CAPÍTULO 6:

Interpolando, extrapolando y comparando las curvas de acumulación de especies basadas en su incidencia

Robert K. Colwell
Department of Ecology and Evolutionary Biology,
University of Connecticut,
Storrs, Connecticut 06269-3043 USA
colwell@uconn.edu

Chang Xuan Mao
Department of Statistics,
University of California,
Riverside, California 92521 USA

Jing Chang
Department of Preventive Medicine,
University of Southern California,
Los Angeles, California 90089 USA

Sobre Diversidad Biológica: El significado de las Diversidades Alfa, Beta y Gamma.

Editores:
Gonzalo Halffter, Jorge Soberón, Patricia Koleff & Antonio Melic

Patrocinadores:
COMISION NACIONAL PARA EL CONOCIMIENTO Y USO DE LA BIODIVERSIDAD (CONABIO) MÉXICO

SOCIEDAD ENTOMOLÓGICA ARAGONESA (SEA), ZARAGOZA, ESPAÑA.

GRUPO DIVERSITAS-MÉXICO

CONSEJO NACIONAL DE CIENCIA Y TECNOLOGÍA (CONACYT) MÉXICO

ISBN: 84–932807–7–1
Dep. Legal: Z–2275–05
m3m: Monografías Tercer Milenio
vol.4, S.E.A., Zaragoza, España
30 Noviembre 2005
pp: 73 – 84.

Información sobre la publicación:
www.sea-entomologia.org/m3m

INTERPOLANDO, EXTRAPOLANDO Y COMPARANDO LAS CURVAS DE ACUMULACIÓN DE ESPECIES BASADAS EN SU INCIDENCIA *

Robert K. Colwell, Chang Xuan Mao & Jing Chang

* Trabajo publicado originalmente en: *Ecology*, **85**(10), 2004, pp. 2717-2727

Resumen: Se propone un modelo mixto binomial general para la función de acumulación de especies basado en la presencia-ausencia (incidencia) de las especies que ocurren en una muestra de cuadros u otras unidades de muestreo. El modelo abarca la interpolación entre cero y el número observado de muestras, así como la extrapolación más allá del conjunto de muestras observadas. En el caso de la interpolación (rarefacción basada en muestras), se desarrollan expresiones de forma cerrada de fácil cálculo mediante el método de momentos, tanto para la riqueza esperada como para sus límites de confianza. Esto elimina completamente la necesidad de utilizar métodos de remuestreo y permite la comparación estadística directa de la riqueza entre conjuntos de muestras. Basada en la incidencia de especies, se desarrolla una variante del modelo de Coleman (ordenación aleatoria) y ésta se compara con el método de interpolación basado en momentos. Para la extrapolación más allá del conjunto de muestras empíricas (y a su vez, como un método alternativo de interpolación), se describe un estimador probabilístico con un intervalo de confianza *bootstrap* basado en un algoritmo secuencial guiado por el Criterio Akaike de Información (CAI) para ajustar los parámetros del modelo mixto. Tanto el estimador probabilístico como el del momento se ilustran con conjuntos de datos para aves de climas templados, y semillas, hormigas y árboles tropicales. El estimador basado en el momento se recomienda confiablemente para interpolación (rarefacción con base en las muestras). Para la extrapolación, el estimador probabilístico se desempeña bien al duplicar o triplicar el número de muestras empíricas, pero no es confiable para estimar la asíntota de la riqueza. Se discute la sensibilidad a la heterogeneidad [*patchiness*] espacial (o temporal) de la rarefacción basada en individuos y la basada en muestras.

Palabras clave: modelo mixto binomial; curva de Coleman; EstimateS; ordenación aleatoria; rarefacción; estimación de la riqueza; extrapolación de la riqueza; curva de acumulación de especies; riqueza de especies.

Interpolating, extrapolating, and comparing incidende-based species accumulation curves

Abstract: A general binominal mixture model is proposed for the species accumualtion function based on presence-absence (incidence) of species in a sambple of quedrats or other sampling units. The model covers interpolation between zero and the observed numbers of samples, as well as extrapolation beyond the observed sample set. For interpolation (sample based rarefaction) easily calculated, closed-form expressions for both expected richness and its confidence limits are developed (using the method of moments) that completely eliminate the need for resampling methods and permit direct statistical comparison of richness between sample sets. An incidence-based form of the Coleman (ramdom-placement) model is developed and compared with the moment-based interpolation method. For extrapolation beyond the empirical sample set (and simultaneously, as an alternative method of interpolation), a likelihood-based estimator with a bootstrap confidence interval is described that relies on a sequential, AIC-guided algorithm to fit the mixture model parameters. Both the moment-based and likelihood-based estimators are illustrayed with data sets for temperate birds and tropical seeds, ants, and trees. The moment-based estimator is confidently recommended for interpolation (sample-based rarefaction). For extrapolation, thelikelihood-based estimator performs well for doubling or tripling the number of empirical samples, but it is nor reliable for estimating the richness asymptote. The sensitivity of individual-based and sample-based rarefaction to spatial (or temporal) patchiness is discussed.

Key words: binomial mixture model, Coleman curve, EstimateS, ramdom placement, rarefaction, richness estimation, richness extrapolation, species accumulation curve, species richness.

Introducción

Los ecólogos y biólogos de la conservación a menudo necesitan determinar el número de especies (riqueza de especies) encontrado en un área dada, o requieren comparar el número de especies entre áreas distintas. Sin embargo, en muchos casos, es poco práctico o aún imposible enumerar directamente a las especies presentes. Por lo tanto es necesario hacer un muestreo. Desafortunadamente, la riqueza de especies observada dentro de hábitats (diversidad alfa) es notablemente dependiente del tamaño de muestra, debido a los efectos de muestreo. Más aún, la riqueza observada depende intrínsecamente del tamaño de muestra cuando los datos de distintos hábitats se agrupan sucesivamente, debido al recambio de especies (cambio en la composición de especies o diversidad beta). El estudio de las relaciones empíricas especies-área (p. ejem. Rosenzweig, 1995; Scheiner, 2003) generalmente se concentran en, o por lo menos contemplan, esta última fuente (es decir, el recambio de especies) de dependencia del tamaño de muestra. En el presente artículo, por lo contrario nos concentramos en la medición de la riqueza de especies a escalas locales, en donde los aspectos del muestreo son sustancialmente más importantes que los del recambio. En términos estadísticos, nos interesan conjuntos de muestras en los que cada muestra razonablemente puede ser considerada una muestra aleatoria del mismo universo. En términos prácticos, esto significa que el orden de las muestras en el tiempo o su arreglo en el espacio dentro de un conjunto de muestras no tiene importancia; de hecho, la no importancia del orden de las muestras es una característica diagnóstica de los tipos de conjuntos de muestras utilizados apropiadamente por los ecólogos en la estimación de la diversidad local (alfa).

Una *curva de acumulación de especies* es la gráfica del número de especies observadas como función de alguna medida del esfuerzo de muestreo requerido para observarlas. (En sentido amplio, las curvas clásicas especies-área que se concentran en la diversidad beta, son así curvas de acumulación de especies, pero las curvas que presentamos en este artículo pueden o no utilizar el área como una medida del esfuerzo de muestreo, y explícitamente representan la diversidad alfa, tal y como se explicó.) La acumulación secuencial de individuos en una sola muestra, o la agrupación sucesiva de muestras de un solo conjunto de muestras, produce una curva de acumulación de especies, pero ésta no será una curva suave debido a la heterogeneidad espacial (o temporal) y efectos estocásticos simples. Para los individuos de una misma muestra, la rarefacción clásica (*basada en individuos*) puede ser utilizada para producir una curva suave que estima el número de especies que se observaría para cualquier número menor de individuos, bajo el supuesto de mezcla aleatoria de individuos (Hurlbert, 1971; Simberloff, 1972; Heck *et al.,* 1975). Para conjuntos de muestras replicados (conjuntos de muestras), el número esperado de especies que sería observado para cualquier número menor de muestras se puede estimar mediante la rarefacción *basada en muestras*, bajo el supuesto de orden aleatorio de muestras (Gotelli y Colwell, 2001). (Aunque la rarefacción basada en muestras es el término más exacto, "suavizando la curva de acumulación de especies" mediante el remuestreo aleatorio es una descripción apropiada para el mismo concepto.)

Una curva de acumulación de especies basada en muestras puede ser construida de cualquier matriz empírica de especies-por-muestra. Las celdas de la matriz empírica pueden contener abundancias de las especies (una *matriz de abundancia*) o simplemente datos de presencia/ausencia (una *matriz de incidencia*). Por supuesto, cualquier matriz de abundancia puede ser transformada a su correspondiente matriz de incidencia al reemplazar cada valor de celda que no sea cero por un uno para indicar la presencia. Las curvas de acumulación de especies basadas en muestras, por su naturaleza dependen únicamente de los datos de incidencia, aún cuando se disponga de datos de abundancia (que no estarán disponibles para algunos tipos de conjuntos de muestras).

Hasta hace poco, las curvas de rarefacción basadas en muestras tenían que construirse por algoritmos de remuestreo computacionalmente intensivos, tales como los utilizados por la aplicación de libre uso EstimateS (Colwell, 1994-2004). La necesidad práctica de tales herramientas se manifiesta por el hecho de que al final del 2003, más de 10,000 copias del programa EstimateS habían sido descargadas por usuarios en alrededor de 100 países y utilizado en decenas de artículos publicados. (Como otro indicador, Google, el motor de búsqueda del Internet www.google.com actualmente registra más de 1000 portales en el Internet que citan EstimateS). Tal y como sucede con la rarefacción basada en individuos, la rarefacción basada en muestras permite la comparación de diferentes ensamblajes a niveles comparables de esfuerzo de muestreo. Desafortunadamente, no ha existido un método adecuado para calcular los intervalos de confianza para curvas de rarefacción basadas en muestras, lo cual seriamente ha limitado su utilidad para la comparación de la riqueza de conjuntos de muestras.

La construcción de curvas de rarefacción basadas en muestras puede ser vista como un proceso de *interpolación* a partir de la riqueza de especies agrupada del conjunto completo de muestras, a la riqueza esperada de un subconjunto de aquellas muestras. El sueño de todo biólogo involucrado en inventarios biológicos es la *extrapolación* rigurosa de las curvas de rarefacción empíricas basadas en muestras para estimar, con intervalos de confianza, cuántas especies serían encontradas en un conjunto de muestras más grande del mismo ensamblaje; idealmente la extrapolación nos daría la asintótica, riqueza "verdadera" del ensamblaje.

En este artículo, presentamos un modelo binomial mixto que es unificado y estadísticamente riguroso para evaluar los patrones de incidencia en ensamblajes multiespecíficos. (El desarrollo estadístico completo del modelo, con sus teoremas de apoyo y pruebas, se presenta en otro lugar [Mao *et al*., 2004]). Con base en el modelo, presentamos fórmulas analíticas simples para las curvas de rarefacción basadas en muestras y sus

Tabla I. Conjuntos de datos empíricos usados para ilustrar los métodos.

Taxa y método	Especies	Muestras	Localidad	Referencias
aves de zonas templadas, inventario	67	50	Inventario de las Aves Nidificantes Norteamericanas, 1998	Dorazio y Royle (2003)
hormigas de la selva, trampas Winkler	197	41	Estación Biológica La Selva, Costa Rica	Longino *et al*. (2002)
banco de semillas de la selva, cuadros	34	121	Estación Biológica La Selva, Costa Rica	Butler y Chazdon (1998)
briznales de árboles (2.5-5.0 cm dap): selva madura, cuadros	60	100	Estación Biológica La Selva, Costa Rica	R.L. Chazdon (datos no publicados)
briznales de árboles árboles (2.5-5.0 cm dap): selva secundaria, cuadros	50	100	Estación Biológica La Selva, Costa Rica	R.L. Chazdon (datos no publicados)

intervalos de confianza (*interpolación*). Estas fórmulas reemplazan completamente los métodos de remuestreo para producir curvas de rarefacción basadas en muestras. Asimismo, por primera vez exploramos un método de extrapolación sin ajuste de curvas, con intervalos de confianza *bootstrap.* Ilustramos tanto la interpolación como la extrapolación utilizando conjuntos de datos para árboles y hormigas de bosque tropical, un banco de semillas tropical y aves de climas templados (Tabla I).

El Modelo

Considere un ensamblaje de especies con una riqueza verdadera desconocida, *S*, muestreado por cuadros, trampas, cebo, cercos, redes de arrastre, redes de niebla, u otras unidades de muestreo replicadas. (Para el desarrollo del modelo, llamaremos a estas unidades de muestreo *cuadros,* para evitar usar la palabra *muestra* como nombre y como verbo.) Los datos para *h* cuadros son expresados como una matriz de incidencia especies-cuadro, *S*-por-*h*, que consiste en los indicadores de presencia Z_{ij}:

Z_{ij} = 1 si la *i*-ésima especie está presente en el *j*-ésimo cuadro

Z_{ij} = 0 si la *i*-ésima especie está ausente en el *j*-ésimo cuadro

Para desarrollar un modelo teórico, asumimos dos supuestos estadísticos: (1) la *i*-ésima especie tiene la misma probabilidad φ_i de estar presente en cada cuadro, y (2) las Z_{ij} son independientes, dada φ_i, para toda *i* y *j*. La función de acumulación de especies, misma que nos da el número esperado de especies observadas en *h* cuadros, es la suma de las probabilidades, para todas las especies, de que cada especie no esté ausente de todos los *h* cuadros:

$$\tau(h) = \sum_{i=1}^{S} [1-(1-\phi_i)^h]. \qquad \text{(Ec. 1)}$$

Las especies con probabilidades de presencia φ idénticas se pueden considerar en conjunto como un grupo. Supongamos que hay *G* de tales *grupos de incidencia* homogéneos (algunos o todos de los cuales pueden contener una sola especie). Para el *k*-ésimo grupo de incidencia, Ψ_k es la *probabilidad de presencia* común (una medida de qué tan rara o común es la especie) y π_k es el *tamaño relativo del grupo*, es decir, el número de especies en el *k*-ésimo grupo dividido por el número total de especies, S. La función de acumulación de especies $\tau(h)$ es, entonces

$$\tau(h) = S\sum_{k=1}^{G} \pi_k [1-(1-\psi_k)^h]. \qquad \text{(Ec. 2)}$$

La asíntota $\tau(\infty)$, el límite de $\tau(h)$ conforme el número de cuadros *h* tiende al infinito, es idéntica a la verdadera riqueza *S*. Notemos que es posible re-escribir la función de acumulación de especies $\tau(h)$ como

$$\tau(h) = S\sum_{k=1}^{G} \pi_k (1-e^{-C_k h}) \qquad \text{(Ec. 3)}$$

donde $C_k = -\log(1-\Psi_k)$. Esta reformulación nos permite mostrar que nuestro modelo es una generalización no paramétrica del modelo clásico exponencial negativo de Holdridge *et al*. (1971) y de Soberón y Llorente (1993). El modelo exponencial negativo supone que todas las especies comparten la misma probabilidad de presencia Ψ_1 y por lo tanto forman un solo grupo de incidencia. De esta manera podemos hacer que $G = 1$, $\pi_1 = 1$, y $C_1 = C$, dando el modelo clásico exponencial:

$$\tau(h) = S(1-e^{-Ch}). \qquad \text{(Ec. 4)}$$

En el modelo expresado por Ec. 2 el número de grupos de incidencia *G* puede tomar cualquier valor, proporciones de grupo π_k pueden variar libremente (con la simple restricción que $\sum_{k=1}^{G} \pi_k = 1$) y el patrón de probabilidades de presencia Ψ_k no tiene restricciones. Por lo tanto, se espera que este riguroso modelo de muestreo-teórico sea aplicable a una amplia gama de taxa con abundancias relativas y patrones de incidencia variadas.

Supongamos que se toma una muestra aleatoria de *H* cuadros, llamada *el conjunto empírico de muestras*. Si $Z_{ij} = 0$ para toda *j* (todos los cuadros), entonces la *i*-ésima especie no se observa en el conjunto empírico de muestras. En la matriz de datos observados, todas las filas (especies) tienen por lo menos un $Z_{ij} > 0$ en la matriz de *S*-por-*H* especies-cuadros. Si s_j representa el número de especies encontradas en exactamente *j* cuadros del conjunto empírico de muestras, entonces los s_j se denominan *conteos* (categorías de la frecuencia de ocurrencia). De esta manera s_0 es el número de especies presente en el ensamblaje "blanco" [*target*] pero no observado en el conjunto empírico de muestras, s_1 es el número de especies encontradas en precisamente un

Tabla II. Rarefacción basada en muestras. El cuadro demuestra que los conteos observados s_1, s_2 . . . s_H son estadísticos suficientes para la rarefacción basada en muestras.

Ejemplo *a*. Desasociación entre especies, baja variación en riqueza entre muestras.

Abundancia Cuadros

	A	B	C	D	Σ
Sp1	6	0	3	0	9
Sp2	0	1	0	5	6
Sp3	1	0	0	0	1
Sp4	0	4	0	0	4
Sp5	0	0	2	0	2
Sp6	0	0	0	3	3
Σ	7	5	5	8	

Incidencia Cuadros

	A	B	C	D	Σ
Sp1	1	0	1	0	2
Sp2	0	1	0	1	2
Sp3	1	0	0	0	1
Sp4	0	1	0	0	1
Sp5	0	0	1	0	1
Sp6	0	0	0	1	1
Σ	2	2	2	2	

Rarefacción basada en muestras
***h* (número de cuadros agrupados)**

1		2		3		4	
A	2	**A+B**	4	**A+B+C**	5	**A+B+C+D**	6
B	2	**A+C**	3	**A+B+D**	5		
C	2	**A+D**	4	**A+C+D**	5		
D	2	**B+C**	4	**B+C+D**	5		
		B+D	3				
		C+D	4				
Riqueza promedio:	2.00		3.75		5.00		6.00

Ejemplo *b*. Asociación entre especies, alta variación en riqueza entre muestras.

Abundancia Cuadros

	A	B	C	D	Σ
Sp1	4	0	5	0	9
Sp2	3	0	3	0	6
Sp3	1	0	0	0	1
Sp4	4	0	0	0	4
Sp5	0	0	2	0	2
Sp6	0	0	3	0	3
Σ	12	0	13	0	

Incidencia Cuadros

	A	B	C	D	Σ
Sp1	1	0	1	0	2
Sp2	1	0	1	0	2
Sp3	1	0	0	0	1
Sp4	1	0	0	0	1
Sp5	0	0	1	0	1
Sp6	0	0	1	0	1
Σ	4	0	4	0	

Rarefacción basada en muestras
***h* (número de cuadros agrupados)**

1		2		3		4	
A	4	**A+B**	4	**A+B+C**	6	**A+B+C+D**	6
B	0	**A+C**	6	**A+B+D**	4		
C	4	**A+D**	4	**A+C+D**	6		
D	0	**B+C**	4	**B+C+D**	4		
		B+D	0				
		C+D	4				
Riqueza promedio:	2.00		3.75		5.00		6.00

Notas: Dado que los dos ejemplos contrastantes comparten los mismos conteos ($s_1 = 4$, $s_2 = 2$) para ambos ejemplos, producen la misma curva de rarefacción basada en muestras (Fig. 5). Cualquier patrón de incidencia que produce los mismos conteos producirá el mismo patrón de riqueza promedio por el proceso de promediación combinatoria. De la misma manera, las curvas de rarefacción basadas en individuos para los dos ejemplos son idénticas una a la otra, a pesar de las diferencias en la abundancia y los patrones de incidencia, dado que comparten el mismo vector de abundancia relativa (9, 6, 1, 4, 2). La Fig. 5 muestra la curva de rarefacción basada en individuos y la que se basa en muestras para los ejemplos, mismos que a la vez se basan respectivamente en las matrices de Abundancia e Incidencia.

cuadro, s_2 es el número de especies encontrados en precisamente dos cuadros, etc. Por lo tanto, la riqueza observada en el conjunto empírico de muestras es $S_{obs} = \sum_{j=1}^{H} s_j$ y el número total de especies, observadas y no observadas, es $S = S_{\text{obs}} + s_0$. Los conteos observados, s_1, s_2 ... s_H, son estadísticos suficientes, dado que contienen toda la información necesaria para estimar la riqueza como una función del esfuerzo de muestreo, $\tau(h)$, como demostramos de manera rigurosa en otra publicación (Mao *et al*., 2004) y mostramos con ejemplos en la próxima sección.

Interpolación (Rarefacción)

Un enfoque intuitivo para la estimación de $\tau(h)$ a $h < H$, un proceso aquí llamado *interpolación*, es enumerar sistemáticamente todos los subconjuntos distintos de h cuadros de H cuadros del conjunto empírico de muestras, encontrar la riqueza observada en cada subconjunto de cuadros y calcular su promedio como estimador de $\tau(h)$. La Tabla II ofrece un ejemplo sencillo de este procedimiento para dos conjuntos de datos hipotéticos contrastantes. Este procedimiento de enumeración sistemática sale caro en términos de cómputo cuando h es grande. El procedimiento de aleatorización usado por EstimateS (Colwell, 1994-2004) es una alternativa aproximada al procedimiento explícito de enumeración. Sin embargo, como ahora demostramos, ni el procedimiento de enumeración ni el de aleatorización son necesarios porque estimadores de forma cerrada y fácilmente calculados están disponibles para $\tau(h)$ a $h < H$, junto con los intervalos de confianza para la asíntota.

Para la interpolación hay un estimador no sesgado $\tilde{\tau}(h)$ para $\tau(h)$ basado en los conteos s_j, debidamente pesados mediante coeficientes combinatorios. Recordando que $S_{obs} = \sum_{j=1}^{H} s_j$, entonces

$$\tilde{\tau}(h) = \sum_{j=1}^{H} (1 - \alpha_{jh}) s_j = S_{obs} - \sum_{j=1}^{H} \alpha_{jh} s_j$$

h= 1, 2,...., H (Ec. 5)

en donde los coeficientes combinatorios α_{jh} se definen por

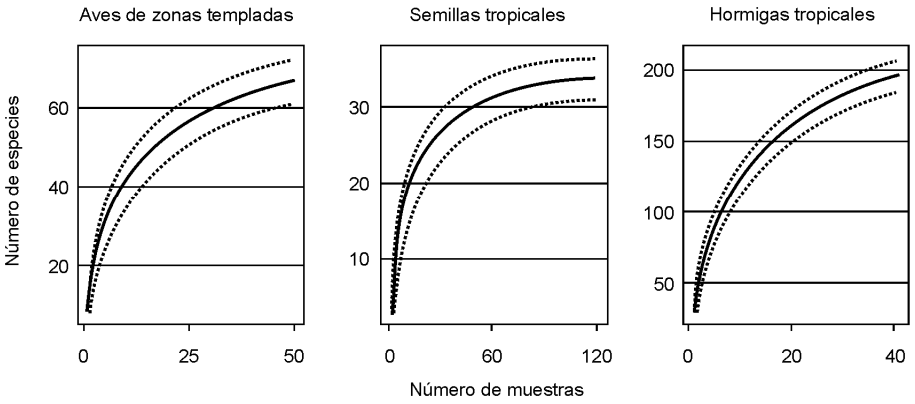

Fig. 1. Curvas de rarefacción basadas en muestras (curvas de acumulación de especies interpoladas) para tres conjuntos empíricos de datos de la Tabla I. Los valores esperados para la riqueza de especies (línea sólida) fueron calculados utilizando el estimador basado en momentos de Ec. 5 con intervalos de confianza de 95% (línea punteada) con base en Ec. 6 y Ec. 7.

$$\alpha_{jh} = \frac{(H-h)!(H-j)!}{(H-h-j)!H!} \text{ para (j + h} \le \text{H)}$$

$$\alpha_{jh} = 0 \text{ para (j +h > H)}$$

Notemos que $\alpha_{jh} = \alpha_{hj}$. Dado que el coeficiente α_{jh} en Ec. 5 es 0 para $h = H$, la riqueza estimada por el conjunto empírico completo $\tilde{\tau}(H) = S_{\text{obs}}$. Consideramos que la riqueza observada S_{obs} se mide con error. Este enfoque es crítico para la derivación de un estimador no condicionado de la varianza para $\tau(h)$ a $h < H$.

Dado que $\tilde{\tau}(h)$ se deriva de estimar momentos (Mao *et al*., 2004), nos referimos a éste como el estimador de la riqueza de especies *basado en momentos* $\tau(h)$ o según el método de los momentos. Es el mejor estimador en el sentido de que $\tilde{\tau}(h)/S$ llega a la varianza mínima de todos los estimadores no sesgados para $\tau(h)/S$. El estimador basado en momentos $\tilde{\tau}(h)$ se puede aproximar con una variable aleatorizada normal con un $\tau(h)$ promedio y una varianza $\sigma^2(h)$ (Mao *et al*., 2004). Por lo tanto, uno puede construir intervalos de confianza aproximados de 95% $\tilde{\tau}(h) \pm 1.96\ \tilde{\sigma}(h)$ para $\tau(h)$ con

$$\tilde{\sigma}^2(h) = \sum_{j=1}^{H} (1-\alpha_{jh})^2 s_j - \tilde{\tau}_2(h)/\tilde{S} \quad \text{(Ec. 6)}$$

en donde $\tilde{S}$ es un estimador de la riqueza total de especies desconocida. Bunge y Fitzpatrick (1993), y Colwell y Coddington (1994) revisaron (y EstimateS [Colwell, 1994-2004] calcula) varios estimadores de la riqueza. Una forma del estimador de riqueza, "Chao2" (Chao, 1989; Colwell, 1994-2004; Colwell y Coddington, 1994; Mao y Lindsay, 2003), ofrece una opción sencilla:

$$\tilde{S} = S_{obs} + \frac{(H-1)s_1^2}{2Hs_2} \quad \text{(Ec. 7)}$$

donde s_1 es el número de especies que ocurre en un solo cuadro y s_2 es el número de especies que ocurre en exactamente dos cuadros. Un enfoque altamente conservador en cuanto a la estimación de $\sigma^2(h)$ es fijar $\tilde{S} = \infty$, para que el segundo término de la Ec. 6 se vuelva insignificante.

Ugland *et al*. (2003) llegaron independientemente a proponer un estimador combinatorial de interpolación que es el equivalente matemático a la Ec. 5, pero no derivaron este resultado como lo esperado de $\tilde{\tau}(h)$ condicionado al conjunto empírico de datos. También presentan un estimador de la varianza para $\tilde{\tau}(h)$ utilizando un enfoque completamente diferente al de la Ec. 6, pero como su estimador es la varianza condicional, no es correcto usarlo para construir intervalos de confianza.

Para ilustrar la interpolación (rarefacción basada en muestras) utilizando las Ec. 5 y Ec. 6, en la Fig. 1 se grafica la riqueza estimada con bandas de confianza aproximadas de 95% para los conjuntos de datos de aves, el banco de semillas y hormigas de la Tabla I.

La teoría de la ordenación aleatoria (Coleman, 1981; Brewer y Williamson, 1994; Colwell y Coddington, 1994) puede parecer un enfoque alternativo a la estimación de la riqueza interpolada $\tau(h)$, aunque, hasta donde sabemos, la teoría de la ordenación aleatoria no se ha aplicado anteriormente a los datos de incidencia. Un estimador de ubicación aleatoria tipo Coleman es

$$\tilde{\tau}_*(h) = S_{obs} - \sum_{j=1}^{H} s_j (1-h/H)^j. \quad \text{(Ec. 8)}$$

Sin embargo, el estimador $\tilde{\tau}_*(h)$ está sesgado. La diferencia entre $\tilde{\tau}(h)$ y $\tilde{\tau}_*(h)$ se vuelve

$$\tilde{\tau}(h) - \tilde{\tau}_*(h) = \sum_{j=1}^{H} s_j [(1-h/H)^j - \alpha_{jh}]. \quad \text{(Ec. 9)}$$

Se puede demostrar que $\alpha_{jh} < (1 - h/H)^j$ tal que $\tilde{\tau}_*(h) < \tilde{\tau}(h)$, aunque la diferencia $\tilde{\tau}(h) - \tilde{\tau}_*(h)$ suele ser pequeña. Además, los estimadores de la varianza de Coleman (1981) son condicionales en el sentido de que la incertidumbre del muestreo no se toma en cuenta (véase también Smith y Grassle, 1977). La Fig. 2 presenta los estimadores de la ubicación aleatoria $\tilde{\tau}_*(h)$ y las diferencias $\tilde{\tau}(h) - \tilde{\tau}_*(h)$ calculadas con base en el conjunto de datos de las hormigas. Las diferencias son notables para h pequeñas, y pueden ser bastante grandes para ciertos conjuntos de datos con valores extremos altos [*outliers*] en los conteos de incidencia (Mao *et al*., 2004).

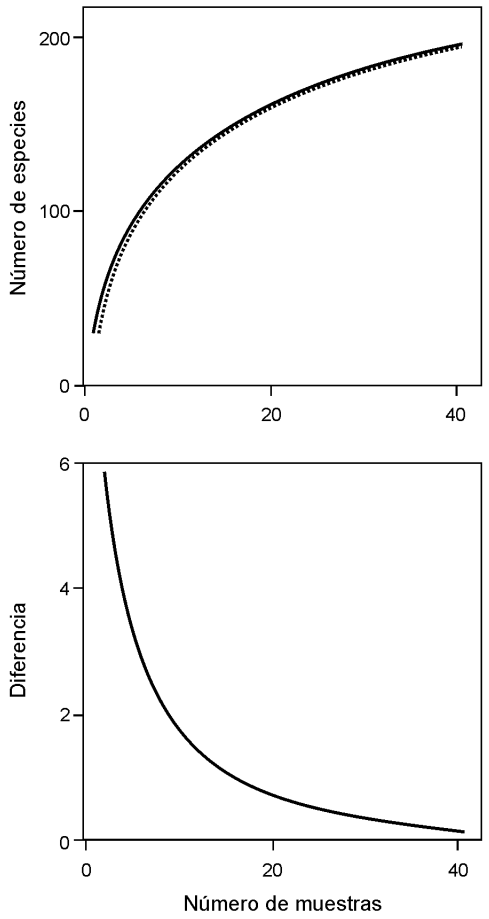

Fig. 2. Estimadores de riqueza para el conjunto de datos de las hormigas (Tabla I), comparando el estimador basado en momentos (Ec. 5) con el estimador de incidencia tipo Coleman (Ec. 8). La gráfica de arriba muestra los estimadores basados en momentos $\tilde{\tau}(h)$ (línea sólida) y los estimadores de Coleman $\tilde{\tau}_*(h)$ (línea punteada) como una función del número de cuadros h. La gráfica de abajo muestra las diferencias $\tilde{\tau}(h) - \tilde{\tau}_*(h)$ como una función de h.

Al principio de la sección anterior (*El Modelo*) aplicamos dos supuestos estadísticos con el fin de simplificar y ahora es momento de volver a examinarlos: (1) la i-ésima especie tiene la misma probabilidad φ_i de estar presente en cada cuadro y (2) los Z_{ij} son independientes para toda i y j. Por medio de ejemplos sencillos pero definitivos, demostraremos que la rarefacción basada en muestras de la Ec. 5 es robusta a estos supuestos. La Tabla II muestra dos ejemplos hipotéticos de conjuntos de muestras empíricos. En cada ejemplo seis especies se distribuyen en cuatro cuadros. Los dos ejemplos comparten la misma distribución de conteos: en ambos casos, $s_1 = 4$ y $s_2 = 2$ (cuatro especies ocurren en solamente un cuadro cada una y dos especies ocurren en precisamente dos cuadros), mientras $s_j = 0$ para toda otra $j > 0$. Así que con Ec. 5 los dos ejemplos tienen que dar las mismas curvas de rarefacción basadas en muestras, la cual depende solamente de los conteos s_j. (Dejaremos a un lado por ahora las matrices de "Abundancia", las cuales se vuelven pertinentes en la *Discusión*).

Ahora, examinamos los patrones de incidencia en los dos ejemplos, mismos que en conjunto abarcan el intervalo de posibilidades. En el Ejemplo *a*, las seis especies están desasociadas de manera no aleatoria (co-ocurren en el número mínimo de cuadros), y no hay variación en la incidencia total (totales de las columnas) entre cuadros. En cambio, en el Ejemplo *b* las seis especies están asociadas al máximo (siempre ocurren juntas) con una distribución heterogénea de ocurrencia global (variación alta en la incidencia total entre cuadros). A pesar de estos patrones extremos, el número promedio de especies es idéntico entre todas las posibles combinaciones de cuadros para $h = 1 \ldots 4$, como se muestra en los cálculos mostrados a la derecha en la Tabla II. (La correspondiente curva de rarefacción basada en muestras se presenta posteriormente en la Fig. 5). Es claro que las curvas de rarefacción basadas en muestras en si son ciegas a las probabilidades de ocurrencia no aleatoria entre cuadros y también a la falta de independencia de ocurrencia entre especies, debido a la promediación combinatoria en la Ec. 5, que se presenta explícitamente a la derecha en la Tabla II. Desde el punto de vista estadístico, es necesario requerir que los cuadros en el conjunto de muestras empírico sean verdaderamente "representativos" del conjunto de todos los cuadros posibles. Por lo tanto, se entiende que la probabilidad de presencia φ_i es el promedio de la probabilidad de presencia para todos los cuadros diferentes para la i-ésima especie.

Comparación de las curvas de rarefacción basadas en muestras

Ahora que podemos estimar intervalos de confianza rigurosos para las curvas de rarefacción basadas en muestras (Fig. 1), la comparación de dos o más de estas curvas para diferentes conjuntos de muestras con esfuerzos de muestreo comparables es sencilla. Por ejemplo, Chazdon y sus colegas (R. L. Chazdon, A. Redondo-Brenes y B. Vilchez-Alvarado, datos no publicados) muestrearon selva madura y selva secundaria en Costa Rica (Tabla I), identificando todos los tallos > 1 cm dap en 100 cuadros (de 10 x 10 m c/u, sobre una cuadrícula de 50 x 250 m) para cada tipo de selva. Ambas gráficas en la Fig. 3 muestran las curvas de rarefacción basadas en muestras con intervalos de confianza de 95% para los briznales más grandes, 2.5-5.0 cm dap. En la gráfica superior, el eje-x se dimensiona con cuadros acumulados mientras las curvas de la gráfica inferior se dimensionan con el número acumulado de tallos individuales conforme se agregan los cuadros. Las dos gráficas son diferentes porque la densidad promedio de briznales es notablemente mayor en la selva secundaria (5.1 tallos/cuadro) que en la selva madura (1.8 tallos/cuadro) donde predominan árboles más grandes.

La gráfica superior compara la densidad de especies entre los dos tipos de selva, pero en la gráfica inferior se compara la riqueza de especies (Gotelli y Colwell, 2001). En la gráfica superior, aunque los estimados de densidad de especies para la selva madura son mayores que para la selva secundaria en todos los niveles de acumulación de cuadros (todo h), las diferencias

son claramente no significativas a $P < 0.05$, dado que los intervalos de confianza se solapan. Al redimensionar las curvas con individuos en la gráfica inferior, la diferencia (en la riqueza de especies) se vuelve fuertemente significativa. Tal y como lo presentan Gotelli y Colwell (2001, y citas ahí presentadas) para la estimación de la riqueza de especies (al contrario de la densidad de especies) suele ser necesario redimensionar las curvas de rarefacción basadas en muestras, mediante individuos, con el fin de hacer el ajuste para densidades de individuos diferentes.

Extrapolación

Suele ser deseable estimar el número de especies que se encontraría (o que se habría encontrado) al colectar más muestras de un ensamblaje. Muchos usos de la extrapolación son posibles, incluyendo la futura e informada asignación de tiempo y recursos limitados, el análisis de los datos históricos cuando ya no es posible obtener más muestras, o la necesidad de hacer más "grandes" en el sentido estadístico los conjuntos de datos pequeños para su comparación con conjuntos más grandes con un esfuerzo de muestreo similar. En los términos de nuestro modelo general, la extrapolación involucra estimar $\tau(h)$ para $h > H$, donde H es el número de cuadros (u otras muestras) en el conjunto de datos empírico. El objetivo se vuelve la estimación del número de especies adicionales, $\tau(h) - \tau(H)$, que se esperaría encontrar en los cuadros adicionales $h - H$. La presentación aquí se simplificó de la de Mao *et al*. (2004) donde se presenta el desarrollo matemático completo de las técnicas de extrapolación.

Tal vez no sea sorpresa que las propiedades estadísticas de $\tau(h)$ para $h > H$, y $\tau(\infty) = S$ sean diferentes de las de $\tau(h)$ para $h \leq H$. Mientras que los datos de la incidencia observada (presencia-ausencia) provean suficiente información para que estimemos $\tau(h)$ para $h \leq H$ utilizando un estimador sencillo y basado en momentos (Ec. 5) sin restricciones sobre el número de grupos de incidencia homogéneos G, no contamos con tal estimador sencillo para $h > H$. Sin embargo, se puede desarrollar un método probabilístico, con la restricción adicional de que $G \leq H/2$ (Mao *et al*., 2004). (En la práctica, es poco probable que esta restricción cauce problemas como los ejemplos empíricos demuestran abajo.)

Nuestra estrategia es desarrollar una función $\theta(h)$ que expresa la diferencia en riqueza proporcional que se espera entre cuadros H y h, tal que

$$\tau(h) = \tau(H)\theta(h) \qquad \text{(Ec. 10)}$$

y, asintóticamente,

$$\tau(\infty) = \tau(H)\theta(\infty). \qquad \text{(Ec. 11)}$$

El modelo que desarrollamos no solo se aplica a la extrapolación ($h > H$; $\theta(h) > 1$), sino también a la interpolación ($h < H$; $\theta(h) < 1$), para la cual el método probabilístico ofrece una alternativa al método basado en momentos descrito en la sección *Interpolación*.

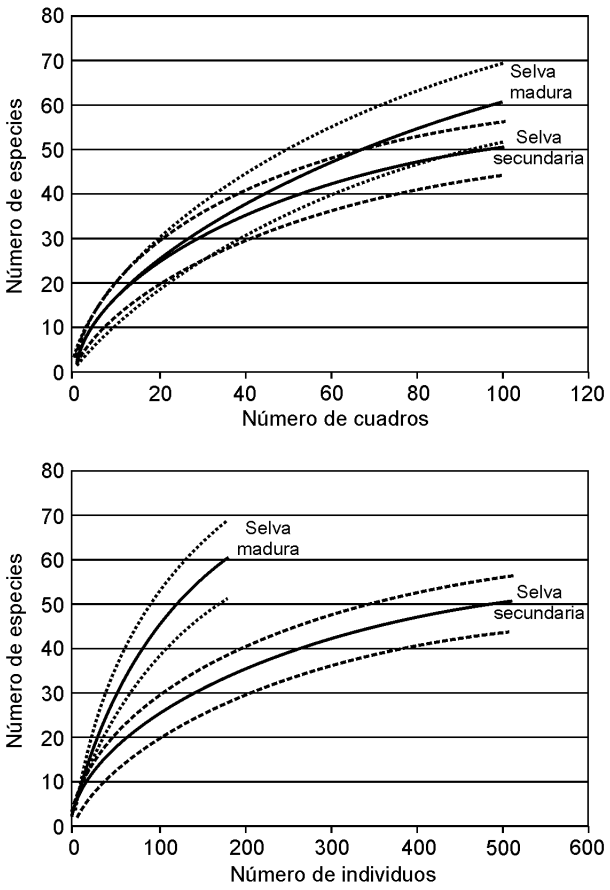

Fig. 3. Comparación de la riqueza de especies entre dos conjuntos de datos. Las dos gráficas muestran los mismos datos para briznales de la selva (Tabla I) en selva madura (línea sólida superior con la línea punteada indicando el intervalo de confianza a 95%) vs. selva secundaria (línea sólida inferior con la línea de guiones indicando el intervalo de confianza a 95%). La gráfica de arriba compara la densidad de especies (porque el eje x está dimensionado por cuadros), para la cual no hay diferencias significativas entre los dos tipos de selva. La gráfica de abajo compara la riqueza de especies (porque el eje x está dimensionado por individuos), para la cual sí hay diferencias significativas entre los dos tipos de selva.

Recuerde que π_k es el tamaño relativo del grupo del k-ésimo grupo de incidencia y Ψ_k es la *probabilidad de presencia* común de las especies dentro del k-ésimo grupo. Trabajando con la Ec. 2, definimos el *peso de mezcla* para el k-ésimo grupo, ω_k, como

$$\omega_k = \frac{\pi_k[1-(1-\psi_k)^H]}{\sum_{m=1}^{G}\pi_m[1-(1-\psi_m)^H]} \qquad \text{(Ec. 12)}$$

k=1, 2,....G.

Ahora se puede especificar la función deseada $\theta(h)$ como la suma pesada de $k = 1, 2, \ldots, G$ términos como sigue:

$$\theta(h) = 1 + \sum_{k=1}^{G}\omega_k \frac{(1-\psi_k)^H - (1-\psi_k)^h}{1-(1-\psi_k)^H}. \qquad \text{(Ec. 13)}$$

Conforme h se vuelve muy grande, la expresión $(1 - \Psi_k)^h$ se aproxima a cero para que

$$\theta(\infty) = 1 + \sum_{k=1}^{G} \omega_k \frac{(1-\psi_k)^H}{1-(1-\psi_k)^H}. \quad \text{(Ec. 14)}$$

Dado que la verdadera riqueza para H cuadros, $\tau(H)$, puede estimarse con S_{obs}, solamente tenemos que estimar $\theta(h)$ y $\theta(\infty)$ o estimar los parámetros Ψ_k y ω_k (probabilidades de presencia y pesos de mezcla) utilizados para definir $\theta(h)$ y $\theta(\infty)$. Para hacerlo, buscamos maximizar el logaritmo de la probabilidad condicional

$$L = l(\{\omega_k, \psi_k\}_{k=1}^{G}) \quad \text{(Ec. 15)}$$

de los conteos empíricos s_1, s_2 ... s_H, dada la riqueza observada S_{obs}. Los métodos que recomendamos para lograr este objetivo van más allá del alcance del presente artículo, pero aparecen completos en Mao *et al.* (2004). Aquí, resumimos la estrategia y luego la aplicamos a los conjuntos empíricos de datos de la Tabla I.

Dado un cierto número de grupos de incidencia G, se puede usar un algoritmo de maximización de lo esperado (ME) para maximizar el logaritmo de la probabilidad L (Ec. 15), resultando en un conjunto de estimadores para Ψ_k y ω_k (k = 1, 2, . . . , G) que están ajustados específicamente para G grupos (Dempster *et al*., 1977). Empezamos con G = 1, y luego continuamos evaluando la bondad del ajuste para las pruebas sucesivas con G = 1,2..., utilizando el método de gráfica de gradientes [*the gradient plot method*] de Lindsay y Roeder (1992) para evaluar en cada paso si el aumentar G resulta en una mejora en el ajuste, y el algoritmo ME para producir nuevos conjuntos de estimados en cada paso para el Ψ_k y ω_k. Un mayor número de grupos G puede aumentar el logaritmo de la probabilidad, pero un G más grande implica que se usan más parámetros para lograr el ajuste mejorado, dado que el número de parámetros independientes para la probabilidad en la Ec. 15 es $2G - 1$. Para llegar a un balance entre la bondad del ajuste y la estimación de un número menor de parámetros, seleccionamos el número de grupos G que minimiza el CAI (Criterio Akaike de Información):

$$AIC(L_G) = 2g - 1 - 2l(\{\omega_{k,G}, \psi_{k,G}\}_{k=1}^{G}). \quad \text{(Ec. 16)}$$

Una vez encontrados los mejores valores estimados para Ψ_k y ω_k con este proceso iterativo, se usan con la Ec. 13 para calcular el estimador $\hat{\theta}(h)$, mismo que se aplica entonces para estimar la riqueza para h cuadros, $\hat{\tau}(h)$, por Ec. 10. Los mismos valores estimados para Ψ_k y ω_k se pueden insertar en la Ec. 14 para estimar $\hat{\theta}(\infty)$, produciendo un estimado de la asíntota de la riqueza $\hat{\tau}(\infty)$ por la Ec. 11. La estimación de los intervalos de confianza para $\hat{\tau}(h)$ se obtiene con B remuestreos tipo *bootstrap* de la probabilidad de los conteos s_1, s_2 ... s_H dada una $\tilde{S}_{obs}$ aleatoria producida como una variable binomial aleatoria con tamaño $\hat{\tau}(\infty)$ y probabilidad $1/(1 + \hat{\theta}(\infty))$. Para cada remuestreo, la riqueza $\hat{\tau}(h)$ (Ec. 10) se calcula, los valores estimados $\hat{\tau}(h)$ se jerarquizan y los valores jerarquizados como $0.025B$ y $0.975B$ se registran como los intervalos de confianza de 95% (Mao *et al*., 2004). Notemos que este enfoque difiere fundamentalmente de los de Norris y Pollack (1996), tanto en teoría como en su cálculo. El enfoque de Norris y Pollack incurre una carga computacional tan masiva que no es un método práctico para construir los intervalos de confianza.

La Fig. 4 muestra los resultados para la extrapolación de la riqueza al triplicar el número empírico de cuadros ($3H$) para los conjuntos de datos empíricos de la Tabla I. Los valores estimados de la riqueza y los intervalos de confianza en la Fig. 3 tanto para la interpolación ($h < H$) como para la extrapolación ($h > H$) fueron producidos usando el procedimiento probabilístico antes mencionado, con B = 1000 para los intervalos de confianza *bootstrap*. La Tabla III muestra los parámetros ajustados Ψ_k y ω_k (probabilidades de presencia y pesos de mezcla) así como el número óptimo de los grupos de incidencia G (guiado por el CAI). Notemos que para los tres ejemplos empíricos, G es muy pequeña comparado con el número de cuadros H; es poco probable que la restricción que obliga a G ser menor que $H/2$ presente algún problema para los niveles razonables de intensidad de muestreo. Dado que las especies se agregan cada vez más lentamente conforme H se vuelve más grande, se espera que también el valor óptimo de G se incrementará mucho más lentamente que H.

También intentamos aplicar el método a la estimación de la riqueza asintótica S por las Ecs. 14 y 11. Desafortunada, pero no sorpresivamente, la extrapolación se vuelve más y más difícil conforme h se vuelve más y más grande. Notemos que los intervalos de confianza se vuelven más y más amplios conforme h se

Tabla III. Grupos de incidencia, probabilidades de presencia y pesos de mezcla para la extrapolación basada en la probabilidad para conjuntos empíricos de datos.

Grupo (k)	Probabilidad de presencia Ψ_k	Peso de mezcla ω_k
Aves de zonas templadas, G = 4		
1	0.0300	0.4589
2	0.1328	0.2787
3	0.2991	0.1401
4	0.5038	0.1224
Banco de semillas de zonas tropicales, G = 4		
1	0.0195	0.2847
2	0.0633	0.4792
3	0.1773	0.0890
4	0.4066	0.1471
Hormigas de zonas tropicales, G = 5		
1	0.0252	0.3904
2	0.0908	0.2874
3	0.2893	0.1849
4	0.5465	0.1218
5	0.8584	0.0155

Notas: Los valores del cuadro fueron calculados para la extrapolación de la curva de acumulación de especies a tres veces el tamaño empírico de la muestra ($3H$), ajustado con el método basado en probabilidad explicado en el texto (*Extrapolación*) con el número de grupos de incidencia (G) optimizado por CAI. Las curvas de acumulación de especies extrapoladas se encuentran en la Fig. 4. Los conjuntos de datos se describen en la Tabla I.

incrementa más allá de *H*. En virtud de que el estimador probabilístico para $\tau(\infty)$ suele ser no confiable (Mao *et al*., 2004), recomendamos que se limite el uso de la extrapolación para extender (es decir, duplicar o triplicar) el número de muestras en conjuntos de datos empíricos, tal y como se mostró en la Fig. 3.

Discusión

El modelo introducido en este artículo ofrece un marco teórico unificado para la conceptualización y análisis de la riqueza de especies en el contexto de muestras de incidencia repetida (presencia-ausencia u ocurrencia) de comunidades biológicas. Para un esquema dado de muestreo, los patrones de incidencia en muestras de las comunidades naturales se ven afectadas por cuando menos tres fuentes mayores de heterogeneidad. La primera y más obvia es la variación entre las especies en cuanto a qué tan comunes o raras son (abundancia relativa), misma que se traduce, en general, en variación entre especies en su frecuencia de ocurrencia. La segunda fuente de heterogeneidad es la variación entre muestras en la abundancia total de individuos (agregación espacial o temporal que es concordante entre especies), que a su vez se traduce en la variación entre muestras en el número total de ocurrencias de las especies. La tercera fuente de heterogeneidad es la asociación o desasociación de especies, entre muestras, la cual se traduce en patrones no aleatorios de co-ocurrencia de especies. El segundo y tercer tipo de heterogeneidad suelen ser difíciles de separar; tomados juntos representan lo que generalmente se caracteriza como heterogeneidad (*patchiness*) en el espacio o en el tiempo entre muestras.

Los modelos de rarefacción basados en individuos explican, explícitamente, la abundancia relativa de las especies (Hurlbert, 1972). Nuestro modelo basado en incidencias permite niveles arbitrarios de heterogeneidad entre especies en su ocurrencia total (abundancia relativa) por tratar a las ocurrencias de las especies como resultado de una mezcla de distribuciones binomiales. En este modelo mixto, se asume que cada especie tiene su propia probabilidad de presencia específica y por lo tanto se supone que sigue su propia distribución binomial en cuanto al registro de presencia y ausencia entre muestras. (Este es el supuesto menos exigente que uno puede hacer para los datos de incidencia.) En efecto, las distribuciones binomiales específicas a las especies son luego "clasificadas" (por el algoritmo de mezcla-ajuste) en grupos de probabilidades de presencia aproximadamente homogéneas. El modelo completo es, entonces, una mezcla de distribuciones binomiales, cada una pesada por el número de especies en su grupo correspondiente (Mao *et al*., 2004).

La rarefacción basada en individuos y la basada en muestras tienen supuestos crucialmente distintos en cuanto a la heterogeneidad, lo cual es entendido de mejor manera al comparar los dos métodos con el mismo conjunto de datos. Dada una matriz empírica de abundancia, tal como las de los ejemplos hipotéticos de la Tabla II (el lado izquierdo de cada ejemplo), el vector de los totales de la fila (especies; 9, 6, 1, 4, 2, 3 en la

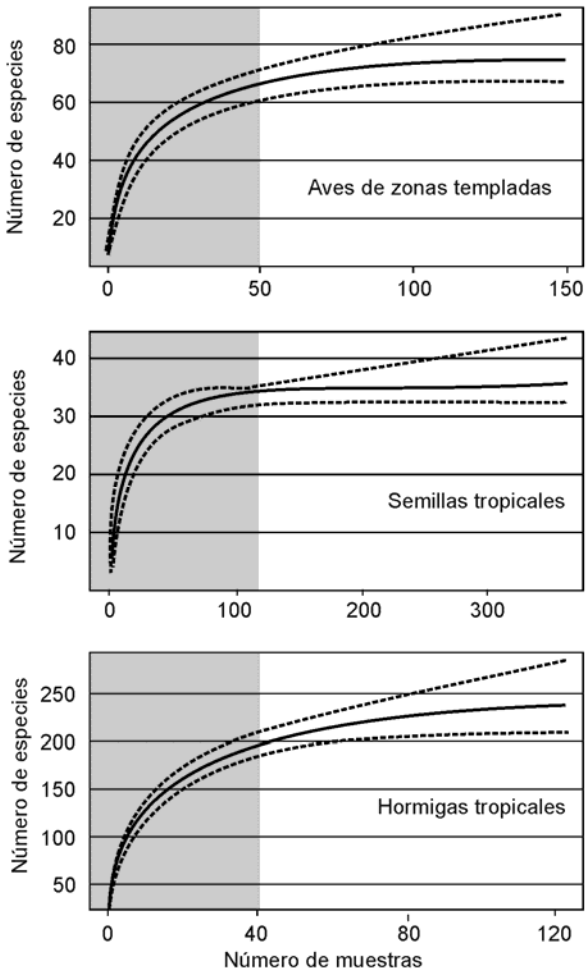

Fig. 4. Extrapolación de las curvas de acumulación de especies para tres conjuntos de datos empíricos (véase Tabla I) a tres veces el tamaño de la muestra empírica. Los valores de la riqueza de especies esperados (líneas sólidas) fueron calculados utilizando el estimador basado en probabilidad de Ec. 10 con intervalos de confianza generados con el método *bootstrap* de 95% (para la interpolación así como la extrapolación; líneas de guiones). El área sombreada indica el número de muestras en el conjunto de datos empírico (*H*).

Tabla II) puede usarse para producir una curva de rarefacción basada en individuos para el conjunto de muestras. La matriz de incidencias correspondiente (la parte de en medio de cada ejemplo en la Tabla II) puede usarse para producir una curva de rarefacción basada en muestras (calculada al lado derecho de cada ejemplo) para el mismo conjunto de muestras.

La Fig. 5 muestra ambos tipos de curvas de rarefacción, basadas en los ejemplos hipotéticos de la Tabla II. Cuando ambos tipos de curva de rarefacción se dimensionan con el número de muestras agrupadas, las dos curvas serán idénticas solamente si los *individuos* de todas las especies ocurren aleatoria e independientemente entre las muestras en el mismo conjunto de muestras. Si los individuos tienden a agregarse (de manera no aleatoria) entre muestras (intra-especie), la curva de rarefacción basada en muestras tiene que encontrarse debajo de la curva de rarefacción basada en abundancia

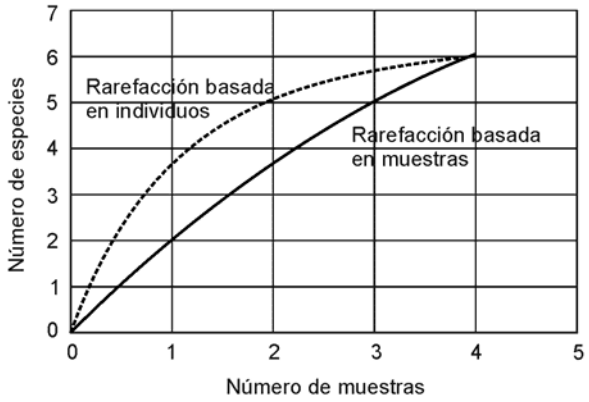

Fig. 5. Curvas de rarefacción basadas en individuos (línea punteada) vs. en muestras (línea sólida) para los datos hipotéticos de la Tabla II. La curva basada en muestras se generó utilizando el estimador basado en momentos (Ec. 5 o los valores idénticos en la Tabla II). La curva basada en individuos se generó utilizando el estimador clásico de rarefacción de Hurlbert (1972).

(como en la Fig. 5) (Coleman, 1981; Colwell y Coddington, 1994; Gotelli y Colwell, 2001). Esto ocurre porque la agregación de los individuos produce una matriz de incidencia con un número menor de registros de presencia y uno mayor de registros de ausencia en comparación con una distribución aleatoria del mismo número de individuos entre muestras, de tal manera que se tienen que agregar más muestras en la curva basada en muestras que en la basada en individuos para llegar a un cierto nivel de riqueza. Para captar esto, imagine lanzar 30 pelotas (individuos de una sola especie) aleatoriamente a 10 cajas, éstas con los números 1 al 10. Algunas cajas pueden quedarse vacías, pero no es probable que la mitad de ellas se quede vacía. Ahora, tome todas las pelotas de las cajas con números pares y distribúyalas en las cajas con números impares. Las pelotas ahora están estadísticamente agregadas y hay un número menor de registros de presencia y un número mayor de registros de ausencia que antes.

En virtud de que la rarefacción basada en individuos no toma en cuenta tal heterogeneidad, generalmente sobre-estima la riqueza esperada para las muestras rarificadas [*rarefied*] (Fig. 5). Nótese que el patrón de la abundancia de especies (totales de las filas en las matrices de Abundancia) en la Tabla II está arreglado para que sea idéntico para los dos ejemplos, mientras la distribución actual de individuos entre los cuadros es sustancialmente diferente. Las curvas de rarefacción basadas en individuos (Fig. 5) para los dos ejemplos son, sin embargo, idénticas.

En cambio, las curvas de rarefacción basadas en muestras reflejan implícitamente los niveles empíricos de agregación de individuos dentro de la especie al considerar únicamente a la incidencia, dando así un estimado realista del número de especies a encontrarse en conjuntos de muestras del mundo real (Colwell y Coddington, 1994; Chazdon *et al*., 1998; Gotelli y Colwell, 2001; Ugland *et al*., 2003). Supongamos que se divide un área de estudio en 1000 cuadros. Muestreamos 50 cuadros al azar y contamos el número de especies que se encuentra en cada uno. Una especie en particular puede encontrarse en algunos de los cuadros muestreados y no en otros, y sus individuos pueden encontrarse agregados de manera no aleatoria entre los cuadros. Solo necesitamos que las 50 unidades estén seleccionadas verdaderamente al azar del total de 1000, para que la estimación empírica de la probabilidad de presencia de una especie en estas 50 unidades sea cercana a la verdadera probabilidad de presencia para dicha especie en los 1000 cuadros, para el tamaño especificado del cuadro y el nivel empírico de agregación individual. Dado que las distribuciones agregadas espaciales (y temporales) son extremadamente comunes, esta propiedad de la rarefacción basada en muestras es muy general.

La estadística fundamental para los estimadores basados en el modelo son los conteos, o frecuencias de ocurrencia, de especies en un conjunto de muestras. Dado que la curva de rarefacción basada en muestras depende solamente de patrones de incidencia (promedios), se puede modelar precisamente con estos conteos para los conjuntos empíricos de datos con cualquier grado de heterogeneidad, tal y como se demostró con los ejemplos en la Tabla II y la Fig. 5. Algunas muestras pueden tener un alto número de ocurrencias y otras pueden tener números bajos (la segunda fuente de la heterogeneidad), pero debido a la promediación combinatoria, lo anterior no tiene un efecto sobre la curva de rarefacción promedio. Por la misma razón, la asociación o desasociación de las ocurrencias de especies (la tercera fuente de la heterogeneidad) no afecta la curva de rarefacción promedio, ni se refleja en los conteos.

Usando el modelo general como marco de referencia, desarrollamos estimadores tanto para la interpolación (o rarefacción basada en muestras) entre cero y la riqueza del conjunto total de muestras en un conjunto empírico de datos, como para la extrapolación, o la proyección de la riqueza más allá del conjunto de datos con el fin de predecir el número esperado de especies en un número mayor de muestras del mismo ensamblaje.

La interpolación se ha llevado a cabo rutinariamente en el pasado, con los datos basados en la incidencia, mediante el submuestreo al azar del conjunto de datos, reteniendo la integridad del muestreo (en vez de agrupando todas las muestras para luego extraer ocurrencias al azar) (Colwell, 1994-2004; Colwell y Coddington, 1994). Hasta hace poco (Ugland *et al*., 2003; Mao *et al*., 2004) no existía un método analítico para estimar con precisión el número de especies en un subconjunto de muestras de un conjunto de datos basado en la incidencia. En cambio, el problema correspondiente para las muestras basadas en la abundancia (rarefacción clásica) se resolvió hace tres décadas (Hurlbert, 1972; Heck *et al*., 1975). Peor aún, el problema de establecer los intervalos de confianza alrededor de las curvas de acumulación basadas en muestras hasta ahora había seguido sin resolverse en absoluto, limitando severamente la comparación de las curvas de diferentes comunidades o tratamientos. Nuestro estimador de riqueza basado en momentos (Ec. 5) para el problema de la interpolación, con su estimador de la varianza (Ec. 6)

atiende a estos problemas de manera rigurosa, y además atiende precisamente la expectativa que se produce al aleatorizar la secuencia de muestras, con intervalos de confianza legítimos para cada punto a lo largo de la curva. Como se mostró en la Fig. 3, estos intervalos de confianza finalmente hacen posible la comparación rigurosa entre curvas de rarefacción basadas en muestras. El estimador basado en momentos para la rarefacción basada en muestras, con intervalos de confianza, se incluye en la versión 7 de EstimateS (Colwell, 1994-2004).

Usando el mismo marco teórico, derivamos la función de ordenación aleatoria (Coleman) para los datos de incidencia (Ec. 8), misma que parece no haber sido examinada anteriormente. La curva Coleman (Coleman, 1981) hasta ahora solamente se ha aplicado al caso de muestras basadas en la abundancia (cuantitativas), para la cual aproxima la curva esperada para la rarefacción basada en individuos (Brewer y Williamson, 1994; Colwell y Coddington, 1994). Asimismo, encontramos que la curva Coleman basada en incidencias aproxima la verdadera curva de acumulación de especies (rarefacción basada en muestras) para muestras de incidencia. Sin embargo, la lógica subyacente a la curva de Coleman basada en incidencia solamente tiene sentido para la interpolación, el estimador es sesgado (de manera notable si alguna especie es altamente dominante, tal y como se indica con valores extremadamente altos para j en los conteos de incidencia s_j), y los estimadores de la varianza disponibles no son apropiados para construir los intervalos de confianza. Por estas razones, para la interpolación preferimos el estimador con base en momentos en vez de la curva de Coleman basada en incidencias.

La extrapolación de las curvas de acumulación de especies solamente se había intentado anteriormente con el ajuste de funciones como la función asintótica Michaelis-Menten o varias funciones no asintóticas (Soberón y Llorente, 1993;Colwell y Coddington, 1994). En el presente artículo, para el problema de la extrapolación, desarrollamos un modelo probabilístico que depende del ajuste de la distribución de los conteos observados para el modelo mixto binomial. El número de grupos de incidencia requeridos por el modelo se optimiza utilizando el Criterio Akaike de Información (CAI) para equilibrar la bondad del ajuste y la complejidad del modelo (el número de parámetros). Los intervalos de confianza *bootstrap* también se pueden calcular como se expuso en la sección *Extrapolación*. Los cálculos y algoritmos para tanto la riqueza esperada como su intervalo de confianza *bootstrap* requieren de computaciones sofisticadas y complejas, pero programas de cómputo para este fin están disponibles de C.X. Mao.

Usando el mismo método probabilístico de extrapolación, uno puede, en teoría estimar el número adicional de especies que un conjunto de muestras infinitamente grande y del mismo ensamblaje rendiría: la asíntota de la curva de acumulación de especies (Ecs. 11 y 14). En la práctica, concluimos que nuestro método probabilístico de extrapolación es muy útil para los problemas de estimación que suponen la duplicación o triplicación del número empírico de muestras. Desafortunadamente, en su forma actual, el método no parece ser una manera confiable para estimar la riqueza asintótica (Mao *et al*., 2004), pero esperamos que nuestros esfuerzos puedan inspirar trabajo futuro en este problema, tan importante y desalentador.

El método probabilístico modela de manera simultánea el problema de la interpolación y la misma técnica *bootstrap* se puede usar para estimaciones de riqueza interpoladas (rarefacción basada en muestras), tal y como se ilustra en la Fig. 4. Sin embargo, el estimador basado en momentos es más sencillo e intuitivo (Ec. 5) y sus intervalos de confianza asociados (basados en la Ec. 7) tienen un desempeño igual de bueno y, por lo tanto, son preferidos.

Agradecimiento

Agradecemos a F. He y dos árbitros anónimos sus comentarios, y a A. M. Ellison por animarnos. Gracias a R.L. Chazdon por compartir sus datos no publicados. Este trabajo fue apoyado por el proyecto DEB-0072702 del US-NSF otorgado a R. K. Colwell.

B. Delfosse, la traductora, agradece a J. Laborde su puntual asesoría.

Bibliografía

Brewer, A. & M. Williamson. 1994. A new relationship for rarefaction. *Biodiversity and Conservation*, **3**: 373-379.

Bunge, J. & M. Fitzpatrick. 1993. Estimating the number of species: a review. *Journal of the American Statistical Association*, **88**: 364-373.

Butler, B. J. & R. L. Chazdon. 1998. Species richness, spatial variation, and abundance of the soil seed bank of a secondary tropical rain forest. *Biotropica*, **30**: 214-222.

Chao, A. 1989. Estimating population size for sparse data in capture-recapture experiments. *Biometrics*, **45**: 427-438.

Chazdon, R. L., R. K. Colwell, J. S. Denslow & M. R. Guariguata. 1998. Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of NE Costa Rica. Pages 285-309 in F Dallmeier and J. A. Comiskey, editors. *Forest biodiversity research, monitoring and modeling: conceptual background and Old World case studies*. Parthenon Publishing, Paris, France.

Coleman, B. D. 1981. On random placement and speciesarea relations. *Mathematical Biosciences,* **54**: 191-215.

Colwell, R. K. 1994-2004. Estimates: statistical estimation of species richness and shared species from samples. (http:// viceroy.eeb.uconn.edu/estimates). [Persistent URL: (http:// purl.oclc.org/estimates).]

Colwell, R. K. & J. A. Coddington. 1994. Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society, Series B*, **345**: 101-118.

Dempster, A. P, N. M. Laird & D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B,* **39**: 1-22.

Dorazio, R. M. & J. A. Royle. 2003. Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics*, **59**: 351-364.

Gotelli, N. & R. K. Colwell. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, **4**: 379-391.

Heck, K. L., Jr., G. van Belle & D. Simberloff. 1975. Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. *Ecology,* **56**: 1459-1461.

Holdridge, L. R., W. G. Grenke, W. H. Hatheway, T. Liang & J. A. Tosi. 1971. *Forest environments in tropical life zones*. Pergamon Press, Oxford, UK.

Hurlbert, S. H. 1971. The nonconcept of species diversity: a critique and alternative parameters. *Ecology*, **52**: 577-586.

Lindsay, B. G. & K. Roeder. 1992. Residual diagnostics for mixture models. *Journal of the American Statistical Association*, **87**: 785-794.

Longino, J., R. K. Colwell & J. A. Coddington. 2002. The ant fauna of a tropical rainforest: estimating species richness three different ways. *Ecology,* **83**: 689-702.

Mao, C. X., R. K. Colwell & J. Chang. 2004. *Estimating species accumulation curves using mixtures*. Technical report, Department of Statistics, University of California, Riverside, California, USA.

Mao, C. X. & B. G. Lindsay. 2003. *Estimating the population size: heterogeneity, nonidentifiability and regular- T ization*. Technical Report, University of California, Riverside, California, USA.

Norris, J. L. & K. H. Pollock. 1996. Nonparameteric MLE under two closed capture-recapture models with heterogeneity. *Biometrics*, **52**: 639-649.

Rosenzweig, M. L. 1995. *Species diversity in space and time*. Cambridge University Press, Cambridge, UK.

Scheiner, S. M. 2003. Six types of species-area curves. *Global Ecology and Biogeography*, **12**: 441-447.

Simberloff, D. 1972. Properties of the rarefaction diversity measurement. *American Naturalist*, **106**: 414-418.

Smith, W. & J. Grassle 1977. Sampling properties of a family of diversity measures. *Biometrics*, **33**: 283-292.

Soberón, M. J. & B. J. Llorente. 1993. The use of species accumulation functions for the prediction of species richness. *Conservation Biology*, **7**: 480-488.

Ugland, K. I., J. S. Gray & K. E. Ellingsen. 2003. The species-accumulation curve and estimation of species richness. *Journal of Animal Ecology*, **72**: 888-897.