ANALISIS DE TRAYECTORIA (PATH ANALYSIS): APLICACION EN LA DEMOGRAFIA SOCIAL CON UN EJEMPLO QUE EMPLEA UN PROGRAMA DISPONIBLE EN CELADE

Johanna de Jong
Arthur M. Conning
CELADE

PATH ANALYSIS IN SOCIAL DEMOGRAPHY WITH AN EXAMPLE USING A PROGRAM EXISTING AT CELADE

SUMMARY

The path analysis of a data set presumes that the investigator has a model in which the causal chains among the variables have been explicitly defined. This article briefly outlines the basic definitions and the assumptions involved in determining the path coefficients of the investigator's model. An example is given of the use of a simple program written in APL for the CELADE computer terminal. The program calculates the path coefficients from an input matrix of zero order correlation coefficients among all the variables specified in the model. In the procedure utilized the resulting path coefficients are standardized regression coefficients.

1. Análisis de trayectoria (path analysis)

1. Con el análisis de trayectoria se analizan los efectos directos e indirectos lineales de unas variables sobre otras dentro de un sistema cerrado, basado en la teoría del investigador. Es una forma de análisis de regresión parcial en el que se usan los puntajes estándares. La ventaja de esto radica en que los coeficientes de todas las relaciones en el sistema son directamente comparables.

La relación entre el coeficiente de regresión parcial (no estandarizado) y el coeficiente de trayectoria (estandarizado) es:

$$P_{ij} = \frac{s_j}{s_i} \cdot c_{ij} \qquad (A)$$

i = variable dependiente
j = variable independiente
s = desviación estándar
c = coeficiente de regresión parcial
P_{ij} = coeficiente de trayectoria de la variable j a la variable i.

2. En el modelo se distinguen varios tipos de variables:

a) *Variables exógenas:* las variables que se suponen son independientes de las demás variables dentro del modelo. Entre las variables exógenas pueden existir correlaciones, pero éstas quedan sin analizar. En la figura 1, Z_1 y Z_2 son variables exógenas.

b) *Variables endógenas:* las que son dependientes de otras variables dentro del sistema. Se supone que toda la varianza de las variables endógenas está determinada por una combinación lineal de la varianza de las variables en el sistema, más las variables residuales. Las variables Z_3, Z_4 y Z_5 en la figura 1 son endógenas.

c) *Variables residuales:* las que no se miden sino que se introducen para explicar los residuos de la varianza de las variables endógenas no explicada por las demás variables dentro del sistema. El valor del coeficiente de trayectoria de la variable residuo r actuando sobre la variable i es, por lo tanto, $P_{ir} = \sqrt{1-R^2}$ donde R es el coeficiente de correlación múltiple de las variables hasta, e inclusive, i. R_a, R_b y R_c en la figura 1 son variables residuales.

3. La aplicación del método requiere de un número de suposiciones importantes; la mayor parte de ellas son las suposiciones de regresión:

a) Relaciones lineales entre las variables.

b) Efectos aditivos, es decir, no debe haber interacciones.

c) Variables medidas en escala de intervalo.

d) Cada variable residual no está relacionada con las demás, como tampoco con las otras variables dentro del modelo.

e) No hay multicolinealidad alta; es decir, no existen intercorrelaciones extremadamente altas entre las variables del modelo.

f) Con excepción de las variables exógenas, cada variable del modelo depende de variables causales anteriores, es decir, no hay trayectorias que regresan a variables anteriores (véase la figura 1). Generalmente, se presume que no existe retro-alimentación ('feedback') ni directamente entre dos variables, como tampoco indirectamente por intermedio de una cadena.

g) Los datos empíricos deben ser medidos con gran confiabilidad.

Para una discusión más detallada sobre las suposiciones y los efectos de su inobservancia, véase Duncan (1966), Land (1969:32ff) y Heise (1969:44ff).

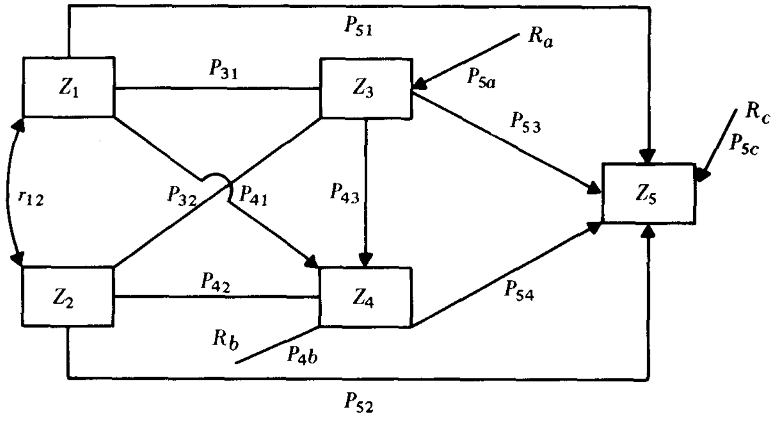
4. En forma gráfica, el diagrama de trayectoria a que nos referimos tiene la forma que se presenta en la figura 1.

5. Entre las variables exógenas se mantienen los coeficientes de correlación porque se las considera como causantes y no influidas por ninguna de las variables endógenas. Se supone que las demás relaciones originales, medidas con coeficientes de correlación, cambian porque están recalculadas tomando en cuenta los demás caminos que influyen en la variable endógena. Ellas se expresan en los coeficientes de trayectoria mencionados.

6. Se puede definir el coeficiente de trayectoria (Reichwein, 1971, citando a Wright, 1934) como el número que indica la fracción de la desviación estándar de una variable dependiente que está explicada en forma directa por la variación de una desviación estándar de la variable independiente. Esta fracción se encontraría si el factor variara en la misma medida de los datos observados, manteniéndose constantes las

Figura 1

MODELO CAUSAL CON TODAS LAS TRAYECTORIAS POSIBLES



en que:

Z_1 y Z_2 = variables exógenas

Z_3, Z_4, Z_5 = variables endógenas

R_a, R_b, R_c = variables residuales

$P_{51}, P_{52}, P_{53}, P_{54}, P_{41}, P_{42}, P_{43}, P_{32}, P_{31}, P_{3a}, P_{4b}, P_{5c}$ = coeficientes de trayectoria

r_{12} = coeficiente de correlación.

demás variables relevantes, incluso las variables residuales. El método normalmente usado para solucionar las ecuaciones del modelo da coeficientes de trayectoria que son iguales a coeficientes beta en una regresión, es decir, a coeficientes de regresión estandarizados.

7. Esta definición se ve reflejada en el teorema básico del análisis de trayectoria, que se puede expresar en la siguiente fórmula:

$$r_{ij} = \sum_{q=1}^{n} P_{iq} \cdot r_{jq} \qquad (B)$$

en que:

r	=	coeficiente de correlación
i	=	variable dependiente
j	=	variable independiente
p	=	coeficiente de trayectoria
$q, ..., n$	=	variables que tienen una trayectoria directa a Z_i.

Esta fórmula muestra que la correlación entre dos variables está compuesta de una serie de efectos directos ($\sum_{q=1}^{n} P_{iq}$) de las variables intervinientes y una serie correspondiente de correlaciones entre la variable más independiente y estas variables intervinientes. Además, muestra que cada coeficiente de trayectoria está calculado tomando en cuenta los demás caminos por los cuales la variable más independiente influye en la variable dependiente.

8. Se advierte a los usuarios del análisis de trayectoria que la validez del modelo usado depende de la teoría empleada. Con excepción de algunos casos, los cálculos matemáticos del análisis de trayectoria no dan una manera definitiva de seleccionar uno de varios posibles ordenamientos de las variables. Por este motivo, Heise (1969:51-52), declara que una suposición del análisis de trayectoria es que las prioridades causales entre las variables están lo suficientemente establecidas como para estar *fuera de discusión.*

Hasta cierto punto, sin embargo, es posible usar los resultados del análisis de trayectoria dado un ordenamiento causal de las variables, para así eliminar empíricamente algunas de las trayectorias que tienen coeficientes "esencialmente de cero" (Heise, 1969:59-64).

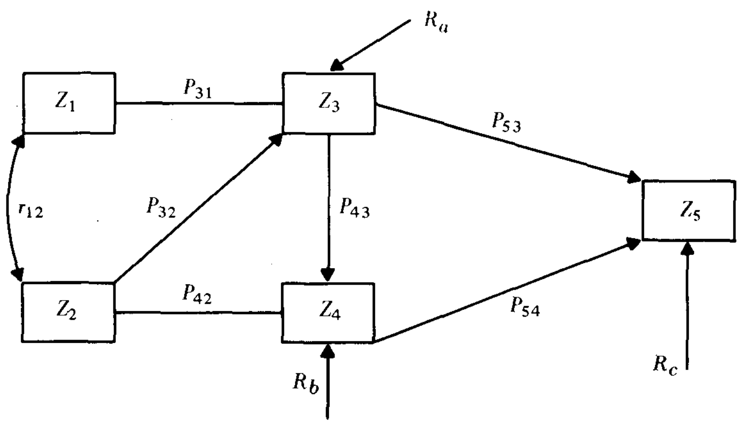
2. *Un ejemplo calculado con el programa de CELADE*

9. CELADE tiene un programa escrito en el lenguaje APL para su terminal de computación. Este programa requiere como entrada la matriz de correlación de orden cero entre todas las variables especificadas en el modelo, y usa el teorema básico para determinar una serie de ecuaciones requeridas por el coeficiente de trayectoria.

10. Para ilustrar el uso del programa, consideramos el modelo de la figura 2; (el ejemplo y el programa de computación están basados, en gran parte, en un artículo de Nygreen, 1971).

Figura 2

MODELO CAUSAL EN QUE SOLAMENTE SE CONSIDERAN LAS RELACIONES POSTULADAS



11. Este modelo es similar a la figura 1, pero el investigador ha eliminado ciertas posibles trayectorias tales como la que va entre Z_1 y Z_4 porque cree que no existen razones teóricas para incluirlas. En efecto, esto implica que el coeficiente de trayectoria P_{41} (etc.), es considerado como igual a cero.
12. La matriz de correlación que se usa aquí, calculada tal vez usando la rutina de correlación del programa SPSS (también disponible en CELADE), se muestra en el cuadro 1.

Cuadro 1

MATRIZ DE CORRELACION

	Z_1	Z_2	Z_3	Z_4	Z_5
Z_1	1,00	0,12	0,41	0,42	0,29
Z_2	0,12	1,00	0,35	0,15	0,31
Z_3	0,41	0,35	1,00	0,56	0,52
Z_4	0,42	0,15	0,56	1,00	0,49
Z_5	0,29	0,31	0,52	0,49	1,00

13. Aunque en la práctica el investigador simplemente provee al terminal con una especificación del modelo y la matriz de correlación, y recibe casi inmediatamente todos los coeficientes de trayectoria, el ejemplo dado se usará para ilustrar cómo se obtienen los coeficientes.
14. Suponiendo que se desea obtener los coeficientes de trayectoria que entren directamente en Z_5, se puede definir una serie de ecuaciones usando el teorema básico con $i = 5$.

Una ecuación de la serie puede ser con $j = 1$:

$$r_{51} = P_{51}\, r_{11} + P_{52}\, r_{21} + P_{53}\, r_{31} + P_{54}\, r_{41} + P_{5a}\, r_{a1}. \qquad (C)$$

pero:

$P_{51} = 0$ y $P_{52} = 0$ de acuerdo con el modelo de la figura 2, porque no hay trayectorias directas a Z_5 desde Z_1 y Z_2

y:

$r_{a1} = 0$ porque se presume (véase el párrafo 3) que no existe correlación entre el residuo y las variables del modelo. Nótese que por este motivo en el teorema básico, q (cuando $i \neq j$) normalmente no es sacado de las variables residuales.

De ahí que *(C)* se transforma en:

$$r_{51} = P_{53}\, r_{31} + P_{54}\, r_{41} \qquad (D)$$

De manera similar las otras ecuaciones se pueden definir para Z_5 dando una serie completa:

$$\begin{aligned} r_{51} &= P_{53}\, r_{31} + P_{54}\, r_{41} \\ r_{52} &= P_{53}\, r_{32} + P_{54}\, r_{42} \qquad (E) \\ r_{53} &= P_{53}\, r_{33} + P_{54}\, r_{43} \\ r_{54} &= P_{53}\, r_{34} + P_{54}\, r_{44} \end{aligned}$$

Por supuesto, r_{33} y r_{44} son iguales a 1.

15. Esta serie de ecuaciones *(E)* debe ser usada para resolver los coeficientes de trayectoria P_{53} y P_{54}. Pero esto requiere solamente dos ecuaciones y nosotros tenemos cuatro. Esto se conoce como "sobreidentificación" ya que cada uno de los posibles seis pares de ecuaciones pueden proporcionar valores únicos de estos dos coeficientes de trayectoria y las soluciones individuales pueden no ser las mismas.
16. La solución al problema de "sobreidentificación" está en usar las dos ecuaciones en las cuales los coeficientes de trayectoria tienen los mismos subíndices que los coeficientes de correlación porque ésta da los coeficientes de trayectoria como definidos en el párrafo 1 (coeficientes de regresión estandarizados). En el ejemplo:

$$\begin{aligned} r_{53} &= P_{53}\, r_{33} + P_{54}\, r_{43} \\ r_{54} &= P_{53}\, r_{34} + P_{54}\, r_{44} \end{aligned} \qquad (F)$$

Listado 1

ENTRADA Y SALIDA DEL PROGRAMA ESCRITO
EN APL CON DATOS Y EJEMPLO DEL TEXTO

```
JCOPY  CELAEJT  IPA
M ←5   5p.01   x100   12   41   42   29   12   100   35   15   31
41   35   100   56   52   42   15   56   100   49   29   31   52  49
100
M

1         0,12      0,41      0,42      0,29
0,12      1         0,35      0,15      0,31
0,41      0,35      1         0,56      0,52
0,42      0,15      0,56      1         0,49
0,29      0,31      0,52      0,49      1
IPA M
INGRESE N° DE ORDEN DE:
VARIABLES EXOGENAS
□ :
   1  2
VARIABLES CAUSALES DE:
VARIABLE 3
□ :
   1  2
VARIABLE 4
□ :
   2  3
VARIABLE 5
□ :
   3  4
COEFICIENTES DE TRAYECTORIA
( ↓ )  CAUSAS/EFECTOS  ( → )
             3          4            5
         1   0,373377   0            0
         2   0,305195   - 0,0524217  0
         3   0          0,578348     0,357809
         4   0          0            0,289627
         5   0          0            0
Residual 0   0,860289   0,827036     0,81977
```

o con las correlaciones obtenidas de la matriz de correlación en el cuadro 1:

$$\begin{aligned} 0{,}52 &= P_{53} + 0{,}56\,(P_{54}) \\ 0{,}49 &= 0{,}56\,(P_{53}) + P_{54} \end{aligned} \qquad (G)$$

17. Estas dos ecuaciones se pueden solucionar fácilmente en forma algebraica para P_{53} y P_{54} pero, naturalmente, en éste y en casos más complejos, el programa de computación dará los resultados de inmediato y calculados en forma más precisa. Los valores de estos dos coeficientes de correlación, como todos los demás en el modelo, se dan en el listado 1, que muestra la entrada y la salida del programa. Por ejemplo, $P_{32} = 0{,}305195$.

18. El programa proporciona, además, los coeficientes de trayectoria residuales que se calculan con la ecuación básica. Esto se realiza de la siguiente manera:

Como la correlación entre cualquier variable y ella misma es 1,0, la ecuación básica para r_{ii} es:

$$r_{ii} = 1{,}0 = \left(\sum_q P_{iq}\, r_{qi}\right) + P_{ia}\, r_{ai} \qquad (H)$$

pero:

$P_{ia} = r_{ai}$ porque se presume que la variable residual R_a no está correlacionada con las demás variables del modelo. De ahí que el

Listado 2

PROGRAMA APL

```
        ∇ IPA [□] ∇
    ∇   IPA M; E; A; F; J; P; B; V
[1]     'INGRESE N° DE ORDEN DE: '  ; SS; 'VARIABLES EXOGENAS'
[2]     P←(2 1 +ρM)ρI←0
[3]     E ←(∧/[1]((ρ, E), F)ρ, (E ←□)∘ . F A)/A ←
        F ←(ρM)[1]
[4]     'VARIABLES CAUSALES DE: '
[5]   I 1: 'VARIABLE'; J ←E [I←I + 1]
[6]     P[2 +F; 1 +J]←(1 - +/(P[1 +B; 1 +J]←∇⌹ M[B;B])
        × V←M[B←□; J]) * 0.5
[7]     →(I<ρE)/I1
[8]     P[1 + A; 1]← P[1; 1 + A]←A
[9]     SS; ' COEFICIENTES DE TRAYECTORIA '; SS; ' (↓) CAUSAS\
        EFECTOS (→)'; SS; P[ ; 1, E + 1]
    ∇
```

coeficiente de regresión parcial estandarizado, P_{ia}, sea el mismo que el coeficiente de correlación r_{ai}.

Por consiguiente:

$$1{,}0 = (\sum_q P_{iq} \, r_{qi}) + (P_{ia})^2 \qquad (1)$$

o:

$$(P_{ia})^2 = 1{,}0 - (\sum_q P_{iq} \, r_{qi})$$

El programa APL sustituye los valores de las correlaciones de la matriz y los coeficientes de trayectoria recientemente calculados, a fin de poder calcular los coeficientes de trayectoria residuales. En el listado 1, estos son: $P_{3a} = 0{,}860289$; $P_{4b} = 0{,}027036$; y $P_{5c} = 0{,}81977$.

BIBLIOGRAFIA

(1) Duncan, O.D., "Path Analysis: Sociological Examples", en *American Journal of Sociology*, 72, 1966, págs. 3-16.

(2) Heise, David R., "Problems in Path Analysis and Causal Inference", en *Sociological Methodology*, editado por E.F. Borgatta, Jossey-Bass, San Francisco, 1969, págs. 38-73.

(3) Land, Kenneth C., "Principles of Path Analysis", en *Sociological Methodology*, editado por E.F. Borgatta, Jossey-Bass, San Francisco, 1969, págs. 3-37.

(4) Nygreen, G.I., "Interactive Path Analysis", en *The American Sociologist*, Vol. 6, 1971, págs. 37-43.

(5) Reichwein, Ch.L.M., *Pad analyse in de sociale wetenschappen*, Utrecht, Holanda, 1971.

(6) Wright, S., "The Method of Path Coefficient", en *Annals of Mathematical Statistics*, 5, 1934, págs. 161-215.