

Modelos Matemáticos para curvas de distribución de especies

Christian Valqui[†], Thomas Valqui[‡]

[†]IMCA-PUCP, [‡]CORBIDI

E-mail: cvalqui@pucp.edu.pe, tvalqu1@lsu.edu

Recibido el 02 de diciembre del 2007; aceptado el 16 de diciembre del 2007

Se describe distintos modelos para medir la biodiversidad a través de las distribuciones de abundancia de especies, luego se discute el método de las listas fijas para recolección de datos y los cálculos necesarios para obtener de los datos recolectados las distribuciones relativas de las especies.

Palabras claves: Abundancia de especies, Modelo de Hubbell, Listas fijas de aves.

We describe different models that measure the biodiversity using the Species abundance distribution. Then we discuss the method of fixed lists for data recollection and the necessary computations to transform the recollected data in relative species abundance distributions.

Keywords: Species abundance, Hubbell's model, Fixed bird lists.

1. Introducción

Uno de los problemas que se plantea en el estudio de los ecosistemas es la manera de medir el impacto de la actividad humana (o de otros factores) sobre estos. Si bien en algunos casos el impacto es obvio, por ejemplo en el caso de la deforestación, es necesario tener una medida más fina para poder prevenir los cambios irreversibles.

Un posible parámetro de medición de la biodiversidad es la cantidad de especies presentes en una comunidad (de plantas, animales, aves, etc.) Este número tiene varias ventajas:

1. Se puede medir cuantitativamente a partir de las observaciones.
2. Todos están de acuerdo en que mide la biodiversidad.
3. Por lo anterior, es fácil de explicar a los entes políticos y administrativos que toman decisiones sobre proteger áreas, otorgar licencias de extracción, etc.

Pero el trabajo necesario para establecerlo es bastante grande, y además requiere un conocimiento experto para identificar las distintas especies. En el caso de plantas de hecho se hace y resulta útil, aunque también allá es un trabajo enorme.

Las aves forman un ecosistema bastante sensible a las pequeñas alteraciones de los ecosistemas, y por lo tanto resultan útiles para determinar cambios, incluso pequeños, en el ecosistema. Sin embargo el trabajo de determinar la cantidad de especies presentes en un área es bastante difícil, y de hecho solamente se ha conseguido en pocos lugares, y para estar más o menos seguros del resultado se debe tomar muestras grandes. Hay modelos matemáticos para estimar la cantidad total a partir de las curvas de acumulación.

Hay otro problema con este parámetro. Veamos los siguientes gráficos correspondientes ambos a distribuciones de abundancias de 2000 especies y 2000 000 de individuos:

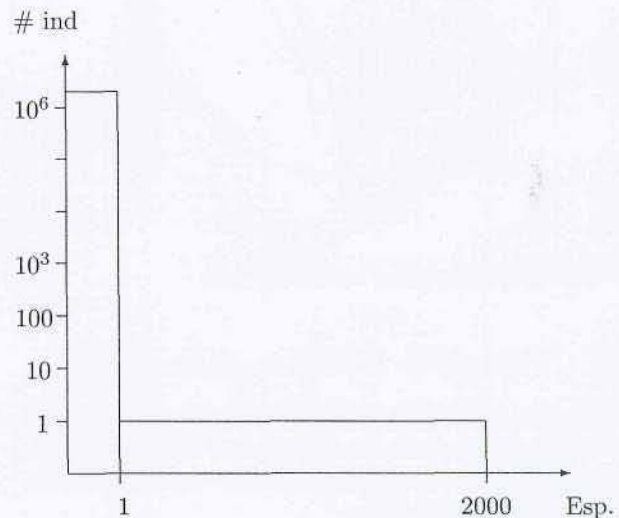


Figura 1. Una especie dominante.

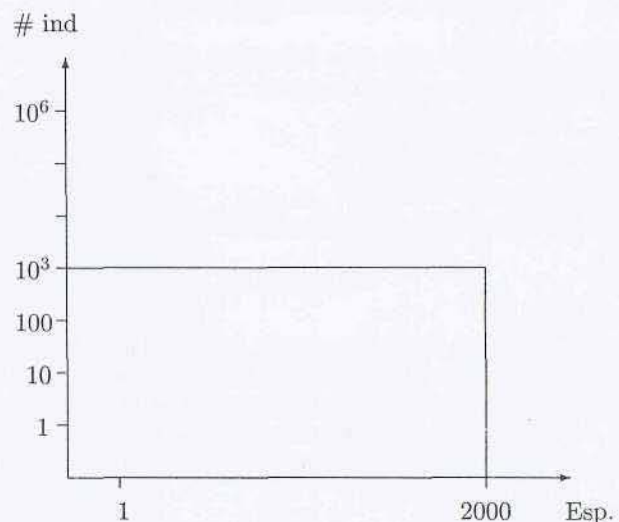


Figura 2. Todas las especies tienen 1000 individuos.

El primer gráfico muestra la distribución correspondiente a 1998001 gorriones y un individuo de cada una de las otras 1999 especies. En el segundo gráfico se con-

Un ejemplo realizable es $\delta_2 = \delta_\epsilon^\theta$ ($0 < \theta \ll 1$), donde

$$\delta_\epsilon : \begin{cases} \frac{dx}{dt} = x^2 + y^2 - \epsilon \\ \frac{dy}{dt} = 2(x + \alpha)y \end{cases} \quad (\epsilon > 0, \quad \alpha + \epsilon < 0)$$

Su diagrama de fase global se muestran en la figura 11.

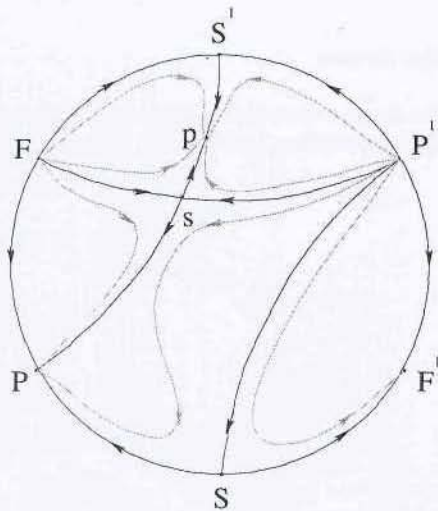


Figura 11. Sistema $s(F, P', p, P), S(P', p)$.

3.10. Tipo d(3). Forma p_1, p_2, s_1, s_2, SPF

Para este puede haber sólo a lo más 30 diferentes estructuras topológicas de los cuales hay 8 tipos los cuales han sido realizados. Los 22 tipos restantes aún están por realizarse o eliminarse [4].

3.11. Tipo e(1). Forma $p_1 p_2 S_1 S_2 F$

Este caso tiene sólo tres diferentes estructuras topológicas, cuyas características son:

$\epsilon_1 : S_1(p_1, p_2), S_2(F, F')$,

$\epsilon_2 : S_1(p_1, F'), S_2(p_2, F')$,

$\epsilon_3 : S_1(p_1, p_2), S_2(p_1, p_2)$.

Todos ellos tienen ejemplos realizables [4].

3.12. Tipo e(2). Forma $p_1, p_2, p_3, s, S_1, S_2, F$

Para esto hay a lo más 9 diferentes tipos de estructuras topológicas. Hay cuatro tipos los cuales han sido realizados, y los cinco restantes aún están por realizarse o eliminarse [4].

Un ejemplo realizable es $\epsilon_1 = \epsilon_\lambda^\theta$ ($0 < \theta \ll 1$), donde

$$\epsilon_\lambda : \begin{cases} \frac{dx}{dt} = x(\lambda + x + 2y) \\ \frac{dy}{dt} = y(-\lambda + 2x + y) \end{cases} \quad (\lambda > 0)$$

Su diagrama de fase global se muestra en la figura 12.

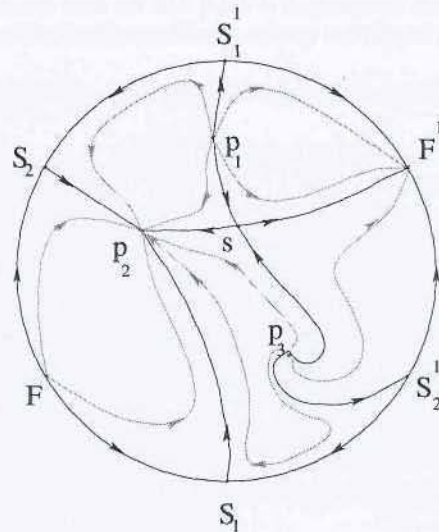


Figura 12. Sistema $s(p_1, p_3, p_2, F'), S_1(p_1, p_2), S_2(p_3, p_2)$.

4. Conclusiones

El análisis realizado en este trabajo se puede resumir en el siguiente teorema:

Teorema 4. *Los sistemas cuadráticos estructuralmente estables sin ciclos límites pueden tener a lo más 60 diferentes tipos de estructuras topológicas, de las cuales al menos 33 pueden ser realizables.*

Recientemente D. M. Zhu ha probado que el número de estructuras topológicas diferentes posibles es menos de 60, pero al menos 36 pueden ser realizados [4].

1. A. A. Andronov, E. A. Leontovich, I. I. Gordon y A. G. Maier; *Qualitative Theory of Second Order Dynamic Systems*. Israel Program for Scientific Translations Ltda. 1973.
 2. L. Perko, *Differential Equations and Dynamical Systems*, Second Edition, Springer-Verlag New York, Inc. 1996

3. G. Sansone and R. Conti. The Macmillan Company. New York. 1964.
 4. Ye Yan-Qian. *Theory of Limit Cycles*. Translations of Mathematical Monographs. Volumen 66. American Mathematical Society. 1986.

sidera una población de 1000 individuos por cada una de las 2000 especies. Es obvio que la segunda distribución representa una comunidad con mayor biodiversidad, a pesar de tener ambas comunidades la misma cantidad de especies y de individuos.

Un parámetro común para medir la biodiversidad diferenciando estos dos casos, es la probabilidad P_{bd} de que al extraer dos individuos, estos sean de especies distintas. En el primer caso la probabilidad es cercana a $1/500$, es decir, casi nunca escogemos distintas especies, mientras que en el segundo caso obtenemos casi siempre (con probabilidad aprox. $1-1/2000$) dos especies diferentes.

También vemos en este ejemplo la manera de representar las curvas de abundancia, una escala logarítmica para el número de individuos y en el eje X las especies, ordenadas desde la más abundante a la menos abundante. De hecho estos son los casos extremos, y en la realidad se dan curvas de abundancia de especies como estas:

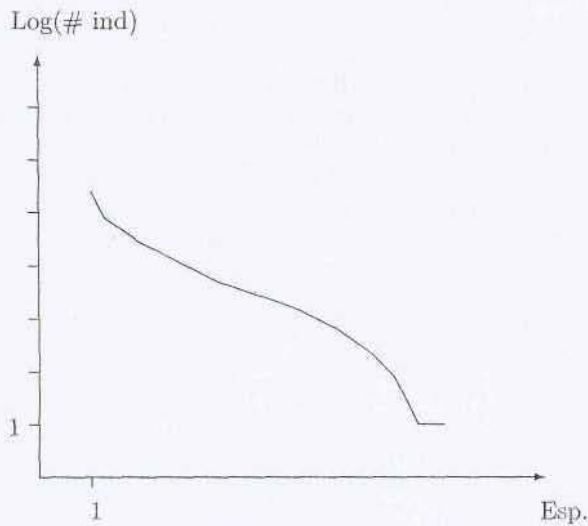


Figura 3. Una curva típica de distribución.

Las curvas de abundancia de especies son un parámetro más fino para medir la biodiversidad y de hecho al conocer las abundancias relativas $\{P_i\}$ podemos determinar

$$P_{bd} = \sum_i P_i(1 - P_i).$$

2. Modelos matemáticos para aproximar curvas de abundancia

Primero veamos los distintos intentos que hubo a lo largo de la historia para aproximar las curvas de abundancia por medio de curvas generadas por pocos parámetros, para poder estimar la biodiversidad de las comunidades reales. El primer intento corresponde a la serie logarítmica de Fisher [2]. Tiene dos parámetros α y x , la cantidad de especies que tienen n individuos es igual a $\alpha x^n/n$. En un gráfico semilogarítmico este modelo produce una recta:

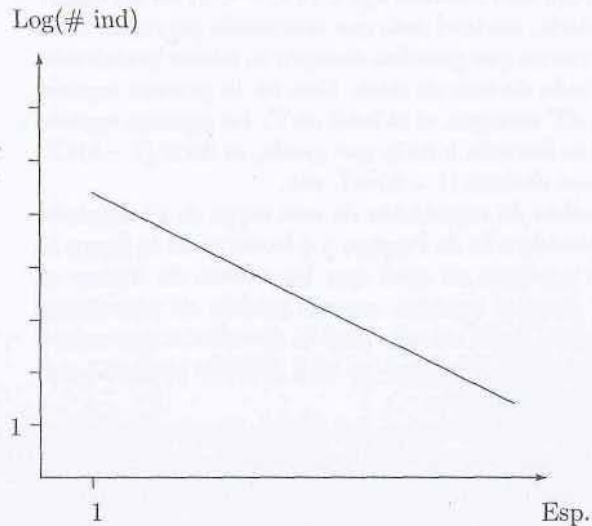


Figura 4. La serie logarítmica de Fisher.

Fisher simplemente buscó una curva que mejor aproximara a sus datos, el conjunto de datos mejor estudiado (por Fisher y muchos otros) es el correspondiente a 8000 especies de mariposas nocturnas (polillas) atrapadas durante un largo lapso de tiempo en una trampa de luz.

Otro biólogo, Preston, que trabajaba mayormente con aves, no lograba acomodar bien sus datos a las curvas propuestas por Fisher y propuso otro conjunto de curvas para aproximar los datos reales, la serie lognormal de Preston [3]. Estas curvas se construyen de la siguiente manera: Luego de aplicar el logaritmo al eje del número de individuos por especie, se supone una distribución normal de las especies que se encuentran en cada intervalo:

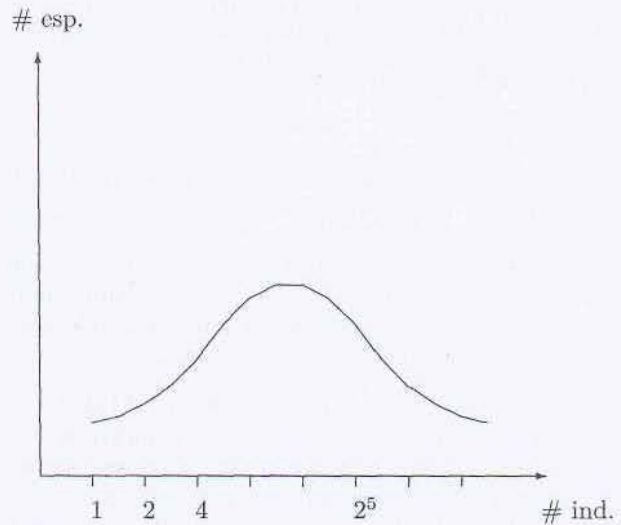


Figura 5. El plot de Preston.

Para obtener las curvas de distribución como antes se ordenan las especies de mayor abundancia a menor abundancia y se obtiene una curva parecida a la de la figura 3. La teoría de Preston está asociada a una teoría de repartición de recursos, llamada la del palo roto, ("broken stick" en inglés) del modo siguiente: Se ordenan las especies por jerarquía o dominancia, y cada especie en su

turna toma una fracción fija k ($0 < k < 1$) de los recursos, es decir, parte el palo que representa los recursos en dos porciones que guardan siempre la misma proporción y se adueña de una de ellas. Esto es, la primera especie obtiene kT recursos, si el total es T . La segunda especie obtiene la fracción k de lo que queda, es decir $(1-k)kT$. La tercera obtiene $(1-k)^2kT$, etc.

Este modelo de repartición da una curva de abundancia muy parecida a la de Preston y a la curva de la figura 3.

Más adelante se notó que las curvas de Fisher se pueden obtener también con un modelo de repartición de recursos, pero en este caso la fracción k que recibe cada especie en su turno es una variable aleatoria con distribución uniforme.

Aquí hay un problema que valdría la pena investigar: ¿Para cada distribución de datos reales es posible determinar la distribución de la fracción k para que un modelo de repartición nos aproxime bien la curva de abundancia?

3. Modelo de Hubbell

Las preferencias de los biólogos por uno u otro modelo se explican por el tipo de comunidades que estaban analizando: Fisher trabajó con una metacomunidad (8000 especies de mariposas), mientras que Preston trabajaba con aves y con comunidades locales (alrededor de 200 especies), y el modelo de Fisher no aproximaba bien a sus datos. El modelo que vamos a describir, debido a Hubbell [1], se basa en una dinámica poblacional y predice que en las comunidades locales se obtiene curvas del tipo log-normal, mientras que en el límite cuando el tamaño de la comunidad tiende al infinito, se obtiene la curva de Fisher.

Vamos a explicar el modelo de Hubbell en tres versiones que van implementando progresivamente más parámetros. El primer modelo es bastante simple y se basa en las siguientes propiedades:

PRIMER MODELO:

1. La comunidad consiste de un número fijo de individuos.
2. Los individuos de las distintas especies son equivalentes, es decir, tienen la misma probabilidad de procrear, la misma expectativa de vida, independientemente de cual especie provenga.
3. Inicialmente se da una distribución $\{P_i\}$ de las especies. Cada cierto tiempo muere un individuo (Todos los individuos tienen la misma probabilidad) y es reemplazado por un recién nacido de otra especie, donde P_i es la probabilidad de que sea de la especie i .

Este modelo alcanza un estado final de equilibrio con una sola especie presente, pues es claro que si el que tiene más individuos también tiene más probabilidad de reemplazar a los otros (aunque también tiene más occisos), y cuando una especie se extingue, no puede regresar. Además al alcanzar la monodominancia esta es un estado estable. Podría ser que la humanidad este en

camino a este estado, al ser la especie que más recursos naturales está consumiendo en la actualidad.

SEGUNDO MODELO:

Este segundo modelo de las comunidades locales tiene las mismas propiedades que el anterior, adicionalmente al determinar la especie del recién nacido, se considera una probabilidad m de que el nuevo miembro de la comunidad provenga de una metacomunidad. Típicamente este caso se da cuando una isla obtiene así miembros de la metacomunidad del continente. Si $\{PM_i\}$ es la distribución de las especies en la metacomunidad, el valor esperado a lo largo del tiempo de cada abundancia relativa P_i en la comunidad local es obviamente igual a PM_i , sin embargo la curva de distribución esperada no va a ser igual. Imaginemos que la migración sea muy baja, de modo que nuestra comunidad local, una isla por ejemplo, siempre este al borde de la monodominancia. Lo que no se sabe es quien va a ser esa especie que este dominando, y cada especie tiene la posibilidad de ser ella la que en algún momento domine, aunque en promedio sigue teniendo una abundancia determinada por PM_i . Solamente si $m=1$ las distribuciones coinciden plenamente. Si suponemos que la metacomunidad tiene la distribución de Fisher, los distintos valores de m generan curvas que coinciden en el primer tramo, pero según el valor de m cortan una parte de la cola de la curva, para parecerse a una de las curvas obtenidas con el método de Preston.

Log(# ind)

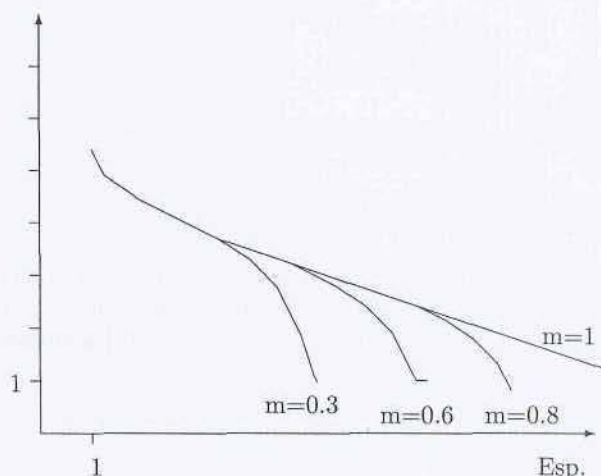


Figura 6. La forma de la curva depende del índice de migración m .

TERCER MODELO:

El modelo final de Hubbell que nos provee de la curva de la metacomunidad tiene como base toda la estructura anterior, pero modifica adicionalmente la dinámica con un parámetro de mutación ν , que determina la probabilidad de que un recién nacido sea de una especie totalmente nueva, es decir, un mutante.

El "número fundamental de biodiversidad" propuesto por Hubbell es $\theta = 2J_M\nu$. En el límite cuando el tamaño

de la metacomunidad J_M tiende al infinito este número corresponde al α de Fisher y la distribución es la serie logarítmica.

VENTAJAS DEL MODELO DE HUBBELL:

1. Se basa en un modelo dinámico.
2. Se obtienen los enfoques anteriores como casos particulares.
3. Buena aproximación a los datos reales.
4. Ofrece un número real θ para poder dialogar con entes administrativos de toma de decisiones.

CRÍTICAS:

1. Ignora la competencia que se sabe existe proponiendo la equivalencia entre especies (se dice que su efecto se elimina en comunidades suficientemente grandes).
2. En datos reales predice una tasa de especiación (mutación) de hasta 30.
3. El proceso de especiación real es más bien a través de separación de comunidades y no puntual.
4. La cantidad de parámetros libres podría permitir aproximar cualquier curva que se parezca lejanamente a lo que se quiere, sin necesidad de tanto cálculo.

En el caso de aves se hizo la aproximación con el propio Hubbell, y se pudo aproximar bien los datos a una curva, pero cuando se planteó la abundancia de cada especie no por cantidad de individuos sino por peso total de cada especie, también se obtuvo una buena aproximación.

Sin embargo este modelo, tal vez mejorado o adaptado, puede dar algunas medidas útiles para el problema planteado inicialmente, evaluar el impacto de la actividad humana.

4. Listas de Aves

Las aves son utilizadas con mucha mayor frecuencia que otros grupos de organismos como objetos de estudios en desarrollo de teorías de poblaciones, de ecología y de impacto ambiental. Las principales razones para esta preferencia son su actividad generalmente diurna, su notoriedad al estar generalmente expuestos al observador, y su manejable, pero significativa diversidad (Generalmente una comunidad tiene entre 10 y 400 especies).

La tarea de describir la comunidad de aves, como un dato comparable en el tiempo y el espacio, ha sido una actividad frecuente, pero poco exitosa. Los conteos de aves en los que se quiere estimar o describir cierta característica de la comunidad de aves para relacionarla al hábitat o para estudiar la comunidad de aves en sí misma, han tenido una serie de protocolos que no son uniformes y por lo tanto no son realmente comparables.

En el caso de metodologías detalladas, el método más usado y aceptado es el de mapeo de territorios. Esto

se consigue apuntando el sitio donde se ubica un ave, tratando de individualizar al ave. Con suficientes réplicas y ayudados con comportamientos como "peleas" en bordes territoriales, se deduce que la agregación de puntos de ubicación representan un territorio y se trazan las fronteras territoriales. Si esto se hace para todas las especies y en áreas donde quepan varios territorios, se puede llegar a un estimado bastante bueno del número de aves por especie que ocurren en una cierta área. Sin embargo para esto se requieren de varias decenas de día hombre de esfuerzo.

Habiendo tantos sitios no estudiados en el Perú y en los estudios de impacto ambiental donde se requiere evaluar grandes áreas y/o muchas de estas, es necesario realizar evaluaciones rápidas. En el caso de estas evaluaciones las dos principales metodologías utilizadas son las de los conteos por puntos y los conteos en transectos.

En ambos casos la unidad de muestreo son el espacio (distancia) y/o el tiempo. Por un lado el tamaño de la unidad es subjetivo y diferente en las distintas evaluaciones y por otro lado los factores como variabilidad climática, variabilidad de esfuerzo por área, diferencias de hábitat entre áreas, y variabilidad debido al efecto del observador causan graves problemas analíticos puesto que influyen directamente el tamaño de la muestra sin haber forma de medir tal influencia. Por ejemplo, si una mañana tiene condiciones adversas para la actividad de las aves, el tiempo o la distancia siguen avanzando y queda a nuestro criterio detener la evaluación o continuar. Se sabe de antemano que la comparación de dos sitios con evaluaciones hechas en dos días con condiciones climáticas distintas es poco confiable. Saber que estamos comparando dos sitios evaluados por 15 horas/hombre no nos da ninguna garantía de igualdad de condiciones para comparar las especies presentes.

El método que analizaremos aquí es el de las "listas fijas como unidades de conteo en evaluaciones rápidas de aves". En este método es muy poco utilizado por tener un marco teórico poco claro y carecer hasta ahora de un marco estadístico. En este método las unidades de muestra son listas de N especies, en las cuales se registran las N primeras especies encontradas, sin repetición de especie. Después de esto se inicia una nueva lista de N especies, independientemente de las especies observadas en la lista anterior, y así consecutivamente. Se evita en todo momento el repetir individuos salvo en repetición de recorridos. La unidad de muestra que es bastante más independiente de la variabilidad causada por diversos factores no controlables que afectan métodos tradicionales. En este método cualquier factor que dificulte la detección de especies, automáticamente disminuirá la rapidez con la que se completa una lista, frenando la adición de unidades muestrales. Es decir, tiene una suerte de regulador intrínseco para comparar dos sitios. En un día lluvioso no haremos tantas listas, como en el soleado por lo que sabemos que debemos añadir esfuerzo en el sitio en el que las condiciones eran adversas. Al comparar los sitios lo que buscamos es que todos tengan igual número de listas. De todas maneras se debe tratar de mantener distancia y condiciones climáticas similares.

Para cada censo se toma los siguientes datos: día,

hora de inicio, hora final, observador. En el censo mismo, se registran todas las especies de aves encontradas en forma visual o auditiva anotando los siguientes datos: especie, número de individuos vistos, hábitat en el que se vio. De esta manera se tiene al final de una evaluación un número de listas por sitio, con los que se puede hacer los siguientes análisis:

- a. Curvas de acumulación de especies.
 - a1. Estimado de lo adecuado del esfuerzo de muestreo. Estimador: pendiente de la curva de acumulación.
 - a2. Estimado del número total de especies en la comunidad evaluada. Estimador: asíntota de la curva.
- b. Estimado de las abundancias relativas de las especies.
 - b1. Curva de abundancias relativas como descripción de la estructura de la comunidad de aves. Estimador: theta de la ecuación de Hubbell.
 - b2. Estimado de la abundancia relativa de especies de interés especial (por ejemplo amenazadas) Estimador: frecuencia de la especie en el número total de listas.

Para estimar la abundancia relativa a partir de los datos obtenidos hay que tener en cuenta que la frecuencia relativa con la que una especie i aparece en una lista es un estimador de la probabilidad Q_i de esta especie de aparecer en una lista, que es distinta a la abundancia relativa P_i , que en un modelo de urna ideal es la probabilidad de extraer un individuo de la especie i . Usando métodos básicos de probabilidad se obtiene para una lista de longitud $N+1$:

$$Q_i = P_i \left(1 + \sum_{j_1 \neq i} \frac{P_{j_1}}{1 - P_{j_1}} \left(1 + \dots \left(1 + \sum_{j_N \notin I_N} \frac{P_{j_N}}{1 - \sum_{s=1}^N P_{j_s}} \right) \dots \right) \right) \quad (1)$$

donde $I_k = \{i, j_1, \dots, j_{k-1}\}$.

Esta fórmula sirve para transformar los datos en bruto (en forma de listas de N especies) en distribuciones $\{P_i\}$, con los cuales se pueden estimar los parámetros de los modelos anteriormente expuestos. Los ornitólogos encargados de recolectar datos típicamente elaboran listas de 5, 10 o 20 especies, dependiendo de la cantidad de especies presentes en la comunidad.

Si bien la fórmula (1) nos da una correspondencia exacta entre los P_i 's y los Q_i 's, cuando se trabaja con los datos reales el tiempo de cómputo crece exponencialmente con el tamaño N de las listas y con el número de especies m . Como además necesitamos varios pasos para implementar el algoritmo de Newton explicado más adelante, los tiempos de cálculo son apenas realizables para listas de 5 especies, con listas de 6 serían calculables en principio con computadoras más potentes, pero listas de 10 ya exceden el tiempo razonable y listas de 20 están totalmente fuera del alcance. Algunos tiempos

de cálculo los vemos en la tabla 1, son necesarios varios pasos de aproximación para calcular el inverso de la función dada por la fórmula (1). En lo que sigue vamos a deducir primeramente la fórmula (1) que define una función $\phi : \{P_i\} \rightarrow \{Q_i\}$. Luego describimos el método de Newton aproximado que usamos para calcular la inversa de ϕ . En la última sección describimos cuatro métodos para aproximar al inverso de ϕ con un tiempo de cómputo mucho menor, que también será adecuado para listas de 20 especies.

Tabla 1. Cuadro comparativo de los tiempos de cómputo para distintas longitudes N de listas y distintas cantidades m de especies.

| N | No. esp. (*) | Tmpo. paso | Tmpo. tot. |
|---|--------------|------------|----------------|
| 2 | < 40 | - | inmediato |
| 3 | < 40 | - | inmediato |
| 4 | 20 a 30 | | 2 a 3 segundos |
| 5 | 15 | 20 seg. | 3 min. |
| 5 | 17 | 1 min. | 20-25 min. |
| 5 | 19 | 2 min. | 40 minutos |
| 5 | 42 | 2½ horas | > 25 h. |
| 6 | 16 | 40 min. | > 20 h. (**) |
| 6 | 38 | 60 horas. | > 1200 h. (**) |

(*) Para el cálculo se considera sólo el número de frecuencias absolutas distintas.

(**) Solamente es un tiempo estimado.

5. La construcción de la función Φ

Suponemos un modelo de urna donde las m distintas especies están representadas por las etiquetas $i = 1, \dots, m$. P_i es la frecuencia relativa de cada especie. Se plantea una serie infinita de experimentos aleatorios independientes que consiste en extraer una bola de la urna, y llamar B_n la etiqueta de la n -ésima bola extraída. Asimismo llamamos L_k a la k -ésima bola distinta. Es decir, $L_1 = B_1$, y L_2 es la etiqueta de la primera bola que sale distinta de B_1 . El experimento concluye una vez que salieron N bolas distintas. Queremos calcular Q_i^N , la probabilidad de que uno de los L_k sea igual a i . Esto corresponde a la probabilidad de la especie i de salir en una lista de longitud N . El objetivo es hallar una fórmula explícita para Q_i^N como función de los P_i . Para aclarar el método supongamos que queremos elaborar listas de 5 especies y que para una cierta lista se tuvo que realizar 9 observaciones:

- 1.Obs: Especie 1
- 2.Obs: Especie 8
- 3.Obs: Especie 4
- 4.Obs: Especie 1
- 5.Obs: Especie 8
- 6.Obs: Especie 2
- 7.Obs: Especie 1
- 8.Obs: Especie 1
- 9.Obs: Especie 5

En este ejemplo tenemos:

$$\begin{aligned} L_1 &= B_1 = 1, \\ L_2 &= B_2 = 8, \\ L_3 &= B_3 = 4, \\ L_4 &= B_6 = 2, \\ L_5 &= B_9 = 5. \end{aligned}$$

Para deducir la fórmula (1) analizaremos los casos de $N = 1, 2, 3$, y luego daremos la fórmula general.

Para el caso $N = 1$ se tiene $Q_i^1 = P_i$. Para $N = 2$ calculemos la probabilidad de la especie i de aparecer en una lista dada:

$$Q_i^2 = P(L_1 = i) + P(L_2 = i)$$

Tenemos que $P(L_1 = i) = P_i$. También tenemos que $P(L_2 = i) = \sum_{j \neq i} P(L_1 = j)P(L_2 = i | L_1 = j)$. Para que pueda ser $L_2 = i$ cuando el primero fue de la especie j ($L_1 = j$), todos los individuos observados entre el primero y el último deben haber sido igual a j , de modo que

$$P(L_2 = i | L_1 = j) = \sum_{n=2}^{\infty} P(B_k = j, 2 \leq k < n)P(B_n = i) = \frac{P_i}{(1 - P_j)}$$

pues $P(B_n = i) = P_i$ y $\sum_{n=2}^{\infty} P(B_k = j, 2 \leq k \leq n - 1) = \sum_{n=2}^{\infty} P_j^{n-2} = \frac{1}{(1 - P_j)}$. Así obtenemos

$$P(L_2 = i) = \sum_{j \neq i} \frac{P_i P_j}{(1 - P_j)}$$

y finalmente

$$Q_i^2 = P_i \left(1 + \sum_{j \neq i} \frac{P_j}{(1 - P_j)} \right) \tag{2}$$

Este resultado también se hubiera podido obtener usando el siguiente lema:

Lema 1. Sean A, B, C tres eventos mutuamente excluyentes cuya probabilidad es respectivamente a, b y c . Además suponemos que $a + b + c = 1$. Si se realiza una serie infinita de experimentos con estos posibles resultados, entonces la probabilidad de que el evento A ocurra antes que el evento C es igual a $\frac{a}{1-b}$.

Demostración: Sean X_i las variables aleatorias independientes relacionadas al experimento. La probabilidad buscada es

$$\begin{aligned} P &= P(X_1 \in A) + P(X_2 \in A)P(X_1 \in B) \\ &\quad + P(X_3 \in A)P(X_k \in B, k < 3) + \dots \\ &= \sum_{n=1}^{\infty} P(X_n \in A)P(X_k \in B, k < n) \\ &= \sum_{n=1}^{\infty} ab^{n-1} \\ &= \frac{a}{1-b} \end{aligned}$$

Usando el lema con

- $A = \{ \text{se extrae la especie } i \},$
 - $B = \{ \text{se extrae la especie } j \}$ y
 - $C = \{ \text{se extrae una especie } k \text{ distinta a } i, j \}$
- se ve que

$$\begin{aligned} P(L_2 = i) &= \sum_{j \neq i} P(L_1 = j)P(A \text{ antes que } C) \\ &= \sum_{j \neq i} \frac{P_j P_i}{1 - P_j} \end{aligned}$$

De aquí se deduce la fórmula (2).

Caso $L = 3$. Tenemos

$$Q_i^3 = P(L_1 = i) + P(L_2 = i) + P(L_3 = i)$$

y además

$$\begin{aligned} P(L_3 = i) &= \sum_{j \neq i} P(L_1 = j) \sum_{k \neq i, j} P(L_2 = k | L_1 = j) \cdot \\ &\quad \cdot P(i \text{ antes de } l, \text{ si } l \neq j, k) \\ &= \sum_{j \neq i} P_j \sum_{k \neq i, j} \frac{P_k}{1 - P_j} \frac{P_i}{(1 - P_j - P_k)} \\ &= \sum_{j \neq i} \sum_{k \neq i, j} \frac{P_j P_k P_i}{(1 - P_j)(1 - P_j - P_k)} \end{aligned}$$

Por lo tanto

$$Q_i^3 = P_i \left(1 + \sum_{j \neq i} \frac{P_j}{1 - P_j} \left(1 + \sum_{k \neq i, j} \frac{P_k}{(1 - P_j - P_k)} \right) \right)$$

En el caso general tenemos:

$$P(L_{r+1} = i) = \sum_{\substack{j_k \notin I_k \\ k=1, \dots, r}} \frac{P_i \prod_{k=1}^r P_{j_k}}{\prod_{k=1}^r (1 - \sum_{s=1}^k P_{j_s})}$$

con $I_k = \{i, j_1, \dots, j_{k-1}\}$, de lo cual se deduce la fórmula (1).

6. El método de Newton

En esta sección establecemos un método efectivo de calcular Φ^{-1} por aproximaciones sucesivas. En general para una función diferenciable $f: \mathbb{R}^m \rightarrow \mathbb{R}^m$ se tiene que $f(x_0 + h) \approx f(x_0) + Df(x_0)h$ donde Df es la matriz de derivadas parciales $\left(\frac{\partial f_i}{\partial x_j} \right)_{i,j}$. Si asumimos además que $f(x_0 + h) = 0$ podemos despejar h y obtenemos $h \approx -Df(x_0)^{-1}(f(x_0))$. Por lo tanto si x_0 está cerca del cero de la función f , podemos hallar un valor más cercano al cero: $x_1 = x_0 + h = x_0 - Df(x_0)^{-1}(f(x_0))$. Esto inicia un proceso recursivo poniendo

$$x_{n+1} = x_n - Df(x_n)^{-1}(f(x_n))$$

con $\lim_n x_n = z$ donde z es una raíz de f . Este método es conocido como el método de Newton.

Notemos además que la función Φ está dada por la matriz

$$\Phi(P) = \begin{pmatrix} K_1 & 0 & \dots & 0 \\ 0 & K_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & K_{m-1} & 0 \\ 0 & \dots & 0 & K_m \end{pmatrix} \begin{pmatrix} P_1 \\ P_2 \\ \vdots \\ P_{m-1} \\ P_m \end{pmatrix}$$

donde

$$K_i = \frac{Q_i}{P_i} = 1 + \sum_{j_1 \neq i} \frac{P_{j_1}}{1 - P_{j_1}} \left(1 + \dots \left(1 + \sum_{j_{N-1} \neq i, j_1, \dots, j_{N-2}} \frac{P_{j_{N-1}}}{(1 - \sum_{s=1}^{N-1} P_{j_s})} \right) \dots \right)$$

Ponemos $f(P) = \Phi(P) - \hat{Q}$, donde $\hat{Q} = (\hat{Q}_i)$ es el vector de las observaciones, que es la frecuencia relativa con la que cada especie i aparece en una lista y por lo tanto el estimador de máxima verosimilitud para Q_i^N . Finalmente aproximamos

$$\frac{\partial Q_j}{\partial P_i} \approx \delta_{ij} K_j \tag{3}$$

La aproximación por el método de Newton nos da la fórmula recursiva:

$$\begin{aligned} P_i^{(n+1)} &= P_i^{(n)} - [D\Phi(P^{(n)})^{-1}(f(P^{(n)}))]_i \\ &= P_i^{(n)} - \sum_{j=1}^m \delta_{ij} K_j^{-1} [\Phi(P^{(n)}) - \hat{Q}]_j \\ &= P_i^{(n)} - \sum_{j=1}^m \delta_{ij} (K_j^{-1} K_j P_j^{(n)} - K_j^{-1} \hat{Q}_j) \\ &= P_i^{(n)} - P_i^{(n)} + K_i^{-1} \hat{Q}_i \\ &= K_i^{-1} \hat{Q}_i \\ &= \frac{P_i^{(n)}}{Q_i^{(n)}} \hat{Q}_i. \end{aligned}$$

El valor inicial canónico es $P^{(0)} = \hat{Q}/N$. Luego se calcula $Q^{(0)} = \Phi(P^{(0)})$, y recursivamente $P_i^{(k+1)} = \frac{P_i^{(k)}}{Q_i^{(k)}} \hat{Q}_i$ con $Q^{(k)} = \Phi(P^{(k)})$. El límite de este proceso nos da una familia $\{P_i\}$ con $\phi(\{P_i\}) = \{\hat{Q}_i\}$, pero no necesariamente se trata de probabilidades. Si en cada paso se corrige todos los P_i a la vez, multiplicándolos por algún factor común de modo que la suma sea igual a 1, el límite es la familia de probabilidades buscada. Este proceso es bastante manejable y da buenos resultados. El tiempo necesario para alcanzar una exactitud de 0.00001 se encuentran en la tabla 1. A mayor cantidad de especies, la cantidad de pasos necesarios aumenta.

7. Métodos alternativos

Como se dijo anteriormente y se ilustró con la tabla de los tiempos requeridos, un programa que calcule la fórmula exacta de $\Phi(P)$ para $N = 10$ o $N = 20$ demandaría por lo menos 100^{20} pasos de cálculo tomando en cuenta que se trata de 100 o más especies. Aún en el caso de tener una computadora mucho más potente, el cálculo exigiría mucho tiempo de cómputo. Se probaron cuatro métodos distintos para aproximar Φ en el caso $N = 5$, y se comparó con el valor exacto.

PRIMER METODO: Aproximación lineal: Se nota que los valores $Q_i^{(1)}, Q_i^{(2)}, Q_i^{(3)}, Q_i^{(4)}$ y $Q_i^{(5)}$ forman aproximadamente una recta. Por lo tanto se calcula $Q_i^{(1)}$ y $Q_i^{(2)}$ y se aproxima $Q_i^{(5)} \approx Q_i^{(1)} + 4\Delta$ donde $\Delta = Q_i^{(2)} - Q_i^{(1)}$. En el caso de listas grandes de 10 y de 20 especies respectivamente, se calcularían los valores de $Q_i^{(1)}$ a $Q_i^{(5)}$, se tomaría la recta más cercana a estos valores y se aproximaría $Q_i^{(10)}$ y $Q_i^{(20)}$. El tiempo de cómputo es mínimo, pero la aproximación no es muy buena.

SEGUNDO METODO: Aproximación cuadrática y polinomial: Se asume que los valores $Q_i^{(1)}, Q_i^{(2)}, Q_i^{(3)}, Q_i^{(4)}$ y $Q_i^{(5)}$ están sobre una parábola. Por lo tanto se calcula $Q_i^{(1)}, Q_i^{(2)}$ y $Q_i^{(3)}$ y se aproxima $Q_i^{(5)}$ con el valor del polinomio de segundo grado que pase por estos tres primeros puntos. En el caso de listas largas se calcularían los valores de $Q_i^{(1)}$ a $Q_i^{(5)}$, se tomaría el polinomio de grado 2 (o 3 o 4) que pase por estos puntos y se aproximaría $Q_i^{(10)}$ y $Q_i^{(20)}$.

El tiempo de cómputo es un poco mayor, la calidad de la aproximación parece ser buena, pero hay que tener en cuenta que para aproximar $Q_i^{(5)}$ se calculó hasta $Q_i^{(3)}$. Hay que hacer pruebas acercándose lo más posible a los valores exactos (para $N=10$ y $N=20$) por simulación, ya que no se pueden hallar con la fórmula exacta.

TERCER METODO: Estimación de la cantidad de observaciones que se tuvo que hacer antes de que se complete la lista. El cálculo de probabilidades elemental nos dice que la probabilidad de que la i -ésima especie aparezca en una lista después de u observaciones es de

$$P(i \in L) = 1 - (1 - P_i)^u$$

Si ponemos $\hat{Q}_i = 1 - (1 - P_i)^u$, aplicando el método de Newton podemos hallar un u tal que $\sum_{i=1}^m \hat{Q}_i = N$. Notemos que las probabilidades de aparecer en una lista suman N y no 1 como podría pensarse. Este método tiene un sesgo sistemático de estimar demasiado alto el Q_i si este ya es alto de por sí, pero el error total es comparable al del segundo método.

CUARTO METODO: Mixto.

Este cuarto método calcula primero las probabilidades hasta Q_i^2 , para luego estimar la cantidad de intentos que faltan para completar la lista y seguir como en el tercer método. En particular se calcula

$$\hat{Q}_i^{(5)} = P_i + \sum_{j \neq i} P_j \left(1 - \left(\frac{1 - P_j - P_i}{1 - P_j} \right)^u \right)$$

aplicando el método de Newton podemos hallar un u tal que $\sum_{i=1}^m \hat{Q}_i^{(5)} = 5$. Este es el método con el menor error de todos, aunque sigue teniendo el mismo error sistemático que el tercer método. En el caso de listas de 10 o de 20 se calcularía hasta $Q_i^{(4)}$ o $Q_i^{(5)}$ dependiendo de la capacidad de cómputo, de la cantidad de especies y de la exactitud requerida.

Se tienen que hacer pruebas en el futuro para confirmar la idoneidad de este método. Asimismo se debe encontrar alguna manera de contrarrestar el sesgo sistemático hallado.

8. Conclusiones

El modelo de Hubbell, tal vez mejorado o adaptado, puede dar algunas medidas útiles para evaluar el impacto de la actividad humana en comunidades ecológicas. Se encontró una manera de transformar los datos recogidos por el método de las listas de tamaño fijo en abundancias relativas. Se deben desarrollar métodos para lograr esta transformación en la práctica, especialmente para listas grandes, de 10 o 20 especies.

Agradecimientos

Los autores agradecen a todos los participantes del coloquio CORBIDI-IMCA por las fructíferas discusiones. El presente trabajo se realizó con apoyo del CONCYTEC a través del proyecto CS-160-2006-OAJ.

1. Hubbell, S.P., The Unified Neutral Theory of Biodiversity and Biogeography (Princeton University Press 2001)
2. Fisher R. A., Corbet, A.S., Williams, C.B. The relation between the number of species and the number of individuals

- in a random sample of an animal population. Journal of animal ecology 12 p.42-58 (1943)
3. Preston, F.W. The commonness and rarity of species. Pitman, London. (1948)