TREBALL DE FI DE GRAU

Grau en Matemàtiques

Títol: Modelització de la biodiversitat d'espècies mitjançant la sèrie geomètrica

Autor: Oriol Ventura Ripoll

Director: Jan Graffelman

Departament: Estadística i Investigació Operativa

Convocatòria: 2015-6

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Modelització de la biodiversitat d'espècies mitjançant la sèrie geomètrica

Oriol Ventura Ripoll

maig 2016

Abstract

The geometric series has been extensively used as a model in studies on species diversity. The main objective of this bachelor's thesis is to investigate and compare the parameter estimators of the geometric model as it is used in diversity studies.

An introduction to biodiversity studies is given, and the most common graphics used in such studies are presented. Basic statistical models in the field (Fisher's log series, the log-normal model, the broken stick model and the geometric series) are discussed.

The thesis focuses on the geometric model, and six different estimators for this model are considered. Three of these estimators currently have widespread use, and the other three are new proposals. The use of these estimators is illustrated with empirical data from ecological diversity studies. Confidence intervals for the geometric parameter based on these estimators are, when possible, derived.

A limited set of simulations is performed in the R environment in order to evaluate the estimators in terms of bias, variance and mean squared error. Most estimators are seen to produce similar estimates, though the maximum likelihood estimator generally produces larger values. The differences in performance of the estimators are seen to be relatively small in the current simulations. The proposed one-parameter least squares estimator has the best performance.

Resum

La sèrie geomètrica és un model molt utilitzat en estudis de diversitat d'espècies. L'objectiu principal d'aquest treball és el d'investigar i comparar diferents estimadors del paràmetre del model geomètric.

Primer es fa una introducció de la biodiversitat, i es presenten les gràfiques més utilitzades per representar les dades. S'introdueixen, també, els models més populars en aquest camp (el model de Fisher, el log-normal, el del bastó trencat i la sèrie geomètrica).

Seguidament, el treball se centra en el model geomètric, i es presenten sis estimadors diferents del paràmetre per aquest model. Tres d'ells ja s'utilitzen actualment, i els altres tres són noves propostes. L'ús d'aquests estimadors s'il·lustra amb dades empíriques d'estudis ecològics. Posteriorment es presenta, en els casos possibles, un interval de confiança per aquests estimadors.

Per últim, es realitzen simulacions de dades amb R amb la finalitat d'avaluar els estimadors en termes de biaix, variància i error quadràtic mig. La majoria d'ells proporcionen estimacions similars, tot i que l'estimador de màxima versemblança proporciona valors més grans. Les diferències de la resta d'estimadors són relativament petites en les simulacions realitzades. Tot i així, l'estimador que proporciona l'error quadràtic mig més petit és l'estimador de mínims quadrats.

Índex

Capítol 1

Introducció

Aquest treball té l'objectiu d'estudiar la biodiversitat d'espècies que hi ha a la Terra, a través de models que expliquin com aquesta es comporta. El concepte de biodiversitat és un concepte relativament nou, que ha anat evolucionant al llarg dels anys, i la comunitat científica cada vegada mostra més interès en ella. Només cal observar el següent gràfic, que ens mostra la quantitat d'articles publicats al llarg dels últims 30 anys que citen la paraula biodiversitat, per adonar-nos d'aquest fet:

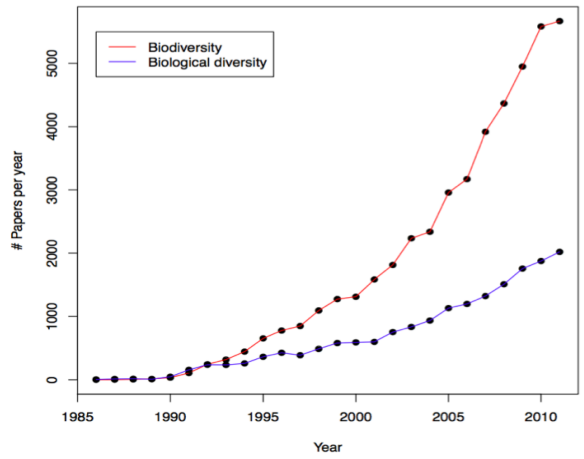

Figura 1.1: Augment de l'us de les paraules "Biodiversitat" i "Diversitat Biològica"en articles científics des de 1985 en endavant. Font: Curs d' estadística per les Biociències UPC: Jan Graffelman

Aquest creixement és degut als problemes medioambientals que han anat apareixent a causa de l'activitat humana (des del canvi climàtic fins a l'extinció de diverses espècies). La biodiversitat és un tema clau per entendre el perquè, i estudiar-la ens pot ajudar a buscar solucions a aquests problemes.

1.1 Definició de biodiversitat

S'entén per biodiversitat, o diversitat biològica, a la varietat i variabilitat de la vida que hi ha a la Terra, distribuïda en les diferents zones que la conformen. Aquesta, tant pot fer referència a la variació genètica de les espècies, com a la variació dels ecosistemes (conjunt d'espècies que comparteixen un medi físic i les relacions que s'hi estableixen), o a la variació de les poblacions d'espècies que habiten un cert ecosistema. Es pot dir que la biodiversitat engloba tota la informació sobre totes les espècies que habiten a la Terra, i com aquestes es distribueixen i es relacionen entre sí.

El concepte de biodiversitat és molt ampli; de fet, hi ha 3 classificacions de biodiversitat diferents: la diversitat d'ecosistemes, la diversitat d'espècies i la diversitat genètica. La diversitat d'ecosistemes té en compte la variabilitat estructural dels diferents ecosistemes que hi ha a la Terra, i la diversitat de processos que hi passen. Exemples d'ecosistemes són els deserts, els boscos tropicals, els parcs urbans, les terres conreades... Com més gran és la biodiversitat d'ecosistemes, més probable és que la vida perduri en el planeta. El concepte de biodiversitat d'ecosistemes és el més ampli dels 3. Seguidament, tenim la biodiversitat d'espècies, que fa referència a la diversitat d'organismes que hi ha en un mateix ecosistema i les relacions que s'hi estableixen entre elles. Per últim, tenim la biodiversitat genètica, que és la diversitat de genomes que es dóna entre els individus d'una mateixa espècie. Tracta sobre l'evolució de l'espècie en qüestió, i de com aquesta s'adapta al medi per sobreviure.

La biodiversitat en global és el resultat d'un procés històric natural de gran antiguitat. Des dels seus inicis fa aproximadament 3.000 milions d'anys, ha estat en constant evolució, buscant sempre l'equilibri dels ecosistemes. S'ha de tenir present que avui en dia es calcula que hi ha, en total, uns 10 milions d'espècies, el que implica que les relacions que s'hi estableixen són molt complexes. Hi ha forts lligams i dependències entre espècies i organismes, i qualsevol alteració dels ecosistemes pot generar un desequilibri enorme que comporta conseqüències rellevants en les poblacions d'espècies.

El gran problema d'aquesta qüestió és que l'ésser humà, des que té raó de pensar, i a mesura que ha anat evolucionant al llarg dels temps, ha anat modificant i destruint la biodiversitat. Activitats com projectes agrícoles, forestals, de transport, industrials, causen un fort impacte sobre ella. Aquesta és delicada i altament inestable, i qualsevol petita alteració pot degenerar a conseqüències rellevants. Per esmentar algun exemple clar de com és de fàcil pertorbar aquest equilibri, històricament hi ha hagut més d'un cas on l'ésser humà ha introduït artificialment una espècie en un ecosistema, i això ha provocat una forta alteració en les poblacions d'espècies i inclús l'extinció d'algunes d'elles. Tot això ens pot arribar a perjudicar molt, ja que l'alteració de la biodiversitat ens afecta directament, en formar part del nostre entorn, i la majoria de vegades l'impacte que ocasiona ens repercuteix negativament.

És per això que l'estudi de la biodiversitat és un tema del qual es té molt interès, i cada vegada més, ja que conèixer-la i entendre-la ens pot ajudar a preservar-la. Des del punt de vista concret de la matemàtica i l'estadística, l'objectiu de l'estudi de la biodiversitat es basa en poder modelitzar la seva complexa estructura. Precisament això és de què tracta aquest projecte: de modelitzar les poblacions d'espècies.

1.2 Aclaracions sobre aquest treball

Abans de desenvolupar el tema, hi ha algunes limitacions i assumpcions que s'han de comentar:

Aquest treball se centra en la biodiversitat d'espècies, i no en la biodiversitat d'ecosistemes o genètica. Dintre de la variació de les poblacions d'espècies hi ha 2 termes involucrats: la riquesa de les espècies i la uniformitat d'aquestes. La riquesa de les espècies és el nombre total d'espècies que habiten una certa zona, mentre que la uniformitat de les espècies és la seva abundància relativa. Per exemple, un bosc on el 80% dels ocells són d'una mateixa espècie i el 20% restant són de les 20 espècies d'ocells restants és una zona on clarament no hi ha uniformitat d'espècies, sinó que n'hi ha una que és dominant. Aquest projecte se centra a estudiar l'abundància relativa de les espècies.

A més a més, en aquest treball s'estudia i comparen espècies que, a part de compartir hàbitat, comparteixen certa relació de parentesc i que exploten els mateixos recursos. No compararem, doncs, plantes amb ocells, o peixos amb insectes. A aquest conjunt d'espècies filogenètic l'anomenarem conjunt característic.

Finalment, hi ha algunes assumpcions que s'han de comentar sobre la mesura de la diversitat.

Primer de tot, totes les espècies són iguals. Això significa que no hi ha espècies que rebin un "tracte especial". L'abundància relativa de cada espècie és l'únic factor que determina la seva importància en la mesura de la diversitat. Les mesures de riquesa no fan distincions entre espècies, i tracten de la mateixa manera tant aquelles espècies que són excepcionalment abundants com aquelles que són extremadament rares.

La segona assumpció és que tots els individus són iguals. Cal fer certs aclariments, però: primer, hi ha certs criteris per determinar quines espècies entren dins de la mateixa categoria o conjunt característic (per exemple, si estem tractant amb arbres, s'han d'establir unes condicions, com podria ser el diàmetre mínim del tronc, per poder classificar les espècies que entren dins d'aquesta categoria i les que no). I dintre d'una mateixa espècie, s'estudiaran els individus catalogats en el mateix cicle de vida. A partir d'aleshores, es tracten a tots els individus igual, sense fer distincions.

Per últim, s'assumeix que la informació sobre l'abundància d'espècies ha estat registrada utilitzant unitats apropiades i comparables. És evident que en un mateix estudi, no és prudent incloure diferents tipus de mesures d'abundància, com ara nombre d'individus i biomassa.

Capítol 2

Conceptes generals

Aquesta secció està dedicada a explicar alguns conceptes bàsics que s'utilitzen posteriorment, així com a explicar quins són els models més freqüents per descriure les poblacions d'espècies que s'estudien.

2.1 Mesura de la biodiversitat

Com bé ja sabem, no hi ha cap entorn on totes les espècies siguin igual de comunes. De fet, generalment en un ecosistema hi trobem espècies molt abundants, d'altres que són moderadament abundants, i la resta -la gran majoria-, que són escasses. Així doncs, el fet que l'abundància d'espècies difereix molt es capta en el concepte d'uniformitat d'espècies. Aquest és simplement una mesura de com les espècies similars són iguals en termes d'abundància. Per tant, un conjunt característic on la majoria d'espècies siguin igualment abundants és una comunitat molt uniforme.

El contrari d'uniformitat és dominància, que significa que una o poques espècies dominen la comunitat. S'associa alta diversitat amb alta uniformitat (o baixa dominància).

La informació d'uniformitat queda captada amb l'índex de Shannon (H') o amb l'índex de Simpson (D). L'índex de Shannon es defineix com:

$$H' = -\sum_{i=1}^{S} p_i ln(p_i),$$

on p_i és la proporció de l'espècie i i S el nombre d'espècies. L'índex de Shannon varia de 0 a $ln(S)$, i assoleix el màxim quan totes les espècies tenen la mateixa freqüència (són equifreqüents). Així doncs, $H'/H'_{max} = H'/ln(S)$ pren valors de 0 a 1, i ens diu com d'uniforme és la comunitat. Com més alt és aquest valor, més uniforme ho és.

D'altra banda, l'índex de Simpson es defineix com:

$$D = \frac{1}{\sum_{i=1}^{S} p_i^2}$$

Aquest varia de 1 a S, i assoleix el màxim quan les espècies són totes equifreqüents. $D/D_{max} = D/S$ pren valors de 0 a 1, i com més gran és aquest valor, més uniforme és la comunitat.

El fet que les espècies poden arribar a variar molt en termes d'abundància és el que va fer promoure als investigadors a desenvolupar models d'abundància d'espècies que expliquessin per què les espècies es distribuïen d'aquesta manera. Des de llavors, s'han anat desenvolupant diferents models, que encaixen millor o pitjor en segons quines comunitats.

2.2 Gràfiques per representar les dades

Sovint els estudis de biodiversitat són difícils de comparar a causa de la gran varietat de mètodes que hi ha per representar les dades. Diferents investigadors visualitzen les distribucions d'abundància d'espècies de maneres diferents. A continuació s'exposen algunes de les gràfiques més freqüents: rang/abundància, percentatge acumulat i individus/espècies.

Una de les gràfiques més conegudes i que millor informa és la gràfica rang/abundància. En l'eix de les x és reparteixen les diferents espècies, de nombre d'organismes major a menor. L'abundància relativa de cada espècie es representa en l'eix de les y en format comunament logarítmic (log_{10}), de manera que les grans diferències d'abundància entre espècies no són un problema a l'hora de representar-ho al gràfic. Aquestes gràfiques també s'anomenen gràfiques de Whittaker, en honor al seu creador. La següent figura n'és un exemple.

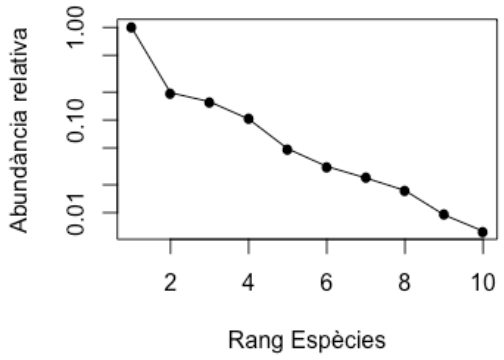

Figura 2.1: Exemple d'una gràfica rang/abundància. En l'eix de les x hi tenim les diferents espècies ordenades segons la seva abundància. A l'eix de les y hi tenim la seva abundància relativa corresponent, a escala logarítmica.

Amb aquesta representació és fàcil contrastar patrons de riquesa de les espècies. A més a més, quan S és relativament petita, tota la informació pertanyent a la seva abundància relativa és visible, cosa que no seria així si la representació fos en format d'histograma. De fet, és recomanable que el que primer faci un investigador amb les dades sigui representar-les en aquest format.

A continuació tenim una representació rang/abundància que il·lustra les formes dels models més utilitzats: el model geomètric, log normal, de Fisher i del bastó trencat.

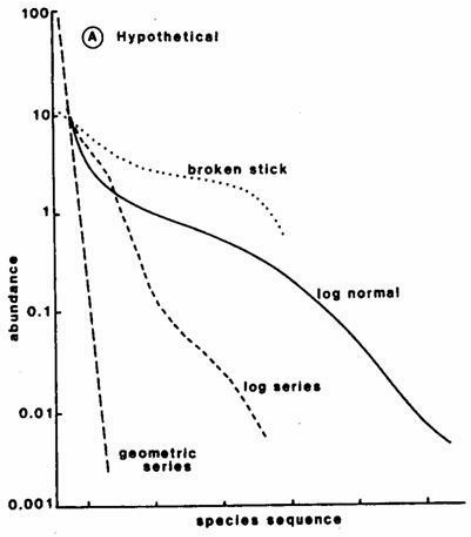

Figura 2.2: Gràfiques dels 4 models més utilitzats utilitzant la representació rang/abundància. Font: www.coastalwiki.org/wiki/Measurements_of_biodiversity

Un altre mètode comunament usat és la representació del percentatge acumulat. En l'eix de les x es distribueixen les espècies de menor a major abundància, i en l'eix de les y la seva abundància relativa més les abundàncies de les espècies més abundants que la en qüestió. D'aquesta manera, la gràfica és sempre creixent. La figura 2.3 ens en mostra un exemple.

Representant les dades d'aquesta manera, obtenim una gràfica que, com més elevada sigui, indica que menys diversa és la comunitat. Així doncs, és fàcil contrastar dades de diferents anys, per comprovar com evoluciona la biodiversitat.

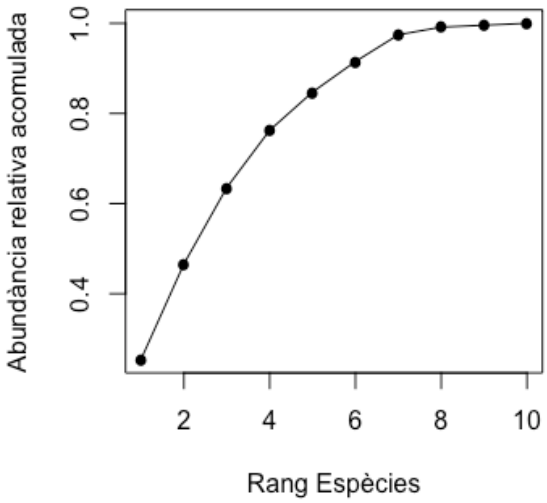

Figura 2.3: Exemple d'una gràfica de percentatge acumulat.

Per últim, un altre mètode també molt utilitzat és representar el nombre d'espècies (en l'eix de les y) en funció del nombre d'individus per espècie (en l'eix de les x). Quan es treballa amb el model log normal, s'acostuma a treballar amb aquest tipus de representació, però representant l'eix de les x en escala logarítmica. Les figures 2.4 i 2.5 exemplifiquen les dues versions.

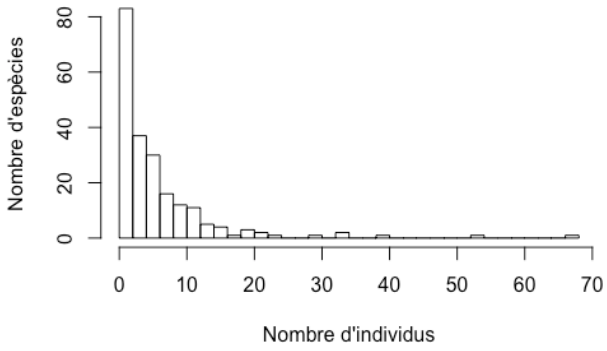

Figura 2.4: Exemple d'una gràfica individus/espècies.

Observem com cada mètode dóna èmfasi a una característica diferent de les dades. La representació convencional dóna èmfasi a l'abundància d'espècies rares, mentre que la transformació logarítmica de l'eix x ens revela el patró de la distribució normal a l'abundància d'espècies. Aquesta segona representació s'utilitza quan treballem amb el model log-normal, que precisament ens mostra

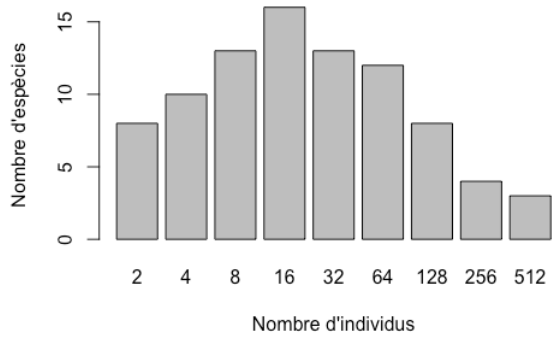

Figura 2.5: Exemple d'una gràfica individus/espècies, en format logarítmic.

com l'abundància d'espècies es comporta com una distribució normal quan hi apliquem l'escala logarítmica en l'eix de les x.

Per tal de facilitar la comparació de diferents dades entre investigadors, fa temps que es vol estandaritzar els mètodes per representar dades. És per això que, des de l'any 1999, la representació rang/abundància ha guanyat molta popularitat.

2.3 Models d'abundància d'espècies

Des que es va començar a investigar sobre la biodiversitat, s'han anat buscant diferents models per descriure la informació sobre l'abundància d'espècies, i avui en dia n'hi ha una gran varietat. Aquests els podem classificar en 2 grans varietats: els models estadístics i els models biològics (o teòrics).

Els models estadístics són aquells que han estat ideats per ajustar empíricament les dades que es té. Són models que intenten descriure els patrons observats. Un dels avantatges que presenten aquest tipus de models és que permeten comparar objectivament diferents conjunts característics. Exemples de models estadístics que veurem a continuació són el model de Fisher i el model log-normal.

Per altra banda, els models biològics intenten explicar, en comptes de descriure, l'abundància relativa d'espècies dins d'un conjunt característic. Per desenvolupar aquests models es necessita saber com es reparteix l'espai de nínxol (o recursos) disponible entre les diverses espècies per preguntar-se a continuació si l'abundància d'espècies observada encaixa amb l'expectativa. Exemples de models biològics que veurem a continuació són el model geomètric i el model del bastó trencat.

2.3.1 El model de Fisher

El model de Fisher,o model de la sèrie logarítmica de Fisher (*Fisher* 1943), va representar un dels primers intents de descriure la relació entre el nombre d'espècies i el nombre d'individus de cada espècie. Donat un paràmetre $p \in (0,1)$ i α (índex de diversitat) i el nombre d'espècies, la sèrie de Fisher es construeix de la següent forma:

$$\alpha p, \frac{\alpha p^2}{2}, \frac{\alpha p^3}{3}, ..., \frac{\alpha p^i}{i}, \quad i \in \mathbb{N}_{>0}$$

on $\frac{\alpha p^i}{i}$ representa el nombre d'espècies que tenen i individus. Com que $0 < p < 1$ i tant p com α són constants, el nombre d'espècies esperat serà més gran en la classe d'abundància més petita, i anirà disminuint com més gran sigui l'abundància d'individus. Com podem observar, per construir la sèrie es necessiten les dades en forma de nombre d'individus. p s'estima resolent iterativament l'equació:

$$S/N = \frac{1-p}{p}[-ln(1-p)]$$

Podem observar que p únicament depèn del quocient S/N. En el següent gràfic s'observa la relació entre p i S/N:

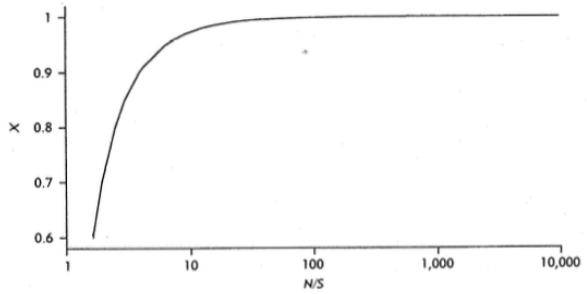

Figura 2.6: Relació entre p i S/N. Ja que a la pràctica aquest quocient és sempre major que 10, obtenim que p és sempre major que 0.9.

Així doncs, 2 paràmetres, α i N determinen completament la distribució, i estan relacionades per:

$$S = \alpha ln(1 + N/\alpha)$$

Com que $p \simeq 1$, α -l'índex de diversitat- representa el nombre d'espècies extremadament rares, on s'espera que només hi hagi un individu (a aquestes espècies les anomenem singletons). Aquest índex s'obté de l'equació:

$$\alpha = \frac{N(1-p)}{p}$$

Tot i que el model de Fisher va ser proposat com un model estadístic, la seva àmplia aplicació ha provocat que molts biòlegs consideressin processos ecològics que poguessin explicar el model des d'un punt de vista teòric. El model de Fisher sovint es relaciona amb la sèrie geomètrica (que posteriorment explicarem), a causa de la seva semblança.

2.3.2 La distribució log normal

La distribució log normal va ser aplicada per primer cop per Preston (*Preston* 1948). Per representar les dades, Preston va utilitzar l'escala logarítmica (log_2) i va dividir l'eix en vuitens. Tot i així, avui en dia es pot utilitzar qualsevol base logarítmica, sent comunes també les escales log_3 i log_{10}. Ara bé, és apropiat treballar amb log_2 si la classe mes abundant no supera els 2000 individus, i treballar amb log_{10} si la classe més abundant supera els 10^5 individus. La distribució s'escriu tradicionalment de la següent forma:

$$S(R) = S_0 exp(-a^2 R^2)$$

On $S(R)$ equival al nombre d'espècies en la vuitena R (classe) de la corba; S_0 la cresta de la curva; i $a = (2\sigma^2)^{-1/2}$ l'amplada inversa de la distribució. En la següent figura observem millor la distribució:

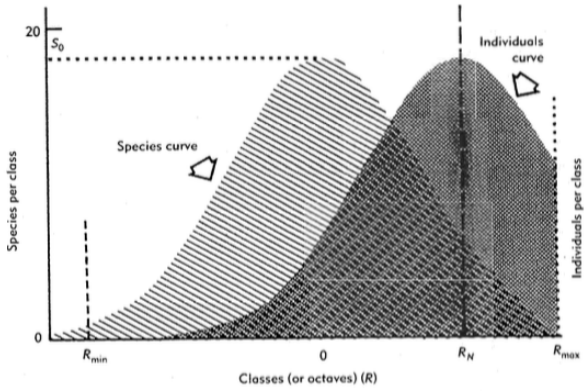

Figura 2.7: Exemple de la distribució log normal. A l'eix de les x hi tenim les diferents classes i en l'eix de les y el nombre d'espècies per classe. Font: Magurran 2004.

Aquest model és aplicat en la gran majoria de conjunts característics amb un volum elevat d'espècies i organismes. S'atribueix aquesta ubiqüitat de la log normal a les propietats matemàtiques dels conjunts de dades grans. De fet,

aquest model és una conseqüència del teorema central del límit. Per això, i com que encaixa tan bé amb els patrons observats, s'ha intentat buscar un significat biològic que expliqui per què les poblacions es distribueixen d'aquesta manera (i per tant s'ha intentat veure el model des d'un punt de vista teòric, en comptes d'estadístic).

Un dels problemes que representa el model log normal és la importància de la mostra observada que es necessita. Com més petita és la mostra, observem que una porció de la part esquerra de la corba (que representa les espècies més rares) no queda retratada. Aquest truncament serà més gran com més petita sigui la mostra observada. Això és degut al fet que, an tenir menys dades, ens passem per alt moltes espècies amb pocs individus. Per això, en estudis de grans conjunts característics, es requereix una mostra elevada (potser a través de diferents mètodes) per tenir en compte el màxim nombre d'espècies possibles.

2.3.3 El model geomètric

El model geomètric és un model que parteix de la següent idea: dins d'un conjunt característic d'espècies i una font de recursos limitada, l'espècie dominant s'endú una proporció k dels recursos; la segona espècie dominant se n'endú una part k del que queda $(1-k)$; la tercera espècie dominant altre vegada agafa una proporció k del sobrant i així successivament, fins que totes les espècies (S) s'han alimentat del recurs. Si es compleix aquesta assumpció, i si l'abundància de les espècies és proporcional a la quantitat de recursos que utilitzen, el patró d'abundància d'espècies seguirà una sèrie geomètrica. Aleshores, l'abundància de cada espècie estarà definida de la següent forma:

$$n_i = NC_k k(1-k)^{i-1},$$

on n_i correspon a l'abundància de cada espècie (ordenades major a menor abundància); N és el nombre total d'individus; k és la proporció de l'espai de nínxol que cada espècie s'endú (k és constant), i $C_k = [1-(1-k)^S)]^{-1}$ és una constant que ajusta les proporcions de tal manera que $\sum n_i = N$.

Com que el ràtio d'abundàncies amb seva predecessora és constant, la seva representació en el format rang/log(abundància) és una recta decreixent. En la figura 2.8 es pot observar aquest fet.

D'aquesta manera, representar les dades que un té en aquest format és una manera de veure visualment si aquestes segueixen o no aquesta distribució.

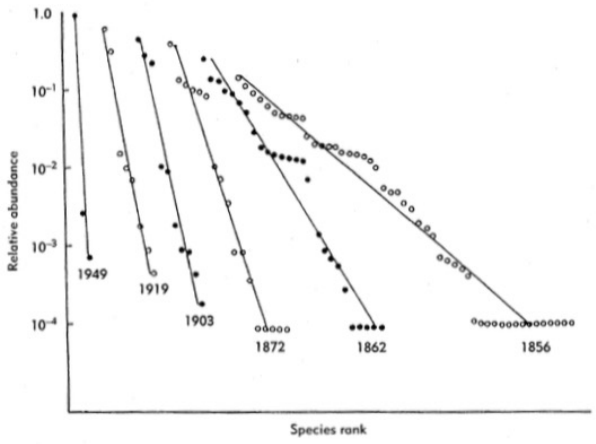

Figura 2.8: Evolució al llarg dels anys d'una comunitat d'espècies que segueixen el model geomètric. Font: Magurran 2004.

El patró de la sèrie geomètrica s'acostuma a veure en entorns poc rics d'espècies (i sovint durs), on els recursos són més aviat escassos. Quan les condicions milloren, és possible que altres models ofereixin una millor descripció de la comunitat. Tot i així, el model ens dóna una visió descriptiva molt interessant que és útil per analitzar els canvis en l'estructura d'una comunitat.

2.3.4 El model del bastó trencat

El model del bastó trencat parteix de la hipòtesi que la divisió de recursos entre espècies es comporta com un bastó trencat aleatòriament en S trossos. La distribució resultant és força uniforme; de fet, de les que ho és més. Com en el cas de la sèrie geomètrica, el model del bastó trencat s'escriu convencionalment en termes de rang d'abundància (de major a menor). El nombre d'individus de la i-èssima espècie més abundant (n_i) ve definit per:

$$n_i = \frac{N}{S} \sum_{n=i}^{S} \frac{1}{n},$$

on n_i és l'abundància de l'espècie i; N és el nombre total d'individus de la comunitat; i S el nombre total d'espècies.

És important observar que aquest model prediu la distribució d'abundància d'espècies esperada, i que per tant no té cap pàrametre, sinó que queda completament definida amb S i N. S'ha de dir que no són molts els casos en els quals aquest model s'ajusta bé a les dades. Tot i així, la importància d'aquest model rau precisament en el fet que ens demostra que els recursos no es reparteixen

aleatòriament, i això ens pot ajudar a buscar altres models que s'ajustin millor que aquest.

Capítol 3

El model geomètric

Com bé s'ha comentat anteriorment, l'objectiu del treball és estudiar la modelització de poblacions d'espècies a través de la sèrie geomètrica. Recordem que aquesta es basa en el repartiment de recursos limitats que hi ha en certa zona a cada una de les espècies, en ordre de dominància decreixent, de manera que, tenint en compte que l'abundància d'espècies és proporcional a la quantitat de recurs del qual disposen, les poblacions d'espècies queden determinades de la següent manera:

$$n_i = NC_k k(1-k)^{i-1},\ i = 1 \div S$$

La idea és que l'espècie més dominant agafa una part k dels recursos disponibles; la segona espècie dominant agafa una part k del que queda $(1-k)$; i així successivament.

3.1 Exemples

A continuació tenim 3 exemples diferents de poblacions d'espècies que podrien quedar modelitzades per la sèrie geomètrica.

3.1.1 Exemple 1: Escarabats

En la taula 3.1 hi tenim les diferents espècies d'escarabats trobades en un cert ecosistema, amb les seves abundàncies respectives trobades.

Obtenim una $S = 16$ (nombre d'espècies), i una $N = 1745$ (nombre d'individus). A la figura 3.1 hi tenim representades les dades en format rang/abundància.

	Espècies	Abundància
1	Onthophagus truncaticornis	897
2	Caccobius meridionalis	339
3	Onthophagus rectecornutus	144
4	Oniticellus cinctus	98
5	Onitis philemon	70
6	Onthophagus dama	63
7	Drepanocerus setosus	62
8	Caccobius unicornis	25
9	Copris indicus	16
10	Oniticellus spinipes	7
11	Onthophagus tarandus	7
12	Liatongus rhadamistus	6
13	Onthophagus catta	5
14	Onthophagus pactolus	2
15	Onthophagus spinifex	2
16	Sisyphys sp.	2

Taula 3.1: Població d'escarabats

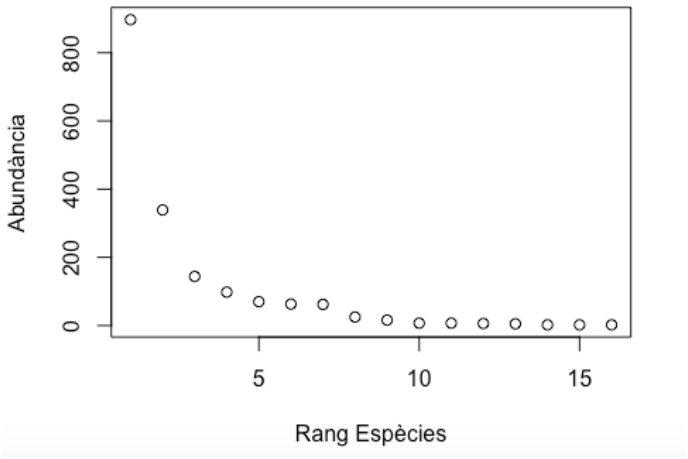

Figura 3.1: Abundància de les diferents espècies d'escarabats.

Podríem dir que hi ha dues espècies dominants, amb una població major de 300 (respecte al total, que és 1745), i que la resta són espècies més aviat rares. Ara, però, representem les mateixes dades utilitzant l'escala logarítmica en l'eix de les y (figura 3.2). Aquesta vegada, les dades estan repartides més uniformement, aproximadament com una recta. És en aquest moment quan s'observa que la sèrie geomètrica podria ser útil per modelitzar les poblacions (en aquest cas d'escarabats). Recordem que una de les propietats de la sèrie geomètrica és que quan representem les dades amb escala logarítmica, les dades formen una

recta, ja que les proporcions d'abundàncies entre espècie i la seva predecessora és constant.

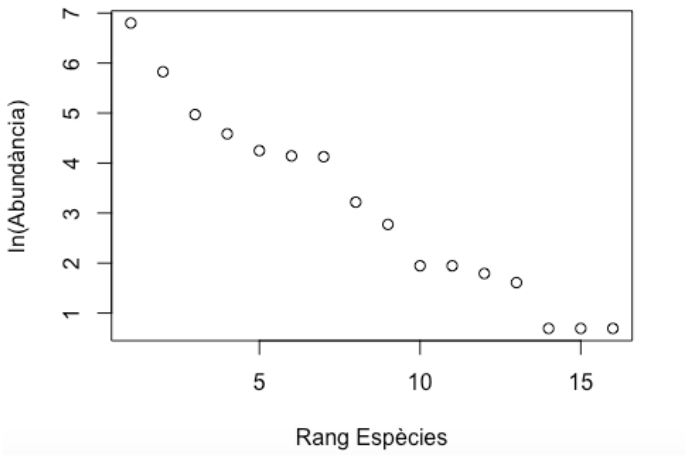

Figura 3.2: Abundància de les diferents espècies d'escarabats a escala logarítmica. S'observa una certa tendència lineal.

3.1.2 Exemple 2: Plantes

Aquesta vegada les espècies corresponen a la flora terrestre que hi ha a la roureda Breen (Breen Oakwood) situada a Irlanda del Nord. En la taula 3.2 hi tenim les diferents espècies amb les seves abundàncies.

	Espècies	Abundància
1	Luzula sylvatica	170
2	Deschampsia flexuosa	140
3	Vaccinium myrtillus	133
4	Oxalis acetosella	63
5	Molinia caerula	52
6	Polytrichum formosum	38
7	Holcus lanatus	37
8	Anthoxanthus odoratum	33
9	Rhytidiadelphus triquetrus	33
10	Pteridium aquilinum	29
11	Potentilla erecta	20
12	Sphagnum acutifolium	15

	Espècies	Abundància
13	Thuidium tamariscinum	15
14	Agrostis tenuis	14
15	Juncus effusus	13
16	Dicranum majus	11
17	Blechnum spicant	10
18	Rhytidiadelphus squarrosus	9
19	Sphagnum palustre	8
20	Calluna vulgaris	7
21	Hocus mollis	6
22	Hypnum cupressiforme	6
23	Dryopteris dilitata	4
24	Rhytidiadelphus loreus	4
25	Carex flexuosa	3
26	Gallium saxatile	3
27	Mnium hornum	3
28	Pseudocleropodium purum	3
29	Poa trivialis	2

Taula 3.2: Plantes de la roureda Breen, Irlanda del Nord

Obtenim una $S = 29$ i una $N = 884$. Representant les dades en el format rang/abundància, i usant l'escala logarítmica, obtenim el següent gràfic:

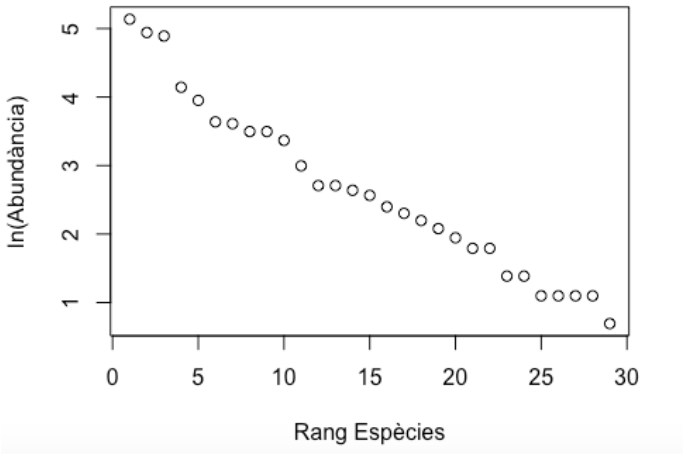

Figura 3.3: Abundància de les espècies de plantes a escala logarítmica.

Novament, observem com les dades es comporten de manera similar a l'anterior exemple. Hi ha una linealitat aproximada, i per tant és probable que puguem utilitzar la sèrie geomètrica per modelitzar les dades.

3.1.3 Exemple 3: Ocells

El següent joc de dades ens dóna informació sobre les espècies d'ocells que hi ha a dues quebrades de les Sierras Chicas, a la província de Córdova, Argentina. Les abundàncies, però, han estat simulades amb R, utilitzant el paràmetre $k = 0.25$, i se'ls ha aplicat un terme soroll posteriorment. La següent taula ens mostra les abundàncies obtingudes per cada espècie:

	Espècies	Abundància
1	Zenaida auriculata	1701
2	Columbfa maculosa	1453
3	Sappho sparganura	1191
4	Furnarius rufus	976
5	Turdus chiguanco	615
6	Myiopsitta monachus	427
7	Heliomaster furcifer	393
8	Colaptes melanochloros	338
9	Cinclodes fuscus	190
10	Synallaxis frontalis	115
11	Pseudoseisura lophotes	101
12	Taraba major	69
13	Elaenia albiceps	48
14	Serpophaga subcristata	46
15	Pitangus sulphuratus	27
16	Notiochelidon cyanoleuca	15
17	Troglodytes aedon	9
18	Cyclarhis gujanensis	6
19	Thraupis bonariensis	5
20	Sporophila caerulescens	5
21	Aimophila strigiceps	3

Taula 3.3: Abundància de les diferents espècies d'ocells trobades a les Sierras Chicas, Córdoba, Argentina.

Obtenim una $N = 7733$ i una $S = 21$. A la figura 3.4 hi ha representades les dades en format rang/abundància, i usant esala logarítmica a l'eix de les y.

Veiem altra vegada com la gràfica és semblant a una recta (aquesta vegada potser ho és més que en els exemples anteriors). Així doncs, utilitzarem la sèrie geomètrica per modelitzar les dades.

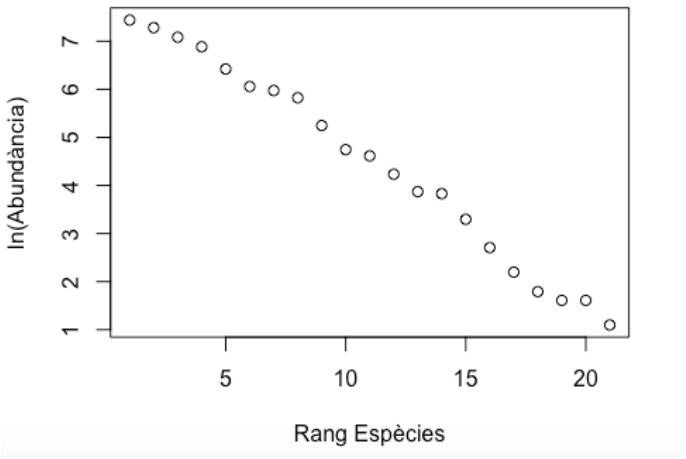

Figura 3.4: Abundància de les espècies d'ocells a escala logarítmica.

Una vegada s'ha reconegut el patró que modela el comportament de les poblacions, la pregunta que es fa a continuació és: sabent que les dades es comporten com una sèrie geomètrica, quin és el paràmetre que millor les descriu? Aquest és un dels principals problemes que es presenten llavors: com podem saber amb exactitud quin és el paràmetre correcte? No podem saber-ho. Existeixen, però, mètodes per estimar aquest paràmetre, per trobar-ne una aproximació que ens serveixi. També es poden buscar intervals de confiança, que ens donen un interval de valors pel paràmetre que volem estimar.

3.2 Mètodes per estimar el paràmetre de la sèrie geomètrica

A continuació s'exposen 3 mètodes diferents utilitzats per estimar el paràmetre k de la sèrie geomètrica (regressió lineal clàssica, el mètode de Newton i el mètode del mínim i el màxim), i 3 mètodes més proposats en aquest treball (una modificació de la regressió lineal clàssica, mínims quadrats i màxima versemblança).

3.2.1 Regressió lineal clàssica

Com bé s'ha comentat abans, una de les característiques de la sèrie geomètrica és que la proporció entre els diferents termes de la successió en format logarítmic és constant. Per això, el mètode de regressió lineal, explicat a continuació, és un dels primers mètodes que ens ve al cap.

Recordem que el nombre d'individus de l'espècie i ve donat per:

$$n_i = \frac{Nk(1-k)^{i-1}}{1-(1-k)^S} \tag{3.1}$$

Aplicant logaritmes a ambdós costats l'expressió esdevé:

$$ln(n_i) = ln(N) + ln(k) + (i-1)ln(1-k) - ln(1-(1-k)^S) =$$
$$= ln\left(\frac{Nk}{1-k-(1-k)^{S+1}}\right) + i \cdot ln(1-k)$$

Obtenim una funció lineal en i (de la forma $x(i) = \beta_0 + i\beta_1$), on β_0 i β_1 són dues constants a determinar, ja que ambdues depenen del paràmetre k. Fent una regressió lineal simple, amb l'ajuda de R, podem trobar fàcilment aquests valors. I un cop trobats, igualant l'expressió de β_1 al valor obtingut trobem una estimació del paràmetre k.

Amb les dades del primer exemple, fem una regressió lineal amb R per obtenir el paràmetre de la nostra sèrie. Sigui y el vector de les abundàncies n_i:

```
> x<-1:S
> lny<-log(y)
> Resultat <-lm(formula=lny~x)
```

Els resultats que s'obtenen són els següents:

Coefficients:				
	Estimate	Std Error	t value	$Pr(>\|t\|)$
Intercept	6.47031	0.18910	34.22	6.78e-15
x	-0.39311	0.01956	-20.10	1.00e-11

Taula 3.4: Regressió lineal amb R

Igualem el terme $\hat{\beta}_1$ obtingut:

$$\hat{\beta}_1 = -0.39311 = ln(1-k) \rightarrow \ \hat{k}_1^1 = 1 - e^{-0.3931} = 0.3251$$

Per tant, obtenim $\hat{k}_1^1 = 0.325$.

Segona versió de la regressió lineal

Fent la regressió lineal, s'ha observat que el terme β_0 també depèn de k. Així doncs, si en comptes d'utilitzar β_1, com ho fa la regressió lineal clàssica, utilitzem β_0, obtenim una altra estimació del paràmetre que busquem:

$$\hat{\beta}_0 = 6.47031 = ln\left(\frac{Nk}{1-k-(1-k)^{S+1}}\right)$$

Aquesta equació es pot resoldre amb el mètode de Newton, que s'explica a continuació. D'altra banda, però, observem que, en el denominador, el sumand $(1-k)^{S+1}$ és tan petit que el podem considerar nul (sense que això afecti gaire al resultat). D'aquesta manera, ens queda el següent:

$$\hat{\beta}_0 \approx ln\left(\frac{Nk}{1-k}\right) \rightarrow \hat{k}_1^2 = \frac{e^{\hat{\beta}_0}}{N + e^{\hat{\beta}_0}}$$

Utilitzant altre cop les dades dels escarabats, obtenim una segona estimació del paràmetre k:

$$\hat{k}_1^2 = \frac{e^{6.47}}{1745 + e^{6.47}} = 0.270$$

A partir d'ara, indicarem amb $\hat{k}_1^1$ l'estimació de k a través del pendent de la recta, i amb $\hat{k}_1^2$ l'estimació de k a través del terme independent.

3.2.2 Mínims quadrats

La idea d'aquest mètode és força semblant a l'anterior. Aquesta vegada, però, partim de la base que les nostres n_i estan definides de la següent manera:

$$n_i = \frac{Nk(1-k)^{i-1}}{1-(1-k)^S}\epsilon_i$$

On ϵ_i és una petita pertorbació. Acceptem, doncs, que les nostres dades segueixen una distribució geomètrica i que estan lleugerament pertorbades. El que volem és trobar el paràmetre k que minimitzi aquest error:

$$ln(n_i) = ln\left(\frac{Nk}{1-k-(1-k)^{S+1}}\right) + i \cdot ln(1-k) + ln(\epsilon_i) \rightarrow$$

$$ln(\epsilon_i) = e_i = ln(n_i) - ln\left(\frac{Nk}{1-k-(1-k)^{S+1}}\right) - i \cdot ln(1-k)$$

Per tant, el nostre objectiu és minimitzar la funció $f(k) = \sum_i e_i^2$. La k que faci aquesta funció mínima, és la k que estem buscant.

La funció *nlm* ens permet estimar el nostre paràmetre k, donant-li un valor inicial aproximat de k. L'avantatge que presenta aquest mètode en comparació amb la regressió lineal és que només necessitem un paràmetre (el propi paràmetre k) per realitzar la nostra estimació, a diferència dels dos paràmetres β_0 i β_1 del qual depenia la nostra recta quan fèiem servir la regressió.

En l'exemple dels escarabats, agafem $k = 0.5$ com a dada inicial. Sigui y el vector d'abundàncies n_i, cridem la funció *nlm* de la següent manera:

```
> x<-1:S
> Resultat<-function(p) {
+     sum((log(y)-log((1745*k)/(1-k-(1-k)^17))-x*log(1-k))^2)
+ }
> nlm(e,p=0.5)
```

El resultats que s'obtenen són els següents:

$minimum
[1] 2.28207
$estimate
[1] 0.343987
$gradient
[1] -2.664535e-09
$iterations
[1] 8

Taula 3.5: Mínims quadrats amb R

Per tant, amb el mètode dels mínims quadrats, obtenim la següent estimació de k:

$$\hat{k}_2 = 0.343987$$

3.2.3 Mètode de Newton

Un altre mètode per estimar el paràmetre k de la sèrie és el conegut mètode de Newton. Aquest serveix per trobar aproximacions dels zeros d'una funció real.

Sigui $f : [a, b] \longrightarrow \mathbb{R}$. Donat un valor inicial x_0 que estigui relativament aprop del zero que estem buscant, desenvolupem la següent successió:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

Aquesta successió convergeix cap al zero que busquem. Tornant al nostre cas, recordem que l'abundància de les espècies estava definida de la següent forma:

$$n_i = \frac{Nk(1-k)^{i-1}}{1-(1-k)^S}$$

Agafant l'abundància de l'espècie més escassa, tenim que:

$$n_{min} = \frac{Nk(1-k)^{S-1}}{1-(1-k)^S}$$

D'aquesta manera, podem definir la nostra funció f com:

$$f(k) := \frac{n_{min}}{N} - \frac{k(1-k)^{S-1}}{1-(1-k)^S}$$

Sigui $g(k)$:

$$g(k) := \frac{f(k)}{f'(k)} = \frac{(1-(1-k)^S)(k(1-k)^{S-1} - \frac{n_{min}}{N}(1-(1-k)^S))}{(1-k)^{S-2}(1-kS)(1-(1-k)^S) - kS(1-k)^{2S-2}}$$

Definim aleshores la nostra successió $\{k_n\}$ com:

$$k_{n+1} = k_n - g(k_n)$$

Així doncs, iterant aquesta expressió obtenim una nova estimació del paràmetre k. Utilitzant les dades dels escarabats, amb una aproximació inicial $k_0 = 0.2$ i amb l'ajut de R obtenim els següents resultats:

[0] 0.2000000
[1] 0.2587406
[2] 0.2954788
[3] 0.3099816
[4] 0.3119481
[5] 0.3119805
[6] 0.3119806
[7] 0.3119806

Observem com, realitzant 6 iteracions del mètode de Newton, obtenim una aproximació amb 7 xifres significatives, i concloem que $\hat{k}_3 = 0.3119806$.

Un dels problemes que presenta aquest mètode, però, és que hi ha vegades que no convergeix. Això depèn de l'aproximació inicial que li donem. En els tres exemples s'ha comprovat que, sempre que agafem una aproximació inicial inferior al resultat final, el mètode convergeix sense problemes ($k_0 \in (0, k)$). En canvi, si comencem amb una aproximació inicial superior al resultat, aquesta convergeix només si no és gaire més gran que k. Una $k_0 = 0.5$ com a aproximació inicial no ens serveix per convergir a una solució en el primer l'exemple dels escarabats.

S'ha de tenir en compte que per a estimar k amb aquest mètode s'ha optat per trobar-la a partir del coeficient n_S (el més petit de tots), però s'hauria pogut trobar a partir de qualsevol altre n_i (i el resultat lògicament variaria). La tria de n_S és completament arbitrària.

3.2.4 Mínim i màxim

Un altre mètode per estimar el paràmetre k, i que té en compte totes les abundàncies d'espècies n_i, és el mètode del mínim i màxim. Observem que:

$$\frac{n_{i-1}}{n_i} = \frac{1}{1-k}$$

Per tant, del següent producte s'obté:

$$(n_1/n_2)(n_2/n_3)...(n_{S-1}/n_S) = (1-k)^{1-S} \Rightarrow$$

$$\frac{n_{min}}{n_{max}} = (1-k)^{1-S} \Rightarrow \hat{k}_4 = 1 - \left(\frac{n_{min}}{n_{max}}\right)^{\frac{1}{S-1}}$$

Aquest mètode és molt simple i a més ens ofereix una expressió tancada per k. Tornant amb el nostre exemple dels escarabats, el mètode del mínim i màxim ens dóna la següent estimació de k:

$$\hat{k}_4 = 1 - \left(\frac{2}{897}\right)^{\frac{1}{15}} = 0.33439$$

3.2.5 Màxima versemblança

L'estimació del paràmetre de les distribucions de probabilitat pel mètode de la màxima versemblança és un dels mètodes més utilitzats. En el cas de la sèrie geomètrica, aquest mètode es basa a identificar la nostra sèrie geomètrica amb una distribució geomètrica truncada. Recordem que la funció de probabilitat per la distribució geomètrica és la següent:

$$Pr(X = i) = k(1-k)^{i-1}, \quad i \in \mathbb{N}_{>0}$$

D'altra banda, la distribució geomètrica té la següent esperança i variància:

$$E[X] = k, \quad Var[X] = \frac{1-k}{k^2}$$

Observem que l'expressió és força semblant a l'expressió de la sèrie geomètrica, tret del terme $(1-(1-k)^S)$, que no deixa de ser un factor lleugerament més petit que 1 que engrandeix els termes de la sèrie. Això és degut al fet que la distribució geomètrica és infinita, mentre que la sèrie és finita, i com que la suma total dels termes ha de ser igual a 1 en ambdós casos, els termes de la sèrie han de ser corregits perquè sumin el mateix que tots els termes de la distribució.

Així doncs, la nostra sèrie no és exactament una distribució geomètrica, però s'hi aproxima. Però quant? Si ens fixem en el terme $(1-(1-k)^S)$, veiem que, com més gran sigui k i com més gran sigui S, aquest més s'aproxima a 1, que és el que volem per poder tractar la sèrie com una distribució.

Això ens porta al següent fet: si les nostres dades són d'un conjunt característic amb moltes espècies, i si veiem que el conjunt és poc uniforme (hi ha molta desigualtat d'abundància entre les diferents espècies), podem suposar que les nostres dades segueixen una distribució geomètrica de paràmetre k.

La construcció és la següent: l'espècie amb major població és el succés $X = 1$ (el succés més probable), la segona espècie més freqüent equival al succés $X = 2$, i així successivament. D'aquesta manera, passem de tenir espècie 1,2,..S amb les diferents abundàncies a tenir un conjunt de $1's, 2's, ...S's$ (on l'abundància de l'espècie i es veu reflectida en el nombre de i's que tenim).

Partint d'aquesta base, ara podem estimar el paràmetre k utilitzant el mètode de la màxima versemblança. Sabent que l'estimació del paràmetre de la distribució geomètrica per aquest mètode és $1/\bar{x}$, obtenim el següent estimador $\hat{k}_5$:

$$\hat{k}_{ML} = \hat{k}_5 = \frac{1}{\bar{x}} = \frac{N}{\sum_{i=1}^{S} i\ n_i}$$

Una altra vegada, aquest mètode ens ofereix una expressió tancada per k. Tornant a l'exemple dels escarabats, l'estimació de k pel mètode de la màxima versemblança és la següent:

$$\hat{k}_5 = \frac{N}{\sum_{i=1}^{S} i\ n_i} = \frac{1745}{4279} = 0.4078$$

Cal tenir en compte, però, que el terme que obtenim $\hat{k}_5$ pot ser lleugerament superior al que realment hauríem d'obtenir. Això és degut al següent fet: estem suposant que els esdeveniments es distribueixen de manera que l'esdeveniment $X = i$ succeirà més vegades que l'esdeveniment $X = i + 1$, ja que $P(X = i) > P(X = i + 1)$, però a la pràctica no té per què ser així. Ens podem trobar en el cas en què passi el contrari (de fet és probable que passi quan i és gran, en algun moment), però en ordenar els esdeveniments en funció de la seva abundància impedim que això passi. I en calcular $\hat{k}_5$, aquest canvi es veu reflectit en què passem de tenir els sumands $in_i + (i + 1)n_{i+1}$ a tenir $in_{i+1} + (i + 1)n_i$, i com que $n_i < n_{i+1}$, obtenim un denominador més gran i per tant una estimació de $\hat{k}_5$ lleugerament inferior.

3.2.6 Resultats

En total, hem obtingut 6 estimacions de k utilitzant 5 mètodes diferents. Realitzem els mètodes amb el segon i tercer exemple. Els resultats obtinguts són a la següent taula.

Estimador	Exemple 1	Exemple 2	Exemple 3
$\hat{k}_1^1$	0.3250	0.1345	0.2828
$\hat{k}_1^2$	0.2687	0.1192	0.3012
$\hat{k}_2$	0.3440	0.1418	0.2582
$\hat{k}_3$	0.3120	0.1367	0.2805
$\hat{k}_4$	0.3344	0.1467	0.2712
$\hat{k}_5$	0.4078	0.1666	0.2543

Taula 3.6: Estimació de p mitjançant els diferents mètodes descrits, i pels diferents exemples comentats.

Primer, cal fer menció a un fet important: així com els 4 primers mètodes busquen l'estimació de k de manera que s'ajusti el millor possible a les proporcions (per tant als logaritmes de n_i), el mètode de la màxima versemblança busca l'estimació de k que millor s'ajusti a les dades reals, sense donar importància a les proporcions. Per veure-ho, observem la figura 3.5, on hi tenim, pel primer exemple i pels estimadors $\hat{k}_1^2$ i $\hat{k}_5$, les gràfiques esperades per aquests 2 paràmetres utilitzant l'escala normal i logarítmica de la representació rang/abundància en comparació amb les dades reals[1].

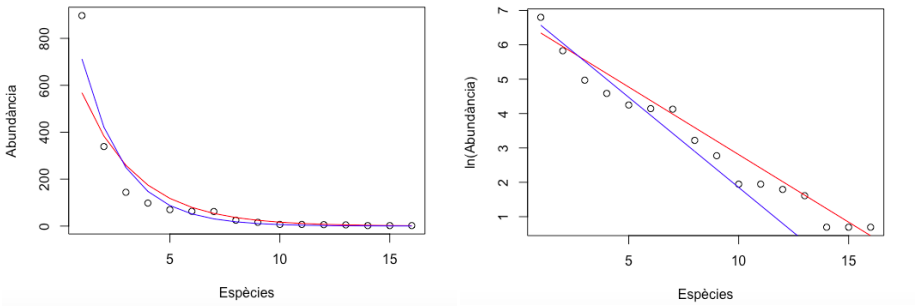

Figura 3.5: A l'esquerra, representació rang/abundància a escala normal; a la dreta, a escala logarítmica. Abundàncies reals a punts; abundàncies esperades per $\hat{k}_1^2$ en vermell, i per $\hat{k}_5$ en blau.

El que veiem és que l'estimador $\hat{k}_1^2$ ens dóna una millor aproximació que $\hat{k}_5$ per les proporcions, mentre que l'estimador $\hat{k}_5$ ens aproxima millor les abundàncies que $\hat{k}_1^2$. Conseqüentment, $\hat{k}_5$ estima millor n_1 que la resta d'estimadors, però aquests aproximen millor la cua dels n_i. Depenent de a què volem donar èmfasi quan volem estimar unes dades, convé doncs agafar o bé l'estimador de màxima versemblança o bé un dels altres. Normalment en estudis no es treballa amb $\hat{k}_5$, ja que es vol una estimació de les proporcions entre espècies, més que de les

[1]Hem agafat $\hat{k}_1^2$ com a representant dels 4 primers mètodes per mostrar la comparació entre el mètode de la màxima versemblança i la resta. Aquesta tria ha estat arbitrària.

mateixes abundàncies.

Com més semblants siguin els resultats obtinguts entre les dues classes d'estimadors, millor s'aproximen les dades a la sèrie geomètrica. En el cas dels escarabats, la diferència entre $\hat{k}_5$ i la resta és considerable, de quasi 0.1. Observem que els 4 primers estimadors ens donen uns valors relativament baixos, aproximadament 0.32. Aquest nombre és compatible amb les proporcions dels n_i, però amb una k tan petita la probabilitat que $n_1 = 897$ és molt baixa (el valor esperat per n_1 seria 580, molt per sota de 897). D'altra banda, una $k = 0.41$ és més coherent amb n_1 i s'aproxima més als primers n_i, però amb un paràmetre tan gran la probabilitat de tenir 16 espècies en total és molt baixa (observant el gràfic de la dreta de la figura 3.5 veiem que el nombre esperat d'espècies és $S = 12$ per $k = \hat{k}_5$. Això ens porta a sospitar que l'exemple dels escarabats no s'adequa completament al model geomètric (tot i que acceptem com a vàlid el model).

Observem, ara, el cas de les plantes: les diferències són més petites. Hi segueix havent diferència entre l'estimador de màxima versemblança i la resta, però en general són tots més semblants entre ells. Aquest segon exemple s'adequa millor a la sèrie geomètrica que el primer. Observant les gràfiques rang/abundància dels 2 casos podem apreciar mínimament aquest fet.

Per últim, en el cas dels ocells, tenim que $\hat{k}_5$ té un valor semblant a la resta. Això és senyal que en general les dades segueixen més rigorosament el model geomètric (tot i que era d'esperar, ja que les dades han estat simulades amb R i pertorbades mínimament). Efectivament, la representació de les dades d'aquest exemple en rang/abundància i escala logarítmica mostren més clarament una recta que els altres 2 exemples.

El que també ens crida l'atenció és que en aquest exemple l'estimador $\hat{k}_1^1$ difereix molt de la resta. L'estimació de k per regressió lineal agafant el terme independent pot variar molt, ja que aquest terme és molt inestable; un mínim canvi de pendent produeix una forta translació en ell. Això ens porta a sospitar que és possible que aquest estimador no sigui gaire fiable. Això ho comprovarem posteriorment en el capítol de simulacions.

Cal no oblidar que els valors de $\hat{k}_i$ no deixen de ser valors possibles que podria prendre el nostre paràmetre real. Però realment no hi ha manera de trobar el paràmetre exacte.

Una altra manera de donar informació sobre el paràmetre que es busca és crear un interval de confiança per aquest mateix, de manera que ja no tindrem un resultat puntual, sinó un interval de valors possibles. En el següent apartat es parla de com trobar aquests intervals per k en el cas d'alguns dels mètodes descrits anteriorment.

3.3 Intervals de confiança

Amb els mètodes descrits anteriorment, hem obtingut una aproximació del paràmetre k de la sèrie geomètrica. L'objectiu d'aquest apartat és de donar un interval de confiança per aquest paràmetre. Aquest el podem calcular quan disposem de la informació sobre quina distribució segueix el nostre estimador. A continuació s'explica com trobar un interval de confiança per k estimada pels mètodes de regressió lineal i màxima versemblança.

3.3.1 Interval de confiança per regressió lineal

Recordem que la regressió lineal ens proporciona dos coeficients β_0 i β_1 que aproximen a través d'una recta els logarítmes de n_i:

$$ln(n_i) = ln\left(\frac{Nk}{1-k-(1-k)^{S+1}}\right) + i \cdot ln(1-k) = \beta_0 + i\beta_1$$

Aquests coeficients s'obtenen a l'hora de fer la regressió lineal amb R (taula 3.4). Seguidament, igualant els termes $\hat{\beta}_0$ i $\hat{\beta}_1$ obtinguts amb les seves definicions en funció de k obtenim les estimacions del nostre paràmetre. Ara el que volem és obtenir un interval de confiança per β_0 i β_1 de nivell $1-\alpha$.

Centrem-nos primer en β_1, el terme dependent. Se sap que $\hat{\beta}_1$ segueix una distribució normal amb:

$$E[\hat{\beta}_1] = \beta_1 \quad i \quad Var[\hat{\beta}_1] = \frac{\sigma^2}{(N-1)s_X^2}$$

Per una variància desconeguda, tenim que:

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{s_R^2}{(N-1)s_X^2}}} \sim t_{N-2}$$

Per tant, l'interval de confiança per β_1 de nivell $1-\alpha$ és:

$$\hat{\beta}_1 \pm t_{N-2,\alpha/2}\sqrt{\frac{s_R^2}{(N-1)s_X^2}}$$

Aquesta arrel és precisament l'error estàndard de $\hat{beta}_1$ que apareix a la taula de regressió. Així doncs, l'expressió de l'interval de confiança per β_1 ve donada per:

$$\hat{\beta}_1 \pm t_{N-2,\alpha/2} \cdot s_{\hat{b}_1}$$

En l'exemple dels escarabats, si volem calcular l'interval de confiança del 90% de β_1, obtenim el següent:

$$IC_{90\%}(\beta_1) = \hat{\beta}_1 \pm t_{14,0.05} \cdot s_{\hat{b}_1} = -0.39311 \pm 1.761 \cdot 0.01965 =$$
$$= (-0.428, -0.358)$$

Utilitzant l'expressió de β_1, obtenim aleshores el següent interval de confiança per k:

$$IC_{90\%}(k) = (1 - e^{-0.358}, 1 - e^{-0.428}) = (0.301, 0.348)$$

Similarment, podríem trobar també un interval de confiança per β_0. El nostre estimador $\hat{\beta}_0$ segueix també una distribució normal amb els següents paràmetres:

$$E[\hat{\beta}_0] = \beta_0 \quad i \quad Var[\hat{\beta}_0] = \frac{\sigma^2}{N} \frac{\sum {x_i}^2}{(N-1)s_X^2}$$

Com que novament no coneixem σ, tenim que:

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{s_R^2}{N} \frac{\sum {x_i}^2}{(N-1)s_X^2}}} \sim t_{N-2}$$

Per tant, el nostre interval de confiança per β_0 ve donat per:

$$\hat{\beta}_0 \pm t_{N-2,a/2} \sqrt{\frac{s_R^2}{N} \frac{\sum {x_i}^2}{(N-1)s_X^2}}$$

Aquesta arrel és precisament l'error estàndard del terme independent $s_{\hat{\beta}_0}$ que tenim a la taula de regressió lineal. L'interval final és, doncs:

$$\hat{\beta}_0 \pm t_{N-2,a/2} \cdot s_{\hat{\beta}_0}$$

Calculem l'interval de confiança del 90% de β_0 en l'exemple dels escarabats:

$$IC_{90\%}(\beta_0) = \hat{\beta}_0 \pm t_{N-2,a/2} \cdot s_{\hat{\beta}_0} = 6.47031 \pm 1.761 \cdot 0.1891 =$$
$$= (6.137, 6.803)$$

Per últim, només ens cal igualar l'expressió de β_0 dependent de k amb aquests dos valors per obtenir l'interval de confiança de k. Així doncs, l'interval és:

$$IC_{90\%}(k) = (0.210, 0.340)$$

Podem observar com l'interval de confiança per k extret d'β_0 és considerablement més ampli que l'interval extret de β_1 en ambdós exemples. Això és degut al fet que el terme independent és més susceptible a petits canvis que el terme dependent, així que és més inestable. A la pràctica, sempre s'estima k a través

de l'expressió de β_1, i per tant si haguéssim de quedar-nos amb un interval, ens quedaríem amb l'interval generat a partir de β_1, que és més limitat (ens interessa trobar intervals de confiança estrets, ja que un interval ampli no té sentit, no ens limita les possibilitats per k). Així doncs, ignorarem l'estimació de k a través del terme independent.

3.3.2 Interval de confiança per màxima versemblança

Recordem que amb el mètode de la màxima versemblança estimàvem el paràmetre k d'aquesta manera:

$$\hat{k} = \frac{1}{\bar{x}} = \frac{n}{\sum_{i=1}^{S} x_i}$$

L'objectiu ara és crear un interval de confiança per a k. Per fer-ho, ens fixem en l'estimador $\bar{x}$. Recordem que el teorema central del límit ens diu que per un conjunt de variables x_i independents i idènticament distribuïdes, amb esperança μ i variància $\sigma^2 \neq 0$, per N suficientment gran es compleix que:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Segueix una distribució normal amb esperança μ i variància $\frac{\sigma^2}{N}$. Així doncs, en el nostre cas, tenim que $\bar{x}$ segueix una distribució normal d'esperança $1/k$ (esperança de la sèrie geomètrica de paràmetre k) i variància $\frac{1-k}{Nk^2}$ (variància d'una distribució geomètrica de paràmetre k dividit per N). Aleshores, estandarditzant l'estimador:

$$\frac{\bar{x} - 1/k}{\sqrt{\frac{1-k}{Nk^2}}} \sim N(0,1)$$

Aquí se'ns presenta un problema: no podem aïllar k. El que podem fer, però, és aproximar la variància com a $(1-\hat{k})/(N\hat{k}^2)$. D'aquesta manera, l'interval de confiança de nivell $1-\alpha$ de $1/k$ ve donat per:

$$\bar{x} \pm Z_{\alpha/2} \sqrt{\frac{1-\hat{k}}{N\hat{k}^2}}$$

Calculem l'interval de confiança del 90% del paràmetre k en l'exemple dels escarabats:

$$IC(\frac{1}{k}, 90\%) = \bar{x} \pm Z_{\alpha/2} \sqrt{\frac{1-\hat{k}}{N\hat{k}^2}} = 2.452 \pm 1.65 \sqrt{\frac{3.561}{1745}} = (2.377, 2.527)$$

I fent el canvi $\hat{k} = \frac{1}{\bar{x}}$ obtenim que:

$$IC_{90\%}(k) = (0.396, 0.421)$$

3.3.3 Resultats

A continuació tenim una taula amb els intervals de confiança del 90% de $\hat{k}_1^1$, $\hat{k}_1^2$ i $\hat{k}_5$, pels 3 exemples:

Estimador	Exemple 1	Exemple 2	Exemple 3
$\hat{k}_1^1$	(0.210,0.340)	(0.104,0.136)	(0.272,0.337)
$\hat{k}_1^2$	(0.301,0.348)	(0.128,0.142)	(0.274,0.292)
$\hat{k}_5$	(0.396,0.421)	(0.159, 0.175)	(0.250,0.259)

Taula 3.7: Intervals de confiança del 90% pels diferents estimadors i exemples.

En un principi ens xoca veure com, en els 3 casos, els intervals de confiança per $\hat{k}_1^1$ i $\hat{k}_1^2$ no tenen intersecció amb l'interval de $\hat{k}_5$. L'explicació d'aquest fet és la que s'explicava anteriorment en l'apartat de resultats dels diferents mètodes. L'estimador $\hat{k}_5$ busca aproximar al màxim les dades reals amb les del model que proporciona, mentre que $\hat{k}_1^2$ pretén que es compleixin les proporcions observades amb les del model proporcionat. Així doncs, des del punt de vista de $\hat{k}_5$, l'estimació de $\hat{k}_1^2$ és impensable, i viceversa.

Depenent del nostre interès a estimar les dades, ens decantarem per un model o altre, i, conseqüent, per un interval o altre.

D'altra banda, s'observa com l'interval de confiança generat per $\hat{k}_1^1$ és més ampli que el de $\hat{k}_1^2$ (com ja s'havia comentat anteriorment). D'aquesta manera, si haguéssim de quedar-nos amb un dels 2 intervals, ens quedaríem amb $\hat{k}_1^2$, ja que ens interessa més tenir un interval el mínim d'ample possible.

En els casos d'estimació de k per mínims quadrats, mètode de Newton i mínim i màxim, com que no tenim la variància d'aquests estimadors, no ens ha estat possible desenvolupar un interval de confiança.

Capítol 4

Simulacions

En aquest capítol volem recrear dades mitjançant simulacions amb R amb l'objectiu de comparar els diferents mètodes descrits, aprofitant que aquesta vegada coneixem el paràmetre real de la sèrie geomètrica. D'aquesta manera, podem calcular el biaix de cada un dels estimadors, així com la seva variància i error quadràtic mig.

Per recrear les dades de la sèrie geomètrica, farem la següent construcció: donat un nombre d'individus N, un nombre d'espècies S, un nombre de trampes M, i el paràmetre de la sèrie geomètrica teòric k, el que volem és recrear les dades que obtindríem si haguéssim posat M trampes en llocs diferents per posteriorment veure quins individus hi han caigut per poder fer un recompte d'abundàncies (en cas que les poblacions d'espècies fossin animals, en el cas de plantes M es pot referir al nombre de zones escrutades). D'aquesta manera, obtindríem una taula com la següent:

	Espècie 1	Espècie 2	Exemple 3	...	Espècie S
Trampa 1 Trampa 2 Trampa 3 ... Trampa M					
Total	n_1	n_2	n_3		n_S

Taula 4.1: En cada trampa es comptabilitzen les diferents espècies trobades, amb les seves respectives abundàncies. Finalment, se sumen les dades de totes les trampes per obtenir les n_i

Els valors teòrics de les n_i vénen donats per l'equació 3.1 (pg. 21), agafant $N' = N/M$, ja que tenim M trampes i al final volem que la suma total d'individus quan sumem les abundàncies de les trampes sigui N. Però en comptes d'afegir aquests valors (arrodonits a l'enter) a la taula, per crear cert soroll i donar a

les dades un punt aleatori, el que fem és generar per a cada abundància de la taula, una Poisson d'esperança n_i. Així doncs, per a cada trampa no hi tenim les mateixes abundàncies. I sumant al final totes les files acabem obtenint les noves abundàncies.

S'ha de tenir present que pot ser que a vegades, normalment per l'espècie més rara, no s'obtingui cap comptabilització, i per tant la S pot variar de la inicial. El mateix passa amb la N: és més que probable que la suma total d'individus no sigui exactament el valor que havíem triat inicialment, tot i que s'hi aproximarà. En la següent taula hi tenim un exemple. Les dades han estat generades amb R, amb una $N = 1000$, $S = 12$, $M = 10$ i $p = 0.35$:

	E1	E2	E3	E4	E1	E2	E3	E4	E1	E2	E3	E4
T1	30	12	14	8	3	2	3	0	0	0	0	0
T2	34	25	15	7	5	6	3	1	0	1	0	1
T3	19	20	13	9	7	4	2	0	0	0	0	0
T4	42	26	13	13	3	5	2	4	0	1	1	0
T5	32	27	15	10	6	1	2	3	1	0	1	0
T6	36	24	16	9	6	2	2	2	2	1	1	3
T7	36	24	16	7	4	11	5	1	2	1	0	0
T8	40	27	12	7	5	5	5	1	1	1	0	1
T9	31	23	13	11	13	5	0	3	0	0	0	0
T10	45	20	18	5	6	4	1	1	2	0	0	2
Total	345	228	145	86	58	45	25	16	8	5	3	7

Taula 4.2: Exemple de simulació de dades, agafant $N = 1000$,$S = 12$, $M = 10$ i $k = 0.35$.

La nova mostra té $N = 971$, i la S es manté igual. Amb aquestes dades, podem procedir a calcular el paràmetre k utilitzant els mètodes descrits al capítol anterior, obtenint així 6 estimacions diferents de k.

Aquest procés el repetim 10.000 vegades, de manera que obtenim 6 vectors de llargada 10.000, cada un amb les estimacions de k per un mètode diferent. D'aquesta manera podem calcular la mitjana i la variància de cada un dels vectors, i, d'aquesta manera, obtenim el biaix i l'error quadràtic mig de cada estimador. Recordem que aquests 2 últims vénen definits per:

$$Biaix(\hat{k}) = E(\hat{k}) - k, \quad EQM(\hat{k}) = Var(\hat{k}) + Biaix(\hat{k})^2$$

En la següent taula hi tenim anotats els resultats obtinguts en la simulació, agafant $N = 2000$, $S = 16$, $M = 20$ i $k = 0.3$:

Estimador	Mitjana	Variància	Biaix	EQM
$\hat{k}_1^1$	0.3017	5.971e-5	1.722e-3	6.267e-5
$\hat{k}_1^2$	0.3036	1.266e-4	3.601e-3	1.396e-4
$\hat{k}_2$	0.3013	4.131e-5	1.272e-3	4.293e-5
$\hat{k}_3$	0.3023	3.157e-4	2.278e-3	3.209e-4
$\hat{k}_4$	0.3020	2.222e-4	2.022e-3	2.263e-4
$\hat{k}_5$	0.3090	1.187e-5	8.951e-3	9.199e-5

Taula 4.3: Mitjana, variància, biaix i error quadràtic mig dels estimadors per cada un dels mètodes proposats.

Ara, realitzarem una simulació semblant, afegint un canvi: quan obtenim vector de les abundàncies sumant les abundàncies de cada trampa, per afegir-hi un soroll més fort, multipliquem cada un d'aquests termes per una variable aleatòria uniforme amb fites 0.6 i 1.4. D'aquesta manera, les abundàncies fluctuen del 60% al 140%, provocant així que les dades siguin molt més irregulars.

En la següent taula hi tenim anotats els resultats obtinguts aquesta vegada, amb els mateixos paràmetres $N = 2000$, $S = 16$, $M = 20$ i $k = 0.3$:

Estimador	Mitjana	Variància	Biaix	EQM
$\hat{k}_1^1$	0.3044	3.563e-4	4.438e-2	3.760e-4
$\hat{k}_1^2$	0.3051	4.733e-4	5.107e-2	4.994e-4
$\hat{k}_2$	0.3045	3.449e-4	4.503e-2	3.651e-4
$\hat{k}_3$	0.3094	7.939e-4	9.433e-2	8.882e-4
$\hat{k}_4$	0.3087	6.891e-4	8.627e-2	7.635e-4
$\hat{k}_5$	0.3217	3.664e-4	2.167e-2	8.363e-4

Taula 4.4: Ara, realitzarem una simulació semblant, afegint un canvi: quan obtenim vector de les abundàncies sumant les abundàncies de cada trampa, per afegir-hi un soroll més fort, multipliquem cada un d'aquests termes per una variable aleatòria uniforme amb fites 0.6 i 1.4. D'aquesta manera, les abundàncies fluctuen del 60% al 140%, provocant així que les dades siguin molt més irregulars.

Observem com l'estimador amb menys biaix tant en la primera prova com en la segona és l'estimador de mínims quadrats $\hat{k}_2$. A més a més, també és l'estimador que menys variància té, i, conseqüentment, l'estimador amb mínim error quadràtic mig.

Per altra banda, en ambdós casos l'estimador més esbiaixat és l'estimador de màxima versemblança $\hat{k}_5$. Mentre que aquest manté una variància relativament petita en comparació amb la resta en la primera prova, en la segona aquesta s'amplifica considerablement.

Per últim, l'estimador amb variància més gran és en els dos casos l'estimador del mètode de Newton $\hat{k}_3$. A més a més, també és l'estimador amb màxim error quadràtic mig.

Capítol 5

Discussió dels resultats i conclusions

Si analitzem les taules obtingudes pels 2 mètodes en el capítol anterior, hi ha algunes observacions a fer al respecte[1]:

En primer lloc, s'observa que l'estimador de màxima versemblança és en ambdós casos lleugerament superior a la resta (i s'allunya més en el segon cas). Si recordem prèviament els resultats obtinguts en el capítol 3, ens trobàvem amb què l'estimador de màxima versemblança $\hat{k}_5$ ens proporcionava un valor del paràmetre que distava considerablement de la resta en els casos on les dades no seguien rigorosament la sèrie geomètrica. Les dues taules obtingudes reflecteixen aquest fet: en el primer cas, en el qual les dades segueixen força rigorosament el model geomètric, l'estimació de $\hat{k}_5$ és molt més bona que en el segon cas, en el qual les dades han estat pertorbades i on el biaix de $\hat{k}_5$ és de més de 0.02.

Així doncs, com que a la pràctica les dades no són perfetes ni ideals, sinó que més aviat al contrari, no tenim cap garantia que $\hat{k}_5$ ens proporcioni una bona estimació de k.

En segon lloc, observem el següent fet: com és de suposar, l'error quadràtic mig és en tots els casos més petit quan no alterem el vector d'abundàncies obtingut de les trampes. El soroll que produïm en la segona prova afecta directament a la variància dels estimadors, que es veu augmentada, fet que fa augmentar també l'error quadràtic mig dels estimadors. Tot i així, les estimacions de k són (en tots els casos excepte la màxima versemblança) molt bones.

Així doncs, descartant $\hat{k}_5$, observem com l'ordre d'estimadors en funció del seu error quadràtic mig (de menor a major) és el mateix en les dues simulacions:

[1] Les dues simulacions han utilitzat els mateixos paràmetres N, S, M i k. Tot i així, les simulacions executades canviant els paràmetres ens mostren que, si bé els ordres de magnitud poden canviar, la tendència entre els estimadors és la mateixa.

1. Mínims quadrats ($\hat{k}_2$)
2. Regressió lineal clàssica ($\hat{k}_1^1$)
3. Regressió lineal modificada ($\hat{k}_1^2$)
4. Mínim i màxim ($\hat{k}_4$)
5. Newton ($\hat{k}_3$)

En general aquests 5 estimadors ens proporcionen una bona aproximació del paràmetre k, però el mètode dels mínims quadrats i la regressió lineal clàssica sembla que ens ofereixen un estimador més bo que la resta.

Tot i així, aquestes simulacions tenen una limitació: si bé hi hem incorporat un factor aleatori per tal d'aconseguir que les dades siguin les més reals possibles, en el fons no deixen d'estar relativament controlades, de manera que segueixen sent força regulars. Aquest fet el podem visualitzar en la següent figura, on hi tenim comparats dos jocs de dades: un provinent de dades empíriques (exemple 3.1.1), i l'altre simulat amb R. Com es pot apreciar, les dades simulades amb R estan lligades al model teòric en certa manera, mentre que les dades empíriques tenen molta més llibertat.

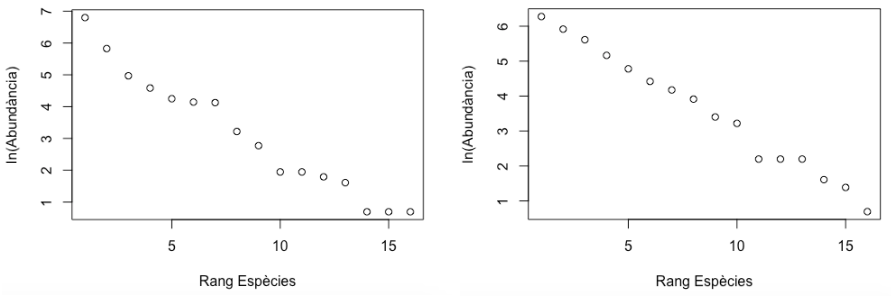

Figura 5.1: A l'esquerra, representació rang/abundància de dades empíriques; a la dreta, representació rang/abundàncie de dades simulades amb R.

Per tant, no tenim garantia absoluta quan les dades no siguin tan perfectes, l'ordre dels estimadors en funció del seu error quadràtic mig es mantingui igual.

Per tal d'afrontar aquest problema, caldria fer un estudi més exhaust amb les simulacions, i intentar recrear unes dades el màxim semblants possibles a les que es tenen a la pràctica. Les simulacions realitzades en el capítol anterior només són un començament. A més a més, caldria també estudiar com varien els resultats de les simulacions en funció del valor del paràmetre k. Malauradament, no s'ha disposat de més temps per trobar un model de simulacions més irregular i que a la vegada respecti el paràmetre original.

Ara bé, seguim considerant bons tots aquests mètodes (havent descartat la màxima versemblança), ja que l'estimació de k que ens proporcionen és coherent i fidel a les dades.

A la pràctica, el mètode més utilitzat clarament és la regressió lineal clàssica. Tot i així, els mètodes de Newton i del mínim i del màxim han estat proposats també com a mètodes per estimar el paràmetre de la sèrie geomètrica. Cal destacar el mètode del mínim i del màxim per la bona aproximació que proporciona a través d'una expressió tancada i simple.

En aquest treball aportem un nou mètode per estimar k, el mètode dels mínims quadrats, com una versió millorada de la regressió lineal clàssica, ja que el mètode dels mínims quadrats depèn d'un únic paràmetre, de manera que tota la informació de la mostra queda retratada en aquest, mentre que la regressió lineal clàssica, que depèn de 2 paràmetres, només té en compte el pendent de la recta de regressió, quan el terme independent també depèn del paràmetre a estimar.

Per a futures investigacions sobre el model geomètric en la biodiversitat d'espècies, es recomana treballar amb la regressió lineal clàssica, tal com s'ha fet fins ara, i en vista dels resultats obtinguts, es proposa fer un estudi del mètode dels mínims quadrats per acabar de veure si l'estimació que pot oferir del paràmetre és tan bona com la que ens ha proporcionat en les simulacions, subjecte a les condicions esmentades.

Bibliografia

[1] Magurran, A. E. (2004), *Measuring Biological Diversity*, Blackwell Publishing.

[2] Fangliang He i Danling Tang (2008), *Estimating the niche preemption parameter of the geometric series*, Elsevier Masson SAS, Acta Oecologica 33, 105-107.

[3] Graffelman, J. (2015), *Species Diversity*, Curs d'Estadística per les biociències, UPC.

[4] Fisher, R. A., Corber, A. S. i Williams, C. B. (1943), *The relation between the number of species and the number of individuals in a random sample of animal population*, Journal of Animal Ecology 12, 42-58.

[5] Preston, F. W. (1948), *The commonness, and rarity, of species*, Ecology 29, 254-283.

[6] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.