ANÁLISIS DE RUTAS EN BIOLOGÍA: ESTADÍSTICA PARA SISTEMAS MULTICAUSALES

RAÚL ORTIZ-PULIDO

El estudio de los sistemas biológicos no es fácil de abordar debido a que diversos factores, internos y externos, los afectan. Por ejemplo, tratar de modelar interacciones entre especies es difícil ya que estas relaciones se llevan a cabo en ambientes que varían de manera temporal y espacial. Si una interacción entre dos especies es clasificada como mutualista en un ambiente, la relación de las mismas especies puede ser catalogada como parasitismo bajo otras condiciones ambientales (Thompson, 1994; 1997).

A pesar de esta complejidad en la naturaleza, nuestra experiencia nos ha indicado que es posible conocerla, que en ella existe causalidad (es decir, unas cosas influyen a otras), que las reglas por las que se rige no son muchas, y que no todas las condiciones originales influyen en cómo se desarrolla un fenómeno (comúnmente se recomienda tomar en cuenta sólo aquellas variables que tienen un efecto notorio en el fenómeno estudiado). Si consideramos válidos estos cuatro puntos, podemos explorar y modelar cómo funciona la naturaleza.

En este trabajo se describe una técnica estadística de exploración y modelación, el análisis de rutas. Por exploración me refiero a la acción de estudiar un sistema con la intención de describirlo y establecer hipótesis de su funcionamiento (hipótesis de causalidades entre variables). Por modelación me refiero a la acción de crear un esquema teórico (matemático) sobre el funcionamiento de un sistema. En el análisis de rutas este esquema se elabora para facilitar la comprensión y estudio del comportamiento del sistema analizado. Al explorar un sistema con análisis de rutas establecemos una hipótesis de causalidades que puede ser puesta a prueba de manera experimental. Al modelar un sistema con análisis de rutas probamos esa hipótesis.

Este artículo está estructurado en cuatro secciones. En la primera trataré sobre la historia de la técnica; en la segunda pretendo demostrar que el análisis de rutas tiene ventajas sobre otras técnicas de exploración y modelación cuando se busca encontrar patrones en sistemas que consideran el transcurso del tiempo entre fenómenos relacionados; en la tercera sección describo cómo realizar esta técnica, con ejercicios prácticos y un ejemplo; y en la cuarta presento otras herramientas que permiten obtener una mayor versatilidad al utilizar el análisis de rutas.

No indicaremos cómo utilizar variables categóricas en el análisis de rutas, porque todavía está en discusión la validez de su uso en este tipo de análisis (Petraitis *et al.*, 1996). Este tipo de variables son comunes en biología y son resultado común de la manipulación experimental. Ejemplos son los experimentos que consideran sitios con y sin depredadores, o con y sin perturbación (Wooton, 1994), o los experimentos con niveles de tratamiento fijos. Su uso puede ser inadecuado en análisis de rutas porque es posible que inflen los estimados de los coeficientes de ruta, pues los intervalos de categorías y tratamientos son mucho más grandes que una desviación estándar (Petraitis *et al.*, 1996).

Hacer un análisis de rutas no es fácil, y menos para los hispanoparlantes, pues prácticamente no existen escritos en nuestro idioma sobre este tema. Además, una gran cantidad de estudios publicados en revistas de prestigio internacional han utilizado esta técnica de manera incorrecta (Petraitis *et al.*, 1996).

El objetivo de este trabajo es ofrecer una primera aproximación a la técnica de análisis de rutas al estudiante de sistemas biológicos, ahorrando horas de búsqueda y traducción. Puede ser una pieza útil para estructurar la investigación (*e.g.*, forma de tomar y ordenar los datos) y podría ser de ayuda a cualquier interesado en estudiar sistemas que tienen variables interconectadas en el tiempo.

Historia

El análisis de ruta fue originalmente desarrollado por Sewall Wright en 1934 como una manera para establecer las relaciones entre diversas variables de un sistema biológico (de manera formal debe decirse que con esta herramienta se intentaba interpretar,

PALABRAS CLAVE / Análisis de Rutas / Análisis de Senderos / Correlación y Regresión Múltiple / Análisis de Sistemas Biológicos Complejos /

Raúl Ortiz-Pulido**. Licenciado en Ecología, Universidad Veracruzana, México. Realiza Doctorado en Instituto de Ecología A. C., México. Editor Jefe de Huitzil, Revista de Ornitología Mexicana. Dirección: Departamento de Ecología Vegetal, Instituto de Ecología, A.C, Apartado 63, Xalapa, Veracruz, 91000, México. e-mail: ortizrau@ecología.edu.mx**

cuantitativamente, las relaciones causales en un sistema de variables correlacionadas). Con esta técnica Wright intentó proponer un "método más flexible" que los existentes hasta el momento (*e.g.*, regresión).

El análisis de ruta fue utilizado inicialmente en las ciencias sociales, pero actualmente se usa con frecuencia en las ciencias naturales. El análisis de ruta se realizaba comúnmente considerando un modelo *a priori* sobre el sistema que se pretendía estudiar. Como veremos más adelante, gracias al desarrollo de las computadoras, considerar un modelo *a priori* ya no es imprescindible, pero sí recomendable. El no necesitar un modelo *a priori* es de gran ayuda, porque los investigadores de sistemas biológicos frecuentemente enfrentamos situaciones en las cuales ni la teoría ni la biología básica están desarrolladas lo suficiente como para permitirnos definir claramente las relaciones causales entre cada uno de los pares de variables en consideración (Shipley, 1997). El uso de nuevos algoritmos, computadoras y programas de cómputo nos permite encontrar modelos viables estadísticamente, lo que es inicialmente de gran ayuda. El papel del biólogo es, en este caso, evaluar si las hipótesis estadísticamente viables tienen sentido biológico. Un ejemplo claro y exitoso donde se usó una herramienta estadística para descubrir las relaciones causales precisas entre dos variables, es el de la relación entre el tamaño del cuerpo y el gasto metabólico basal en animales (Shipley, 1997). A pesar de ello aún se sugiere que el investigador que usa análisis de rutas tenga un conocimiento previo del sistema que estudiará, para que pueda proponer las variables relevantes a considerar en un sistema, y pueda evaluar la factibilidad biológica de las relaciones predichas por los modelos estadísticos entre esas variables.

Al principio también se sugirió que una regla para realizar un análisis de ruta era tener, al menos, de 10 a 20 observaciones por cada variable considerada en el modelo original. Así, si se tenía un modelo donde se consideraban cinco variables era bueno tener de 50 a 100 observaciones para cada variable (equivalente a medir las características de 50 a 100 individuos de una población). Actualmente, gracias al desarrollo de técnicas de iteración (definidas como un re-muestreo de los mismos datos), esto ya no es imprescindible.

La técnica de análisis de rutas, también llamada análisis de senderos, fue revisada y mejorada por el mismo Wright (1968), Li (1975), Pedhazur (1982), Kingsolver y Schemske (1991), Mitchell (1992), Petraitis *et al.* (1996) y Shipley (1997). En español sólo conozco un escrito de circulación limitada (Parra, 1995) y otro no publicado de M. H. Reyes, de la Universidad Autónoma Agraria "Antonio Narro", México. Algunos de los programas de computo actualmente disponibles para realizar todos o varios de los pasos de un análisis de rutas son: AMOS, CALIS, COSAN, EPA, EQS, LINCS, LISCOMP, LISREL, MECOSA, Mx, PLS, RAM, RAMONA, SEPATH, STREAMS y TETRAD II. Una versión gratis de varios de estos programas (*e.g.*, AMOS, EPA y LISREL) puede ser obtenida de Internet.

Comparación con otras herramientas estadísticas

Las técnicas estadísticas "comunes" generalmente consideran el efecto o relación de una o diversas variables causales (independientes) sobre otra que depende de éstas (dependiente). Ejemplos de estas herramientas son la correlación y la regresión simple y múltiple. Cuando se considera la temporalidad en los sistemas, el análisis de rutas da frecuentemente mejores resultados que otros análisis porque:

- Permite construir rutas causales (*i.e.*, efectos entre las variables o hipótesis de causalidad) con un *x* número de variables independientes y dependientes.
- Su estructura no es fija, pues permite que una variable dependiente (efecto) se convierta en una variable independiente (causa) y viceversa.
- Provee un medio para descomponer la correlación entre dos variables cualesquiera (independiente y dependiente) en componentes que representen contribuciones causales y no-causales (Kingsolver y Schemske 1991). Las contribuciones causales incluyen dos elementos, llamados efectos directo e indirecto, y las contribuciones no-causales otros dos, conocidos como efecto azaroso y no analizado (Kingsolver y Schemke, 1991). Estos últimos elementos no se miden, pero se consideran en el modelo.
- Facilita la aproximación a los sistemas biológicos considerando la secuencia temporal que estos presentan.

Supuestos. El análisis de rutas supone (sensu Kingsolver y Schemske, 1991) que:

- Las relaciones entre las variables son lineares, aditivas y causales.
- Los residuales de cada variable no están correlacionados con las variables que la preceden en el modelo.
- No hay causación recíproca.
- Las variables son medidas sobre una escala de intervalos.
- Las variables son medidas sin error.

Debido a que en biología es común que muchas relaciones entre variables no sean del tipo mencionado, el usuario del análisis de rutas debe verificar que sus variables tengan este tipo de relaciones entre sí. Si no es el caso, el usuario puede emplear las técnicas utilizadas en regresión para intentar lograrlo (*e.g.*, transformación, eliminación de variables para evitar colinearidad). En este sentido, las limitantes del análisis de rutas son las mismas que se aplican a los análisis de regresión.

Fundamentos. Las bases de la técnica de análisis de ruta están dentro de las desarrolladas para los métodos de modelación estructural de ecuaciones (SEM, por sus siglas en inglés). Los métodos de regresión múltiple y análisis de factores también pertenecen a SEM (Rigdon, 1998). Al contrario que la regresión múltiple, el análisis de rutas permite la partición de las correlaciones causales y no causales de cualquier par de variables (Parra, 1995).

Cómo realizar un análisis de rutas

Construcción de un diagrama de flujo. En el diagrama se representan las relaciones entre las variables del sistema considerado (ver Figura 1). Para su desarrollo debe considerarse dos aspectos principales:

a) El diagrama debe ser ordenado temporalmente, existirán variables "tempranas" que influirán sobre variables "tardías" (*e.g.*, en la Figura 1, A afecta a B y ésta a Y). La relación entre cada variable debe estar determinada por una hipótesis biológica viable.

b) El sistema tiene variables desconocidas, y por lo tanto no medibles, que lo afectan (U en la Figura 1).

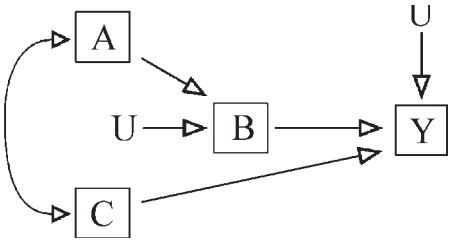

Figura 1. Diagrama de flujo para el análisis de rutas.

Gráficamente las variables están encerradas dentro de cajas que están conectadas por líneas rectas y curvas. Las líneas rectas (*e.g.*, entre B e Y) indicarán una relación causal entre las dos variables conectadas (en este caso B afecta a Y). Las líneas curvas (*e.g.*, entre

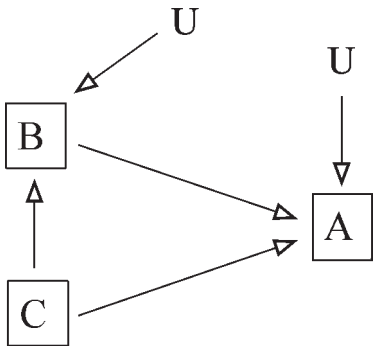

Figura 2. Diagrama de flujo representando las relaciones hipotéticas entre las variables A, B y C de la Tabla I.

A y C) indicarán una correlación sin causamiento entre las dos variables (*i.e.*, donde no se sabe con certeza cual de las dos variables es el efecto y cual es la causa o si existe una tercera variable que afecta a las dos). Una flecha al final de la línea indicará cual de las dos variables es la dependiente y cual la independiente (o la que causa a la dependiente). En el caso de las líneas curvas, debido a que ambas variables se están afectando, habrá flechas en ambos extremos de la línea. En cada variable que se considere como dependiente en el modelo debe haber una línea con una flecha apuntando a ella. La letra U en un extremo de la línea indicará variables que no fueron medidas que determinan la varianza de los datos no explicada por el modelo en ese punto.

Toma de datos en campo. Como en el análisis de rutas se está considerando la evaluación de un proceso temporal, se deben tomar los datos en unidades bien definidas (e.g., individuos, poblaciones, objetos). Es común que estas unidades sean marcadas porque en cada una se realizan medidas a través del tiempo. Por ejemplo, si se pretende evaluar la producción de frutos en una especie de planta considerando producción de flores, características de las flores y visita de polinizadores, los individuos medidos deben ser marcados, pues en cada uno de ellos mediremos las características de sus flores, el número de visitas de polinizadores y su producción de frutos a través del tiempo. Otro ejemplo, si queremos conocer los patrones de ovoposición de las hembras de una especie dada, dependiendo de las características climáticas y ambientales, las hembras deberán ser marcadas (pues seguramente en ellas se medirá tamaño o peso, número de huevos producidos y sitio en que cada hembra deposita los huevos, así como el dato de qué hábitat y macho "usa").

Una muestra mayor a 20 individuos es conveniente para realizar un análisis de rutas. Por ejemplo, en el estudio de producción de frutos deberán marcarse cuando menos 20 plantas. Otro punto recomendable es tratar, en lo posible, de obtener todos los datos completos para cada unidad marcada (es decir, que no falte medir el tamaño de las flores de un individuo) pues no hacerlo así ocasionaría que todos los datos de esa unidad sean casi inservibles para el análisis.

Preparación de los datos. Es necesario estandarizar los datos obtenidos en campo para evitar problemas de magnitud entre las variables consideradas. Cada dato en cada variable puede ser estandarizado según la siguiente fórmula:

Valor estandarizado = $(x_i - \bar{x}) / d.e.$

Donde: x_i = dato original en la variable

$\bar{x}$ = media de los datos en la variable considerada

d.e. = desviación estándar de los datos en la variable considerada

Ejemplo: Se desea estudiar un sistema biológico y después de observar dicho sistema se cree que el modelo que lo puede definir está representado gráficamente por el diagrama de flujo presentado en la Figura 2. En el campo obtenemos los cinco valores representados en la Tabla I para cada variable (presento cinco por simplicidad, pues ya se indicó que se necesitan cuando menos 20).

TABLA I
VALORES OBTENIDOS

# Dato	C	B	A
1	36	5,6	101
2	41	7,8	120
3	62	9,3	153
4	43	7,9	117
5	45	8,0	180
$\bar{x}$	45,4	7,72	134,2
d.e.	8,82	1,19	28,56

El valor estandarizado de C1 (primer valor de la columna C en la Tabla I) es:

C1 estandarizado = (36 – 45,4) / 8,82 = -1,06.

Si hacemos este procedimiento para todos los datos, los valores estandarizados provenientes de la Tabla I quedan como se observa en el Tabla II.

TABLA II
VALORES ESTANDARIZADOS

# Dato	C	B	A
1	-1,06	-1,77	-1,16
2	-0,49	0,06	-0,49
3	1,88	1,32	0,66
4	-0,27	0,15	-0,60
5	-0,04	0,23	1,60

Utilizando la estandarización de los datos obtenemos variables con media = 0 y d.e. = 1. Para más información sobre la ecuación utilizada en esta sección puede consultarse Kingsolver y Schemske (1991) y Petraitis *et al.* (1996).

Obtención de la influencia de una variable sobre otra. El siguiente paso es conocer la influencia cuantitativa entre las variables, denominada "coeficiente de ruta" y representada con un valor numérico. Este valor indicará la intensidad con que una variable independiente afecta a otra dependiente en nuestro modelo. Este valor será colocado junto a la línea que conecta dos variables en el modelo de flujo (*e.g.*, números en la Figura 3A). Es común que en vez de números se coloquen líneas de diferente grosor (como en 3B), que dependerá del valor del "coeficiente de ruta", lo que permite que el diagrama pueda ser evaluado visualmente con facilidad. En caso de que una variable afecte negativamente a otra (*i.e.*, que entre ellas exista un "coeficiente de ruta" negativo) esto puede

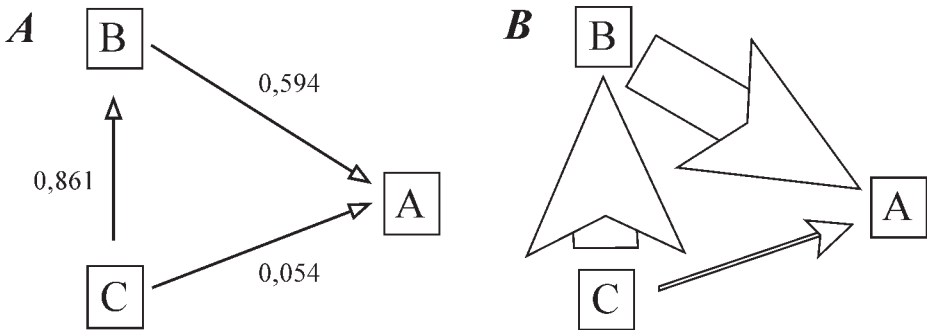

Figura 3. Diagrama de flujo con coeficientes de ruta indicados por números (A) o líneas (B).

ser representado gráficamente con una línea discontinua.

Los "coeficientes de ruta" son iguales a los "coeficientes parciales estandarizados de la regresión" (*i.e.*, a las pendientes de las rectas de regresión estandarizadas) cuando se realiza una regresión múltiple con las variables estandarizadas (*i.e.*, convirtiendo cada valor de las variables interactuantes a su valor de z [z-scores]). Estos valores son fácilmente obtenidos en muchos programas estadísticos de computadora. La forma específica de hacerlo puede ser consultada en Zar (1996).

Ejemplo: Consideramos el sistema representado en la Figura 2 y los datos estandarizados en la Tabla II para obtener los coeficientes de ruta para este sistema. Realizando una regresión múltiple para cada variable dependiente (en este caso la variable B, porque es determinada por la influencia de C, y la variable A, porque es determinada por C y B) obtenemos los coeficientes de la Figura 3.

Para obtener el efecto de dos variables que tienen correlación sin causamiento (efecto entre A y C en la Figura 1), se puede seguir el mismo procedimiento indicado arriba, pero en vez de regresión múltiple se realiza una correlación y se toma el "coeficiente de correlación" (comúnmente representado por "r") como el "coeficiente de ruta" entre las dos variables.

En un sistema en que tres o más variables tienen correlación sin causamiento se recomienda realizar una correlación parcial para obtener el "coeficiente de ruta" para cada línea curva. El "coeficiente de ruta" en este caso será el "coeficiente de correlación parcial" obtenido para cada interacción entre dos variables (Zar, 1996).

Para conocer el efecto que tiene una o más variables desconocidas (representadas por las U en las Figuras 1 y 2) sobre una variable dependiente, descontamos el efecto que tienen en ella las variables conocidas. Para ello se usa la fórmula $U = (1-R^2)^{1/2}$ donde R^2 es la variabilidad explicada del sistema cuando se consideran como causas las variables conocidas. Este valor es obtenido cuando se realiza, por ejemplo, un análisis de varianza o regresión múltiple.

Cuando se tienen datos en forma de medidas repetidas en el tiempo, es común que cada unidad de tiempo se analice independientemente de las otras, *i.e.*, no se mezclan datos repetidos en el tiempo. Así, si un sistema fue observado durante dos años, para cada año se hará un análisis de rutas (Schemske y Horvitz, 1988).

TABLA III
EFECTOS ENTRE LAS VARIABLES DE LA TABLA II SEGÚN EL MODELO PROPUESTO EN LA FIGURA 2

Var. dependiente	Var. independiente	Efectos directos	Efectos indirectos	Efectos totales
A	B	0,594	—	0,594
	C	0,054	0,511	0,565
B	C	0,861	—	0,861

Definición de efectos directos, indirectos y totales de una variable sobre otra. Hemos visto como una variable independiente puede afectar "directamente" a una dependiente, es decir, conocemos como obtener el "coeficiente de ruta" entre ambas variables. Pero... ¿Cómo evaluar el efecto "indirecto" de una variable afectando a otra a través de una tercera (en la Figura 2, el caso del efecto de C sobre A a través de B)? y ¿Cómo conocer el efecto "total" de una variable sobre otra (es decir, considerando el efecto "directo" e "indirecto")? Para hacerlo se debe hacer el siguiente procedimiento:

a) Definición de efectos directos: se toma el valor del "coeficiente de ruta". Ejemplo: el efecto directo de B sobre A en la Figura 3 es:

$$B \rightarrow A = 0{,}594$$

b) Definición de efectos indirectos: se multiplican los "coeficientes de ruta" por los que "pasa" el efecto. Ejemplo: el efecto indirecto de C sobre A en la Figura 3 es:

$(C \rightarrow B) \times (B \rightarrow A) = 0{,}861 \times 0{,}594 = 0{,}511$

c) Definición de efectos totales: se suman los efectos directos e indirectos. Ejemplo: el efecto total de C sobre A en la Figura 3 es:

$(C \rightarrow A) + ((C \rightarrow B) \times (B \rightarrow A)) = 0{,}054 + 0{,}511 = 0{,}565$

Una manera de representar de manera compacta todos los efectos mencionados es la ejemplificada en la Tabla III. Si se pretende determinar los efectos reales entre las variables es conveniente conocer todos los efectos en el modelo propuesto para dicho sistema.

Si consideramos las variables de la Tabla II y exploramos o modelamos sus relaciones como en el modelo de la Figura 2, observaremos (Tabla III) que sobre A la variable B es la que tiene un efecto total mayor (0,594), que sobre B la variable C tienen un efecto directo fuerte (0,861) y que sobre A la variable C tiene un efecto indirecto también fuerte (0,511). Esto último condiciona que el efecto total de C sobre A sea también fuerte (0,565) (y no ligero, como haría pensar si se tomara en cuenta sólo el efecto directo de C sobre A (0,054).

Evaluación del modelo propuesto. Si consideramos que en un sistema de cinco variables es posible obtener 59.000 modelos diferentes, es imprescindible tratar de conocer cual de estos modelos es biológica y estadísticamente viable. La validez biológica de un modelo debe ser evaluada explícitamente por cada investigador y la estadística puede ser valorada por diferentes procedimientos. El tema de cómo se realiza la evaluación de los modelos propuestos es enorme (ver Mitchell, 1993), por lo que sólo lo consideraré brevemente aquí. Además, los procedimientos estadísticos para evaluar un modelo son realizados automáticamente por diversos programas de cómputo. Uno de los programas más amigables y accesibles que hace esto es "EPA" (Exploratory path analysis), distribuido gratis por su autor (Shipley, 1997).

Existe una manera estadística de ver cuán bien el modelo predice lo observado, es decir, describe de manera correcta la estructura correlacional observada entre las variables medidas en campo. Para llevarla a cabo se realiza una prueba de bondad de ajuste entre la matriz de correlación predicha por el modelo y la observada con los datos obtenidos en campo. "El estadístico de prueba (para hacer esto) se define a partir de las diferencias entre las matrices de correlación esperadas y observadas, las cuales son cuantificadas por una función de probabilidad. Esta probabilidad es usada para generar una función estadística que se distribuye aproximadamente como una χ^2 "(Parra, 1995), que puede ser descrita con la siguiente ecuación (Kingsolver y Schemske, 1991; Mitchell, 1992):

χ^2 = - 2 ln (probabilidad del modelo) / (probabilidad del modelo perfecto), siendo el "modelo perfecto" el modelo hipotético que reproduce perfectamente la matriz de correlación.

Mientras más similitud exista entre la matriz de correlación observada y esperada, mejor será el ajuste del modelo de rutas propuesto. Modelos con χ^2 no significativos (es decir, sin di-

TABLA IV
COMPARACIÓN ANIDADA DE MODELOS ALTERNATIVOS
AL REPRESENTADO EN LA FIGURA 1

	Bondad de ajuste			Comparación anidada con modelo 1		
Modelo	χ^2	gl	P	χ^2	gl	P
1 (ver Figura 1)	1,38	2	0,50	—	—	—
2 (como Figura 1, sin efecto de A a B)	4,10	3	0,25	2,72	1	0,10
3 (como modelo 2, sin efecto de C a Y)	13,27	4	0,01	11,89	2	0,005

ferencias estadísticas significativas entre la matriz esperada y la observada para los datos de campo) son descriptores adecuados de los datos, ya que reflejan la estructura correlacional de estos (Mitchell, 1992; 1993; Parra, 1995).

Los grados de libertad (gl) de esta prueba se obtienen al restar el número de elementos únicos en la matriz de correlación al número de coeficientes estimados, incluyendo los valores residuales y las relaciones no analizadas (Mitchell, 1992; Parra, 1995). Por ejemplo, en la Figura 1 sería posible tener seis correlaciones, dos efectos de variables desconocidas (valores de U) y dos valores residuales (varianza restante en C y A), es decir diez, pero como sólo tenemos ocho (pues no se consideran las rutas entre A y Y, y entre C y B), la resta da 2 gl.

La comparación entre dos modelos alternativos puede ser hecha usando métodos de máxima verosimilitud para estimar coeficientes de rutas. Esto es debido a que la diferencia entre sus dos valores de χ^2 también se distribuye como una χ^2, con grados de libertad iguales a la diferencia de los grados de libertad de los dos modelos (Mitchell, 1992; Parra, 1995). Por ejemplo, si comparamos tres modelos y estos tienen los valores presentados en la Tabla IV, después de hacer las operaciones correspondientes vemos que los modelos 1 y 2 representan adecuadamente la estructura correlacional entre las variables medidas (porque P = 0,50 y 0,25 respectivamente), no así el modelo 3 (porque P = 0,01). Cuando comparamos el modelo 1 con los modelos 2 y 3 observamos que el modelo 2 no difiere significativamente del 1 (porque χ^2 = 4,1-1,38 = 2,72, gl = 3-2 = 1, y P = 0,10), pero que el 3 sí (porque χ^2 = 13,27-1,38 = 11,89, gl = 4-2 = 2, y P = 0,005). Si nos guiamos por la parsimonia deberíamos seleccionar el modelo 2, pues es menos complejo que el 1, pero la parsimonia no siempre funciona en la naturaleza y debemos evaluar la factibilidad biológica de los dos modelos viables (1 y 2) antes de decidir.

El porcentaje de variación para cada modelo se puede obtener mediante el cálculo de su valor de R^2 (que es igual a: 1 - II (Ui^2), donde Ui es el valor de cada variable residual (Parra, 1995). Otros índices que indican el ajuste de los modelos, que además son buenos para tamaños de n pequeños (Steiger, 1989), son: Gamma (Population index gamma) y RMS (Steiger-Lind adjusted RMS index)

Un análisis de rutas "real"

A continuación desarrollaré un ejemplo donde se indica cómo hacer e interpretar un análisis de rutas. Analizaré el efecto que tienen algunas variables conductuales, alométricas y fisiológicas en la cantidad de esperma eyaculado por humanos. A la fecha las pocas observaciones realizadas sobre este tema indican que el tamaño de la hembra humana está relacionado con la cantidad de esperma que es depositado en ella por su pareja. Existen dos teorías que intentan explicar este hecho. La primera señala que el macho humano evalúa la calidad que tiene la hembra para tener descendientes y, como consecuencia de ello deposita mayor cantidad de esperma en hembras de mayor calidad que en hembras de menor calidad. En este caso se ha sugerido (Baker y Bellis 1993a; b) que el peso de la hembra es un buen indicador de su calidad (Figura 4, modelo 1; Baker y Bellis, 1993a; b). La segunda teoría (Figura 4, modelo 2) considera que el volumen de esperma depositado en una hembra es resultado de un proceso más complejo donde primero intervienen factores conductuales (e.g., selección de pareja, donde no es claro quién selecciona a quién o si lo hacen ambos), después alométricos (e.g., machos de mayor tamaño tendrán mayor volumen de testículos) y por último fisiológicos (e.g., machos con mayor volumen de testículos eyaculan más esperma y a mayor tiempo transcurrido desde la última eyaculación menor cantidad de esperma eyaculado). Los datos originales (n = 8 parejas) fueron tomados de Baker y Bellis (1993a; 1993b) y fueron complementados artificialmente por mí. Los nuevos datos "artificiales" (n = 12) se crearon con la

TABLA V
VARIABLES RELACIONADAS CON LA EYACULACIÓN EN HUMANOS

Pareja	Peso (Kg)		Volumen testículos	Tiempo desde última eyaculación	Espermas eyaculados por cópula
	Hembra	Macho	(cm^3)	(horas)	(individuos x 10^6)
A	64	95	28	60	570
B	58	79	20	149	219
F	59	73	18	168	516
L	52	67	7	32	60
m	54	57	12	48	282
N	54	73	17	56	455
T	56	64	11	54	109
U	52	68	15	32	295
X1	58	74	15	25	263
X2	58	75	16	45	560
X3	58	73	15	25	263
X4	58	75	16	20	216
X5	59	72	14	25	242
X6	59	72	15	70	844
X7	60	76	19	24	333
X8	61	76	18	32	436
X9	57	71	17	38	496
X10	57	71	14	25	242
X11	56	70	12	37	322
X12	56	70	13	25	220

Se presentan datos para 20 parejas. Las primeras ocho parejas son reales (Tomadas de Baker y Bellis, 1993a), los datos de las parejas señaladas con una "x" fueron inventados por el autor con la intención de desarrollar este ejemplo.

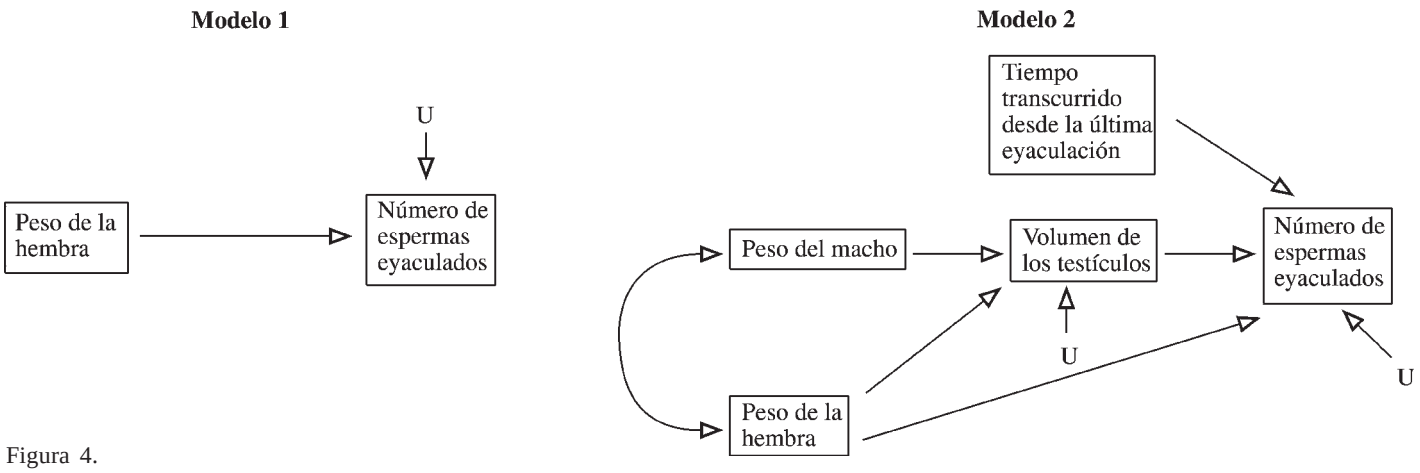

Figura 4.

sola intención de desarrollar este ejemplo, pero respetaron la estructura correlacional establecida entre las variables en los datos originales. Los datos completos son presentados en la Tabla V.

Después de estandarizar los datos, obtener la influencia entre variables (considerando la estructura indicada en la Figura 4, modelos 1 y 2) y la validez de los modelos probados, hacemos los diagramas de flujo resultantes (Figura 5). En ambos modelos observamos que la relación entre las variables es generalmente fuerte, pero el modelo 1 no representa adecuadamente la estructura correlacional de los datos (porque χ^2 = 50,5; P < 0,0005). El modelo 1 no es viable y lo desechamos como una explicación válida que represente la relación entre las variables consideradas. Nos queda el modelo 2, que es viable (porque χ^2 = 3,9; P < 0,419), y que simplificaré en los modelos 3 y 4 con la intención de señalar como realizar la comparación entre modelos.

Los modelos 3 y 4 (Figura 6) están anidados en el modelo 2 (Figura 5), porque pueden ser obtenidos de éste después de borrar varias rutas. Estos modelos son diferentes del 2 porque no consideran las rutas débiles de este último. No toman en cuenta la relación directa entre peso de la hembra y número de espermas eyaculados y, además, el modelo 4 no considera las relaciones directas entre peso de la hembra y volumen de testículos, ni entre tiempo transcurrido desde la última eyaculación y número de espermas eyaculados. Los valores obtenidos para los modelos 3 y 4 pueden ser observados en la Figura 6.

Antes de analizar los efectos directos, indirectos y totales de cada variable sobre otra, comparemos los cuatro modelos propuestos. Los valores de χ^2 resultantes se presentan en la Tabla VI. Recordemos que todos los modelos pueden ser anidados en el modelo 2 y con él los comparamos. De igual manera, el modelo 1 está anidado en todos los otros, de tal manera que puede ser comparado con ellos. De la Tabla VI es claro que los modelos 2, 3 y 4 representan adecuadamente la estructura correlacional de los datos originales (P > 0,05 en la primera columna) y que estos tres modelos no son estadísticamente distintos entre sí (P > 0,25 en la segunda columna), pero que el modelo 1 es diferente al 2 (P < 0,001). En base a esta comparación podemos eliminar el modelo 1 y seleccionar con criterios biológicos a cualquiera de los otros tres.

Para ver los efectos entre las variables supongamos que de los tres modelos restantes el modelo 4 es el adecuado biológicamente. Estadísticamente este modelo es el más adecuado por que tiene un mejor ajuste a los datos (P = 0,515) y es menos complejo, lo que se refleja en un mayor gl = 7. Para este modelo los valores de los efectos entre variables se representan en la Tabla VII. Los efectos directos y totales más grandes ocurren entre peso del macho y volumen de testículos (0,87) y entre los pesos de hembra y macho (0,77). El efecto indirecto más grande es el establecido entre peso de la hembra y volumen de testículos (0,67).

La interpretación biológica del sistema analizado puede ser la siguiente. Del análisis de los resultados podemos concluir que la relación entre "tamaño de la hembra" y "número de espermas eyaculados en ella por su pareja"

TABLA VI
COMPARACIÓN ANIDADA DE MODELOS ALTERNATIVOS AL MODELO 2 PRESENTADO EN LA FIGURA 5

	Bondad de ajuste			Comparación anidada con modelo 1		
Modelo	χ^2	gl	P	χ^2	gl	P
2 (Figura 4)	3,91	4	0,419	—	—	—
1 (Figura 4)	50,5	12	<0,0005	46,59	8	<0,001
3 (Figura 6)	4,40	5	0,493	0,59	1	>0,25
4 (Figura 6)	6,22	7	0,515	2,31	3	>0,25

TABLA VII
EFECTOS ENTRE LAS VARIABLES DEL MODELO 4 PROPUESTO EN LA FIGURA 6

Variable Dependiente	Independiente	Efectos Directos	Indirectos	Totales
Peso de la hembra	Peso del macho	0,77	—	0,77
	Volumen de testículos	—	0,67	0,67
	Número de espermas eyaculados	—	0,34	0,34
Peso del macho	Volumen de testículos	0,87	—	0,87
	Número de espermas eyaculados	—	0,44	0,44
Volumen de testículos	Número de espermas eyaculados	0,51	—	0,51

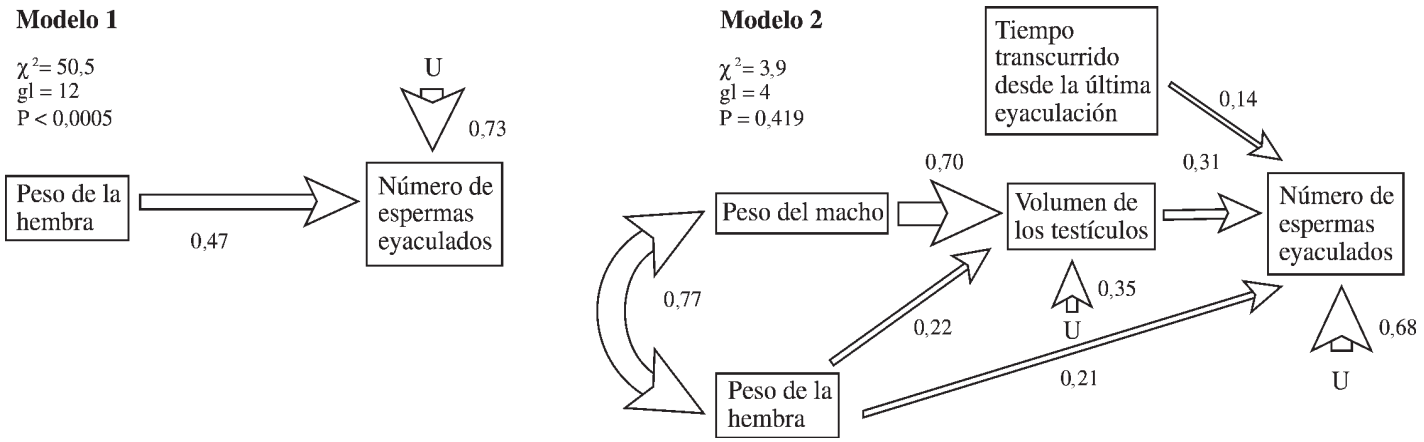

Figura 5.

está determinada por un conjunto de variables conductuales, alométricas y fisiológicas involucradas en el proceso. Los datos sugieren que hembras y machos escogen parejas de tamaño similar (factor conductual), lo que ocasiona que, por ejemplo, hembras grandes tengan de pareja a machos grandes. Esto determina a su vez que hembras grandes tengan machos con un volumen testicular relativamente grande (factor alométrico en el macho), y que estos machos eyaculen mayor cantidad de esperma (factor fisiológico). El efecto del tiempo pasado desde la última eyaculación no tuvo un efecto marcado en los modelos analizados. La relación directa "peso de la hembra - número de espermas eyaculados en ella" no fue viable cuando se consideró sola en un modelo, o fue de baja intensidad (coeficiente de ruta < 0,23) cuando se consideró en otros modelos.

Herramientas que mejoran la aplicación de un análisis de rutas

Hasta hace unos años dos de las principales limitantes para utilizar el análisis de ruta, en las ciencias biológicas y sociales, eran los cálculos estadísticos complicados y laboriosos y el número reducido de repeticiones con las que comúnmente se cuenta. Estas limitantes se han visto resueltas parcialmente con el desarrollo de computadoras más veloces y nuevas técnicas estadísticas que permiten manejar muestras pequeñas.

Una de las técnicas estadísticas que ha facilitado el uso del análisis de ruta es la de iteraciones. Las iteraciones son básicamente un re-muestreo de los datos con que se cuenta. Para entender la importancia de esta técnica en las ciencias biológicas es necesario dar un ejemplo. Utilizaremos uno de ecología, donde pocas veces podemos medir a todos los individuos en una población y generalmente sólo tenemos datos para una pequeña muestra de esta población. Esto nos imposibilita tener los valores "reales" de los parámetros poblacionales (e.g., media, desviación estándar) con los cuales estimar, por ejemplo, semejanzas entre poblaciones vecinas o similitud en los tamaños de los individuos más depredados en dos o más poblaciones. A través de la técnica de iteraciones podemos muestrear nuestros datos muchas veces y obtener un espacio probabilístico en que se puede encontrar nuestro parámetro poblacional (Shipley 1997). Existen varias técnicas de iteración (e.g., Montecarlo y Jaknife) pero aquí únicamente describiré la de Bootstrap porque, dentro de las existentes, es la que permite realizar más iteraciones (J. Navarro, com. pers.).

Consideremos tener la variable C para la cual tenemos un tamaño de muestra n = 5 con los valores representados en la Tabla I. Bootstrap lo que hace es un muestreo al azar con re-emplazamiento (es decir, que toma de la muestra n número de valores (en este caso cinco), pero pudiendo repetir un valor hasta n veces en esa muestra). Con los valores disponibles en la Tabla I para la variable C podemos obtener 36, 36, 43, 45, 45 en la primera iteración; 45, 41, 41, 62, 36 en la segunda; y 41, 41, 41, 41, 41 en la tercera. Para cada iteración podemos obtener, por ejemplo, una media, que en este caso podemos llamar media de la iteración ($\bar{x}_b$), pero también podemos obtener otros estimadores como desviación estándar, intervalo o error estándar. Si tenemos 100 iteraciones, podemos graficar cómo se distribuye este parámetro o estimador de la población. La media original de los datos de la variable C en la Tabla I es 45,4. Con las tres iteraciones que realizamos sabemos ahora que la media se puede encontrar entre 41 y 47,

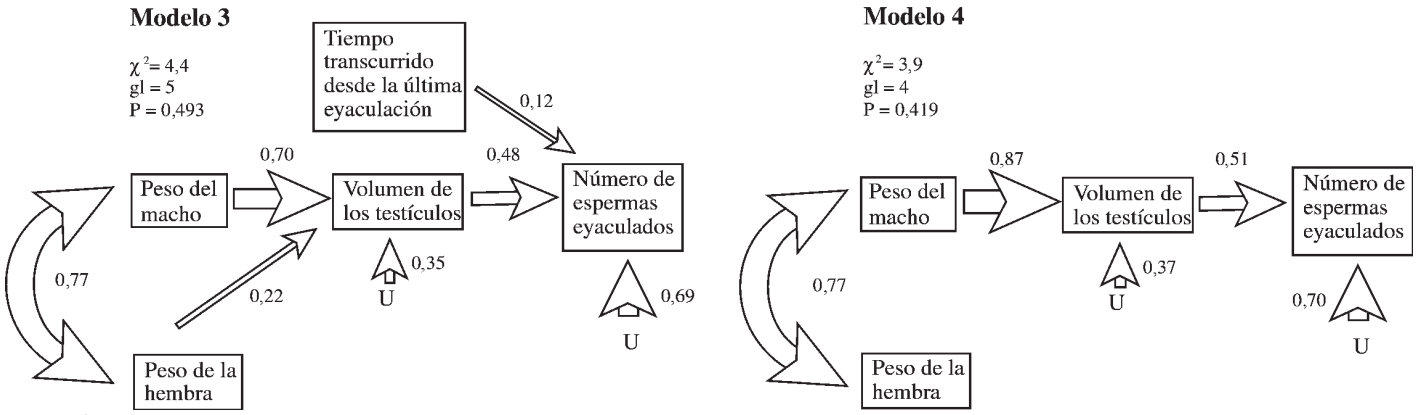

Figura 6.

es decir, en vez de tener un solo valor tendremos una zona en la que con cierta probabilidad se encuentra este valor.

Otro de los usos de las iteraciones es para determinar la distribución exacta de nuestros datos y no suponerla, como se hace comúnmente cuando se usa una prueba paramétrica que presume una distribución normal de los datos. Aun si pudiéramos tener todos los valores de, digamos, altura para una especie de árbol neotropical, es poco probable que las frecuencias de dichos datos se ajusten perfectamente a una distribución normal u otra bien estudiada (e.g., binomial, Poisson), que son las que comúnmente se asume que tienen nuestros datos, por cuestión de simplicidad. Así, cuando modelamos los datos considerando *a priori* que tienen distribuciones predeterminadas, nos enfrentamos al problema, o más bien, a la posibilidad, de que nuestros datos tengan otra distribución.

Con la técnica bootstrap se resuelve, cuando menos más convincentemente que antes, dos problemas: el bajo tamaño de muestra y el supuesto de una distribución que no es (normal, binomial, Poisson, etc.). Al realizar las iteraciones puede obtenerse n^{10} poblaciones de datos, lo que permite a la vez modelar una distribución más "real" de los datos a través de ellos mismos. De hecho eso es lo que pretende decir "bootstrap". Es un juego de palabras en inglés que puede ser traducido como "levantar las botas por las agujetas". Eso es lo que hace esta técnica: construye la distribución de los datos a partir de ellos mismos.

En este punto tal vez surja la pregunta ¿Cuál es la aplicación de las iteraciones en un análisis de rutas? Una de sus varias aplicaciones es que nos pueden ayudar a determinar el espacio en que se encuentran nuestros "coeficientes de ruta". Como ya vimos los valores de los "coeficientes de ruta" entre dos variables en un análisis de ruta son los "coeficientes parciales estandarizados de la regresión". Si consideramos dos variables, digamos, C y B, de la Tabla I y hacemos tres iteraciones, podemos obtener:

Primera iteración:	41	41	36	45	45
	7,8	7,8	5,6	8	8
Segunda iteración:	43	36	62	43	41
	7,9	5,6	9,3	7,9	7,8
Tercera iteración:	41	41	41	41	62
	7,8	7,8	7,8	7,8	9,3

y si realizamos regresiones múltiples con estos datos (después de estandarizarlos), podemos obtener el rango en el que es más probable que se encuentre el "coeficiente de ruta", en este caso del efecto de la variable C sobre B. Si la relación que encontramos inicialmente fue al azar en este punto lo notaremos, pues podremos determinar si ese valor fue un estimado no común dentro de la distribución del mismo estimado. Por ejemplo, los datos originales indican que el "coeficiente de ruta" entre las variables C y B es 0,861. Los "coeficientes de ruta" obtenidos en las tres iteraciones son 0,891; 0,851 y 1. Dado que nestro "coeficiente de ruta" original se encuentra en el rango de lo obtenido en las tres iteraciones, podemos concluir que es un estimado común dentro de su distribución.

El uso de iteraciones para determinar coeficientes de ruta se ha hecho en estudios biológicos. En Costa Rica, Langen *et al.* (1998) utilizaron, para sus modelos que tratan de explicar la territorialidad grupal del ave *Calocitta formosa*, los coeficientes parciales estandarizados de la regresión máximos posibles. Obtuvieron coeficientes de ruta muy altos entre las variables que analizaron, lo que les permitió a su vez "probar la hipótesis de que el tamaño y cualidad del territorio determinan el tamaño del grupo" en esta especie.

Otra aplicación de la técnica de iteraciones es la posibilidad de evaluar qué rutas son las más comunes entre un número "x" de variables. La evaluación de rutas se logra quitando y poniendo "líneas de datos" (es decir, todos los datos de un individuo; el caso de los valores 36; 5,6 y 101 para el individuo número 1 en la Tabla I) y evaluando para cada iteración la probabilidad de que exista o no una línea que conecta dos variables. La probabilidad de la existencia de esta línea la podemos determinar, por ejemplo, considerando sólo aquellas que tengan un coeficiente de ruta mayor a 0,3 que sean estadísticamente significativas y que se presentan en más del 40% de las iteraciones. Esta aproximación nos permite obtener diferentes modelos alternativos que son posibles de probar contra los datos originales o con nuevas series de datos obtenidas implícitamente para ello.

AGRADECIMIENTOS

A los alumnos de los cursos de análisis de rutas que he coordinado en el Instituto de Ecología, A.C. y en la Facultad de Biología de la Universidad Veracruzana, y a dos revisores anónimos por los comentarios que permitieron mejorar este trabajo. A R. Dirzo por sugerirme este análisis para evaluar "mis" sistemas de estudio. A J. Navarro por sus comentarios sobre iteraciones. A M. Carvallo por correcciones a las distintas versiones de este escrito. Este trabajo fue parcialmente apoyado por una beca-crédito de doctorado de CONACyT (9352171).

REFERENCIAS

Baker RR, Bellis MA (1993a) Human sperm competition: ejaculate adjustment by males and the function of masturbation. *Animal Behavior 46*: 861-885.

Baker RR, Bellis MA (1993b) Human sperm competition: ejaculate manipulation by females and a function for female orgasm. *Animal Behavior 46*: 887-909.

Kingsolver JG, Schemske DW (1991) Path analysis of selection. *Trends in Ecology and Evolution 6*: 276-280.

Langen TA, Vehrencamp SL (1998) Ecological factors affecting group and territory size in White-throated Magpie-Jays. *Auk 115*: 327-339.

Li CC (1975) *Path analysis - a primer*. Boxwood Press, California. pp. 100-134.

Mitchell R J (1992) Testing evolutionary and ecological hypothesis using path analysis and structural equation modeling. *Functional Ecology 6:* 123-129.

Mitchell RJ (1993) Path analysis: pollination. In Scheiner SM, Gurevitch J (Eds.) *Design and analysis of ecological experiments.* Chapman & Hall. pp. 211-231.

Parra V (1995) *Factores ecológicos limitantes de la fecundación y selección natural en características florales de* Ipomea wolcottianna *Rose (Colvolvulaceae).* Tesis de doctorado. Centro de Ecología, UACPyP-CCH, UNAM. pp. 86-114.

Pedhazur EJ (1982) *Multiple regression in behavioral research: explanation and prediction.* Holt, Rinehart & Winston, New York.

Petraitis PS, Dunham AE, Niewiarowski PH (1996) Inferring multiple causality: the limitations of path analysis. *Functional Ecology 10*: 421-431.

Rigdon E (1998) *What is structural equation modeling*? http://www.gsu.edu/~mkteer/sem.html. Consultado el 5 de mayo de 1998.

Shemske DW, Horvitz CC (1988) Plant-animal interactions and fruit production in a neotropical herb: a path analysis. *Ecology 69*: 1128-1137.

Shipley B (1997) Exploratory path analysis with applications in ecology and evolution. *American Naturalist 149*: 1113-1138.

Steiger (1988) *EZPATH: a supplementary module for SYSTAT and SYGRAPH.* Evanston, Illinois, USA.

Thompson JN (1994) *The coevolutionary process*. The Chicago University Press. Chicago, USA. 376 pp.

Thompson JN (1997) Evaluating the dynamics of coevolution among geographically structured populations. *Ecology 78*: 1619-1623.

Wooton JT (1994) Predicting direct and indirect effects: an integrated approach using experiments and path analysis. *Ecology 75*: 151-165.

Wright S (1968) Evolution and the genetics of populations. Vol. 1. *Genetics and Biometric foundations.* The University of Chicago Press, Chicago. pp. 299-324.

Zar JH (1996) *Biostatistical analysis*. Prentice-Hall Inc., New Jersey. 718 pp.