x_1 + p_{4w}\, e_w$$

e $$x_4 = p_{43}\, x_3 + p_{42}\, x_2 + p_{41}\, x_1 + p_{4w}\, e_w$$

Iniciemos na última especificação e estimemos pelo método dos mínimos quadrados a equação que contém p_{42}.

Para testar a hipótese nulo H_0: $p_{42} = 0$, calcule a razão

$$t = \frac{\hat{p}_{42}}{\text{E.p.}(\hat{p}_{42})}$$

onde E.p. $(\hat{p}_{42})$ é o erro padrão da estimativa do coeficiente de ligação p_{42}, e compare tal valor de t com o da distribuição de t, com os graus de liberdade apropriados. Se a amostra for razoavelmente grande, quanto $|t| \geqslant 2{,}0$, podemos concluir que, com um risco de erro não maior do que 5%, que a hipótese nula é falsa. Neste caso rejeitaríamos a restrição superidentificante no modelo e, provavelmente, reespecificaríamos o modelo e introduziríamos um valor não nulo de p_{42}.

No caso de não rejeição da hipótese nula — i.é, se a razão t não é estatisticamente significante — a situação é intrinsecamente ambígua. Esclarecendo, o investigador não é obrigado a ACEITAR a hipótese nula a menos que exista uma razão suficiente *a priori* para fazer isto. Poderia acontecer, por exemplo, que o verdadeiro valor de p_{42} fosse positivo mas pequeno, de tal forma que nossa amostra não seria suficiente para detectar confiavelmente o efeito. Se nossa teoria garantisse seguramente que $p_{42} \neq 0$, a despeito do resultado do teste, poderíamos manter p_{42} na equação. De qualquer forma, é de boa prática indicar os erros padrões de todos os coeficientes, de tal forma que se possa ter uma idéia da precisão nas estimativas dos coeficientes.

7.2 — Decomposição da Covariação da Relação Bivariada

No modelo restrito a tabela de decomposição é um tanto diferente da apresentada para o modelo geral. Considerando o diagrama da figura 8, a decomposição da covariação é apresentada na tabela 4.

TABELA 4

Decomposição da covariação no modelo restrito

RELAÇÃO BIVARIADA X_i X_j	COVARIAÇÃO ρ_{ij}	CAUSA				EFEITO TOTAL $c_{ij} = p_{ij} + I$	IMPLÍCITO PELO MODELO RESTRITO $c_{ij} + CD + CI$ (A)	CORRELAÇÃO NÃO EXPLICADA PELO MODELO $\rho_{ij} - A$	NÃO CAUSAL $A - c_{ij} = CD + CI$
		Direta p_{ij}	Indireta I	Comum					
				Direta (CD)	Indireta (CI)				
X_2 X_1	ρ_{21}	p_{21}	—	—	—	ρ_{21}	ρ_{21}	—	—
X_3 X_1	ρ_{31}	p_{31}	$p_{32}\,p_{21}$	—	—	ρ_{31}	ρ_{31}	—	—
X_3 X_2	ρ_{32}	p_{32}	—	$p_{32}\,p_{21}$	—	p_{32}	ρ_{32}	—	$p_{32}\,p_{21}$
X_4 X_1	ρ_{41}	p_{41}	$p_{43}\,p_{31} + p_{43}\,p_{32}\,p_{21}$	—	—	$p_{41} + p_{43}\,p_{31} + p_{43}\,p_{32}\,p_{21}$	c_{41}	$\rho_{41} - c_{41}$	—
X_4 X_2	ρ_{42}	—	$p_{43}\,p_{21}$	$p_{41}\,p_{21}$	$p_{43}\,p_{31}\,p_{21}$	$p_{43}\,p_{32}$	$c_{42} + p_{41}\,p_{21} + p_{43}\,p_{31}\,p_{21}$	$\rho_{42} - [c_{42} + p_{41}\,p_{21} + p_{43}\,p_{31}\,p_{21}]$	$p_{41}\,p_{21} + p_{43}\,p_{31}\,p_{21}$
X_4 X_3	ρ_{43}	p_{43}	—	$p_{41}\,p_{31}$	$p_{41}\,p_{21}\,p_{32}$	p_{43}	$c_{43} + p_{41}\,p_{31} + p_{41}\,p_{21}\,p_{32}$	$\rho_{43} - [c_{31} + p_{41}\,p_{31} + p_{41}\,p_{21}\,p_{32}]$	$p_{41}\,p_{31} + p_{41}\,p_{21}\,p_{32}$

8 — INFERÊNCIA ESTATÍSTICA

Quando a análise de trajetória é desenvolvida em dados amostrais e as conclusões devem ser generalizadas para uma dada população, a variação devido à amostragem deve ser considerada.

A estimação dos coeficientes de trajetórias de uma população exige simplesmente uma série de regressões de mínimos quadrados, onde se considera uma variável de cada vez como variável dependente e todas as outras variáveis de ordem inferior como variáveis independentes. Portanto, no caso do modelo geral, é necessário resolver (n — 1) equações de regressão, se o modelo contiver n variáveis explícitas.

8.1 — Significância dos Coeficientes de Trajetórias

Na seção 7 apresentamos o teste t para averiguar a possibilidade de um dado coeficiente ser nulo na população, apresentamos agora o teste F que também pode ser utilizado com este propósito.

A estratégia usual neste teste dos coeficientes envolve a decomposição da soma de quadrados explicada em componentes atribuíveis a cada variável independente na equação. Existem dois métodos de decomposição, que indicaremos por (1) o MÉTODO PADRÃO e (2) o MÉTODO HIERÁRQUICO. No *método padrão* cada variável é tratada como se fosse agregada à equação de regressão em um estágio em separado após todas as outras variáveis terem sido incluídas. O incremento em R^2 (ou na soma de quadrados explicada) devido ao acréscimo de uma dada variável é tomado como a componente da variação atribuível àquela variável. No *método hierárquico* as variáveis são acrescentadas à equação de regressão em uma ordem predeterminada pelo investigador. As variáveis são incorporadas seqüencialmente, e o incremento em R^2 (ou na soma de quadrados explicada) em cada estágio, é considerado como a componente da variação atribuível à variável particular acrescentada naquele estágio.

O método padrão e o hierárquico fornecerão incrementos diferentes atribuíveis às várias variáveis independentes, e a razão empregada no teste de significância dos coeficientes diferirá. O pesquisador deverá escolher entre as duas estratégias. O critério essencial neste caso é se o pesquisador considera as correlações entre as variáveis independentes como sendo CAUSAL ou NÃO CAUSAL. Se, por exemplo, não existir uma ordem intrínseca nas relações entre as variáveis e as correlações entre estas variáveis são consideradas como não causais, então a estratégia padrão será a adequada. Por outro lado, se as correlações entre estas variáveis são resultados de uma ordenação causal, então a estratégia hierárquica é a mais apropriada.

8.1.1 — O teste padrão

Para este teste a razão F é dada por

$$F = \frac{\text{incremento na SQ devido a } x_i/1}{SQ_{\text{residual}}/(n - k - 1)} \qquad (2)$$

com 1 e (n — k — 1) graus de liberdade, onde n é o tamanho da amostra k é o número de variáveis independentes na equação e SQ é soma de quadrados. Se o valor de F calculado for maior do que o valor crítico

tabelado para um dado nível de significância, digamos 0,05, a hipótese nula $c_i = 0$ será rejeitada. Caso contrário, se concluirá que o coeficiente observado não é significante ao nível de 0,05, i.é, não podemos rejeitar a hipótese nula $c_i = 0$.

Como por exemplo consideremos o seguinte diagrama

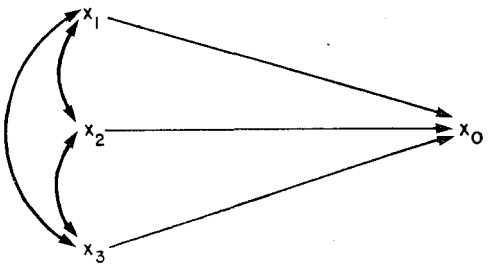


FIGURA 9

Suponhamos o seguinte resultado para este diagrama, admitindo-se $n = 100$

VARIÁVEL	C_{0i}	p_{0i}	F
X_1	1,2963	0,03889	14,563
X_2	0,0889	0,1778	2,717
X_3	0,0185	0,0556	0,297
Constante	2,9889		

Tais resultados são informados como saída do subprograma SPSS REGRESSION. Para completar o teste deve-se comparar o valor de F com o do F tabelado para 1 e 96 graus de liberdade. O leitor pode verificar que o único c_{0i} significante ao nível de 0,05, é o da variável X_1.

8.1.2 O teste hierárquico

O método hierárquico, ao invés de determinar a contribuição de cada variável pela suposição de que foi a última a ser acrescentada à equação, exige que o pesquisador indique a ordem de inclusão. Para a primeira variável a ser testada utiliza-se a razão F,

$$F = \frac{r^2_{y1} / 1}{(1 - R^2_{y.12,\ldots,k}) / (n - k - 1)} \tag{3}$$

O segundo coeficiente é testado pelo valor F calculado por

$$F = \frac{\text{incremento na SQ devido a } x_2/1}{(1 - R^2_{y.12,\ldots,k}) / (n - k - 1)} \tag{4}$$

ou,

$$F = \frac{(R^2_{y.12} - r^2_{y1}) / 1}{(1 - R^2_{y.12,\ldots,k}) / (n - k - 1)}$$

Para a terceira variável na ordem, o teste estatístico é

$$F = \frac{\text{incremento na SQ devido a } x_3/1}{(1 - R^2_{y.12,\ldots,k})/(n-k-1)} \qquad (5)$$

ou,

$$F = \frac{(R_{y.123} - R_{y.12})/1}{(1 - R^2_{y,12,\ldots,k})/(n-k-1)}$$

e assim sucessivamente até a k-ésima variável na ordem preestabelecida. Todos estes valores calculados de F devem ser comparados com o valor tabelado de F para 1 e (n — k — 1) graus de liberdade. Nestas fórmulas estamos considerando y como variável dependente e x_1, x_2, ... x_k como variáveis independentes.

A fim de ilustrar o procedimento do teste hierárquico, vamos supor que o pesquisador tenha estabelecido a ordem causal apresentada nas variáveis da figura 4. Se o método padrão fosse empregado, os testes para os coeficientes refletiriam somente as trajetórias diretas entre as variáveis dependentes e x_4. Já o procedimento hierárquico envolve ajustamentos somente para aquelas variáveis que precedem uma dada variável na ordem hierárquica, e, portanto, reflete a *Influência Total* de cada variável. Por exemplo, como x_1 é a primeira na hierarquia, será testada sem efetuar ajuste em relação a x_2 e x_3. E a soma de quadrados atribuível a x_1 não somente incluirá a parte devido a sua influência direta em x_4, mas também a parte devido à sua influência indireta através da trajetória $x_2 \rightarrow x_4$, $x_3 \rightarrow x_4$ e $x_2 \rightarrow x_3 \rightarrow x_4$. Da mesma forma, a parte da soma de quadrados atribuível a x_2 refletirá sua influência direta mais a sua influência indireta via $x_3 \rightarrow x_4$.

A tabela a seguir mostra um grupo de estatísticas que podem ser obtidas como parte da saída do subprograma SPSS *Regression*. Esta tabela não fornece os valores de F calculado para o teste hierárquico, mas poucos cálculos são necessários para obter o F calculado a partir desta tabela:

VARIÁVEL	R. MÚLTIPLO	R^2	INCREMENTO EM R^2
x_1	0,5000	0,2500	0,2500
x_2	0,5292	0,2800	0,0300
x_3	0,5312	0,2822	0,0022

Aplicando-se as fórmulas de (3) a (5) às quantidades da tabela, temos

Para x_1:

$$F = \frac{0,2500/1}{(1 - 0,2822)/96} = 33,4359$$

Para x_2:

$$F = \frac{0,0300/1}{(1 - 0,2822)/96} = 4,012$$

Para x_3:

$$F = \frac{0,0022/1}{(1 - 0,1811)/96} = 0,2942$$

comparando-se estes F calculados com os valores tabelados de F com 1 e 96 graus de liberdade, verificamos que os coeficientes para x_1 e x_2 são significantes ao nível de 0,05, enquanto que o coeficiente de x_3 não é significante.

9 — COEFICIENTES PADRONIZADOS E ABSOLUTOS

Até agora temos somente utilizado coeficientes padronizados na análise de trajetória, isto devido a conveniências de interpretação, como foi explicado na seção 4. Contudo, existem duas sérias desvantagens nos coeficientes padronizados. Concluiremos nossa discussão sobre análise de trajetória com um exame breve destas desvantagens.

Admitamos uma estrutura causal bem simples na qual uma variável X_0 está completamente determinada por três variáveis causais X_1, X_2 e X_3. Suponhamos também que todas estas variáveis causais sejam não correlacionadas, tal estrutura pode ser representada por

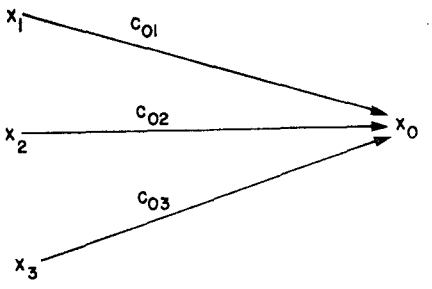


FIGURA 10

Admitamos, sem perda de generalidade, que todas as variáveis causais sejam medidas a partir de suas respectivas médias.

Pela figura 10, X_0 é determinada por

$$X_0 = b_{01} X_1 + b_{02} X_2 + b_{03} X_3$$
$$= c_{01} X_1 + c_{02} X_2 + c_{03} X_3$$

onde b_{ij} é um coeficiente absoluto da regressão e c_{ij} é o coeficiente do efeito. A variância em X_0 é dada por

$$\text{VAR}(X_0) = \text{VAR}(b_{01} X_1) + \text{VAR}(b_{02} X_2) + \text{VAR}(b_{03} X_3)$$
$$= b_{01}^2 \text{VAR}(X_1) + b_{02}^2 \text{VAR}(X_2) + b_{03}^2 \text{VAR}(X_3)$$

ou

$$\sigma_0^2 = b_{01}^2 \sigma_1^2 + b_{02}^2 \sigma_2^2 + b_{03}^2 \sigma_3^2 \qquad (6)$$

Agora dividindo-se σ_0^2 vem

$$\frac{\sigma_0^2}{\sigma_0^2} = b_{01}^2 \frac{\sigma_1^2}{\sigma_0^2} + b_{02}^2 \frac{\sigma_2^2}{\sigma_0^2} + b_{03}^2 \frac{\sigma_3^2}{\sigma_0^2}$$

ou

$$1 = p_{01}^2 + p_{02}^2 + p_{03}^2 \qquad (7)$$

A expressão (7) é bem mais simples que (6), e com (7) e os desvios padrões das variáveis a expressão (6) pode ser reconstruída. No entanto, com a padronização das variáveis perde-se uma parte crucial da informaçção. Por exemplo, a importância relativa de p_{01} em (7) depende não só de c_{01} (ou b_{01}) como de σ_1, já que $p_{0I} = b_{01}\ \sigma_1/\sigma_0$. Admitamos que os coeficientes de efeito sejam iguais, i.é., $c_{01} = c_{02} = c_{03}$, mas que $\sigma_1 = 2\sigma_2 = 2\sigma_3$, quer dizer a variância de X_1 é duas vezes a variância de X_2 e X_3. Então p_{01} será o dobro de p_{03}. Se tivéssemos utilizados coeficientes absolutos, teríamos percebido a igualdade dos coeficientes de efeito, mas se tivéssemos utilizado só os coeficientes padronizados não teríamos notado esta semelhança crucial. Por outro lado, se as variáveis X_1, X_2 e X_3 não estiverem sendo medidas na mesma unidade, não tem sentido a comparação de coeficientes não padronizados na mesma população.

Mas admita-se que estejamos interessados nos valores relativos dos c's para a mesma variável em duas amostras ou populações distintas. Mais especificamente, suponha que estejamos interessados nas similaridades ou diferenças na estrutura causal da discriminação salarial no Norte e no Sul. Suponha, ainda, que X_1 corresponda às remunerações mensais, X_2 a uma variável *dummy* que representa sexo, e X_3 a uma variável *dummy* que representa raça. Finalmente, admita-se a hipótese de ser a estrutura causal como a da figura 11, e que o salário do empregado no Norte e no Sul depende de só dois fatores e que o efeito de cada fator é o mesmo em ambos os lugares.

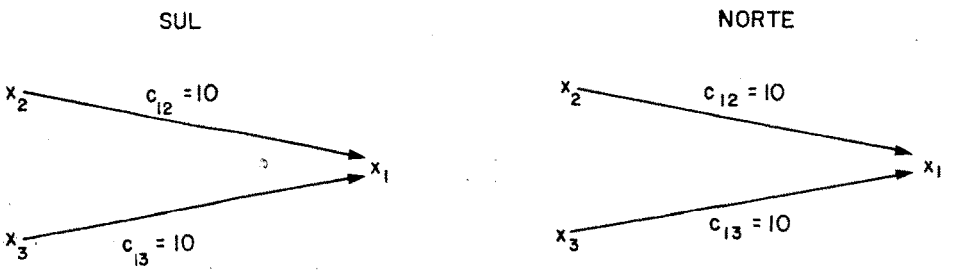


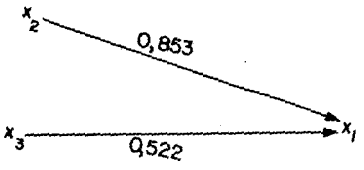
FIGURA 11

Portanto, poderíamos dizer que a estrutura subjacente à discriminação por sexo e raça é a mesma nos dois lugares. Os coeficientes ABSOLUTOS revelarão tais igualdades a respeito das diferenças que possam existir na variação de X_2 e X_3. Por exemplo, obteríamos os mesmos coeficientes absolutos no Sul e no Norte pela regressão

$$X_1 = B_{12}\ X_2 + B_{13}\ X_3$$
$$= 10\ X_2 + 10\ X_3$$

onde, por simplicidade, fizemos de Cr$ 10,00 as diferenças nas remunerações mensais entre homens e mulheres e negros e brancos. Contudo, os coeficientes padronizados mascaram as similaridades se a variância em qualquer uma das variáveis (por exemplo, a proporção de mulheres empregadas no Sul) for diferente das variâncias de outra (por exemplo, a proporção de mulheres empregadas no Norte). Examinemos alguns exemplos hipotéticos. Na figura 12(a) estamos admitindo que a proporção de mulheres no Norte (40%) é maior do que a do Sul (10%), enquanto que a proporção de negros é a mesma nas duas regiões (10%). Em 12(b) estamos supondo que a proporção de mulheres é a

SUL



$V_1 = 3300$

$V_2 = 24$

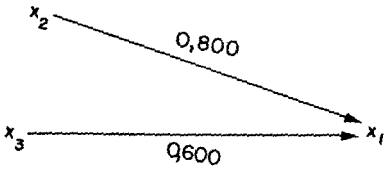
$V_3 = 9$

$N = 100$

$$P_{12}^2 = 100 \frac{24}{3300}$$

$$P_{13}^2 = 100 \frac{9}{3300}$$

NORTE



$V_1 = 2500$

$V_2 = 16$

$V_3 = 9$

$N = 100$

$$P_{12} = 10 \frac{4,90}{57,45} = 0,853$$

$$P_{13} = 10 \frac{3}{57,45} = 0,522$$

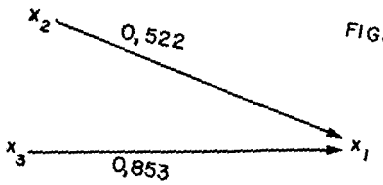
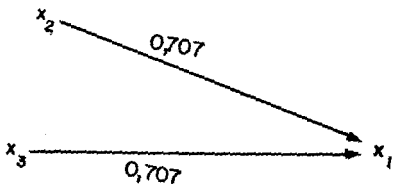


FIGURA 12 (a)



$V_1 = 3300$

$V_2 = 9$

$V_3 = 24$

$N = 100$

$V_1 = 1800$

$V_2 = 9$

$V_3 = 9$

$N = 100$

FIGURA 12 (b)

mesma nas duas regiões (10%), enquanto que a proporção de negros é maior no Sul (40%) do que no Norte (10%).

Portanto, a menos que se compare os coeficientes absolutos para as duas populações, não poderemos dizer se as diferenças observadas são devido à operação de processos causais diferentes, ou meramente devido às diferenças na variabilidade nas variáveis independentes na população.

O coeficiente padronizado combina a importância relativa de c_{ij} e a importância relativa da variância de X_j em um único dado que contém informação mais resumida do que o coeficiente *absoluto,* mas exatamente por ser um valor sintético é que não consegue diferenciar entre dois aspectos conceitualmente distintos na estrutura linear.

Em resumo, se o objetivo é o montante relativo de variância explicada em Y, para uma dada amostra ou população, por várias variáveis independentes, os coeficientes padronizados são mais apropriados. Se as variáveis independentes são medidas em unidades distintas e o interesse principal está em se obter o efeito total de uma variável sobre outra, na mesma amostra ou população, então os coeficientes padronizados serão uma solução mais inteligente. No entanto, deve-se dar preferência aos coeficientes absolutos caso se esteja interessado na descoberta de *leis causais* ou *processos causais* e/ou na comparação dos parâmetros de uma população com os de outra.

10 — RESUMO DOS PROCEDIMENTOS DA ANÁLISE DE TRAJETÓRIA

O fluxograma a seguir ilustra os passos de uma análise de trajetória.

11 — APLICAÇÃO

Nos dois primeiros exemplos [1] que serão apresentados utilizou-se o SUBPROGRAMA SPSS REGRESSION para a determinação dos coeficientes estruturais, de trajetória e residuais das equações de regressão extraídas dos diagramas causais.

O primeiro exemplo visa a familiarizar o leitor com os cálculos expressos algebricamente para a obtenção das tabelas de decomposição de um modelo geral. Tanto as variáveis quanto as hipóteses que participam do diagrama são subconjuntos das variáveis e hipóteses empregadas no segundo exemplo.

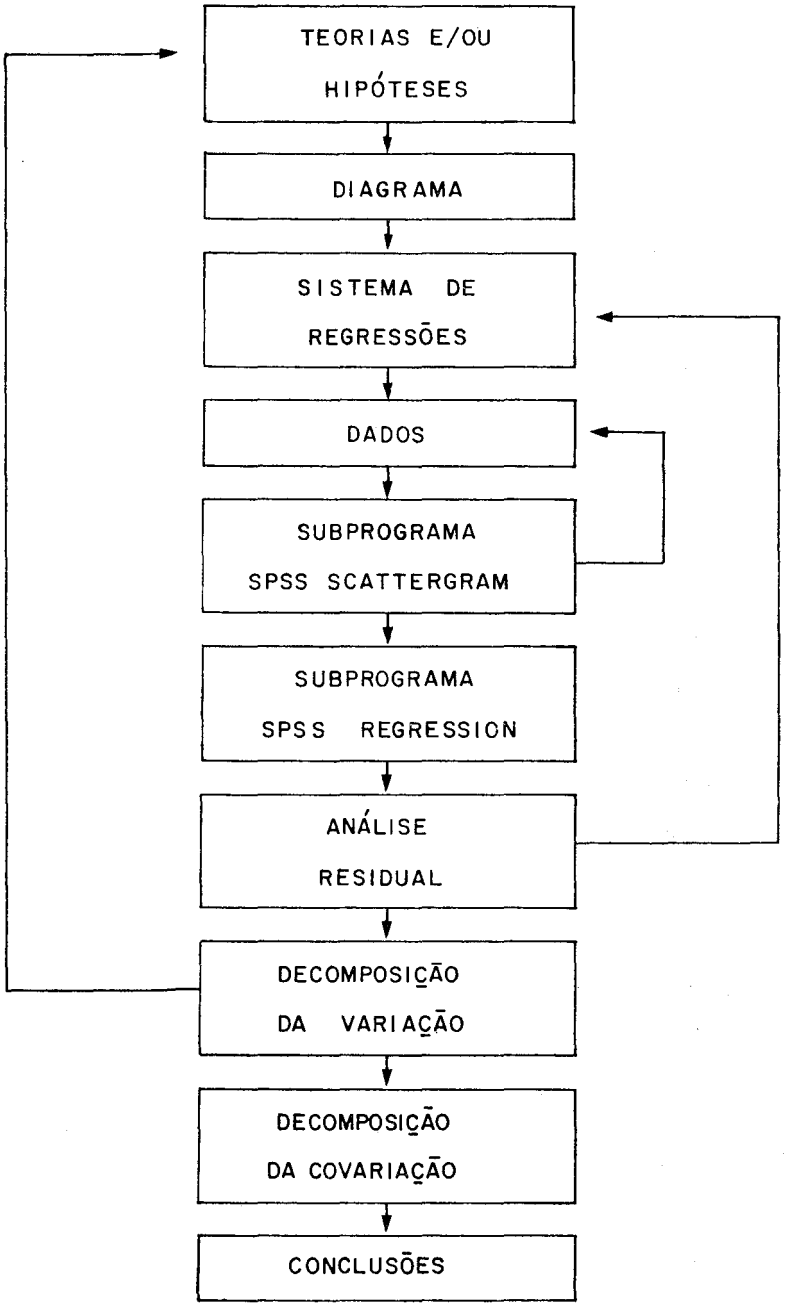
O segundo exemplo [2] consiste em primeira fase de aplicação de Análise Causal aos estudos que vêm sendo desenvolvidos pela Divisão de Estudos Urbanos do Departamento de Geografia, sobre Aglomerações Urbanas no Brasil. Neste exemplo, as implicações lógicas indicadas pelos coeficientes e pelas tabelas de decomposição serão analisadas.

Os dois últimos exemplos ilustram a aplicação da Análise de Trajetória na Demografia e na Genética.

---

1 A preparação das tabelas de decomposição destes exemplos contou com a colaboração de Maria Cristina Moreira Safadi — Estatística do DEGEO.

2 A formulação das hipóteses e a análise das tabelas deste exemplo são de autoria da Prof.ª Olga Maria Buarque de Lima Fredrich — Geógrafa do DEGEO.



11.1 — Modelo Geral

Para ilustrar os procedimentos de cálculos necessários à preparação das tabelas finais da análise de trajetória de um modelo geral, utilizou-se o diagrama da figura 2. O diagrama é reapresentado na figura 13 já com os valores dos coeficientes de trajetória. Deve-se notar que as variáveis estão na forma padronizada.

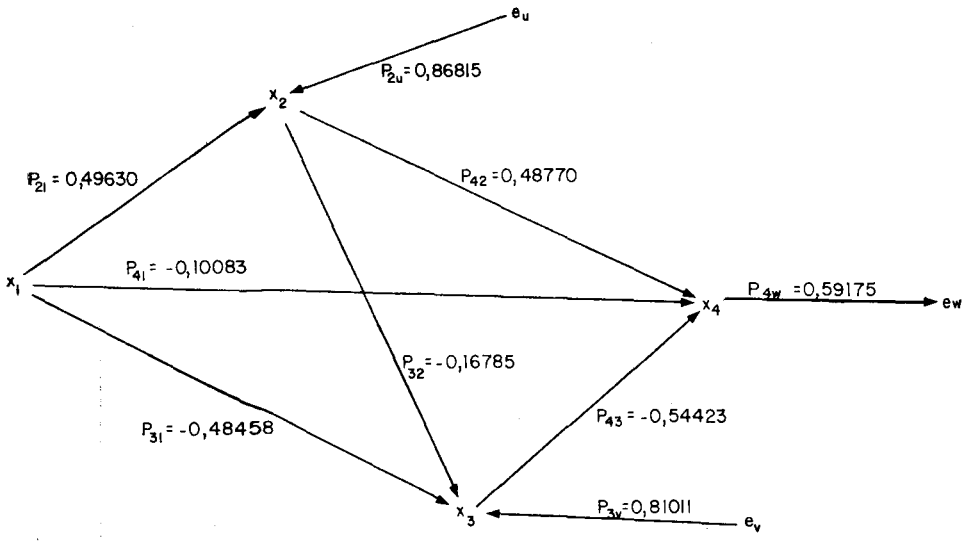


FIGURA 13

As equações de trajetória do modelo são

(x_1 exógena)

$$x_2 = p_{21}\, x_1 + p_{2u}\, e_u$$

$$x_3 = p_{32}\, x_2 + p_{31}\, x_1 + p_{3v}\, e_v$$

$$x_4 = p_{43}\, x_3 + p_{42}\, x_2 + p_{41}\, x_1 + p_{4w}\, e_w$$

onde

X_1 — NÍVEL DE INSTRUÇÃO 1

% da população de 25 anos e mais com curso médio ou curso superior completo.

X_2 — RENDIMENTO DA PEA

% da população economicamente ativa com rendimento mensal de Cr$ 201,00 a Cr$ 400,00.

X_3 — ESTRUTURA ETÁRIA

% da população no grupo etário 0 ⊢ 14 anos em relação a população total.

X_4 — NÍVEL DE INSTRUÇÃO 2

% da população de 12 a 14 anos de idade com curso elementar completo.

O resultado das regressões do modelo estão na tabela 5. Neste exemplo, bem como nos dois seguintes, os coeficientes estruturais são apresentados porque são mais convenientes para a comparação dentro da própria estrutura, já que as variáveis estão na mesma unidade de medida (%).

TABELA 5

Coeficientes das equações estruturais e de trajetória

RELAÇÃO BIVARIADA X_i X_j	COEFICIENTE ESTRUTURAL b_{ij}	COEFICIENTE DE TRAJETÓRIA p_{ij}
X_2 X_1	0,77569	0,49630
X_2 e_u	—	0,86815
X_3 X_1	—0,45883	—0,48458
X_3 X_2	—0,10158	—0,16785
X_3 e_v	—	0,81011
X_4 X_1	—0,44492	—0,10083
X_4 X_2	1,37685	0,48770
X_4 X_3	—2,53885	—0,54423
X_4 e_w	—	0,59175

A tabela mostra que em cada equação a ordenação das variáveis independentes segundo seus efeitos, medidos pelos coeficientes estruturais, se mantém nos coeficientes de trajetória. Vemos também que na equação $-X_3$ os coeficientes estruturais indicam que a variável de maior efeito causal direto em X_3 é a variável X_1, tal fato também é acompanhado nos coeficientes de trajetória. Na equação $-X_4$ a variável de maior efeito causal direto em X_4 é a variável X_3, note-se que tal efeito é negativo, indicando que tal variável causal age provocando uma redução, i.é, quanto maior o aumento na variável X_3 tanto menor será o nível da variável X_4. Esse comportamento no relacionamento entre as variáveis X_3 e X_4 também é acompanhado nos coeficientes de trajetória.

A fórmula de determinação completa (1) permite a decomposição percentual da variação da variável dependente em cada equação. Os cálculos necessários seguem a tabela 2, os resultados são apresentados na tabela 6.

TABELA 6

Decomposição da variação

VARIÁVEL DEPENDENTE	VARIAÇÃO					
	Determinada Diretamente pelos Fatores Imediatos		Determinada pelas Intercorrelações entre as Variáveis Independentes		Determinada pelo Resíduo	
	Termo	%	Termo	%	Termo	%
X_2	p^2_{21}	24,6				
	Total	24,6	—	—	p^2_{2u}	75,4
X_3	p^2_{31}	23,5				
	p^2_{32}	2,8				
	Total	26,3	$2p_{31}p_{32}r_{12}$	8,1	p^2_{3v}	65,6
X_4	p^2_{41}	1,0				
	p^2_{42}	23,8				
	p^2_{43}	29,6				
	Total	54,4	$2[p_{41}p_{42}r_{12} + p_{41}p_{43}r_{13} + p_{42}p_{43}r_{23}]$	10,6	p^2_{4w}	35,0

Vemos por esta tabela que a variável X_4 é a que tem a maior variação explicada pelo modelo (75%), e isto se deve principalmente à contribuição das variáveis X_2 e X_3. Apesar de a variável X_1 ter grande influência nas variáveis X_2 e X_3, tal não ocorre com relação a variável X_4.

A partição das correlações entre as variáveis da estrutura em componentes causais e espúrias é feita na tabela 8. As expressões utilizadas nos cálculos são as constantes da tabela 1.

TABELA 7

Decomposição da covariação bivariada

RELAÇÃO BIVARIADA X_i X_j	COVARIAÇÃO TOTAL p_{ij}	CAUSA				EFEITO TOTAL $c_{ij} = p_{ij} + I$	NÃO CAUSAL CD + CI
		Direta p_{ij}	Indireta (I)	Comum			
				Direta (CD)	Indireta (CI)		
X_2 X_1	0,49630	0,49630	—	—	—	0,49630	—
X_3 X_1	—0,56788	—0,48458	—0,08330	—	—	—0,56788	—
X_3 X_2	—0,40834	—0,16785	—	—0,24049	—	—0,16785	—0,24043
X_4 X_1	0,45027	—0,10083	0,5511	—	—	0,45027	—
X_4 X_2	0,65989	0,48770	0,09135	—0,05004	0,13088	0,57905	0,08084
X_4 X_3	—0,68612	—0,54423	—	—0,03300	—0,10889	—0,54423	—0,14189

Esta tabela exibe algumas características interessantes do modelo. Como foi demonstrado na tabela 1, as correlações entre as variáveis X_2, X_3 e X_4 com a X_1 devem ser completamente explicadas em termos causais, a tabela 8 confirma aqueles resultados. Contudo, nota-se uma distinção forte na atuação dessas causas, por exemplo, na relação (X_4 X_1) vemos que o efeito causal indireto da variável X_1 na X_4 é tão intenso que supera a própria correlação total r_{14}. Na relação (X_3 X_1) ocorre o contrário, visto que o efeito direto p_{13} quase alcança o valor da correlação r_{13}.

11.2 — Análise Causal das Aglomerações Urbanas

As variáveis deste exemplo estão sendo utilizadas em um estudo sobre a estrutura econômica e social das aglomerações urbanas brasileiras. Dada a suposição de uma estrutura de causação nestas variáveis, propôs-se a execução de uma análise de trajetória como uma primeira abordagem para o problema.

Este exemplo ilustra a análise de um modelo restrito. Apesar da estrutura causal nas variáveis ser bem distinta daquela apresentada na figura 8, as características básicas da análise desenvolvida em termos algébricos se mantém nas tabelas que serão calculadas.

(1) VARIÁVEIS

As variáveis que constam deste exemplo são:

X_1 — NÍVEL DE INSTRUÇÃO 1

% da população de 25 anos e mais com curso médio ou curso superior completo.

X_2 — MIGRAÇÃO

% de população não natural do município onde reside com tempo de permanência até 5 anos, em relação à população total.

X_3 — RENDIMENTO MENSAL DA PEA

% da PEA com rendimento mensal de Cr$ 201,00 a Cr$ 400,00.

X_4 — ESTRUTURA ETÁRIA

% da população no grupo etário 0 ⊢ 14 anos em relação a população total.

X_5 — NÍVEL DE INSTRUÇÃO 2

% da população de 12 a 14 anos de idade com curso elementar completo.

X_6 — CONSUMO DE BENS DURÁVEIS

% de domicílios urbanos cujos moradores possuem automóvel.

X_7 — INFRA-ESTRUTURA

% de domicílios urbanos com instalações sanitárias ligadas a rede geral ou com fossa séptica.

(2) UNIDADES DE OBSERVAÇÃO

As variáveis foram medidas para 90 unidades de observação, correspondentes a aglomerações urbanas e a municípios com cidades de população igual ou superior a 50 mil habitantes não incluídos em aglomerações.

(3) HIPÓTESES

Admite-se que existam entre as variáveis as seguintes relações:

1. Considera-se o nível de instrução a um tempo causa e efeito da renda da população. De uma maneira simplista, pressupõe-se:

(i) que o nível de renda da população seria afetado pelo nível de instrução (PROXY da qualificação profissional) da população adulta ($X_1 \rightarrow X_3$).

(ii) que o nível de instrução da população pertencente aos grupos etários mais jovens dependeria dos rendimentos da população em idade adulta, ou seja, que o *status* econômico dos pais influenciaria o nível de escolaridade dos filhos ($X_3 \rightarrow X_5$).

2. Pressupondo-se que a renda dos migrantes recentes seja, em média, inferior à da população residente, um aumento relativamente grande da população devido à migração poderia, pelo menos a curto prazo, concorrer para diminuir a renda média da mesma. Em outras palavras, os migrantes, a curto prazo, concorreriam para aumentar o tamanho dos estratos de menor renda ($X_2 \rightarrow X_3$).

3. Considerando-se que a propensão para migrar é maior em uma determinada faixa de idade, um aumento relativamente grande da população devido à migração poderia, pelo menos a curto prazo, afetar a estrutura etária da população de uma dada área ($X_2 \rightarrow X_4$).

4. Em algumas aglomerações urbanas a deficiência em infra-estrutura estaria, em parte, ligada a um crescimento muito forte da população, a uma defasagem entre a rapidez com que se processa o crescimento populacional e o tempo requerido para a implantação de

infra-estrutura — especialmente no que concerne saneamento básico; e uma participação mais expressiva de migrantes recentes caracterizaria as áreas de forte crescimento populacional ($X_2 \rightarrow X_7$).

5. Pelo maior poder de decisão e de pressão dos estratos de população que dispõem de maior renda, os investimentos públicos em infra-estrutura tenderia a ser mais intensos nos lugares onde é maior a proporção de população residente enquadrada nas faixas de renda mais elevadas. Por outro lado, a renda da população refletiria, em parte, a capacidade de gerar recursos das atividades desenvolvidas nas diferentes aglomerações urbanas, portanto, a capacidade de investir de cada uma delas ($X_3 \rightarrow X_7$).

6. A intensidade e diversificação do consumo de bens duráveis varia em função da renda da população ($X_3 \rightarrow X_6$).

7. Um aumento da renda geraria uma diminuição relativa dos grupos etários mais jovens, através de uma diminuição da taxa de natalidade, sobretudo, mas também por um aumento da esperança de vida ($X_3 \rightarrow X_4$).

8. Um peso maior dos grupos etários mais jovens, aumentando a taxa de dependência, diminui o poder de consumo da população economicamente ativa ($X_4 \rightarrow X_6$).

9. Um peso maior dos grupos etários mais jovens, aumentando a taxa de dependência, limitaria a capacidade da população adulta em investir na educação da população em idade escolar ($X_4 \rightarrow X_5$).

10. Supõe-se *a priori*, que $p_{41} = p_{51} = p_{61} = p_{71} = p_{52} = p_{62} = p_{74} = p_{65} = p_{75} = 0$. Com tal suposição indicamos que não há nenhuma evidência de que exista causação direta entre as variáveis que os coeficientes de trajetória acima relacionam.

11. Considera-se como não causal a correlação entre os fatores primários X_1 e X_2.

12. Os resíduos são não correlacionados com as variáveis predeterminadas.

(4) DIAGRAMA

A teoria verbal nas hipóteses sobre o sentido da causação nas variáveis é transcrita a seguir para uma forma diagramática na figura 14.

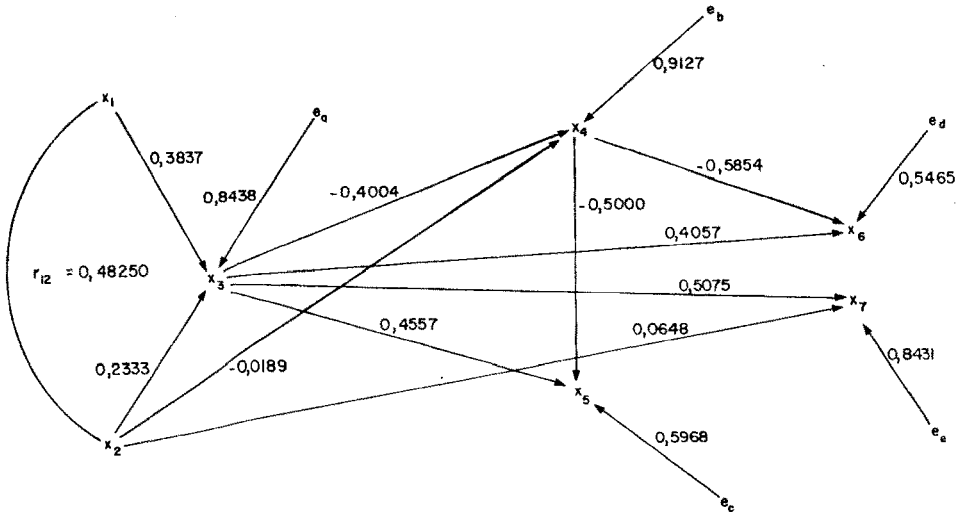


FIGURA 14

(5) EQUAÇÕES DE REGRESSÃO

As equações analíticas representativas do diagrama da figura 14 são,

$$x_3 = p_{31}\, x_1 + p_{32}\, x_2 + p_{3a}\, e_a$$

$$x_4 = p_{42}\, x_2 + p_{43}\, x_3 + p_{4b}\, e_b$$

$$x_5 = p_{53}\, x_3 + p_{54}\, x_4 + p_{5c}\, e_c$$

$$x_6 = p_{63}\, x_3 + p_{64}\, x_4 + p_{6d}\, e_d$$

$$x_7 = p_{72}\, x_2 + p_{73}\, x_3 + p_{7e}\, e_e$$

(6) MULTICOLINEARIDADE

A matriz de correlação das variáveis é mostrada na tabela 12.

TABELA 8

Matriz de correlação

VARIÁVEL	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1 Nível de Instrução	1	0,48250	0,49630	—0,56788	0,45027	0,54813	0,49411
X_2 Migração	0,48250	1	0,41846	—0,18643	0,36444	0,36992	0,27715
X_3 Rendimento Mensal da PEA	0,49630	0,41846	1	—0,40834	0,65989	0,64479	0,53458
X_4 Estrutura Etária	—0,56788	—0,18643	—0,40834	1	—0,68612	—0,75108	—0,79469
X_5 Nível de Instrução 2	0,45027	0,36444	0,65989	—0,68612	1	0,78934	0,76630
X_6 Consumo de Bens Duráveis	0,54813	0,36992	0,74479	—0,75108	0,78934	1	0,76461
X_7 Infra-Estrutura	0,49411	0,27715	0,53458	—0,79469	0,76630	0,74461	1

A matriz de correlação indica a inexistência de multicolinearidade grave nas variáveis independentes das equações de regressão do modelo.

(7) ANÁLISE DE DISPERSÃO

Os gráficos gerados pelo *Subprograma SPSS Scattergram* correspondentes a cada variável dependente versus as respectivas independentes não apresentaram afastamento da linearidade, não havendo, portanto, necessidade de transformação nas variáveis.

(8) ANÁLISE RESIDUAL

Os gráficos dos resíduos padronizados versus os valores estimados padronizados da variável dependente para cada equação de regressão, fornecidos pelo *Subprograma SPSS Regression,* não se caracterizaram por nenhum comportamento sistemático.

(9) RESULTADOS

Os coeficientes determinados a partir das equações de regressão são os constantes da tabela 9.

TABELA 9

Coeficientes das equações estruturais e de trajetória

RELAÇÃO BIVARIADA X_i X_j	COEFICIENTE ESTRUTURAL b_{ij}	COEFICIENTE DE TRAJETÓRIA P_{ij}
X_3 X_1	0,59975	0,38373
X_3 X_2	0,25179	0,23331
X_3 e_a	—	0,84376
X_4 X_2	—0,01232	—0,01886
X_4 X_3	—0,24234	—0,40045
X_4 e_b	—	0,91267
X_5 X_3	1,28652	0,45571
X_5 X_4	—2,33268	—0,50003
X_5 e_c	—	0,59683
X_6 X_3	0,33733	0,40575
X_6 X_4	—0,80421	—0,58539
X_6 e_d	—	0,54653
X_7 X_2	0,21806	0,06480
X_7 X_3	1,58233	0,50746
X_7 e_e	—	0,84307

Na tabela 10 apresenta-se a decomposição da variação total das variáveis dependentes.

TABELA 10

Decomposição da variação

VARIÁVEL DEPENDENTE	VARIAÇÃO					
	Determinada Diretamente pelos Fatores Imediatos		Determinada pelas Intercorrelações entre as Variáveis Independentes		Determinada pelo Resíduo	
	Termo	%	Termo	%	Termo	%
X_3	p^2_{31}	14,7				
	p^2_{32}	5,4				
	Total	20,1	$2p_{32}p_{31}r_{12}$	8,6	p^2_{3a}	71,3
X_4	p^2_{42}	0,0				
	p^2_{43}	16,0				
	Total	16,0	$2p_{42}p_{43}r_{23}$	0,6	p^2_{4b}	83,4
X_5	p^2_{53}	20,8				
	p^2_{54}	25,0				
	Total	45,8	$2p_{53}p_{54}r_{34}$	18,6	p^2_{5c}	35,6
X_6	p^2_{63}	16,5				
	p^2_{64}	34,2				
	Total	50,7	$2p_{63}p_{64}r_{34}$	19,4	p^2_{6d}	29,9
X_7	p^2_{72}	0,4				
	p^2_{73}	25,8				
	Total	26,2	$2p_{73}p_{72}r_{23}$	2,7	p^2_{7e}	71,1

A decomposição da variação total das variáveis dependentes mostra que, em alguns casos, a proporção de variação explicada pelo modelo foi bastante significativa: variáveis X_5 e X_6. Em outros, ao contrário, a variação determinada pelo resíduo é muito alta: variáveis X_3, X_4 e X_7. Uma análise mais detalhada da variação destas últimas indica que, nos três casos, e mais especialmente no das variáveis X_4 e X_7, esta situação deve-se ao fato de a migração explicar pouco ou quase nada da variação das variáveis em pauta, quando deveria fazê-lo segundo as hipóteses formuladas. Considere-se, neste particular, a fraqueza dos coeficientes de trajetória p_{43} e p_{72} (tabela 9).

As hipóteses referentes à migração foram, pois, mal formuladas. Em primeiro lugar, partiu-se do montante de migrantes, isto é, considerou-se apenas o movimento para dentro e não para fora, quando o mais correto teria sido considerar a migração líquida. Além disso, as hipóteses referentes à migração só poderiam ter sido testadas corretamente a partir da situação no tempo t + 1 de unidades de observação que tivessem, no que concerne às variáveis dependentes, uma situação semelhante no tempo t. Com efeito, a capacidade de atração de cada uma das aglomerações varia em função de determinadas características. Assim, por exemplo, a imigração para uma dada área poderia ser vista como ligada à capacidade efetiva ou percebida de seu setor produtivo em gerar empregos, e os rendimentos mensais da PEA, por sua vez, refletiriam a estrutura produtiva. Um aumento dos estratos de menor renda por efeito de migração não seria suficiente para colocar algumas aglomerações em situação de inferioridade em relação a outras aglomerações que tivessem de início uma estrutura mais desfavorável. Do mesmo modo, em uma certa medida, a melhor qualidade de infra-estrutura existente em uma dada área pode funcionar como fator de atração para a população e a análise dos efeitos de migração sobre a infra-estrutura deveria ter levado este fato em consideração.

A análise da tabela 11 também indicou que, exclusive as relações que envolviam a variável imigração (X_2), a decomposição das correlações nos efeitos direto, indiretos e totais, confirmou as hipóteses formuladas, permitindo, assim, uma quantificação destes efeitos.

11.3 — Decomposição de uma Variável Dependente [1]

Muitas das variáveis estudadas em uma investigação são ou podem ser interpretadas como compostas. Assim, o crescimento da população é a soma do crescimento natural e da migração líquida; cada uma das últimas variáveis pode ser decomposta, sendo o crescimento natural a diferença entre nascimento e morte, e a migração líquida a diferença entre migração para dentro e para fora. Onde for possível esta decomposição, é de interesse:

1 — determinar as contribuições relativas das componentes na variação da variável composta, e

2 — determinar como as causas que afetam a variável composta são transmitidas via suas respectivas componentes.

---

1 Otis D. Duncan, "Path Analysis: Sociological Examples", *Biometrics*, 1960, vol. 16, pp. 189-202.

TABELA 11

Decomposição da covariação bivariada

RELAÇÃO BIVARIADA X_i X_j	COVARIAÇÃO r_{ij}	CAUSA				EFEITO CAUSAL TOTAL $c_{ij} = p_{ij} + I$	IMPLÍCITO PELO MODELO RESTRITO $c_{ij} + CD + CI$ (A)	CORRELAÇÃO NÃO EXPLICADA PELO MODELO $r_{ij} - A$	NÃO CAUSAL CD + CI
		Direta p_{ij}	Indireta (I)	Comum					
				Direta (CD)	Indireta (CI)				
X_3 X_1	0,49630	0,38373	—	—	—	0,38373	0,38373	0,11257	—
X_3 X_2	0,41846	0,23331	—	—	—	0,23331	0,23331	0,18515	—
X_4 X_1	—0,56788	—	—0,15366	—	—	—0,15366	—0,15366	—0,41422	—
X_4 X_2	—0,18643	—0,01886	—0,09343	—	—	—0,11229	—0,11229	—0,07414	—
X_4 X_3	—0,40834	—0,40045	—	—0,00440	—	—0,40045	—0,40485	—0,00349	—0,00440
X_5 X_1	0,45027	—	0,25171	—	—	0,25171	0,25171	0,19856	—
X_5 X_2	0,36444	—	0,16247	—	—	0,16247	0,16247	0,20197	—
X_5 X_3	0,65989	0,45571	0,20024	—	—	0,65595	0,65955	0,00394	—
X_5 X_4	—0,68612	—0,50003	—	—0,18249	—	—0,50003	—0,68252	—0,00360	—0,18249
X_6 X_1	0,54813	—	0,24565	—	—	0,24565	0,24565	0,30248	—
X_6 X_2	0,36992	—	0,16040	—	—	0,16040	0,16040	0,20952	—
X_6 X_3	0,64479	0,40575	0,23442	—	0,00258	0,64017	0,64275	0,00204	0,00258
X_6 X_4	—0,75108	—0,58539	—	—0,16248	—	—0,58539	—0,74787	—0,00321	—0,16248
X_7 X_1	0,49411	—	0,19473	—	—	0,19473	0,19473	0,29938	—
X_7 X_2	0,27715	0,06480	0,11840	—	—	0,18320	0,18320	0,09395	—
X_7 X_3	0,53458	0,50746	—	0,01512	—	0,50746	0,52258	0,01200	0,01512

Um exemplo retirado do trabalho de Winsborough [1] ilustra o caso de componentes multiplicativas, transformadas por logaritmização em componentes aditivas. Winsborough notou, ao estudar a variação na densidade da população de sententa e quatro áreas da comunidade Chicago (exclusive o distrito central de negócios), que a densidade, definida como a razão entre a população e a área, pode ser expressa como:

$$\frac{\text{População}}{\text{Área}} = \frac{\text{População}}{\text{N.º de Domicílios}} \times \frac{\text{N.º de Domicílios}}{\text{N.º de Construções}} \times \frac{\text{N.º de Construções}}{\text{Área}}$$

Fazendo-se $X_0 = \log \frac{\text{População}}{\text{Área}}$

$$X_1 = \log \frac{\text{População}}{\text{N.º de Domicílios}}$$

$$X_2 = \log \frac{\text{N.º de Domicílios}}{\text{N.º de Construções}}$$

$$X_3 = \log \frac{\text{N.º de Construções}}{\text{Área}}$$

então

$$X_0 = X_1 + X_2 + X_3$$

Expressando-se estas variáveis na forma padrão temos,

$$X_0 = p_{01} \times x_1 + p_{02} \times x_2 + p_{03} \times x_3$$

onde

$p_{01} = 0{,}132$

$p_{02} = 0{,}468$

$p_{03} = 0{,}821$

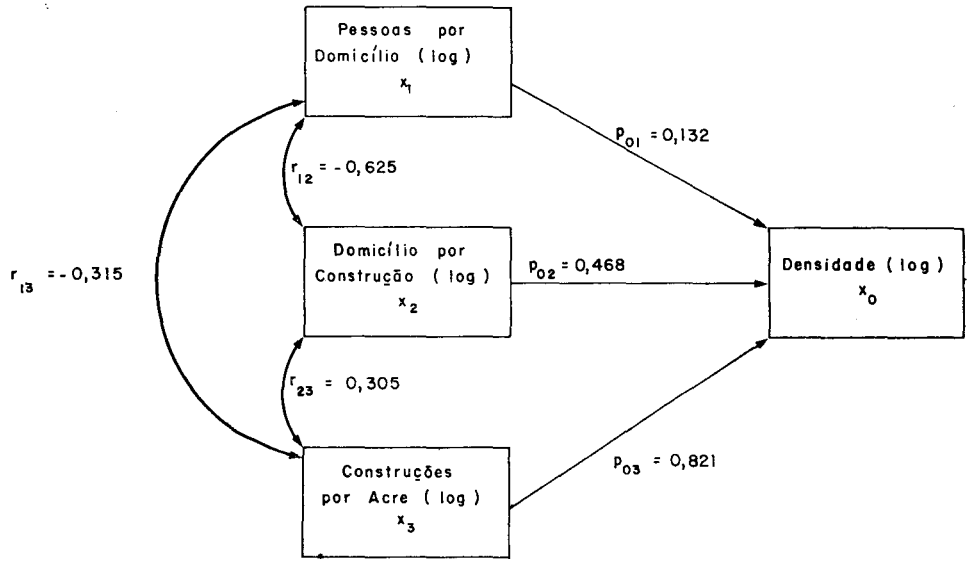
Matriz de correlação para os logaritmos da densidade e suas componentes, e duas variáveis indepedentes

Áreas da Comunidade de Chicago — 1940

VARIÁVEL	X_1	X_2	X_3	W	Z
X_0 densidade (log)	—0,419	0,636	0,923	—0,663	—0,390
X_1 pessoas por domicílio (log)	—	—0,625	—0,315	0,296	0,099
X_2 pessoas por construção (log)	—	—	0,305	—0,594	—0,466
X_3 construções por acre (log)	—	—	—	—0,517	—0,226
W distância do centro	—	—	—	—	0,549
Z crescimento recente	—	—	—	—	—

2 Hal H. Winsborough, "City Growth and City Structure", *Journal of Regional Science*, IV, (Winter, 1962), 35 — 39.

As intercorrelações das componentes expostas acima são usadas para completar o diagrama de trajetória abaixo.



As correlações da variável dependente com suas componentes podem ser calculadas a partir do Teorema Fundamental.

VARIÁVEL CAUSAL	CORRELAÇÃO TOTAL =	EFEITO DIRETO + EM X_0	CORRELAÇÃO DEVIDO A CAUSAS COMUNS E/OU CORRELACIONADAS
X_1	$r_{01} = -0,419$	$p_{01} = 0,132$	$p_{02}r_{12} + p_{03}r_{13} = -0,551$
X_2	$r_{02} = 0,636$	$p_{02} = 0,468$	$p_{01}r_{12} + p_{03}r_{23} = 0,168$
X_3	$r_{03} = 0,923$	$p_{03} = 0,821$	$p_{01}r_{13} + p_{02}r_{23} = 0,102$

A análise não só tornou visível a ordenação nas três componentes, em termos de sua importância relativa que é dada pelos coeficientes de trajetórias, mas mostrou também que uma das componentes é, na realidade, correlacionada negativamente com a variável composta, devido a sua correlação negativa com as outras duas componentes.

Winsborough considerou duas variáveis independentes como fatores produtores de variação na densidade: distância do centro da cidade e crescimento recente (percentagem de domicílios construídos em 1920 ou posteriormente). O diagrama pode ser elaborado a fim de indicar como operam estes fatores via as componentes do log da densidade.

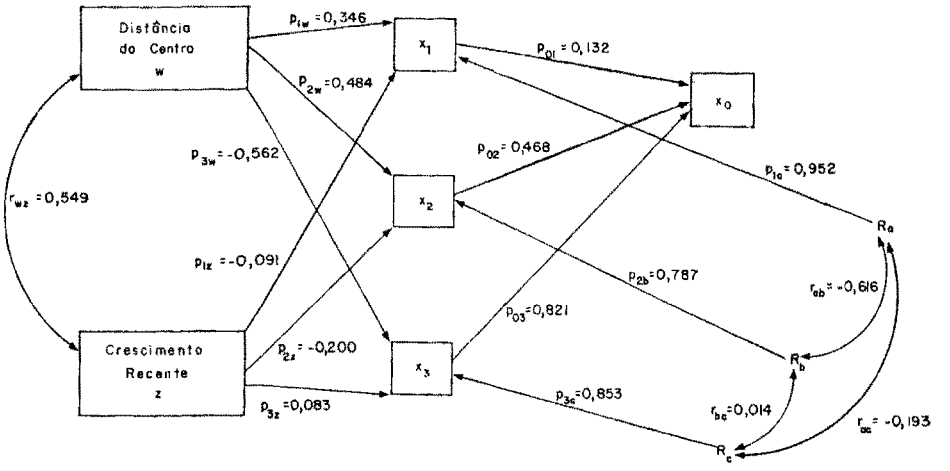
O primeiro passo é calcular os coeficientes de trajetória para as relações de cada componente com as duas variáveis independentes na forma padrão.

$$x_1 = p_1 w^w + p_1 z^z + p_{1a} R_a$$

$$x_2 = p_2 w^w + p_2 z^z + p_{2b} R_b$$

$$x_3 = p_3 w^w + p_3 z^z + p_{3c} R_c$$

Calculando-se as regressões acima, os coeficientes resultantes são colocados no diagrama que se segue.



É interessante que a distância w tem em cada componente da densidade efeitos nitidamente maiores que o crescimento recente Z. As duas variáveis não explicam a variação total em qualquer das componentes, como pode ser visto pelo tamanho dos resíduos p_{1a}, p_{2b} e p_{3c}. As correlações entre os resíduos nos fornecem outras informações importantes. São obtidas a partir do teorema básico, escrevendo-se, por exemplo,

$$r_{23} = p_{2w}r_{3w} + p_{2z}r_{3z} + p_{2b}p_{3c}r_{bc}$$

que resulta em $r_{bc} = 0{,}014$. Desta forma, as correlações entre os resíduos nada mais são que correlações parciais convencionais; assim $r_{ab} = r_{12.wz}$, $r_{ac} = r_{13.wz}$ e $r_{bc} = r_{23.wz}$. As correlações parciais, que geralmente tem pouca utilidade na análise de trajetória, tornam-se apropriadas quando o problema em questão é se um conjunto de variáveis independentes *explica* a correlação entre duas variáveis dependentes. No presente exemplo, enquanto $r_{23} = 0{,}305$, achamos que $r_{bc} = r_{23.wz} = 0{,}014$. Assim, a correlação entre os logaritmos dos domicílios por estrutura (X_2) e estrutura por acre (X_3) é satisfatoriamente explicada pelas relações respectivas destas duas componentes com a distância (w) e o crescimento recente (z). O mesmo não é verdade para as correlações que envolvem pessoas por domicílio (X_1), mas felizmente esta é a componente de menor importância na densidade.

Embora as correlações entre os resíduos sejam necessários para a complementação do diagrama e, em certo sentido, para avaliar a adequação das variáveis explicativas, estas correlações não entram nos cálculos que respondem a pergunta final. Como são transmitidos à variável dependente os efeitos das variáveis independentes (via suas componentes)? O efeito total de w é dado por

$$p_{01}p_{1w} + p_{02}p_{2w} + p_{03}p_{3w} = 0{,}046 - 0{,}226 - 0{,}461 = -0{,}641$$

e o z é dado por

$$p_{01}p_{1z} + p_{02}p_{2z} + p_{03}p_{3z} = -0{,}012 - 0{,}094 + 0{,}068 = -0{,}038$$

A densidade (X_0) é relacionada negativamente com a distância (w) e o crescimento recente (z), sendo os efeitos transmitidos via a primeira

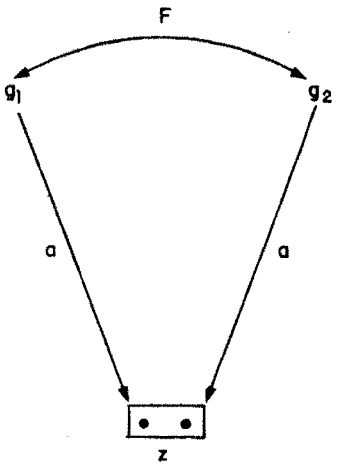
componente bem pequenos. A distância diminui a densidade em princípio via seu efeito intermediário nas estruturas por acre (X_3), em segundo via os domicílios por construção.

O problema da densidade serve como exemplo para uma estratégia geral raramente empregada na pesquisa: o desmembramento de uma variável complexa em suas componentes antes de iniciar uma pesquisa sobre suas causas. Contudo, deve-se evitar um erro extraordinário: dar a mesma base de tratamento para as componentes e causas. Por este caminho pode-se chegar ao resultado enganoso de que a migração líquida é a *causa* mais importante do crescimento populacional do que a mudança na produção industrial. Deve-se fazer forte objeção a um esquema causal construído na premissa: *se as variáveis econômicas e demográficas ajudam a explicar o crescimento urbano, então podemos obter melhor compreensão do processo de crescimento reunindo-as.* Pelo contrário, a compreensão exigiria uma distinção clara entre as COMPONENTES demográficas e as CAUSAS econômicas que pudessem afetar o crescimento via uma ou outra de suas componentes.

11.4 — Aplicação da Análise de Trajetória em Genética [1]

A análise de trajetória é utilizada em genética na determinação das várias relações existentes entre os pais e seus descendentes, e na determinaçção das relações existentes entre cruzamentos de indivíduos correlacionados.

De uma maneira sucinta, encontramos todos esses resultados em Li.[2]



z (zigoto) = g_1 (espermatozóide + g_2 (ovo)

1 Achcar, J. A. — *Análise de Trajetórias,* Inst. de Matemática e Estatística, Universidade de S. Paulo.

2 Li, C. C. — *Population Genetics,* (1955), pp. 174.

Apenas citando um exemplo, sem entrar em detalhes, sobre as relações existentes entre os pais e seus descendentes, que estão apresentados no livro acima citado, tomamos a trajetória do zigoto determinada pela união de dois gametas.

O zigoto é considerado como linearmente, completamente e igualmente determinado pela união de dois gametas. Esta trajetória pode ser representada pelo esquema causal simples abaixo.

Desde que z é igualmente determinado por g_1 e g_2 para genes *autossomais*, os dois coeficientes, um do ovo para o zigoto e o outro do espermatozóide para o zigoto, devem ser iguais. Usando a notação de Wright, seja *a* o coeficiente de trajetória do gameta (ovo ou espermatozóide) para o zigoto e seja F o coeficiente de correlação entre os gametas. Pela fórmula de determinação completa, temos,

$$r_{zz} = p^2_{zg_1} + p^2_{zg_2} + 2\ p_{zg_1}\ r_{g_1 g_2}\ p_{zg_2}$$

como

$$r_{zz} = 1$$

$$p_{zg_1} = p_{zg_2} = a,$$

e

$$r_{g_1 g_2} = F$$

então, substituindo estes valores na expressão anterior, temos que

$$1 = 2\ a^2 + 2\ a^2\ F$$

Assim,

$$a^2 = \frac{1}{2\ (1 + F)}$$

e

$$a = \sqrt{\frac{1}{2\ (1 + F)}}$$

Assim encontramos por meio de uma aplicação da análise de trajetória uma relação genética importante.

BIBLIOGRAFIA

ON-KIM, Jae & KOHOUT, Frank J. Path. Analysis and Causal Interpretation. In: —. *Statistical Package for the Social Sciences.* 2 ed., 1975. p. 383-97.

ON-KIM, Jae & KOHOUT, Frank J. Multiple Regression Analysis: Subprogram Regression. In: —. *Statistical Package for the Social Sciences.* 2 ed. 1975. p. 320-67.

TURNES, Malcoln E. & STEVENS, Charles D. The Regression Analysis of Causal Paths. *Biometrics,* v. 15, p. 236-58, 1959.

WRIGHT, Sewall. Path Coefficients and Path Regressions: Alternative or Complementary Concepts? *Biometrics,* v. 16, p. 189-202, 1960.

DUNCAN, Otis Dudley. Path Analysis: Sociological Examples. *American Journal of Sociology.* v. 72, p. 1-16, 1966.

DUNCAN, Otis Dudley. *Introduction to Structural Equation Models.* 1975.

BLALOCK JR., H. M. Causal Inferences, Closed Populations and Measures of Association. *American Political Science Review,* v. 61, p. 130-6, 1967.

BLALOCK JR., H. M. *Causal Inferences in Nonexperimental Research.* 1972.

BLALOCK JR., H. M. *Causal Models in the Social Sciences.* 1978.

HEISE, David R. *Causal Analysis.* 1975.

KENDAL, M. G. & O'MUIRCHEARTAIGH, C. A. Path Analysis and Model Building. *World Fertility Survey-Technical Bulletins.* 1977 (n. 2/ TECH. 414).

JOHNSTON, R. J. *Multivariate Statistical Analysis in Geography.* 1978.

ACHCAR, Jorge Alberto. *Análise de Trajetórias.* Instituto de Matemática e Estatística da Universidade de São Paulo. 1976. Dissertação para obtenção do grau de mestre.

DRAPER, N. R. & SMITH, H. *Applied Regression Analysis.* 1966.

JOHNSTON, J. *Métodos Econométricos.* 1963.

SUMMARY

This article tries to present, in an operational way, the so-called *Trajectory Analysis* method. We owe its bases to the geneticist Sewall Wright. The method has been rather discussed by other investigators since the first works on it, in 1918, what contributed very much to its improvement. Nowadays, the trajectory analysis is applied to several fields of human knowledge. In the geographical context this method appears as a powerful instrument for a realistic approach of varied systems, where the existence of causation is admitted. Nevertheless, Wright himself stated that the trajectory analysis "isn't restricted at all to the relations that can be described as those of cause and effect. It can be applied to purely mathematical linear systems and it is incorporated with the multiple regression method."

RÉSUMÉ

Cet article essaie de présenter, d'une manière opérationnelle, la méthode appelée *Analyse de la Trajectoire,* dont les fondements on doit au généticien Sewall Wright. La méthode a été discutée par les investigateurs dès la présentation des premiers travaux, en 1918, jusqu'aujourd'hui, ce qui a beacoup contribué pour son perfectionnement. L'analyse de la trajectoire est appliquée, actuellement, aux plus diverses branches de la connaissance humaine. Dans le contexte géographique cette méthode se présente comme un puissant instrument pour une perspective réaliste des systèmes multivariés, où l'on admit l'existence de la causation. Cependant, comme Wright lui-méme a affirmé, l'analyse de la trajectoire "ne se restreint pas du tout aux relations qui peuvent étre décrites comme de cause et effet. Elle peut étre appliquée aux systèmes linéaires uniquement mathématiques et elle se fond avec la méthode de régression multiple."