What does a zero mean? Understanding false, random and structural zeros in ecology

Anabel Blasco-Moreno[a,c], Marta Pérez-Casany[b], Pedro Puig[c], Maria Morante[d],

Eva Castells[d,e,*]

[a]Servei d'Estadística Aplicada, Univ. Autònoma de Barcelona, Cerdanyola del Vallès 08193, Spain

[b]Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, Spain

[c]Departament de Matemàtiques, Univ. Autònoma de Barcelona, Cerdanyola del Vallès 08193, Spain

[d]Departament de Farmacologia, Terapèutica i Toxicologia, Univ. Autònoma de Barcelona, Cerdanyola del Vallès 08193, Spain

[e]CREAF, Cerdanyola del Vallès 08193, Spain

*Corresponding author

Email addresses: anabel.blasco@uab.cat (Anabel Blasco-Moreno), marta.perez@upc.edu (Marta Pérez-Casany), ppuig@mat.uab.cat (Pedro Puig), maria.morante@uab.cat (Maria Morante), eva.castells@uab.cat (Eva Castells)

Abstract

1. Zeros (i.e. events that do not happen) are the source of two common phenomena in count data: overdispersion and zero-inflation. Zeros have multiple origins in a dataset: false zeros occur due to errors in the experimental design or the observer; structural zeros are related to the ecological or evolutionary restrictions of the system under study; and random zeros are the result of the sampling variability. Identifying the type of zeros and their relation with overdispersion and/or zero inflation is key to select the most appropriate statistical model.

2. Here we review the different modelling options in relation to the presence of overdispersion and zero inflation, tested through the dispersion and zero inflation indices. We then examine the theory of the Zero-inflated (ZI) models and the use of the score tests to assess overdispersion and zero inflation over a model.

3. In order to choose an adequate model when analyzing count data we suggest the following protocol: Step 1) classify the zeros and minimize the presence of false zeros; Step 2) identify suitable covariates; Step 3) test the data for overdispersion and zero-inflation; and Step 4) choose the most adequate model based on the results of step 3 and use score tests to determine whether more complex models should be implemented.

4. We applied the recommended protocol on a real dataset on plant-herbivore interactions to evaluate the suitability of six different models (Poisson, NB and their zero-inflated versions –ZIP, ZINB–). Our data was overdispersed and zero-inflated, and the ZINB was the model with the best fit, as predicted.

5. Ignoring overdispersion and/or zero inflation during data analyses caused biased estimates of the statistical parameters and serious errors in the interpretation of the results. Our results are a clear example on how the conclusions of an ecological hypothesis can change depending on the model applied. Understanding how zeros arise in count data, for example identifying the potential sources of structural zeros, is essential to select the best statistical design. A good model not only fits the data correctly but also takes into account the idiosyncrasies of the biological system.

Keywords: Enemy release hypothesis, false zeros, overdispersion, plant-herbivore interactions, random zeros, structural zeros, zero inflation, zero-inflated models

Resum

1. Els zeros (és a dir, successos que no s'esdevenen) són la font de dos fenòmens comuns en les dades de recompte: la sobredispersió i la zero inflació. L'origen dels zeros pot ser divers: els zeros falsos són el resultat d'errors en el disseny experimental o en l'observador; els zeros estructurals es relacionen amb les restriccions ecològiques o evolutives del sistema d'estudi; i els zeros aleatoris s'esdevenen per la variabilitat en el mostreig. Identificar els tipus de zeros i la seva relació amb la sobredispersió i/o la zero inflació és clau per seleccionar el model estadístic més apropiat.

2. En aquest article hem revisat les diferents opcions per modelar dades amb sobredispersió i zero inflació, característiques que hem determinat mitjançant els índex de sobredispersió i zero inflació. Hem revisat la teoria dels models zero-inflats (ZI) i l'ús dels score tests per determinar sobredispersió i zero inflació sobre un model.

3. Per tal d'escollir el model estadístic més adequat quan analitzem dades de recomptes, suggerim aplicar el següent protocol: Pas 1) classificar els zeros i minimitzar la presència de zeros falsos; Pas 2) identificar les covariables adequades; Pas 3) comprovar si les dades estan sobredispersades o zero inflades; i Pas 4) escollir el model estadístic més adequat en base als resultats obtinguts al pas 3, i aplicar els score tests per determinar si cal implementar altres models més complexes.

4. Hem aplicat el protocol recomanat en unes dades reals d'interaccions planta-herbívor per avaluar l'adequació de sis models diferents (Poisson, NB i les seves versions zero-inflades–ZIP, ZINB–). Les dades estaven sobredipersades i zero inflades, i el model ZINB oferia el millor ajust, tal com preveiem.

5. Quan ignoravem la sobredipsersió i/o la zero inflació en l'anàlisi de les dades l'estima dels paràmetres estadístics resultava esbiaixada, fet que provocava errors seriosos en la interpretació dels resultats. Els nostres resultats són un clar exemple de com les conclusions d'una hipòtesi ecològica poden canviar depenent del model estadístic aplicat. Per seleccionar el millor disseny estadístic és essencial entendre com es generen els zeros, per exemple identificant fonts potencials de zeros estructurals. Un bon model estadístic no només s'ha d'ajustar a les dades correctament sinó que també ha de contemplar les idiosincràsies del sistema biològic.

1. Introduction

Count data is common in ecology (e.g. the number of plants infected by a disease, the number of eggs in a nest, or the number of individuals in a plot) and frequently exhibit two characteristics: first, a variance significantly larger than the mean, known as *overdispersion* (Bliss & Fisher, 1953); and second, an excess of zero values in comparison to those expected from a classical count probability distribution, known as *zero inflation* (Heilbron, 1994). These two phenomena are related to each other because the excessive number of zeros also contributes to the data overdispersion. Ignoring overdispersion and zero inflation entails biased parameter estimates and overestimation of standard errors (Lambert, 1992; MacKenzie et al., 2002) resulting in the selection of excessively complex models and poor ecological inference.

A zero value may have different meanings in a dataset. The literature differentiates between *true* and *false* zeros (Martin et al., 2005; Kuhnert et al., 2005; Zuur et al., 2009). False zeros correspond to observer errors (e.g. sampling errors due to poor experience of the observer) or to errors in the experimental design (e.g. sampling at the wrong time or place) (Fig. 1). An extensive knowledge of the studied system together with a good experimental design minimizes such zeros. Nevertheless, these zeros should be monitored and removed from the exploratory data analyses when possible. On the other hand, true zeros are related to the nature of the process. These can be classified as two types: *structural* zeros and *random* zeros (He et. al., 2014; Tang et. al., 2018). Structural zeros are zeros associated to the ecological or evolutionary restrictions of the biological system under study (e.g. an herbivore does not interact with a plant species because it prefers to feed on another host), and are the source of zero inflation (Fig. 1). Finally, random zeros emerge due to the sampling variability and include those events that could potentially occur but do not (e.g. a host plant is not attacked by an herbivore when both species are co-occurring because the herbivore population is not large enough to interact with all individual plants). On occasions, the difference between a false zero due to design errors and a structural zero can be subtle. For example, consider the absence of an insect species on a plant that is no host. Are these zeros false or structural? Indeed, both options could be possible depending on the previous knowledge of the biological system and the hypotheses being tested. Only when the study includes the possibility of a zero value as part of the hypotheses (e.g. the study aims to test whether the interaction is occurring) the resulting zeros would be structural. Hence, identifying the source of zeros will be key in order to select the most appropriate statistical model.

A common but inadequate approach to the analysis of count data is to log-transform the response variable in order to reduce data overdispersion, and apply multiple linear regression or ANOVA models. Log transformation should be avoided because it does not guarantee that the key distributional assumptions (normality and homoscedasticity, among others) are satisfied and it fails to improve the distribution (O'Hara and Kotze, 2010). A more adequate procedure is to use a Generalized Linear Model (GLM; McCullagh & Nelder, 1989) with a Poisson distribution (Table 1). However, the disadvantage of using the Poisson is the restrictive assumption of the *mean-variance relationship*: for a Poisson distributed random variable (r.v.) Y with parameter λ, the variance and the mean are equal , $Var(Y) = E(Y) = \lambda$. By definition, the Poisson cannot deal with overdispersion.

The usual alternative to the Poisson when data is overdispersed is the negative binomial (NB) distribution (McCullagh & Nelder, 1989; Hilbe, 2011) (Table 1). The NB relaxes the variance assumption by modelling the variance as a function to the mean: $Var(Y) = \lambda + \alpha\lambda^p$ where α is a scalar parameter and *p* a specified constant. Type 1 NB sets *p* = 1 and assumes that the variance is a multiple of the mean: $Var(Y) = (1 + \alpha)\lambda$. Type 2 NB sets *p* = 2 and assumes that the variance is a quadratic function of the mean: $Var(Y) = \lambda + \alpha\lambda^2$. The extra NB parameter α, called the *dispersion parameter*, allows increasing the variance with respect to the Poisson distribution with the same mean. Additionally, because the expected number of zeros for the NB is higher than for the Poisson distribution with the same mean, NB can also deal with a moderate excess of random zeros. Other distributions, such as the Hermite or the Neyman Type A, or the Quasi-Poisson method, which is a variance-corrected Poisson regression, might be suitable alternatives to the NB, although they have their own assumptions to be met and are rarely implemented in statistical software.

A simple way to test if a r.v. is overdispersed with respect to a Poisson distribution is through the *dispersion index* (Fisher, 1950), defined as

$$d = Var(Y)/E(Y)$$

A variable is overdispersed when $d > 1$. The dispersion index also allows to discriminate between type 1 and type 2 NB distributions. Given the sample index, $\hat{d} = S^2/\bar{Y}$, k different data groups established by the experimental design (e.g. species, sampling locations, etc.) the type 1 NB will be the suitable distribution when $\hat{d}$ tends to be constant across all groups. Otherwise, type 2 NB should be considered.

When zeros are too abundant to be adjusted by a NB researchers should consider whether the excess zeros are the result of ecological or evolutionary restrictions of the biological system under study, i.e. whether the excess zeros are structural. If the NB do not suffice and there is no reasonable explanation for the presence of structural zeros, that is, if all zeros have to be considered random, other distributions with greater probability at zero should be applied (e.g. Hermite, Polya Aeppli, etc). Otherwise, if the existence of structural zeros is coherent with the biological system the use of Zero-Inflated (ZI) models will be more adequate (Table 1). The ZI models (Zero-Inflated Poisson or ZIP, Lambert, 1992; Zero-Inflated Negative Binomial or ZINB, Greene, 1994) are a mixture of two distributions: a reference count distribution, which models random zeros through a Poisson or NB distribution, and a degenerate distribution at zero, which models the structural zeros. Therefore, in the absence of structural zeros the use of ZI model would not be justified. ZIP models provide a way to model overdispersion due to structural zeros and ZINB also makes possible to model overdispersion not only caused by zero inflation. Another approach to modelling excess zeros is the application of zero-altered (ZA; Mullahy, 1986) models, also called Hurdle models. ZA are two-part models where the zero and non-zero counts are modelled separately, and therefore they are only adequate when the counting process cannot generate zero values.

The presence of excess zero values can be assessed by the *zero inflation index* (Puig & Valero, 2006), defined as $zi_P = 1 + log(p_0)/\lambda,$ being p_0 the probability of the zero value and $E(Y) = \lambda$. If Y is Poisson distributed, then $zi_P = 0$. When the sample index is significantly larger than zero, $\widehat{zi}_P > 0,$ the variable is zero-inflated with respect to a Poisson distribution. We extended this index to a new zero-inflated index applied to two-parameter distributions as follows:

$$zi_{NB} = 1 + \frac{(\sigma^2 - \lambda)log(p_0)}{\lambda^2 log(\frac{\sigma^2}{\lambda})}$$

where λ and σ^2 correspond to the mean and variance of the r.v. If Y is NB distributed then $zi_{NB} = 0$, and if $\sigma^2 = \lambda$, zi_{NB} reduces to zi_P. Values of $\widehat{zi}_{NB} > 0$ correspond to a zero-inflated distribution with respect to the NB and should be adapted by zero-inflated models, either ZIP or ZINB, if zero inflation is due to structural zeros (Table 1). In recent years, several studies have addressed the need to consider zero inflation when modelling count data in ecology, especially when assessing occupancy-abundance relationships (Martin et al., 2005; Sileshi, Hailu & Nyadzi, 2009; Smith, Anderson & Millar, 2012; Sadykova et al., 2017; Williams et al., 2017).

Here we first review the theory of ZI models. Second, we present a protocol to assist with the selection of an adequate statistical model when analyzing count data. Third, we apply the recommended protocol on a real dataset on plant-herbivore interactions to evaluate the suitability of six different models (Poisson, NB, two versions of ZIP and two versions of ZINB) in accordance with the presence of overdispersion and zero inflation. Finally, we discuss the consequences of adjusting suboptimal models.

2. Theory of Zero-inflated mixture models

ZIP and ZINB models

ZI models arise as a mixture of two distributions: a degenerate distribution at zero, f_0, and a reference count distribution, f_R. A r.v. Y with a ZI distribution has a probability mass function (pmf) equal to,

$$f(\mathrm{y}) \coloneqq f(\mathrm{y}|\omega, \theta_R) = \omega f_0 + (1 - \omega) f_R(\mathrm{y}|\theta_R), \qquad \text{eqn 1}$$

where θ_R and ω are model parameters to be estimated. The mixture parameter ω is a weight that represents the probability of observation y coming from the degenerate distribution f_0, i.e. of the observation corresponding to a structural zero. From Johnson and Kotz (2005), we know that eqn 1 can still be sensibly interpreted for $\frac{-\mathrm{f_R}(0|\theta_\mathrm{R})}{1-\mathrm{f_R}(0|\theta_\mathrm{R})} \leq \omega \leq 1$. Negative values of ω result in a zerodeflected distribution while positive values result in a zero-inflated distribution compared with the reference distribution. Here we only consider models with ω ≥ 0.

When f_R is the pmf of a Poisson distribution with parameter λ we obtain, by definition, the ZIP model, which has a pmf equal to:

$$Pr(Y = y|\omega, \lambda) = \begin{cases} \omega + (1-\omega)\, exp(-\lambda), & \text{if } y = 0 \\ (1-\omega)\, exp(-\lambda)\lambda^y/y!, & \text{if } y > 0 \end{cases} \qquad \text{eqn 2}$$

where $0 \leq \omega \leq 1$ and $\lambda > 0$. The ZIP model has $E(Y) = \lambda(1-\omega) = \mu$ and $Var(Y) = \mu + \frac{\omega}{1-\omega}\mu^2$. Since the variance is larger than the mean, the ZIP distribution is overdispersed with respect to the Poisson, and it will be appropriate when overdispersion is due to a large number of zeros.

When there are other sources of overdispersion different from the excess of zeros, a ZINB model could be more appropriate. By the eqn 1, taking f_R the pmf of a type 2 NB, the pmf of the ZINB is equal to,

$$Pr(Y = y|\omega, \alpha, \lambda) = \begin{cases} \omega + (1-\omega)\left(\frac{1}{1+\lambda\alpha}\right)^{\frac{1}{\alpha}}, & \text{if } y = 0 \\ (1-\omega)\frac{\Gamma\left(y+\frac{1}{\alpha}\right)}{\Gamma\left(\frac{1}{\alpha}\right)\cdot\Gamma(y+1)}\left(\frac{1}{1+\lambda\alpha}\right)^{\frac{1}{\alpha}}\left(\frac{\lambda\alpha}{1+\lambda\alpha}\right)^{y}, & \text{if } y > 0 \end{cases} \qquad \text{eqn 3}$$

where $0 \leq \omega \leq 1$, and $\alpha > 0$ and $\lambda > 0$ are the dispersion and mean parameters of the underlying type 2 NB distribution, respectively, and Γ corresponds to a Gamma function. The mean and the variance for the ZINB are equal to: $E(Y) = \lambda(1-\omega) = \mu$ and $Var(Y) = \mu + \left(\frac{\omega}{1-\omega} + \frac{\alpha}{1-\omega}\right)\mu^2$. Note that the overdispersion comes from the ratio $\frac{\omega}{1-\omega}$, related with the proportion of structural zeros, and it also comes from the dispersion parameter of the underlying NB distribution which is related to $\frac{\alpha}{1-\omega}$, both effects being additive.

Incorporating covariates into the model

When we assume a Poisson or a NB or a ZIP or a ZINB distribution for the response variable, the parameters should be considered as a function of the covariates. If λ_i is the expected value of the response variable under the i-th experimental conditions, it can be modelled by the logarithmic link function, then

$$\lambda_i = exp(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_r X_{ri}), \quad i = 1,..,n \qquad \text{eqn 4}$$

where $\boldsymbol{X} = (X_1, \ldots, X_r)$ with $X_j = (X_{j1}, \ldots, X_{jn})^T, j = 1, \ldots, r$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_r)$ are respectively, the covariate matrix and the parameter vector.

Regarding the mixture parameter of ZIP and ZINB models, ω, and given that it is a probability, it makes sense to model it through the covariates using the logistic link, and thus to consider that

$$\omega_i = \frac{exp(\gamma_0 + \gamma_1 Z_{1i} + \ldots + \gamma_k Z_{ki})}{1 + exp(\gamma_0 + \gamma_1 Z_{1i} + \ldots + \gamma_k Z_{ki})}, \quad i = 1,..,n \qquad \text{eqn 5}$$

where $\boldsymbol{Z} = (Z_1, \ldots, Z_k)$ with $Z_j = (Z_{j1}, \ldots, Z_{jn})^T, j = 1, \ldots, k$, and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_k)$ are, respectively, the covariate matrix and parameter vector of the logistic regression. Frequently, the covariates of the logistic expression are a subset of the covariates of the mean parameter. The dispersion parameter, α, is considered constant among the covariates. The relation between the dispersion parameter α of the NB and the dispersion index d is d=α+1 for type 1 NB and d=1+αλ for type 2 NB. Consequently, the dispersion index for type 2 NB is a linear function of the mean (where α is the slope). Estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$, as well as the NB dispersion parameter, can be obtained by maximum likelihood (MLE).

These models can be fitted using SAS and R. However, the available functions or packages in each program and their limitations should be carefully considered before analyses (see Table S1).

Score tests: assessing overdispersion and zero inflation over a model

The dispersion and the zero inflation indices are useful to detect overdispersion or zero inflation over data, respectively, but not over a model with covariates because these can mimic overdispersion and/or zero inflation. When covariates are present, the score tests are a convenient tool. The use of the score tests to ascertain if a zero-inflated model is required is common in other disciplines (e.g. Oliveira et al., 2016). These tests are performed on the dispersion parameter (α) or the zero inflation parameter (ω) without the need to fit overdispersed or zero-inflated models.

A score test for testing Poisson against a NB regression model was developed by Dean and Lawless (1989) by testing $H_0\colon \alpha = 0$ versus $H_1\colon \alpha > 0$. The statistic is defined as

$$T_1 = \sum\nolimits_{i=1}^{n} \left\{ \left(y_i - \hat{\lambda}_i \right)^2 - y_i \right\} \Big/ \left(2 \sum\nolimits_{i=1}^{n} \hat{\lambda}_i^2 \right)^{1/2}$$

where $\hat{\lambda}_i = \lambda_i\left(\boldsymbol{x}_i; \widehat{\boldsymbol{\beta}}\right)$ with $\widehat{\boldsymbol{\beta}}$ the MLE of $\boldsymbol{\beta}$ under the Poisson model. Large positive values of T_1 indicate overdispersion and large negative values underdispersion, relative to a Poisson distribution. Under H_0, this statistic converges in distribution to a N(0,1) as $n \to \infty$.

Similar score test exists for testing a Poisson against a ZIP regression model, van der Broek (1995). The score statistic for testing $H_0\colon \frac{\omega}{1-\omega} = 0$, is defined as

$$S\left(\widehat{\boldsymbol{\beta}}\right) = \frac{\left\{ \sum_{i=1}^{n} \left(\frac{I_{\{y_i=0\}}}{e^{-\widehat{\lambda}_i}} - 1 \right) \right\}^2}{\left\{ \sum_{i=1}^{n} \left(\frac{1}{e^{-\widehat{\lambda}_i}} - 1 \right) \right\} - \hat{\boldsymbol{\lambda}}^{\mathrm{T}} \mathbf{X} \left[\mathbf{X}^{\mathrm{T}} \mathbf{diag}(\hat{\boldsymbol{\lambda}}) \mathbf{X} \right]^{-\mathbf{1}} \mathbf{X}^{\mathrm{T}} \hat{\boldsymbol{\lambda}}}$$

where $\boldsymbol{X}$ is the covariate matrix, $\hat{\lambda}_i = \lambda_i\left(\boldsymbol{x}_i; \widehat{\boldsymbol{\beta}}\right)$ as before and, $I_{\{y_i=0\}}$ is an indicator variable taking the value one when $y_i = 0$ and zero otherwise. Under H_0, this statistic has an asymptotic χ_1^2 distribution.

Finally, Ridout et al. (2001) provide a score test for testing ZIP regression model against a ZINB alternative. The score statistic for testing $H_0\colon \alpha = 0$, is defined as

$$S\left(\widehat{\boldsymbol{\beta}}\right) = \frac{1}{2} \sum\nolimits_{i=1}^{n} \hat{\lambda}_i^{c-1} \left\{ \left[(y_i - \hat{\lambda}_i)^2 - y_i \right] - I_{\{y_i=0\}} \hat{\lambda}_i^2 \widehat{\omega}_i / \hat{p}_{0,i} \right\}$$

where symbols with hats denote estimates under null hypothesis and $\hat{p}_{0,i} = Pr(Y_i = 0)$ from the zero-inflated model. For c=0, the underlying distribution is NB type 1 and, for c=1, the NB type 2.

All these tests require that the mean is modeled through a log-link function. Calculations were performed using SAS System ® v9.4.

Goodness of fit tests

To compare two nested models, such as Poisson versus NB or ZIP versus ZINB, we can use the Likelihood Ratio test (LR), defined as the ratio between the likelihoods of the two models. To compare non-nested models, such as Poisson versus ZIP or NB versus ZINB (the null value -no zero inflation- is on the boundary of the feasible space), we can use the Akaike Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criterion (BIC; Schwarz, 1978). According to these criteria, models with smaller AIC and BIC are considered preferable, although they do not provide a statistical test of comparison. To compare the performance of two different non-nested models the Vuong (Vuong, 1989) and/or the Clarke (Clarke, 2007) tests can be used. A large, positive test statistic provides evidence of the superiority of model 1 over model 2, while a large, negative test statistic is evidence of the superiority of model 2 over model 1. However, be aware that the use of the Vuong test for testing zero-inflation is controversial, because the null value -no zero inflation- is on the boundary of the parameter space (Wilson, 2015). When comparing a model with its zero-inflated version using this test, the rejection of the null hypothesis indicates that models do not perform equally, but not necessarily that data is zero-inflated.

3. How to analyse count data

In order to choose an adequate model when analysing count data that is suspicious to be zero inflated, we recommend the following protocol:

- **Step 1**: Detect and classify the zeros. Minimize the presence of false zeros in the dataset, i.e. those zeros that arise due to design errors and observer errors during data collection (Fig. 1). Remove them from the dataset if identified. Consider the context of the biological system to forecast whether structural zeros can be present, i.e. true zeros related to ecological or evolutionary restrictions. Notice that the difference between a false zero obtained from a design error and a structural zero depends on whether the hypothesis being tested includes the possibility of obtaining a zero. A good knowledge of the biological system under study, a well-defined hypothesis and an extensive sampling are essential to avoid the presence of false zeros and detect the presence of structural zeros.
- **Step 2**: Identify suitable covariates to model count data, including zeros. The selection of covariates will depend on the particularities of the experimental design.
- **Step 3**: Test the data for overdispersion and zero inflation, for example by calculating the

dispersion index and the zero inflation index, either for the overall data or separately for each of the covariates. A dispersion index larger than 1 indicates that data is overdispersed respect to a Poisson distribution. A zero inflation index (zi_P or zi_{NB}) larger than 0 indicates that data is zero-inflated respect to a Poisson or a NB distribution, respectively.

- **Step 4**: Based on the overdispersion and zero-inflation results from step 3 choose the most adequate model following the suggestions of Table 1. Use the score tests to determine whether more complex models should be implemented.

4. Example on plant-herbivore interactions: testing the Enemy Release Hypothesis

The Enemy Release Hypothesis (Keane & Crawley, 2002) postulates that exotic plants will experience a decrease in their levels of herbivory compared with the native species in the novel range, by leaving behind the enemies from their area of origin. This hypothesis predicts a large number of undamaged plants (absence of herbivory) in the invaded habitat, and, consequently, data might be zero-inflated and probably overdispersed. We analysed a dataset obtained from an observational field study comparing the levels of herbivory in two native plant species, *Senecio vulgaris* and *S. lividus*, and two exotic plant species, *S. inaequidens* and *S. pterophorus*. The Enemy Release Hypothesis would be supported if the two native species had higher levels of herbivory than the two exotic species.

The survey was conducted in Montseny Natural Park (60 Km NE of Barcelona, 2°16'E 41°42' N). We delimited six locations ("Vallfornés", "Can Bosc", "Can Perepoc", "Can Tarrer", "Fogueres" and "Santa Susanna") of a diameter of 600m that contained the four *Senecio* species. In each location, we labelled 4 to 32 individuals per species, depending on their availability, with a total of 475 plants (129 *S. vulgaris*, 164 *S. lividus*, 100 *S. Inaequidens* and 82 *S. pterophorus*). Individuals were labelled before blooming and surveyed every 10-15 days throughout their entire reproductive season, starting in April for the earliest flowering plants (*S. vulgaris*) and ending in December for plants with the longest reproductive period (*S. inaequidens*), with a total of 22 to 26 visits. During each visit, we counted the number of fructified flower heads, collected them, and evaluated any signs of herbivory damage or the presence of insects feeding within. Herbivory damage (*Y*) was defined as the number of damaged heads in an individual plant. We also obtained the total number of heads produced by each individual plant during the entire reproductive period. Further details on the sampling design, plant phenology and insect community can be found in Castells et al. (2014).

We followed the suggested steps to analyse the count data obtained in our study:

Step 1: *Identify false and structural zeros in the* Senecio *dataset*

Due to the nature of the hypothesis tested (i.e. exotic plants are released from enemies in the introduced range) we anticipated that zeros would be highly frequent in our dataset, especially for the two exotic species *Senecio inaequidens* and *S. pterophorus*. Preliminary surveys showed the presence of a tephritid fly *Sphenella marginata* feeding on flower heads of the natives *S. vulgaris* and *S. lividus*, with an approximate development time of the larval stage of two-three weeks. We were also aware that plant phenology varied among *Senecio* species although the exact timing was unknown. In the light of this information, we designed a survey to minimize the presence of false zeros (see recommendations in Fig. 1). We performed a frequent year-round survey at different locations sampling well-developed flower heads to maximize the presence of large larvae or pupae, and herbivory was recorded as the presence of an insect or signs of herbivory damage. Because we aimed to determine whether exotic plants were consumed by local herbivores in comparison with the natives, the absence of herbivory was considered a structural zero.

Step 2: *Select the covariates*

Based on the experimental design we selected two covariates: Species and Location. Let Y_{ijk} be the r.v. that describes the number of damaged flower heads in plant i, $i = 1,..,n_{jk}$, of species j, $j =1,..,4$, from location k, $k = 1,..,6$, where n_{jk} is the number of individual plants of species j in location k. Let $heads_{ijk}$ be the total number of flower heads per plant i of species j in location k. Therefore, we modelled the response r.v. Y_{ijk} as a function of these two categorical covariates. The variable $heads_{ijk}$ was added to the model as an offset variable, as it varied from plant to plant and it was related to herbivory (Castells et al., 2017).

Step 3: *Determining whether data is overdispersed and zero-inflated*

The exotic species *S. inaequidens* and *S. pterophorus* produced a much larger number of heads than the native *S. vulgaris* and *S. lividus* (Table 2). The relatively high variance compared to the mean suggested that the data was overdispersed. Indeed, the global dispersion index was d = 77.66 and the dispersion index for each species was $d_{S.vulgaris}$ = 54.16, $d_{S.lividus}$ = 62.39, $d_{S.inaequidens}$ = 78 and $d_{S.pterophorus}$ = 138.78. Thus, data from all the species was overdispersed with respect to a Poisson. We grouped the plants into 24 categories defined by the four species and the six locations. The relationship between the index of dispersion for each category (d_k) and its mean was not constant (Fig. 3), which suggested that data would better fit a type 2 NB distribution.

The number of undamaged plants (plants with zero damaged flower heads) was 261 from a total of 475, which corresponds to 54.95% of the observations. When modelling the data with a Poisson distribution without covariates, and excluding the counts larger than 30, we obtained a Poisson parameter estimate of $\hat{\lambda} = 3.007$. Thus, the corresponding probability of zero was $p_0 = e^{-3} \cong 0.05$, i.e. the probability of undamaged plants was expected to be 5%. The high proportion of zeros compared to the predicted value is a strong indication of zero inflation. Zeros were highly represented in all species, particularly in *S. pterophorus* and *S. vulgaris* where the proportion of undamaged plants were 84% and 83% respectively (Fig. 2, Table 2).

The zero inflation index for all the *Senecio* species together was zi_P = 0.921, and the 95% confidence interval (*CI*) by bootstrap techniques using 10,000 samples with replacement of the original sample was $CI_{95\%}(zi_P)$ = (0.895, 0.939). Therefore, our data was zero-inflated respect to a Poisson distribution because the *CI* did not contain the zero value. The new zero-inflated index was $zi_{\mathrm{NB}} = -0.0069$ and $CI_{95\%}(zi_{\mathrm{NB}}) = (-0.0297, 0.0183)$, and in this case *CI* included the zero value meaning that data was not zero-inflated respect to a NB distribution. When we obtained these indexes by species and localities, we observed that the zero inflation depended on both covariates (supplementary Table S2).

Step 4: *Modelling the* Senecio *dataset*

As shown in step 3, our data was overdispersed and zero-inflated. According to the suggestions on Table 1 the best model would be a ZINB. We aimed to confirm this, as well explore the consequences of fitting suboptimal models.

We considered four probability distributions for Y_{ijk},: Poisson, type 2 NB, ZIP and ZINB. The parameter λ_{ijk} denotes the expected number of damaged heads of plant *i* of species *j* and location *k* and, it is modelled by logarithmic link function (eqn 4),

$$\lambda_{ijk} = exp\big(\beta_0 + \beta_{1j}Species_j + \beta_{2k}Location_k + \log{(heads_{ijk})}\big), \qquad \text{eqn 6}$$

for $i = 1,..,n_{jk},\ j = 1..4,\ k = 1,..,6$, where β_{1j} measures the change from the reference species (*S. vulgaris*) to *j*th species and β_{2k} measures the change from the reference location ("Vallfornés") to the *k*th location. We did not expect an interaction between species and location and thus we only included the main effects.

The weight parameter ω was modelled by a logistic regression (eqn 5). We contemplated two possibilities. First, we assumed that ω was constant across species and locations:

$$(I) \qquad \omega = \frac{exp(\gamma_0)}{1+exp(\gamma_0)}. \qquad \text{eqn 7}$$

Second, we assumed that ω was constant across locations but differed among species:

$$(II)\ \ \omega_j = \frac{exp\big(\gamma_0+\gamma_{1j}Species_j\big)}{1+exp\big(\gamma_0+\gamma_{1j}Species_j\big)}, \qquad \text{eqn 8}$$

where γ_{1j} measures the change from the reference species (*S. vulgaris*) to *j*th species in the weight parameter. Therefore, we analysed six different models: Poisson; NB; ZIP with a weight defined as in eqn 7, i.e. ZIP(I); ZIP with a weight defined as in eqn 8, i.e. ZIP(II); ZINB with a weight model defined as in eqn 7, i.e. ZINB(I); and 6) ZINB with a weight model defined as in eqn 8, i.e. ZINB(II). We used SAS 9.4 to fit zero-inflated models using the NLMIXED procedure.

All statistical models predicted that *S. lividus* was the species with the highest number of damaged heads, as shown by its positive parameter together with the negative or non-significant parameters in the other species (Table 3). However, models differed in predicting the least damaged species, which corresponded to *S. pterophorus* in the non-zero-inflated models (Poisson and NB) and to *S. inaequidens* in all zero-inflated models. ZIP models predicted less damage in *S. pterophorus* than in *S. vulgaris*, while no differences between these two species were found by the ZINB models (Table 3). The estimation of the dispersion parameter, $\hat{\alpha}$, in the NB models was significantly different from zero, which indicates that data was overdispersed and that the Poisson and ZIP models should be rejected in favour of the NB and ZINB models. Additionally, the dispersion parameter in the non-zero-inflated NB model was higher compared with both ZINB models (Table 3), which suggests that the NB models incorporated some of the overdispersion related to the excess of zeros.

The zero inflation component, $\widehat{\omega}$, did not change significantly between the ZIP(I) and ZINB(I) models. Therefore, both models predicted a similar probability of extra zeros, 47% and 40% respectively, which is similar to the expected probability of 50% obtained from eqn 7 when $\hat{\gamma}_0$ is close to 0, that is $\widehat{\omega} \cong 1/2$. For the ZIP(II) and ZINB(II) models, we observed major differences in the zero inflation component across species: the mixture parameter of *S. vulgaris* ($\widehat{\omega}_{S.\ vulgaris}$ = 77%) did not differ from *S. pterophorus* ($\widehat{\omega}_{\mathrm{S.\ pterophorus}}$ = 84%) but it was different from *S. lividus* ($\widehat{\omega}_{S.\ lividus}$ = 29%) and *S. inaequidens* ($\widehat{\omega}_{\mathrm{S.\ inaequidens}}$ = 11%) (Table 3). The mixture parameter in the ZINB(II) model was smaller than in ZIP(II) for all species, to the extent that zero inflation became not significant in *S. inaequidens* (Table 3).

The score test for comparing a Poisson versus a NB model given by Dean and Lawless was T_1 = 186.733 and p < 0.001. Therefore, the data was clearly overdispersed and a NB model was preferred to a Poisson. Given the value of the score test by Ridout the ZIP model was rejected in favour of the ZINB model: T = 6929.276 and p <0.001 for comparing ZIP(I) versus ZINB(I) and T = 6238.159 and p < 0.001 for comparing ZIP(II) versus ZINB(II). The comparison between Poisson and ZIP regression (Van der Broek score test) was not needed because both models were already rejected.

Based on AIC and BIC the best model was ZINB(II) (Table 4). Models with a zero-inflated distribution ranked above their non-zero-inflated counterparts, and models using an NB distribution ranked above those using a Poisson distribution (Table 4). The Poisson model provided the worst fit. The Vuong and Clarke tests rejected the Poisson and NB models in favour of their zero-inflated versions, and the LR test rejected the Poisson and ZIP models in favour of the NB and ZINB models, respectively (Table 5). When comparing the models in terms of predicted percentage of zeros, the NB and ZI versions performed the best (Fig. S1).

4. Discussion

Our data (i.e. number of damaged heads) was overdispersed and zero-inflated when considering all the *Senecio* species together. This result is consistent with the expected properties of count data (Martin et al., 2005; Sileshi, Hailu & Nyadzi, 2009; Smith, Anderson & Millar, 2012). After adjusting the standard models (Poisson and NB) and their zero-inflated (ZI) versions we rejected the Poisson and ZIP distributions because the dispersion parameter of the NB and ZINB distributions was significantly different from zero (Table 3).We found that the overdispersion parameter dropped significantly in ZINB models compared with the NB model. This result indicates that most of the overdispersion came from the excess zeros, and consequently ZI models were more adequate than non-ZI models. Finally, we found that zero inflation differed across species (Fig. 3), and thus the second specification of ZI models (II), which considered the Species as a covariate, was more adequate than the first specification (I). In conclusion, the best model was ZINB(II), as confirmed by the goodness of fit tests (Table 4 and Table 5, Fig. S1). These results are coherent with the recommended models according to the presence of overdispersion and zero inflation presented in Table 1.

One advantage of using zero-inflated models, besides the more accurate modelling of data, is the distinction between random and structural zeros, which helps to interpret the results from a biological perspective. Random zeros emerge from the sampling variability and have no ecological explanation, while structural zeros are related to ecological events (Fig. 1). We applied Bayes' theorem to calculate the probability of an observed zero not being random. This probability was 0.99 for *S. pterophorus* (i.e. nearly all zeros in this species were structural), 0.84 for *S. vulgaris* and 0.6 for

S. lividus. The probability of non-random zeros was close to zero for *S. inaequidens* (i.e. all zeros were random), which is consistent with the absence of zero-inflation in *S. inaequidens* under the ZINB(II) model. The predominance of structural zeros in the exotic *S. pterophorus* may be explained by its short reproductive stage and its relatively short time since invasion, which may have reduced the probability of plant-herbivore encounters and the adaptation of the local fauna to the new host (Castells et al., 2013; Castells et al., 2014). The interactions between herbivores and the native *S. vulgaris* were constricted by the absence of coincident phenology with herbivores; most of the *S. vulgaris* individuals are already senescent when herbivores appear in spring (Castells et al., 2014). In contrast, the absence of herbivory on the exotic *S. inaequidens* was mostly due to random events. *S. inaequidens* has a longer plant invasion history compared with *S. pterophorus* (100 years and 40 years, respectively) and an extended plant phenology compared with the other *Senecio* species, even during periods when no native *Senecio* species are available. Both factors could have favoured the associations between *S. inaequidens* and local herbivores and the low number of zero values (Castells et al., 2013; Castells et al., 2014).

Failure to model zeros correctly, i.e. ignoring zero inflation and overdispersion, can lead to serious errors in the interpretation of the results from an ecological point of view. In our example, all zero-inflated models predicted that *S. inaequidens* was the least damaged species, but according to the non-zero-inflated models it was *S. pterophorus*. Similar problems occurred when overdispersion was ignored. Under a Poisson distribution (Poisson and ZIP models) herbivory in the two native species was higher than in the two exotic species, which supports the Enemy Release Hypothesis. In contrast, the NB distribution predicted no differences in herbivory between *S. vulgaris* and the exotic *S. inaequidens* (NB and ZINB(I) models) and between the native *S. vulgaris* and the exotic *S. pterophorus* (ZINB(I) and ZINB(II) models) (Table 3). Because the best model for our data was ZINB, we rejected the Enemy Release Hypothesis. This is a clear example on how choosing the wrong statistical model may affect the outcome of a hypothesis.

Conclusions

Count data can be overdispersed and zero-inflated. When these traits occur, the standard statistical models are no longer valid. Here we have described a protocol to assist with the selection of an adequate model. Briefly, we propose to identify the type of zeros present in the dataset in relation to the biological system, the experimental design and the hypotheses formulated, and to select a model based on the presence of overdispersion and zero inflation. We finally want emphasize that the ecological context of the study must always be considered when selecting a statistical model. An understanding of how zeros arise in count data, for example identifying the potential sources of structural zeros that may cause zero-inflation, is essential in order to choose the most adequate statistical design, which will be a model that not only fits the data correctly, but that also takes into account the idiosyncrasies of the biological system under study.

Acknowledgements

This research was conducted thanks to the financial support provided to E.C. (grants GCL2008-02421/BOS, GCL2011-29205 and 2017SGR-1001), to P.P. (grant MTM2015-69493-R) and to M.P. (grants MTM2013-43992-R and MTM2017-89244-12) by the Spanish Ministry of Economy and Competitiveness and to M.P. via recognition 2017SGR-856 from AGAUR (Generalitat de Catalunya). The authors thank S. Baena-Mirabete (Departament de Matemàtiques, Univ. Autònoma de Barcelona), Jordi Martínez-Vilalta (CREAF) i Lucía Galiano (Univ. Autònoma de Barcelona) for additional suggestions and helpful comments.

Data accessibility

Dataset and SAS codes (dispersion and ZI indices, ZI mixture models and score tests) deposited in the repository of Universitat Autònoma de Barcelona: https://ddd.uab.cat/record/194390

Authors' contributions

A.B., E.C., M.P. and P.P. conceived the ideas and designed the methodology; M.M. and E.C. collected the data; A.B. analysed the data; A.B. and E.C. led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

References

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

Bliss, C., & Fisher, R. (1953). Fitting the negative binomial distribution to biological data. Biometrics 9, 176–200.

Castells, E., Morante, M., Blanco-Moreno, J.M., Sans, F.J., Vilatersana, R., & Blasco-Moreno, A. (2013). Reduced seed predation after invasion supports enemy release in a broad biogeographical survey. Oecologia 173(4), 1397–1409.

Castells, E., Morante, M., Goula, M., Pérez, N., Dantard, J., & Escolà, A. (2014). Herbivores on native and exotic *Senecio* plants: is host switching related to plant novelty and insect diet breadth under field conditions? Insect Conservation and Diversity 7, 420–431.

Castells, E., Morante, M., Saura-Mas, S., & Blasco-Moreno, A. (2017). Plant-herbivore assemblages under natural conditions are driven by plant size, not chemical defenses. Journal of Plant Ecology 10, 1012–1021.

Clarke, K. (2007). A simple distribution-free test for non nested model selection. Political Analysis 15 (3), 347–363.

Dean, C., & Lawless, J. (1989). Test for detecting overdispersion in poisson regression models. Journal of the American Statistical Association 84, 467–472.

Fisher, R. (1950). The significance of deviations from expectation in a poisson series. Biometrics 6 (1), 17–24.

Greene, W.H. (1994). Accounting for excess zeros and sample selection in poisson and negative binomial regression models. New York University, Department of Economics Working Paper No. EC-94-10, http://ssrn.com/abstract=1293115.

He, H., Tang, W., Wang, W., & Crits-Christoph, P. (2014). Structural zeroes and zero-inflated models. Shanghai archives of psychiatry 26(4), 236–242.

Heilbron, D. (1994). Zero-altered and other regression models for count data with added zeros. Biometrical Journal 36, 531–547.

Hilbe, J. (2011). Negative Binomial Regression. Cambridge.

Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). Univariate discrete distributions. Hoboken, N.J: Wiley.

Keane, R., & Crawley, M. (2002). Exotic plant invasions and the enemy release hypothesis. Trends in Ecology and Evolution 17, 164–170.

Kuhnert, P., Martin, T., Mengersen, K., & Possingham, H. (2004). Assessing the impacts of grazing levels on bird density in woodland habitat: a Bayesian approach using expert opinion. Environmetrics 16, 1–31.

Lambert, D. (1992). Zero-inflated poisson regression with an application to defects in manufacturing. Technometrics 34,1–14.

MacKenzie, D., Nichols, J., Lachman, G., Droege, S., Royle, J., & Langtimm, C. (2002). Estimating site occupancy rates when detection probabilities are less than one. Ecology 83, 2248–2255.

Martin, T., Wintle, B., Rhodes, J., Kuhnert, P., Field, S., Low-Choy, S., Tyre, A., & Possingham, H. (2005). Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecology Letters 8, 1235–1246.

McCullagh, P., & Nelder, J. (1989). Generalised Linear Models. Chapman and Hall, 2nd ed.

Mullahy, J. (1986). Specification and testing of some modified count data models. Journal of Econometrics 33, 341–365.

O'Hara, R., & Kotze, D. (2010). Do not log-transform count data. Methods in Ecology & Evolution 1, 118–122.

Oliveira, M., Einbeck, J., Higueras, M., Ainsbury, E., Puig, P. & Rothkamm, K. (2016), Zero-inflated regression models for radiation-induced chromosome aberration data: A comparative study. Biometrical Journal 58, 259–279. doi:10.1002/bimj.201400233

Puig, P., & Valero, J. (2006). Count data distributions: some characterizations with applications. Journal of the American Statistical Association 101 (473), 332–340.

Ridout, M., Hinde, J., & Demetrio, C. (2001). A score test for testing a zero-inflated poisson regression model against zero-inflated negative binomial alternatives. Biometrics 57, 219–223.

Sadykova, D., Scott, B. E., De Dominicis, M., Wakelin, S. L., Sadykov, A., & Wolf, J. (2017). Bayesian joint models with INLA exploring marine mobile predator–prey and competitor species habitat overlap. Ecology and Evolution 7, 5212–5226.

Schwarz, G. E. (1978). Estimating the dimension of a model. Annals of Statistics 6 (2), 461–464.

Sileshi, G., Hailu, G., & Nyadzi, G. (2009). Traditional occupancy abundance models are inadequate for zero-inflated ecological count data. Ecological Modelling 220 (15), 1764 –1775.

Smith, A. N. H., Anderson, M. J., & Millar, R. B. (2012) Incorporating the intraspecific occupancy-abundance relationship into zero-inflated models. Ecology 93 (12), 2526–2532.

Tang, W., He, H., Wang, W.J., & Chen, D.G. (2018) Untangle the structural and random zeros in statistical modelings, Journal of Applied Statistics, 45:9, 1714–1733, doi: 10.1080/02664763.2017.1391180

van der Broek, J. (1995). A score test for zero inflation in a poisson distribution. Biometrics 51, 738–743.

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57, 307–334.

Williams, P. J., Hooten, M. B., Womble, J. N., Esslinger, G. G., Bower, M. R., & Hefley, T. J. (2017). An integrated data model to estimate spatiotemporal occupancy, abundance, and colonization dynamics. Ecology 98, 328–336.

Wilson, P. (2015). The misuse of the Vuong test for non-nested models to test for zero-inflation. Economics Letters 127, 51-53.

Zuur, A. F., Ieno, E. I., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). Mixed Effects Models and Extensions in Ecology with R. Springer.

Table 1. Source of zeros in count data and recommended modelling approaches according to the presence of overdispersion and zero inflation, as proposed by literature. A Poisson distribution is adequate in absence of overdispersion and zero inflation. The NB distribution can deal with overdispersion and a moderate excess of random zeros. Zero-inflated models (Zero-Inflated Poisson, ZIP; Zero-Inflated Negative Binomial, ZINB) are adequate when the counting process can generate zero values and the existence of structural zeros is coherent with the biological system under study. Zero-Altered models (Zero-Altered Poisson, ZAP; Zero-Altered Negative Binomial, ZANB) are only adequate when the counting process cannot generate zero values, as these models do not discriminate between the different types of zeros. See more details in the text.

Type of zeros	**Source**	**Generator process**	**Over-dispersion**	**Zero inflation**	**Modelling approach**
False zeros	Design errors	Poor experimental design	-	-	Remove before analysis
	Observer errors	Lack of experience	-	-	Remove before analysis
True zeros	Random	Sampling variability	No	No	Poisson
			Yes	No	NB[1]
	Structural	Outside the count process	No	Yes	ZIP[2] or ZAP[3]
			Yes	Yes	ZINB[2] or ZANB[3]

[1] NB, negative binomial (McCullagh & Nelder, 1989).

[2] ZIP, Zero-Inflated Poisson (Lambert, 1992) and ZINB, Zero-Inflated Negative Binomial (Greene, 1994).

[3] ZAP, Zero-Altered Poisson and ZANB, Zero-Altered Negative Binomial (Mullahy, 1986).

Table 2. Percentage of plants with no damaged heads (*Y* = 0), total head production per plant, and number and percentage of damaged heads per plant when including all individuals (plants with *Y* ≥ 0) and only including the damaged individuals (plants with *Y* > 0) for four *Senecio* species across all locations (Mean (SD)).

Species	**% of undamaged plants**	**Number of heads per plant**	**Number of damaged heads per plant (*Y*)**		**Rate of damaged heads per plant (%)**	
			All plants	Damaged plants	All plants	Damaged plants
S. vulgaris	83%	29.8 (49.8)	2.92 (12.58)	17.14 (26.64)	2.51 (7.48)	14.73 (12.33)
S. lividus	33%	28.0 (39.9)	9.20 (23.96)	13.84 (28.31)	23.20 (24.46)	34.90 (22.15)
S. inaequidens	30%	262.4 (675.8)	12.91 (31.73)	18.44 (36.62)	6.67 (11.45)	9.53 (12.67)
S. pterophorus	84%	1069.4 (1897.9)	5.37 (27.29)	33.85 (63.11)	1.94 (8.00)	12.26 (17.17)

Table 3. Estimated parameters and their standard errors for herbivory damage on two native species (*Senecio vulgaris* and *S. lividus*) and two exotic species (*S. inaequidens* and *S. pterophorus*) sampled at six locations, using Poisson, NB, ZIP and ZINB models. ZIP and ZINB were modelled under two different assumptions: (I) all species had the same level of zero inflation, and (II) species had different levels of zero inflation.

	Poisson	**NB**	**ZIP(I)**	**ZIP(II)**	**ZINB(I)**	**ZINB(II)**
COUNT COMPONENT						
(Intercept)	**-2.03 (0.05)**	**-2.88 (0.24)**	**-1.50 (0.05)**	**-1.46 (0.05)**	**-2.05 (0.18)**	**-1.74 (0.21)**
Species (*S. vulgaris*)						
S. lividus	**1.29 (0.06)**	**1.94 (0.25)**	**0.86 (0.06)**	**0.82 (0.06)**	**1.46 (0.19)**	**1.08 (0.22)**
S. inaequidens	**-0.86 (0.06)**	0.46 (0.29)	**-1.40 (0.06)**	**-1.45 (0.06)**	-0.23 (0.21)	**-0.67 (0.24)**
S. pterophorus	**-2.95 (0.07)**	**-0.67 (0.32)**	**-0.67 (0.07)**	**-0.70 (0.07)**	0.26 (0.29)	0.10 (0.33)
Location (Vallfornés)						
Can Bosc	**-0.59 (0.06)**	-0.42 (0.31)	**-0.18 (0.06)**	**-0.17 (0.06)**	**-0.46 (0.19)**	**-0.40 (0.20)**
Can Perepoc	**-0.93 (0.08)**	**-0.75 (0.32)**	**-0.62 (0.08)**	**-0.64 (0.08)**	**-0.79 (0.20)**	**-0.87 (0.21)**
Can Tarrer	**-0.43 (0.06)**	**-0.65 (0.30)**	**-0.45 (0.07)**	**-0.44 (0.07)**	**-0.73 (0.20)**	**-0.69 (0.20)**
Fogueres	**0.24 (0.05)**	-0.17 (0.32)	**0.40 (0.05)**	**0.40 (0.05)**	-0.34 (0.20)	-0.34 (0.20)
Santa Susanna	**-1.09 (0.06)**	**-1.09 (0.30)**	**-1.03 (0.07)**	**-1.03 (0.07)**	**-1.09 (0.19)**	**-1.08 (0.20)**
Dispersion parameter ($\hat{\alpha}$)		**3.12 (0.29)**			**0.60 (0.09)**	**0.64 (0.11)**
ZERO INFLATION COMPONENT						
(Intercept)			-0.10 (0.10)	**1.23 (0.25)**	**-0.40 (0.13)**	**0.81 (0.30)**
Species (*S. vulgaris*)						
S. lividus				**-2.11 (0.31)**		**-2.24 (0.41)**
S. inaequidens				**-3.36 (0.50)**		**-3.90 (1.18)**
S. pterophorus				0.41 (0.39)		0.80 (0.43)
Mixture parameter						
$\hat{\omega}$			**0.47 (0.03)**		**0.40 (0.03)**	
$\hat{\omega}_{S.vulgaris}$				**0.77 (0.04)**		**0.69 (0.06)**
$\hat{\omega}_{S.lividus}$				**0.29 (0.04)**		**0.19 (0.05)**
$\hat{\omega}_{S.inaequidens}$				**0.11 (0.04)**		0.04 (0.05)
$\hat{\omega}_{S.pterophorus}$				**0.84 (0.04)**		**0.83 (0.04)**

NB: negative binomial
ZIP: Zero-Inflated Poisson
ZINB: Zero-Inflated negative binomial
The reference levels for Species and Location are indicated in parenthesis.
Statistically significant parameters at $P < 0.05$ are shown in bold.

Table 4. Goodness of fit statistics (log *L*, AIC and BIC) for the Poisson, NB, ZIP and ZINB models.

	Poisson	**NB**	**ZIP(I)**	**ZIP(II)**	**ZINB(I)**	**ZINB(II)**
# parameters	9	10	10	13	11	14
log *L*	-3154.6	-978.0	-1587.1	-1517.0	-941.7	-877.1
AIC	6327.1	1976.0	3194.1	3060.1	1905.4	1782.2
BIC	6364.6	2017.6	3235.8	3114.2	1951.2	1840.5

NB: negative binomial

ZIP: Zero-Inflated Poisson

ZINB: Zero-Inflated negative binomial

log L: log-likelihood

AIC: Akaike Information Criterion

BIC: Bayesian Information Criterion

Table 5. Model comparison: LR and Vuong/Clarke tests (without correction). Statistical significance is shown in parenthesis.

Model comparison	Test performed	Preferred Model
NB *vs* Poisson	4353.1 (P<0.001) [1]	NB
ZIP(II) *vs* Poisson	2.407 (P=0.016) / 71.5 (P<0.001) [2]	ZIP(II)
ZINB(II) *vs* NB	7.049 (P<0.001) / 114.5 (P<0.001) [2]	ZINB(II)
ZINB(II) *vs* ZIP(II)	1279.8 (P<0.001) [1]	ZINB(II)

[1] Likelihood Ratio test (LR)

[2] Vuong/Clarke tests

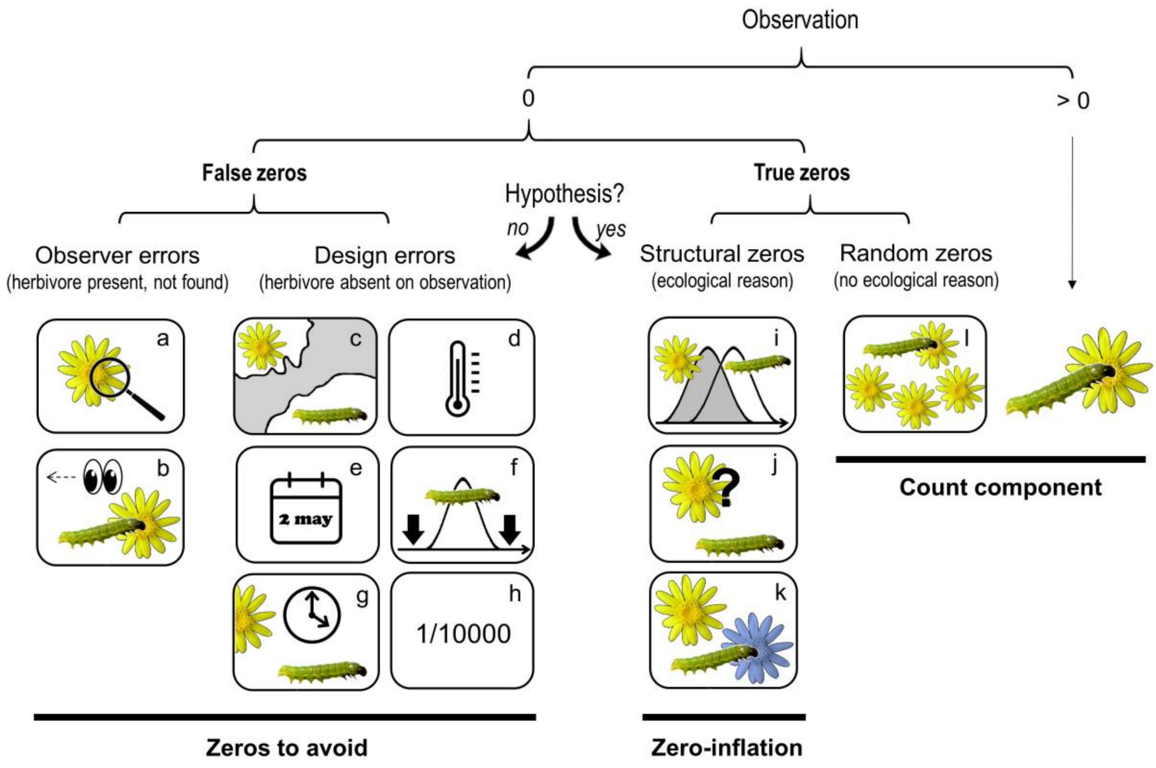

Figure 1. Different sources of zeros that could emerge in count data. The example shows the presence (>0) or absence (0) of herbivores on a plant species. Zeros due to the lack of experience of the observer (a-b) or resulting from a poor experimental design (c-h) are called *False Zeros* and should be minimized when performing the experiment. Structural Zeros, that is, zeros related to the ecological system under study (i-k), and Random Zeros emerging from the sampling variability (l) are known as *True Zeros*. Classifying a zero as a design error or structural zero depends on whether the event is part of the hypotheses tested. Only when the study includes the possibility of a zero value as part of the hypotheses (e.g. the study aims to test whether the interaction is occurring) the resulting zeros would be structural and should be included in the statistical analysis. The following text explains different scenarios that would result in a zero value, and, in brackets, how errors due to false zeros can be minimized: (*a*) the insects or the damage exerted are so small that the observer cannot detect them [sample when the insects are expected to be well developed]; (*b*) the observer does not see the herbivore (e.g. it is mistaken for a seed) or the damage is associated to other causes not related to herbivory (e.g. mechanical damage during sampling, pathogens, etc.) [the observer should be trained properly]; (*c)* the distributional areas of herbivores and plants are not coincident [know the species

distribution before sampling]; (*d*) a herbivore is not present in a certain location within its distributional area, for example due to the microclimatic conditions [sample in habitats with adequate environmental conditions for a herbivore, or perform replicate surveys in different areas]; (*e*) a single survey is conducted, and is not coincident with the herbivore phenology [know the herbivore life cycle or perform long-term surveys]; (*f*) a long-term survey is conducted, but the low sampling frequency does not enable capture of the presence of the herbivore [sample on a more frequent basis]; (*g*) herbivores are not found because they are absent at the time of sampling [record plant damage instead of the presence of insects]; (*h*) herbivores are so infrequent that the design cannot capture their presence [perform extensive sampling with a high number of replicates]; (i) phenology of plants and herbivores are not completely coincident at a temporal level; (*j*) herbivores do not recognize a plant as a potential host; (*k*) herbivores recognize a plant as a host but prefer to feed on another species; and (*l*) the herbivore population is not large enough to saturate the available plant resources.

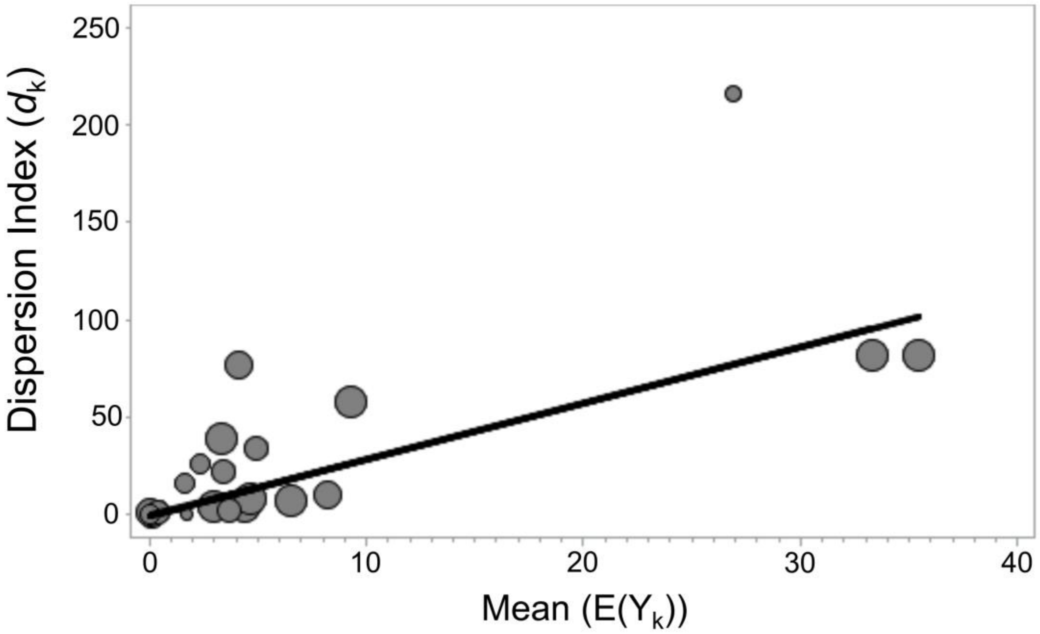

Figure 2. Relationship between the mean and the dispersion index for herbivory damage in *Senecio*. The dispersion index (d_k) is estimated as the sample variance divided by the sample mean within each group. The diameters of the circles are proportional to the number of averaged values for each group (k= 24). The line is the weighted-regression fit.

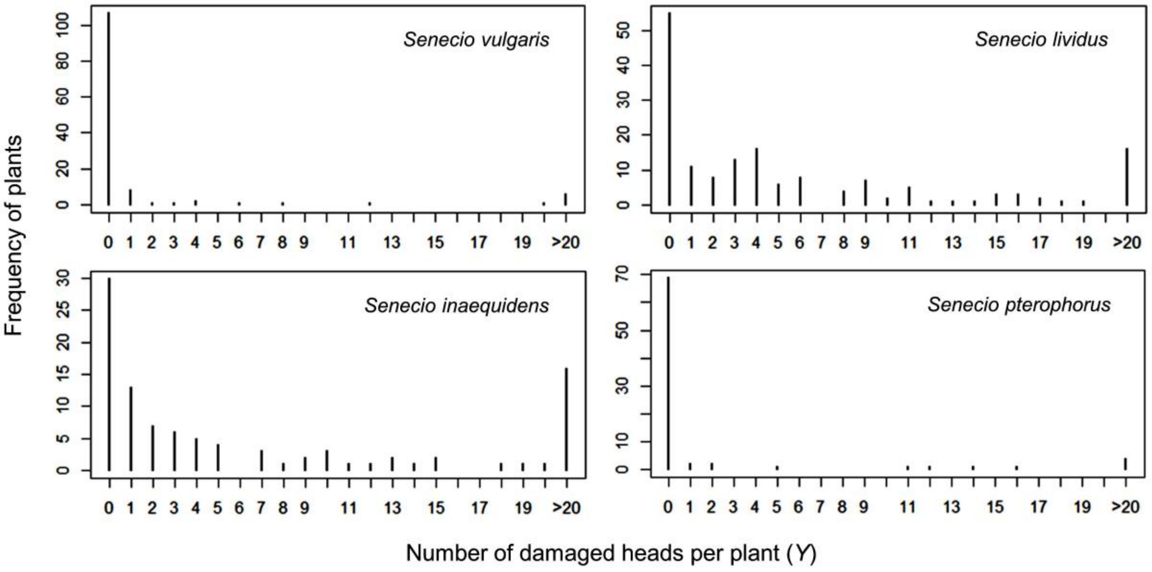

Figure 3. Frequency of plants for the response variable "Herbivory damage" defined as the number of damaged heads per plant in four *Senecio* species across six locations.