1-Introducción

El *Isolation By Distance* (IBD) (Wright 1943), se define como el aumento de la distancia genética entre individuos de una especie con el aumento de la distancia geográfica que los separa. Esta relación se puede medir como la regresión de una variable de *distancia* genética (que denotaremos por Y), sobre una variable de distancia espacial (que denotaremos por X). Sus formas se definen basándose en la relación entre el flujo genético y la deriva genética, generalmente no se debe asumir que la relación es lineal Hutchinson y Tempelton (1999). Por ello, debemos buscar un modelo más flexible, pero es de gran importancia que los parámetros del modelo escogido tengan una clara interpretación biológica, por lo que se considerará un modelo paramétrico. Aunque este patrón es un fenómeno muy conocido, no conocemos trabajos que aborden su modelización de manera formal, sin embargo, hay otro patrón que sí que está muy estudiado y modelizado, y que guarda estrecha relación con nuestro problema, el *distance decay of similarity.* Se define como el descenso de la similitud entre comunidades (es decir, el conjunto de especies que viven en una localidad) con el aumento de la distancia geográfica que las separa (Nekola y White 1999). La relación entre la similitud de las comunidades y la distancia espacial se puede modelar también mediante la regresión de una variable de similitud entre comunidades (índice de similitud) (denotada por Y), sobre una variable de distancia espacial (denotada por X). Dado el paralelismo entre los procesos de IBD y *distance decay,* es lícito pensar que los modelos que se usan para modelar el *distance decay* también pueden ser útiles para IBD.

Unas de las primeras modelizaciones matemáticas formal se encuentra en Nekola y White (1999). Estos autores proponen la modelización mediante modelos lineales. Concluyeron que el modelo más adecuado era:

$$\log(S) = \log(s_0) - xc + \varepsilon. \qquad (1.1)$$

donde S es la similitud entre comunidades, s_0 la similitud a distancia 0, x es la distancia geográfica que separa a las comunidades, c la pendiente de *distance decay* y ε un error aleatorio. Los inconvenientes de este modelo son que las observaciones con un valor de similitud 0 no se pueden transformar con el logaritmo, y su eliminación puede alterar el resultado del ajuste. Millar *et al*. (2011) propusieron como solución que los ajustes se realizasen con un modelo lineal generalizado (GLM, del inglés General Linear Model).

$$E(Y|X) = g^{-1}(X\beta),$$

con Y la variable dependiente, X la variable explicativa, β una serie de parámetros a estimar y g una función enlace. En este trabajo utilizaron como función enlace el logaritmo, asumiendo para la variable dependiente una distribución binomial. De esta forma se puede modelizar la relación (1.1) sin necesidad de eliminar ni transformar los ceros, y sin tener que transformar las variables.

Más recientemente Nekola y McGill (2014) comprobaron a pequeñas escalas la relación se aproxima con una función de forma power-law, mientras que a grandes escalas predomina la relación de forma exponencial.

En la mayor parte de estos modelos el intercepto se toma como indicador de la similitud a distancias cortas, es decir, a pequeñas escalas, mientras que la pendiente indica la tasa de substitución de especies (*turnover*) en las comunidades por unidad de distancia (Soininen *et al.* 2007, Steinbauer *et al.* 2012, Qian y Ricklefs 2012).

Por tanto, tomamos el GLM como el modelo más utilizado para modelar el patrón de disminución de la similitud biológica ente comunidades con la distancia y en consecuencia se tomará como referencia para los modelos de *distance decay*.

$$Y = e^{\beta X} + \varepsilon = e^{\beta_0} e^{\beta_1 X} + \varepsilon = a e^{bX} + \varepsilon \text{ . con } Y \sim B(n, p).$$

Los inconvenientes de esta aproximación son que los ajustes con estos modelos pueden tener un intercepto mayor que 1, lo que supone un desajuste con la realidad, ya que los índices de similitud toman valores entre 0 y 1, y que la disminución de la similitud con la distancia se produciría siempre desde el principio, y no tiene porqué ser así. Por tanto, sería apropiado contar con un modelo que pudiera ajustarse a estas dos situaciones. Tampoco se ha descrito un modelo para el IBD que pueda tener en cuenta estos factores, por eso en este trabajo se ha explorado una modelización de IBD y *distance decay* con una función que sí lo pueda hacer , la función de Gompertz. Este modelo se usa en varios campos del estudio biológico, incluso en alguna ocasión se ha usado para ajustar patrones de *distance decay* (Brownstein *et al.* 2012), pero el citado trabajo puede considerarse una excepción.

La función de Gompertz presenta la siguiente forma:

$$Y = a e^{-b e^{-cx}} + \varepsilon, \qquad (1.9)$$

donde *a* es la asíntota del modelo, *b* refleja el desplazamiento de la curva en el eje *x* (relacionado con el intercepto) y *c* la tasa de crecimiento. Este es un modelo diseñado especialmente para modelar patrones de crecimiento, por o que se usará para modelos de IBD, para un decrecimiento se usará (1.10).

$$Y = 1 - (a e^{-b e^{-cx}}) + \varepsilon,$$

Sin embargo este modelo puede tomar valores menores que cero, por lo que, para que pueda cumplir la exigencia de que a grandes distancias las similitudes han de ser 0 se puede forzar a que la asíntota inferior sea 0 con la formulación (1.11).

$$Y = 1 - (1 e^{-b e^{-cx}}) + \varepsilon,$$

Este es el modelo que se utilizará para el *distance decay*. Se puede hacer fácilmente la comprobación de que toma valores entre 0 y 1, teniendo en cuenta que el intercepto es $1 - (1e^{-b})$, de la siguiente forma:

$$\lim_{x \to \infty} 1 - \left(1 e^{-b e^{-cx}}\right) = 0\,, \qquad \lim_{b \to \infty} 1 - (1 e^{-b}) = 1.$$

De esta manera tenemos un nuevo modelo con dos parámetros que toma valores en el mismo rango que nuestra variable dependiente, y que permite ajustes a situaciones en las que la similitud pequeñas distancias es muy elevada.

Datos

Para *distnce decay* se han utilizado 21 *datasets* que representan a 21 taxones (grupo de especies relacionadas) de coleópteros europeos de donde se han tomado datos de presencia-ausencia de las especies que los componen en 16 países de la zona sur. Se dispone de

información sobre características biológicas de las especies (porcentaje de especies ápteras, número de especies y tamaño medio de las espcies) y el tamaño medio de los rangos de distribución. El índice de similitud se ha calculado con la función *beta.pair* del paquete *betapart* (Baselga *et al.* 2018) de R (R core team 2019). Las distancias geográficas entre países se han calculado como las distancias euclídeas entre los centroides de los polígonos que forma cada país. Se han realizado los ajustes del patrón con un GLM y con un modelo Gompertz con asíntota en 0, que se evaluarán con el criterio de información de Akaike (AIC). Se ha estudiado la correlación entre las características biológicas de las especies de cada taxón con los parámetros de un modelo de Gompertz forzado para explorar la interpretación biológica de esos parámetros.

Para comprobar el comportamiento del modelo de Gompertz en datos reales se han tomado 9 datasets con información genética (cox-1, gen neutral) y espacial de 4 especies de coleópteros y 5 especies de arañas distribuidas por todo el mundo. Estos *datasets* han sido obtenidos de Bold[1] (Ratnasingham & Hebert 2007) y de los trabajos de Smith y Farrell (2005) y Maroja *et al.* (2007). Las distancias geográficas se han calculado como las distancias euclídeas entre los puntos de muestreo. Posteriormente se han calculado las distancias genéticas entre secuencias con la función *dis.aligment* del paquete *seqinr* (Charif y Lobry 2007).

Debido al interés que la interpretación de los parámetros de un modelo tiene en un contexto biológico, es de gran utilidad contar con una herramienta de comparación de parámetros en un contexto como el nuestro, donde los datos presentan una dependencia espacial poco que no es la habitual, y donde no funcionan las herramientas estadísticas clásicas.

Por tanto, en este trabajo se pretende comprobar las ventajas prácticas de la función de Gompertz para ajustar los patrones de *isolation by distance* y *distance decay*, explorar el significado biológico de sus parámetros y construir un estadístico que permita compararlos.

2-Marco de dependencia: problemas y alternativas

En el tipo de datos con los que aquí trabajamos, cada valor de disimilitud es el resultado de la comparación de dos observaciones, y por tanto, cada observación interviene en el cálculo de varios valores de disimilitud. De esta forma, los resultados de las comparaciones derivadas de un mismo punto no serán independientes. Así, tanto en la simulación de datos como en la aplicación y calibrado de un test, se requerirá la aplicación de técnicas que sean adecuadas para poder recoger y analizar esta dependencia. Para la simulación de datos los modelos geoestadísticos clásicos no parecen adecuados, por lo que actualmente se trabaja con construcción de una matriz de varianzas-covarianzas que recoja la correlación de los procesos espaciales. A partir de esta matriz se podrán obtener variables que tengan una correlación similar a la que queremos testar. Por otro lado, para poder aproximar las distribuciones de estadísticos bajo la hipótesis nula considerada o construir vectores de réplicas de las estimaciones de los parámetros se utilizan técnicas de remuestreo. En este contexto la técnica más efectiva es que Bootstrap por bloques, tomando como bloques todas as observaciones derivadas de un mismo punto.

Página web: http://boldsystems.org/

3-Estadístico para comparación de parámetros en modelos biológicos

Considerando un modelo de Gompertz, y suponiendo que tenemos dos grupos de datos $(\mathrm{X}_{\mathrm{i}}^{1}, \mathrm{Y}_{\mathrm{i}}^{1})$, (X_i^2, Y_i^2), y que al ser un modelo paramétrico depende de un vector de parámetros $\theta_1 = (a_1, b_1, c_1)$ y $\theta_2 = (a_2, b_2, c_2)$ queremos testar si:

$$H0\colon \theta_1 = \theta_2 \text{ frente a } H1\colon \theta_1 \neq \theta_2,$$

donde f_1 y f_2 son las funciones de regresión de cada grupo. Equivalentemente se puede escribir:

$$H0\colon (a1, b1, c1) = (a2, b2, c2) \text{ frente a } H1\colon (a1, b1, c1) \neq (a2, b2, c2).$$

Se considerará un estadístico utilizado en el contexto biológico para la comparación de parámetros. Este estadístico compara las estimaciones de los parámetros del primer grupo con un vector de remuestras de las estimaciones de parámetros del segundo grupo, de la siguiente forma:

$$T = \frac{1}{M}\sum_{m=1}^{M} I\left[\left(\widehat{a_1}, \widehat{b_1}, \widehat{c_1}\right) > \left(\widehat{a_2^m}, \widehat{b_2^m}, \widehat{c_2^m}\right)\right],$$

siendo I la función indicadora y M la longitud del vector de remuestras de las estimaciones de parámetros.

Si consideramos que el vector de parámetros tiene una única componente este estadístico sigue una distribución $\mathrm{N}\left(0.5, \frac{0.25}{\mathrm{M}}\right)$. Equivalentemente, y para que siga una distribución estándar, lo podemos escribir como:

$$T = \frac{1}{\sqrt{\mathrm{M}}\,\sqrt{0.5(1\text{-}0.5)}}\left(\sum_{m=1}^{M} I(\widehat{a_1} - \widehat{a_2^m} > 0) - \mathrm{M}{*}0.5\right) \rightarrow \mathrm{N}\,(0,1).$$

El inconveniente es que las estimaciones para el segundo grupo no son independientes, pero los resultados de las simulaciones revelan que el test funciona bien y puede ser de utilidad para las comparaciones de parámetros que aquí nos interesan.

Para el caso de vectores de parámetros con dos o más componentes, la correlación entre las funciones indicadoras hacen que el test pierda utilidad, aunque si la tiene para la comparación de parámetros individuales.

4-Simulaciones

Para comprobar el desempeño del test propuesto se ha hecho un estudio de simulación en que se ha estimado su aproximación al nivel y su potencia. Para ello se ha definido una variable explicativa *x* como 200 valores contenidos en [0,1] y se han calculado las variables

respuesta y_i e y_j como un modelo de Gompertz con parámetros predefinidos, al que se le ha añadido un error aleatorio con distribución N(0,σ) con σ 0.1 y 0.2. Con estos modelos se ha calculado la estimación de los parámetros del primer grupo y un vector de 200 remuestras de las estimaciones de los parámetros del segundo grupo y se ha calculado 1000 valores del estadístico y calculando el p-valor correspondiente, finalmente se ha calculado el porcentaje de rechazos de la hipótesis nula en cada caso.

También se ha querido ver el desempeño del test con datos con dependencia espacial, aunque no sea como la que tienen nuestros datos, el proceso ha sido similar al anterior, solamente que se han simulado 100 datos y los errores añadidos son datos en una rejilla de 10x10 datos en el cuadrado unidad [0,1]x[0,1] donde se generan valores de un proceso espacial Gaussiano de media cero con estructura de dependencia exponencial.

Tras el estudio de simulación el test propuesto parece tener un buen desempeño general tanto en términos de calibrado (porcentaje de rechazos bajo la hipótesis nula, donde dichos porcentajes se aproximan al valor nominal considerado, del 1%, 5% o 10%) como en términos de potencia. Los resultados tienden a ser algo mejores en el caso de variables independientes, aunque el desempeño general del test es bueno en ambos casos.

5-Resultados y aplicación

***Distance decay**:* En 14 de los 21 casos estudiados, ambos modelos son equivalentes (diferencias de AIC <2). En 5 casos el modelo Gompertz forzado presenta un mejor ajuste que el modelo exponencial y en 2 casos es el modelo exponencial es el que tiene un mejor ajuste.

El parámetro *b* se relaciona con el intercepto y es indicador de similitudes y capacidad de dispersión de las especies a pequeñas distancias.

El parámetro *c* es el único que se relaciona con el porcentaje de especies ápteras, por lo que indica las similitudes y capacidad de dispersión a grandes distancias.

***Isolation by distance**:* En 7 de los 9 casos estudiados, el modelo de Gompertz presenta un mejor ajuste que el modelo exponencial, en 1 caso es el modelo exponencial es el que tiene un mejor ajuste y en el caso restante ambos modelos son equivalentes.

El parámetro *a* indica la distancia genética máxima que alcanzan los individuos de una población, y el punto dónde se alcanza es el puno donde el flujo genético se ha interrumpido completamente.

El parámetro *b* se relaciona con el intercepto y es indicador de similitudes a pequeña distancia y el flujo genético que se establece en las mismas.

El parámetro *c*, indica el efecto que tiene la distancia espacial sobre la distancia genética, la capacidad de dispersión de los individuos y por tanto la capacidad de establecer flujo genético a grandes distancias.

En su aplicación a datos reales el test ha presentado un buen desempeño, tanto en el caso de *distance decay* como de IBD.

6-Conclusiones

Como se ha podido ver a lo largo de este trabajo, la función de Gompertz resulta muy útil para modelar patrones biogeográficos como el *Isolation by Distance* y el *distance decay*. Por su parte el test propuesto para la comparación de parámetros parece presentar un buen desempeño en la detección de desviaciones de la hipótesis nula de igualdad de parámetros entre dos modelos y en su aplicación a datos reales, cuando se trabaja con parámetros individuales, aunque posteriormente también se podrían considerar otros estadísticos de contraste.