TREBALL DE FI DE MÀSTER

Màster en Estadística i Investigació Operativa

Títol: Estudi de l'herbivorisme del *Senecio* mitjançant models zero-inflats

Autora: Anabel Blasco Moreno

Directora: Marta Pérez Casany

Departament: Matemàtica Aplicada II

Convocatòria: 9 de juny de 2014

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Universitat Politècnica de Catalunya
Facultat de Matemàtiques i Estadística

Treball de fi de Màster

Estudi de l'herbivorisme del *Senecio* mitjançant models zero-inflats

Anabel Blasco Moreno

Directora: Marta Pérez Casany
Departament de Matemàtica Aplicada II

A Sergio, als meus pares Valentín i Àngela, a la meva germana Sònia i a Diego.

Los modestos estadísticos han cambiado nuestro mundo - no descubriendo nuevos hechos o desarrollos tecnológicos, sino cambiando nuestra forma de razonar y experimentar, y formando nuestras opiniones acerca de todo ello.

Ian Hacking
Filòsof i historiador de la ciència

Prefaci

L'origen d'aquest projecte es troba en una consulta aparentment innocent que va arribar al Servei d'Estadística Aplicada a la UAB del qual formo part. L'escrit que ve a continuació correspon al post publicat a *freshbiostats.wordpress.com* amb data de novembre de 2012 i relata l'evolució que es produeix a partir de la presa de contacte entre l'investigador amb dades i l'estadístic amb ganes d'anar una mica més enllà.

Appearances can be deceiving

I'm a statistical consultant. While developing my job, I have assessed many applied researchers: from botanists to andrologists, and performed many different statistical analyses: from a simple t-test, to more sophisticated analyses which are resolved through advanced statistical modelling. In order to evaluate the needs of researchers, I find necessary to meet him and let him explain the study goal, show the available data and detail of their statistical doubts. After the meeting, I usually know what kind of analysis is required.

At this point, I think we should not underestimate any study despite of what it may seem at first sight, and I think it is a serious mistake to do it. Let me explain.

As a statistical researcher, I like to work with data that test my analytical abilities while trying to extract its maximum profit. However, not always a high-level analysis is required; sometimes the simplest analysis satisfies researcher needs and expectations. Only sometimes, some seemingly harmless data, conceal a sophisticated statistical analysis that initially had gone unnoticed.

Some months ago, I had a meeting with two biologists. Their study dealt with predation of certain type of plant by some insects in different regions. They tried to use a simple ANOVA test, compare the number of plants affected by predation among regions. But, the test did not give statistically significant results. A statistician realizes quickly what is wrong: "Maybe, you are not taking into account the variability among regions and, of course, you don't have normal data because you are dealing with counts".

Homogeneity of variances and normal distribution are two important hypotheses in the ANOVA test. To solve the problem of non-constant variances, different alternatives are possible, for example using transformations. The most common data transformations are the proposed by Ascombe (1944) and the Box-Cox transformations (1964). These transformations not only solve the problem of non-homogeneity, but they also reduce data anomalies such as non-additivity and non-normality. Transform the data is a good solution but we can go even further. In 1972 John Nelder and Robert Wedderburn formulated the generalized linear model (GLM), a flexible

generalization of the linear regression model allowing for response variables having other than a normal distribution.

Since we are evaluating counts, a GLM using Poisson distribution could be applied. The result remained the same: statistically non significant differences in count predation among regions. We started with ANOVA, then transformed the data obtaining variables with theoretically nice properties, estimated a GLM with Poisson distribution and, at the end, we were at the same point. There was something wrong. In fact, there was a subtile difference among regions: one of which had much more zero counts in contrast to other regions. These zero data could be treated in a more proper way.

The response to this problem appeared in the nineties: zero inflated Poisson models. These models are a way of dealing with overdispersion. The model assumes that the data is a "mixture"of two sorts of individuals: one group whose counts are generated by a standard Poisson regression model, and another group whose individuals have a large frequency of 0. Thus, this approach can take into account the excess in zero counts. Therefore, a zero inflated Poisson model (ZIP) was claimed to solve our problem. Moreover, in this setting, not only a Poisson can be assumed, but a Negative Binomial distribution could also be assumed (ZINB). This led me to further investigation, comparing ZIP and ZINB models with GLM with Poisson and NB distributions by using appropriate tests. The decision of using one or other model not only can be done from a statistical point of view but also using the biological interpretation. In this case, we saw that a ZINB model could model not only the count process for the data predation but also the process for zero predation.
The lesson of this story is that sometimes a simpler study can require a quite sophisticated analysis. Never underestimate the difficulty of a simple experiment because appearances can be often (and very often) deceiving.

By Anabel Blasco
Statistical Consultant
Servei d'Estadística Aplicada de la UAB

Agraïments

En primer lloc, vull agrair a l'Eva Castells i a la Maria Morante que m'hagin permès portar a terme aquest projecte. A l'Eva per confiar-me les seves dades i permetre'm realitzar aquest treball i a la Maria, li vull agrair en especial la seva compressió i dedicació a l'hora d'explicar-me els conceptes biològics i l'ajuda prestada per plasmar-ho en aquest treball.

Vull agrair a la meva directora de projecte, la Dra. Marta Pérez, la seva dedicació a aquest treball així com els suggeriments realitzats i el suport facilitat. Gràcies per trobar un moment en la teva agenda per mi i per la teva amabilitat.

Agrair també a tots els meus companys del Servei d'Estadística Aplicada (Ana, Oliver, Ester, Anna i Llorenç) la revisió realitzada d'aquest document així com a en Sergio per la seva revisió i aportacions al mateix.

Finalment, agrair als meus pares que des de petita van despertar en mi la curiositat per tot allò que m'envolta, curiositat que cada dia fa créixer el meu interès per a aprendre.

Resum

Paraules clau: Excés de zeros; sobredispersió; distribució binomial negativa; models zero inflats; distribució Zipf

MSC2000: Primari: 62J12; Secundari: 60E05

Els estudis ecològics que involucren recomptes d'abundància o de presència-absència sovint produeixen dades amb un excés de zeros. Una forma de tractar amb l'excés de zeros és considerar els models lineals generalitzats zeros inflats. Aquests models assumeixen que l'excés de zeros prové d'un procés amb dues parts que es poden modelar de forma conjunta emprant una mixtura de distribucions. Per una banda, es modela el procés de recompte mitjançant una distribució de Poisson o Binomial Negativa on queden recollits una part dels zeros observats. Per altra banda, es modelen els zeros addicionals a través d'un model logit (Lambert, 1992; Hall, 2000). En aquest treball aquesta metodologia es compara amb un model de regressió amb distribució Zipf (Zipf, 1949). Aquesta distribució de probabilitat exhibeix una elevada probabilitat al primer valor, té una cua pesada i mostra un comportament lineal en l'escala log-log.

Les dades analitzades es van obtenir d'un estudi de camp consistent en analitzar el dany ocasionat pels herbívors en plantes de la familia dels *Senecio*. L'herbivorisme es defineix com el nombre total de capítols menjats pels insectes al llarg de l'estudi. En total es van incloure 475 plantes de quatre espècies diferents de *Senecio*, dues natives i dues exòtiques. Les dades es van recollir en sis localitats del Parc Natural del Montseny a Catalunya entre abril i desembre de 2009. L'objectiu de l'estudi és determinar la influència de l'espècie i de la localització de mostreig en l'herbivorisme. D'altra banda, també és important determinar si les espècies natives i les espècies exòtiques són o no atacades pels insectes de forma similar.

La principal conclusió de l'anàlisi que s'ha realitzat és que s'obté el millor ajust assumint un model de regressió Binomial Negatiu Zero Inflat (ZINB). Tots els models que han estat considerats assenyalen l'existència d'una espècie nativa que pateix un herviborisme significativament més gran que la resta. El model Zipf s'ha realitzat després de trasladar una unitat la variable de recompte que recull l'herbivorisme. Aquest model ha demostrat ser útil per adaptar correctament la probabilitat del zero però no ajusta bé la cua de la distribució. S'arriba a la conclusió que els models zeros inflats funcionen millor que el model de Zipf i, a més, permeten explicar l'existència de dos tipus diferents de zeros que resulta d'interès des del punt de vista dels biòlegs.

Abstract

Keywords: Excess zeros; overdispersion; negative binomial distribution; zero-inflated models; Zipf distribution

MSC2000: 2000Primary 62J12; secondary: 60E05

Ecological studies involving counts of abundance or presence-absence often produce data having an excess of zeros. One method of dealing with the excess zeros is to consider the class of univariate zero-inflated generalized linear models. These models assume that the excess zeros can be generated by a two-steps process that can be modeled using a mixture distributions. On the one hand, modeling the counts from the counting process assuming a Poisson or a Negative Binomial regression model. On the other hand, modeling the extra zeros by means of a logit model (Lambert, 1992; Hall, 2000). In this work this methodology is compared with a regression model with a response variable that follows a Zipf distribution (Zipf, 1949). This probability distribution exhibits a large probability at the first value, has a heavy-tail and shows a linear behavior in the log-log scale.

The data used were obtained from a field study analyzing the damage performed by herbivores on *Senecio* plants. Herbivory damage was defined as the total number of inflorescence eaten by insects along the study. A total of 475 plants from four different *Senecio* species, two native and two exotic, were collected at six different locations in Montseny Natural Park, Catalonia, from April to December 2009. The aim of the study is to determine the influence of *Species* and *Location* in the herbivory damage. Moreover, it is also important to determine if the native and the exotic species are or not affected by the insects in a similar way.

The main conclusion of the analysis that has been performed is that the better fit is obtained assuming a Zero-inflated Negative Binomial regression model. All the models that have been considered point out that there exits a native specie that suffers an herbivory significatively larger than the rest. The Zipf model has been fitted after applying a translation of one unit to the observed data. It has proved to be useful to adapt correctly the large probability at zero, but it did not prove to be appropriate to adapt the heavy-tail. We conclude that zero-inflated models worked better than the Zipf model, and allowed to explain the existence of two different types of zeros that can be of interest for biologists.

Notació i abreujaments

β_i, γ_i	Paràmetres del model de regressió
CV	Coeficient de variació
$E[\cdot]$	Esperança matemàtica
$G(\cdot)$	Funció generatriu de probabilitats
H_0	Hipòtesi nul·la
H_1	Hipòtesi alternativa
$L(\cdot)$	Funció de versenblança
$l(\cdot)$	Logaritme de la funció de versenblança
Min	Mínim
Med	Mediana
Max	Màxim
μ	Experança matemàtica
N	Número d'observacions
$Q1$	Primer quartil
$Q3$	Tercer quartil
S	Desviació estàndard
p	P valor
π	Probabilitat que modela l'excés de zeros
$P(\cdot)$	Funció de probabilitat
Pr	Probabilitat
$V[\cdot]$	Variància matemàtica
χ^2	Estadístic Khi-quadrat
$\bar{Y}$	Mitjana mostral d'Y
$ZeroF$	Zero fals
$ZeroPR$	Zero del procés de recompte

AIC	Criteri d'informació d'Akaike
$ANOVA$	Anàlisi de la Variància
BIC	Criteri d'informació Bayesià
BN	Binomial Negativa
CB	Can Bosc
CP	Can Perepoc
CT	Can Tarrer
$EPPO$	Organització Europea i Mediterrànea per la Protecció Fitosanitària
ERH	Hipòtesi d'alliberament d'enemics naturals
FM	Fogueres de Montsoriu
H	Nivell de predació alt
L	Nivell de predació baix
$M1$ i $M2$	Nivells de predació intermedis
SI	*Senecio Inaequidens*
SL	*Senecio Lividus*
SP	*Senecio Pterophorus*
SS	Santa Susanna
SV	*Senecio Vulgaris*
TC	Nombre Total de Capítols Produïts
TLL	Nombre Total de Llavors Produïdes
TP	Nombre Total de Capítols Predats
VF	Vallforners
ZA	*Zero Altered*
ZI	Zero Inflat
ZIB	Model Binomial Zero Inflat
$ZIBN$	Model Binomial Negatiu Zero Inflat
ZIP	Model Poisson Zero Inflat

Índex general

Capítol 1
Introducció

L'home al llarg del temps ha modificat la distribució natural de molts animals i plantes, tant de forma deliberada com accidental, a partir de les diferents activitats humanes. Amb la seva activitat ha introduït noves espècies, tant animals com vegetals, en d'altres hàbitats. Algunes de les espècies introduïdes són capaces d'establir-se en els nous territoris i proliferar donant lloc a les anomenades *invasions biològiques*.

Les invasions biològiques es produeixen quan espècies que són introduïdes de forma intencionada o accidental en un nou territori, s'estableixen en aquest i es produeix un increment, sovint descontrolat, de les seves poblacions, ocasionant importants perjudicis a les espècie autòctones i als ecosistemes natius. No obstant, el terme espècie introduïda no es sinònim d'espècie invasora. No totes les espècies introduïdes acaben esdevenint invasores. Només ho seran aquelles que s'estableixin de forma estable i acabin esdevenint una amenaça per les espècies autòctones de la seva nova ubicació. Així doncs, les *espècies invasores* són aquells animals, plantes o altres organismes transportats i introduïts en llocs fora de la seva àrea de distribució natural, que han aconseguit establir-se i reproduir-se amb èxit en la nova regió, on poden provocar canvis importants en la composició, l'estructura o els processos dels ecosistemes naturals posant en perill la diversitat biològica nativa (*http://es.wikipedia.org/wiki/Especie_invasora*).

En el cas de les plantes, l'èxit de la invasió d'una nova regió depèn de diferents factors, un dels més rellevants es basa en les relacions (*interaccions* en ecologia) que les plantes introduïdes poden establir amb el nou medi. Els insectes tenen un paper clau en el control de les poblacions vegetals, tant des d'un aspecte positiu (pol·linització, dispersió de llavors) com negatiu (consum de llavors i/o fulles). Per tant, les interaccions que les plantes puguin establir amb els insectes de la regió envaïda podran condicionar l'èxit de la invasió.

Segons la *Hipòtesi d'alliberament d'enemics naturals* (*The Enemy Release Hypothesis* (ERH), Keane i Crawley, 2002) es considera que les espècies exòtiques quan són introduïdes en el nou hàbitat quedaran alliberades dels enemics naturals (insectes entre d'altres) que a la zona d'origen els hi provocaven danys consumint les seves llavors o menjant les seves fulles. Quan es queden alliberades d'aquesta pressió en les noves regions, es preveu que les plantes exòtiques puguin produir més quantitat

de llavors, créixer millor i d'aquesta manera expandir-se i envair el nou territori. No obstant, si a la regió envaïda on s'han establert les espècies exòtiques també hi ha plantes autòctones de característiques similars (mateixa família de plantes, composició química similar, etc) es podria produir un canvi d'hoste per part dels insectes locals cap a les espècies exòtiques. És a dir, que els insectes del lloc d'invasió a més d'atacar les espècies autòctones, passin a consumir també les espècies introduïdes. Per tal de corroborar la ERH es va dissenyar un estudi observacional on es comparaven quatre espècies de plantes de la mateixa família amb dues espècies autòctones i dues espècies exòtiques.

En aquest Treball Final de Màster (TFM) s'analitza el dany ocasionat per insectes a quatre espècies de plantes del gènere *Senecio*: com s'ha comentat abans, dues espècies exòtiques d'origen sud-africà (*S. Inaequidens* i *S. Pterophorus*) i dues espècies autòctones com a control (*S. Vulgaris* i *S. Lividus*) que coexisteixen en sis localitats diferents al Parc Natural del Montseny. D'aquesta manera, avaluant el dany que ocasionen els insectes en les espècies exòtiques i en les autòctones, es podrà discutir si les espècies exòtiques queden alliberades del dany ocasionat per predadors i si això facilita o no la seva expansió i, per tant, invasió de la nova regió.

L'*herbivorisme*, que és el dany causat pels insectes predadors, es va recollir al llarg del temps que cobria tot el període reproductiu de les quatre espècies. L'herbivorisme s'ha mesurat distingint quin era l'insecte causant del dany. En concret, s'han distingit dos tipus d'insectes: *Tephritids* (dipters) i *Pyralids* (lepidòpters). Aquests dos tipus d'insectes predadors s'alimenten de les llavors que estan en formació a l'interior de la flor madura o *capítol*, que és com s'anomena a la flor madura en el cas de les plantes de la família dels *Senecis*. Ara bé, en aquest estudi no es distingeix entre els tipus d'insectes i, en conseqüència, les dades s'han analitzat agrupades.

Aquestes espècies d'insectes fan la posta a l'interior d'aquests capítols i, a mesura que la larva es va desenvolupant, s'alimenta de les llavors encara no dispersades a l'exterior, que haurien de garantir l'èxit reproductiu de la planta. D'aquesta manera, ocasionen un impacte negatiu en la reproducció i en l'expansió de les plantes evitant l'adequada dispersió de les seves llavors.

Així doncs, l'herbivorisme podria constituir una possible manera de frenar les invasions d'espècies vegetals exòtiques, sempre i quan els insectes de la zona envaïda reconeguin les espècies exòtiques com a hostes i, aquestes noves interaccions planta - insecte causin un impacte suficient. En la creació d'aquestes noves interaccions, la presència d'espècies natives genèticament relacionades amb les exòtiques pot afavorir-ne el procés.

Per tal de conèixer si les espècies exòtiques s'alliberen dels enemics (insectes) de la zona envaïda i avaluar quin és el paper dels insectes en aquestes espècies i quins factors podrien influir en les noves interaccions planta - insecte, l'objectiu principal de l'estudi consisteix en explicar com varia l'herbivorisme causat per insectes sobre les quatre especies considerades.

Les dades analitzades en aquest TFM són propietat de les investigadores Eva Castells i Maria Morante de la Unitat de Toxicologia del Departament de Farmacologia, Terapèutica i Toxicologia de la Facultat de Veterinària de la Universitat Autònoma

de Barcelona i en particular, formen part de la tesis doctoral de la Maria. El paràgraf anterior recull la hipòtesis de treball principal d'ambdues investigadores. Al Capítol 2 es presenta de forma més detallada els diferents objectius formulats en aquest TFM així com les hipòtesis de treball.

Per quantificar l'herbivorisme es va recomptar el nombre de capítols danyats (menjats pels insectes) present a cada planta al llarg dels mesos de floració de la mateixa. Al tractar-se d'espècies diferents, el temps de seguiment no va ser el mateix per a les quatre espècies en estudi, existint grans diferències d'una espècie a una altra. En el Capítol 3 s'explica amb més detall la naturalesa de les dades i com s'ha efectuat la recollida de les mateixes. A continuació, en el Capítol 4 es realitza l'anàlisi exploratòria de les variables relatives a l'herbivorisme.

L'anàlisi plantejada inicialment per les dues investigadores del Dept. de Farmacologia va ser una anàlisi de la variància (ANOVA), donat que es volia comparar la mitjana de l'herbivorisme entre les diferents espècies de plantes avaluades. Tot i la robustesa de la prova ANOVA, quan la variable d'estudi és un recompte no té massa sentit assumir normalitat de la variable resposta, d'aquí que l'anàlisi més adient consisteix en fer servir un Model Lineal Generalitzat per a dades de recomptes, com ara un model de Poisson o Binomial Negatiu. Cal esmentar però, que sovint s'utilitzen els models lineals assumint normalitat amb les dades transformades logarítmicament. Ara bé, en O'Hara (2010) es veu que utilitzar el Model Lineal Normal aplicant la transformació logaritmica implica biaixos importants respecte d'aplicar un Model Lineal Generalitzat per a dades de recomptes. D'aquí que en aquest treball s'hagi obviat l'anàlisi mitjançant un Model Lineal Normal amb dades transformades. D'altra banda, la descriptiva de les dades va revelar l'existència d'un nombre molt elevat de recomptes igual a zero. Aquest excés de zeros va donar peu a investigar models alternatius al Poisson o al Binomial Negatiu que poguessin adaptar-se a aquesta situació.

Trobar-se amb dades amb un elevat número de zeros és un fenomen molt habitual en múltiples disciplines com ara l'epidemiologia, l'economia, les assegurances, etc. Des de finals de la dècada dels noranta han aparegut una bona quantitat de metodologies estadístiques desenvolupades per fer front a aquest tipus de dades. El Capítol 5 està dedicat als models zeros inflats (*Zero Inflated models*, ZI) proposats en la literatura. Es comença descrivint el model teòric i a continuació es realitza l'ajust a les dades en estudi. Els resultats de l'ajust d'aquests models es troben en el Capítol 6.

Com es veurà en el Capítol 5, els models amb excés de zeros incorporen l'assumpció que les dades provenen de dos processos diferenciats: el procés que genera l'excés de zeros i el procés pròpiament de recompte. Si no es vol fer aquesta diferenciació, s'ha de trobar una distribució que otorgui un pes important al primer valor i que decaigui ràpidament generant una cua pesada a la dreta. En el Capítol 7, es veu la distribució Zipf que compleix aquests requisits i permet realitzar l'ajust sense necessitat d'assumir l'existència de dos processos diferenciats.

Finalment, en els Capítols 8 i 9 es presenten, respectivament, les conclusions de l'estudi realitzat i les línies futures de treball.

Capítol 2
Objectius de l'estudi

En aquest capítol es detallen els objectius plantejats en aquest TFM. S'enuncia primer la motivació i les hipòtesis de treball per, a continuació, plantejar en base a aquestes, els objectius de l'estudi.

2.1. Motivació i Hipòtesi de treball

Les meva tasca com a assessora estadística del Servei d'Estadística Aplicada de la UAB em permet estar en contacte amb investigadors i investigadores de molts diversos àmbits. Aquest TFM recull la investigació duta a terme arran d'una consulta estadística realitzada per la professora Eva Castells i la seva doctorant Maria Morante de la Unitat de Toxicologia del Departament de Farmacologia, Terapèutica i Toxicologia de la Facultat de Veterinària de la UAB.

Un dels aspectes avaluats en la tesis doctoral de la Maria és l'estudi de l'*herbivorisme*, definit a partir del dany causat pels insectes en diferents espècies de plantes del gènere dels *Senecio*. Bàsicament, es vol analitzar si el nivell d'herbivorisme depèn de l'espècie a la qual pertany la planta. Es van triar dues espècies autòctones i dues d'exòtiques per validar la hipòtesi d'*alliberament d'enemics*, és a dir, les espècies exòtiques queden alliberades dels enemics naturals en el nou hàbitat. Tot i que enunciat el problema d'aquesta forma sembla senzill de resoldre, les dades recollides van presentar certes dificultats d'anàlisi que van propiciar investigar més enllà dels models estadístics habituals. La dificultat principal d'anàlisi d'aquestes dades radicava en l'elevat nombre de zeros que va presentar la variable de mesura de l'herbivorisme. Aquest fet ha comportat l'aplicació de metodologies d'anàlisis alternatives no estàndards per a tractar aquest tipus de dades. Per tal de seleccionar el model més adient, s'han comparat els resultats dels diferents models implementats i s'ha triat aquell que ofereix el millor ajust, tant des del punt de vista estadístic com des del punt de vista biològic.

La motivació des del punt de vista ecològic és de gran importància per a les investigadores donat que, com es descriurà més detalladament en el Capítol 3, les plantes en estudi d'origen exòtic poden esdevenir plantes invasores i la seva expansió pot repercutir negativament en l'equilibri actual de l'ecosistema on es troben.

Així doncs, l'estudi es va dissenyar amb l'ànim de respondre la següent hipòtesis de treball:

Els insectes predadors mostren preferències per les espècies autòctones en detriment de les espècies exòtiques?

Aquesta qüestió té un impacte directe sobre la capacitat reproductiva de la planta i, per tant, sobre la seva supervivència. D'aquí l'elevat interès a donar-li resposta.

2.2. Objectius

En base a la hipòtesi de treball anterior es van plantejar en aquest TFM els següents objectius:

1) Portar a terme una anàlisi descriptiva detallada de les dades obtingudes en funció de l'espècie.
2) Detectar si el nivell d'herbivorisme entre les espècies de *Senecio* considerades és estadísticament diferent. En particular interessa detectar diferències entre les espècies autòctones i les exòtiques.
3) Comparar els models discrets de Poisson i Binomial Negatiu amb les seves corresponents versions zero inflades, per tal de determinar quin d'aquests models és el més escaient per a aquest tipus de dades.
4) Ajustar les dades mitjançant un model Zipf i comparar l'ajust obtingut amb el corresponent als models zero-inflats.
5) Determinar la informació que aporta el model més apropiat, respecte al comportament dels predadors de les plantes de *Senecio*.

Capítol 3
Material i Mètodes

El primer apartat d'aquest capítol està dedicat a explicar les espècies d'estudi que es volen comparar així com els insectes d'interès en l'herbivorisme. En el segon apartat s'explica la recollida de dades que es va portar a terme i finalment, el tercer apartat recull les variables utilitzades en aquest TFM.

3.1. Descripció de les espècies i insectes depredadors d'interès

La descripció de les espècies de plantes estudiades així com dels insectes s'ha realitzat amb les aportacions de la Maria Morante i es pot trobar una descripció més detallada a la seva tesis doctoral: *Paper dels alcaloides de pirrolizidina del gènere Senecio com a reguladors de les interaccions entre plantes i herbivors.*

Com s'ha explicat al Capítol d'Introducció, l'interés de l'estudi radica en comparar l'herbivorisme, és a dir, el dany causat pels insectes depredadors, en quatre espècies diferents del gènere *Senecio*, dues espècies exòtiques d'origen sud africà (*S. Inaequidens* i *S. Pterophorus*) i dues espècies autòctones com a control (*S. Vulgaris* i *S. Lividus*).

El gènere *Senecio* pertany a la família de les *Asteràcies* o *Compostes*. Les plantes d'aquest grup es caracteritzen perquè les seves flors en realitat estan composades per una agrupació de petites flors. Les flors d'aquesta família reben el nom de *flors compostes o capítols*.

En totes les plantes de la família de les compostes els capítols es desenvolupen al llarg del temps passant per les següents fases:

- Capítol en borró: fase inicial de desenvolupament. Es manté tancat com un *botonet* de color verd.
- Capítol en flor: el capítol està a punt per ser pol·linitzat pels insectes perquè la planta pugui reproduir-se. Té pètals que li donen una coloració groguenca.
- Capítol madur: el capítol té les llavors a dins a l'espera d'acabar-se de formar i posteriorment escampar-se amb el vent. Sovint, és en aquesta fase on es troben les larves d'insectes consumint les llavors.

- Capítol en *aqüeni* o dispersat: és quan el capítol no ha estat atacat per cap larva d'insecte i té totes les llavors (aquenis) per ser escampades o bé, només es troben restes perquè les llavors ja han estat escampades. Podem dir que el capítol s'ha reproduït amb èxit.

FIGURA 3.1. Etapes reproductives d'una planta de *Senecio Vulgaris*
a) Capítols en borró b) Capítols en flor c) Capítols madurs.

A la Figura 3.1 es mostren les imatges de les diferents etapes de desenvolupament d'un capítol.

3.1.1. Espècies objecte d'estudi

Les espècies exòtiques *S. Pterophorus* i *S. Inaequidens* com s'ha mencionat anteriorment, són d'origen sud-africà mentre que el *S. Lividus* i el *S. Vulgaris* són espècies autòctones a Catalunya. La Figura 3.2 mostra una imatge de cadascuna de les espècies avaluades que permet fer-se una idea visual de les mateixes.

El *S. Vulgaris* i el *S. Lividus* són espècies anuals, és a dir, el seu cicle de vida és màxim d'un any, però poden durar només uns mesos. Creixen en llocs amb poca vegetació com marges de camins, zones enjardinades, marges de boscos i àrees naturals obertes. Floreixen des del novembre fins al juny, i entre l'abril i el juliol, respectivament.

El *S. Inaequidens* i el *S. Pterophorus* són arbustos llenyosos, perennes i originaris del sud d'Àfrica. El *S. Inaequidens* floreix de maig a desembre i el *S. Pterophorus* de juny al juliol però esporàdicament pot tenir floració en alguns individus a la tardor.

El *S. Inaequidens* es va introduir accidentalment al voltant de 1889 des del Nord d'Europa a partir del comerç amb llana d'ovella. Durant les últimes dècades s'ha estès amb èxit a la majoria de països europeus (veure EPPO Panel[1] sobre espècies exòtiques invasores 2012) representant una gran amenaça per a la biodiversitat. Com a espècie oportunista, el *S. Inaequidens* té la capacitat de colonitzar una àmplia varietat d'hàbitats, incloent zones com ara terrenys erms, vores de camins,

[1] *European and Mediterranean Plant Protection Organization*, EPPO. Organització intergovernamental responsable de la cooperació europea en matèria de salut de les plantes. Fundada el 1951 per 15 països europeus, l'EPPO té ara 50 membres, que abasten gairebé tots els països de la regió europea i mediterrània. Els seus objectius són la protecció de les plantes, el desenvolupament d'estratègies internacionals contra la introducció i propagació de plagues perilloses i promoure mètodes de control efectius i segurs. Pàgina web: http://www.eppo.int/

FIGURA 3.2. Espècies de *Senecio* en estudi. En la part superior esquerra *S. Pterophorus*, a la dreta *S. Inaequidens*. En la part inferior esquerra *S. Vulgaris*, a la dreta *S. Lividus*.

cultius, terra cremada i pastures, on es poden formar grans taques que esgoten la flora local. Va ser citada per primera vegada a la Península Ibèrica el 1984 a partir de les poblacions del sud de França.

El *S. Pterophorus* ha estat introduït més recentment a Europa. Aquesta espècie, però, té una distribució força restringida. Es van trobar algunes poblacions durant el segle XX al Regne Unit prop de les indústries de llana, però en l'actualitat només es troba a Catalunya i a Gènova. A l'Europa continental el *S. Pterophorus* es va detectar per primera vegada durant els anys 80 a prop de Sabadell a la zona tèxtil industrial més important a Espanya durant els segles XIX i XX. El *S. Pterophorus* creix preferentment en els llits dels rius i àrees poc transitades com ara vores de camins i terrenys erms, però també s'ha estès a les comunitats naturals i semi naturals fora de les zones urbanitzades, com les pastures i marges de boscos. S'ha estès amb èxit al sud-oest d'Austràlia on es considera una mala herba nociva. No obstant, a Europa el comportament invasiu del *S. Pterophorus* no ha estat encara demostrat.

L'herbivorisme en realitat està causat no directament per l'insecte adult, sinó per les larves d'aquest que en el seu desenvolupament es mengen les llavors de la planta. La planta té l'objectiu de reproduir-se i l'insecte aprofita el procés reproductiu de la mateixa per garantir aliment a la seva descendència i així, el seu propi èxit reproductiu. En l'estadi inicial, Figura 3.1 a), la planta està produint capítols en

FIGURA 3.3. Capítol madur amb larva d'insecte (*Sphenella marginata*) consumint les seves llavors.

desenvolupament o borrons que esdevindran posteriorment capítols en flor (Figura 3.1 b)). Els insectes identifiquen les flors com a aliment per a la seves larves i hi fan la posta. Quan la flor es panseix, el capítol ja és madur (Figura 3.1 c)) i pot contenir al seu interior una larva de l'insecte alimentant-se de les llavors, truncant així l'èxit reproductiu de la planta però completant amb èxit la reproducció de l'insecte. La Figura 3.3 mostra una larva d'insecte a l'interior d'un capítol madur.

3.1.2. Insectes depredadors d'interès

Els insectes identificats en aquest estudi com a depredadors naturals del *Senecio* pertanyen a dues famílies en concret: els *Tephritids* (dipters) i els *Pyralids* (lepidòpters) La Figura 3.4 mostra la imatge d'un parell d'insectes pertanyents a aquestes dues famílies de depredadors. Concretament, la *Sphenella marginata* pertany a la família dels tefrítids i la *Phycitodes albatella* als piràlids.

FIGURA 3.4. Espècies depredadores d'interès: esquerra *Sphenella marginata*, dreta *Phycitodes albatella*.

Les espècies de *Senecio* estudiades són d'especial interès ja que són plantes tòxiques que solen créixer en zones de pastures, marges de camins, alzinars, ... i poden arribar a ocasionar la mort als animals que se les mengin. Tanmateix, aquests insectes en particular estan adaptats als compostos tòxics (alcaloides) produïts per aquestes espècies essent els depredadors naturals d'aquest tipus de plantes. Per aquest motiu es van triar aquests dos tipus d'insectes per a l'estudi. Aquests insectes doncs, regulen la capacitat per expandir-se de la planta donat que, un cop l'han identificat com a hoste, deixen la seva posta (introdueixen els ous a l'interior del capítol en fase

de borró o flor, (Figura 3.1 a) i 3.1 b)) i impedeixen que els capítols esdevinguin noves llavors. Així doncs, perquè l'insecte pugui interactuar amb la planta, aquesta ha d'estar en el seu període reproductor, és a dir, en floració. Per tant, és interessant conèixer el moment de floració de cada espècie ja que l'herbivorisme dependrà de la quantitat de menjar disponible que trobi en tot moment l'insecte depredador

3.2. Disseny de l'estudi i temporització

L'estudi es va realitzar al Parc Natural del Montseny, situat a 60 kilòmetres al nord-est de Barcelona. Aquest lloc es va triar perquè és un dels pocs espais naturals a Europa, on el *S. Inaequidens* i el *S. Pterophorus* creixen simultàniament en les mateixes comunitats. Es van seleccionar el *S. Vulgaris* i el *S. Lividus* com a espècies autòctones en base a la informació recollida pel Banc de Dades de la Biodiversitat de Catalunya[2] (Font 2012) que recull que ambdues espècies són presents a la zona.

Primerament, es va dur a terme un estudi inicial per determinar les zones on coexistien les quatre varietats de *Senecio* simultàniament. El *S. Inaequidens* abundava en la zona sud occidental del Parc del Montseny, en la seva majoria formant agrupacions denses, encara que es podien trobar plantes més disperses en zones del nord. El *S. Pterophorus* es limitava a la vessant sud i era més abundant a la part oriental del Parc. El *S. Vulgaris* i el *S. Lividus* van ser, en general, molt abundants en totes les regions investigades i no van suposar cap limitació en l'elecció de les àrees de mostreig. Com a resultat de l'exploració preliminar, es van seleccionar 6 àrees circulars de 600 metres de diàmetre distribuïdes per tot el parc que contenien almenys 4 individus de cadascuna de les espècies d'interès amb l'objectiu de seleccionar 20 plantes de cada espècie a cada localització. Es van seleccionar àrees de mostreig no gaire extenses per mirar de garantir que les plantes autòctones i exòtiques estiguessin exposades a la mateixa comunitat d'enemics. Els llocs van ser distribuïts d'oest a est al vessant sud del Parc del Montseny i separats per almenys 2 km. Els hàbitats predominants dins de les àrees eleccionades van ser boscos de *Quercus Ilex*, *Q. Suber* i *Castanea sativa*, marges de boscos i pastures abandonades.

Les 6 zones de mostreig seleccionades van ser:

- Localització Vallforners (VF): el seu nom prové de la vall on s'ubica i de la casa senyorial que hi ha amb el mateix nom.
- Localització Can Bosc (CB): el seu nom prové del nom del petit restaurant que hi ha perdut en aquesta zona de la muntanya.
- Localització Can Tarrer (CT): el seu nom prové de la casa abandonada que hi ha en la zona de mostreig.
- Localització Santa Susanna (SS): el seu nom prové de l'ermita que hi ha a pocs metres de la zona de mostreig.

[2] *El Banc de dades de biodiversitat vol ser una recopilació de tota la informació sobre la biodiversitat de Catalunya. Es desenvolupa mitjançant la informatització de totes les citacions disponibles de les espècies que es fan al territori català. La informatització de citacions es complementa amb altres tipus de dades com són ara: Biologia, distribució, ecologia, etc. Pàgina web: http://biodiver.bio.ub.es/biocat/*

- Localització Can Perepoc (CP): el seu nom prové de la casa de pagès que hi ha prop de la zona de mostreig.
- Localització Fogueres de Montsoriu (FM): el seu nom prové de la urbanització que hi ha a tocar de la zona de mostreig.

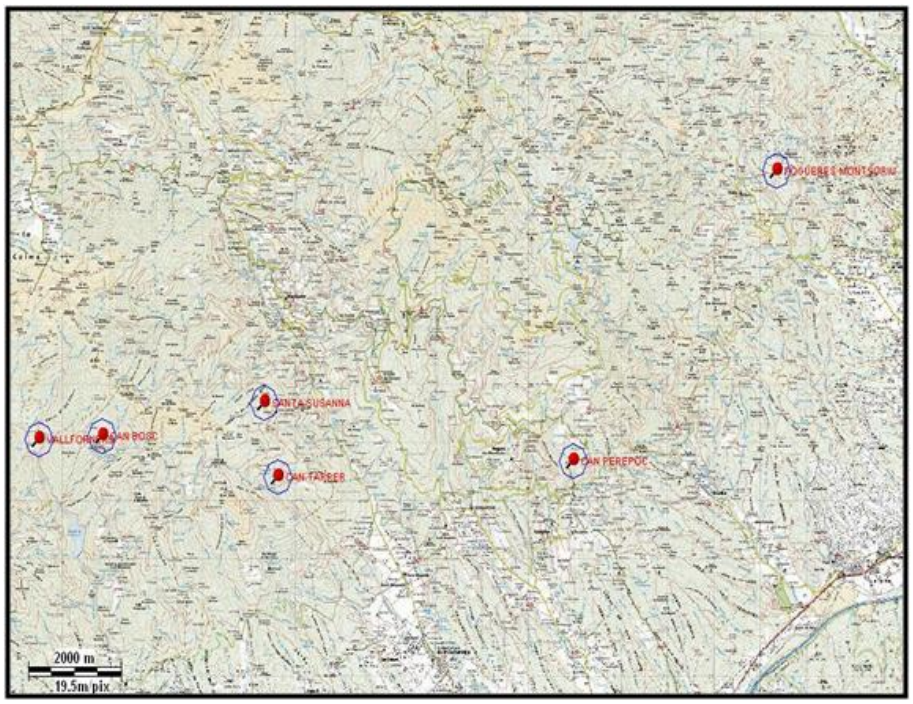


FIGURA 3.5. Mapa geogràfic de la de la serralada del Montseny. Els punts vermells corresponen a les zones de mostreig.

Les localitzacions VF i CB queden ubicades sobre la zona de la població de Cànoves, a prop de Cardedeu. Les localitzacions CT i SS es troben ubicades prop de la població de Sant Esteve de Palautordera, a prop de Sant Celoni. La localització CP es troba a la zona de la població de Campins, ubicada cap a l'est de Sant Celoni. Finalment, la localització FM queda ubicada a la zona de Fogueres de Montsoriu mateix, a prop de la població de Breda. Es pot veure la geolocalització concreta al mapa de la Figura 3.5.

La planta és la unitat experimental en estudi. A cada zona de mostreig es van seleccionar aquelles plantes que estaven sanes i semblava que podrien arribar a reproduir-se amb èxit. Aquestes plantes havíen de presentar capítols en borrons, és a dir, trobar-se en la fase preliminar de la seva etapa reproductiva i per tant era d'esperar que en el futur desenvolupessin flors candidates a ser atacades pels insectes.

Des del mes d'abril fins al maig de 2009, es van seleccionar en total 475 plantes de *Senecio* de les quatre varietats d'estudi. Es va tenir cura d'etiquetar totes les plantes en la seva etapa reproductiva inicial, quan els caps de les flors (borrons)

s'estaven desenvolupant, per tal de cobrir tot el període de floració. Com que l'objectiu era determinar si les espècies exòtiques presentaven nivells més baixos d'herbivorisme per part dels insectes locals, només va interessar el seguiment de les mateixes durant la seva etapa reproductiva. Per tant, el fi del seguiment es produïa quan la planta deixava de fer flors, bé per què acabava morint, bé per què entrava en l'estadi de no floració.

Es va fer un seguiment de les plantes en les 6 zones de mostreig seleccionades cada 10 o 15 dies entre els mesos d'abril i desembre de 2009. En total, es van realitzar entre 22 i 26 visites, depenent de la zona de mostreig.

3.3. Recollida de dades i variables d'estudi

A cada visita, es comptaven el nombre de capítols a la planta i el nombre de capítols dispersats, aquells dels quals només quedava la base del capítol seca a la planta perquè el vent ja havia dispersat totes les seves llavors indicador de que la planta s'havia reproduït amb èxit. Per avaluar el nivell d'herbivorisme, tots els capítols en flor van ser recollits, disseccionats *in situ* o portats al laboratori per dissecció per veure si contenien la larva d'insecte en el seu interior (Figura 3.3). En total es van disseccionar 30.085 capítols en flor.

Cal esmentar però, que el nombre de capítols produïts i l'herbivorisme per algunes plantes de les espècies *S. Pterophorus* i *S. Inaequidens* es va aproximar prenent una submostra per a cada planta i extrapolant el resultat obtingut a la planta sencera. Es va haver de procedir d'aquesta forma perquè l'envergadura de les plantes de les espècies *S. Pterophorus* i *S. Inaequidens*, que pot arribar a més de dos metres, va impossibilitar en ocasions poder mostrejar tota la planta sencera. En aquestes situacions es va mostrejar una branca de la planta escollida a l'atzar i el resultat es va multiplicar pel nombre de branques que presentava la planta. D'aquesta forma s'aproximava el valor real de capítols produïts per aquestes plantes.

Els capítols recollits, com s'ha comentat abans, van ser disseccionats longitudinalment per determinar la presència de qualsevol insecte desenvolupant-se en l'interior del mateix. Es van establir les següents categories per tal de distingir els diferents estats del capítol:

i) Capítols intactes.
ii) Capítols danyats però sense presència del depredador. Aquest tipus de dany que no pot ser assignat a cap espècie d'insecte en particular va ser anomenat *Indeterminat*.
iii) Presència d'un embolcall buit. Això vol dir que hi havia una larva d'insecte però que ja havia arribat a adult i havia pogut marxar amb èxit.
iv) Presència d'una pupa. Això indica que la larva de l'insecte està a punt de fer-se adulta.
v) Presència d'una larva.

Si els capítols estaven intactes, situacó *i)*, es va concloure que no hi havia hagut predació. En qualsevol de les altres situacions, hi havia existit predació per part dels insectes.

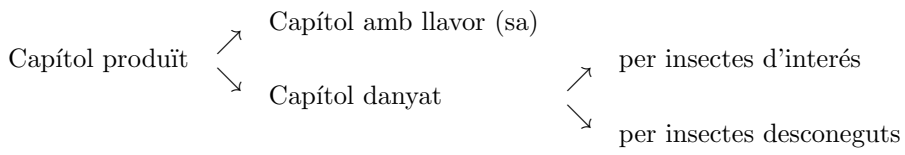
El nombre total de capítols danyats es va calcular tenint en compte tots els tipus de dany (de *ii)* a *v)* enunciats anteriorment), independentment de l'estadi de desenvolupament de l'insecte (pupa, larva, etc). Les pupes i les larves recuperades durant les disseccions van ser identificades per estudiar si es corresponien a les espècies d'insectes d'interès.

Finalment, es va comptabilitzar el nombre total de capítols produïts per cada planta com la suma de capítols amb llavor més capítols danyats.

Els capítols amb llavor eren aquells que:

- presentaven llavors a punt de dispersar-se (Figura 3.1 c)) o,
- presentaven restes de capítols secs sense llavors perquè ja s'han dispersat o,
- capítols que al ser disseccionats al laboratori estaven sans i només contenien llavors (situació *i)*).

L'esquema següent exemplifica les diferents situacions en que es pot trobar el capítol:



Així doncs, es va recollir el nombre de capítols segons fossin

- danyats per tipus d'insecte,
- amb llavor i,
- total (la suma de tots plegats).

A banda, també es va recollir el nombre de borrons i el nombre de flors presents a la planta en cada seguiment. Tota aquesta informació es va guardar i emmagatzemar dins d'una base de dades Excel.

Les variables de la base de dades es poden agrupar segons el paper que desenvolupen en l'estudi. Existeix una única variable resposta principal i dues variables resposta secundàries. Hi ha també diverses variables explicatives. Entre les variables explicatives destaca la variable *Espècie* com a factor principal d'estudi.

Variable resposta principal:

Nombre Total de Capítols Predats (TP): suma, per totes les visites realitzades, del nombre de capítols disseccionats amb evidències d'haver estat predats pels insectes. Es calcula com,

$$TP = \sum_j TP_j,$$

on TP_j correspon al nombre de capítols disseccionats amb evidències d'haver estat predats pels insectes en la visita j-èssima.

Variables resposta secundàries:

Nombre Total de Capítols Predats per Tephritids: suma, per totes les visites realitzades, del nombre de capítols disseccionats dins els quals hi havia un tefrítid (en algun dels seus estadis: larva, pupa, ...).

Nombre Total de Capítols Predats per Pyralids: suma, per totes les visites realitzades, del nombre de capítols disseccionats dins els quals hi havia un piràlid (només el podíem trobar en fase de larva, per què fan la pupa al terra).

Variable explicativa principal:

Espècie: espècie de *Senecio* a la qual pertany la planta. Presenta les categories: SP, SI, SV i SL.

Variables explicatives secundàries:

Localització: lloc de mostreig. Correspon a les sis zones de mostreig esmentades anteriorment: CB, CP, CT, FM, SS i VF.

Variables de control:

Nombre Total de Capítols Produïts (TC): suma, per totes les visites realitzades, del nombre de capítols produïts per la planta. Correspon al nombre de capítols que han sortit de la planta, és a dir, la suma de: 1) els que tenien llavor, dispersada o no, al camp, 2) els que es van disseccionat i estaven predats i 3) els que es van disseccionar i estaven sense predar.

Nombre Total de Llavors Produïdes(TLL): suma, per totes les visites realitzades, del nombre de capítols amb llavor que ha fet la planta. És la suma dels capítols amb llavor comptabilitzats al camp (els que contenia la planta quan es va mostrejar) i al laboratori (capítols diseccionats sospitosos d'estar predats però en realitat tenien llavor).

Data: Dia que es va mostrejar la planta.

Finalment, cal fer una apreciació entre la variable resposta principal, *Nombre Total de Capítols Predats* i la variable *Nombre Total de Capítols Produïts* per la planta. Concretament, es dona la següent relació:

Si TC $= 0$ i TP $= 0 \Longrightarrow$ No hi ha capítols, per tant no hi pot haver predació.

Si TC $\neq 0$ i TP $= 0 \Longrightarrow$ Hi ha capítols però no hi ha predació.

Si TC $\neq 0$ i TP $> 0 \Longrightarrow$ Hi ha capítols i hi ha predació.

Aquest fet fa que tingui sentit que la variable *Nombre Total de Capítols Produïts* (TC) faci el paper de terme *offset* en el model. Per tant, en totes els models que s'ajustaran en aquest treball s'assumirà que el coeficient que acompanya a aqust terme és igual a la unitat.

Capítol 4
Anàlisi Exploratòria

En aquest capítol s'explora la base de dades fent ús de l'estadística descriptiva. És important realitzar una bona anàlisi descriptiva de les dades per detectar errors, dades faltants, dades influents sobre la mostra, observar tendències i extreure'n conclusions preliminars. Els estadístics descriptius emprats en aquesta anàlisi han estat els següents: número d'observacions (N), mitjana mostral ($\overline{Y}$), desviació estàndard (S), coeficient de variació (CV), mediana (Med), primer i tercer quartils ($Q1$ i $Q3$), mínim i màxim ($Min.$ i $Max.$). Com a representació gràfica de les dades s'han triat gràfics de freqüències, diagrames de barres i diagrames de caixa. Tanmateix, es realitza també un estudi de la homogeneïtat de les espècies en les diferents zones de mostreig mitjançant la prova d'homogeneïtat de la χ^2 (Pearson, 1928).

En primer lloc, es mostra una taula resum amb el nombre de plantes recollides per cada espècie en cada localització. Després, es descriu la variable principal de l'estudi, *Nombre Total de Capítols Predats*. Seguidament s'estudia el *Nombre Total de Capítols Produïts* que permetrà estudiar les diferències existents entre les espècies i, a continuació, es defineix i es descriu la variable *Taxa de predació* definida com el quocient entre el nombre de capítols predats i el nombre de capítols produïts. Finalment, es porta a terme la mateixa anàlisi descriptiva però només per aquelles plantes per les quals s'ha observat predació, es a dir, restringint l'anàlisi a la població de plantes que han mostrat algun tipus de predació.

A l'Apèndix A es troba el codi R utilitzat per realitzar aquesta anàlisi exploratòria.

4.1. Distribució de la mostra per espècie i localització

Tot i que, tal i com s'ha explicat en el Capítol 3, inicialment s'havia contemplat recollir 20 plantes per espècie en cadascuna de les 6 localitzacions seleccionades, en total 120 plantes per espècie, l'estudi ha resultat desbalancejat donada la naturalesa del mateix. En els estudis de camp on l'investigador no pot exercir un control directe sobre l'experiment es dóna sovint aquest tipus de desviaments respecte al protocol de selecció i recollida de dades establert inicialment.

Com s'ha esmentat en l'apartat 3.2, es va recollir informació per a un total de 475 plantes de *Senecio*. La Taula 4.1 mostra quantes plantes es van recollir finalment de cada espècie en global i per localització. Es mostren també els percentatges per fila i columna.

	Espècie				
Localització	*S. Inaequidens*	*S. Lividus*	*S. Pterophorus*	*S. Vulgaris*	Total
Can Bosc	21	20	19	16	76
	27,63%	26,32%	25%	21,05%	100%
	21%	12,20%	23,17%	12,40%	16%
Can Perepoc	14	32	11	12	69
	20,29%	46,38%	15,94%	17,39%	100%
	14%	19,51%	13,41%	9,30%	14,53%
Can Tarrer	19	27	22	23	91
	20,88%	29,67%	24,18%	25,27%	100%
	19%	16,46%	26,83%	17,83%	19,15%
Fogueres	15	29	9	20	73
de Montsoriu	20,55%	39,73%	12,33%	27,4%	100%
	15%	17,68%	10,98%	15,50%	15,37%
Santa Susanna	4	29	11	29	73
	5,48%	39,73%	15,07%	39,73%	100%
	4%	17,68%	13,41%	22,48%	15,37%
Vallforners	27	27	10	29	93
	29,03%	29,03%	10,75%	31,18%	100%
	27%	16,46%	12,20%	22,48%	19,58%
Total	100	164	82	129	475
	21,05%	34,53%	17,26%	27,16%	
	100%	100%	100%	100%	

TAULA 4.1. Taula resum del nombre d'observacions, percentatge fila i percentatge columna per espècie i localització de mostreig

L'espècie més representada correspon al *S. Lividus* (34,53% de la mostra), una de les espècies autòctones i, la menys representada correspon al *S. Pterophorus* (17,26% de la mostra), una de les espècies exòtiques. Respecte a les localitzacions de mostreig, allà on es va recollir major mostra va ser a *Vallforners* i *Can Tarrer* i, on menys mostra hi ha és a *Can Perepoc*. El nombre de plantes recollides en cada localització no ha estat homogeni, havent-hi moltes plantes de *S. Lividus* a *Can Perepoc*, poques plantes de *S. Pterophorus* a *Fogueres*, *Santa Susana* i *Vallforners* i només 4 plantes de *S. Inaequidens* a *Santa Susana*. La prova d'homogeneïtat de la χ^2 confirma que, efectivament, la distribució de la mostra no ha estat homogenia per localització i espècie (χ^2=36,81, $p = 0{,}0013$).

4.2. Descriptiva variable *Nombre Total de Capítols Predats* (TP)

4.2.1. Descriptiva sense covariables

Com s'ha esmentat a la introducció, en ecologia és habitual trobar-se amb dades que presenten un excés de zeros. Un senzill gràfic de freqüències serveix per posar de manifest l'excés de zeros de la variable TP com s'il·lustra en la Figura 4.1. Aquesta figura mostra la freqüència del nombre total de capítols predats pels insectes. Concretament, de les 475 plantes analitzades, 261 no teniem cap capítol predat, 34 plantes només tenien un capítol predat i així successivament. És a dir, més de la meitat de les observacions, el 54,95%, prenen el valor zero i hi ha poques observacions amb un nombre molt elevat de capítols predats.

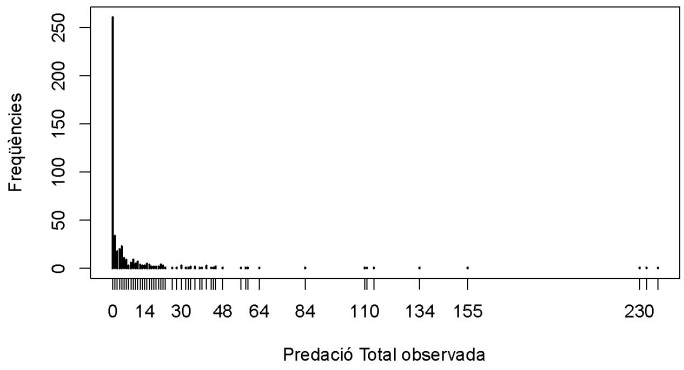


FIGURA 4.1. Diagrama de freqüències per a la variable *Nombre Total de Capítols Predats*. Hi ha 261 zeros, 34 uns, 18 dosos, 20 tresos, 24 quatres, etc.

El patró de les dades observat a la Figura 4.1 es pot correspondre, per una banda, a la distribució d'una variable de recompte amb un excés de zeros però també, a la distribució d'una variable discreta amb una probabilitat molt elevada en el primer valor i amb una cua asimètrica per la dreta que decau lentament. Aquest patró coincideix amb una distribució de Pareto discreta, també coneguda com distribució *Zipfian*. En el Capítol 7 s'explicará aquesta distribució amb més detall i s'implementará el model corresponen en base a utilitzar la distribució *Zipfian*.

4.2.2. Descriptiva amb covariables i prova d'homogeneïtat

L'objectiu principal de l'estudi és determinar si existeixen diferències a nivell de predació atribuïbles al factor *Espècie* al qual pertany la planta. Resulta interessant doncs, desagregar les dades anteriors segons aquest factor d'interès. El resultat es mostra a la Figura 4.2.

Com s'observa a la Figura 4.2, l'elevat recompte del zero es dóna en totes les espècies però principalment en les espècies *S. Pterophorus* i *S.Vulgaris*. Important observar que d'aquestes dues espècies, el *S. Pterophorus* es tracta d'una espècie exòtica

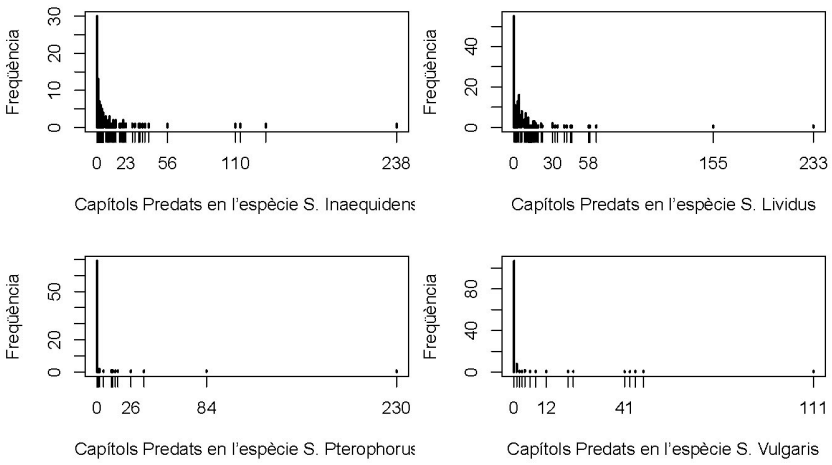


FIGURA 4.2. Diagrama de freqüències per a la variable *Nombre Total de Capítols Predats* segons *Espècie*. Hi ha 30 zeros per l'espècie SI, 55 per l'espècie SL, 69 per SP i 107 per SV.

mentre que el *S. Vulgaris* es considerada una espècie nativa o autòctona. S'observen 30 zeros per l'espècie *S. Inaequidens* que correspon al 30% de les observacions, 55 zeros per a l'espècie *S. Lividus* que correspon al 33% del total d'observacions, 69 zeros corresponen a l'espècie *S. Pterophorus* que representen més del 84% de les observacions i, finalment s'observen 107 zeros en les plantes de *S. Vulgaris* que representen pràcticament el 83% de les observacions. Així doncs, el factor *Espècie* jugarà un paper rellevant almenys dins de la detecció de l'excés de zeros. D'altra banda a les quatre figures es visualitza la forma característica de la distribució *Zipfian* que correspon al dibuix d'una L.

Com s'ha comentat anteriorment, les dades es van recollir en sis localitzacions diferents que, a priori, es consideraven homogènies. Per comprovar aquesta idea d'homogeneïtat entre localitzacions, la Taula 4.2 recull la suma del nombre total de capitols predats creuant les covariables *Localització* i *Espècie*. També recull els percentatges fila i columna.

La Taula 4.2 permet visualitzar clarament les grans diferències existents entre les espècies avaluades. Les espècies *S. Inaequidens* i *S. Lividus* han experimentat una predació gairebé tres vegades superior a la de les altres dues espècies. Cal remarcar, que l'especie *S. Lividus* és nativa mentre que l'espècie *S. Inaequidens* és exòtica. D'altra banda, es detecta que en funció de la localització l'espècie més predada és una o una altra. A la localització de *Vallforners* és on es va detectar la predació més elevada d'entre totes les localitzacions avaluades per tres de les quatre espècies en estudi. Les espècies més atacades en aquesta localització van resultar la *S. Inaequidens* i la *S. Lividus*. Es pot hipotitzar amb el fet que en aquesta localització en concret es donava l'hàbitat més adequat pels insectes depredadors. No obstant, la informació relativa a la quantitat d'insectes presents en l'ambient no va ser recollida i per tant no es pot corroborar la hipòtesis plantejada. De la resta de localitzacions

Localització	Espècie				Total
	S. Inaequidens	*S. Lividus*	*S. Pterophorus*	*S. Vulgaris*	
Can Bosc	172	65	65	3	305
	56,39%	21,31%	21,31%	0,98%	100%
	13,32%	4,31%	14,77%	0,80%	8,43%
Can Perepoc	5	140	26	6	177
	2,82%	79,10%	14,69%	3,39%	100%
	0,39%	9,28%	5,91%	1,59%	4,89%
Can Tarrer	94	79	91	3	267
	35,21%	29,59%	34,08%	1,12%	100%
	7,28%	5,24%	20,68%	0,80%	7,38%
Fogueres	55	135	242	1	433
de Montsoriu	12,70%	31,18%	55,89%	0,23%	100%
	4,26%	8,95%	55,00%	0,27%	11,97%
Santa Susanna	7	190	0	96	293
	2,39%	64,85%	0,00%	32,76%	100%
	0,54%	12,59%	0,00%	25,46%	8,10%
Vallforners	958	900	16	268	2.142
	44,72%	42,02%	0,75%	12,51%	100%
	74,21%	59,64%	3,64%	71,09%	59,22%
Total	1.291	1.509	440	377	3.617
	35,69%	41,72%	12,16%	10,42%	
	100%	100%	100%	100%	

TAULA 4.2. Suma del nombre total de capítols predats de cada espècie per localització. Percentatge fila i percentatge columna.

destaca el nivell molt elevat de predació en l'espècie *S. Pterophorus* detectat a la localització de *Fogueres de Montsoriu* i la predació del *S. Inaequidens* a *Can Bosc*.

Per avaluar l'homogeneïtat de la predació en les diferents localitzacions, es porta a terme la prova de la Khi-quadrat. El resultat de la mateixa indica que la suma del nombre de capítols predats en cada espècie no és independent de la localització (χ^2=1.764,71, $p < 0{,}001$).

La Taula 4.3 recull els estadístics descriptius de la variable TP en funció de l'espècie. Com es pot observar, l'espècie de *Senecio* amb major nombre de capítols predats, en mitjana, correspon a l'espècie *S. Inaequidens*, seguida del *S. Lividus*, el *S. Pterophorus* i el *S. Vulgaris*. No obstant, la variabilitat present a les dades és molt elevada com es desprén del valor observat per la desviació estàndard. L'elevat nombre de zeros detectat en la Figura 4.2 té com a conseqüència que la mediana de les espècies *Pterophorus* i *Vulgaris* sigui igual a 0.

Tenir moltes observacions iguals a 0 comporta que la mitjana sigui pròxima a 0 però, presentar una cua pesada indica a la seva vegada l'existència de valors molt extrems. Tot això es recull en el coeficient de variació que pren valors molt elevats. Donat que aquestes dades són recomptes i que a més presenten molta dispersió, aleshores té sentit pensar que, efectivament, la mitjana no serà l'estadístic que millor les representi.

Espècie	*S. Inaequidens*	*S. Lividus*	*S. Pterophorus*	*S. Vulgaris*	Total
N	100	164	82	129	475
$\overline{Y}$	12,91	9,20	5,37	2,92	7,61
S	31,73	23,96	27,29	12,58	24,32
CV	245,80%	260,40%	508,56%	430,48%	319,36%
$Med.$	2,5	3	0	0	0
$Q1$	0	0	0	0	0
$Q3$	11,25	9	0	0	4,5
$Min.$	0	0	0	0	0
$Max.$	238	233	230	111	238

TAULA 4.3. Taula d'estadístics descriptius per a la variable *Nombre Total de Capítols Predats* per espècie i globalment

La Figura 4.3 mostra el diagrama de caixa per a la variable TP. Es veu que existeixen força dades anòmales en totes les espècies estudiades. Aquestes dades anòmales corresponen a les plantes que presenten un elevadíssim nombre de capítols predats. També es detecta que la predació més baixa es dóna a les espècies de *S. Pterophorus* (amb excepció d'una planta) i *S. Vulgaris*. Un altre fet que es pot apreciar és la assimetria de les distribucions. La banda de la mediana està desplaçada cap a la part inferior de la caixa mentre que la cua és més extensa en la part superior de la caixa com a conseqüència clara de l'elevada probabilitat del zero. Finalment, aquest gràfic és un indicador clar de l'elevada dispersió existent en les dades.

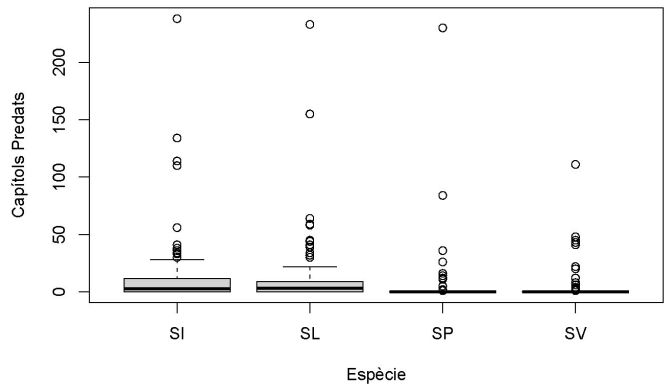


FIGURA 4.3. Diagrama de caixa de la variable *Nombre total de Capítols Predats*.

Donada la falta d'homogeneïtat de la predació de cada espècie en les localitzacions, a continuació la Taula 4.4 recull els mateixos estadístics descriptius de la Taula 4.3 desglossats per localització.

Espècie	***Localització***	*Can Bosc*	*Can Perepoc*	*Can Tarrer*	*Fogueres de Montsoriu*	*Santa Susanna*	*Vallforners*
S. Inaequidens	N	21	14	19	15	4	27
	$\overline{Y}$	8,19	0,36	4,95	3,67	1,75	35,48
	S	9,34	0,74	12,92	2,94	1,26	53,85
	CV	114,05%	208,58%	261,16%	80,29%	71,90%	151,77%
	$Med.$	5	0	1	3	2	15
	$Q1-Q3$	0-13	0-0	0-2,5	1,5-5	1,5-2,25	5,5-35
	$Min-Max$	0-30	0-2	0-56	0-10	0-3	0-238
S. Lividus		20	32	27	29	29	27
		3,25	4,38	2,93	4,66	6,55	33,33
		3,73	4,36	3,77	6,42	6,82	52,20
		114,64%	99,67%	128,90%	137,82%	104,05%	156,61%
		3	3,5	1	1	4	16
		0-4,25	1-5,25	0-4,5	0-6	2-8	0-42,5
		0-13	0-16	0-15	0-21	0-22	0-233
S. Pterophorus		19	11	22	9	11	10
		3,42	2,36	4,14	26,89	0,00	1,60
		8,81	7,84	17,87	76,27	0,00	5,06
		257,67%	331,66%	432,04%	283,65%	–	316,23%
		0	0	0	0	0	0
		0-1	0-0	0-0	0-0	0-0	0-0
		0-36	0-26	0-84	0-230	0-0	0-16
S. Vulgaris		16	12	23	20	29	29
		0,19	0,50	0,13	0,05	3,31	9,24
		0,54	1,17	0,34	0,22	11,37	23,08
		290,08%	233,55%	264,00%	447,21%	343,49%	249,78%
		0	0	0	0	0	0
		0-0	0-0,25	0-0	0-0	0-0	0-4
		0-2	0-4	0-1	0-1	0-45	0-111

TAULA 4.4. Taula d'estadístics descriptius per a la variable *Nombre Total de Capítols Predats* per espècie i localització.

En general, l'espècie *S. Inaequidens* està poc predada en totes les localitzacions amb excepció de a *Vallforners*, tant en termes de mitjana com de mediana. Per l'espècie *S. Lividus* la predació més elevada es detecta també a *Vallforners*. Per l'espècie *S. Pterophorus* aquesta es dóna a *Fogueres de Montsoriu* i, finalment per l'espècie *S. Vulgaris* la predació més elevada també s'observa a *Vallforners*. Així doncs, la localització de *Vallforners* presenta les condicions idònies per a la predació del Senecis amb excepció de l'espècie *S. Pterophorus*. Aquest por guardar relació amb les colònies d'insectes presents a la zona. Els insectes de la família dels piràlids tenen preferència per l'espècie *S. Pterophorus* mentre que l'altra família d'insectes, el tefrítids, no fan la posta en aquesta espècie de Seneci.

Finalment, en el Capítol 3 s'ha fet notar que les espècies a comparar són molt diferents entre elles. Aquí, el nombre total de capítols que produeix cada espècie pot estar influenciant aquests valors de predació, pot haver més predació pel simple fet d'haver més capítols. Per tant, es descriu a continuació el nombre total de capítols realitzats per les plates per detectar aquestes diferències.

4.3. Descriptiva variable *Nombre Total de Capítols produïts* (TC)

4.3.1. Descriptiva sense covariables

La distribució de la variable TC es veu a la Figura 4.4. En aquest gràfic s'ha truncat fins al valor 300 per poder apreciar millor aquesta distribució. S'han exclòs en total 45 observacions.

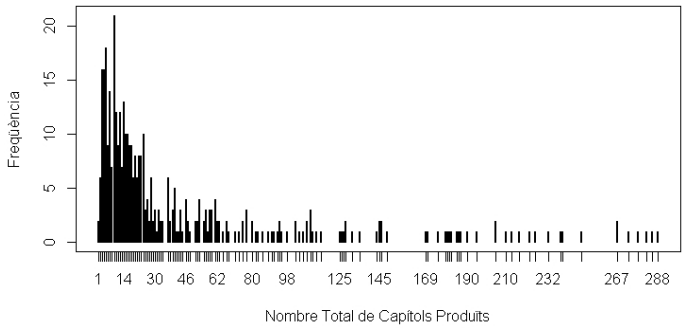


FIGURA 4.4. Diagrama de freqüències per a la variable *Nombre Total de Capítols Produïts* (valor màxim truncat a 300).

La majoria d'observacions es concentren en els primers valors entre 1 i 40 capítols produïts. Es tracta d'una distribució molt asimètrica amb una llarga cua a la dreta. En realitat s'estan barrejant les quatre espècies de Senecis considerades que, com es veurà a continuació, presenten produccions de capítols molt diferents.

4.3.2. Descriptiva amb covariables i prova d'homogeneïtat

La Figura 4.5 desglossa la Figura 4.4 per espècie. Ara es pot visualitzar clarament les grans diferències existents en la producció de capítols. Les espècies autòctones, d'envergadura molt més petita, tenen una producció de capítols molt inferior a les espècies exòtiques de major mida. El valor màxim per l'espècie *S. Inaequiden* s'ha truncat a 600. El valor màxim per les espècies *S. Lividus* i *S. Vulgaris* s'ha truncat a 150 i, el màxim per l'espècie *S. Pterophorus* s'ha truncat en 1600.

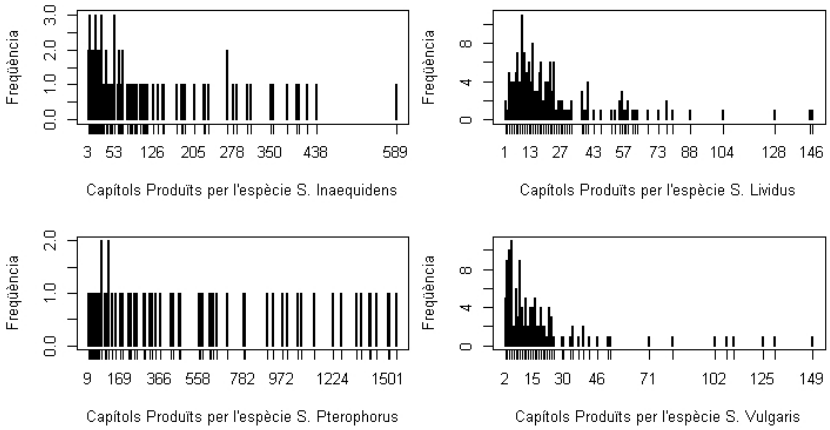


FIGURA 4.5. Diagrama de freqüències per a la variable *Nombre Total de Capítols Produïts* segons *Espècie* (valors màxims truncats a 600 (SI), 150 (SL i SV) i 1600 (SP)).

La Taula 4.5 recull els estadístics descriptius per la variable TC per espècie de *Senecio* i globalment.

Espècie	*S. Inaequidens*	*S. Lividus*	*S. Pterophorus*	*S. Vulgaris*	Total
N	100	164	82	129	475
$\overline{Y}$	262,40	28,02	1069,40	29,80	257,62
S	675,76	39,90	1897,86	49,80	926,54
CV	257,53%	142,38%	177,47%	167,13%	359,65%
$Med.$	64	17	426	12	24
$Q1$	23,75	9,00	101,00	5,00	10,00
$Q3$	209,25	29,25	1115,50	23,00	102,00
$Min.$	3	1	9	2	1
$Max.$	4854	393	9856	273	9856

TAULA 4.5. Taula d'estadístic descriptius per a la variable *Nombre Total de Capítols Produïts* per espècie i globalment

Com es pot apreciar a la Taula 4.5, les espècies exòtiques *S. Pterophorus* i *S. Inaequidens* van presentar una producció de capítols molt més elevada que les espècies autòctones les quals van presentar una producció similar de capítols. Si comparem les medianes, l'espècie amb una producció més petita correspon al *S. Vulgaris*

seguida del *S. Lividus*. El *S. Inaequidens* quintuplica i quadruplica aquestes produccions respectivament i finalment, el *S. Pterophorus* multiplica per 35 i 25 les produccions de les autòctones i per 7 la producció del *S. Inaequidens*. Notar que el mínim de capítols produïts és estrictament major a 0 donat que només es van seleccionar aquelles plantes que van produir capítols susceptibles de ser predats pels insectes.

El diagrama de caixa (Figura 4.6) il·lustra les diferències detectades en la Taula 4.5 i permet detectar una sèrie de plantes que són considerades *outliers* per presentar valors molt allunyats dels valors centrals. La presència d'aquestes plantes en l'estudi es tradueix en més variabilitat.

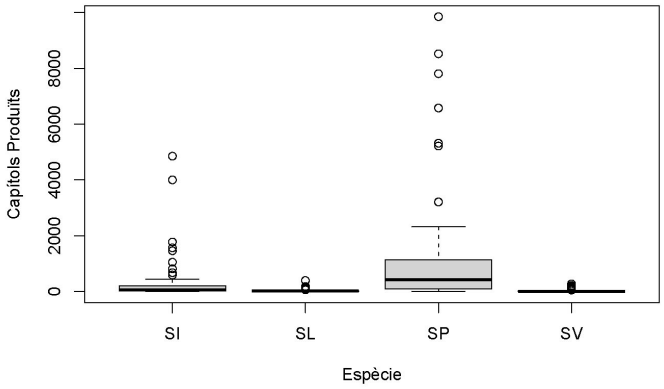


FIGURA 4.6. Diagrama de caixa de la variable *Nombre Total de Capítols Produïts*.

L'elevat nombre de capítols produïts per l'espècie *S. Pterophorus* distorsiona l'eix i no permet visualitzar correctament el nombre de capítols produïts en les espècies *Lividus* i *Vulgaris* on la seva caixa apareix concentrada en el valor 0.

La Taula 4.6 recull la suma del nombre total de capítols produïts per espècie i localització. D'aquesta forma, es pretén detectar si va existir alguna localització on la producció de les plantes va ser més elevada que en la resta.

Efectivament, tot i que les localitzacions es creien inicialment homogènies, s'observa que en la localitat de *Vallforners* les condicions van resultar més favorables pels *Senecios*, experimentant totes les espècies la producció de capítols més elevada en aquesta localització. En canvi, a la localització de *Fogueres* es van donar les produccions més baixes per a totes les espècies. Un altre fet destacable es que en la localització de *Santa Susanna* les espècies autòctones presenten una bona producció de capítols mentre que per les espècies exòtiques és una localització de poca producció de capítols.

La prova Khi-quadrat permet afirmar que existeix relació entre aquests dos factors, depenent la producció de capítols d'una espècie de la localització on es troba (χ^2=20.477,97, $p < 0{,}001$).

Per explicar el motiu de les diferències en les produccions de capítols necessitaríem conèixer les característiques pròpies de les diferents localitzacions però aquestes

Localització	Espècie				Total
	S. Inaequidens	*S. Lividus*	*S. Pterophorus*	*S. Vulgaris*	
Can Bosc	2.655	368	26.347	364	29.734
	8,93%	1,24%	88,61%	1,22%	100%
	10,12%	8,01%	30,05%	9,47%	24,30%
Can Perepoc	1.115	612	11.090	120	12.937
	8,62%	4,73%	85,72%	0,93%	100%
	4,25%	13,32%	12,65%	3,12%	10,57%
Can Tarrer	3.430	302	8.385	124	12.241
	28,02%	2,47%	68,50%	1,01%	100%
	13,07%	6,57%	9,56%	3,23%	10,00%
Fogueres	971	452	6.260	216	7.899
de Montsoriu	12,29%	5,72%	79,25%	2,73%	100%
	3,70%	9,83%	7,14%	5,62%	6,45%
Santa Susanna	242	1.404	5.522	1.135	8.303
	2,91%	16,91%	66,51%	13,67%	100%
	0,92%	30,55%	6,30%	29,53%	6,79%
Vallforners	17.827	1.458	30.087	1.885	51.257
	34,78%	2,84%	58,70%	3,68%	100%
	67,94%	31,72%	34,31%	49,04%	41,89%
Total	26.240	4.596	87.691	3.844	122.371
	21,44%	3,76%	71,66%	3,14%	
	100%	100%	100%	100%	

TAULA 4.6. Producció total de capítols (suma del nombre de capítols produïts) per espècie i localització. Percentatge fila i percentatge columna.

es va descartar recollir-les inicialment donat que no s'esperaven observar grans diferències entre les mateixes.

La Taula 4.7 d'estadístics descriptius per localització i espècie, confirma el que s'ha observat de forma global a la Taula 4.6. S'observa que la localització on les espècies han produït més capítols ha estat *Vallforners* i, aquest resultat es pot interpretar com que en aquesta localització les condicions per a la producció de capítols són idònies per a totes les espècies de *Senecio*. Respecte a la localització on la producció de capítols ha estat més baixa, en termes de mediana de capítols produïts, correspon a la localització de *Can Tarrer*. Esmentar que, la localització de *Santa Susanna* ha estat més favorable per a la producció de cápítols de les espècies autòctones, per tant, es pot deduir que en aquesta localització es donen condicions favorables per a la producció en les espècies autòctones en detriment de les espècies exòtiques. De forma similar succeeix amb la localització de *Can Tarrer* on l'espècie exòtica *S. Inaequidens* presenta molta producció de capítols. Per tant, a la localització de *Can Tarrer* semblen donar-se les condicions favorables per a l'expansió d'aquesta espècie exòtica.

Espècie	***Localització***	*Can Bosc*	*Can Perepoc*	*Can Tarrer*	*Fogueres de Montsoriu*	*Santa Susanna*	*Vallforners*
	N	21	14	19	15	4	27
	$\overline{Y}$	126,43	79,64	180,53	64,73	60,50	660,26
	S	185,55	95,42	398,16	63,12	32,34	1168,22
S. Inaequidens	CV	146,76%	119,81%	220,56%	97,50%	53,45%	176,93%
	$Med.$	80	33,5	28	48	57	306
	$Q1 - Q3$	15-144	19,75-100,25	9,5-204	37,5-65	41,75-75,75	59,5-513,5
	$Min - Max$	3-818	6-312	4-1777	11-278	26-102	11-4854
		20	32	27	29	29	27
		18,40	19,13	11,19	15,59	48,41	54,00
		12,03	11,34	8,13	11,65	39,26	77,78
S. Lividus		65,37%	59,27%	72,72%	74,71%	81,10%	144,04%
		16,5	18	9	12	38	39
		12,75-22	10,7524,25	6-14,5	8-22	23-56	11-60
		4-62	1-55	1-40	3-56	6-146	3-393
		19	11	22	9	11	10
		1386,68	1008,18	381,14	695,56	502,00	3008,70
		2657,24	622,70	472,48	813,98	426,74	3234,96
S. Pterophorus		191,63%	61,76%	123,97%	117,03%	85,01%	107,52%
		312	1069	119	239	420	1425,5
		56,5-679	591,5-1433,5	62,25-630	205-1266	207,5-620,5	709,5-4793,25
		9-8525	12-1995	22-1413	14-2322	40-1505	145-9856
		16	12	23	20	29	29
		22,75	10,00	5,39	10,80	39,14	65,00
		23,22	6,74	4,20	11,57	51,13	77,43
S. Vulgaris		102,06%	67,42%	77,84%	107,17%	130,63%	119,12%
		16	7,5	4	7	16	24
		12-25,25	4,5-16,25	3-5	4-12,75	11-42	12-111
		5-102	2-21	2-20	2-52	5-195	5-273

TAULA 4.7. Taula d'estadístics descriptius per a la variable *Nombre Total de Capítols Produïts* per espècie i localització.

Donades les diferències observades en termes de producció de capítols entre les espècies, enlloc d'estudiar la predació en termes absoluts, es pot estudiar la taxa de predació observada, es a dir, expressar l'herbivorisme en termes de percentatges.

4.4. Descriptiva variable *Taxa de Capítols Predats*

Es defineix la *Taxa de Capítols Predats* com el quocient entre el nombre de capítols predats i el nombre total de capítols produïts per la planta, es a dir

$$\text{Taxa de Capítols Predats} = \frac{\text{Nombre Total de Capítols Predats}}{\text{Nombre Total de Capítols Produïts}} \cdot 100\%$$

Aquesta taxa és d'interès perquè, com es veurà més endavant en el capítols de modelització, utilitzar un model per a una variable de recompte amb una variable *offset* es pot interpretar en termes de modelització d'una taxa. A més, les Taules 4.3 i 4.5 mostren grans diferències entre espècies i per tant es addient relativitzar la quantitat de capítols predats a la quantitat de capítols disponibles.

4.4.1. Descriptiva sense covariables

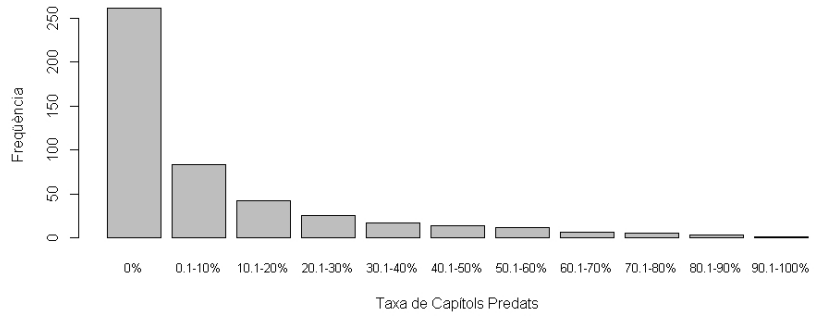


FIGURA 4.7. Diagrama de barres per les freqüències de la variable *Taxa de predació* agrupades. La primera barra representa la freqüència del 0%, les següents s'agrupen de 10% en 10%.

Com en els casos anteriors, s'estudia primer la variable *Taxa* sense tenir en compte els factors *Espècie* i *Localització*. La Figura 4.7 correspon al diagrama de barres de la variable *Taxa* on la primera barra correspon exclusivament a la freqüència de 0 i les barres següents agrupen en franges de 10%.

Es veu el mateix comportament que en la variable TP: una elevada freqüència en el 0 i a continuació una cua que decau lentament.

4.4.2. Descriptiva amb covariables

La Figura 4.8 mostra el diagrama de barres de la variable *Taxa de Capítols Predats* per a les quatre espècies en estudi. En aquesta figura es poden apreciar les diferents

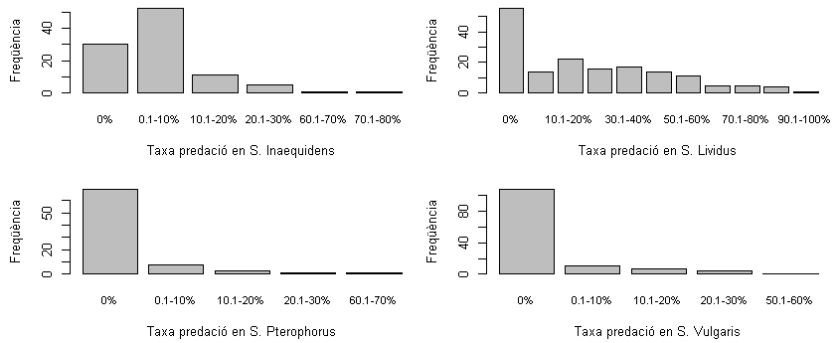


FIGURA 4.8. Diagrama de barres per les freqüències de la variable *Taxa de predació* agrupades. La primera barra representa la freqüència del 0%, les següents s'agrupen de 10% en 10% (no es dibuixen totes les barres).

distribucions de les taxes de predació que es donen en cada espècie, essent les espècies menys predades el *S. Pterophorus* i el *S. Vulgaris*. Per l'espècie *S. Lividus*, hi ha un nombre força elevat de plantes amb taxa del 0% però, a la vegada, també presenta una llarga cua que arriba fins a taxes de predació molt elevades. En canvi, l'espècie *S. Inaequidens* presenta menys nombre de plantes amb taxa igual a 0% que l'espècie anterior però, el gruix de plantes es troba en taxes relativament petites de predació, d'entre el 0,1 i el 10%, presentant molt poques plantes amb una taxa que superin el 30% de predació.

Espècie	*S. Inaequidens*	*S. Lividus*	*S. Pterophorus*	*S. Vulgaris*	Total
N	100	164	82	129	475
$\overline{Y}$	6,67%	23,20%	1,94%	2,51%	10,43%
S	11,45%	24,46%	8,00%	7,48%	18,66%
CV	171,69%	105,45%	411,52%	297,55%	178,86%
$Med.$	3,43%	17,83%	0%	0%	0%
$Q1$	0%	0%	0%	0%	0%
$Q3$	8,18%	38,57%	0%	0%	12,11%
$Min.$	0%	0%	0%	0%	0%
$Max.$	76,47%	100%	62,84%	60,00%	100%

TAULA 4.8. Taula d'estadístic descriptius (en %) per a la variable *Taxa de Capítols Predats* per espècie i globalment.

La Taula 4.8 recull l'estadística descriptiva de la variable *Taxa de capítols predats*. S'observa que l'espècie més predada correspon a l'espècie *S. Lividus* on, en mitjana el 23,2% dels capítols han estat predats per les larves dels insectes, no obstant, la variabilitat és força elevada i si es mira la mediana d'aquest percentatge es redueix fins al 17,83%. La segona espècie amb major predació és l'espècie *S. Inaequidens* que presenta una taxa mitjana de predació del 6,67% també amb una elevada variabilitat i amb un valor per la mediana que es redueix al 3,43%. Les espècies *S. Pterophorus* i *S. Vulgaris* presenten molt poca predació, com ja s'havia observat abans, i en

particular, el valor de la mediana per a aquestes dues espècie és igual a 0%. Això s'explica, com s'ha indicat abans, perquè més del 80% de les observacions són zero.

Relativitzar la predació observada respecte el total de capítols produïts per la planta, permet detectar quina és realment l'espècie més afectada per la predació un cop eliminat l'efecte producció (o grandària) de la planta. Això és el que posa de manifest el diagrama de caixa de la Figura 4.9. S'observen clares diferències entre les espècies de Senecis considerades. La mediana per a les espècies *S. Pterophorus* i *S. Vulgaris* és situa en el valor 0, mentre que aquesta és una mica superior per a l'espècie *S. Inaequidens* i pren un valor molt elevat en el cas de l'espècie *S. Lividus*.

Com s'ha comentat anteriorment, la predació total observada per cada espècie depenia de la localització de mostreig però, això també succeeix per a la producció total de capítols, per tant, es descriu a continuació la taxa de predacció observada de cada espècie a cada localització de mostreig.

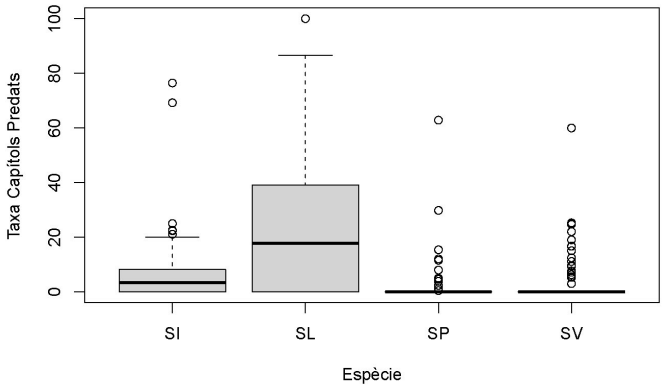


FIGURA 4.9. Diagrama de caixa de la variable *Taxa de Capítols predats*.

En base als estadístics de la Taula 4.9, quan es comparen les mitjanes aritmètiques de les taxes de predació es veu que a la localitat de *Vallforners* és on la taxa de predació és més gran en totes les espècies amb excepció del *Pterophorus*. Aquesta espècie en concret presenta la taxa de predació més elevada en la localitat de *Fogueres*. En aquesta localitat l'espècie *Vulgaris* presenta molt poca predació. Per tant, la taxa de predació per a les dues espècies que presentaven els percentatges més elevats de zeros és molt diferent en funció de la localitat.

Espècie	***Localització***	*Can Bosc*	*Can Perepoc*	*Can Tarrer*	*Fogueres de Montsoriu*	*Santa Susanna*	*Vallforners*
	N	21	14	19	15	4	27
	$\overline{Y}$	7,19	0,76	6,27	7,01	2,55	10,03
	S	7,79	1,98	16,10	5,38	1,80	14,89
S. Inaequidens	CV	108,30%	261,39%	256,71%	76,72%	70,84%	148,40%
	$Med.$	3,89	0	0,59	5,97	2,96	6,11
	$Q1 - Q3$	0-12,15	0-0	0-3,36	3,33-9,20	2,21-3,30	3,25-9,43
	$Min - Max$	0-25,00	0-7,14	0-69,23	0-18,75	0-4,26	0-76,47
		20	32	27	29	29	27
		18,43	22,60	23,55	21,46	12,85	40,06
		18,55	18,15	22,86	23,88	11,74	37,23
S. Lividus		100,64%	80,29%	97,08%	111,26%	91,36%	92,93%
		19,81	20,87	20	16,67	10,71	47,62
		0-31,41	8,61-35,40	0-43,65	0-40,91	3,64-19,61	0-74,68
		0-52,94	0-61,54	0-66,67	0-70,83	0-41,67	0-100
		19	11	22	9	11	10
		2,92	1,40	0,88	7,54	0	0,11
		7,24	4,64	2,72	20,80	0	0,36
S. Pterophorus		247,55%	331,66%	307,65%	275,91%	–	316,23%
		0	0	0	0	0	0
		0-1,18	0-0	0-0	0-0	0-0	0-0
		0-29,79	0-15,38	0-12,07	0-62,84	0-0	0-1,14
		16	12	23	20	29	29
		1,48	2,55	1,88	0,28	1,80	5,83
		4,04	5,65	5,71	1,24	6,07	12,42
S. Vulgaris		273,79%	222,04%	303,80%	447,21%	337,74%	213,09%
		0	0	0	0	0	0
		0-0	0-1,32	0-0	0-0	0-0	0-7,69
		0-12,50	0-19,05	0-25,00	0-5,56	0-24,73	0-60

TAULA 4.9. Taula d'estadístics descriptius (en %) per a la variable *Taxa de capítols predats* per espècie i localització.

4.5. La influència dels zeros

Donat que els recomptes iguals a 0 tenen una influència clara sobre aquests estadístics, un exercici interessant es calcular aquests mateixos estadístics però eliminant les observacions iguals a 0, es a dir, restringir la població d'estudi a aquelles plantes que han experimentat algun tipus de predació. D'altra banda, calcular aquest percentatge permetrà fer-se una idea de la intensitat del dany causat pels insectes.

La Taula 4.10 mostra el nombre de zeros i el percertatge que representen dins de cada espècie.

Espècie	N	$Nzeros$	%
S. Inaequidens	100	30	30%
S. Lividus	164	55	33%
S. Pterophorus	82	69	84%
S. Vulgaris	129	107	83%

TAULA 4.10. Nombre d'observacions iguals a 0 i percentatge per espècie.

A continuació, la Taula 4.11 recull el nombre d'observacions amb valor igual a 0 per espècie i localització i, es calcula el percentatge que representen aquestes observacions respecte el nombre total de plantes mostrejat per espècie i localització.

	Espècie				
Localització	*S. Inaequidens*	*S. Lividus*	*S. Pterophorus*	*S. Vulgaris*	Total
Can Bosc	6	8	14	14	42
	28,57%	40%	73,68%	87,5%	55,27%
Can Perepoc	11	6	10	9	36
	78,57%	18,75%	90,91%	75%	52,17%
Can Tarrer	7	10	18	20	55
	36,84%	37,04%	81,82%	86,96%	60,44%
Fogueres	2	14	7	19	42
de Montsoriu	13,33%	48,28%	77,77%	95%	57,53%
Santa Susanna	1	6	11	26	44
	25%	20,69%	100%	89,66%	60,27%
Vallforners	3	11	9	19	42
	11,11%	40,74%	90%	65,52%	45,16%
Total	30	55	69	107	261
	30%	33,54%	84,15%	82,95%	54,95%

TAULA 4.11. Nombre d'observacions iguals a 0 recollides per espècie i localització. Percentatge respecte el nombre total d'observacions per cel·la.

Com ja s'havia observat en les Taules 4.4 i 4.7 semblaven existir localitzacions més favorables a les espècies autòctones i localitzacions més favorables a les espècies invasores. Per a l'espècie *S. Inaequidens* la localitat on s'ha observat un major nombre de zeros ha estat en *Can Perepoc* amb un 78,57% de zeros. Per a l'espècie

S. Lividus ha estat la localitat de *Fogueres de Montsoriu* amb un 48,28% i, per les espècies *S. Pterophorus* i *S. Vulgaris* la localitat de *Santa Susanna* amb el 100% i 89,66% de plantes amb zero predació, respectivament. La prova de la Khi-quadrat conclou que la distribució dels zeros no es homogènia dins les localitzacions entre les espècies considerades (χ^2=33,22, p=0,004). No obstant, una de les limitacions d'aquest estudi es no poder conèixer més profundament les característiques de les localitzacions i tampoc es disposa d'informació relativa a la quantitat d'insectes depredadors presents a les diferents localitzacions. Aquesta falta de coneixement implica que no podem saber la causa de perquè hi ha localitzacions més adients per a unes espècies en concret. En general però, si s'observa el percentatge de zeros total per localitat no existeixen diferències tant evidents com les que es donen entre espècies.

A continuació, la Taula 4.12 recull els estadístics descriptius de la variable *Nombre Total de Capítols Predats* per a les plantes amb algun tipus de predació, és a dir sense tenir en compte els zeros.

Espècie	*S. Inaequidens*	*S. Lividus*	*S. Pterophorus*	*S. Vulgaris*	Total
N	70	109	13	22	214
$\overline{Y}$	18,44	13,84	33,85	17,14	16,90
S	36,62	28,31	63,11	26,64	34,03
CV	198,58%	204,51%	186,45%	155,44%	201,36%
$Med.$	6	6	12	4	6
$Q1$	2	3	2	1	3
$Q3$	18,75	13	26	21,5	15,75
$Min.$	1	1	1	1	1
$Max.$	238	233	230	111	238

TAULA 4.12. Taula d'estadístics descriptius per a la variable *Nombre Total de Capítols Predats* seleccionant les plantes amb predació > 0, per espècie i globalment

El primer que s'observa a la taula anterior és que s'ha reduit molt la mostra en les espècies S. Pterophorus i S. Vulgaris. Com ja s'havia observat en la Figura 4.2 aquestes espècies presentaven més d'un 80% d'observacions iguals a 0.

Respecte la Taula 4.3 on les plantes sense predació eren presents, ara s'observa que l'espècie amb major predació correspon al *S. Pterophorus* quan abans ocupava la tercera posició respecte a predació observada. L'espècie *S. Vulgaris* s'equipara amb l'espècie *S. Inaequidens* en termes de mitjana. Aquesta última era l'espècie més predada si es consideraven totes les observacions. Respecte l'espècie *S. Lividus*, abans ocupava la segona posició i ara és l'espècie menys predada si només es consideren les plantes que han patit algun tipus de predació, en termes de mitjana. Si ens fixem en la mediana, és l'espècie *S. Vulgaris* la menys predada, seguida de les espècies *S. Inaequidens* i *S. Lividus* i, l'espècie *S. Pterophorus* és la més predada amb diferència respecte les altres tres. Això sembla indicar que, en el cas que es produeixi la predació, la intensitat de la mateixa és molt diferent entre les espècies.

A continuació es presenta la descriptiva de la variable *Nombre Total de Capítols Produïts* per veure si es detecten diferències entre les produccions de totes les plantes i les produccions només de les plantes que han experimentat la predació.

Espècie	*S. Inaequidens*	*S. Lividus*	*S. Pterophorus*	*S. Vulgaris*	Total
N	70	109	13	22	214
$\overline{Y}$	355,07	35,60	318,62	89,68	162,85
S	790,03	46,25	381,22	86,27	484,42
CV	222,50%	129,93%	119,65%	96,19%	297,46%
$Med.$	104	23	210	36	39
$Q1$	49,25	13	52	16,5	18
$Q3$	283,25	40	366	173,75	122,5
$Min.$	4	3	25	4	3
$Max.$	4854	393	1408	273	4854

TAULA 4.13. Taula d'estadístics descriptius per a la variable *Nombre Total de Capítols Produïts* seleccionant les plantes amb predació > 0, per espècie i globalment.

Si es comparen aquests resultats de la Taula 4.13 amb els observats a la Taula 4.5, es conclou que per les espècies *S. Inaequidens* i *S. Lividus* no hi han grans diferències. Respecte les plantes de *S. Pterophorus*, les plantes amb predació han produït menys capítols que quan es consideren totes les plantes mostrejades, no obstant, continuen sent, juntament amb la varietat *Inaequidens*, les plantes més grosses. Finalment, les plantes predades de l'espècie *S. Vulgaris* han produït més capítols respecte a quan es consideren totes les plantes mostrejades. Sembla doncs que, els insectes es senten atrets per aquelles plantes amb major producció de capítols, amb excepció del *S. Pterophorus*. O bé, que els insectes depredadors es localitzen allà on es donen les condicions més favorables per a la floració de les espècies de *Senecio*.

Per acabar, s'observa quin és l'efecte de treure les plantes sense predació en la variable *Taxa de Capítols Predats*. La Taula 4.14 recull els estadístics descriptius.

Espècie	*S. Inaequidens*	*S. Lividus*	*S. Pterophorus*	*S. Vulgaris*	Total
N	70	109	13	22	214
$\overline{Y}$	9,53	34,90	12,26	14,73	23,15
S	12,67	22,15	17,17	12,33	21,88
CV	132,97%	63,47%	140,05%	83,70%	94,49%
$Med.$	5,75	32,14	5,02	11,11	16,20
$Q1$	3,16	18,27	2,37	6,47	6,01
$Q3$	11,25	50	12,07	18,48	33,33
$Min.$	0,44	2,50	0,48	3,03	0,44
$Max.$	76,47	100	62,84	60	100

TAULA 4.14. Taula d'estadístics descriptius per a la variable *Taxa de Capítols Predats* seleccionant les plantes amb predació > 0, per espècie i globalment

Si es comparen els resultats de la Taula 4.14 amb els resultats de la Taula 4.8 s'observa que l'espècie més predada continua sent l'espècie *S. Lividus* però ara, tot

i que la mida de la mostra s'ha reduït bastant, les espècies *S. Pterophorus* i *S. Vulgaris* superen en termes de taxa mitjana de predació a l'espècie *S. Inaequidens*. Per tant, com s'ha apuntat abans, la intensitat de la predació allà on s'ha produït és diferent segons l'espècie considerada i, no sembla està lligada amb el fet que hi hagi hagut més o menys predació.

4.6. Conclusions preliminars

La raó per mostrar les mitjanes calculades amb i sense zeros de la variable *Taxa de Capítols Predats*, és visualitzar la influència d'aquests zeros sobre l'anàlisi descriptiva. En el cas de les espècies amb un elevat nombre de zeros, la taxa de predació que presenten és molt baixa però si es calcula la mitjana excloent les plantes que no han presentat cap capítol predat, òbviament la taxa de predació esdevé més elevada però, l'aspecte important és que no és manté l'ordre en la predació entre les espècies. Per exemple, l'espècie *S. Inaequidens* presenta una taxa de predació més elevada que les espècies *S. Pterophorus* i *S. Vulgaris* quan es tenen en compte els zeros però, aquest resultat es capgira si els càlculs es realitzen sense els zeros (Figura 4.10 dreta). Això es podria interpretar com que les plantes de les espècies *S. Pterophorus* i *S. Vulgaris* són poc atacades pels depredadors, però quan es dóna aquest atac és bastant agressiu. D'altra banda, l'espècie *S. Lividus* es tracta d'una espècie força atacada pels depredadors, només s'ha detectat un 30% de plantes amb predació igual a 0, però les plantes atacades no han presentat una predació tant agressiva com en les espècies *S. Pterophorus* i *S. Vulgaris* en termes de mitjana de predació (Figura 4.10 esquerra).

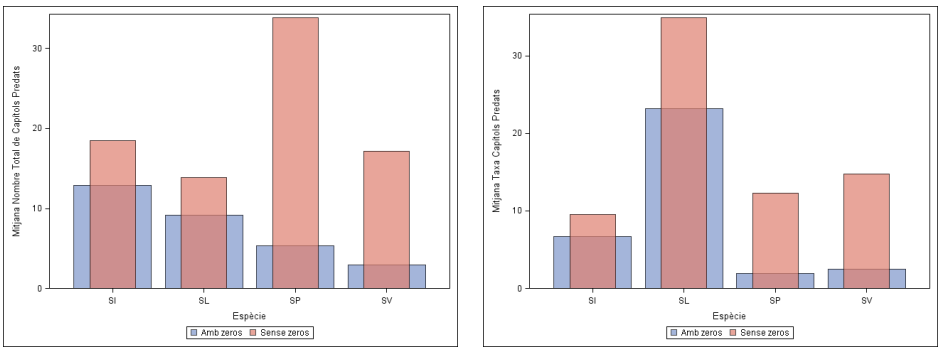


FIGURA 4.10. Diagrames de barres on l'alçada representa la mitjana de *Nombre Total de Capítols Predats* (esquerra) i *Taxa de Capítols Predats* (dreta), amb zeros (barra blava) i sense zeros (barra vermella) per espècie.

Realitzar aquesta comparativa entre la predació en totes les plantes i la predació només en aquelles plantes que l'han experimentat permet visualitzar dos aspectes diferenciadors en la predació:

- que hi ha espècies que pateixen més que d'altres l'atac dels insectes i,
- que hi ha espècies on l'atac, quan es produeix, és molt més agressiu que en altres espècies,

és a dir, hi ha dos vessants diferenciades en la predació: la freqüència i la intensitat del dany que ocasiona.

Capítol 5
Models *Zero Inflats*

En aquest capítol es presenten dos models que permeten fer front a l'excessiva quantitat de zeros, el model d'excés de zeros amd distribució de Poisson (ZIP, *Zero Inflated Poisson model*) i el model d'excés de zeros amb distribució binomial negativa (ZIBN, *Zero Inflated Negative Binomial model*). Aquests models van començar a popularitzar-se la darrera dècada atès que, en els darrers anys, es poden trobar implementats en diferents paquets estadístics. Aquests models són emprats també en molts àmbits fora de l'ecologia, per exemple, en els camps de les ciències socials, estudis d'accidents de trànsit, l'econometria, la psicologia, etc.

Al llibre d'en Zuur *et al.* (2007) es dedica un capítol del mateix a descriure en detall els models amb excés de zeros. També es fa una petita menció del model ZIP a Johnson *et al.* (2005). Aquestes dues referències són les que han estat utilitzades com a base de les seccions següents.

5.1. L'excés de zeros

L'excés de zeros associat a una mostra, significa que s'observen molts més zeros del que s'esperaria assumint una distribució de probabilitat inicialment raonable donada la naturalesa del problema. Si s'observa de nou la Figura 4.1, si només es consideren les freqüències entre el 0 i el 30, sota una distribució de Poisson, no és raonable obtenir 261 zeros, el 56,74% de les observacions, ja que $P(Y = 0) = 0,05$ si $Y \sim Poisson(\lambda = 3)$.

Fer cas omís de l'excés de zeros pot tenir dues conseqüències importants: en primer lloc, els paràmetres estimats i els errors estàndards poden estar esbiaixats (Lambert 1992; MacKenzie *et al.* 2002) i, en segon lloc, el nombre excessiu de zeros pot provocar sobre-dispersió en les dades, que vol dir que mostren una variància superior a la que s'esperaria sota un model teòric concret.

És important plantejar-se la següent pregunta: per què es donen tants zeros?. Poder identificar diferents fonts de zeros, és a dir, poder identificar diferents orígens dels mateixos és important per distingir-ne la naturalesa. A la literatura es distingeix entre els *zeros vertaders* i els *zeros falsos*, tot i que no hi ha un consens general sobre com fer aquesta distinció. Des d'un punt de vista estadístic, direm que els zeros

vertaders són aquells que es poden atribuir al procés de recompte, mentre que els zeros falsos són els que estan provocant l'excés de zeros i la seva explicació és aliena al procés de recompte. A la següent subsecció s'aprofundeix més en aquests dos conceptes. D'ara en endavant, denotarem per *ZeroF* els zeros falsos i per *ZeroPR* els zeros corresponents al procés de recompte.

5.1.1. Fonts de zeros

Com s'ha comentat abans, si s'assumeix que les nostres dades provenen d'una distribució de Poisson, no s'esperaria observar a la Figura 4.1 un percentatge tant elevat de valors iguals a zero. Per tant, una petita porció de la barra vertical del zero correspon als zeros de la distribució de Poisson tots els altres, són zeros de més. Es proposen a continuació diferents fonts d'error per explicar els zeros:

1 *Errors estructurals*: existeix una raó estructural que origina el zero. Per exemple, la planta ja es morta quan l'insecte apareixen en l'ambient o bé, l'insecte no reconeix la planta com a font potencial d'aliment.
2 *Errors de disseny i de mostreig*: són deguts a un mal disseny de l'experiment o bé a errors de mostreig. Per exemple, mostrejar en un període massa curt de temps o en una àrea massa petita.
3 *Error de l'observador*: si l'observador no està acostumat a aquest tipus d'estudi pot confondre la predació dels insectes d'interès amb la d'altres insectes i no comptabilitzar-la o bé, simplement el dany és petit i no el detecta.
4 *L'atzar*: el capítol és ferm candidat a ser atacat per l'insecte però l'insecte no l'ataca.

Poden existir d'altres fonts de zeros. Per exemple, mostrejar en un habitat inadequat, allà on no es troba l'insecte depredador. En aquest cas, aquests zeros han de ser eliminats de l'estudi. Existeix encara una altra font per explicar un excés de zeros: els *outliers*. La presència de dades molt anòmales desplacen la mitjana de la distribució i això, a la seva vegada, provoca que la probabilitat del valor zero disminueixi. Es possible encara trobar d'altres fonts per explicar l'excés de zeros.

Els zeros corresponents al punt quatre són els zeros que pertanyen al procés de recompte en estudi, mentre que els tres primers corresponen a zeros *falsos* que no pertanyen pròpiament al procés de recompte, en base a la notació que fem servir. Dels zeros falsos d'aquest estudi la majoria són estructurals con es veurà més endavant. Els zeros estructurals a la literatura també s'anomenen zeros *vertaders*, en el sentit que sempre que es repliqui l'experiment aquests hi seran ja que la raó per ser-hi és estructural.

Els insectes locals, per evolució, estant adaptats a les espècies autòctones i per tant existeix una certa sincronia entre el moment de reproducció de l'insecte i la floració de la planta autòctona. Per tant, si les plantes exòtiques presenten un moment de floració molt diferent al de les plantes autòctones, aleshores l'insecte depredador ja no es troba en l'estadi reproductiu quan les plantes exòtiques produeixin els capítols. Els zeros observats produïts per aquesta falta de sincronia correspondrien a errors estructurals, donat que els investigadors poc poden millorar en el disseny

per aconseguir que la floració de les plantes i l'època reproductiva de l'insecte coincideixin en el temps. Aquests zeros formarien la major part dels zeros falsos. Una altra raó que pot originar zeros estructurals és que els insectes locals no identifiquen les plantes exòtiques com a font potencial de predació. En aquests casos no hi hauria interacció planta-insecte.

D'errors de disseny en aquest estudi no és d'esperar que n'hi hagi o, almenys seran mínims. El temps d'estudi considerat, d'abril a desembre, és prou gran per captar tot el temps reproductiu tant de les plantes com dels insectes. Tanmateix, totes les localitzacions es van seleccionar de forma que fos possible trobar les espècies de plantes estudiades.

Seguidament, també es contemplen els zeros a causa d'errors d'observació ja que, aparentment, no és sempre fàcil de detectar la predació pels insectes d'interés en estudi i es pot confondre amb predació d'un altre tipus d'insecte. Així que aquests són també zeros falsos, però d'aquest tipus s'espera que n'hi hagi molts menys donada l'expertesa de la investigadora involucrada en l'estudi.

L'altre tipus de zeros, els zeros reals del procés de recompte, vénen de plantes que han estat en contacte amb els insectes en qüestió però, per cap raó en particular, no van ser atacades.

5.2. Models generals

En aquesta secció es defineixen els models teòrics utilitzats en aquest treball. Préviament es comenten breument les distribucions que els originen i algunes de les seves propietats més importants.

5.2.1. Resenya històrica dels models *Zero Inflats*

Les distribucions bàsiques Poisson o binomial negativa són les més habituals a l'hora de modelar dades de recomptes en absència d'excés de zeros (McCullagh i Nelder, 1989). No obtant, tal com ja s'ha comentat, en presència d'un excés de zeros el model de Poisson no ajusta bé aquest excés present a les dades.

Davant d'aquesta situació, es fan servir els models zero inflats, com ara el model en dues parts, també conegut com a model *condicional* o *hurdle model* o bé, el model de mixtura de distribucions (Lambert, 1992; Welsh *et al.*, 1996) que és el que es veu en aquest treball.

L'enfoc d'un model estructurat en dues parts o *hurdle model* és adequat per a l'anàlisis de dades de recomptes amb excés de zeros degut a zeros estructurals. Apareix com un model on la primera part correspon a un model binari (Bernoulli), i la segona part a un model de recompte truncat en el zero (Cameron i Trivedi, 1998). Per exemple, es fa servir un model logístic pels zeros i un model de Poisson truncat al zero pels recomptes (Welsh *et al.*, 1996). Així doncs, aquest enfocament assumeix que sorgeixen zeros a partir d'un únic procés i un conjunt de covariables.

L'ús d'aquest enfocament permet estimar primer la probabilitat de presència i després, atès que està present, estimar l'esperança del procés de recompte. Un dels avantatges d'aquest tipus de models és la seva programació: són relativament fàcils d'ajustar i d'interpretar.

Els models de mixtures de distribucions (ZIP, ZIBN) són adequats per l'anàlisi de dades de recomptes on l'excés de zeros pot provenir tant de zeros estructurals com de zeros del procés de recompte. Aquests models combinen distribucions de probabilitats escollides per la seva capacitat de presentar dos o més processos reals. El model de mixtura de distribucions utilitzat per modelar dades de recomptes és una barreja d'una distribució amb tota la massa en el zero i una distribució de Poisson o binomial negativa. Amb aquest enfocament, els zeros poden sorgir d'un dels dos processos, un procés on només s'observen zeros i un procés de recompte on també es pot observar el valor zero i, les seves respectives covariables relacionades (Lambert, 1992). Quan hi ha excés de zeros i sobredispersió causada, entre altres raons, per valors molt elevats en els recomptes, l'ús d'un model de mixtures de distribucions utilizant la distribució binomial negativa (ZIBN) és més adequat que la distribució de Poisson (Welsh *et al.*, 2000). La interpretació d'aquests models no és tan senzilla com en el cas dels *hurdle models*. L'ajust de models zero inflats Binomial (ZIB) o Poisson (ZIP) sense covariables, es pot trobar a la literatura des de fa temps (Johnson i Kotz 1969). Lambert (1992) estableix la forma general de regressió ZIP amb covariables en el model per al recompte de defectes en un procés de fabricació. A partir d'aquí, Heilbron (1994), Welsh *et al.* (1996) o Hall (2000), entre d'altres, van desenvolupar models més específics per a dades de recomptes amb excés de zeros. Més recentment, s'aplica estadística Bayesiana a la inferència estadística per aquest tipus de dades en Angers i Biswas (2003), Martin *et al.* (2005) i Kuhnert *et al.* (2005). El desenvolupament dels models zero inflats per a variables contínues també ha estat objecte d'estudi, Aitchison (1955), Stefansson (1996) o Fletcher *et al.* (2005), però cau fora de l'objectiu d'aquest treball.

Finalment, el model Binomial Negatiu també s'ha emprat per modelar conjunts de dades amb excés de zeros per la seva capacitat per adaptar la sobredispersió. No obstant, Welsh *et al.* (1996) i Hall (2000) demostren que l'excés de zeros sovint excedeix el nombre esperat de zeros sota la distribució binomial negativa.

Fins on arriba el nostre coneixement, no es troba a la literatura un consencs sobre com modelar dades amb excés de zeros. Aquest és encara un problema obert en modelització estadística. Com s'ha mencionat abans, realitzar l'enfoc des de la inferència bayesiana pot donar bons resultats, on el model incorpora informació relativa als zeros mostrals com a informació a priori. Una revisió d'altres mètodes per a modelar recomptes amb excés de zeros es troba a Ridout *et al.* (1998).

5.2.2. Desenvolupament teòric

Sigui Y_i una variable aleatòria que representa el recompte d'un fenomen d'interès, per exemple la variable aleatòria *Nombre Total de Capítols Predats* en la planta i-èssima. S'assumeix que,

$$Pr(Y_i = 0) = Pr(Y_i = ZeroF) + (1 - Pr(Y_i = ZeroF))\cdot Pr(Y_i = ZeroPR). \quad (5.1)$$

El component $Pr(Y_i = ZeroF)$ correspondria a una part de la barra dels zeros del gràfic de la Figura 4.1. La segona component prové de la probabilitat de que no és tracti d'un zero fals multiplicada per la probabilitat que sigui un zero del procés de recompte. Bàsicament, dividim la població mostrejada en dos grups. El primer grup conté només zeros (els zeros falsos). Aquest grup se l'anomena també observacions amb massa zero. El segon grup correspon als zeros del procés de recompte, que pot produir zeros així com valors més grans que zero. Remarcar que les dades no es divideixen en dos grups pròpiament, és només una suposició que es fa sobre l'existència d'aquests dos grups en la població. Donada una observació igual a zero no es coneix a quin grup pertany. Tot el que es coneix és que els valors diferents a zero (els recomptes positius) es troben en el segon grup. En realitat és com assumir que la població mostrejada consta de dues subpoblacions. Una que donarà sempre lloc a que el valor observat sigui zero, la que correspon als capítols que no poden ser predats, i l'altre que correspon a la població objecte d'interés.

El terme $Pr(Y_i = ZeroF)$ i $1 - Pr(Y_i = ZeroF)$ indica un procés Bernoulli i, de fet, això és el que s'implementa. Si s'assumeix que la probabilitat que Y_i sigui un fals zero és π, aleshores automàticament obtenim que la probabilitat que Y_i no sigui un zero fals és $1 - \pi$. L'equació (5.1) es pot reescriure com:

$$Pr(Y_i = 0) = \pi + (1 - \pi)\cdot P^*_Y(Y_i = ZeroPR). \quad (5.2)$$

El següent pas és substituir $P^*_Y(Y_i = ZeroPR)$ per allò que correspongui depenent de quin sigui el model de probabilitat teòric que es prengui com a base. En aquests casos, es pot assumir que els recomptes segueixen una distribució de Poisson o bé una distribució binomial negativa (BN), o bé la distribució geomètrica. Aquesta és la diferència entre el model zero inflat de Poisson (ZIP) i el model zero inflat Binomial Negatiu (ZIBN). Donat que la distribució geomètrica és un cas particular de la BN, no rep cap nom especial.

Els models zero inflats són un cas particular de models de mixtura (*Mixture Models*) donat que els zeros es modelen com provinents de dos processos diferents: el procés Poisson o BN i el procés constant igual a zero. El pes emprat en aquesta mixtura és el paràmetre d'una Bernoulli. Donat que es fa servir un model Bernoulli, es podran utilitzar també covariables dins d'aquest model. Això passarà quan s'assumeixi que π depèn d'unes covariants, fet que es veurà més endavant. Una representació esquemàtica d'un model zero inflat es dóna a la Figura 5.1.

En realitat es tracta d'un model on es modelitza conjuntament les següents probabilitats:

- la probabilitat d'ocurrència - no ocurrència del succés d'interès: $Pr(Y_i = 0)$
- el recompte de successos condicionat a l'observació del mateix: $Pr(Y_i = y_i \mid Y_i > 0)$

S'ajusta una única distribució,

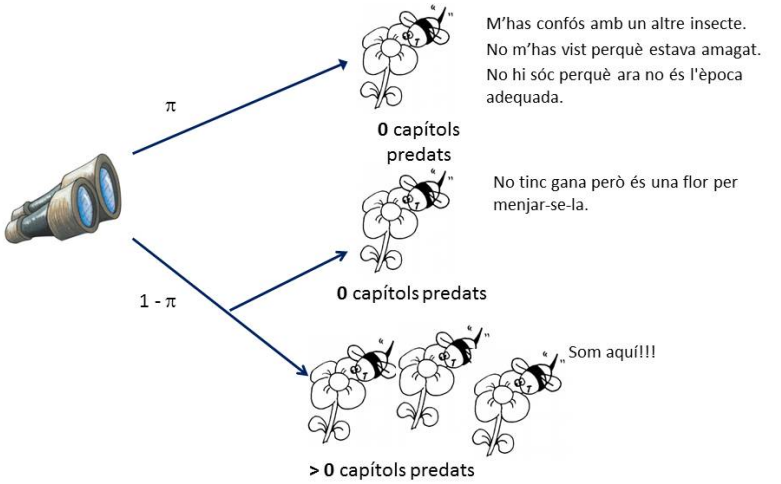


FIGURA 5.1. Esquema del principi subjacent del models ZIP i ZIBN contextualitzat a les dades de capítols predats pels insectes.

$$Pr(Y_i = k) = \begin{cases} \pi + (1-\pi)Pr^*(Y_i = 0) & \text{si} \quad k = 0 \\ (1-\pi)Pr^*(Y_i = k) & \text{si} \quad k > 0, \end{cases} \tag{5.3}$$

on $Pr^*(Y_i = k)$ és la funció de probabilitat teòrica que s'assumeix (Poisson o binomial negativa), π (*inflated probabilitiy*) és la proporció de zeros falsos i $(1 - \pi)Pr^*(Y_i = 0)$ la proporció de zeros vertaders.

La definició de mixtura discreta (Johnson *et al.*, 2005) diu que la distribució d'una variable aleatòria X és una mixtura de distribucions P_j amb pesos $p_j \in (0, 1)$, on $\sum_{j\geq 1} p_j = 1$, quan es compleix que

$$Pr(X = x) = \sum_{j\geq 1} p_j P_j(X).$$

Aleshores l'equació (5.3) es pot interpretar com la mixtura de dues distribucions: la distribució degenerada en el zero i la distribució dels recomptes P_Y^*.

5.2.3. Models Poisson i *Zero Inflated Poisson*

Les dades analitzades corresponen a plantes de *Senecio* que produeixen capítols succeptibles de ser atacats (predats) pels insectes. Des del punt de vista estadístic, direm que per a la planta i-èssima, que ha produït n_i capítols, el recompte de

capítols predats segueix una distribució Binomial de paràmetres n_i i p_i on p_i correpon a la probabilitat de predació. Ara bé, si n_i es fa gran i p_i es fa petit, la Binomial es pot aproximar mitjançant la distribució de Poisson de paràmetre $\mu_i = n_i \cdot p_i$. Per aquest motiu enlloc d'utilitzar la Binomial es fa servir la distribució de Poisson. Assumint doncs, que les plantes mostrejades han fet molts capítols i que la probabilitat que un capítol concret d'una planta concreta sigui predat és molt petita, aleshores la funció de probabilitat d'una variable aleatòria amb distribució de Poisson és:

$$P_Y^*(y_i; \mu_i \mid y_i \geq 0) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!},$$

d'on la probabilitat del zero ve donada per

$$P_Y^*(Y_i = 0) = e^{-\mu_i}.$$

Substituint aquesta probabilitat a l'equació (5.2) s'obté que, si Y_i segueix un model ZIP, llavors

$$Pr(Y_i = 0) = \pi + (1 - \pi)e^{-\mu_i}.$$

És a dir, la probabilitat d'observar un 0 és igual a la probabilitat d'un fals zero més la probabilitat que no sigui un fals zero multiplicada per la probabilitat d'observar un zero verdader.

La probabilitat d'observar un valor diferent de zero ve donada per:

$$Pr(Y_i = y_i) = (1 - \pi)P_Y^*(Y_i = y_i) = (1 - \pi)\frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}.$$

Aleshores, la funció de distribució de probabilitats P_Y d'un procés ZIP és:

$$P_Y(y_i; \mu_i \mid y_i \geq 0) = \begin{cases} \pi + (1 - \pi)e^{-\mu_i} & \text{si} \quad y_i = 0 \\ (1 - \pi)\frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} & \text{si} \quad y_i > 0. \end{cases}$$

L'esperança i la variància d'un procés ZIP es poden trobar a partir de l'equació (5.3) i corresponen a:

$$\begin{aligned} E[Y_i] &= \eta_i = (1 - \pi) \cdot \mu_i \\ V[Y_i] &= \mu_i \cdot (1 - \pi) \cdot (1 + \mu_i \cdot \pi) = \eta_i + \frac{\pi}{1 - \pi} \cdot \eta_i^2. \end{aligned}$$

Com es veu, el model ZIP relaxa l'assumpció d'igualtat estre esperança i variància del model de Poisson. D'altra banda, si $\pi > 0$ incorpora sobredispersió a les dades on el quocient $\frac{\pi}{1-\pi}$ és la contrapartida del paràmetre de dispersió en la binomial negativa.

Si es volen introduir covariables en el model aleshores, s'assumeix que el valor esperat de la distribució de Poisson evoluciona com a funció d'aquestes covariables de manera que:

$$\mu_i = e^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\cdots+\beta_q X_{iq}}, \tag{5.4}$$

Això és conseqüència del fet que en un model lineal generalitzat el *link* canònic és el logaritme, de manera que la relació entre les covariables i l'esperança tranformada, mitjançant el logarítme, és lineal. β_0 és el terme independent i $\beta_1, ..., \beta_q$ són els paràmetres objecte d'estimació que acompanyen a les covariables.

Pel que respecta a la probabilitat π, donat que el procés dels zeros es modela fent servir una distribució Bernoulli, l'enfoc més senzill és utilitzar la transformació logística amb només el terme independent. Es a dir, assumir que

$$\pi = \frac{e^{\gamma_0}}{1+e^{\gamma_0}}.$$

Ara bé, no hi ha cap inconvenient en incloure covariables per modelar la part dels zeros i, en aquest cas, les probabilitats variarien en funció de les condicions experimentals i s'assumiria:

$$\pi_i = \frac{e^{\gamma_0+\gamma_1 Z_{i1}+\gamma_2 Z_{i2}+\cdots+\gamma_r Z_{ir}}}{1+e^{\gamma_0+\gamma_1 Z_{i1}+\gamma_2 Z_{i2}+\cdots+\gamma_r Z_{ir}}}, \tag{5.5}$$

on γ_0 és el terme independent i $\gamma_1, ..., \gamma_r$ els paràmetres que acompanyen a les covariables del model per a π.

Cal esmentar que les covariables que intervenen en cadascuna de les parts a modelar no tenen perquè coincidir. Les covariables per a modelar l'esperança de la distribució de Poisson s'han denotat per X_j i les covariables utilitzades per modelar l'excés de zeros s'han denotat per Z_j. Pot passar que determinades covariables siguin adequades per modelar el procés de recompte i d'altres covariables per a la proporció del zeros.

El fet d'estimar la probabilitat π mitjançant la transformació logística implica que el model de Poisson no es pot considerar niat dins del model ZIP (Greene, 1994) per tant, els estadístics de bondat d'ajust clàssics no serviran per comparar aquests dos models.

Un cop establert el model i per tal d'estimar-ne els paràmetres, es construeix a continuació la funció de versemblança del mateix. Sigui $y = (y_1, y_2, \ldots, y_n)$ una mostra que prové d'un model ZIP i y_i s'ha recollit sota unes determinades condicions expressades en els vectors de covariants $(x_{i1}, x_{i2}, \ldots, x_{iq})$ i $(z_0, z_{i1}, z_{i2}, \ldots, z_{ir})$. La funció de versemblança és, per definició, igual a:

$$L(\pi_i, \mu_i; y) = \prod_{\substack{i \\ y_i=0}} \left[\pi_i + (1-\pi_i)e^{-\mu_i}\right] \cdot \prod_{\substack{i \\ y_i\neq 0}} (1-\pi_i)\frac{\mu_i^{y_i}e^{-\mu_i}}{y_i!}.$$

Així doncs, prenent logaritmes es tindrà que:

$$\begin{aligned} l(\pi_i, \mu_i; y_i) = & \sum_{\substack{i \\ y_i=0}} log(\pi_i + (1-\pi_i)e^{-\mu_i}) \\ & + \sum_{\substack{i \\ y_i \neq 0}} \left[log(1-\pi_i) + y_i log(\mu_i) - \mu_i - log(y_i!) \right], \end{aligned} \tag{5.6}$$

Tenint en compte les equacions (5.4) i (5.5), s'introdueixen les covariables en la funció de versemblança. Per trobar els estimadors màxim versemblants dels paràmetres es deriva aquesta expressió i s'iguala a 0. En aquest treball s'ha programat l'equació (5.6) en SAS 9.22 mitjançant el procediment *NLMIXED* per trobar les estimacions màxim versemblants dels paràmetres.

D'altra banda, el model de Poisson i, per extensió, el model ZIP té sentit utilitzarlos, com ja s'ha comentat, per modelar la mitjana dels recomptes però pot interessar modelar la taxa relativa a la unitat d'estudi enlloc del recompte. Aquest problema s'aborda introduint un terme *offset*.

Si s'assumeix que el logaritme de la taxa és lineal amb les covariants i que sota les condicions experimentals i-èssimes es tenen n_i observacions, té sentit assumir que:

$$log(\frac{\mu_i}{n_i}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_q X_{iq}$$

En el cas que ens ocupa, μ_i és l'esperança de la variable *Nombre Total de Capítols Predats* i n_i correspon al nombre total de capítols produits per la planta i-èssima.

L'equació anterior també es pot expressar com:

$$log(\mu_i) = log(n_i) + \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_q X_{iq}$$

Els recomptes n_i s'introdueixen en el model amb la transformació logarítmica i el coeficient que l'acompanya es fixa igual a 1. S'anomena *offset* perquè és una covariant de la qual no s'ha d'estimar el coeficient.

Finalment, volem fer notar que la distribució de Poisson compleix que la variància és igual al valor esperat. Aquesta propietat, davant de la presència d'un nombre molt elevat de zeros, és difícil que es doni. Més aviat passa tot el contrari, l'excés de zeros fa augmentar la dispersió de les dades i apareix el que s'anomena sobredispersió. En les nostres dades, com s'ha vist en el Capítol 4, la variància mostral és molt més elevada que la mitjana mostral. De fet, són la mitjana i la desviació típica mostral aquelles que coincideixen en ordre. En aquesta situació, assumir distribució de Poisson per a la variable resposta no és correcte. El fet de considerar un model ZIP ja fa augmentar la variància perquè, com s'ha vist, és una mixtura. Ara bé, aquest augment sovint no és suficient.

La distribució binomial negativa relaxa el supòsit d'igualtat entre esperança i variància i permet controlar la sobredispersió present a les dades. Per aquest motiu és la distribució que habitualment es fa servir enlloc de la Poisson quan hi ha sobredispersió i per això, s'ha considerat en aquest treball. Com es veurà a continuació, en realitat la distribució binomial negativa es pot considerar, a la seva vegada, una mixtura de Poisson. En les mixtures de Poisson la variància excedeix a la mitjana (Johnson *et al.*, 2005), de manera que aquest tipus de distribucions són més apropiades per modelar dades que presenten sobredispersió. Finalment, es pot afegir el procés associat a l'excés de zeros i obtenir el model ZIBN capaç de captar la sobredispersió i l'excés de zeros.

5.2.4. Models Binomial Negatiu i *Zero Inflated Negative Binomial*

La distribució binomial negativa (BN) com s'ha esmentat abans, s'empra en substitució de la distribució de Poisson quan les dades presenten sobredispersió. No obstant, les propietats teòriques i el seu origen no són tant coneguts i, per aquest motiu, es realitza una petita introducció a aquesta distribució per a conèixer-la més en profunditat. Més endavant es veurà que en el model ZIBN la distribució de Poisson es substitueix per la distribució BN.

Un dels resultats més coneguts de la distribució BN, enunciat per Greenwood and Yule (1920), és el següent: la BN és una mixtura de Poisson, *Mixed Poisson*. La variable d'interès es distribueix segons una distribució de Poisson amb esperança μ on s'assumeix que aquesta esperança no és constant sino que a la seva vegada segueix una distribució Gamma. Aquest resultat és el que es fa servir més habitualment per a definir la distribució BN.

D'altra banda, com s'ha comentat en el Capítol 3, el tractament estadístic de les dades s'ha realitzat agregant la informació a nivell temporal, és a dir, s'han sumat tots els registres recollits al llarg de l'estudi per cada individu. D'aquesta forma s'ha obtingut la suma del nombre total de capítols produïts i la suma del nombre total de capítols predats d'abril a desembre que constitueixen dues de les variables principals de l'estudi.

La suma de variables aleatòries de recomptes generen els processos de mixtures de distribucions. La variable aleatòria *Suma* és el resultat de barrejar un nombre diferent de poblacions que en el nostre cas, és el nombre de capítols predats en cada seguiment on, el nombre esperat de seguiments realitzats varia per cada planta. Lüders (1934) va enunciar que la BN es podia derivar d'una suma de N variables aleatòries independents i idènticament distribuïdes segons la distribució Logarítmica (veure Johnson *et al.*, 2005), on N segueix una distribució de Poisson. Aquesta derivació va ser anomenada més endavant per Boswell i Patil (1970) com a distribució *Poisson-Stopped Sum* de variables aleatòries logarítmiques. Formalment, X segueix una distribució BN quan

$$X = X_1 + X_2 + ...X_N,$$

essent N una variable aleatòria amb distribució de Poisson i X_i variables independents identicament distribuïdes amb distribució Logarítmica.

Més recentment, Valero *et al.* (2013) enuncien les propietats adients per tal que una distribució *Poisson-Stopeed Sum* sigui també una distribució *Mixed Poisson* com és el cas de la distribució BN.

Així doncs, acabem de justificar que la distribució BN admet ésser enunciada com a una *Mixed Poisson* així com una distribució *Poisson-Stopeed Sum*. Il·lustrem a continuació aquest resultat en el cas que ens ocupa.

Es defineix primer la funció generatriu de probabilitats (*fgp*) d'una variable aleatòria discreta, ja que la necessitem més endavant. Donada una variable aleatòria discreta X que pren valors en els enters no negatius amb probabilitats:

$$p_j = P(X = j), \quad j = 0, 1, ...$$

la seva funció generatriu de probabilitats es defineix com:

$$G(z) = \sum_{j=0}^{\infty} p_j z^j = E\left[z^X\right].$$

Aquesta sèrie de potències convergeix absolutament almenys per tots els nombres complexos z tals que $|z| = 1$ (veure Johnson *et al.*, 2005).

Els individus (les plantes de *Senecio*) són seguits durant totes les setmanes de floració. Cada setmana es recull el nombre total de capítols produïts i predats presents a la planta, entre d'altres (veure Materials i Mètodes). Donat que es recullen tots, en el següent moment del mostreig la planta haurà tornat a generar nous capítols. Podem suposar, doncs, que es dóna independència entre el nombre total de capítols recollits en cada seguiment realitzat.

Si es denota per TP_j la variable aleatòria *Nombre de capítols predats en la visita j-èssima*, variable aleatòria discreta no negativa, amb funció generatriu de probabilitats $G_2(z)$, aleshores la variable que recull el nombre total de capítols predats per planta es pot calcular com:

$$TP = TP_1 + TP_2 + ...TP_m,$$

on m representa el nombre de setmanes de seguiment propi de cada planta, d'on es desprèn que el nombre de seguiment també correspon a una variable aleatòria discreta no negativa. Si $G_1(z)$ és la funció generatriu de probabilitats de la variable *Nombre de seguiments*, aleshores la *fgp* de la distribució de la variable TP ve donada per:

$$E[z^{TP}] = E_m\left[E[z^{TP}|m]\right] = E_m[G_2(z)] = G_1(G_2(z)).$$

Si es denota per F_2 la funció de distribució de les variables de recompte i per F_1 la funció de distribució de la variable indicadora del nombre de seguiments que s'han realitzat, Gurland (1957) es refereix a aquest tipus de distribucions com a *generalitzades* i empra la següent notació:

$$TP \sim F_1 \bigvee F_2.$$

que es llegeix *distribució* F_1 *generalitzada per la distribució* F_2.

La distribució BN es representa com

$$Binomial\, negativa \sim Poisson(\mu) \bigvee Logaritmica(\theta).$$

Si assumim que les variables TP_j són iid amb distribució logarítmica de paràmetre θ, la seva *fgp* és igual a:

$$G_2(z) = \frac{ln(1-\theta z)}{ln(1-\theta)} -$$

Assumint que m segueix una distribució de Poisson de paràmetre μ independent de les TC_j li correspon la *fgp*

$$G_1(z) = e^{\mu(z-1)}.$$

Aleshores, la *fgp* de la variable TP serà igual a la composició de G_1 i G_2 i per tant serà igual a:

$$E\left[z^{TP}\right] = G_1\left(G_2(z)\right) = exp\left[\mu\left(\frac{ln(1-\theta z)}{ln(1-\theta)} - 1\right)\right] = \left(\frac{1-\theta z}{1-\theta}\right)^{\frac{-\mu}{ln(1-\theta)}},$$

que correspon a la *fgp* d'una distribució BN de paràmetres $\frac{-\mu}{ln(1-\theta)}$ i θ. Per tant, acabem de provar que la distribució BN és una distribució *Poisson-Stopped-Sum*.

Finalment, hi ha diverses formes de parametritzar la distribució BN, aquí s'ha triat la parametrització més emprada en el models lineals generalitzats (veure Hilbe, 2011). La funció de probabilitat de la distribució BN és la següent:

$$P_Y^*(Y=y) = \frac{\Gamma(y+k)}{\Gamma(k)+\Gamma(y+1)} \cdot \left(\frac{k}{\mu+k}\right)^k \cdot \left(1-\frac{k}{\mu+k}\right)^y, \quad y = 0,1,2,... \quad (5.7)$$

on μ i k són els paràmetres ambdós estrictament positius. $1/k$ es coneix com el paràmetre de dispersió d'una distribució BN donat que permet ajustar, en part, la sobredispersió de les dades. No s'ha de confondre aquest paràmetre amb el paràmetre de dispersió ϕ del model lineal generalitzat.

L'esperança i la variància amb aquesta parametrització corresponen respectivament a:

$$E[Y] = \mu, \qquad V[Y] = \mu + \frac{\mu^2}{k}$$

És fàcil comprovar que la parametrització proposta en l'equació (5.7) és equivalent a la parametrització més comuna per a definir la distribució BN que podem trobar, entre d'altres, a Johnson *et al.* (2005), i que és la següent:

$$P_Y^*(Y = y) = \binom{k+y+1}{k-1} \left(\frac{P}{Q}\right)^y \left(1 - \frac{P}{Q}\right)^k, \quad k > 0,\ 1 > P > 0,\ Q = 1 - P,$$

on P correspon a la probabilitat d'obtenir un èxit, Q és la probabilitat complementaria a l'èxit (fracàs) i k el nombre d'intents necessaris per a l'obtenció d'y èxits.

Per passar d'una parametrització a l'altra, només cal substituir P per $\frac{\mu}{k}$.

Com s'ha comentat anteriorment, la distribució BN té sobredispersió respecte una distribució de Poisson amb la mateixa esperança. Veiem-ho fent ús de la primera parametrització.

Siguin $Y \sim BN(\mu, k)$ i $X \sim Po(\mu)$ amb $E[Y] = \mu$, $E[X] = \mu$, $V[Y] = \mu + \frac{\mu^2}{k}$ i $V[X] = \mu$. Només cal veure que $V[Y] > V[X]$. Per això, s'ha de comprovar que,

$$V[Y] = \mu\left(1 + \frac{\mu}{k}\right) > \mu = V[X].$$

Ara bé, això és veritat atès que,

$$\frac{\mu}{k} > 0,$$

ja que la distribució BN es defineix per a dades no-negatives, $k > 0$ i $\mu > 0$.

En el cas de la distribució BN la probabilitat del zero ve definada per:

$$P_Y^*(Y_i = 0) = \left(\frac{k}{\mu_i + k}\right)^k \tag{5.8}$$

Aleshores, incorporant les equacions (5.7) i (5.8) a l'equació (5.3) s'obté que la funció de distribució de probabilitats P_Y d'un procés ZIBN és:

$$P_Y(y_i; k, \mu_i \mid y_i \geq 0) = \begin{cases} \pi + (1-\pi) \cdot \left(\frac{k}{\mu_i+k}\right)^k & y_i = 0 \\ (1-\pi) \cdot \frac{\Gamma(y_i+k)}{\Gamma(k)+\Gamma(y_i+1)} \cdot \left(\frac{k}{\mu_i+k}\right)^k \cdot \left(1 - \frac{k}{\mu_i+k}\right)^{y_i} & \text{si} \quad y_i > 0. \end{cases}$$

L'esperança i la variància d'un procés ZIBN venen donades per:

$$\begin{aligned} E[Y_i] &= \eta_i = (1-\pi)\cdot\mu_i \\ V[Y_i] &= \eta_i + \left(\frac{\pi}{1-\pi} + \frac{k}{1-\pi}\right)\cdot\eta_i^2. \end{aligned} \tag{5.9}$$

Observem que en el cas del procés ZIBN la sobredispersió prové de dos fonts independents $\frac{\pi}{1-\pi}$ i $\frac{k}{1-\pi}$, és més, els efectes són acumulatius (sempre i quan $\pi > 0$).

Les covariables del model s'introdueixen de la mateixa forma que s'ha vist en el model ZIP. La funció enllaç entre el valor esperat i les covariants segueix tenint sentit que sigui el logaritme. El paràmetre k s'assumeix constant per a totes les observacions, és a dir, no depèn de les covariants. Les probabilitats π_i depenen de les covariants seguint el model logístic expressat en l'equació (5.5) igual que en el model ZIP.

Un cop es té la funció de distribució, es pot trobar la funció de versemblança corresponent. En aquest cas, el logaritme de la funció de versemblança és:

$$\begin{aligned} l(\pi_i, k, \mu_i; y_i) = &\sum_{\substack{i \\ y_i=0}} log\left(\pi_i + (1-\pi_i)\cdot\left(\frac{k}{\mu_i+k}\right)^k\right) \\ &+ \sum_{\substack{i \\ y_i\neq 0}} log(1-\pi_i) + k\cdot log\left(\frac{k}{\mu_i+k}\right) + y_i\cdot log\left(\frac{\mu_i}{\mu_i+k}\right) \\ &+ log(\Gamma(y_i+k)) - log(\Gamma(k)) - log(\Gamma(y_i+1)). \end{aligned} \tag{5.10}$$

Incorporant les equacions (5.4) i (5.5), derivant i igualant a zero es pot trobar el màxim d'aquesta funció que dóna els estimadors màxims versemblants dels paràmetres del model. Com en el model ZIP, s'ha programat l'equació (5.10) mitjançant el procediment *NLMIXED* de SAS v9.22 per trobar l'estimador màxim-versemblant.

5.3. Proves per detectar l'excés de zeros

Existeixen diferents test d'hipòtesis per provar l'existència d'excés de zeros a les dades sota la hipòtesi nul·la que les dades segueixen una distribució de Poisson. Aquests és formulen de la forma següent:

$$\begin{aligned} H_0 &: \quad \pi = 0 \\ H_1 &: \quad \pi \neq 0. \end{aligned}$$

Rebutjar la hipòtesis nul·la implica que el model de Poisson no és adequat però no vol dir que un model ZIP o ZIBN sigui el correcte.

El-Shaarawi (1985) proposa fer servir els dos estadístics següents per tal de contrastar la hipòtesi nul·la anterior:

$$C\,test: \quad \frac{n_0 - Ne^{-\overline{y}}}{\left[Ne^{-\overline{y}}(1-e^{-\overline{y}}) - \overline{y}e^{-\overline{y}}\right]^{\frac{1}{2}}}$$

$$R\,test: \quad \frac{n_0 - N\left[\frac{N-1}{N}\right]^T}{\left[N\left[\frac{N-1}{N}\right]^T - N^2\left[\frac{N-1}{N}\right]^{2T} + N(N-1)\left[\frac{N-2}{N}\right]^T\right]^{\frac{1}{2}}},$$

on n_0 és el nombre d'observacions amb valor igual a zero, N és el nombre d'observacions totals, $\overline{y}$ és la mitjana de totes les observacions i T és la suma de totes les observacions. Per a N gran la distribució d'aquests estadístics és Normal amb mitjana 0 i desviació 1.

5.4. Models considerats en aquest treball

Definim la següent notació per a les variables que intervenen en la modelització. Denotem per:

- TC: la variable *Nombre Total de Capítols Produïts*,
- Esp: el factor *Espècie* i,
- Loc: el factor *Localització*

Donada la variable resposta principal *Nombre Total de Capítols Predats*, atès que les covariables són *Espècie* i *Localització* i que l'*offset* correpon a la variable *Nombre Total de Capitols Produïts*, els models que s'han implementat en aquest treball són els següents:

$$\text{Model 1:} \quad log(E[Y_i^r|\mathbf{X}]) = \beta_0 + \beta_1 Esp_i + log(TC_i^r)$$

$$\text{Model 2:} \quad log(E[Y_{ij}^r|\mathbf{X}]) = \beta_0 + \beta_1 Esp_i + \beta_2 Loc_j + log(TC_{ij}^r),$$

on $i = 1 \div 4$ i $j = 1 \div 6$. En el cas dels models P1 i ZIP1 $r = 1 \div n_i$ i, en els models P2 i ZIP2 $r = 1 \div n_{ij}$. Y_i^r és el nombre observat de capítols predats de la r-èssima planta de l'espècie i-èssima que ha produït un nombre total de capítols igual a TC_i^r. Pel Model 2, Y_{ij}^r és el nombre observat de capítols predats de la r-èssima planta de l'espècie i-èssima a la localització j-èssima que sabem ha produït un nombre total de capítols igual a TC_{ij}^r. $\mathbf{X}$ denota el conjunt de covariables.

Les distribucions de probabilitat considerades són la distribució Poisson, la distribució BN, la distribució ZIP i la distribució ZIBN.

Poisson: $Y \sim Po(\mu)$
Binomial negativa: $Y \sim BN(k, \mu)$
Poisson Zero-Inflat: $Y \sim ZIP(\pi, \mu)$
Binomial Negatiu Zero-Inflat: $Y \sim ZIBN(\pi, k, \mu)$.

La combinació dels dos models juntament amb les quatre distribucions proposades implica l'ajust de 8 models diferents. A més, per a les distribucions zero inflades es consideraran els següents dos models per a l'excés de zeros:

$$\text{Model 3:} \quad log\left(\frac{\pi}{1-\pi}\right) = \gamma_0$$

$$\text{Model 4:} \quad log\left(\frac{\pi_i}{1-\pi_i}\right) = \gamma_0 + \gamma_1 \cdot Esp_i, \text{ on } i = 1 \div 4$$

Per tant, en total s'ajusten 12 models diferents: 2 amb distribució de Poisson, 2 amb distribució BN, 4 amb distribució ZIP i 4 amb distribució ZIBN.

El paper del factor *Espècie* és d'especial interès ja que, com s'ha esmentat al Capítol 3 les espècies són força diferents entre elles. És d'esperar que les espècies autòctones presentin un dany més elevat que les espècies exòtiques donat que els insectes estan adaptats a aquestes espècies. Sobre la localització on van ser recollides les dades, d'entrada no s'haurien d'observat diferències entre les diferents zones de mostreig. No obstant, el Capítol 4 recull que la predació a les sis localitzacions és estadísticament diferent (proves Khi-quadrat). Al següent capítol es presenten els resultats d'ajustar aquests 12 models i es treuen conclusions sobre la importància de les dues covariants respecte la predació.

Capítol 6
Resultats obtinguts

En aquest capítol es presenten els resultats obtinguts a l'ajustar els models enunciats en l'Apartat 5.4 del capítol anterior. Tots els paràmetres dels models ajustats s'han estimat fent servir el mètode de la màxima versemblança. A les equacions (5.6) i (5.10) es troben les funcions corresponents als models zero inflats que s'han maximitzat mitjançant el mètode de Newton-Raphson.

Per a comparar els models estimats, es presenta el valor del logaritme de la funció de versemblança, la Deviança, el criteri d'informació d'Akaike (AIC) i el criteri d'informació bayesià (BIC).

En l'anàlisi estadística s'ha utilitzat el software SAS System v.9.22, SAS Institute Inc., Cary, NC, USA. Les decisions estadístiques s'han portat a terme fixant com a nivell de significació el valor 0,05. El codi implementat es pot trobar a l'Apèndix A.

6.1. Criteris de bondat d'ajust

S'han emprat tres criteris diferents per a comparar la bondat d'ajust dels diferents models ajustats. Aquests criteris serveixen per a seleccionar entre models no necessàriament aniuats. Es basen en la comparació de la log-versemblança i alguns penalitzen aquells models amb major nombre de paràmetres. Els criteris emprats són els següents:

Deviança (McCullagh i Nelder, 1989): Quan es treballa amb models lineals generalitzats és útil disposar d'una quantificació que pugui ser interpretada com la generalització de les sumes de quadrats dels residus quan es fa servir mínims quadrats ordinaris amb variable resposta Normal. Aquesta quantitat és la *deviança* del model i es calcula com:

$$D(y) = 2\left[ln(L_{max}) - ln(L)\right] \cdot \phi,$$

on L_{max} denota el màxim de la funció de versemblança per al model saturat (model que té tants paràmetres com observacions) i L és el màxim de la funció de versemblança per al model proposat. ϕ és el paràmetre d'escala, en el cas del model de Poisson aquest és fixe i igual a 1 i pel model Normal aquest és igual a la variància. El model BN només és família exponencial si el paràmetre k associat a la dispersió de la distribució BN és considerat fix. En el cas dels models ZIP i ZIBN no està definida la Deviança donat que es tracten de models de mixtures de distribucions i no es consideren models lineals generalitzats. Per tant, per als models zero-inflats no es posible realitzar el càlcul d'aquest estadístic.

Criteri d'informació d'Akaike (Akaike, 1974): El *criteri d'informació d'Akaike* (AIC) és una mesura de la qualitat relativa d'un model estadístic per a un conjunt de dades. L'AIC ofereix una estimació relativa de la informació que es perd quan es fa servir un model determinat per a representar el procés que genera les dades. Es calcula com:

$$AIC = -2 \cdot ln(L) + 2 \cdot p,$$

on p és el número de paràmetres en el model estadístic i L és el màxim valor de la funció de versemblança del model.

Criteri d'informació bayesià (Schwarz, 1978): El *criteri d'informació bayesià* (BIC) es sustenta, igual que els anteriors, en la funció de versemblança i està estretament relacionat amb l'AIC. Com l'AIC introdueix un terme de penalització que té en compte el nombre de paràmetres en el model, el terme de penalització és més gran en el BIC que en l'AIC

Es calcula com:

$$BIC = -2 \cdot ln(L) + p \cdot ln(n),$$

on p és el número de paràmetres en el model estadístic, n el nombre d'observacions i L és el màxim valor de la funció de versemblança del model.

6.2. Test C i Test R per a detectar l'excés de zeros

En el capítol anterior s'han enunciat aquestes dues proves per contrastar si existeix excés de zeros. En aquest cas, el càlcul de l'estadístic C ha donat un valor de 539,94 i el resultat de l'estadístic R ha estat molt similar, 542,10. Donat que la distribució que segueixen aquests dos estadístics és la distribució normal estàndard, fixat un nivell de significació de 5%, el valor de contrast d'una distribució Normal per una prova d'hipòtesis bilateral és 1,96, per tant, tant l'estadístic C com l'estadístic R estarien indicant que es rebutja la hipòtesis nul·la de que *No hi ha excés de zeros.*

Aquests tests doncs, indiquen la presencia d'un excés de zeros i descartarien el model de Poisson. No obstant, això no vol dir que el model correcte correspongui al model ZIP o ZIBN.

Tot i que aquestes proves descarten d'entrada el model de Poisson, no està de més ajustar-lo per comparar-lo amb la resta de models ajustats i constatar el resultat d'aquesta secció.

6.3. Models clàssics i inflats amb π constant

Els primers models que es presenten són el model clàssic, Poisson o Binomial Negatiu, i es comparen amb els models inflats amb π constant. Això vol dir que s'assumeix que no intervenen variables en la modelització de la probabilitat del zero. És a dir, pot existir excés de zeros però no s'explica a partir de cap covariable observada en l'estudi.

6.3.1. Model Poisson i Model ZIP

En els models d'aquesta secció s'assumeix que la distribució de la variable resposta *Nombre Total de Capítols Predats* segueix una distribució de Poisson o bé una distribució Poisson Zero Inflada. En base a aquestes distribucions i donades les covariables en estudi, s'ajusten els models de la Taula 6.1.

Model P1:	$\mu_i^r = e^{\beta_0+\beta_1\cdot Esp_i+ln(TC_i^r)}$	
Model P2:	$\mu_{ij}^r = e^{\beta_0+\beta_1\cdot Esp_i+\beta_2\cdot Loc_j+ln(TC_{ij}^r)}$	
Model ZIP1:	$\mu_i^r = e^{\beta_0+\beta_1\cdot Esp_i+ln(TC_i^r)}$	$\pi = \frac{e^{\gamma_0}}{1+e^{\gamma_0}}$
Model ZIP2:	$\mu_{ij}^r = e^{\beta_0+\beta_1\cdot Esp_i+\beta_2\cdot Loc_j+ln(TC_{ij}^r)}$	$\pi = \frac{e^{\gamma_0}}{1+e^{\gamma_0}}$

TAULA 6.1. Models ajustats per a explicar la variable *Nombre Total de Capítols Predats* assumint distribució Poisson o Poisson Zero Inflada.

on $i = 1 \div 4$ i $j = 1 \div 6$. En el cas dels models P1 i ZIP1 $r = 1 \div n_i$ i, en els models P2 i ZIP2 $r = 1 \div n_{ij}$.

La Taula 6.2 conté les estimacions dels paràmetres dels quatre models formulats en les línies anteriors, així com les seves desviacions típiques entre parèntesis. La categoria de referència per al factor *Espècie* és *SV*. La categoria de referència per al factor *Localització* és *VF*. En negreta s'assenyalen els paràmetres estadísticament significatius.

Respecte al factor principal d'interès, l'espècie a la qual pertany la planta, es veu que és significativa en tots quatre models i presenta un comportament similar en tots ells. En els models amb distribució Poisson, P1 i P2, s'observa que el valor esperat de predació en les espècies *S. Inaequidens* i *S. Pterophorus* és inferior a la predació esperada en l'espècie de referència *S. Vulgaris*. Per contra, l'estimació de la predació per l'espècie *S. Lividus* resulta molt superior al *S. Vulgaris*. Aquestes

	Models			
Paràmetre	P1	P2	ZIP1	ZIP2
Intercept	**-2,322 (0,051)**	**-2,029 (0,053)**	**-1,709 (0,054)**	**-1,502 (0,054)**
Espècie				
SI	**-0,690 (0,058)**	**-0,863 (0,059)**	**-1,253 (0,060)**	**-1,401 (0,061)**
SL	**1,208 (0,058)**	**1,295 (0,058)**	**0,760 (0,059)**	**0,860 (0,061)**
SP	**-2,973 (0,070)**	**-2,955 (0,071)**	**-0,537 (0,072)**	**-0,667 (0,075)**
SV	0	0	0	0
Localització				
CB		**-0,593 (0,062)**		**-0,176 (0,063)**
CP		**-0,927 (0,079)**		**-0,623 (0,082)**
CT		**-0,426 (0,065)**		**-0,453 (0,067)**
FM		**0,236 (0,054)**		**0,398 (0,055)**
SS		**-1,093 (0,065)**		**-1,034 (0,066)**
VF		0		0
Zero Model				
Intercept			-0,061 (0,102)	-0,100 (0,103)
$\widehat{\pi}$			**0,485 (0,025)**	**0,475 (0,026)**

TAULA 6.2. Estimació (desviació estàndard) dels paràmetres dels 4 models de la Taula 6.1 emprats per a modelar la variable *Nombre Total de Capítols Predats*. La part superior correspon als paràmetres del model per a l'esperança. La part central al model per a π i, la línea inferior correspon a la estimació de la probabilitat π.

estimacions concorden bastant amb allò observat a la Taula 4.8 on, si es mira la taxa mitjana de predació, es veu que l'espècie *S. Lividus* és la que presenta la taxa de predació més elevada, seguida del *S. Inaequidens* i a continuació el *S. Vulgaris* i el *S. Pterophorus*. Aquest comportament s'observa en els quatre models ajustats.

En els models ZIP no obstant, hi ha hagut un lleuger canvi en les magnituds. La diferència en la mitjana esperada de predació entre l'espècie de referència *S. Vulgaris* i l'espècie *S. Pterophorus* s'ha reduït molt, mentre que augmenta la diferència entre l'espècie de referència respecte l'espècie *S. Inaequidens*. De fet, ara és aquesta espècie, el *S. Inaequidens*, la que presenta la predació esperada més petita. L'espècie *S. Lividus*, continua al capdavant com a espècie amb predació esperada més elevada. Aquests resultats concorden amb les mitjanes de la taxa de predació sense zeros recollides a la Taula 4.14 del Capítol 4.

Respecte al factor *Localització*, aquest ha resultat significatiu indicant l'existència de diferències estadísticament significatives entre les diferents zones de mostreig. Malgrat aquest resultat, la comparació de zones de mostreig no forma part de les hipòtesis de treball donat que, tot i existir diferències, és d'esperar que aquestes repercuteixin de forma similar en les plantes avaluades ja que es van mostrejar totes les espècies en totes les localitzacions. Per explicar les diferències entre les diferents localitzacions caldria disposar de dades referides a la meteorologia local de la zona (precipitació per m^2, factor ultravioleta, exposició hores de sol, etc) i disposar de les dades geogràfiques (latitud, longitud, pendent, orientació, etc). Sobre les dades meteorològiques hi ha molt poca informació disponible donat que, al tractar-se de

zones molt properes entre sí, no es disposa de dades tan individualitzades. Respecte a les dades del terreny, sí hauria estat possible obtenir-ne alguna informació però aquesta no va ser recollida. Finalment, també caldria disposar de la informació relativa a les colònies d'insectes presents a les diferents localitzacions, ja que aquests mostren preferències per algunes espècies en detriment d'altres.

Si s'observa el valor del terme independent de la part zero-inflada del model ZIP, es veu que aquest no ha resultat significatiu. No obstant això, sí que es troba un excés de zeros significatiu, quantificable en un 48% aproximadament. Si el terme independent no és significatiu, aquest es pot considerar igual a 0, aleshores, subtituint a l'equació del model (Taula 6.1) s'obté una estimació per π del 50% que efectivament concorda amb l'estimació proporcionada pel model.

La Taula 6.3 conté les mesures de bondat d'ajust enunciats anteriorment per als quatre models ajustats.

	Models			
Criteri	P1	P2	ZIP1	ZIP2
Log likelihood	-3451,97	-3154,56	-1826,52	-1587,08
Deviance	6098,63	5503,80	–	–
AIC	6911,94	6327,12	3663,03	3194,15
BIC	6928,60	6364,59	3683,85	3235,79
graus de llibertat	471	466	470	465

TAULA 6.3. Estadístics de bondat d'ajust dels 4 models ajustats per a modelar la variable *Nombre Total de Capítols Predats*.

Donat que el model de Poisson no està aniuat dins el model ZIP (Greene, 1994) es comparen els dos models de Poisson i els dos models ZIP. En tots dos casos, en base als estadístics de bondad d'ajust, es veu que incloure la covariable *Localització* millora l'ajust. En el cas dels models de Poisson, millora en un 8,62% la log-versemblança, en un 9,75% la deviança i en un 8,46% i 8,14% l'AIC i BIC, respectivament. En els models ZIP aquesta reducció és major i suposa un 13,11% en la log-versemblança i la deviança i d'un 12,08% i 12,16% en el AIC i BIC, respectivament.

Finalment, s'enumeren a continuació les principals diferències detectades en avaluar els models Poisson i ZIP amb una i dos covariants i π constant.

1) La inclusió de la localització no canvia gaire les estimacions del terme independent i dels coeficients del factor principal d'interés l'*Espècie*. Això s'observa tant pel model Poisson com per al model ZIP.
2) Diferències importants en incloure el paràmetre π. Els coeficients associats al factor *Espècie* es redueixen significativament, tant si es consideren dues covariants com si només se'n cosidera només una. És a dir, la tendència en el factor *Espècie* és la mateixa però les magnituds canvíen de forma important: SP i SV passen a ser més semblans i SI i SV més diferents.
3) La millora substancial s'obté al canviar de distribució, no a l'incloure el factor *Localització* com a covariant.

6.3.2. Model Binomial Negatiu i Model ZINB

S'ajusten a continuació els mateixos models de la Taula 6.1 però substitutint la distribució de Poisson per la distribució Binomial Negativa. Com abans, s'ajusten quatre models aniuats dos a dos.

A les Taules 6.4 i 6.5 es troben les estimacions dels paràmetres dels 4 models ajustats i els estadístics de bondat d'ajust, respectivament. La categoria de referència per al factor *Espècie* és *SV*. La categoria de referència per al factor *Localització* és *VF*. En negreta els paràmetres estadísticament significatius.

	Models			
Paràmetre	BN1	BN2	ZIBN1	ZIBN2
Intercept	**-3,346 (0,193)**	**-2,883 (0,244)**	**-2,577 (0,170)**	**-2,048 (0,184)**
Espècie				
SI	**0,621 (0,276)**	0,456 (0,291)	-0,010 (0,208)	-0,230 (0,212)
SL	**1,917 (0,243)**	**1,943 (0,252)**	**1,440 (0,188)**	**1,461 (0,187)**
SP	**-0,608 (0,289)**	**-0,672 (0,325)**	0,251 (0,293)	0,264 (0,294)
SV	0	0	0	0
Localització				
CB		-0,421 (0,311)		**-0,458 (0,195)**
CP		**-0,747 (0,319)**		**-0,790 (0,203)**
CT		**-0,654 (0,304)**		**-0,728 (0,201)**
FM		-0,1742 (0,316)		-0,341 (0,199)
SS		**-1,095 (0,302)**		**-1,086 (0,193)**
VF		0		0
Zero Model				
Intercept			**-0,443 (0,134)**	**-0,405 (0,128)**
Dispersió $1/k$	3,324 (0,307)	3,124 (0,293)	0,787 (0,114)	0,603 (0,089)
$\widehat{\pi}$			**0,391 (0,032)**	**0,400 (0,031)**

TAULA 6.4. Estimació (desviació estàndard) dels paràmetres dels 4 models ajustats per a modelar la variable *Nombre Total de Capítols Predats*. La part superior correspon als paràmetres del model per a l'esperança. La part central correspon als paràmetres del model per π i, les línies inferiors corresponen a les estimacions dels paràmetres $1/k$ i π.

Els models amb distribució BN introdueixen un paràmetre més per controlar la sobredispersió de les dades. En aquest cas, s'observa que aquest paràmetre pren un valor molt elevat en el cas que no es modeli la probabilitat del zero i, força menys elevat quan es modela aquesta. Això és degut a que part de l'excés de variabilitat contingut en les dades prové de mostrejar la població constant igual a zero i la població amb distribució BN. Un cop distingides les poblacions, l'excés de variabilitat que ha d'adaptar el paràmetre de la BN és menor. Com es va veure al capítol anterior, la variància dels models ZIBN es descompon en aquests dos efectes additius. Això implica també que alguns paràmetres deixin de ser estadísticament significatius.

Per al model BN1, es veu que la taxa més elevada de predació correspon a l'espècie *S. Lividus*, seguida de l'espècie *S. Inaequidens*, l'espècie de referència *S. Vulgaris* i finalment l'espècie *S. Pterophorus*. Aquest resultat concorda amb allò observat a les dades (Taula 4.8). Per a aquest mateix model, quan s'inclou el factor *Localització*, l'efecte de l'espècie *S. Inaequidens* deixa de ser significatiu i, per tant, s'equipara a l'espècie de referència *S. Vulgaris*. De les zones de mostreig seleccionades, apareixen diferències estadísticament significatives entre les localitzacions *CP*, *CT* i *SS* respecte la localització de referència *VF*. No obstant això, i com s'ha comentat anteriorment, comparar les diferents zones de mostreig no és l'objecte principal d'aquest estudi.

En els models ZIBN el paràmetre associat a la dispersió és considerablement més petit que en els models BN, això vol dir que part de la variabilitat de les dades, que en el model BN es carrega al paràmetre de dispersió, és explicada pel paràmetre π en el model ZIBN tal com s'ha explicat anteriorment. A més es detecta significació estadística en el terme independent per al model de la part dels zeros. Sembla doncs que la sobredispersió de les dades prové en gran part, de l'excés de zeros en les mateixes. Tant el model amb una covariant com el de dos, estimen en un 40% d'excés de zeros. Finalment, en aquests models ZIBN es perd la significació de diversos paràmetres associats al factor *Espècie*, trobant únicament diferències entre l'espècie *S. Lividus*, que presenta la taxa de predació més elevada, i l'espècie *S. Vulgaris*.

	Models			
Criteri	BN1	BN2	ZIBN1	ZIBN2
Log Likelihood	-986,01	-978,01	-959,35	-941,70
Deviance	398,64	399,75	–	–
AIC	1982,03	1976,02	1930,69	1905,40
BIC	2002,84	2017,65	1955,67	1951,19
graus de llibertat	471	466	470	465

TAULA 6.5. Estadístics de bondat d'ajust dels 4 models ajustats per a modelar la variable *Nombre Total de Capítols Predats*.

Si es comparen els dos models BN i els dos models ZIBN en base als valors de l'AIC, el BIC i la versemblança, no hi ha unes diferències tant evidents com en el cas dels models amb distribució de Poisson. Concretament, diriem que modificar la probabilitat del zero en el model BN és menys rellevant que modificar-la en el model de Poisson. Això es degut a que certa dispersió de les dades és absorvida pel paràmetre de dispersió de la BN. En base a aquests estadístics, el model a seleccionar seria el model ZIBN amb *Espècie* i *Localització* com a factors fixes.

Finalment, s'enumeren les principals diferències detectades en avaluar els models BN i ZIBN amb una i dos covariants i π constant:

1) En tots els models, clàssics i zero inflats, passar d'assumir resposta Poisson a assumir resposta BN suposa una millora molt important de l'ajust.
2) En el model BN afegir la covariable *Localització* pràcticament només afecta al terme independent del model que s'incrementa, ja que les estimacions dels

coeficients de les espècies pràcticament no varien. També fa que l'espècie *S. Inaequidens* passi de ser significativament diferent de l'espècie *S. Vulgaris* a no ser-ho.

3) Quan s'inclou el paràmetre π, l'única espècie que passa a ser significativament diferent de la resta és la *S. Lividus* tant si es té en compte la localització en el model com si no. Les diferències que hi havien entre la resta d'espècies passen a desapareixer en estimar la probabilitat del zero a part.
4) El paràmetre de dispersió en els models ZIBN és molt més petit que en els models BN. Aquest paràmetre disminueix en considerar resposta ZIBN com a conseqüència de que el paràmetre π recull part important de la dispersió de les dades.

6.4. Models amb π depenen de l'espècie

En aquest apartat s'introdueix el factor *Espècie* com a covariable dins del model per ajustar la probabilitat del zero. Els nous models ajustats són els següents:

Model ZIP3:	$\mu_i^r = e^{\beta_0+\beta_1\cdot Esp_i+ln(TC_i^r)}$	$\pi_i = \frac{e^{\gamma_0+\gamma_1\cdot Esp_i}}{1+e^{\gamma_0+\gamma_1\cdot Esp_i}}$
Model ZIP4:	$\mu_{ij}^r = e^{\beta_0+\beta_1\cdot Esp_i+\beta_2\cdot Loc_j+ln(TC_{ij}^r)}$	$\pi_i = \frac{e^{\gamma_0+\gamma_1\cdot Esp_i}}{1+e^{\gamma_0+\gamma_1\cdot Esp_i}}$
Model ZIBN3:	$\mu_i^r = e^{\beta_0+\beta_1\cdot Esp_i+ln(TC_i^r)}$	$\pi_i = \frac{e^{\gamma_0+\gamma_1\cdot Esp_i}}{1+e^{\gamma_0+\gamma_1\cdot Esp_i}}$
Model ZIBN4:	$\mu_{ij}^r = e^{\beta_0+\beta_1\cdot Esp_i+\beta_2\cdot Loc_j+ln(TC_{ij}^r)}$	$\pi_i = \frac{e^{\gamma_0+\gamma_1\cdot Esp_i}}{1+e^{\gamma_0+\gamma_1\cdot Esp_i}}$

TAULA 6.6. Models ajustats per a explicar la variable *Nombre Total de Capítols Predats* amb probabilitat del zero depenent de l'espècie.

on $i = 1 \div 4$ i $j = 1 \div 6$. En el cas dels models ZIP31 i ZIBN3 $r = 1 \div n_i$ i, en els models ZIP4 i ZIBN4 $r = 1 \div n_{ij}$.

Les Taules 6.7 i 6.8 contenen, respectivament, les estimacions dels paràmetres i els estadístics de bondat d'ajust dels models ZIP i ZIBN que incorporen la covariable *Espècie* en el model per a la probabilitat del zero. La categoria de referència per al factor *Espècie* és *SV*. La categoria de referència per al factor *Localització* és *VF*. En negreta els paràmetres estadísticament significatius.

Si es compara el model ZIP1 (Taula 6.2) amb el model ZIP3, on la innovació es troba en introduir *Espècie* en el model per π, les estimacions referides al procés de recompte no varien gaire. L'espècie més predada continua sent SL, seguida de SV, SP I SI. Si es comparen ara els models ZIP2 (Taula 6.2) i ZIP4, es veu que tampoc hi han gaires diferencies en les estimacions dels paràmetres per l'esperança. El que han permès els models ZIP3 i ZIP4 és distingir les diferències en l'excés de zeros entre les diferents espècies.

	Models			
Paràmetre	ZIP3	ZIP4	ZIBN3	ZIBN4
Intercept	**-1,671 (0,052)**	**-1,465 (0,053)**	**-2,247 (0,214)**	**-1,740 (0,213)**
Espècie				
SI	**-1,297 (0,059)**	**-1,446 (0,060)**	-0,469 (0,244)	**-0,675 (0,238)**
SL	**0,719 (0,058)**	**0,819 (0,060)**	**0,993 (0,228)**	**1,085 (0,217)**
SP	**-0,572 (0,071)**	**-0,702 (0,074)**	0,117 (0,341)	0,098 (0,326)
SV	0	0	0	0
Localització				
CB		**-0,173 (0,063)**		**-0,399 (0,198)**
CP		**-0,641 (0,083)**		**-0,872 (0,206)**
CT		**-0,444 (0,067)**		**-0,688 (0,204)**
FM		**0,404 (0,055)**		-0,345 (0,204)
SS		**-1,029 (0,066)**		**-1,076 (0,201)**
VF		0		0
Model en π				
Intercept	**1,288 (0,246)**	**1,232 (0,249)**	**0,784 (0,308)**	**0,813 (0,300)**
Espècie				
SI	**-3,180 (0,457)**	**-3,360 (0,505)**	**-3,743 (1,112)**	**-3,902 (1,182)**
SL	**-2,181 (0,309)**	**-2,113 (0,309)**	**-2,630 (0,524)**	**-2,244 (0,410)**
SP	0,363 (0,390)	0,408 (0,392)	0,824 (0,432)	0,805 (0,427)
SV	0	0	0	0
Dispersió $1/k$			**0,872 (0,145)**	**0,640 (0,106)**
$\widehat{\pi}_{SI}$	**0,131 (0,044)**	**0,106 (0,042)**	0,049 (0,051)	0,044 (0,048)
$\widehat{\pi}_{SL}$	**0,290 (0,038)**	**0,293 (0,038)**	**0,136 (0,054)**	**0,193 (0,048)**
$\widehat{\pi}_{SP}$	**0,839 (0,041)**	**0,838 (0,042)**	**0,833 (0,042)**	**0,834 (0,042)**
$\widehat{\pi}_{SV}$	**0,784 (0,042)**	**0,774 (0,043)**	**0,687 (0,066)**	**0,693 (0,064)**

TAULA 6.7. Estimació (desviació estàndard) dels paràmetres dels 4 models ajustats per a modelar la variable *Nombre Total de Capítols Predats*. La part superior correspon als paràmetres del model per a l'esperança. La part central correspon als paràmetres del model per π i, la part inferior correspon a les estimacions del paràmetre de dispersió de la distribució BN i de l'excés de zeros segons espècie.

Es compara a continuació el model ZIBN1 (Taula 6.4) amb el model ZIBN3. En aquest cas s'observa que les estimacions dels paràmetres del factor *Espècie* són més petites, disminuint així les diferències observades entre les estimacions. Per exemple, l'estimació per l'espècie SL passa de 1,44 a 0,99 i l'estimació per l'espècie de referència, SV, passa de -2,58 a -2,25, es retalla la diferència de 4,02 a 3,24. Disminueix la diferència observada, en termes de mitjana estimada, entre les espècies SL i SV, SL i SP i es manté aquesta diferència entre SL i SI. No obstant això, les significacions estadístiques no canvien.

Finalment, es comparen els models ZIBN2 (Taula 6.4) i ZIBN4. Respecte del factor *Localització*, no s'observen diferències rellevants en les estimacions. Les tendències són les mateixes i les magnituds molt semblants. Respecte del factor *Espècie* sí que hi ha canvis importants. El més rellevant és que l'estimació del paràmetre per l'espècie SI passa a ser significatiu, això és, es troben diferències estadístiques

significatives, en termes de mitjana estimada, entre l'espècie SI i la resta d'espècies que en el model ZIBN2 no s'havien observat. Com en el cas anterior, es retallen les diferències entre les estimacions dels paràmetres del factor *Espècie*. No obstant, es mantenen les diferències estadístiques significatives observades en el model ZIBN2 entre SL i SI, SV i SP en el model ZIBN4, i es troben de noves entre SI i la resta d'espècies. En realitat, ara només presenten la mateixa predació les espècies SP i SV.

Com en el cas dels models ZIP3 i ZIP4, el que permeten aquests models es poder distingir les diferències en la proporció estimada de zeros entre les diferents espècies.

La millora substancial d'introduir el factor *Espècie* com a covariant per modelar els zeros radica en poder diferenciar el percentatge de zeros estimat a cada espècie en concret. Els models ZIP3 i ZIP4 troben que aquest percentatge és significatiu per a totes les espècies. Estimen un 13,1% i un 10,6% d'excés de zeros, respectivament, per l'espècie SI. Aquest percentatge augmenta fins al 29% per l'espècie SL, a poc més del 83% per l'espècie SP i fins al 78%, aproximadament, per l'espècie SV, en tots dos models. Si es comparen aquests valors amb el percentatge observat de zeros (Taula 4.10) es veu que són molt propers. Aquest percentatge però, correspon a l'excés i per tant queda un petit percentatge de zeros que es modelen dins del procés de recompte.

Per als models ZIBN, no es troba un excés de zeros estadísticament significatiu en l'espècie SI. Per aquesta espècie tots els zeros es consideren zeros del procés de recompte. L'excés de zeros per l'espècie SL és del 13,6% en el model ZIBN3 i del 19,3% en el model ZIBN4. La inclusió del factor *Localització* fa incrementar l'excés de zeros per a aquesta espècie en particular. A la Taula 4.11 ja s'havia observat la falta d'homogeneitat del número de zeros al llarg de les localitzacions de mostreig. L'espècie SP continua sent aquella que presenta un excés més elevat de zeros, una mica més del 83% per a tots dos models. Finalment, l'espècie SV presenta al voltant d'un 69% d'excés de zeros. La Taula 4.10 recollia que tant per l'espècie SV com SP s'havia observat un 83% i 84% respectivament de zeros. Els models ZINB3 i ZINB4 estan indicant que per l'espècie SP tots els zeros són zeros falsos, alients al procés de recompte, mentre que per l'espècie SV, hi ha un petit percentatge de zeros propis del procés de recompte.

En termes generals, els models ZIP estimen un excés de zeros superior als models ZIBN. És a dir, la distribució BN atorga més pes al zero del procés de recompte que la distribució de Poisson.

	Models			
Criteri	ZIP3	ZIP4	ZIBN3	ZIBN4
Log likelihood	-1756,34	-1517,03	-893,52	-877,11
AIC	3528,67	3060,06	1805,04	1782,22
BIC	3561,98	3114,18	1842,51	1840,51
graus de llibertat	467	462	467	462

TAULA 6.8. Estadístics de bondat d'ajust dels 4 models ajustats per a modelar la variable *Nombre Total de Capítols Predats*.

Els valors observats als criteris de bondat d'ajust, mostrats a la Taula 6.8, deixen clar que fer servir una distribució ZIBN millora molt l'ajust respecte d'utilitzar la distribució ZIP.

Tal com s'ha fet anteriorment, enumerem les principals diferències detectades en els models ZIP i ZIBN amb una i dues covariants i π depenent de l'espècie.

1) Incloure el factor *Espècie* en l'ajust de la probabilitat del zero resulta estadísticament significatiu i implica canvis importants en l'estimació de l'excés de zeros respecte els models ZIP i ZIBN sense covariant en π. Ara es possible observar les diferències en les proporcions estimades de zeros entre espècies.
2) En el model ZIP la inclusió de la covariant *Espècie* en la probabilitat del zero no implica canvis importats en la resta de coeficients del model. Per contra, en el model ZIBN incloure la covariant *Espècie* en la modelització de π, implica que el coeficient de l'espècie SI passi de ser no significatiu a significatiu.
3) L'aportació de la covariant *Localització* és la mateixa que en els models anteriors.

6.5. Comparativa dels models

A continuació comparem els models P2, NB2, ZIP4 i ZIBN4 basant-nos en les probabilitat predites i els test de Vuong i Clarke.

Gràficament, es pot comparar la bondat d'ajust d'aquests models mitjançant la comparació de la probabilitat predita pel model amb la probabilitat observada per a un valor concret (Figura 6.1).

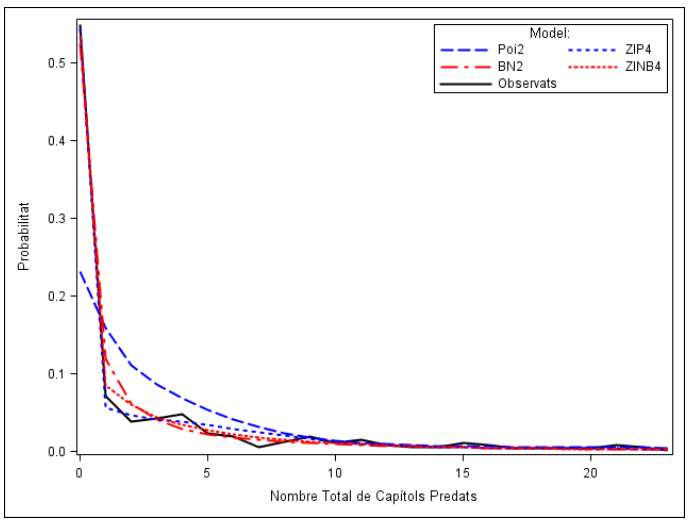


FIGURA 6.1. Gràfic de probabilitats estimades i empíriques

El recompte predit per a una observació donada és la corresponen estimació de μ per als models de Poisson i BN o de $\mu \cdot (1 - \pi)$ per als models ZIP i ZIBN. Per als models Poisson i BN, la probabilitat estimada per a un recompte donat y_i,

$p(y_i) = P(Y = y_i|x_i)$ s'obté directament de la funció de massa de probabilitat de la Poisson o BN, utilitzant la mitjana estimada per al recompte y_i, $\widehat{\mu}_i$. Per als models ZIP i ZIBN, la probabilitat predita d'un recompte donat y_i és: $I_0\widehat{\pi} + (1 - \widehat{\pi}) \cdot p(y_i)$ on $I_0 = 1$ si $y_i = 0$ i 0 en cas contrari, utilitzant la proporció de zeros estimada $\widehat{\pi}$,

Un cop trobades les probabilitats estimades per a totes les observacions, les probabilitats empíriques dibuixades a la Figura 6.1 s'han trobat com segueix: per $l = 1 \div 25$ s'ha calculat la mitjana de les probabilitats estimades en l sota el model corresponen, per a totes aquelles observacions que han estat iguals a l. Per exemple, pel model P2, s'ha calculat:

$$\frac{1}{\sum_{i=1}^{4}\sum_{j=1}^{6}\sum_{r=1}^{n_{ij}} I_{y_{ij}^r=l}} \cdot \sum_{i=1}^{4}\sum_{j=1}^{6}\sum_{r=1}^{n_{ij}} Pr(y_{ij}^r = l), \quad \text{on} \quad y_{ij}^r \sim Po(\widehat{\mu}_{ij}^r).$$

Es pot apreciar que el model de P2 subestima la probabilitat del zero, com indicaven els contrastos C i R per a detectar l'excés de zeros, el model de Poisson no és adequat per a aquest tipus de dades. El model ZIP4 arregla aquesta infraestimació però subestima les probabilitats dels primers recomptes per a continuació sobreestimar les probabilitats dels recomptes intermitjos. El model NB2 funciona prou bé, capta l'excés de zeros però sobreestima la probabilitat de l'1. Finalment, el model ZIBN4, capta correctament l'excés de zeros i és el que dóna unes estimacions dels primers recomptes més equilibrades sense sobreestimar en excés uns o subestimar uns altres.

Els criteris d'informació, com l'AIC, el BIC poden ser utilitzats per comparar els models no niats però no proporcionen una prova de comparació. En conseqüència no poden indicar si un model és significativament millor que un altre. Per comparar dos models niats que s'han ajustat mitjançant màxima versemblança s'utilitza la raó de versenblança (*Likelihood Ratio test*). Aquest estadístic es distribueix segon una Khi-quadrat de graus de llibertat igual a la diferència entre el nombre de paràmetres dels models. Aquesta prova no es pot utilitzar per comparar models que no són niats, per exemple models amb diferent funció de distribució.

Les proves de Vuong (Vuong, 1989) i Clarke (Clarke, 2007) són les adequades quan es volen comparar models no niats. La hipòtesi nul·la d'aquestes proves és que tots dos models són igualment distants del veritable model. La hipòtesi alternativa és bilateral i estableix que un dels models és més proper al model real. Si l'estadístic de la prova és positiu i significatiu, la hipòtesi nul·la es rebutja en favor de l'alternativa i, el primer model és més proper al veritable model. Si l'estadístic de la prova és negatiu i significatiu, la hipòtesi nul·la es rebutja en favor de l'alternativa i, el segon model és més proper al veritable model. Ambdues proves es basen en els criteris d'informació de Kullbach-Leibler (KLIC). El KLIC és una mesura de la *distància* entre el model objecte d'estudi i el model real. Ambdues suposen que els models que es comparen són estrictament no niats.

Test de Vuong (Vuong, 1989): Siguin L_1 i L_2 dos models amb p_1 i p_2 paràmetres respectivament. Denotem per β_{ML_1} i β_{ML_2}les estimacions màxim versemblants dels seus paràmetres. Suposis que $y_i, y_2, ..., y_n$ són les observacions obtingudes de la variable resposta Y, de forma que y_i s'ha observat sota les condicions experimentals x_i.

El test de Vuong es formula com segueix:

$$Z = \frac{\sum_{i=1}^{n} LR_i(\beta_{ML_1}, \beta_{ML_2})}{\sqrt{n} \times \omega},$$

on

$$LR_i(\beta_{ML_1}, \beta_{ML_2}) = \begin{cases} l_i & \text{sense correcció} \\ l_i - \frac{p_1 - p_2}{n} & \text{correcció d'Akaike} \\ l_i - \frac{p_1 - p_2}{2n} \cdot ln(n) & \text{correcció d'Schwarz,} \end{cases}$$

essent

$$l_i = l_{1_i} - l_{2_i} = ln \frac{L_1(y_i|x_i, \beta_{ML_1})}{L_2(y_i|x_i, \beta_{ML_2})}.$$

L_1 i L_2 corresponen als valors puntuals de la funció de versemblança quan els paràmetres són els màxim versemblants per als dos models. El terme del denominador de l'expresssió per Z es defineix fixant ω^2 igual a la variància empírica dels valors l_i.

Aquest estadístic Z segueix una distribució normal estàndard sota la hipòtesi nul·la de que els dos models disten de forma estadísticament equivalent del model real.

El ***Test de Clarke*** es formula de forma molt similar al test de Vuong. Realitza la prova no paramètrica dels Signes sobre els termes l_i per contrastar la hipòtesi nul·la de que la mediana d'aquestes diferències és igual a 0. És a dir,

$$\begin{cases} H_0 : Pr\left(ln \frac{L_1(y_i|x_i, \beta_{ML_1})}{L_2(y_i|x_i, \beta_{ML_2})} > 0\right) = 0,5 \\ H_1 : \neg H_0. \end{cases}$$

Els detalls es poden consultar a l'article d'en Clarke del 2007.

Si es comparen els models P2, NB2, ZIP4 i ZIBN4 en base a les proves Vuong i Clarke sense correcció, s'obtè que el model que està mès aprop del veritable model, tal com cabia esperar, és el model ZIBN4. La Taula 6.9 recull els resultats dels tests.

Comparativa	Vuong Test	Clarke Test	Model seleccionat
BN2 vs P2	2,747 (0,006)	71,5 (<,001)	BN2
ZIP4 vs P2	2,407 (0,016)	71,5 (<,001)	ZIP4
ZIBN4 vs BN2	7,049 (<,001)	114,5 (<,001)	ZIBN4
ZIBN4 vs ZIP4	3,520 (0,004)	8,5 (0,463)	ZIBN4

TAULA 6.9. Resultats de la prova de Vuong i Clarke sense correcció: estadístic i significació entre parèntesis.

Si s'ha de triar entre un model de recomptes amb distribució de Poisson o bé distribució Binomial Negativa, és aquesta última la més adient. Anàlogament, si es compara el model de Poisson amb el model ZIP, resulta millor el model ZIP. Si es compara el model Binomial Negatiu amb el model ZIBN, és el model ZIBN aquell que es considera més proper al model veritable, així com si es compara aquest model amb el model ZIP on també surt més proper al model veritable. No obstant, el test de Clarke per als dos últims models no surt significatiu i, per tant, consideraria que tots dos models són igualment propers al veritable model. Tot i això, donat que l'estadístic és positiu, es decanta pel primer model que correspon al model ZIBN.

6.6. Anàlisi de residus

Tots els models ajustats en aquest capítol es sustenten en distribucions de recomptes no normals. Aquest fet fa difícil establir com avaluar la bondat d'ajust dels mateixos fent servir les tècniques emprades en els models lineals clàssics: contrastar els residus contra la distribució normal. McCullagh i Nelder (pàg. 398, 1989) recomanen utilitzar els residus de la Deviança (*Deviance residuals*) per tal d'avaluar l'ajust de models lineals generalitzats, ja que aquests tenen propietats distribucionals que estan més a prop dels residus d'un model de regressió lineal Gaussià. De fet, la deviança coincideix amb la suma de quadrats residuals pel cas Normal. Els models ZI no es consideren model lineals generalitzats (Long, 1997), no obstant això, s'analitzaran els residus tenint en compte que no s'està buscant normalitat sinó detectar una possible falta d'ajust i la recerca de patrons en els residus.

Atès que la BN zero inflada no és familia exponencial de probabilitat, com s'ha comentat abans, l'anàlisi dels residus proporcionada en la bibliografia de models lineals generals no és aplicable aquí. Per portar a terme l'anàlisi dels residus s'ha decidit calcular els residus estandaritzats i dibuixar aquests en funció dels valors ajustats i de les covariants, per tal d'esbrinar si existeixen o no tendències o patrons.

A la vista dels resultats de l'apartat anterior, el model que millor ajusta les dades correspon al model ZIBN4 amb dos covariants i amb π depenen del factor *Espècie* i, per tant, s'analitzaran els residus associats a aquest model.

El nombre esperat de capítols predats d'una planta d'una espècie determinada en una localització determinada, que ha produït TC_{ij}^r capítols, es calcularà per mitjà de l'expressió

$$\widehat{\mu}_{ij}^r = e^{\widehat{\beta}_0 + \widehat{\beta}_1 Esp_i + \widehat{\beta}_2 Loc_j + ln(TC_{ij}^r)},$$

on $\widehat{\beta}_0$, $\widehat{\beta}_1$ i $\widehat{\beta}_2$ prenen els valors que figuren a la Taula 6.7 segons correpongui a cada espècie i localització.

A partir d'aquest valor estimat, els residus crus s'obtindran mitjançant la diferència entre y_{ij}^r i $\widehat{\eta}_{ij}^r$ on $\widehat{\eta}_{ij}^r$ correspon a l'esperança de la ZIBN que en l'equació (5.9) s'ha vist que es troba a partir de l'expressió $(1-\widehat{\pi}_i)\cdot\widehat{\mu}_{ij}^r$. A partir de $\widehat{\eta}_{ij}^r$ podem calcular la variància de y_{ij}^r tal com s'explicita a la mateixa equació (5.9), com:

$$\widehat{V}[y_{ij}^r] = \widehat{\eta}_{ij}^r + \left(\frac{\widehat{\pi}_i}{1-\widehat{\pi}_i} + \frac{\widehat{k}}{1-\widehat{\pi}_i} \right) \cdot \widehat{\eta}_{ij}^{r\,2},$$

essent $\widehat{\pi}_i$ les estimacions de les probabilitats corresponents a l'excés de zeros per a cada espècie, i $\widehat{k}$ el valor estimat del paràmetre k de la distribució BN.

Així doncs, els residus estandarditzats es calcularan com

$$e_{ij}^r = \frac{y_{ij}^r - \widehat{\eta}_{ij}^r}{\sqrt{\widehat{V}[y_{ij}^r]}}.$$

Per a validar aquest model, cal traçar els residus estandarditzats contra els valors ajustats així com contra cada variable explicativa i no s'hauria de detectar cap patró en els gràfics resultants. També és útil representar les dades originals en comparació amb les dades ajustades amb la intenció de veure una línia recta a la bisectriu del quadrant.

A la Figura 6.2 es mostren els residus estandarditzats envers els valors predits d'aquest model.

Donat que hi ha un total de 475 dades, si considerem que un 5% d'elles poden representar valors extrems o anòmals, això ens diu que no ens hem de preocupar gaire si hi ha al voltant de 25 valors que estan relativament lluny de zero. Això és el que s'observa en la Figura 6.2 (gràfic superior) on el nombre de residus fora de l'interval $(-2, 2)$ és igual a 15 i això representa menys del 5% del total d'observacions.

Pel cas que ens ocupa, les observacions anòmales corresponen majoritàriament a espècies molt predades atès que el valor observat és molt superior al predit pel model.

D'altra banda, donat que estem maximitzant la versemblança, és d'esperar que els residus estiguin centrats al voltant del zero. El que s'observa a la Figura 6.2 (gràfic superior) és que els residus semblen quedar per sota del valor zero. Fer un zoom de la regió on els valors predits ≤ 30 i residus ≤ 2, permet visualitzar millor aquest comportament. No obstant, l'ajust mitjançant un suavitzat lineal, revela que aquests residus estan centrats en 0 com era d'esperar. Afegir el factor *Espècie* per indentificar les observacions i, representar-les mitjançant el valor que pren la variable *Nombre Total de Capítols Predats* (gràfic inferior de la Figura 6.2), permet observar que els punts amb residus per sota del zero corresponen precissament a observacions amb valor igual a 0. El model quan realitza l'ajust atorga un valor positiu a l'estimació del valor i, per petit que sigui aquest valor, el residu quedarà sempre en negatiu.

Té sentit observar aquest patró de corbes exponencials negatives, per cada combinació d'espècie i localització, ja que, el valor predit $\widehat{\eta}_{ij}^r$ és igual a $(1-\widehat{\pi}_i) \cdot TC_{ij}^r \cdot e^{\widehat{\beta}_0 + \widehat{\beta}_1 Esp_i + \widehat{\beta}_2 Loc_j}$ i, si s'assumeix que els capítols produïts per cada espècie a cada localització és similar, això és un múltiple d'una exponencial negativa. El fet

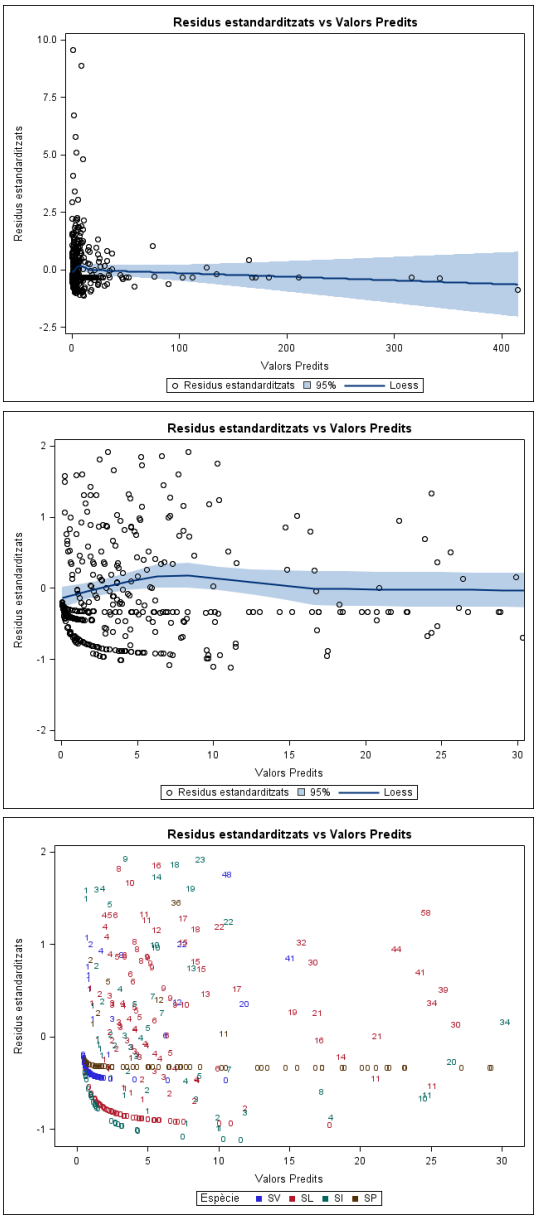


FIGURA 6.2. Diagrames de dispersió dels residus del model ZIBN4. A dalt, residus estandarditzats vers valors predits amb suavitzat lineal (*loess*). Enmig, zoom del gràfic superior. A sota, zoom del gràfic superior calorejant les observacions segons espècie de pertinença i identificant-les segons el valor observat.

de que es tracti d'una exponencial negativa es deriva de que les estimacions dels paràmetres són gairebé totes negatives (Taula 6.7).

A la Figura 6.3 es representen els valors predits vers els valors observats. A la dreta es troba un zoom dels primers valors d'aquest gràfic.

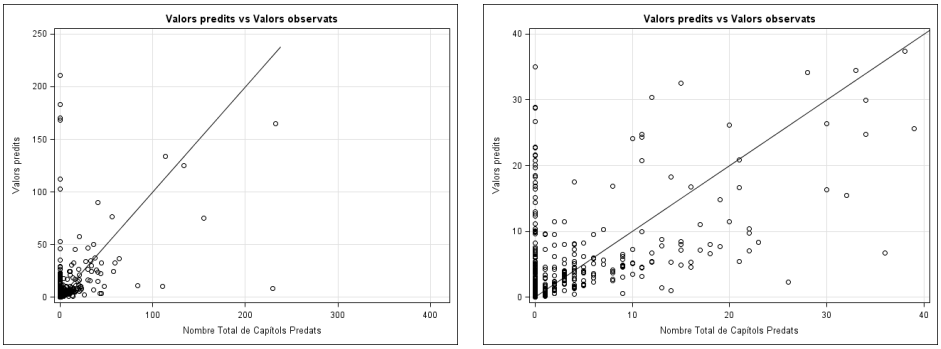


FIGURA 6.3. Diagrama de dispersió dels valors predits vers els valors observats. A l'esquerra, es mostren totes les observacions. A la dreta, hi figura el zoom dels primers valors. La recta correspon a la bisectriu.

El que s'observa a la Figura 6.3 és com, en general, els valors predits i els valors observats es disposen en la diagonal del gràfic tot i que els predits semblen anar sempre una mica per sota. El model estima una predació positiva superior a zero en moltes observacions on s'ha observat una predació igual a zero. Aquestes observacions corresponen a plantes de la família *S. Pterophorus* que presenten un nombre total de capítols produïts molt gran i, per tant, s'esperaria observar una predació positiva, però s'ha observat zero capítols predats.

Finalment, es poden realitzar els gràfics per contrastar els residus respecte les covariables del model i avaluar visualment una possible falta d'ajust. Donat que les dues covariables que intervenen són factors, s'utilitzen gràfics de caixa per a representar-les.

El més apreciable en els gràfics de la Figura 6.4 és la presència de residus molt elevats que indica la presència de valors anòmals. Aquests valors corresponen a casos partículars de plantes molt predades en determinades espècies i localitzacions. Per poder apreciar millor les caixes, s'ha fet un zoom d'aquestes tallant l'eix Y en el valor 2.

Dels gràfics sense observacions anòmales de la Figura 6.4 es pot apreciar el següent:

(1) Els residus estan centrats en zero
(2) La dispersió dels residus és més gran per a les espècies SL i SI atès que, per a les altres espècies el valor predit és molt proper a zero, independentment del número de capítols produïts per la planta ja que tenen més d'un 80% d'observacions iguals a zero (Taula 4.10).
(3) El comportament dels residus en les diferents àrees d'estudi és força més semblant essent SS la de mínima dispersió i CP la de màxima dispersió.

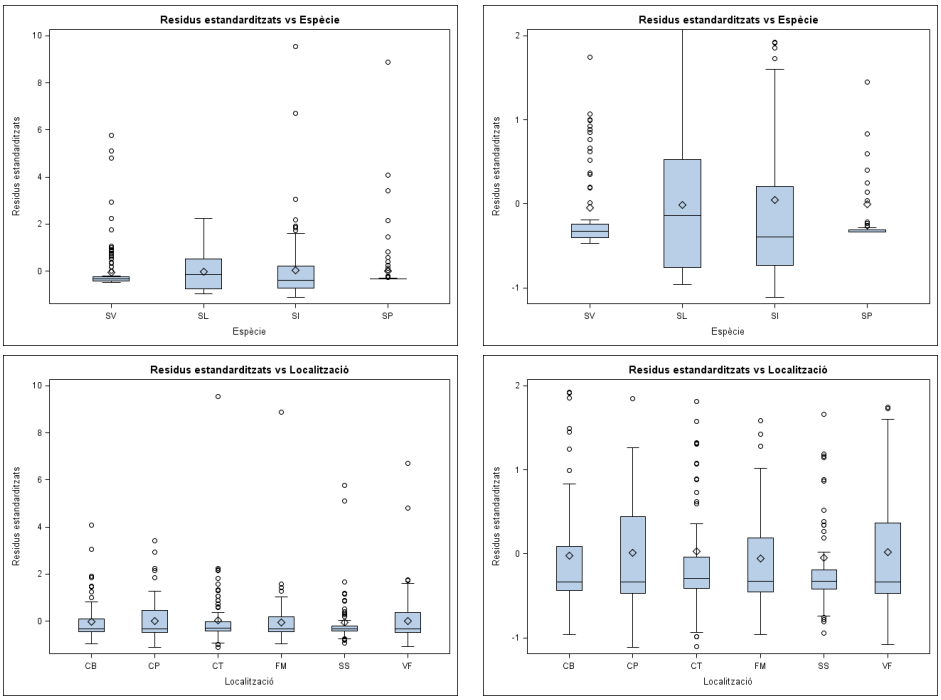


FIGURA 6.4. Diagrama de caixa dels residus estandarditzats segons les categories definides per les covariables. A la dreta, zoom de les caixes.

Capítol 7
Anàlisi mitjançant el model *Zipfian*

En el Capítol 4 s'ha vist que les dades estudiades presenten molta assimetria, una elevada probabilitat al primer valor, el zero, i una cua molt llarga que decreix lentament. Aquest patró pot associar-se a la llei de Zipf. En aquest capítol es presenta una alternativa als models zero inflats emprats en el capítol anterior fent ús d'aquesta distribució per ajustar les dades.

S'ajusten dos models diferents: un model amb només la covariant *Espècie* i un altre model amb *Espècie* i *Localització*. En tots dos models s'introdueix també la variable *Nombre Total de Capítols Produïts* com a offset. Per veure quin model proporciona un millor ajust es realitza el test de la raó de versemblances (LRT, *Likelihood Ratio Test*), donat que es tracten de models niats l'un en l'altre i atès que la Zipfian és família exponencial de probabilitat. Sota la hipòtesis nul·la, aquest estadístic es distribueix segon una χ^2 amb 5 graus de llibertat. Fixant el nivell de significació al 5%, el valor crític de contrast amb el qual s'ha de comparar l'estadístic és 11,07.

El model Zipfian s'ha estimat mitjançant el software R v.3.0.2. Les decisions estadístiques s'han portat a terme fixant com a nivell de significació el valor 0,05. El codi implementat es pot trobar a l'Apèndix A.

7.1. Introducció a la distribució Zipf

La distribució Zipf pertany a la família de distribucions de probabilitat de potències discreta (*Power Law*). La llei porta el nom del lingüista nord-americà George Kingsley Zipf (1902-1950), qui va proposar per primera vegada la llei Zipf (1935, 1949) quan estudiava la freqüència d'aparició de les paraules en un text, tot i que el taquígraf francès Jean-Baptiste Estoup (1868-1950) també es va adonar de la regularitat d'aquesta llei abans de Zipf, per exemple, Johnson *et al.* pàgina 527 (2005) esmenta aquesta llei com a *Zipf-Estoup law*.

Les principals característiques de la distribució Zipf són que té per suport els enters positius majors o iguals a 1, que la unitat presenta una probabilitat en general força elevada i, que aquesta probabilitat decreix ràpidament però generant una cua molt llarga. A més, si es representa gràficament el logaritme de la probabilitat empírica com a funció del logaritme del valor observat, el resultat és una línia recta.

Donada Y v.a. tal que $Y \sim Zipf(\alpha)$, la seva funció de probabilitat és igual a:

$$Pr(Y = y) = \begin{cases} \frac{y^{-\alpha}}{\zeta(\alpha)} & \text{si} \quad y = 1, 2, 3, 4, ... \\ 0 & altrament, \end{cases} \tag{7.1}$$

on $\alpha > 1$ i $\zeta(\alpha)$ és la funció Zeta de Riemann, és a dir:

$$\zeta(\alpha) = \sum_{k=1}^{\infty} k^{-\alpha},$$

que convergeix només si $\alpha > 1$.

Prenent logaritmes a totes dues bandes de l'equació (7.1) es comprova que el logaritme de la probabilitat és lineal respecte del logaritme del valor observat:

$$lnPr(Y = y) = -\alpha lny - ln\zeta(\alpha) \qquad y = 1, 2, 3, 4, ...$$

En quan als moments de la distribució, l'esperança de la Zipf és igual a:

$$E[Y] = \frac{\zeta(\alpha - 1)}{\zeta(\alpha)}.$$

En general, el moments d'ordre r de la distribució Zipf es calculen com:

$$E[Y^r] = \frac{\zeta(\alpha - r)}{\zeta(\alpha)},$$

d'on es desprèn que la variància, calculada a partir del primer i segon moment, és:

$$Var[Y] = \frac{\zeta(\alpha - 2)\zeta(\alpha) - \zeta(\alpha - 1)^2}{\zeta(\alpha)^2}.$$

L'esperança de la distribució Zipf només existirà per a valors d'$\alpha > 2$ i la variància per a valors d'$\alpha > 3$ donat que, com s'ha esmentat abans, la funció Zeta només convergeix per a valors d'α majors a 1. Així doncs, la distribució Zipf amb $\alpha \in (1, 2)$ existeix però no té ni esperança ni variància finites. Si $\alpha \in (2, 3)$, té esperança finita però no variància i, si $\alpha > 3$ existeixen tots els moments.

Per estimar el valor del paràmetre α de la distribució es poden fer servir diferents metodologies. Fent ús del mètode de la màxima versemblança, donada una mostra aleatòria simple $y_1, y_2, ...y_N$ d'una v.a. amb distribució Zipf(α), el logaritme de la funció de versemblança és:

$$l(\alpha; y_1, y_2, ...y_N) = -Nln\zeta(\alpha) - \alpha \sum_{i=1}^{N} lny_i.$$

Per estimar el paràmetre de la distribució s'ha programat una rutina en R que maximitza aquesta log-versemblança fent ús d'un algoritme d'optimització.

És important esmentar que

$$\frac{\zeta'(\alpha)}{\zeta(\alpha)} = E[log(Y)],$$

que existeix per tot $\alpha > 1$. El màxim versemblant coincideix amb el valor d'α que iguala els moments logarítmics teòric i empíric. Veieu-m'ho,

$$\begin{aligned}
&\frac{(\partial l)}{\partial \alpha} = -N\frac{\zeta'(\alpha)}{\zeta(\alpha)} - \sum_{i=1}^{N} lny_i \\
&\frac{(\partial l)}{\partial \alpha} = 0 \Leftrightarrow \frac{\zeta'(\alpha)}{\zeta(\alpha)} = \frac{1}{N}\sum_{i=1}^{N} lny_i \\
&\Leftrightarrow E[log(Y)] = \frac{1}{N}\sum_{i=1}^{N} lny_i
\end{aligned}$$

Com en els models anteriors, si es volen introduir les covariants *Espècie* i *Localització* en el model, assumint que el valor esperat de la distribució Zipf evoluciona com a funció lineal d'aquestes, llavors:

$$g(E[\boldsymbol{Y}|\boldsymbol{X}]) = \boldsymbol{X}\boldsymbol{\beta}, \tag{7.2}$$

on g és la funció *link*. Es aquest cas la funció d'enllaç que s'ha emprat és la funció logaritme.

D'altra banda, l'esperança de la distribució Zipf s'ha vist que és funció de la funció Zeta de Riemann, per tant l'equació (7.2) queda com segueix,

$$\frac{\zeta(\alpha-1)}{\zeta(\alpha)} = e^{\boldsymbol{X}\boldsymbol{\beta}}. \tag{7.3}$$

De l'equació (7.3) es desprén que per trobar les estimacions màxim versemblants dels paràmetres $\boldsymbol{\beta}$ i del paràmetre α caldrà fer servir un mètode iteratiu. En aquest treball s'ha emprat el software R per solucionar aquest problema. Aquest software contempla multitud de funcions, encapsulades en *packages*, que implementen aquests tipus de rutines. En aquest cas, s'ha emprat la funció *vglm* del paquet VGAM (*Vector Generalized Linear and Additive Models*, Yee 2008) d'R per ajustar tant la distribució Zipf i comparar-la amb la rutina programada, com per realitzar els models de regressió pertinents. La funció *vglm* utilitza el mètode IRLS (*Iteratively Reweighted Least Squares*) per trobar les estimacions màxim versemblants dels paràmetres.

7.2. Ajustos sense covariants

Per portar a terme l'ajust d'un model *Zipfian*, les dades s'agrupen en una taula de freqüències de freqüències. En aquest tipus de taula es donen els valors observats per la variable resposta *Nombre Total de Capítols Predats* i la seva freqüència en la mostra. En el nostre cas, les dades comencen en el valor 0 i, per tant, cal traslladar una unitat les mateixes. És a dir, és suma 1 a totes les observacions. Les probabilitats no canvien, únicament s'ha desplaçat l'escala dels valors observats. La probabilitat de l'1 és molt elevada, en contrast amb les probabilitats dels valors més grans que són molt petites. Per tant, té sentit ajustar una distribució Zipf.

A partir d'aquest moment es treballa amb la variable traslladada $TP^* = TP + 1$. Cal fer notar que és l'esperança de la variable TP^* la que és modela i, aquesta és

$$E[TP^*] = E[TP + 1] \Rightarrow \mu^* = \mu + 1$$

La Figura 7.1 mostra el gràfic de les dades aplicada la transformació log-log i l'ajust obtingut. S'observa que la distribució Zipf ajusta prou bé la probabilitat de l'1 (del 0 sense transformar) però, en canvi no ajusta correctament els valors de la cua. La distribució Zipf, al tractar-se d'una distribució uniparamètrica, és poc flexible. El fet d'ajustar bé el primer valor sembla determinant en la falta d'ajust de la cua. Finalment, donat que el paràmetre de la distribució és més petit que 2, $\alpha = 1{,}67$, no hi ha ni esperança ni variància finites.

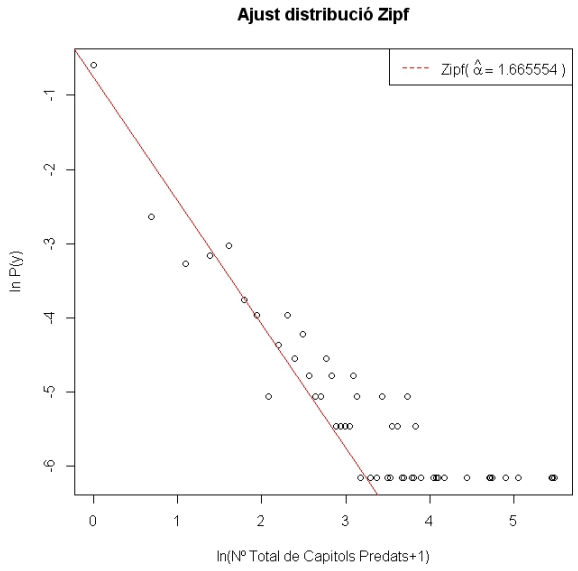


FIGURA 7.1. Gràfic, en eixos logarítmics, de la funció de probabilitat empírica del *Nombre Total de Capítols Predats* + 1 juntament amb l'ajust obtingut mitjançant la distribució Zipf.

La Taula 7.1 recull els estadístics de bondat d'ajust de la distribució Zipf a les dades d'estudi.

	$\hat{\alpha}$	Log-versemblança	AIC	BIC
Zipf	1,67	-1106,16	2214,32	2216,25

TAULA 7.1. Estimació del paràmetre de la distribució Zipf i estadístics de bondat d'ajust.

En base als valors observats en els criteris d'informació AIC i BIC, l'ajust no és del tot dolent. No obstant, com s'ha observat a la Figura 7.1 els valors de la cua no s'ajusten del tot bé.

TP+1	o_i	e_i	TP+1	o_i	e_i
1	261	228.02	31	3	0.71
2	34	78.22	33	1	0.64
3	18	40.26	34	1	0.61
4	20	24.83	35	2	0.58
5	23	16.98	37	2	0.53
6	11	12.41	39	1	0.48
7	9	9.52	40	1	0.46
8	3	7.55	42	3	0.42
9	6	6.16	44	1	0.39
10	9	5.13	45	1	0.37
11	5	4.34	46	2	0.36
12	7	3.73	49	1	0.32
13	4	3.25	57	1	0.25
14	3	2.85	59	1	0.23
15	3	2.53	60	1	0.23
16	5	2.26	65	1	0.2
17	4	2.03	85	1	0.12
18	2	1.84	111	1	0.08
19	2	1.68	112	1	0.08
20	2	1.53	115	1	0.07
21	2	1.41	135	1	0.06
22	4	1.3	156	1	0.04
23	3	1.2	231	1	0.02
24	1	1.12	234	1	0.02
27	1	0.91	239	1	0.02
29	1	0.8			
$\chi^2 = 58{,}34$ g.ll. $= 10$ $p < 0{,}001$					

TAULA 7.2. Taula de freqüència de freqüències. o_i freqüència observada del nombre de plantes + 1. e_i freqüència esperades sota una distribució Zipf. χ^2 estadístic Khi-quadrat de bondat d'ajust amb 10 graus de llibertat (g.ll.).

La Taula 7.2 correspon a la taula de freqüències de freqüències esmentada inicialment. A més, es trobem els valors esperats sota la distribució Zipf ajustada. També hi figura el valor de l'estadístic χ^2 de bondat d'ajust on, per a calcular-ho, s'han agrupat totes les categories de recompte ≥ 12 per tal que els valors esperats de cada categoria siguin suficientment grans.

La prova de bondat d'ajust de la χ^2 indica que es rebutja la hipòtesi nul·la, les dades no segueixen una distribució Zipf. Com es desprèn de les freqüències esperades, es subestima la freqüència del primer valor i es sobreestima la freqüència del segon. D'altra banda, subestima totes les freqüències per sobre del valor 30.

Augmentar un paràmetre addicional la distribució Zipf permet flexibilitzar l'ajust. La distribució Zipf Estesa o MOEZipf és una distribució biparamètrica que generalitza la distribució Zipf a partir de la transformació de Marshall-Olkin (Pérez-Casany i Caselles, 2013). Aquesta distribució permet millorar l'ajust dels primers valors respecte l'ajust que proporciona la distribució Zipf. Ara bé, tot i que s'ha portat a terme el seu ajust, s'ha comprovat que no millora significativament l'ajust proporcionat per la distribució Zipf i, per tant, no s'ha inclòs en aquesta memòria.

7.3. Ajustos amb covariants

Assumint que la distribució de la variable resposta segueix una llei Zipf i emprant la mateixa notació que en el Capítol 5, s'ajusten els següents models:

Model Zipf1:	$\mu_i^{r*} = e^{\beta_0+\beta_1\cdot Esp_i+ln(TC_i^r)}$
Model Zipf2:	$\mu_{ij}^{r}{}^{*} = e^{\beta_0+\beta_1\cdot Esp_i+\beta_2\cdot Loc_j+ln(TC_{ij}^r)}$,

TAULA 7.3. Models ajustats per a explicar la variable *Nombre Total de Capítols Predats* + 1 assumint distribució Zipf.

on $i = 1 \div 4$ i $j = 1 \div 6$. En el cas del model Zipf1 $r = 1 \div n_i$ i, en el model Zipf2 $r = 1 \div n_{ij}$.

Hem de tenir en compte que aquí s'assumeix que l'esperança de la distribució existeix i, per tant que $\alpha > 2$.

La Taula 7.4 recull les estimacions dels paràmetres dels dos models anteriors, així com les seves desviacions típiques entre parèntesis. En negreta es ressalten els paràmetres que han resultat estadísticament significatius.

Donat que s'ha traslladat la variable una unitat i es modela $TP + 1$ és necesari fer la següent apreciació. Per al model Zipf1, es té:

$$\mu_i^{r*} = e^{\beta_0+\beta_1\cdot Esp_i+ln(TC_i^r)} \Rightarrow \frac{\mu_i^{r*}}{TC_i^r} = e^{\beta_0+\beta_1\cdot Esp_i} \Rightarrow \frac{\mu_i^r + 1}{TC_i^r} = e^{\beta_0+\beta_1\cdot Esp_i}$$

Per tant, els resultats obtinguts del model s'han de comparar amb la taxa observada un cop s'ha traslladat la variable *Nombre Total de Capítols Predats* una unitat, és a dir, no són comparables directament amb les taxes observades a la Taula 4.8. Ara bé, com més gran sigui el valor de TC_i^r, més properes seran aquestes dues taxes. Per al model Zipf2, es deriva de forma anàloga.

	Models	
Paràmetre	Zipf1	Zipf2
Intercept	**-3,487 (0,067)**	**-4,550 (0,109)**
Espècie		
SI	**-4,508 (0,216)**	**-4,174 (0,129)**
SL	**-0,236 (0,089)**	**-1,183 (0,084)**
SP	**-1,659 (0,112)**	**-3,442 (0,112)**
SV	0	0
Localització		
CB		**3,195 (0,144**)
CP		**3,122 (0,142)**
CT		**3,506 (0,132)**
FM		**3,301 (0,139)**
SS		**1,405 (0,134)**
VF		0

TAULA 7.4. Estimació (desviació estàndard) dels paràmetres dels dos models ajustats per a modelar la variable *Nombre Total de Capítols Predats* + 1 assumint distribució Zipf.

Com s'ha comentat abans, l'estimació de la mitjana proporcionada pel model s'ha de comparar amb la taxa de predació *traslladada* una unitat, és a dir, es calcula com $(TP + 1)/TC$. Les mitjanes observades per a aquesta taxa traslladada són: 10,84% per SI, 32,52% per SL, 2,91% per SP i 15,10% per SV. L'espècie que més s'ha modificat respecte a la taxa real observada és l'espècie *S. Vulgaris*. Aquesta espècie presentava molts zeros i poca producció de capítols. Aleshores, traslladar les dades una unitat ha fet incrementar molt la taxa de predació. L'espècie *S. Pterophorus* presentava també molts zeros però és una espècie amb molta producció de capítols i, per tant, no s'ha vist tant afectada com l'anterior. Anàlogament, l'espècie *S. Inaequidens* també ha produït molts capítols. Donat que ha presentat més predació que les anteriors, augmentar la predació una unitat ha fet incrementar una mica aquesta taxa. Finalment, l'espècie *S. Lividus* és l'espècie que ha presentat més predació i, es tractava d'una espècie amb poca producció de capítols. Es veu també afectada per la addició d'una unitat a la predació però no tant com ha estat el cas del *S. Vulgaris*.

De la Taula 7.4 es desprèn que, en el model Zipf1 es troben diferencies estadístiques significatives entre totes les espècies i l'espècie de referència, el *S. Vulgaris*. El valor esperat de la taxa de capítols predats per a l'espècie *S. Vulgaris* és 3,06% ($e^{-3,487}$), aquest valor es troba molt allunyat de la taxa traslladada, 15,10%. Per l'espècie *S. Pterophorus* s'estima un valor esperat de 0,58% quan la taxa traslladada és del 2,91%. Aquestes dues espècies es caracteritzen per presentar un elevat percentatge de zeros, 83% i 84% respectivament i, a més, presenten baixes taxes de predació. D'altra banda, per a les espècies *S. Lividus* i *S. Inaequidens* l'ajust és força dolent. Per l'espècie *S. Lividus* s'estima una taxa esperada de predació del 2,42% quan la taxa traslladada calculada és del 32,52%. Per l'espècie *S. Inaequidens*, la taxa esperada de predació és del 0,03% quan s'ha calculat una taxa traslladada del 10,84%. Aquestes dues espècies corresponen a les espècies més predades i amb un

percentatge de zeros més petit, 33% i 30%. Aquest model no sembla l'adequat per aquestes dades.

Respecte al model Zipf2, la inclusió del factor *Localització* no millora les estimacions del model Zipf1. Per exemple, les taxes de predació esperades en les espècies *S. Inaequidens* i *S. Lividus* en la localització de Vallforners (VF) són pràcticament del 0%, quan les taxes traslladades de predació en aquesta localització són 11,59% i 46,80% respectivament. Com en el cas del model anterior, la distribució Zipf no sembla la més adequada per modelar la taxa de predació a partir de les covariants *Espècie* i *Localització*.

La Taula 7.5 recull les mesures de bondat d'ajust que s'han vist en el Capítol 6.

	Models	
Criteri	Zipf1	Zipf2
Log likelihood	-2352,63	-1887,44
AIC	4713,26	3792,89
BIC	4729,91	3830,36

TAULA 7.5. Estadístics de bondat d'ajust dels dos models ajustats per a modelar la variable *Nombre Total de Capítols Predats* + 1 assumint distribució Zipf.

Del fet de passar de l'ajust del model Zipf1 al model Zipf2, s'aconsegueix una reducció del 19,77% en el logaritme de la versemblança. L'estadístic LRT pren el valor 930,38 cosa que porta a rebutjar la hipòtesis nul·la en favor de l'alternativa, és a dir, el model Zipf2 millora l'ajust del model Zipf1.

Si es compara l'ajust del model Zipf1 amb els models ZIP3 i ZIBN3 en termes d'AIC i BIC, els models zero inflats ofereixen un millor ajust. El model zero inflat amb distribució de Poisson presenta un AIC de 3529 i quan s'assumeix una distribució BN l'AIC val 1805, menys de la meitat que el valor del model Zipf1. Respecte al model Zipf2, si es compara amb els models zero inflats ZIP4 i ZIBN4 s'observa la mateixa situació: l'AIC quan s'assumeix distribució de Poisson és de 3060 i si s'assumeix distribució BN és de 1782 enfront de 3792 que pren aquest criteri en el model Zipf2. Si es comparen les versemblances dels models, el model ZIP3 millora un 25,35% la versemblança del model Zipf1 i el model ZIBN3 la millora en un 62,02%. Respecte la versemblança del model Zipf2, el model ZIP4 la millora en un 18,62% i el model ZIBN4 en un 53,53%. En base a aquest resultats, es preferible realitzar l'ajust mitjançant models zeros inflats enlloc de fer servir la distribució Zipf.

Els valors dels estadístics de bondat d'ajust observats a la Taula 7.5 són més grans que els de l'ajust sense covariants. En l'ajust de la Figura 7.1 no es contemplava ni les covariants ni l'*offset* i es modelava la variable *Nombre Total de Capítols Predats*. A l'ajust de la Taula 7.5 la variable que es modela és la *Taxa de Capítols Predats*.

És probable que els ajustos emprant la distribució Zipf milloressin si el que s'assumeix lineal amb les covariants és el valor esperat del logaritme de la variable enlloc del valor esperat de la variable. Això pensem que és així perquè l'esperança del

logaritme de la variable existeix sempre i perquè la transformació logarítmica és la que estabilitza la variància de les dades. Aquest tipus d'anàlisi però no es porta a terme en aquest treball però marca una línia futura de recerca.

Capítol 8
Conclusions

En aquest treball s'ha analitzat l'herbivorisme del *Senecio* en quatre espècies diferents. Per a fer-ho, s'han utilitzat diferents models de regressió assumint resposta Poisson, binomial negativa, les seves respectives versions zero inflades, i també assumint resposta Zipf.

Des del punt de vista estadístic, els resultats obtinguts permeten concloure, en primer lloc, que el model de Poisson proporciona un ajust molt pobre donat que no es capaç de captar l'elevada sobredispersió present a les dades i el considerable nombre de zeros. El model Binomial Negatiu s'adapta millor a aquesta situació però són els models ZIBN els que proporcionen el millor ajust. D'altra banda, introduir la covariable *Espècie* per modelar l'excés de zeros en el model ZIBN ha permès detectar diferències entre espècies, no només en termes de valors esperats sinó també en l'excés de zeros. El model ZIBN en contemplar dos fonts de variabilitat additives, una per l'excés de zeros i una altra per la sobredispersió, s'han mostrat més eficient per capturar aquestes dues fonts de variabilitat.

La localització de mostreig ha resultat rellevant al llarg de tot el treball però la falta d'informació respecte les diferents localitzacions fa difícil justificar l'efecte observat en cada localització en concret.

Respecte el model Zipf, si no es tenen en compte covariants, el model és prou bo per modelar la variable *Nombre Total de Capítols Predats* traslladada una unitat. És capaç d'ajustar satisfactòriament la probabilitat de la unitat i dels valors intermitjos amb només un paràmetre. Ara bé, la falta de flexibilitat ocasionada per tenir només un paràmetre fa que no adapti prou bé els valors de la cua.

Respecte els models Zipf amb covariants per a modelar la taxa de predació, queda clar que els ajustos no són tant satisfactòris com els obtinguts amb el model ZIBN. Ara bé, en el primer cas, el model té només 9 paràmetres i, en el segon se n'estimen 14.

Des del punt de vista biològic, fer servir models amb excés de zeros ha permés detectar dues components diferenciades: la *freqüència de dany*, corresponen al percentatge de plantes atacades (complementari a l'excés de zeros), i la *intensitat del dany* mesurada com la predació esperada (estimada a partir del procés de recompte).

Aquesta diferenciació ha portat a les investigadores a elaborar quatre escenaris diferents per a descriure l'herbivorisme: *escenari L*, causat per un baix percentatge de plantes atacades i amb una intensitat baixa de predació; *escenari M1*, determinat per un baix percentatge de plantes atacades però amb una intensitat de predació alta; *escenari M2*, caracteritzat per un alt percentatge de plantes atacades però amb baixa intensitat de predació i, finalment l'*escenari H*, causat per un alt percentatge de plantes atacades amb una elevada predació. La Figura 8.1 il·lustra els quatre escenaris definits.

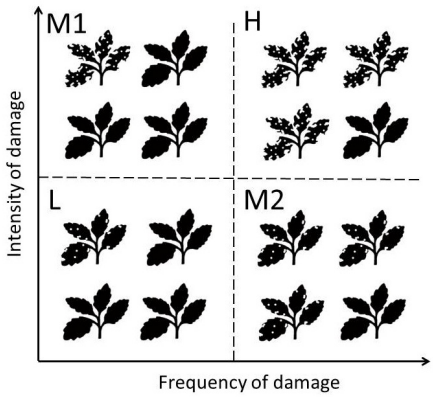


FIGURA 8.1. Visualització dels quatre escenaris de predació detectats: nivell de predació baix (L), nivell de predació mitjà (M1 y M2) i nivell de predació alt (H). La intensitat del dany es refereix a la taxa de predació en les plantes atacades (estimada a partir del procés de recompte) i la freqüència del dany es refereix al percentatge de plantes atacades. Imatge elaborada per Maria Morante.

En aquests quatre escenaris detectats, l'espècie *S. Lividus* es situaria a l'escenari d'elevada predació (H), tant en taxes com en percentatge d'afectació. L'espècie *S. Inaequidens* es situaria en l'escenari M2, es tractaria d'una espècie força atacada pels insectes però amb poca intesitat de predació. Finalment, les espècies *S. Pterophorus* i *S. Vulgaris*, que inicialment diríem que estan molt poc predades, les situaríem en un escenari de nivells mitjos de predació. S'ha observat que efectivament hi ha poques plantes atacades però, aquelles plantes que han estat atacades pels insectes, presenten una predació força elevada (escenari M1).

De l'ajust obtingut amb el model Binomial Negatiu Zero Inflat (ZIBN4), amb la probabilitat del zero depenent de l'espècie, es deriva que l'espècie *S. Lividus* ha resultat ser la més predada. Efectivament això s'explica perquè es tracta d'una espècie autòctona a la qual estan habituats els insectes i, a més, existeix sincronia entre l'època de floració de la planta i l'època de reproducció de l'insecte. Aquesta espècie floreix entre abril i juliol i els insectes són més actius els mesos de juny i juliol. Per l'altre espècie autòctona, la *S. Vulgaris* s'ha detectat un important excés de zeros i una predació força alta quan es produeix. Això és degut a que els insectes reconeixen la planta però aquesta floreix abans que l'espècie *S. Lividus* per tant hi ha pocs insectes en l'ambient quan aquesta planta està en flor. El *S. Vulgaris* floreix generalment entre els mesos de maig i juny.

El model estima un elevat excés de zeros per a l'espècie *S. Pterophorus* però amb una elevada predació quan aquesta es produeix. L'època de floració d'aquesta espècie coincideix amb l'època reproductiva dels insectes, de fet és coetània amb l'espècie *S. Lividus*. En aquest escenari es pot dir que els insectes encara no estan habituats a la seva presència i no la reconeixen com a potencial font d'aliment, prefereixen el *S. Lividus* en detriment d'aquesta altra. Respecte l'altre espècie exòtica, la *S. Inaequidens*, aquesta pot florir des del maig fins al desembre. De fet ha estat l'única espècie per a la qual s'ha detectat predació en els mesos de novembre. Davant de la presència de l'espècie autòctona *S. Lividus*, els insectes s'estimen més aquesta espècie autòctona en detriment de l'espècie exòtica. D'altra banda, aquesta espècie ofereix flors quan les altres espècies ja no ho estan fent i, per tant, els insectes que encara pugui haver-hi en l'ambient només tenen aquesta opció per fer la posta, és a dir, aquesta espècie és aquella que trien quan no hi ha cap altra alternativa.

Així doncs, sembla confirmar-se la hipòtesis d'alliberament dels enemics per part de les espècies exòtiques: ja per motius d'adaptació al cicle temporal (els insectes estan adaptats als cicles de floració de les espècies autòctones) o bé per motius de no ésser reconegudes com a fonts d'aliment, els insectes prefereixen l'espècie autòctona *S. Lividus* en detriment de les dues espècies exòtiques.

Capítol 9
Linies de futur

En aquest capítol es presenten algunes directrius del que podriem considerar-se línies futures d'aquest treball.

Per començar es poden provar d'altres distribucions alternatives a la Poisson o la BN en el model ZI (Famoye i Singh, 2006). Mitjançant el paquet GAMLSS (*Generalized Additive Models for Location, Scale and Shape*, www.gamlss.org) del software de lliure distribució R es poden ajustar models ZI amb distribucions que són mixtura Poisson i no són la binomial negativa. Per exemple, es podria considerar la distribució Poisson-Inversa Gaussiana.

Una altra alternativa, seria implementar un model *Zero Altered* (ZA), o també conegut com *Hurdle Models* (Zuur, 2009 i Johnson *et al.*, 2005). Com s'ha introduït en el Capítol 5, els models ZA es consideren models en dues parts. Per una banda, es modela un procés binari de zeros davant no zeros mitjançant un model binomial i, per una altra banda, les observacions diferents de zero es modelen mitjançant distribucions de recompte truncades en zero. Com les distribucions són zero-truncades, s'assumeix que el procés de recompte no pot produir zeros (Martin *et al.*, 2005). Des d'aquest punt de vista, les conclusions biològiques són sensiblement diferents donat que no s'avalua directament el dany ocasionat sinó el dany quan aquest s'ha produït.

D'altra banda, també es pot fer servir un model lineal amb distribució Neyman Type A (Dobbie i Welch, 2001), que és també mixtura de Poisson i alhora, *Poisson-Stopped Sum*.

Respecte la distribució Zipf, una altra possibilitat és ajustar un model Zipf amb covariants però, imposant que l'esperança del logaritme i no l'esperança de la distribució sigui qui evoluciona linealment com a funció de les covariants.

Finalment, com a alternativa a l'estimació màxim versemblant, es poden fer servir eines bayesianes on s'obtenen les distribucions dels paràmetres estimats, en base a assumir unes distribucions a priori sorgides del coneixement dels experts (Martin *et al.*, 2005).

Annexe A: Codis R i SAS

Codi R: Anàlisi Exploratòria

En aquest apèndix es mostra el codi R utilitzat per generar els resultats presentats en el capítol ***Anàlisi Estadística Descriptiva***.

```
# DESCRIPTIVA DADES
# ANABEL BLASCO-MORENO
# TFM

library("ggplot2")
library("lattice")
library("MASS")
library("Deducer")

dades <- read.csv("Dades agregat.csv",header=TRUE,sep=";")
attach(dades)

dades$PP <- (dades$TP_suma/dades$TC_suma)*100
dades$PP_tep <- (dades$pred_Tep_suma/dades$TC_suma)*100
dades$PP_pyr <- (dades$pred_Pyr_suma/dades$TC_suma)*100
# Taxes per TP + 1
dades$PP_1 <- ((dades$TP_suma + 1)/dades$TC_suma)*100


# Resum de la BBDD
summary(dades)

# Nombre observacions per especie i individu
print(tabla.esp.pob <- table(dades$especie, dades$poblacio))

# Porcentajes fila
print(pct.pob <- prop.table(tabla.esp.pob,1)*100 ,3)

# Porcentajes columna
```

```
print(pct.pob <- prop.table(tabla.esp.pob,2)*100 ,3)


# Descriptiva per especie

desc <- function(var, clas){
n <- tapply(var,clas, length)
m <- tapply(var,clas, mean)
s <- tapply(var,clas, sd)
cv <- s/m
med <- tapply(var,clas, median)
q1 <- tapply(var,clas,quantile, probs=0.25)
q3 <- tapply(var,clas,quantile, probs=0.75)
min <- tapply(var,clas, min)
max <- tapply(var,clas, max)
cbind(N=n, Mitjana=m, "Desviació Estàndard"=s, CV = cv, Mediana=med,
"1er Quartil"=q1, "3er quartil"=q3, Mínim=min, Màxim=max)
}

# Descriptiva per especie i poblacio

dades.CB <- subset(dades,poblacio=="CB")
dades.CP <- subset(dades,poblacio=="CP")
dades.CT <- subset(dades,poblacio=="CT")
dades.FM <- subset(dades,poblacio=="FM")
dades.SS <- subset(dades,poblacio=="SS")
dades.VF <- subset(dades,poblacio=="VF")

descriptive <- function(loc, var, clas){
cat("Localització", loc ,"\n")
n <- tapply(var,clas, length)
m <- tapply(var,clas, mean)
s <- tapply(var,clas, sd)
cv <- s/m
med <- tapply(var,clas, median)
q1 <- tapply(var,clas,quantile, probs=0.25)
q3 <- tapply(var,clas,quantile, probs=0.75)
min <- tapply(var,clas, min)
max <- tapply(var,clas, max)
cbind(N=n, Mitjana=m, "Desviació Estàndard"=s, CV = cv, Mediana=med,
"1er Quartil"=q1, "3er quartil"=q3, Mínim=min, Màxim=max)
}


##############################################################################
#  DESCRIPTIVA TOTAL PREDACIO
##############################################################################
```

```
#Taula de frequencies per TP
table(dades$TP_suma)

#Diagrama de la taula de frequencies
I1 <- is.na(dades$TP_suma) |
is.na(dades$especie) |
is.na(dades$poblacio) |
is.na(dades$TC_suma)
dades2 <- dades[!I1, ]
win.graph()
par(mfrow=c(1,1),mai=c(0.85, 0.8, 0.05, 0.05))
plot(table(dades2$TP_suma),
ylab = "Freqüència",
xlab = "Nombre Total de Capítols Predats")

#Diagrama de la taula de frequencies de TP per especie
I1 <- is.na(dades$TP_suma) |
is.na(dades$especie) |
is.na(dades$poblacio) |
is.na(dades$TC_suma)
dades2 <- dades[!I1, ]
dades2.SI <- subset(dades2,especie=="SI")
dades2.SL <- subset(dades2,especie=="SL")
dades2.SP <- subset(dades2,especie=="SP")
dades2.SV <- subset(dades2,especie=="SV")
win.graph()
par(mfrow=c(2,2),mai=c(0.7, 0.7, 0.3, 0.05))
plot(table(dades2.SI$TP_suma),
ylab = "Freqüència",
xlab = "Capítols Predats en l'espècie S. Inaequidens")
plot(table(dades2.SL$TP_suma),
ylab = "Freqüència",
xlab = "Capítols Predats en l'espècie S. Lividus")
plot(table(dades2.SP$TP_suma),
ylab = "Freqüència",
xlab = "Capítols Predats en l'espècie S. Pterophorus")
plot(table(dades2.SV$TP_suma),
ylab = "Freqüència",
xlab = "Capítols Predats en l'espècie S. Vulgaris")

#, mar=c(2, 2, 2.5, 0.5)

#Diagrama de la taula de frequencies només fins a 15 de TP
dades3 <- subset(dades2, dades2$TP_suma < 16)
win.graph()
plot(table(dades3$TP_suma),
ylab = "Freqüència",
xlab = "Nombre Total de Capítols Predats")
```

```
#Diagrama de la taula de frequencies per TC truncat a 300.
I2 <- is.na(dades$TC_suma) |
is.na(dades$especie) |
is.na(dades$poblacio) |
is.na(dades$TP_suma)
dades2 <- dades[!I2, ]
win.graph()
par(mfrow=c(1,1),mai=c(0.85, 0.8, 0.05, 0.05))
dades3 <- subset(dades2, dades2$TC_suma < 300)
plot(table(dades3$TC_suma),
ylab = "Freqüència",
xlab = "Nombre Total de Capítols Produïts")

win.graph()
dades3 <- subset(dades2, dades2$TC_suma >= 300)
plot(table(dades3$TC_suma),
ylab = "Freqüència",
xlab = "Nombre Total de Capítols Produïts")
summary(dades3)

#Diagrama de la taula de frequencies per la variable TAXA.

dades2$PP_cat <- recode.variables(dades2$PP, "0 -> '0%';
0.000001:10 -> '0.1-10%';
10.000001:20 -> '10.1-20%';
20.000001:30 -> '20.1-30%';
30.000001:40 -> '30.1-40%';
40.000001:50 -> '40.1-50%';
50.000001:60 -> '50.1-60%';
60.000001:70 -> '60.1-70%';
70.000001:80 -> '70.1-80%';
80.000001:90 -> '80.1-90%';
else -> '90.1-100%';")
win.graph()
par(mfrow=c(1,1),mai=c(0.85, 0.8, 0.05, 0.05))
barplot(table(dades2$PP_cat), xlab="Taxa observada de predació",
 ylab="Freqüència", cex.names=0.7 )

plot(table(dades2$PP),
ylab = "Freqüència",
xlab = "Taxa Capítols Predats")

win.graph()
dades3 <- subset(dades2, dades2$TC_suma >= 300)
plot(table(dades3$TC_suma),
ylab = "Freqüència",
xlab = "Capítols Predats")
```

```
summary(dades3)


#dades SENSE zeros
dades2_sense0 <- subset(dades2,TP_suma>0)
desc(dades2_sense0$PP,dades2_sense0$especie)
dades2_sense0$glob <- "Total"

########################################################################

# Mitjana del Nº PREDATS per especie i global: tots els casos

desc(dades2$TP,dades2$especie)

dades2$glob <- "Total"
desc(dades2$TP,dades2$glob)

#Diagrama de caixa per Total Predacio
win.graph()
par(mfrow=c(1,1),mai=c(0.9, 0.9, 0.15, 0.15))
boxplot(TP_suma ~ especie, data = dades, col = "lightgray",
ylab = "Capítols Predats",
xlab = "Espècie")

# Mitjana del Nº PREDATS per especie i global: sense els zeros
desc(dades2_sense0$TP,dades2_sense0$especie)
desc(dades2_sense0$TP,dades2_sense0$glob)

# Mitjana del total de capitols PREDATS per especie i poblacio
descriptive("CB",dades.CB$TP_suma,dades.CB$especie)
descriptive("CP",dades.CP$TP_suma,dades.CP$especie)
descriptive("CT",dades.CT$TP_suma,dades.CT$especie)
descriptive("FM",dades.FM$TP_suma,dades.FM$especie)
descriptive("SS",dades.SS$TP_suma,dades.SS$especie)
descriptive("VF",dades.VF$TP_suma,dades.VF$especie)

# Distribucio del total capitols predatas (TP) per especie i poblacio
win.graph()
dotplot(especie~TP_suma|poblacio,dades)

########################################################################

# Mitjana del Nº CAPITOLS PRODUITS per especie i global: tots els casos

desc(dades2$TC_suma,dades2$especie)

dades2$glob <- "Total"
desc(dades2$TC_suma,dades2$glob)
```

```
#Diagrama de caixa
win.graph()
par(mfrow=c(1,1),mai=c(0.9, 0.9, 0.15, 0.15))
boxplot(TC_suma ~ especie, data = dades, col = "lightgray",
ylab = "Capítols Produïts",
xlab = "Espècie")

# Mitjana del Nº CAPITOLS PRODUITS per especie i global: sense els zeros
desc(dades2_sense0$TC,dades2_sense0$especie)
desc(dades2_sense0$TC,dades2_sense0$glob)

# Mitjana del total de capitols (TC) per especie i poblacio

descriptive("CB",dades.CB$TC_suma,dades.CB$especie)
descriptive("CP",dades.CP$TC_suma,dades.CP$especie)
descriptive("CT",dades.CT$TC_suma,dades.CT$especie)
descriptive("FM",dades.FM$TC_suma,dades.FM$especie)
descriptive("SS",dades.SS$TC_suma,dades.SS$especie)
descriptive("VF",dades.VF$TC_suma,dades.VF$especie)

# Distribucio del total capitols (TC) per especie i poblacio
win.graph()
dotplot(especie~TC_suma|poblacio,dades)

#################################################################

# Mitjana del % PREDATS per especie i global: tots els casos
# dades2$PP <- dades2$PP/100

desc(dades2$PP,dades2$especie)

dades2$glob <- "Total"
desc(dades2$PP,dades2$glob)

#Diagrama de caixa per Total Predacio
win.graph()
par(mfrow=c(1,1),mai=c(0.9, 0.9, 0.15, 0.15))
boxplot(PP ~ especie, data = dades, col = "lightgray",
ylab = "Taxa Capítols Predats",
xlab = "Espècie")

# Mitjana del % PREDATS per especie i global: sense els zeros
desc(dades2_sense0$PP,dades2_sense0$glob)

# Mitjana del % de capitols PREDATS per especie i poblacio
descriptive("CB",dades.CB$PP,dades.CB$especie)
descriptive("CP",dades.CP$PP,dades.CP$especie)
```

```
descriptive("CT",dades.CT$PP,dades.CT$especie)
descriptive("FM",dades.FM$PP,dades.FM$especie)
descriptive("SS",dades.SS$PP,dades.SS$especie)
descriptive("VF",dades.VF$PP,dades.VF$especie)

# Distribucio del % capitols predatas per especie i poblacio
win.graph()
dotplot(especie~PP|poblacio,dades)

################################################################

# Mitjana del % TP + 1 per especie i global: tots els casos
# dades2$PP_1 <- dades2$PP_1/100

desc(dades2$PP_1,dades2$especie)

dades2$glob <- "Total"
desc(dades2$PP_1,dades2$glob)

#Diagrama de caixa per Total Predacio
win.graph()
par(mfrow=c(1,1),mai=c(0.9, 0.9, 0.15, 0.15))
boxplot(PP_1 ~ especie, data = dades, col = "lightgray",
ylab = "Taxa Capítols Predats",
xlab = "Espècie")

# Mitjana del % PREDATS per especie i global: sense els zeros
desc(dades2_sense0$PP_1,dades2_sense0$glob)

# Mitjana del % de capitols PREDATS per especie i poblacio
descriptive("CB",dades.CB$PP_1,dades.CB$especie)
descriptive("CP",dades.CP$PP_1,dades.CP$especie)
descriptive("CT",dades.CT$PP_1,dades.CT$especie)
descriptive("FM",dades.FM$PP_1,dades.FM$especie)
descriptive("SS",dades.SS$PP_1,dades.SS$especie)
descriptive("VF",dades.VF$PP_1,dades.VF$especie)

# Distribucio del % capitols predatas per especie i poblacio
win.graph()
dotplot(especie~PP_1|poblacio,dades)
```

Codi SAS: Models Zero Inflats

En aquest apèndix es mostra el codi SAS utilitzat per generar els resultats presentats en el capítol ***Models amb excés de zeros***.

```
* MODEL ZIP I ZIBN;
* ANABEL BLASCO-MORENO;
* TFM;

libname ruta "E:\PFC Anabel\TFM\SAS";
%let ruta=E:\PFC Anabel\TFM\SAS;

option nodate nonumber;
ods noproctitle;
title;

proc format;
value pob 1="CB"
2="CP"
3="CT"
4="FM"
5="SS"
6="VF";
value esp 1="SV" 2="SL" 3="SI" 4="SP";
run;

data dades;
set ruta.dades_agregat;
run;

data dades;
set dades;
if poblacio= "CB" then pob2=1;
if poblacio= "CP" then pob2=2;
if poblacio= "CT" then pob2=3;
if poblacio= "FM" then pob2=4;
if poblacio= "SS" then pob2=5;
if poblacio= "VF" then pob2=6;
espSI=0;
espSP=0;
espSV=0;
espSL=0;
if especie="SI" then do; espSI=1; esp2=3; end;
if especie="SP" then do; espSP=1; esp2=4; end;
if especie="SV" then do; espSV=1; esp2=1; end;
if especie="SL" then do; espSL=1; esp2=2; end;
pobCB=0;
pobCP=0;
pobCT=0;
pobFM=0;
pobSS=0;
pobVF=0;
```

```
if poblacio="CB" then pobCB=1;
if poblacio="CP" then pobCP=1;
if poblacio="CT" then pobCT=1;
if poblacio="FM" then pobFM=1;
if poblacio="SS" then pobSS=1;
if poblacio="VF" then pobVF=1;
log_TC = log(TC_suma);
format pob2 pob.;
format esp2 esp.;
run;

proc sort data=dades;
by esp2 pob2;
run;




ods rtf;

/*********************************************************************/
*Gràfic barres de freqüències;
/*********************************************************************/

ods graphics on;
proc freq data=dades;
   table TP_suma / plots(only)=freqplot(scale=freq); /*percent*/
label TP_suma = "Total capítols predats";
run;

/*********************************************************************/
*POISSON;
/*********************************************************************/

*1) Només especie; *P1;
ods graphics on;
proc genmod data=dades plots= RESCHI(index xbeta);
class esp2 ;
model TP_suma = esp2  / offset=log_TC dist = poisson type3;
lsmeans esp2 / diff adjust=tukey ilink;
run;

*2) especie i localització; *P2;
proc genmod data=dades plots= RESCHI(index xbeta);
class esp2 pob2;
model TP_suma = esp2 pob2 / offset=log_TC dist = poisson type3;
lsmeans esp2 / diff adjust=tukey ilink;
run;
```

```
*3) especie i localització amb interacció;
proc genmod data=dades plots= RESCHI(index xbeta);
class esp2 pob2;
model TP_suma = esp2 pob2 esp2*pob2/ offset=log_TC dist = poisson type3;
run;


/*****************************************************************/
*BINOMIAL NEGATIU;
/*****************************************************************/

*1) Només especie; *BN1;
proc genmod data=dades plots= RESCHI(index xbeta);
class esp2 ;
model TP_suma = esp2  / offset=log_TC dist = NEGBIN type3;
lsmeans esp2 / diff adjust=tukey ilink;
run;

*2) especie i localització; *BN2;
proc genmod data=dades plots= RESCHI(index xbeta);
class esp2 pob2;
model TP_suma = esp2 pob2 / offset=log_TC dist = NEGBIN type3;
lsmeans esp2 / diff adjust=tukey ilink;
run;

*3) especie i localització amb interacció;
proc genmod data=dades plots= RESCHI(index xbeta);
class esp2 pob2;
model TP_suma = esp2 pob2  esp2*pob2/ offset=log_TC dist = NEGBIN type3;
run;


/*****************************************************************/
*ZIP;
/*****************************************************************/

*Versió SAS 9.2;
*1) només especie; *ZIP1;
proc genmod data=dades plots= RESCHI(index xbeta);
class esp2 ;
model TP_suma = esp2 /  offset=log_TC dist = zip type3;
zeromodel / link= logit;
lsmeans esp2 / diff adjust=tukey ilink;
run;

*Alternativa: programacio manual;
proc nlmixed data = dades tech=NEWRAP;
```

```
parms b0=-2.105558 pf=1 pe1=-0.771373 pe2=1.074283 pe3=-0.136343
      a0=0.104418;
bounds 0.99999 < pf < 1.0001;
logit0 = a0;
prob0 = 1/(1+exp(-logit0));
mu = exp(b0 + pe1*espSI + pe2*espSL + pe3*espSP + pf*log_TC);
if TP_suma = 0 then
ll = log(prob0 + (1- prob0)*exp(-mu));
else
ll = TP_suma*log(mu) + log(1- prob0) - mu - lgamma(TP_suma + 1);
model TP_suma ~ general(ll);
estimate 'inflation p' prob0;
run;

*2) especie i localització; *ZIP2;
proc genmod data=dades  plots= RESCHI(index xbeta);
class esp2 pob2;
model TP_suma = esp2 pob2/  offset=log_TC dist = zip type3;
zeromodel / link= logit;
lsmeans esp2 / diff adjust=tukey ilink;
run;

*Alternativa: programacio manual;
proc nlmixed data = dades tech=NEWRAP;
parms b0= -1.5020 pf=1 pe1=2.241209 pe2=2.196204 pe3=0.201141
pob1=-0.1763 pob2= -0.6227 pob3=-0.4531 pob4=0.3983 pob5= -1.0377
      a0= -0.1003;
    bounds 0.99999 < pf < 1.0001;
logit0 = a0;
prob0 = 1/(1+exp(-logit0));
mu = exp(b0 + pe1*espSI + pe2*espSL + pe3*espSP + pf*log_TC +
pob1*pobCB + pob2*pobCP + pob3*pobCT + pob4*pobFM + pob5*pobSS);
if TP_suma = 0 then
ll = log(prob0 + (1- prob0)*exp(-mu));
else
ll = TP_suma*log(mu) + log(1- prob0) - mu - lgamma(TP_suma + 1);
model TP_suma ~ general(ll);
estimate 'inflation p' prob0;
run;

*3) especie   +  especie als zeros; *ZIP3;
proc genmod data=dades  plots= RESCHI(index xbeta);
class esp2 ;
model TP_suma = esp2 /  offset=log_TC dist = zip type3;
zeromodel esp2 / link= logit;
lsmeans esp2 / diff adjust=tukey ilink;
run;
```

```
*Alternativa: programació manual;
proc nlmixed data = dades tech=NEWRAP;
parms b0= -1.6713 pf=1 pe1=-1.2975 pe2=0.7186 pe3=-0.5716
  a0=1.2876 zpe1= -3.1802 zpe2=-2.1810 zpe3=0.3631 ;
bounds 0.99999 < pf < 1.01;
logit0 = a0 + zpe1*espSI + zpe2*espSL + zpe3*espSP;
prob0 = 1/(1+exp(-logit0));
mu = exp(b0 + pe1*espSI + pe2*espSL + pe3*espSP + pf*log_TC);
if TP_suma = 0 then
ll = log(prob0 + (1- prob0)*exp(-mu));
else
ll = TP_suma*log(mu) + log(1- prob0) - mu - lgamma(TP_suma + 1);
model TP_suma ~ general(ll);
estimate 'Inflation SI' 1/(1+exp(-(a0 + zpe1)));
estimate 'Inflation SL' 1/(1+exp(-(a0 + zpe2)));
estimate 'Inflation SP' 1/(1+exp(-(a0 + zpe3)));
estimate 'Inflation SV' prob0;
run;


*4) especie i localització  +  especie als zeros; *ZIP4;
proc genmod data=dades  plots= RESCHI(index xbeta);
class esp2 pob2;
model TP_suma = esp2 pob2/  offset=log_TC dist = zip type3;
zeromodel esp2 / link= logit;
lsmeans esp2 / diff adjust=tukey ilink;
run;

*Alternativa: programació manual;
proc nlmixed data = dades tech=NEWRAP;
parms b0= -1.4652 pf=1 pe1=-1.4461 pe2=0.8188 pe3=-0.7023
pob1=-0.1735 pob2=-0.6414 pob3= -0.4436 pob4= 0.4036 pob5=-1.0291
      a0=1.2325 zpe1=-3.3601 zpe2=-2.1133 zpe3=0.4082 ;
bounds 0.99999 < pf < 1.01;
logit0 = a0 + zpe1*espSI + zpe2*espSL + zpe3*espSP;
prob0 = 1/(1+exp(-logit0));
mu = exp(b0 + pe1*espSI + pe2*espSL + pe3*espSP + pf*log_TC +
pob1*pobCB + pob2*pobCP + pob3*pobCT + pob4*pobFM + pob5*pobSS);
if TP_suma = 0 then
ll = log(prob0 + (1- prob0)*exp(-mu));
else
ll = TP_suma*log(mu) + log(1- prob0) - mu - lgamma(TP_suma + 1);
model TP_suma ~ general(ll);
estimate 'Inflation SI' 1/(1+exp(-(a0 + zpe1)));
estimate 'Inflation SL' 1/(1+exp(-(a0 + zpe2)));
estimate 'Inflation SP' 1/(1+exp(-(a0 + zpe3)));
estimate 'Inflation SV' prob0;
run;
```

```
/*****************************************************************/
*ZINB;
/*****************************************************************/

*Versió SAS 9.22;
*1) només especie; *ZIBN1;
proc genmod data=dades  plots= RESCHI(index xbeta);
class esp2 ;
model TP_suma = esp2 /  offset=log_TC dist = zinb type3;
zeromodel / link= logit;
lsmeans esp2 / diff adjust=tukey ilink;
run;

*Alternativa: programació manual;
proc nlmixed data = dades tech=NEWRAP;
parms b0=-2.576647 pf=1 pe1=-0.009946 pe2=1.440169 pe3= 0.250516
      a0=-0.443428 alpha=0.787072;
  bounds 0.9 < pf < 1.0001;
logit0 = a0;
prob0 = 1/(1+exp(-logit0));
mu = exp(b0 + pe1*espSI + pe2*espSL + pe3*espSP + pf*log_TC);
    m = 1/alpha;
    p = 1/(1+alpha*mu);
    if TP_suma = 0 then
     ll = log(prob0 + (1-prob0)*(p**m));
    else ll = log(1-prob0) + lgamma(m + TP_suma) - lgamma(TP_suma + 1)
 - lgamma(m) + m*log(p) + TP_suma*log(1-p);
model TP_suma ~ general(ll);
estimate 'inflation p' prob0;
run;

*2) especie i localització; *ZIBN2;
proc genmod data=dades  plots= RESCHI(index xbeta);
class esp2 pob2;
model TP_suma = esp2 pob2/  offset=log_TC dist = zinb type3;
zeromodel / link= logit;
lsmeans esp2 / diff adjust=tukey ilink;
run;

*Alternativa: programació manual;
proc nlmixed data = dades tech=NEWRAP;
parms b0=-2.0481 pf=1 pe1=-0.230339 pe2=1.461367 pe3=0.264191
pob1=0.457879 pob2= -0.7904 pob3=-0.7280  pob4=-0.3412 pob5= -1.086
      a0=-0.405262 alpha=0.602843;
bounds 0.99 < pf < 1.000001;
logit0 = a0;
```

```
prob0 = 1/(1+exp(-logit0));
mu = exp(b0 + pe1*espSI + pe2*espSL + pe3*espSP + pf*log_TC +
pob1*pobCB + pob2*pobCP + pob3*pobCT + pob4*pobFM + pob5*pobSS);
    m = 1/alpha;
    p = 1/(1+alpha*mu);
    if TP_suma = 0 then
     ll = log(prob0 + (1-prob0)*(p**m));
    else ll = log(1-prob0) + lgamma(m + TP_suma) - lgamma(TP_suma + 1)
 - lgamma(m) + m*log(p) + TP_suma*log(1-p);
model TP_suma ~ general(ll);
estimate 'inflation p' prob0;
run;

*3) especie  +  especie als zeros; *ZIBN3;
proc genmod data=dades  plots= RESCHI(index xbeta);
class esp2 ;
model TP_suma = esp2 /  offset=log_TC dist = zinb type3;
zeromodel esp2 / link= logit;
lsmeans esp2 / diff adjust=tukey ilink;
run;

*Alternativa: programació manual;
proc nlmixed data = dades tech=NEWRAP;
parms b0= -2.2467 pf=1 pe1=-0.4695 pe2=0.9927 pe3=0.1167
  a0=0.7842 zpe1=-3.7427 zpe2=-2.6301 zpe3=0.8241 alpha= 0.8720;
bounds 0.99999 < pf < 1.00001;
logit0 = a0 + zpe1*espSI + zpe2*espSL + zpe3*espSP;
prob0 = 1/(1+exp(-logit0));
mu = exp(b0 + pe1*espSI + pe2*espSL + pe3*espSP + pf*log_TC);
    m = 1/alpha;
    p = 1/(1+alpha*mu);
    if TP_suma = 0 then
     ll = log(prob0 + (1-prob0)*(p**m));
    else ll = log(1-prob0) + lgamma(m + TP_suma) - lgamma(TP_suma + 1)
 - lgamma(m) + m*log(p) + TP_suma*log(1-p);
model TP_suma ~ general(ll);
estimate 'Inflation SI' 1/(1+exp(-(a0 + zpe1)));
estimate 'Inflation SL' 1/(1+exp(-(a0 + zpe2)));
estimate 'Inflation SP' 1/(1+exp(-(a0 + zpe3)));
estimate 'Inflation SV' prob0;
run;

*4) especie i localització  +  especie als zeros; *ZIBN4;
proc genmod data=dades  plots= RESCHI(index xbeta);
class esp2 pob2;
model TP_suma = esp2 pob2/  offset=log_TC dist = zinb type3;
zeromodel esp2 / link= logit;
lsmeans esp2 / diff adjust=tukey ilink;
```

```
run;

*Alternativa: programació manual;
proc nlmixed data = dades tech=NEWRAP;
parms b0= -1.7400 pf=1 pe1=-0.6754 pe2=1.0848 pe3=0.0983
pob1=-0.3990 pob2=-0.8718 pob3=-0.6880 pob4=-0.3448 pob5=-1.0759
      a0=0.8127 zpe1=-3.9023 zpe2=-2.2445 zpe3=0.8046 alpha= 0.6399;
bounds 0.99999 < pf < 1.00001;
logit0 = a0 + zpe1*espSI + zpe2*espSL + zpe3*espSP;
prob0 = 1/(1+exp(-logit0));
mu = exp(b0 + pe1*espSI + pe2*espSL + pe3*espSP + pf*log_TC +
pob1*pobCB + pob2*pobCP + pob3*pobCT + pob4*pobFM + pob5*pobSS);
    m = 1/alpha;
    p = 1/(1+alpha*mu);
    if TP_suma = 0 then
     ll = log(prob0 + (1-prob0)*(p**m));
    else ll = log(1-prob0) + lgamma(m + TP_suma) - lgamma(TP_suma + 1)
 - lgamma(m) + m*log(p) + TP_suma*log(1-p);
model TP_suma ~ general(ll);
estimate 'Inflation SI' 1/(1+exp(-(a0 + zpe1)));
estimate 'Inflation SL' 1/(1+exp(-(a0 + zpe2)));
estimate 'Inflation SP' 1/(1+exp(-(a0 + zpe3)));
estimate 'Inflation SV' prob0;
run;

/*****************************************************************/
*MODELS AMB POBLACIO COM A FACTOR ALEATORI;
/*****************************************************************/

*poisson;
proc sort data=dades;
by pob2;
proc glimmix data=dades;
class esp2 pob2;
model TP_suma = esp2  / offset=log_TC dist = poisson s;
random intercept / subject = pob2 s;
run;


*binomial negatiu;
proc sort data=dades;
by pob2;
proc glimmix data=dades;
class esp2 pob2;
model TP_suma = esp2  / offset=log_TC dist = negbin s;
random intercept / subject = pob2 s;
run;
*no converge;
```

```
*ZIP;
proc sort data=dades;
by pob2;
proc nlmixed data = dades tech=newrap;
parms b0=-1.7794 pf=1 pe1= -1.4441 pe2= 0.8184 pe3= -0.6992 sigma2=0.2
  a0= 1.2328 zpe1= -3.3572 zpe2=-2.1137 zpe3=  0.4080;
  bounds 0.99999 < pf < 1.01;
logit0 = a0 + zpe1*espSI + zpe2*espSL + zpe3*espSP;
prob0 = 1/(1+exp(-logit0));
mu = exp(b0 + u + pe1*espSI + pe2*espSL + pe3*espSP + pf*log_TC);
if TP_suma = 0 then
ll = log(prob0 + (1- prob0)*exp(-mu));
else
ll = TP_suma*log(mu) + log(1- prob0) - mu - lgamma(TP_suma + 1);
model TP_suma ~ general(ll);
random u ~ normal(0, sigma2) subject=pob2;
estimate 'Inflation SI' 1/(1+exp(-(a0 + zpe1)));
estimate 'Inflation SL' 1/(1+exp(-(a0 + zpe2)));
estimate 'Inflation SP' 1/(1+exp(-(a0 + zpe3)));
estimate 'Inflation SV' prob0;
estimate 'random effect' sigma2;
run;


*ZINB;
proc sort data=dades;
by pob2;
proc nlmixed data = dades tech=newrap;
parms b0=-1.7794 pf=1 pe1= -1.4441 pe2= 0.8184 pe3= -0.6992 sigma2=0.2
  a0= 1.2328 zpe1= -3.3572 zpe2=-2.1137 zpe3=0.4080 alpha=1;
  bounds 0.99 < pf < 1.00001;
logit0 = a0 + zpe1*espSI + zpe2*espSL + zpe3*espSP;
prob0 = 1/(1+exp(-logit0));
mu = exp(b0 + u + pe1*espSI + pe2*espSL + pe3*espSP + pf*log_TC);
    m = 1/alpha;
    p = 1/(1+alpha*mu);
    if TP_suma = 0 then
     ll = log(prob0 + (1-prob0)*(p**m));
    else ll = log(1-prob0) + lgamma(m + TP_suma) - lgamma(TP_suma + 1)
 - lgamma(m) + m*log(p) + TP_suma*log(1-p);
model TP_suma ~ general(ll);
random u ~ normal(0, sigma2) subject=pob2;
estimate 'Inflation SI' 1/(1+exp(-(a0 + zpe1)));
estimate 'Inflation SL' 1/(1+exp(-(a0 + zpe2)));
estimate 'Inflation SP' 1/(1+exp(-(a0 + zpe3)));
estimate 'Inflation SV' prob0;
    estimate 'random effect' sigma2;
```

```
estimate 'Sobredispersió' alpha;
run;

ods rtf close;

/*****************************************************************/
*TEST PER A DETECTAR ZERO INFLACIO;
/*****************************************************************/

*Calculs C Test i R Test;
*Nº de zeros;
proc freq data=dades;
where TP_suma=0;
tables TP_suma;
run;
*Mitjana, total obs i suma total;
proc means data=dades N sum mean;
var TP_suma;
run;


/*****************************************************************/
*VALIDACIÓ DEL MODEL SELECCIONAT: ANALISI DELS RESIDUS;
/*****************************************************************/


*MODEL SELECCIONAT: ZINB4;
ods graphics on;
proc genmod data=dades  plots=RESCHI(index xbeta);
class esp2 pob2;
model TP_suma = esp2 pob2/  offset=log_TC dist = zinb type3;
zeromodel esp2 / link= logit;
output out=res RESCHI=RESCHI PRED=PRED RESRAW=res PZERO=PZERO ;
ods output ParameterEstimates=ParameterEstimates
ZeroParameterEstimates=ZeroParameterEstimates;
run;
ods graphics off;

* % residus fora del (-2, -2);
data res2;
set res;
if reschi >=2 or reschi <= -2 then output;
n= _n_;
run;

proc sgplot data=res;
title "Residus estandarditzats vs Valors Predits";
scatter  x=PRED y=RESCHI ;
```

```
*loess x=PRED y=RESCHI / interpolation=linear;
xaxis label="Valors Predits";
yaxis label="Residus estandarditzats";
run;

*zomm;
proc sgplot data=res;
where pred<40;
title "Residus estandarditzats vs Valors Predits";
scatter  x=PRED y=RESCHI ;
*loess x=PRED y=RESCHI / interpolation=linear;
xaxis label="Valors Predits";
yaxis label="Residus estandarditzats";
run;

proc sgplot data=res;
where pred<30 and reschi< 2;
title "Residus estandarditzats vs Valors Predits";
scatter  x=PRED y=RESCHI / group=esp2 MARKERCHAR=TP_suma;
*loess x=PRED y=RESCHI / interpolation=linear;
xaxis label="Valors Predits";
yaxis label="Residus estandarditzats";
label esp2 = "Espècie";
run;

proc sgplot data=res;
title "Histrograma i corba de densitat Residus estandarditzats";
histogram RESCHI;
density RESCHI;
density RESCHI / type=kernel;
keylegend / location=inside position=topright;
xaxis label="Residus estandarditzats";
yaxis label="Percentatge";
run;

proc sgplot data=res;
title "Valors predits vs Valors observats";
scatter  x=TP_suma y=PRED;
yaxis label="Valors predits";
xaxis label="Nombre Total de Capítols Predats";
run;

proc sgplot data=res;
where pred<40 and TP_suma < 50;
title "Valors predits vs Valors observats";
scatter  x=TP_suma y=PRED;
yaxis label="Valors predits";
xaxis label="Nombre Total de Capítols Predats";
```

```
run;

proc sgplot data=res;
scatter  x=esp2 y=RESCHI;
xaxis label="Valors ajustats";
yaxis label="Residus estandarditzats";
run;

proc sgplot data=res;
scatter  x=pob2 y=RESCHI;
xaxis label="Valors ajustats";
yaxis label="Residus estandarditzats";
run;

proc sgplot data=res;
title "Residus estandarditzats vs Espècie";
Vbox RESCHI / category=esp2;
xaxis label="Espècie";
yaxis label="Residus estandarditzats";
run;
proc sgplot data=res;
where RESCHI < 2 ;
title "Residus estandarditzats vs Espècie";
Vbox RESCHI / category=esp2;
xaxis label="Espècie";
yaxis label="Residus estandarditzats";
run;

proc sgplot data=res;
title "Residus estandarditzats vs Localització";
Vbox RESCHI / category=pob2;
xaxis label="Localització";
yaxis label="Residus estandarditzats";
run;
proc sgplot data=res;
where RESCHI < 2 ;
title "Residus estandarditzats vs Localització";
Vbox RESCHI / category=pob2;
xaxis label="Localització";
yaxis label="Residus estandarditzats";
run;

/*****************************************************************/
*COMPARATIVA DE MODELS;
/*****************************************************************/

%inc "probcounts.sas";
```

```
*POISSON;
* la macro probcounts no admet variables class;
proc genmod data=dades  plots=RESCHI(index xbeta);
model TP_suma = espSI espSL espSP pobCB pobCP pobCT pobFM pobSS /
  offset=log_TC dist =poisson type3;
output out=respoi RESCHI=RESCHI PRED=PREDPOI XBETA =XBETA ;
ods output ParameterEstimates=ParameterEstimates ;
run;

proc corr data=respoi;
var TP_suma XBETA;
run;

%probcounts(version, data=dades, inmodel=ParameterEstimates,
 proc=GENMOD, pred=p, counts=%str(0 to 200), modeloffset=log_TC, out=probcounts)

proc means data=probcounts mean std;
var Pr0-Pr200;
output out=means_prob mean= / autoname;
run;

data poi;
set means_prob;
drop _freq_ _type_;
run;


*NB;
* la macro probcounts no admet variables class;
proc genmod data=dades  plots=RESCHI(index xbeta);
model TP_suma = espSI espSL espSP pobCB pobCP pobCT pobFM pobSS /
  offset=log_TC dist =negbin type3;
output out=resnb RESCHI=RESCHI PRED=PREDNB XBETA =XBETA;
ods output ParameterEstimates=ParameterEstimates ;
run;

proc corr data=resnb;
var TP_suma XBETA;
run;

%probcounts(version, data=dades, inmodel=ParameterEstimates,
 proc=GENMOD, pred=p, counts=%str(0 to 200), modeloffset=log_TC, out=probcounts)

proc means data=probcounts mean std;
var Pr0-Pr200;
output out=means_prob mean= / autoname;
run;
```

```
data negbin;
set means_prob;
drop _freq_ _type_;
run;

*ZIP;
* la macro probcounts no admet variables class;
proc genmod data=dades  plots=RESCHI(index xbeta);
model TP_suma = espSI espSL espSP pobCB pobCP pobCT pobFM pobSS /
  offset=log_TC dist = zip type3;
zeromodel espSI espSL espSP / link= logit;
output out=reszip RESCHI=RESCHI PRED=PREDZIP PZERO=PZEROZIP XBETA =XBETA;
ods output ParameterEstimates=ParameterEstimates
ZeroParameterEstimates=ZeroParameterEstimates;
run;

proc corr data=reszip;
var TP_suma XBETA;
run;

%probcounts(version, data=dades, inmodel=ParameterEstimates,
inzeromodel=ZeroParameterEstimates,
 proc=GENMOD, pred=p, counts=%str(0 to 200), modeloffset=log_TC,
zerolink=logistic, out=probcounts)

proc means data=probcounts mean std;
var Pr0-Pr200;
output out=means_prob mean= / autoname;
run;

data zip;
set means_prob;
drop _freq_ _type_;
run;


*ZINB;
* la macro probcounts no admet variables class;
proc genmod data=dades  plots=RESCHI(index xbeta);
model TP_suma = espSI espSL espSP pobCB pobCP pobCT pobFM pobSS /
  offset=log_TC dist = zinb type3;
zeromodel espSI espSL espSP / link= logit;
output out=reszinb RESCHI=RESCHI PRED=PREDZIP PZERO=PZEROZIP XBETA =XBETA;
ods output ParameterEstimates=ParameterEstimates
ZeroParameterEstimates=ZeroParameterEstimates;
run;

proc corr data=reszinb kendall;
```

```
var TP_suma XBETA;
run;

%probcounts(version, data=dades, inmodel=ParameterEstimates,
 inzeromodel=ZeroParameterEstimates,
 proc=GENMOD, pred=p, counts=%str(0 to 200), modeloffset=log_TC,
zerolink=logistic, out=probcounts)

proc means data=probcounts mean std;
var Pr0-Pr200;
output out=means_prob mean= / autoname;
run;

data zinb;
set means_prob;
drop _freq_ _type_;
run;

 proc freq data=dades;
    table TP_suma / out=obs;
 run;

data means;
set poi negbin zip zinb;
run;

proc transpose data=means out=tmeans;
run;

data tmeans;
set tmeans ;
if _n_ > 24 then delete;
run;

 data allpred;
    merge obs(where=(TP_suma<=23)) tmeans;
    obs=percent/100;
 run;

*Grafic de probabilitats;
proc sgplot;
 title;
    yaxis label='Probabilitat';
    xaxis label='Nombre Total de Capítols Predats';
    series y=obs  x=TP_suma / name='obs' legendlabel='Observats'
       lineattrs=(color=black thickness=2px);
    series y=col1 x=TP_suma / name='Poi2' legendlabel='Poi2'
       lineattrs=(color=blue thickness=2px);
```

```
    series y=col2 x=TP_suma/ name='NB2' legendlabel='BN2'
       lineattrs=(color=red thickness=2px);
    series y=col3 x=TP_suma/ name='ZIP4' legendlabel='ZIP4'
       lineattrs=(color=blue pattern=2 thickness=2px);
    series y=col4 x=TP_suma/ name='ZINB4' legendlabel='ZINB4'
       lineattrs=(color=red pattern=3 thickness=2px);
    discretelegend 'Poi2' 'ZIP4' 'NB2' 'ZINB4' 'obs' / title='Model:'
       location=inside position=ne across=2 down=3;
 run;


*Test de Vuong per a models no aniuats;
 %inc "vuong.sas";

*COMPARATIVA POI2 VS NB2;
proc genmod data=dades  plots=RESCHI(index xbeta);
model TP_suma = espSI espSL espSP pobCB pobCP pobCT pobFM pobSS /
  offset=log_TC dist =poisson type3;
output out=respoi RESCHI=RESCHI PRED=PREDPOI XBETA =XBETA ;
ods output ParameterEstimates=ParameterEstimates ;
run;
proc genmod data=respoi  plots=RESCHI(index xbeta);
model TP_suma = espSI espSL espSP pobCB pobCP pobCT pobFM pobSS /
  offset=log_TC dist =negbin type3 ;
output out=resnb RESCHI=RESCHI PRED=PREDNB XBETA =XBETA;
ods output ParameterEstimates=ParameterEstimates ;
run;

%vuong(data=resnb, response=TP_suma,
       model1=negbin,  p1=PREDNB, dist1=nb, scale1=3.1236,
       model2=Poisson, p2=predpOI,   dist2=poi,
       nparm1=10, nparm2=9)


*COMPARATIVA POI2 VS ZIP4;
proc genmod data=dades  plots=RESCHI(index xbeta);
model TP_suma = espSI espSL espSP pobCB pobCP pobCT pobFM pobSS /
  offset=log_TC dist =poisson type3;
output out=respoi RESCHI=RESCHI PRED=PREDPOI XBETA =XBETA ;
ods output ParameterEstimates=ParameterEstimates ;
run;
proc genmod data=respoi  plots=RESCHI(index xbeta);
model TP_suma = espSI espSL espSP pobCB pobCP pobCT pobFM pobSS /
  offset=log_TC dist = zip type3;
zeromodel espSI espSL espSP / link= logit;
output out=reszip RESCHI=RESCHI PRED=PREDZIP PZERO=PZEROZIP XBETA =XBETA;
ods output ParameterEstimates=ParameterEstimates
ZeroParameterEstimates=ZeroParameterEstimates;
```

```
run;

%vuong(data=reszip, response=TP_suma,
       model1=zip,  p1=predzip, dist1=zip, pzero1=PZEROZIP,
       model2=Poisson, p2=predp0I,   dist2=poi,
       nparm1=13, nparm2=9)


*COMPARATIVA NB2 VS ZINB4;
proc genmod data=dades  plots=RESCHI(index xbeta);
model TP_suma = espSI espSL espSP pobCB pobCP pobCT pobFM pobSS /
  offset=log_TC dist =negbin type3 ;
output out=resnb RESCHI=RESCHI PRED=PREDNB XBETA =XBETA;
ods output ParameterEstimates=ParameterEstimates ;
run;
proc genmod data=resnb  plots=RESCHI(index xbeta);
model TP_suma = espSI espSL espSP pobCB pobCP pobCT pobFM pobSS /
  offset=log_TC dist = zinb type3;
zeromodel espSI espSL espSP / link= logit;
output out=reszinb RESCHI=RESCHI PRED=PREDZINB PZERO=PZEROZINB XBETA =XBETA;
ods output ParameterEstimates=ParameterEstimates
ZeroParameterEstimates=ZeroParameterEstimates;
run;

%vuong(data=reszinb, response=TP_suma,
       model1=zinb,  p1=predzinb, dist1=zinb, scale1=0.6399, pzero1=PZEROZINB,
       model2=NegBin, p2=PREDNB, scale2= 3.1236, dist2=nb,
       nparm1=14, nparm2=10)

*COMPARATIVA ZIP4 VS ZINB4;
proc genmod data=respoi  plots=RESCHI(index xbeta);
model TP_suma = espSI espSL espSP pobCB pobCP pobCT pobFM pobSS /
  offset=log_TC dist = zip type3;
zeromodel espSI espSL espSP / link= logit;
output out=reszip RESCHI=RESCHI PRED=PREDZIP PZERO=PZEROZIP XBETA =XBETA;
ods output ParameterEstimates=ParameterEstimates
ZeroParameterEstimates=ZeroParameterEstimates;
run;
proc genmod data=reszip  plots=RESCHI(index xbeta);
model TP_suma = espSI espSL espSP pobCB pobCP pobCT pobFM pobSS /
  offset=log_TC dist = zinb type3;
zeromodel espSI espSL espSP / link= logit;
output out=reszinb RESCHI=RESCHI PRED=PREDZINB PZERO=PZEROZINB XBETA =XBETA;
ods output ParameterEstimates=ParameterEstimates
ZeroParameterEstimates=ZeroParameterEstimates;
run;

%vuong(data=reszinb, response=TP_suma,
```

```
model1=zinb,  p1=predzinb, dist1=zinb, scale1=0.6399, pzero1=PZEROZINB,
model2=zip,  p2=predzip, dist2=zip, pzero2=PZEROZIP,
nparm1=14, nparm2=13)
```

Codi R: Models Zipfian

En aquest apèndix es mostra el codi R utilitzat per generar els resultats presentats en el capítol 7.

```
# MODELS ZIPFIAN
# ANABEL BLASCO-MORENO
# TFM

library("VGAM")
library("VGAMdata")
library("pscl")
library("sqldf")
library("tcltk")

dades <- read.csv("Dades agregat.csv",header=TRUE,sep=";")
summary(dades)
dades$especie <- factor(as.character(dades$especie), levels=c("SV","SI","SL","SP"))
dades$poblacio <- factor(as.character(dades$poblacio),
levels=c("VF","CB","CP","CT","FM","SS"))
summary(dades)

dades$Num <- seq(1:475)
dades$TP_1 <- dades$TP_suma + 1
dades$TC_1 <- dades$TC_suma + 1
dades$lnTC_1 <- log(dades$TC_suma+1)
dades$weight<-1   #Pesos individuals

#Taula de frequencies: frequencies dels valors de la variable TP_suma +1
n_dades<-as.data.frame(table(dades[,13]))
x<-sort(unique(rev(dades[,13])))
Freq<-as.numeric(n_dades[,2])
Freq2<-tabulate(as.numeric(dades[,13]))

#Prenem logaritme dels valors i les probabilitats
Prob<-Freq[which(Freq>0)]/sum(Freq)
log_x <- log(x)
log_Prob <-log(Prob)

#Diagrama de dispersió en escala log-log
plot(log_x,log_Prob,xlab="ln(Nº Total de Capitols Predats+1)",ylab="ln P(y)")
```

```
#AJUST DE LA DISTRIBUCIÓ ZIPF MANUAL

N<-sum(Freq) # mida mostra

#Logaritme de la versemblança d'una Zipf per dades agregades
fnZ<-function(alpha,x,Freq){
n<-sum(Freq)
-n*log(zeta(alpha))-alpha*sum(Freq*log(x))
}

#Maximitzem el logaritme de la versemblança d'una Zipf
l_Z<-optimize(fnZ,x=x,Freq=Freq, c(1, 100), maximum=T, tol=0.00001)
l_Z
# parametre estimat
a_Z<-l_Z$maximum
a_Z

#AJUST DE LA DISTRIBUCIÓ ZIPF AMB EL PAQUET VGAM
?vglm.fit

fit <- vglm(x ~ 1, zipf(link = loge, N=N),weight=Freq, trace = TRUE, log = FALSE)
fit
coef(fit)
coef(summary(fit))[,"Std. Error"]
fit@misc$N
shat <- Coef(fit)
shat
AIC(fit)
BIC(fit)


#Gràfic en escala log-log
#Probabilitats observades vs probabilitats esperades per a la distribució Zipf
plot(log_x,log_Prob, xlab="ln(Nº Total de Capitols Predats+1)",
ylab="ln P(y)", main="Ajust distribució Zipf")
abline(a=-log(zeta(shat)), b=-shat, col=2)
legend("topright", legend=c(as.expression(bquote("Zipf(" ~ hat(alpha)
==.(shat)~ ")"))), col=2,lty=2)


#COMPROVACIÓ: sense agregar les dades, dóna el mateix resultat.
#Dades sense agregar, manual
l_Z<-optimize(fnZ,x=dades$TP_1,Freq=dades$weight, c(1, 100), maximum=T, tol=0.00001)
l_Z
# parametre estimat
a_Z<-l_Z$maximum
a_Z
```

```
#Amb la funció vglm de R
fit <- vglm(TP_1 ~ 1, zipf(link = loge, N=N), weight=weight, dades, trace = TRUE,
 log=FALSE)
fit
fit@misc$N
(shat <- Coef(fit))
with(dades, weighted.mean(TP_1 , weight))
fitted(fit, matrix = FALSE)

#Funcio de probabilitat d'una distribucio Zipf(alpha)
zeta_x<-function(alpha,x){
aux<-0
if(x==1) {
zeta(alpha)
} else{
zeta(alpha)-sum((1:(x-1))^(-alpha))
}}

#Càlcul freqüències esperades
PY<-function(a,x){
x^(-a)*zeta(a)/((zeta(a)-zeta_x(a,x))*(zeta(a)-zeta_x(a,x+1)))
}

taula<-matrix(c(rep(NA,max(x)*4)),ncol=4)

for(i in 1:(max(x))){
auxX<-N*PY(a_Z,i+1)
taula[i,]<-c(i,Freq2[i],auxX,(Freq2[i]-auxX)^2/auxX)
}

#Bondat d'ajust
#S'agrupen les categories més petites, a partir de la freqüència 25
taula2<- matrix(c(rep(NA,25*4)),ncol=4)
taula2[1:(24),]<-taula[1:(24),]
taula2[25,]<-c(25,sum(taula[25:max(x),2]),sum(taula[25:max(x),3]),
(sum(taula[25:max(x),2])-sum(taula[25:max(x),3]))^2/sum(taula[25:max(x),3]))
colnames(taula2)<-colnames(taula)

#Test Khi-quadrat de Pearson
gdlX<-dim(taula2)[1]-1-1
X2X<-sum(taula2[,4])
X2X; 1-pchisq(X2X,gdlX); 1-pnorm(X2X,gdlX,sqrt(2*gdlX))

#S'agrupen les categories més petites, a partir de la freqüència 11
taula2<- matrix(c(rep(NA,12*4)),ncol=4)
taula2[1:(11),]<-taula[1:(11),]
taula2[12,]<-c(12,sum(taula[12:max(x),2]),sum(taula[12:max(x),3]),
```

```
(sum(taula[12:max(x),2])-sum(taula[12:max(x),3]))^2/sum(taula[12:max(x),3]))
colnames(taula2)<-colnames(taula)

#Test Khi-quadrat de Pearson
gdlX<-dim(taula2)[1]-1-1
X2X<-sum(taula2[,4])
X2X; 1-pchisq(X2X,gdlX); 1-pnorm(X2X,gdlX,sqrt(2*gdlX))


##########################################
# MODELS ZIPFIAN #
##########################################

#Model amb només terme independent

N<-sum(dades$weight)
fit1 <- vglm(TP_1 ~ 1, zipf(link = loge, N=N), weight=weight, trace = TRUE,
log=FALSE, dades)
fit1
coef(summary(fit1))
shat1 <- Coef(fit1)
shat1
AIC(fit1)
BIC(fit1)

#Amb variable offset.
fit1 <- vglm(TP_1 ~ 1, offset=lnTC_1, zipf(link = loge, N=N), weight=weight,
trace = TRUE, log=FALSE, dades)
fit1
coef(summary(fit1))
shat1 <- Coef(fit1)
shat1
AIC(fit1)
BIC(fit1)

#Modelo amb només especie i offset
fit2 <- vglm(formula = TP_1 ~ especie, offset=lnTC_1,
family = zipf(link = loge,N=N), data = dades, weights = weight,
   trace = TRUE)
fit2
coef(summary(fit2))
AIC(fit2)
BIC(fit2)

#Modelo amb especie i poblacio i offset
fit3 <- vglm(formula = TP_1 ~ poblacio + especie, offset=lnTC_1 ,
family = zipf(link = loge,N=N), data = dades, weights = weight,
   trace = TRUE)
```

```
fit3
coef(summary(fit3))
AIC(fit3)
BIC(fit3)
```

Bibliografia

[1] Aitchison, J. *On the distribution of a positive random variable having a discrete probability mass at the origin*. Journal of the American Statistical Association, 1955, 50, 901-908.

[2] Akaike, H. *A new look at the statistical model identification*. IEEE Transactions on Automatic Control, 1974, 19 (6), 716-723.

[3] Angers, J.F.; Biswas, A. *A Bayesian analysis of zero-inflated generalized Poisson model*. Computational Statistics & Data Analysis,2003, 42, 37-46.

[4] Boswell, M.T.; Patil, G.P. *Chance mechanisms generating the negative binomial distribution, Random Counts in Scientific Work*, Vol. 1: *Random counts in Models and Structures*, University Park: Pennsylvania State University Press, 1970.

[5] Cameron, A.C.; Trivedi, P.K. *Regression Analysis of Count Data*. Cambridge University Press, 1998.

[6] Clarke, K.A. *A Simple Distribution-Free Test for Nonnested Hypotheses*. Political Analysis, 2007, 15:3.

[7] Dobbie, M.J.; Welsh, A.H. *Models for zero-inflated count data using the Neyman type A distribution*. Statistical Modelling, 2001, 1, 65-80.

[8] El-Shaarawi, A.H. *Some Goodness-of-Fit Methods for the Poisson Plus Added Zeros Distribution*. Applied and Environmental Microbiology, 1985, 1304-1306.

[9] Famoye, F.; Signh, K.P. *Zero-Inflated Generalized Poisson Regression Model with an Application to Domestic Violence Data*. Journal of Data Science, 2006, 4, 117-130.

[10] Fletcher, D.; Mackenzie, D.I.; Villouta, E. *Modelling skewed data with many zeros: a simple approach combining ordinary and logistic regression*. Environmental and Ecological Statistics, 2005, 12, 45-54.

[11] Greene, W. H. *Accounting for excess zeros and sample selection in Poisson and negative binomial regression models*. New York University, Department of Economics Working Paper, 1994, 94-10.

[12] Greenwood, M.; Yule, G.U. *An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the ocurrence of multiple attacks of disease or of repeated accidents*. Journal of the Royal Statistical Society, Series A 1920, 83, 255-279.

[13] Gurland, J. *Some interrelations among compound and generalized distributions*. Biometrika, 1957, 44, 265-268.

[14] Hall, D.B. *Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study*. Biometrics, 2000, 56, 1030-1039.

[15] Heilbron, D.C. *Zero-altered and other regression models for count data with added zeros*. Biometrical Journal, 1994, 36, 531-547.

[16] Hilbe, J.M. *Negative Binomial Regression*. Cambridge, 2011.

[17] Johnson, N.L.; Kotz, S. *Distributions in Statistics: Discrete Distributions*. Haughton Mifflin, 1969.

[18] Johnson, N.L.; Kotz, S.; Kemp, A. W. *Univariate Discrete Distributions*. John Wiley & Sons, 2005.

[19] Keane, R.M.; Crawley, M.J. *Exotic plant invasions and the enemy release hypothesis*. Trends in Ecology and Evolution, 2002, 17, 164-170.

[20] Kuhnert, P.M.; Martin, T.G.; Mengersen, K.; Possingham, H.P.; *Assessing the impacts of grazing levels on bird density in woodland habitat: a Bayesian approach using expert opinion.* Environmetrics, 2005, 16, 1-31.

[21] Lambert, D. *Zero-inflated Poisson regression with an application to defects in manufacturing.* Technometrics, 1992, 34, 1-14.

[22] Long, J. S. *Regression Models for Categorical and Limited Dependent Variables.* Thousand Oaks, CA: Sage Publications, 1997.

[23] Lüders, R. *Die Statistik der seltenen Ereignisse.* Biometrika, 1934, 26, 108-128.

[24] MacKenzie, D.I.; Nichols, J.D.; Lachman, G.B.; Droege, S.; Royle, J.A.; Langtimm, C. *Estimating site occupancy rates when detection probabilities are less than one.* Ecology, 2002, 83, 2248-2255.

[25] Marshall, A.W.; Olkin, I. *A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families*, Biometrika, 1997. 84, 641-652.

[26] Martin, T.G.; Kuhnert, P.M.; Mengersen, K.; Possingham, H.P. *The power of expert opinion in ecological models using bayesian methods: impact of grazing on birds.* Ecological Applications, 2005, 15(1), 266-280.

[27] Martin, T.G.; Wintle, B.A.; Rhodes, J.R.; Kuhnert, P.M.; Field, S.A.; Low-Choy, S.J.; Tyre, A.J.; Possingham, H.P. *Zero tolerance ecology: improving ecological inference by modelling the source of zero observations.* Ecology Letters, 2005, 8, 1235-1246.

[28] McCullagh, P.; Nelder, J.A. *Generalised Linear Models.* Chapman and Hall, 2nd ed. 1989.

[29] O'Hara, R.B.; Kotze, D.J. *Do Not Log-transform Count Data.* Methods in Ecology & Evolution 2010, 1, 118-122.

[30] Neyman, J.; Pearson, E.S. *On the use and interpretation of certain test criteria for purposes of statistical inference.* Biometrika, 1928, 20A, 263-294.

[31] Pérez-Casany, M.; Casellas, A. *Marshall-Olkin Extended Zipf Distribution*, arXiv:1304.4540, 2013.

[32] Ridout, M.; Demetrio, C. G. B.; Hinde, J. *Models for count data with many zeros.* Invited paper presented at the Nineteenth International Biometric Conference, Cape Town, South Africa, 1998, 179-190.

[33] Russell, D.W. *Fiting Nonlinear Mixed Models with the New NLMIXED Procedure.* SAS Institute Inc., 1999.

[34] Stefansson, G. *Analysis of ground fish survey abundance data: combining the GLM and delta approaches.* ICES Journal of Marine Science, 1996, 53, 577-588.

[35] Schwarz, G. E. *Estimating the dimension of a model.* Annals of Statistics, 1978, 6 (2), 461-464

[36] Valero, J.; Pérez-Casany, M.; Ginebra, J. *On Poisson-Stopped-Sums that are Mixed Poisson.* Statistics and Probability Letters, 2013, v.83 i.8, 1830-1834.

[37] Vuong, Q. *Likelihood ratio tests for model selection and non-nested hypotheses.* Econometrica, 1989, 57, 307-334.

[38] Welsh, A.H.; Cunningham, R.B.; Donnelly, C.F.; Lindenmayer, D.B. *Modelling the abundance of rare species: statistical models for counts with extra zeros.* Ecological Modelling, 1996, 88, 297-308.

[39] Welsh, A.H.; Cunningham, R.B.; Chambers, R.; *Methodology for estimating the abundance of rare animals: seabird nesting on North East Herald Cay.* Biometrics, 2000, 56, 22-30.

[40] Yee, T. W.; *The VGAM package.* R News, 2008, 8 (2), 28-39. URL http://CRAN.R-project.org/doc/Rnews/.

[41] Zuur, A. F.; Ieno, E. I.; Walker, N. J.; Saveliev, A. A.; Smith, G. M. *Mixed Effects Models and Extensions in Ecology with R.* Springer, 2009.