CAPÍTULO 7:

Un nuevo método estadístico para la evaluación de la similitud en la composición de especies con datos de incidencia y abundancia

Anne Chao[1],

Robin L. Chazdon[2*],

Robert K. Colwell[2]

Tsung-Jen Shen[1]

[1]. Institute of Statistics, National Tsing Hua University, Hsin-Chu, Taiwan

[2]. Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, USA

* chazdon@uconn.edu

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: 85 – 96.

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

UN NUEVO MÉTODO ESTADÍSTICO PARA LA EVALUACIÓN DE LA SIMILITUD EN LA COMPOSICIÓN DE ESPECIES CON DATOS DE INCIDENCIA Y ABUNDANCIA*

Anne Chao, Robin L. Chazdon,
Robert K. Colwell & Tsung-Jen Shen

* Traducción del artículo publicado en*: Ecology Letters* (2004), **8**: 148-159.

Resumen: Los índices clásicos Jaccard y Sørensen de similitud en la composición de especies (y otros índices que dependen de las mismas variables) son notoriamente sensibles al tamaño de la muestra, especialmente aquellos ensamblajes con numerosas especies raras. Además, dado que estos índices se basan exclusivamente en datos presencia-ausencia, no hay estimadores precisos para ellos. Ofrecemos una derivación probabilística para las formas clásicas basadas en incidencia de estos índices y extendemos este método para formular nuevos índices tipo Jaccard o Sørensen basados en datos de la abundancia de especies. Luego proponemos estimadores para estos índices, los cuales incluyen el efecto de las especies compartidas no vistas y que se basan en datos de muestreos de incidencia o de abundancia (replicados). En las simulaciones de muestreo, estos nuevos estimadores demuestran ser menos sesgados que los índices clásicos cuando falta una proporción sustancial de especies en las muestras. Utilizando conjuntos de datos empíricos y ricos en especies, demostramos como la incorporación del efecto de especies compartidas pero no vistas no solamente incrementa la exactitud, sino también puede afectar la interpretación de los resultados.

Palabras Clave: datos de abundancia, diversidad beta, biodiversidad, complementariedad, datos de incidencia, especies compartidas, estimadores de similitud, índice de similitud, solapamiento de especies, sucesión.

A new statistical approach for assessing similarity of species composition with incidence and abundance data

Abstract: The classical Jaccard and Sørensen indices composotopnal similarity (and others indices that depend upon the same variables) are anotoriously sensitive to sample size, especially for assemblages with numerous rare species. Furthers, because these indices are based solely on presence-absence data, accurate estimators for them are unattainable. We provide a probabilistic derivation for the classic, incidende-based forms of these indices and extend this approach to formulate new Jaccerd-type of Sørensen-type indices based on species abundance data. We then propose estimators for these indices thatinclude the effect of unseen shared species, based on either (replicated) indicende- or abundance- based sample data. In sampling simulations, these new estimator prove to be considerably less biased than classic indices when a substantial proportion od species are missing from samples. Based on species-rich empirical datasets, we show how incorporating the effect of unseen shared species not only increases accuracy but also can change the interpretation of results.

Key words: Abundance data, beta diversity, biodiversity, complementarity, indicence data, shared species, similarity stimators, similarity index, species overlap, succession.

Introducción

Los ecólogos quienes llevan a cabo el registro de la riqueza de especies, desde hace mucho se han dado cuenta de que es casi imposible detectar a todas las especies y determinar sus abundancias relativas con un número limitado de muestras o una intensidad limitada de muestreo. Limitaciones de muestreo crean retos para estimar con precisión la diversidad alfa, el número de especies dentro de ensamblajes locales y aproximadamente homogéneos, particularmente para los ensamblajes con una riqueza específica alta y una fracción grande de especies raras (Colwell y Coddington, 1994; Chazdon *et al.,* 1998; Colwell *et al.,* 2004; Magurran, 2004). Para enfrentar este reto, se han desarrollado varios métodos para estimar la riqueza de especies a partir

de los datos de muestreo, o por la extrapolación de las curvas de acumulación de especies o con la aplicación de métodos no paramétricos (véanse reseñas por Bunge y Fitzpatrick, 1993; Colwell y Coddington, 1994; Magurran, 2004; Chao, en prensa). Este último enfoque involucra la estimación de *las especies no vistas* (las especies que probablemente están presentes en una muestra homogénea y más grande del ensamblaje, pero que no se encuentran en los datos de la muestra actual). Dado que los estimados de las especies no vistas se basan en el número de especies raras observadas dentro de las muestras (Colwell y Coddington, 1994; Chazdon *et al*., 1998), para estimar la riqueza se requiere de datos de la abundancia o de muestras de la incidencia replicadas. En los estimadores de la riqueza más sencillos (p. ejem. Chao1, Chao2 o los estimadores *jack-knife*), las especies raras se clasifican como especies con una abundancia total de 1 (*singletons*) o 2 (*doubletons*) en una muestra basada en la abundancia, y se encuentran en solamente una unidad de muestreo (únicos = *uniques*) o en exactamente dos unidades de muestreo (duplicados = *duplicates*) en los datos de incidencia con muestreo replicado. El estimador de cobertura basada en abundancias (*abundance-based coverage estimator*: ACE) utiliza información adicional basada en aquellas especies con diez o menos individuos en la muestra (Chao *et al*., 1993) y el correspondiente estimador basado en incidencia (*incidence-based coverage estimator*: ICE) se basa en las especies que ocurren en diez o menos unidades de muestreo (Lee y Chao, 1994; Chazdon *et al*., 1998; Magurran, 2004).

Las mismas limitaciones que se aplican a la estimación de la diversidad alfa de los ensamblajes de especies, se aplican de igual manera a la estimación de la diversidad beta o la disimilitud (complementariedad, recambio o distancia) entre dos ensamblajes. El índice Jaccard de similitud y el muy relacionado índice Sørensen son los dos más viejos y ampliamente utilizados para la valoración de la similitud en la composición de los ensamblajes (a veces llamado 'solapamiento de especies') y, por lo tanto, su complemento, la falta de similitud. Ambos se basan en la presencia/ausencia de especies en ensamblajes pareados, y son cálculos sencillos (Magurran, 2004). Existen muchos otros índices de la similitud que se basan en la misma información: el número de especies compartidas por dos muestras y el número de especies únicas en cada muestra (Legendre y Legendre, 1998), y nuevos índices siguen apareciendo (p. ejem. Lennon *et al*., 2001). Una versión modificada del índice Sørensen fue desarrollada por Bray y Curtis (1957), con base en datos de abundancia (también conocido como el índice Sørensen de la abundancia; Magurran, 2004), y un gran número de otros índices basados en abundancias se han desarrollado (Legendre y Legendre, 1998), incluyendo el ampliamente aplicado índice Morisita–Horn (Magurran, 2004).

A pesar de su amplia aplicación en los estudios ecológicos, los índices clásicos de Jaccard y Sørensen, cuando calculados con datos de muestreo, tienen un desempeño pobre como medidas de la similitud entre ensamblajes diversos que incluyen una fracción sustanciosa de especies raras (Wolda, 1981; Colwell y Coddington, 1994; Plotkin y Muller-Landau, 2002), dado que se asume que los datos de muestreo (usualmente erróneamente) son representaciones verdaderas y completas de la composición del ensamblaje. [En efecto, con pocas excepciones (p. ejem. Grassle y Smith, 1976; MacKenzie *et al*., 2004), casi todos los métodos actuales para medir la similitud parten de este supuesto.] En general, como demostraremos con simulaciones, es probable que estas medidas subestimen severamente la verdadera similitud entre dos ensamblajes (genuinamente similares) que contienen numerosas especies raras. Dado que muchas especies quedan fuera de la muestra, es probable que las especies raras que aparecen en una muestra sean diferentes a las que aparecen en la otra muestra, aun cuando todas estén realmente presentes en ambos ensamblajes. Problemas similares surgen al comparar dos muestras de tamaños notablemente diferentes: sencillamente porque la muestra más pequeña tiene un número menor de individuos o de unidades de muestreo, puede que no tenga especies que aparecen en la muestra más grande. En breve, la subestimación de la similitud ocurre por no tomar en cuenta las especies compartidas pero *no vistas*.

En principio, la sobre-estimación de la similitud también puede ocurrir al comparar comunidades submuestreadas de dominancia alta en las cuales las especies comunes están ampliamente distribuidas y en donde las especies raras tienden a ser endémicas localmente. En este caso, dos muestras pueden dar las mismas pocas especies comunes, pero no revelan las especies raras que diferenciarían los ensamblajes de contar con muestras más grandes (Colwell y Coddington, 1994; Ruokolainen y Tuomisto, 2002 discuten un posible ejemplo). Sin embargo, en casi todos los casos que hemos examinado cuantitativamente, la rareza (o por naturaleza o por tratarse de un tamaño de muestra pequeño) incrementa la posibilidad de que una especie esté erróneamente ausente de una muestra pero no de otra, introduciendo equivocadamente así un sesgo negativo a los índices de similitud. [Fisher (1999, Fig. 8) llega a la misma conclusión para varios conjuntos de datos, basado en pruebas de rarefacción.] Además, para los nuevos índices que presentamos aquí, se puede demostrar teóricamente que el sesgo de muestreo, cuando presente, siempre es negativo. [Los autores demuestran el sesgo negativo esperado matemáticamente (A. Chao, R. L. Chazdon, R. K. Colwell y T.-J. Shen, datos no publicados); se puede probar para cualquiera de los modelos de abundancia dados en Magurran (2004) y en Plotkin y Muller-Landau (2002).]

Recientemente, se ha intensificado el interés en el desarrollo y evaluación de los índices para medir la diversidad beta, o la tasa de recambio, de ensamblajes de especies (Duivenvoorden, 1995; Lennon *et al*., 2001; Arita y Rodríguez, 2002, 2004; Condit *et al*., 2002; Plotkin y Muller-Landau, 2002; Koleff *et al*., 2003; Rodríguez y Arita, 2004), subrayando la necesidad de estimadores estadísticos robustos para poder inferir la similitud de la composición a partir de los datos de muestreo. El aumento en el recambio de especies (simi-

litud decreciente) conforme se incrementa la distancia entre sitios puede reflejar patrones espaciales de dispersión o podrían resultar del aumento en la heterogeneidad ambiental a escalas mayores (Harte *et al*., 1999; Hubbell, 2001; Balvanera *et al*., 2002; Chave y Leigh, 2002; Condit *et al*., 2002; Duivenvoorden *et al*., 2002; Ruokolainen y Tuomisto, 2002; Rodríguez y Arita, 2004; Valencia *et al*., 2004). Desafortunadamente, la mayoría de los índices de diversidad beta dependen de la misma información que los índices clásicos de Jaccard y Sørensen, y comparten las limitaciones arriba mencionadas.

Con este problema en mente, Plotkin y Muller-Landau (2002) desarrollaron un índice de similitud tipo Sørensen para conteos de abundancia utilizando un enfoque 'paramétrico' que depende de la distribución gama para caracterizar la estructura de las abundancias de las especies. Condit *et al*. (2002) adoptan un método para medir la diversidad beta usando el índice de 'codominancia' *F* (*codominance index* F) de Leigh *et al*. (1993); la probabilidad de que dos individuos seleccionados al azar, cada uno de un diferente ensamblaje sean la misma especie. Aunque esta medida está basada en los datos de la abundancia, *F*, en si, no es un índice de similitud estadísticamente válido. Para dos ensamblajes idénticos con muchas especies, *F* tiende a 0. Además, es posible para cualquier par de ensamblajes tener un valor de *F* de 0 a 1, dependiendo de cuántas especies están presentes y de los patrones de la abundancia relativa. Sin embargo, es posible normalizar *F* para producir un índice de similitud válido. Chave y Leigh (2002) señalan que el índice Morisita–Horn es una versión normalizada de *F*.

Empezamos por desarrollar un nuevo método probabilístico para los índices clásicos Jaccard y Sørensen basados en la incidencia. Posteriormente, extendemos este enfoque para formular índices tipo Jaccard y tipo Sørensen que consideran las abundancias de las especies. A cambio de Plotkin y Muller-Landau (2002), adoptamos una estrategia no paramétrica que no requiere de ningún supuesto en cuanto a las distribuciones de la abundancia de las especies. Luego, proponemos un método para estimar tanto los índices Jaccard y Sørensen basados en incidencia y abundancia a partir de datos de muestreo, incorporando el efecto de las especies compartidas *no vistas*.

Después, llevamos a cabo simulaciones de muestreo con conjuntos de datos empíricos con el fin de evaluar el desempeño de los índices clásicos de Jaccard y Sørensen; sus nuevas contrapartes Jaccard y Sørensen basadas en abundancias; y los correspondientes estimadores Jaccard y Sørensen. Demostramos que la incorporación del efecto de las especies no vistas disminuye sustancialmente el sesgo de tamaño de muestra de estos estimadores y mejora su utilidad para inferir la similitud (o su complemento, la disimilitud) entre ensamblajes hiper-diversos en los cuales una porción grande de sus especies no se registra en las muestras. Finalmente, ilustramos una aplicación del nuevo índice Jaccard basada en abundancias y el estimador Jaccard basada en abundancias, usando datos de un estudio sucesional de las abundancias de árboles, plántulas y briznales para especies del dosel. Con base en conjuntos de datos para ensamblajes ricos de insectos y plantas tropicales, demostramos como la incorporación del efecto de especies compartidas no vistas no solamente incrementa la exactitud, pero también puede cambiar la interpretación de los resultados.

El desarrollo de los nuevos índices y estimadores

Los índices clásicos de Sørensen y Jaccard

Los índices clásicos de Sørensen y Jaccard dependen de tres sencillos conteos de incidencia: el número de especies compartidas por dos ensamblajes y el número de especies únicas en cada ensamblaje. Se ha vuelto tradición referirse a estos conteos como *A*, *B* y *C*, respectivamente (Tabla I). Los índices clásicos Jaccard y Sørensen para los conteos de incidencia entonces son

$$J_{clas} = \frac{A}{A+B+C} \qquad (1) \quad \text{y}$$

$$L_{clas} = \frac{2A}{2A+B+C} \qquad (2)$$

(Usamos *L* para el índice de Sørensen para evitar la confusión con la *S* para especies.) Hay una relación cercana monotónica entre los dos índices: $L_{clas} = 2J_{clas}/(J_{clas} + 1)$ y $J_{clas} = 1/(2/L_{clas} - 1)$.

Asuma que hay S_1 especies en el Ensamblaje 1 y S_2 especies en el Ensamblaje 2. Que el número de especies compartidas sea S_{12}. Entonces, los conteos de incidencia *A, B, C* en la Tabla I corresponden a: $A = S_{12}$, $B = S_1 - S_{12}$, y $C = S_2 - S_{12}$. Sustituyendo estas expresiones en las ecuaciones 1 y 2 tenemos una manera alternativa de escribir los índices clásicos que serán requeridos para los próximos pasos en el desarrollo de los nuevos índices:

$$J_{clas} = \frac{A}{A+B+C} = \frac{S_{12}}{S_1+S_2-S_{12}} \qquad (3) \quad \text{y}$$

$$L_{clas} = \frac{2A}{2A+B+C} = \frac{2S_{12}}{S_1+S_2} \qquad (4)$$

Tabla I. Conteos de clasificación de especies utilizados en los índices clásicos

	Ensamblaje 2	
	Presente	**Ausente**
Ensamblaje 1		
Presente	*A*	*B*
Ausente	*C*	-

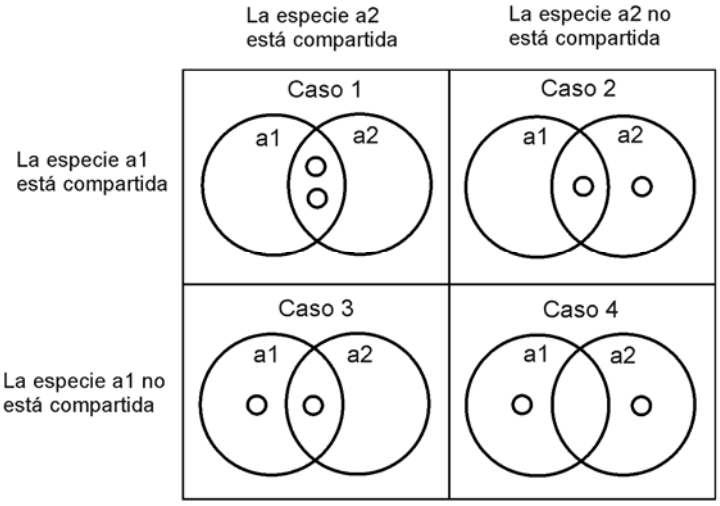

Fig. 1. Una representación gráfica del significado de especies compartidas para dos ensamblajes. El Ensamblaje 1 (a1) es gris, el Ensamblaje 2 (a2) es blanco. El punto gris representa una especie seleccionada al azar del Ensamblaje 1 y el punto blanco representa una especie seleccionada al azar del Ensamblaje 2. El Caso 1 es el único caso en el que ambas especies están compartidas (pero no necesariamente la misma especie). En el Caso 2, la especie seleccionada al azar del Ensamblaje 1 es una especie compartida, pero la especie seleccionada del Ensamblaje 2 no está compartida con el Ensamblaje 1. El opuesto ocurre en el Caso 3. En el Caso 4, ninguna de las especies seleccionadas está compartida. Estos patrones se describen matemáticamente en la Tabla II.

Un enfoque probabilístico de los índices clásicos de Jaccard y Sørensen

Los índices clásicos de Jaccard y Sørensen solamente consideran la presencia o ausencia (incidencia) de especies. Dos pares de ensamblajes, uno compartiendo las especies abundantes pero no las raras y el otro compartiendo las especies raras pero no las comunes, darán el mismo valor para el índice. Desde el punto de vista de la similitud global de los ensamblajes, llevar la similitud de la composición del ensamblaje al nivel del individuo suele ser sensato (Magurran, 2004). Nuestro próximo objetivo es extender los índices de incidencia para que tomen en cuenta la abundancia relativa de las especies, un prerrequisito para el desarrollo de estimadores del índice para datos de muestreo que toman en cuenta las especies raras no vistas.

Primero, tenemos que proveer una derivación probabilística de los índices clásicos de incidencia Jaccard y Sørensen. Suponga que seleccionamos al azar una *especie* del Ensamblaje 1 y una *especie* del Ensamblaje 2 y luego clasificamos a cada miembro de este par de acuerdo con si se trata de una especie compartida o no. Las probabilidades correspondientes se muestran gráficamente en la Fig. 1 y se especifican en la Tabla II.

Aunque las probabilidades en la Tabla II no son conteos, pueden ser considerados como 'conteos normalizados' dado que suman la unidad (1). Sustituyendo estas probabilidades en las ecuaciones 1 y 2, entonces tenemos:

$$J_{clas} = \frac{A}{A+B+C} = \frac{\left[(S_{12}/S_1)(S_{12}/S_2)\right]}{\left[(S_{12}/S_1)(S_{12}/S_2)\right]+\left[(S_{12}/S_1)(1-(S_{12}/S_2))\right]+\left[(1-(S_{12}/S_1))(S_{12}/S_2)\right]} = \frac{S_{12}}{S_1+S_2-S_{12}}$$

que es exactamente la ecuación 3. Asimismo, tenemos

$$L_{clas} = \frac{2A}{2A+B+C} = \frac{2\left[(S_{12}/S_1)(S_{12}/S_2)\right]}{2\left[(S_{12}/S_1)(S_{12}/S_2)\right]+\left[(S_{12}/S_1)(1-(S_{12}/S_2))\right]+\left[(1-(S_{12}/S_1))(S_{12}/S_2)\right]} = \frac{2S_{12}}{S_1+S_2}$$

que es la misma que la ecuación 4.

Tabla II. Derivación probabilística de conteos de especies para los índices clásicos

	Seleccione cualquier *especie* del Ensamblaje 2	
	Compartida	**No compartida**
Seleccione cualquier *especie* del Ensamblaje 1		
Compartida	$A = \frac{S_{12}}{S_1}\frac{S_{12}}{S_2}$ (Caso 1)	$B = \frac{S_{12}}{S_1}\left(1-\frac{S_{12}}{S_2}\right)$ (Caso 2)
No compartida	$C = \left(1-\frac{S_{12}}{S_1}\right)\frac{S_{12}}{S_2}$ (Caso 3)	$D = \left(1-\frac{S_{12}}{S_1}\right)\left(1-\frac{S_{12}}{S_2}\right)$ (Caso 4)

Puede parecer que no hemos avanzado, pero este método probabilístico establece la base para desarrollar índices basados en abundancias, que a su vez permiten la estimación de índices que toman en cuenta el efecto de las especies compartidas no vistas. Nótese que, utilizando este método, también podemos calcular la probabilidad de que ambas especies seleccionadas al azar sean especies no compartidas (Caso 4, presentado en la Fig. 1 y la Tabla II). Sin embargo, el concepto fundamental para los índices Jaccard y Sørensen se basa solamente en información para las otras tres celdas (Casos 1-3).

Extendiendo el enfoque probabilístico a los índices basados en abundancias

Dejemos que las probabilidades de que las especies sean descubiertas (mismas que dependen principalmente de la abundancia relativa, asumiendo una mezcla aleatoria y detectabilidad equiparable) en los Ensamblajes 1 y 2 sean denotadas por (p_1, p_2, ..., p_{S1}) y (π_1, π_2, ..., π_{S2}), respectivamente, donde $p_i > 0$, $\pi_i > 0$ y

$$\sum_{i=1}^{S_1} p_i = \sum_{i=1}^{S_2} \pi_i = 1 \cdot$$

Ya no tratamos a todas las especies de manera igual dado que algunas son comunes y otras son raras. En cambio, la idea básica para manejar los conteos de la abundancia es que tratemos a todos los *individuos* idénticamente. Adaptando el método de la sección anterior, seleccionamos al azar un *individuo* del Ensamblaje 1 y un *individuo* del Ensamblaje 2. Para cada individuo del par, notamos si pertenece a una especie compartida o no.

Derivamos ahora las fórmulas generales para las versiones basadas en abundancias de los índices Jaccard y Sørensen. Sin perder la generalidad, asumimos que las primeras S_{12} especies son especies compartidas, es decir las especies compartidas están indexadas por 1,2,...,S_{12}. En el Ensamblaje 1, dejemos que U denote la suma de las abundancias relativas de individuos que pertenecen a las especies *compartidas*, $U = p_1 + p_2 + ... + p_{S12}$. Asimismo, en el Ensamblaje 2, dejemos que V denote la suma de las abundancias relativas de individuos que pertenecen a las especies *compartidas*, $V = \pi_1 + \pi_2 + ... + \pi_{S12}$. La Tabla III muestra las probabilidades de que dos individuos, uno de cada ensamblaje, representen cada una de las cuatro categorías usuales.

Con base en las ecuaciones 1 y 2 para las tres probabilidades (A, B y C en la Tabla III), obtenemos los siguientes índices basados en la abundancia en términos de U y V:

$$J_{abd} = \frac{A}{A+B+C} = \frac{UV}{U+V-UV} \qquad (5) \text{ y}$$

$$L_{abd} = \frac{A2}{2A+B+C} = \frac{2UV}{U+V} \qquad (6)$$

Dado que U y V representan las abundancias totales de las especies *compartidas* en Ensamblajes 1 y 2, respectivamente, vemos que ambos índices tienden a 1 para ensamblajes idénticos y tienden hacia 0 para ensamblajes disimilares. En este último caso, por ejemplo $L_{abd} = 2/[(1/U) + (1/V)]$ tiende hacia 0 conforme tanto U como V se acercan a 0.

Tabla III Probabilidades para conteos de especies basados en individuos

	Seleccione cualquier *individuo* del Ensamblaje 2	
	Compartido	No compartido
Seleccione cualquier *individuo* del Ensamblaje 1		
Compartido	$A = UV$	$B = U(1 - V)$
No compartido	$C = (1 - U)V$	$D = (1 - U)(1 - V)$

Estimación de los índices basados en abundancia a partir de datos de muestreo

Hasta ahora, solamente hemos considerado las especies y los individuos *observados* en dos ensamblajes. Tanto las versiones clásicas de Jaccard y Sørensen como la nueva versión basada en abundancias, asumen total y completo conocimiento de los dos ensamblajes que estamos comparando. En la práctica, necesitamos estimar los índices de similitud usando datos de muestreo, una tarea que realizamos ahora. Nuestro enfoque es no paramétrico en el sentido de que no necesitamos postular ninguna distribución de abundancia de especies en particular para derivar los estimadores, mismos que por lo tanto son válidos bajo muchos modelos estadísticos de la abundancia (p. ejem. log-normal, vara rota, gamma, etc.). La derivación sí asume que el número de especies es finito por lo que las probabilidades de descubrimiento de especies tienen un límite. [Los autores demuestran que los estimadores son válidos bajo muchos de los modelos estadísticos de la abundancia (A. Chao, R. L. Chazdon, R. K. Colwell y T.-J. Shen, datos no publicados) (p. ejem. log-normal, exponencial, gamma, binomial negativo, Zipf-Mandelbrot, modelos de vara rota, etc.) que aparecen en Magurran (2004, Tabla 2.1) o en Plotkin y Muller-Landau (2002, Tabla 1]

Una muestra aleatoria de n individuos (Muestra 1) se toma del Ensamblaje 1 y una muestra aleatoria de m individuos (Muestra 2) se toma del Ensamblaje 2. Denote las frecuencias de las especies en las *muestras* por (X_1, X_2, ..., X_{S1}) y (Y_1, Y_2, ..., Y_{S2}), respectivamente. (Nótese que si una especie falta en una muestra, X_i o Y_i será igual a cero.) Así, el par de frecuencias para las especies S_{12} verdaderamente compartidas por los dos *ensamblajes* son (X_1, Y_1)(X_2, Y_2)...(X_{S12}, Y_{S12}). Asuma que D_{12} de las S_{12} especies compartidas disponibles de hecho se observan en ambas muestras, y que sus frecuencias son los primeros D_{12} pares. De esta manera, S_{12} - D_{12} especies adicionales se comparten entre los dos ensamblajes, pero están ausentes de una o dos de las muestras. Conforme las frecuencias de las especies raras compartidas sean mayores, se incrementa la probabilidad de que especies compartidas adicionales estén presentes en ambos ensamblajes, pero ausentes de una o ambas muestras. Nos referimos a éstas como especies *compartidas, no vistas*.

Para incorporar a las probabilidades de la Tabla III el efecto de las especies compartidas pero no vistas, usamos las frecuencias de las especies raras observadas y compartidas para estimar el término de ajuste apropiado para U y V para tomar en cuenta las especies compartidas *no vistas*. Primero definimos la función indicadora I(expresión) tal como $I = 1$ si 'la expresión' es verdadera y $I = 0$ si 'la expresión' es falsa. Dejemos que

$$f_{1+} = \sum_{i=1}^{D_{12}} I\left[X_i = 1, Y_i \geq 1\right]$$

sea el número observado de las especies *compartidas* que ocurren una sola vez [*singletons*] ($X_i = 1$) en la Muestra 1 (estas especies tienen que estar presentes en la Muestra 2, pero pueden tener cualquier abundancia). Ahora, dejemos que f_2+ sea el número observado de especies compartidas que ocurren dos veces [*doubletons*] ($X_i = 2$) en la Muestra 1. De igual manera, definimos f_{+1} y f_{+2} como el número observado de especies compartidas que ocurren, respectivamente, una sola vez ($Y_i = 1$) y dos veces ($Y_i = 2$) en la Muestra 2.

Entonces, el estimador propuesto para U es

$$\hat{U} = \sum_{i=1}^{D_{12}} \frac{X_i}{n} + \frac{(m-1)}{m} \frac{f_{+1}}{2f_{+2}} \sum_{i=1}^{D_{12}} \frac{X_i}{n} I\left(Y_i = 1\right) \quad (7)$$

Nótese que el primer término al lado derecho de la ec. 7 denota el total observado de las frecuencias asociadas con las especies observadas compartidas; el segundo término representa el efecto estimado de las especies compartidas no vistas. De manera similar, tenemos

$$\hat{V} = \sum_{i=1}^{D_{12}} \frac{Y_i}{m} + \frac{(n-1)}{n} \frac{f_{1+}}{2f_{2+}} \sum_{i=1}^{D_{12}} \frac{Y_i}{m} I\left(X_i = 1\right) \quad (8)$$

Cuando $f_{+2} = 0$ o $f_{2+} = 0$, reemplace f_{+2} y f_{2+} en los denominadores por $f_{+2} + 1$ o $f_{2+} + 1$, respectivamente. Si el valor de $\hat{U}$ o $\hat{V}$ es mayor a 1 (que rara vez ocurre), entonces se reemplaza por 1. Los estimadores Jaccard y Sørensen basados en abundancias que proponemos son

$$\hat{J}_{abd} = \frac{\hat{U}\hat{V}}{\hat{U} + \hat{V} - \hat{U}\hat{V}} \quad (9) \text{ y}$$

$$\hat{L}_{abd} = \frac{2\hat{U}\hat{V}}{\hat{U} + \hat{V}} \quad (10)$$

Las varianzas para estos dos estimadores se pueden derivar por el método *bootstrap*. (La derivación completa de las ecuaciones 7 y 8 y los detalles del procedimiento *bootstrap* para calcular los estimadores de la varianza para las ecuaciones 9 y 10 están disponibles previa solicitud al primer autor.)

Estimación de los índices de similitud a partir de frecuencias de incidencia

En virtud de que la información acerca de las frecuencias y las identidades de las especies raras, provee información crítica para ajustar los índices de similitud para que tomen en cuenta el efecto de las especies compartidas no vistas, no es posible usar una sencilla lista de las especies presentes en dos ensamblajes (datos de incidencia), aun en principio, para ajustar los índices de similitud para el efecto de las especies no vistas. Por otro lado, el método basado en estimación se puede extender a los datos de incidencia (presencia-ausencia) *replicados*.

Suponga que tomamos un conjunto de w muestras de incidencia replicadas del Ensamblaje X y un conjunto de z muestras de incidencia replicadas del Ensamblaje Y. Para ambos conjuntos de muestras *combinados*, hay S especies. El número de muestras en las que se encuentra una especie en el Ensamblaje X o Y es la *frecuencia* para esta especie en dicho conjunto de muestras. Las frecuencias para la especie i entonces se definen como

$$X_i = \sum_{j=1}^{w} x_{ij} \quad \text{y} \quad Y_i = \sum_{j=1}^{z} y_{ij}$$

donde x_{ij} y y_{ij} representan la presencia (1) o ausencia (0) de la especie i en la muestra j.

Nótese que X_i o Y_i será cero para algunas especies, a menos que todas las especies estén compartidas y observadas.

Bajo el supuesto que las muestras replicadas de la incidencia son estadísticamente homogéneas (*dentro de* cada ensamblaje), la probabilidad de que una especie esté presente en una muestra dada es proporcional a su abundancia relativa en el ensamblaje, y los vectores de frecuencia X_i o Y_i así representan estadísticamente la abundancia relativa de las especies en los Ensamblajes X y Y (p. ejem. Chao, 2004; Colwell *et al*., 2004). Por ello, con cambios menores, las ecuaciones 7 y 8 pueden usarse para calcular las probabilidades ajustadas de que una incidencia seleccionada al azar (detección de especies) de cada uno de los dos ensamblajes representarán especies compartidas (aunque no necesariamente la misma especie compartida).

Para los datos de incidencia replicados, f_{1+} es el número de especies compartidas observadas que ocurre en exactamente una muestra ($X_i = 1$) en X y f_{2+} es el número de especies compartidas observadas que ocurre en exactamente dos muestras ($X_i = 2$) en X; f_{+1} y f_{+2} son los números correspondientes para la matriz de muestras Y. Definamos la suma de las frecuencias de incidencia para las matrices como

$$n = \sum_{i=1}^{S} X_i \quad \text{y} \quad m = \sum_{i=1}^{S} Y_i$$

Entonces los estimadores propuestos son

$$\hat{U}_{inc} = \sum_{i=1}^{D_{12}} \frac{X_i}{n} + \frac{(z-1)}{z} \frac{f_{+1}}{2f_{+2}} \sum_{i=1}^{D_{12}} \left[\frac{X_i}{n} I\left(Y_i = 1\right) \right] \quad (11) \text{ y}$$

$$\hat{V}_{inc} = \sum_{i=1}^{D_{12}} \frac{Y_i}{m} + \frac{(w-1)}{w} \frac{f_{1+}}{2f_{2+}} \sum_{i=1}^{D_{12}} \left[\frac{Y_i}{m} I\left(X_i = 1\right) \right] \quad (12)$$

(Las mismas modificaciones descritas para las ecuaciones 7 y 8 pueden aplicarse aquí si $f_{+2} = 0$ o $f_{2+} = 0$.) De

esta manera, nuestros estimadores Jaccard y Sørensen basados en incidencia son

$$\hat{J}_{inc} = \frac{\hat{U}_{inc}\hat{V}_{inc}}{\hat{U}_{inc} + \hat{V}_{inc} - \hat{U}_{inc}\hat{V}_{inc}} \quad (13) \text{ y}$$

$$\hat{L}_{inc} = \frac{2\hat{U}_{inc}\hat{V}_{inc}}{\hat{U}_{inc} + \hat{V}_{inc}} \quad (14)$$

Evaluación de desempeño: índices clásicos *vs.* índices nuevos

Índices evaluados

Llevamos a cabo pruebas de desempeño para: (1) los índices clásicos de Jaccard y Sørensen (ecuaciones 1 y 2); (2) los nuevos índices Jaccard y Sørensen basados en incidencia (ecuaciones 5 y 6); (3) los estimadores para los índices basados en abundancia (ecuaciones 9 y 10); y (4) los estimadores de incidencia replicada para los índices basados en abundancias (ecuaciones 13 y 14).

Conjuntos de datos utilizados en las pruebas

Llevamos a cabo pruebas de desempeño con un conjunto de datos grande para hormigas de la selva, rico en especies (Longino *et al*., 2003), coleccionado mediante varias técnicas de colecta masiva replicada en la Estación Biológica La Selva, en Costa Rica. Aquí presentamos resultados representativos para tres métodos de colección: la extracción de muestras de suelo Berlese (217 muestras, 4318 individuos, 117 especies de las cuales 19 ocurrieron una sola vez), muestras de trampas Malaise para insectos voladores y rastreros (62 muestras, 1660 individuos, 103 especies de las cuales 35 ocurrieron una sola vez) y muestras obtenidas con la aspersión de insecticida (fumigación del dosel) (459 muestras, 26302 individuos, 165 especies de las cuales 19 ocurrieron una sola vez). [Los diagramas de abundancia relativa aparecen en Longino *et al.* (2002).] Tal y como Longino *et al.* (2002) señalan, estos tres métodos muestrean, a propósito, diferentes pero solapados segmentos de la fauna local de hormigas. Mientras la suma bruta de especies para los tres métodos sería 117 + 103 + 165 = 385 especies, el número actual de especies capturadas por estos tres métodos solamente fue de 276 especies. Pruebas paralelas para otros conjuntos de datos ricos en especies, incluyendo los datos de especies de árboles de la selva presentados más adelante en este artículo, dieron resultados concordantes (A. Chao, R. L. Chazdon, R. K. Colwell y T.-J. Shen, datos no publicados).

Las pruebas

Aunque los índices clásicos de Jaccard y Sørensen y nuestros nuevos índices miden todas la 'similitud,' su propósito es medir aspectos diferentes de esta construcción: los índices clásicos miden ostensiblemente la similitud en la composición de especies mientras hacen caso omiso de la abundancia relativa (aunque son fuertemente afectados por ella, cuando se trata de muestrear), mientras nuestros nuevos índices [y muchos otros (Legendre y Legendre, 1998; Magurran, 2004)] consideran explícitamente la abundancia relativa. De esta manera, para cualquier conjunto de datos en particular, las diferencias en la magnitud *absoluta* de los valores Jaccard o Sørensen basados en incidencia *vs.* abundancia (o bien, las diferencias entre la mayoría de los otros índices de similitud) en sí, carecen de sentido.

No obstante, los índices de similitud en la composición pueden compararse en términos de su desempeño en pruebas de su sensibilidad al submuestreo. Utilizando los datos de las hormigas, ilustramos tres pruebas: (1) Prueba 1: muestras de igual tamaño de un solo conjunto de datos (rarefacción dentro del mismo ensamblaje); (2) Prueba 2: muestras de tamaño desigual de un solo conjunto de datos; y (3) Prueba 3: muestras de igual proporción de dos conjuntos de datos (rarefacción entre ensamblajes). Para los fines de estas pruebas, tratamos a los datos de cada uno de los métodos de colecta de hormigas (Berlese, Malaise o la fumigación) como un 'ensamblaje' completo y por separado, al cual nos referimos aquí como el *agrupamiento de muestreo* (*sampling pool*). Muestras de tamaños especificados (en términos de los números de individuos) fueron entonces seleccionadas al azar, *con reemplazo*, de estos agrupamientos. Desde luego, no todas las especies presentes en un agrupamiento de muestreo están representadas en las muestras de menor tamaño. Sin embargo, dado que el muestreo se hizo con reemplazo, no todas las especies están presentes aun cuando el número de individuos seleccionado es el mismo que el número de individuos en el agrupamiento.

Resultados

Prueba 1: Muestras de igual tamaño de un solo conjunto de datos

Todos los índices de similitud rinden un valor verdadero de 1 cuando un agrupamiento de muestreo completo (ensamblaje) es comparado con si mismo. ¿Qué pasa cuando un índice de similitud se calcula para dos muestras aleatorias de un solo agrupamiento de muestreo? Si un índice no está sesgado por el tamaño de la muestra, debe dar un valor de 1 cuando se aplica a muestras de cualquier tamaño. Primero, muestreamos individuos al azar (con reemplazo) del agrupamiento de datos de hormigas para un solo método de colecta para producir pares de muestras con el mismo número de individuos que los agrupamientos mismos (muestras completas). Luego, al azar seleccionamos muestras más pequeñas, cada una con la mitad del número de individuos en el agrupamiento de muestreo original. Seguimos repitiendo este procedimiento para un par de muestras, cada una 1/4 del tamaño del agrupamiento original, luego un par 1/8 del tamaño del agrupamiento, etc., sucesivamente dividiendo la muestra a la mitad hasta quedar con 1/64 del número original de individuos. (Nótese que es una prueba muy severa del sesgo del submuestreo, aun para

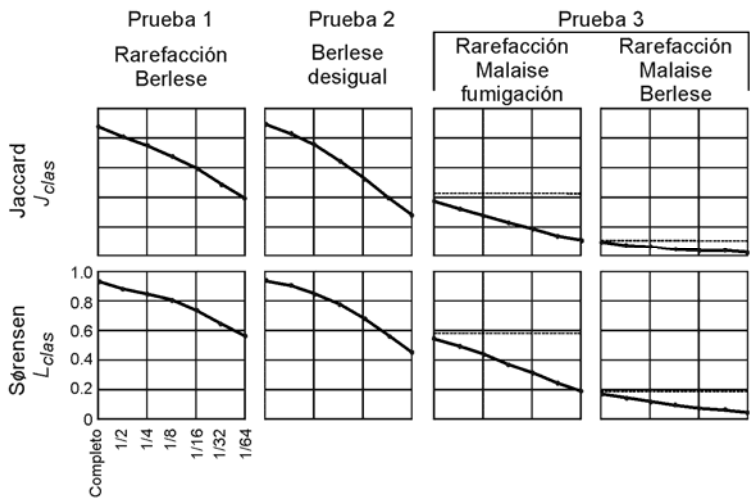

Fig. 2. Pruebas de muestreo aleatorio de los índices de solapamiento clásicos de Jaccard (J_{clas}, ec. 1) y Sørensen (L_{clas}, ec. 2). Las gráficas muestran el efecto sobre cada índice al considerar muestras aleatorias compuestas de 1/1 (Completo), 1/2, 1/4, ..., 1/64 de las abundancias o los equivalentes en incidencia en los agrupamientos de muestreo, muestreados con reemplazo. (Las etiquetas de la gráfica inferior a la izquierda se aplican a todas las gráficas). La columna 1 (Prueba 1: Berlese rarefacción) muestra valores del índice de similitud para pares de muestras del mismo tamaño del conjunto de datos Berlese de hormigas. La columna 2 (Prueba 2: Berlese desigual) muestra los valores del índice para comparaciones de muestras de tamaño decreciente *vs.* una muestra del mismo tamaño del conjunto completo de datos Berlese de hormigas. La columna 3 (Malaise-fumigación rarefacción) muestra valores del índice de similitud para pares de muestras de igual proporción (Prueba 3) de los conjuntos de datos de hormigas Malaise *vs.* fumigación, una comparación de alta similitud. La columna 4 (Malaise-Berlese rarefacción) muestra los valores del índice de similitud para pares de muestras de igual proporción (Prueba 3) del conjunto de datos de hormigas Berlese *vs.* Malaise, una comparación de baja similitud. El verdadero valor de cada índice para los agrupamientos de muestreo se indica con líneas punteadas horizontales en las columnas para la Prueba 3 (rarefacción Malaise-fumigación y Malaise-Berlese). El verdadero valor del índice para la Prueba 1 y la Prueba 2 es 1.0, es decir, la parte superior de las gráficas.

estos agrupamientos muy grandes.) Este proceso completo se repitió 1000 veces y promedios fueron calculados para cada prueba de cada índice, y para cada uno de los tres métodos de colectar las hormigas.

La Figura 2 presenta los resultados representativos para esta prueba para los índices clásicos de Jaccard y Sørensen (primera columna, Prueba 1: rarefacción Berlese). Claramente ambos índices fueron muy sensibles al submuestreo. La Figura 3 (primera columna) presenta los resultados correspondientes a los nuevos índices para esta prueba. Los nuevos índices Jaccard y Sørensen basados en abundancia, sin ajustar por las especies compartidas no vistas (J_{abd} y L_{abd}), fueron también sensibles al tamaño de la muestra. En cambio, los estimadores Jaccard y Sørensen, mismos que incluyen el efecto estimado de las especies compartidas, no vistas, resultó menos sensible al submuestreo, con valores notablemente más cercanos a 1 aun para muestras pequeñas (Fig. 3). Esto fue cierto tanto para los estimadores basados en abundancia ($\hat{J}_{abd}$ y $\hat{L}_{abd}$) como para los estimadores basados en los datos de incidencia replicados ($\hat{J}_{inc}$ y $\hat{L}_{inc}$).

Prueba 2: Muestras de diferentes tamaños de un solo conjunto de datos

Un índice de similitud idealmente debería ser robusto en cuanto al tamaño de muestra no solamente para muestras de igual tamaño, sino también para muestras de diferentes tamaños. Para poner a prueba esta propiedad calculamos los índices de similitud para muestras de tamaños sucesivamente más pequeños, *vs.* muestras 'completas', con un número de individuos igual al número en el agrupamiento de muestreo correspondiente. Tal y como se vio en la primera prueba, un índice ideal debe mantener un valor de 1, sin importar discrepancias en los tamaños de las muestras. Las Figuras 2 y 3 (segunda columna, Prueba 2: Berlese desigual) muestran una prueba así para los datos de hormigas de la muestra Berlese, utilizando muestras creadas por el mismo esquema indicado en el primer método. Aun más que en la primera prueba, los índices clásicos de Jaccard y Sørensen (Fig. 2) se vieron fuertemente afectados por el tamaño de la muestra, causando un sesgo negativo y severo cuando una muestra era mucho más pequeña que la muestra completa. En cambio, los nuevos estimadores de Jaccard y Sørensen (Fig. 3, segunda columna) resultaron notablemente resistentes al submuestreo, incluyendo tanto los estimadores basados en abundancia ($\hat{J}_{abd}$ y $\hat{L}_{abd}$) y los basados en datos de incidencia replicados ($\hat{J}_{inc}$ y $\hat{L}_{inc}$).

Muestras de igual proporción de dos conjuntos de datos

Está bien que un índice de la similitud sea robusto al tamaño de muestra al comparar muestras pareadas que provienen del mismo agrupamiento, pero un índice es de poca utilidad si no retiene esta robustez al comparar conjuntos de datos distintos, a la vez que detecta exitosamente las diferencias en la composición entre ellos. Llevamos a cabo los mismos procedimientos de comparación de tamaño de muestra descritos para el primer conjunto de pruebas, pero en vez de comparar pares de muestras del mismo agrupamiento de muestreo, comparamos pares de muestras sucesivamente más pequeños de los conjuntos de datos Malaise y fumigación [de alta similitud (Longino *et al*., 2002)], y de los conjuntos de datos de Malaise y Berlese (baja similitud). Los resultados para los índices clásicos de Jaccard y Sørensen se

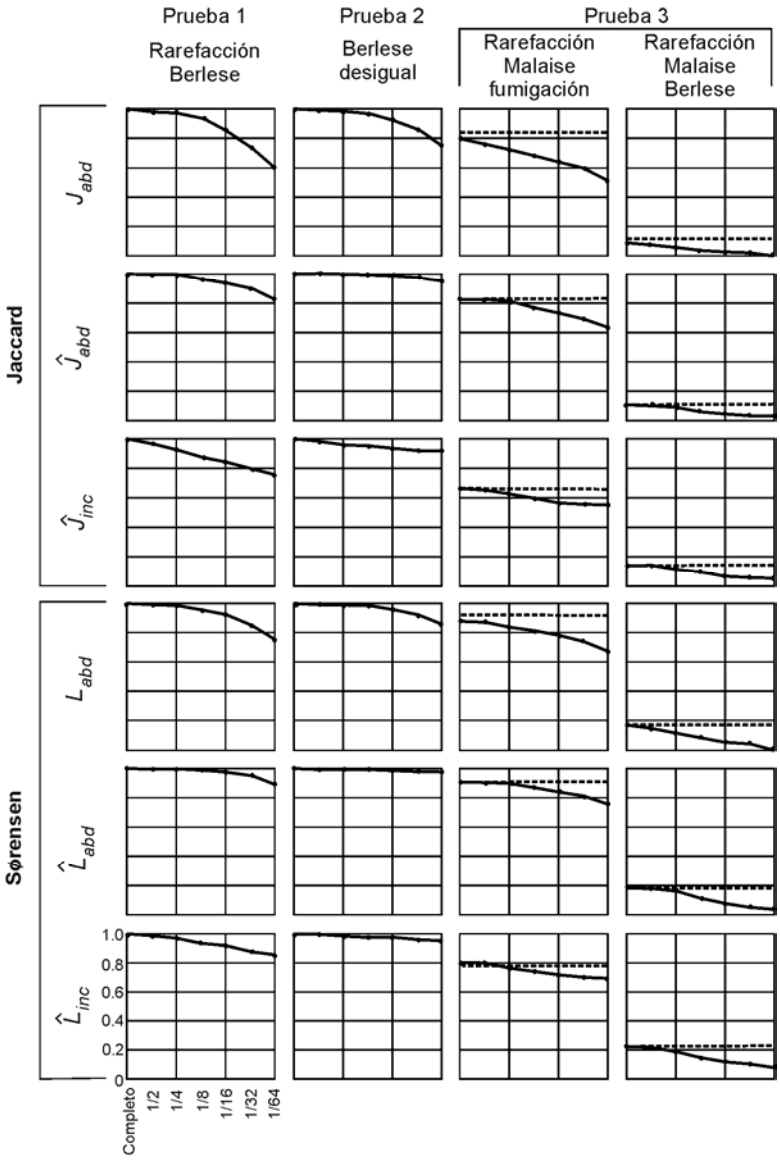

Fig. 3. Pruebas de muestreo aleatorio de los nuevos índices de solapamiento. Para cada índice las gráficas muestran el efecto al considerar muestras aleatorias compuestas de 1/1 (*Completo*), 1/2, 1/4, ..., 1/64 de las abundancias o equivalentes de incidencia en los agrupamientos de muestreo, muestreados con reemplazo. (Las etiquetas de la gráfica inferior se aplican a todas las gráficas.) Las columnas se describen en la leyenda de la Figura 2. Índices de Jaccard: J_{abd} es el nuevo índice basado en abundancia, no ajustado para las especies no vistas, calculado con la ecuación 5. $\hat{J}_{abd}$ es el estimador basado en abundancias correspondiente que toma en cuenta las especies no vistas, calculado con la ecuación 9. El estimador basado en datos de incidencia replicados, $\hat{J}_{inc}$, se calcula con la ecuación 13. Los índices Sørensen: L_{abd} es el nuevo índice Sørensen basado en abundancias, no ajustado para las especies no vistas, y calculado con la ecuación 6. $\hat{L}_{abd}$ es el estimador basado en abundancias que toma en cuenta las especies no vistas, calculado con la ecuación 10. El estimador basado en los datos de incidencia replicados, $\hat{L}_{inc}$, se calcula con la ecuación 14. El verdadero valor de cada índice para los agrupamientos de muestreo considerados se indica con líneas punteadas horizontales en las columnas para la Prueba 3 (Rarefacción Malaise–fumigación y Malaise–Berlese). El verdadero valor del índice para la Prueba 1 y la Prueba 2 es 1.0, es decir, la parte superior de las gráficas. Para permitir la comparación válida entre los estimadores basados en incidencia ($\hat{J}_{inc}$ y $\hat{L}_{inc}$) y los estimadores basados en abundancias correspondientes ($\hat{J}_{abd}$ y $\hat{L}_{abd}$, respectivamente), el eje X para cada estimador basado en incidencia se ajustó para que el número mínimo de incidencias corresponda con la abundancia mínima del estimador basado en abundancias, igualando así la cantidad de información estadística.

presentan en la tercera y cuarta columna de la Fig. 2. Un índice ideal daría y mantendría el verdadero valor calculado para los agrupamientos completos (la línea punteada y horizontal en cada caja) en el proceso de rarefacción. Los índices clásicos de Jaccard y Sørensen resultaron muy sensibles al submuestreo en esta prueba (Fig. 2). Los nuevos índices Jaccard y Sørensen basados en abundancia y sin corregir por las especies no vistas (J_{abd} y L_{abd} en la tercera y la cuarta columna de la Fig. 3), también sufrieron del sesgo del submuestreo, pero el sesgo se redujo para sus contrapartes basadas en la abundancia y corregidas para las especies no vistas ($\hat{J}_{abd}$ y $\hat{L}_{abd}$ en la tercera y cuarta columna de la Fig. 3) así como para los estimadores basados en los datos correspondientes de incidencia replicados ($\hat{J}_{inc}$ y $\hat{L}_{inc}$ en la tercera y cuarta columna de la Fig. 3).

Aplicación

A manera de ejemplo de los nuevos índices, aplicamos el índice clásico Jaccard (ec. 1), el nuevo índice Jaccard basado en abundancias (ec. 5) y su estimador (ec. 9) a datos provenientes de dos sitios de selva madura y cuatro sitios de selva secundaria en Costa Rica. Examinamos la similitud en la composición entre las especies de árboles ≥ 25 cm diámetro a la altura del pecho (dap; especies de árboles del dosel), briznales de las especies del dosel (1 – 5 cm dap) y plántulas de las especies del dosel (altura > 20 cm, dap < 1 cm) en cuatro selvas secundarias con diferente tiempo transcurrido desde su abandono como pastizal y en dos selvas maduras en la misma área de estudio. Durante las etapas tempranas de las sucesión, cuando el dosel empieza a cerrarse, las especies arbóreas colonizadoras de rápido crecimiento

Tabla IV. Patrones observados de riqueza de especies arbóreas para plántulas, briznales e individuos del dosel para cuatro cuadros de una ha de selva secundaria y dos cuadros de selva madura en el año 2000

Sitio	Edad	S_{obs} plántulas	S_{obs} briznales	S_{obs} árboles del dosel
LSUR	15	45	68	12
TIR	18	49	74	16
LEP	23	47	67	24
CR	28	57	91	33
LSUR selva madura	>200	47	101	37
LEP selva madura	>200	69	102	43

Todos los árboles y los briznales fueron marcados y su diámetro medido dentro de un cuadro de una ha en cada selva. Las plántulas fueron muestreadas en 144 cuadros que medían 1 x 5 m dentro del cuadro de una ha, resultando en un área muestreada de 0.072 ha. En estos análisis, incluimos solamente especies arbóreas; se excluyeron los arbustos, arbolitos (*treelets*) y árboles del dosel medio. Nótese que los sitios jóvenes tienen un número bajo de especies arbóreas del dosel por hectárea (individuos ≥ 25 cm dap) y un número menor de briznales comparado con la selva madura, pero para el caso de las plántulas las diferencias en la riqueza de especies fueron menos notables.

que no toleran la sombra están presentes en el dosel y también se encuentran como briznales y plántulas en el sotobosque. Conforme avanza el tiempo y el sotobosque se vuelve más sombreado, las especies que no toleran la sombra desaparecen del agrupamiento de plántulas y briznales y las especies que toleran la sombra rápidamente colonizan estas clases de tamaño pequeñas. Estas especies tolerantes a la sombra están representadas en los briznales y las plántulas, pero tienen pocos o ningún árbol en el dosel, gradualmente aumentando la riqueza de las especies arbóreas conforme la selva madura (Guariguata *et al.,* 1997; Tabla IV). De esta manera, predeciríamos que, a la medida que la selva secundaria madura, la similitud en la composición entre especies de árboles inicialmente sería alta pero rápidamente disminuiría a un mínimo durante las etapas intermedias de la sucesión y luego empezaría a aumentarse más tarde en la sucesión cuando los árboles tolerantes a la sombra alcanzan la madurez reproductiva y producen plántulas que pueden establecerse, crecer y sobrevivir.

El índice clásico Jaccard (ec. 1) mostró baja similitud en la composición entre árboles y plántulas para las cuatro selvas secundarias en comparación con las selvas maduras, con la similitud disminuyendo un poco a mayor edad entre las cuatro selvas secundarias (Fig. 4). La similitud entre árboles y briznales, en cambio, mostró aumentos graduales de la selva más joven a la selva secundaria mayor, continuando con esta tendencia a las selvas maduras (Fig. 4).

El índice Jaccard basado en abundancias (ec. 5) mostró un patrón marcadamente diferente para los seis sitios de selva. La similitud en la composición entre los ensamblajes de plántulas y árboles, y entre los de los briznales y los árboles fue inicialmente alta en la selva más joven, tal y como se predijo. Conforme la selva va madurando, los agrupamientos de plántulas y briznales se enriquecen con las especies tolerantes a la sombra que no están representadas en el dosel, y esto resulta en la disminución de la similitud en la composición que llegó a su mínimo en la selva LEP de 23 años de edad (Fig. 4). Esta similitud mínima representa un punto en la sucesión de la selva de máxima limitación de reclutamiento tanto para plántulas como para briznales. En el sitio con la selva secundaria de mayor edad, CR, el índice Jaccard basado en abundancias empezó a aumentar, reflejando el reclutamiento de las especies tolerantes a la sombra en cada una de las tres clases de tamaño (Fig. 4). El índice de similitud continuó aumentando y se estabilizó en 0.4 – 0.5 en las dos selvas maduras. Con la excepción de uno de los sitios de selva madura, los índices de similitud fueron más altos para plántulas *vs.* árboles que para briznales *vs.* árboles. En la escala de cuadros de una ha, la similitud en la composición entre las clases de tamaño árboles, plántulas y briznales (especies del dosel) en selvas maduras fue comparable a lo observado en una selva secundaria de 15 años de edad, pero mayor a lo observado en selvas secundarias de edad intermedia. Por diseño, el índice Jaccard basado en abundancias responde sensiblemente a cambios en las abundancias relativas totales de especies compartidas durante la sucesión en selvas.

El estimador Jaccard basado en abundancias (ec. 9), el cual incorpora los efectos de las especies compartidas no vistas, mostró tendencias generales similares para todas las selvas cuando se comparó con el índice Jaccard basado en abundancias (Fig. 4). La selva secundaria de 28 años de edad, sin embargo, tuvo estimados de similitud casi comparables con las dos selvas maduras, sugiriendo que el estimador está respondiendo a especies raras o infrecuentes que están compartidas entre las clases de tamaño (Fig. 4). El estimador para la similitud entre briznales y árboles fue más alto que el obtenido para plántulas *vs.* árboles en el sitio de selva secundaria TIR, indicando que este sitio tiene un mayor número de especies raras compartidas en briznales que en plántulas.

Conclusiones

En virtud de que *la similitud* es una construcción cualitativa humana, no tiene una definición matemática precisa. No obstante, el medir 'la similitud' depende de índices cuantitativos diseñados para el propósito, y en la práctica, podemos esperar que los índices de la similitud cumplan con criterios razonables para su comportamiento matemático (Legendre y Legendre, 1998). Dados índices que tengan sentido matemáticamente, lo que nos concierne aquí es su desempeño estadístico en el contexto de la realidad del muestreo de campo, particularmente para aquellos taxa ricos en especies para los cuales es poco práctico o incluso imposible llevar a cabo inventarios completos.

Utilizando simulaciones de muestreo aplicadas a conjuntos de datos de campo representativos, confirmamos que dos de los índices clásicos más ampliamente usados, Jaccard y Sørensen, sufren de sesgo negativo bajo condiciones de submuestreo, a menudo un sesgo muy fuerte (Fig. 2). Nuestro objetivo fue desarrollar

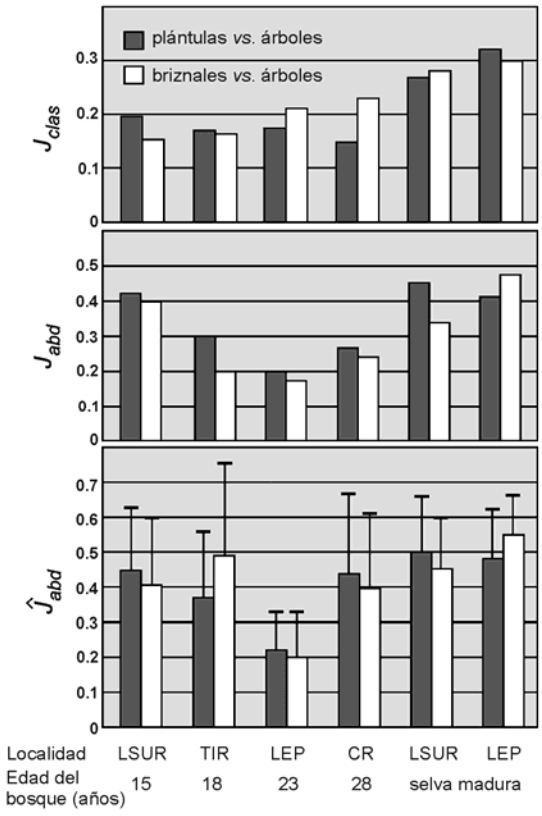

Fig. 4. La similitud en composición entre los árboles del dosel y las plántulas y los árboles del dosel y los briznales en cuatro cuadros de selva secundaria de diferentes edades y en dos cuadros de selva madura. Los resultados se presentan para J_{clas}, el índice clásico de Jaccard (ec. 1 gráfica superior), para el nuevo índice Jaccard basado en abundancias J_{abd} (ec. 5) sin ajustar por las especies no vistas (gráfica de en medio), y para $\hat{J}_{abd}$, el nuevo estimador Jaccard basado en abundancias que toma en cuenta las especies no vistas (ec. 9; las barras de error muestran 1 error estándar, calculado con el procedimiento *bootstrap*; detalles disponibles del primer autor; A. Chao, R. L. Chazdon, R. K. Colwell & T.-J. Shen, datos no publicados). Estos análisis incluyen solamente especies arbóreas del dosel; se excluyeron los arbustos, arbolitos (*treelets*) y árboles del dosel medio.

nuevos índices basados en la probabilidad que reduce el sesgo introducido por el submuestreo mediante la estimación y compensación por los efectos de las especies compartidas no vistas. Basamos un nuevo índice de similitud en la probabilidad de que dos individuos seleccionados al azar, cada uno proveniente de una de dos muestras, pertenezcan a cualquiera de las especies compartidas por las dos muestras [no necesariamente a la misma especie compartida, la base de F (Chave y Leigh, 2002; Condit *et al*., 2002) y el índice Morisita-Horn]. Este enfoque abrió el camino al paso crítico, el ajuste de esta probabilidad para tomar en cuenta la posibilidad de que muestras más grandes revelarían una proporción mayor de las especies compartidas. Como se anticipó, los nuevos índices redujeron de manera consistente el sesgo del submuestreo en las pruebas de desempeño, notablemente en la mayoría de las circunstancias. Inevitablemente, todavía hay algo de sesgo, especialmente cuando se trata del submuestreo severo y para las muestras altamente disimilares. Bajo tales condiciones, existe relativamente poca información para guiarnos en la reducción del sesgo.

Los ecólogos distinguen dos aspectos de la similitud en la composición de los ensamblajes de especies: la similitud en las listas de especies (incidencia) y la similitud en las abundancias relativas de las especies. Los índices clásicos basados en abundancias (p. ejem. Morisita-Horn o Bray-Curtis) aparean abundancias, especie por especie. Nuestros nuevos índices toman un camino intermedio, evaluando la probabilidad de que individuos pertenezcan a especies compartidas *vs.* no compartidas, sin importar a qué especies pertenecen. Desafortunadamente, para muchos estudios, datos puros de incidencia, no replicados (listas pareadas de especies) no proveen información que se pueda usar para estimar el número de especies compartidas pero no vistas. En principio, puede ser posible derivar estimadores que utilizan datos de abundancia para corregir índices de similitud puros de incidencia para las especies no vistas, pero actualmente es estadísticamente difícil para datos biológicamente realistas. Sin embargo, recomendamos los nuevos índices para cualquier aplicación en la que no solamente el apareamiento de especies sino también la similitud de las abundancias relativas sean de interés. Más aún, estos nuevos índices son más apropiados que los índices clásicos correspondientes para la evaluación de la similitud en la composición entre muestras de diferentes tamaños, así como en situaciones reales o sospechadas de submuestreo, y cuando es probable que las muestras tengan numerosas especies raras.

Agradecimiento

Agradecemos a tres árbitros anónimos sus comentarios y sugerencias. Este trabajo recibió el apoyo del Consejo Nacional de Ciencias de Taiwan (Contrato NSC92-2118-M007-013) otorgado a A. Chao y T.-J. Shen, y un apoyo del Andrew W. Mellon Foundation otorgado a R. L. Chazdon, y por US-NSF proyecto DEB-0072702 otorgado a R. K. Colwell. Damos las gracias a Jorge Leiva por compartir sus datos de vegetación para las especies arbóreas en selvas maduras. Los nuevos estimadores presentados en este artículo se incluyen en la versión 7.5 de ESTIMATES (Colwell 200ón 7.5 de ESTIMATES (Colwell 2004) y el programa SPADE (Chao y Shen 2003), que serán difundidos al publicar este artículo. La derivación completa de las ecuaciones 7 y 8 y los estimadores de la varianza para las ecuaciones 9 y 10 son disponibles, previa solicitud, del primer autor. Los conjuntos completos de los datos de las hormigas están disponibles de RKC.

La traductora, Bianca Delfosse, agradece a Javier Laborde su puntual asesoría.

Bibliografía

Arita, H. T. & P. Rodriguez. 2002. Geographic range, turnover rate and the scaling of species diversity. *Ecography*, **25**: 541-550.

Arita, H. T. & P. Rodriguez. 2004. Local-regional relationships and the geographical distribution of species. *Global Ecol. Biogeogr*., **13**: 15-21.

Balvanera, P., E. Lott, G. Segura, C. Siebe & A. Islas. 2002. Beta diversity patterns and correlates in a tropical dry forest of Mexico. *J. Veg. Sci.,* **13**: 145-158.

Bray, J. R & J. T. Curtis. 1957. An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr*., **27**: 325-349.

Bunge, J. & M. Fitzpatrick. 1993. Estimating the number of species: a review. *J. Am. Stat. Assoc*., **88**: 364-373.

Chao, A. (in press). Species richness estimation. In: *Encyclopedia of Statistical Sciences*, 2nd edn (eds. Balakrishnan, N., Read, C.B. & Vidakovic, B.). Wiley Press, New York, NY, USA.

Chao, A. & T. J. Shen. 2003. Program SPADE (Species Prediction and Diversity Estimation). Program and User's Guide available at http://chao.stat.nthu.edu.tw.

Chao, A., M.-C. Ma & M. C. K Yang. 1993. Stopping rules and estimation for recapture debugging with unequal failure rates. *Biometrika*, **80**: 193-201.

Chave, J. & E. G. Leigh. 2002. A spatially explicit neutral model of beta-diversity in tropical forests. *Theor. Pop. Biol*., **62**: 153-168.

Chazdon, R. I., 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. In: *Forest Biodiversity Research Monitoring and Modeling. Conceptual Background and Old World Case Studies*. (eds Dallmeier, F. & Comiskey, J.). Parthenon Publishing, París, France, pp. 285-309.

Colwell, R. K. 2004. ESTIMATES: Statistical Estimation of Species Richness and Shared Species from Samples, Version 7.5. Available at 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. *Phil Trans. R. Soc. Lond. B. Biol. Sci*., **345**: 101-118.

Colwell, R.K., C. X. Mao & J. Chang. 2004. Interpolating, extrapolating, and comparing incidence-based species accumulation curves. *Ecology*, **85**: 2717-2727.

Condit, R., N. Pitman, E. G. Leigh Jr, J. Chave, J. Terborgh, R. B. Foster *et al.* 2002. Beta-diversity in tropical forest trees. *Science*, **295**: 666-669.

Duivenvoorden, J. F. 1995. Tree species composition and rain forest-environment relationships in the. middle Caquetá area, Colombia, NW Amazonia. *Vegetatio*, **120**: 91-113.

Duivenvoorden, J. F., J.-C. Svenning & S. J. Wright. 2002. Beta diversity in tropical forests. *Science*, **295**: 636-637.

Fisher, B. L. 1999. Improving inventory efficiency: a case study of leaf-litter ant diversity in Madagascar. *Ecol. Appl*., **9**: 714-731.

Grassle, J. R. & W. Smith. 1976. A similarity measure sensitive to the contribution of rare species and its use in investigation of variation in marine benthic communities. *Oecologia*, **25**: 13-22.

Guariguata, M. R., R. L. Chazdon, J. S. Denslow, J. M.Dupuy & L. Anderson. 1997. Structure and floristics of secondary and old-growth forest stands in lowland Costa Rica. *Plant Ecology*, **132**: 107-120.

Harte, J., A. Kinzig & J. Green. 1999. Self-similarity in the distribution and abundance of species. *Science*, **284**: 334-336.

Hubbell, S. P. 2001. *A Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, Princeton, NJ.

Koleff, P., K. J. Gaston & J. J. Lennon. 2003. Measuring beta diversity for presence-absence data. *J. Anim. Ecol*., **72**: 367-382.

Lee, S.-M. & A. Chao. 1994. Estimating population size via sample coverage for closed capture-recapture models. *Biometrics*, **50**: 88-97.

Legendre, P. & L. Legendre. 1998. *Numerical Ecology*. Elsevier, Amsterdam.

Leigh, E.G., S. J. Wright, F. E. Putz & E. A. Herre. 1993. The decline of tree diversity on newly isolated tropical islands: a test of a null hypothesis and some implications. *Evol. Ecol*., **7**: 76-102.

Lennon, J. J., P. Koleff, J. J. D. Greenwood & K. J. Gaston. 2001. The geographical structure of British bird distributions: diversity, spatial turnover and scale. *J. Anim. Ecol*., **70**: 966-979.

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

MacKenzie, D. I., L. Bailey & J. D. Nichols. 2004. Investigating species co-occurrence patterns when species are detected imperfectly. *J. Anim. Ecol*., **73**: 546-555.

Magurran, A. E. 2004. *Measuring Biological Diversity*. Blackwell, Oxford.

Plotkin, J. B. & H. C. Muller-Landau. 2002. Sampling the species composition of a landscape. *Ecology*, **83**: 3344-3356.

Rodriguez, P. & H. T. Arita. 2004. Beta diversity and latitude in North American mammals: testing the hypothesis of covariation. *Ecography*, **27**: 1-11.

Ruokolainen, K. & H. Tuomisto. 2002. Beta-diversity in tropical forests. *Science*, **297**: 1439a.

Valencia, R, R. B. Foster, G. Villa, R. Condit, J.-C. Svenning, C. Hernández, C. *et al*. 2004. Tree species distributions and local habitat variation in the Amazon: large forest plot in eastern Ecuador. *J. Ecol*., **92**: 214-229

Wolda, H. 1981. Similarity indices, sample size and diversity. *Oecologia*, **50**: 296-302.