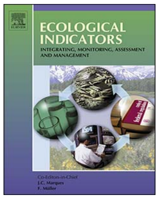
Ecological Indicators 85 (2018) 502–508

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind



Research paper

# A study on the effects of unbalanced data when fitting logistic regression models in ecology

Christian Salas-Eljatib[a,*], Andres Fuentes-Ramirez[a], Timothy G. Gregoire[b], Adison Altamirano[c], Valeska Yaitul[a]

[a] *Laboratorio de Biometría, Departamento de Ciencias Forestales, Universidad de La Frontera, Temuco, Chile*
[b] *School of Forestry and Environmental Studies, Yale University, New Haven, CT 065111, USA*
[c] *Laboratorio de Ecología del Paisaje Forestal, Departamento de Ciencias Forestales, Universidad de La Frontera, Temuco, Chile*

## ARTICLE INFO

*Keywords:*
Statistical inference
Model prediction
Logit model
Binary variable
Bias
Precision

## ABSTRACT

Binary variables have two possible outcomes: occurrence or non-occurrence of an event (usually with 1 and 0 values, respectively). Binary data are common in ecology, including studies of presence/absence, alive/dead, and change/no-change. Logistic regression analysis has been widely used to model binary response variables. Unbalanced data (i.e., an extremely larger proportion of zeros than ones) are often found across a variety of ecological datasets. Sometimes the data are balanced (i.e., same amount of zeros and ones) before fitting the model, however, the statistical implications of balancing (or not) the data remain unclear. We assessed the statistical effects of balancing data when fitting a logistic regression model by studying both its statistical properties of the estimated parameters and its predictive capabilities. We used a base forest-mortality model as reference, and by using stochastic simulations representing different configurations of 0/1 data in a sample (unbalanced data scenarios), we fitted the logistic regression model by maximum likelihood. For each scenario we computed the bias and variance of the estimated parameters and several prediction indexes. We found that the variability of the estimated parameters is affected, with the balanced-data scenario having the lowest variability, thus, affecting the statistical inference as well. Furthermore, the prediction capabilities of the model are altered by balancing the data, with the balanced-data scenario having the better sensitivity/specificity ratio. Balancing, or not, the data to be used for fitting a logistic regression models may affect the conclusion that can arise from the fitted model and its subsequent applications.

## 1. Introduction

Data of occurrence/non-occurrence of a phenomenon of interest are vastly found across several disciplines (Alberini, 1995; Arana and Leon, 2005; Bell et al., 1994). This type of variable is known as binary or dichotomous, and it represents whether an event occurs or not. This event is represented by the random variable $Y$, and we usually record occurrence by $Y = 1$ and non-occurrence by $Y = 0$. In ecology, binary variables arise when studying the presence of a species in a geographic area (Bastin and Thomas, 1999; Phillips and Elith, 2013; Hastie and Fithian, 2013) or the occurrence of mortality at the tree or forest level (Davies, 2001; Wunder et al., 2008; Chao et al., 2009; Young et al., 2017). Meanwhile in landscape ecology, binary variables are used to represent the occurrence of fire within a given area (Bigler et al., 2005; Mermoz et al., 2005; Dickson et al., 2006; Vega-García and Chuvieco, 2006; Palma et al., 2007; Bradstock et al., 2010); deforestation (Wilson et al., 2005; Schulz et al., 2011; Kumar et al., 2014; Hu et al., 2014); and in general the change from one land use category to another (Seto and Kaufmann, 2005; Leyk and Zimmermann, 2007; Lander et al., 2011).

Logistic regression analysis is the most frequently used modelling approach for analyzing binary response variables. If we need to model a binary variable, to statistically relate it to predictor variable(s) or covariate(s), one of the most used approaches for pursuing this task in ecology is to use logistic regression models (Warton and Hui, 2011). These models belong to group of the generalized linear models (GLM). In a GLM, three compartments must be specified (Lindsey, 1997; Schabenberger and Pierce, 2002): a random component, a systematic component, and a link function. A logistic regression model uses: a binomial probability density function as the random component; a linear predictor function $\mathbf{X}'\boldsymbol{\beta}$ (where $\mathbf{X}$ is a matrix with the covariates and $\boldsymbol{\beta}$ is a vector with the parameters or coefficients) as the systematic component; and a logistic equation as the link function. One of the key advantages of using logistic regression models in ecology is that the

[*] Corresponding author. Tel.: +56 45 2325652.
*E-mail address:* christia.salas@ufrontera.cl (C. Salas-Eljatib).

http://dx.doi.org/10.1016/j.ecolind.2017.10.030
Received 7 April 2017; Received in revised form 9 August 2017; Accepted 16 October 2017

probability of the binary response variable is directly modelled, thereby accounting explicitly for the random nature of the phenomenon of interest.

In many applications when dealing with binary data in ecology, it happens that the number of observations with ones ($Y = 1$) is much smaller than the number of observations with zeros ($Y = 0$) or vice versa. We simply term this situation as *unbalanced data*, but other terms have been also used for this situation, including disproportionate sampling (Maddala, 1992) or rarity events (King and Zeng, 2001). Based on our review of scientific applications of logistic regression to model ecological phenomena, the proportion of zeros in datasets ranges between 80% and 95%. Therefore, having balanced data (i.e., equal numbers of observations of zeros and ones) is more the exception than the rule.

Both unbalanced and balanced data have been used for fitting logistic regression models. In ecological studies, some researchers have adopted the practice of balancing the data before carrying out the analyses (e.g., Vega-García et al., 1995; Vega-García et al., 1999; Lloret et al., 2002; Brook and Bowman, 2006; Vega-García and Chuvieco, 2006; Jones et al., 2010; Rueda, 2010). Balancing data means to select, by some rule (usually at random), the same amount of observations with ones and zeros from the originally available dataset. Therefore, a balanced dataset or balanced sample is created, where a 50–50% proportion of zero and one values is met. After the balanced dataset is built, the logistic regression model is fitted (i.e., its parameters are estimated) by maximum likelihood (ML). An example of this practice in ecological applications is the option for balancing data before fitting a logit model when conducting analyses of land use changes in the software IDRISI (Eastman, 2006). On the other hand, it is important to point out that unbalanced data have been also used in ecological studies (Wilson et al., 2005; Echeverria et al., 2008; Kumar et al., 2014; Young et al., 2017). Therefore, unbalanced data in applied ecological studies has been considered as not having important effects into the models being fitted. Moreover, to date, no studies have addressed the effect of balancing data when fitting logistic regression models in ecological analyses, and just a handful have explored some statistical implications in ecological applications (Qi and Wu, 1996; Wu et al., 1997; Cailleret et al., 2016).

The applied statistical implications of unbalanced data in logistic regression are not well described nor realized for applied researchers. Although balancing the data seems to be an accepted practice, the reasons that justify its use are not well explained. The most immediate effect of balancing the data is to greatly reduce the sample size available for fitting purposes, therefore decreasing the precision with which the parameters of the model are estimated. Among the statistical studies on logistic regression and unbalanced data, we highlight the following: Schaefer (1983) and Scott and Wild (1986) pointed out that the maximum likelihood estimates (MLE) of a logit model are biased only for small sample sizes. On the other hand, Xie and Manski (1989) stated that unbalanced data only affect the intercept parameter of a logit model, specifically being biased estimated according to Maddala (1992). King and Zeng (2001), advocated that all the MLE of the logit parameters are biased. Schaefer (1983) and Firth (1993) proposed correction for the bias of the MLE of the logistic regression model parameters. McPherson et al. (2004) conducted one of the few related analysis when fitting presence-absence species distribution models in ecology, but only focusing in the prediction capabilities of the fitted models. Maggini et al. (2006) assessed the effect of weighting absences when modelling forest communities by generalized additive models. Recently, Komori et al. (2016) indicated that logistic regression suffer poor predictive performance, and proposed an alternative model to improve predictive performance. Komori et al. (2016) approach involves to add a new parameter to the original structure of a logistic regression model, and fitted it in a mixed-effects modelling framework, therefore their approaches becomes a different type of statistical model. From above, we can infer that: (a) most of the statistical studies on logistic regression and unbalanced data have focus on the bias of the MLE parameters (a topic that has been rarely taking into account in ecological applications); (b) much less attention has been put into the prediction performance; and (c) no study has dealt with the effects of unbalanced data in the variance of the MLE parameters.

In this study we aimed at assessing the effect of using unbalanced data when fitting logistic regression models by analyzing both the statistical properties (i.e., bias and variance) of the estimated parameters and the predictive capabilities of the fitted model.

## 2. Methods

### 2.1. Base model

The binary variable ($Y$) is the occurrence of a phenomenon of interest, where $Y = 1$ denotes occurrence and $Y = 0$ otherwise. In a modelling framework, we seek to model the probability of the response variable being $Y = 1$, given the values of the predictor variables, this is $\Pr(Y = 1|\mathbf{X})$, that we can more easily represent by $\pi_{\mathbf{y}|\mathbf{X}}$.

In our analysis we used a logistic regression equation with five predictor variables, as a base model for carrying out our analysis. This model served as a reference for assessing the statistical effects of unbalanced data on fitting logistic regression models. The binary variable of forest mortality occurrence ($Y$), given the analyses of Young et al. (2017) in the state of California, USA, is modeled as a function of climate and biotic variables, as follows:

$$\ln\left[\frac{\pi_{y_i|x_i}}{1-\pi_{y_i|x_i}}\right] = \text{logit}[Y_i = 1] = \beta_0 + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \beta_3 X_{3_i} + \beta_4 X_{4_i} + \beta_5 X_{5_i}, \tag{1}$$

where $Y_i$ is the occurrence of forest mortality (i.e., 1 for occurrence, 0 for non-occurrence) at the $i$th pixel), meanwhile the predictor variables $X_{1i}$, $X_{2i}$, $X_{3i}$, $X_{4i}$, and $X_{5i}$ represent the: mean climatic water deficit (CWD) or simply *Defnorm*$_i$; basal area of live trees (*BA*$_i$); BA$_i^2$; CWD anomaly (*Defz*0$_i$); and *Defnorm*$_i$ × *BA*$_i$ for the $i$th pixel, respectively. We have used the nomenclature for the variables as in the study of Young et al. (2017) and only the available data for year 2012. Notice that we could more easily represent model (1) as:

$$\ln\left[\frac{\pi_{\mathbf{y}|\mathbf{X}}}{1-\pi_{\mathbf{y}|\mathbf{X}}}\right] = \text{logit}[\mathbf{y}{=}1] = \mathbf{X}'\boldsymbol{\beta}, \tag{2}$$

where $\mathbf{y}$ is the vector with the binary variable, $\mathbf{X}$ is the matrix with the predictor variables (and a first column of 1), and $\boldsymbol{\beta}$ is the vector of parameters $[\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2}, \widehat{\beta_3}, \widehat{\beta_4}, \widehat{\beta_5}]$.

In the sequel, we shall use Eq. (2) as the mean function in various scenarios of unbalanced data. It is important to point out that we are not interested in finding the best model, but rather on studying the effects of using several unbalanced data scenarios on a reference model. Furthermore, we want to remark that we are not pursuing to assess different alternative statistical models for unbalanced data (e.g. as in, Warton and Hui, 2011; Hastie and Fithian, 2013). We also want to mention that the zero-inflated models are those focusing on modelling count variables (Schabenberger and Pierce, 2002; Zuur et al., 2010), such as the prediction of the amount of tree mortality (e.g., Affleck, 2006). These models are not part of our study, since we are dealing with modelling a binomial variable.

### 2.2. Unbalanced data scenarios

We use data of forest mortality occurrence from Young et al. (2017), in California during 2012 as our population, containing 11763 total observations ($N$), with 2985 cases of mortaltity occurrence ($N_1$) and 8778 cases of non-occurrence ($N_0$). In order to assess the effects of unbalanced data on the statistical properties of the logit model (Eq.

(2)), we examined different sample strategies from the population, where each has a different proportion of occurrence and non-occurrence of mortality (1 and 0 values, respectively). We fixed the sample size in $n = 1000$ in all scenarios, and the number of cases with zeros and ones for the response variable that the sample should contain, across scenarios ranging from 10% to 90%. In this way, we constrained the sample to containing different cases with zeros ($n_0$) and ones ($n_1$), but the same sample size ($n = 1000$). In order to achieve each of the proportion of 0/1 values, which has a fixed sample size of 0 and 1 (i.e., $n_0$ and $n_1$, respectively), we (i) drew a random sample without replacement of size ($n_0$) from the sub-population (with size $N_0$) of cases containing zero in the response variable; (ii) drew a random sample without replacement of size $n_1$ from the sub-population (with size $N_1$) of cases containing ones in the response variable; and (iii) merge the randomly selected $n_0$ and $n_1$ cases in a sample of size $n$ (*i. e*., $n = n_0 + n_1$).

### 2.3. Statistical assessment

We assessed the statistical properties of the fitted logistic regression model by stochastic simulations (i.e., Monte Carlo simulation). We carried out $S = 100,000$ simulations so that the sampling error of the simulation itself is negligibly small. A similar analysis to justify the number of simulation has been conducted by Gregoire and Schabenberger (1999), in agreement with the amount of simulations conducted in other statistical simulation studies (e.g. Gregoire and Salas, 2009; Salas and Gregoire, 2010). For each simulated sample, we fitted the logistic regression model (Eq. (2)) by maximum likelihood using the `glm` function implemented in R (R Development Core Team, 2016).

Based on the simulations, we examined the empirical distribution of the estimated parameters and prediction indexes. Our assessment was divided and focused in: (a) the statistical properties of estimated model parameters, and (b) the accuracy of predictions from the fitted model.

(a) *Statistical properties of the estimated parameters*. In order to assess how the accuracy of the estimated parameters is affected by unbalanced data, we computed the empirical bias ($B_{MC}$) of each parameter being estimated, $\widehat{\theta}$, as follows:

$$B_{MC}[\widehat{\theta}] = \theta - E[\widehat{\theta}], \tag{3}$$

where $\theta$ is the respective parameter value and $E[\widehat{\theta}]$ is the empirical expected value of the estimated parameter. The former was obtained from the maximum likelihood estimate (MLE) of $\theta$ using the population available, and the latter is approximated from the average of the $S$ values of the estimated parameter $\widehat{\theta}$. Notice that $\widehat{\theta}$ in Eq. (3) is replaced by each parameter of the model (i.e., $\widehat{\beta_0}$, $\widehat{\beta_1}$, $\widehat{\beta_2}$, $\widehat{\beta_3}$, $\widehat{\beta_3}$, and $\widehat{\beta_5}$).

In order to assess how the precision of the estimated parameters is affected by unbalanced data, we computed the empirical variance ($V_{MC}$) of each estimated parameter $\widehat{\theta}$ as follows:

$$V_{MC}[\hat{\theta}] = \frac{1}{S}\sum_{j=1}^{S}(\hat{\theta}_j - E[\hat{\theta}])^2, \tag{4}$$

where $\widehat{\theta}$ is the MLE of $\theta$ for the $j$th simulation. Finally, we compute the empirical mean square error ($ECM_{MC}$) of each $\widehat{\theta}$ by:

$$ECM_{MC}[\widehat{\theta}] = V_{MC}(\widehat{\theta}) + [B_{MC}(\widehat{\theta})]^2 \tag{5}$$

We represented the variance and mean square error in the same units of the estimated parameters by taking their square root, thus obtaining the standard error ($SE[\widehat{\theta}]$) and their root mean squared error ($RMSE[\widehat{\theta}]$).

(b) *Prediction capabilities*. For each simulation and unbalanced data scenario we computed prediction indexes of the logistic regression model. In order to do so, we calculated the predicted probability of mortality occurrence for the $i$th observation ($\widehat{\pi}_{y_i=1|\mathbf{X}_i}$), as follows:

$$\widehat{\pi}_{y_i=1|\mathbf{X}_i} = \frac{1}{1 + e^{-\mathbf{X}'_i\widehat{\beta}}}, \tag{6}$$

where $\mathbf{X}_i$ is the matrix of predictor variables for the $i$th case and $\widehat{\beta}$ is the vector of estimated parameters. We use a probability threshold of 0.5 for occurrence, that is to say, if $\widehat{\pi}_{y_i=1|\mathbf{X}_i} \geq 0.5$ we assume that the event occurs, and non-occurrence otherwise (Jones et al., 2010). Based on these predicted probabilities, we computed the following eight prediction indexes: commission error (proportion of $n$ cases in which the model erroneously predicts occurrence); commission accuracy (proportion of $n$ cases in which the model correctly predicts occurrence); omission error (proportion of $n$ cases in which the model erroneously predicts non-occurrence); omission accuracy (proportion of $n$ cases in which the model correctly predicts non-occurrence); sensitivity (proportion of the total cases of occurrence where the model correctly predicts occurrence); and specificity (proportion of the total cases of non-occurrence where the model correctly predicts non-occurrence).

We have also carried out all the above analyses (i.e., simulation and statistical properties assessment) for a different dataset. We used data of forest fire occurrence in central-Chile, as a way of representing how our findings could change in a forest fire model, and the main results are shown in Supplementary Material.

## 3. Results

The proportion of 0/1 in the data used for fitting a logistic regression model affects the distribution of the estimated parameters. The variability of the estimated parameters tends to increase with an extreme proportion of zero (or ones) in the data (Fig. 1).

Unbalanced data affects on the bias of the estimated parameters. All the parameters estimates were nearly unbiased for the proportion of zeros data assessed that is closer to the proportion of zeros in the entire population (First row panel of Fig. 1). However, for the other unbalanced data scenarios, all parameters are biasedly estimated (Fig. 1). The bias increases as the proportion of zeros in the data decreases both in nominal units (Fig. 1), as well as in percentage (Fig. 2a). The bias is larger for the estimated intercept-parameter than for the other parameters, regardless the unbalanced data scenario. The only exception to this trend is the estimate of the parameter $\beta_2$, being also heavily biased, which could be a result of its higher variability compared to the other parameter estimates (Fig. 2b). More importantly, the greatest precision of all estimated parameters occurs with balanced data (Fig. 2b), as well as the lowest root mean squared error (Fig. 2c). The reported greatest precision of the estimated parameter for the balanced-data scenario was even more pronounced for the forest fire model (Fig. 4). This can be a result of a stronger relationship among the response and the predictor variables, than we found in the forest mortality model. Besides, the forest fire model (Eq. (7) in Supplementary Material) has a lower number of parameters, therefore multicollinearity should be a minor problem than in a model with two more parameters (Eq. (1)). In fact, in the mortality model there are two parameters representing function of variables already present in the model (i.e., $BA_i^2$ and $Defnorm_i \times BA_i$), therefore the model is affected by multicollinearity.

The prediction capabilities of the logit model are greatly affected by the different proportions of zeros and ones. Both overall error (i.e., sum of omission and commission errors) and overall accuracy (i.e., sum of omission and commission accuracy) tend to be better, with a decreasing and increasing trend, respectively, when extreme proportions of zeros (or ones) are used for fitting the model (Table 1). Moreover, the larger is the proportion of zeros in the data, the better is the prediction of non-occurrence (i.e., higher values of omission accuracy). A similar trend, but not completely linear, is found when the omission errors are used as reference. On the contrary, the larger is the proportion of ones in the data, the better is the prediction of occurrence (i.e., higher values of commission accuracy). A similar trend is found, when the commission errors are used as reference (Table 1). A clear pattern is observed if

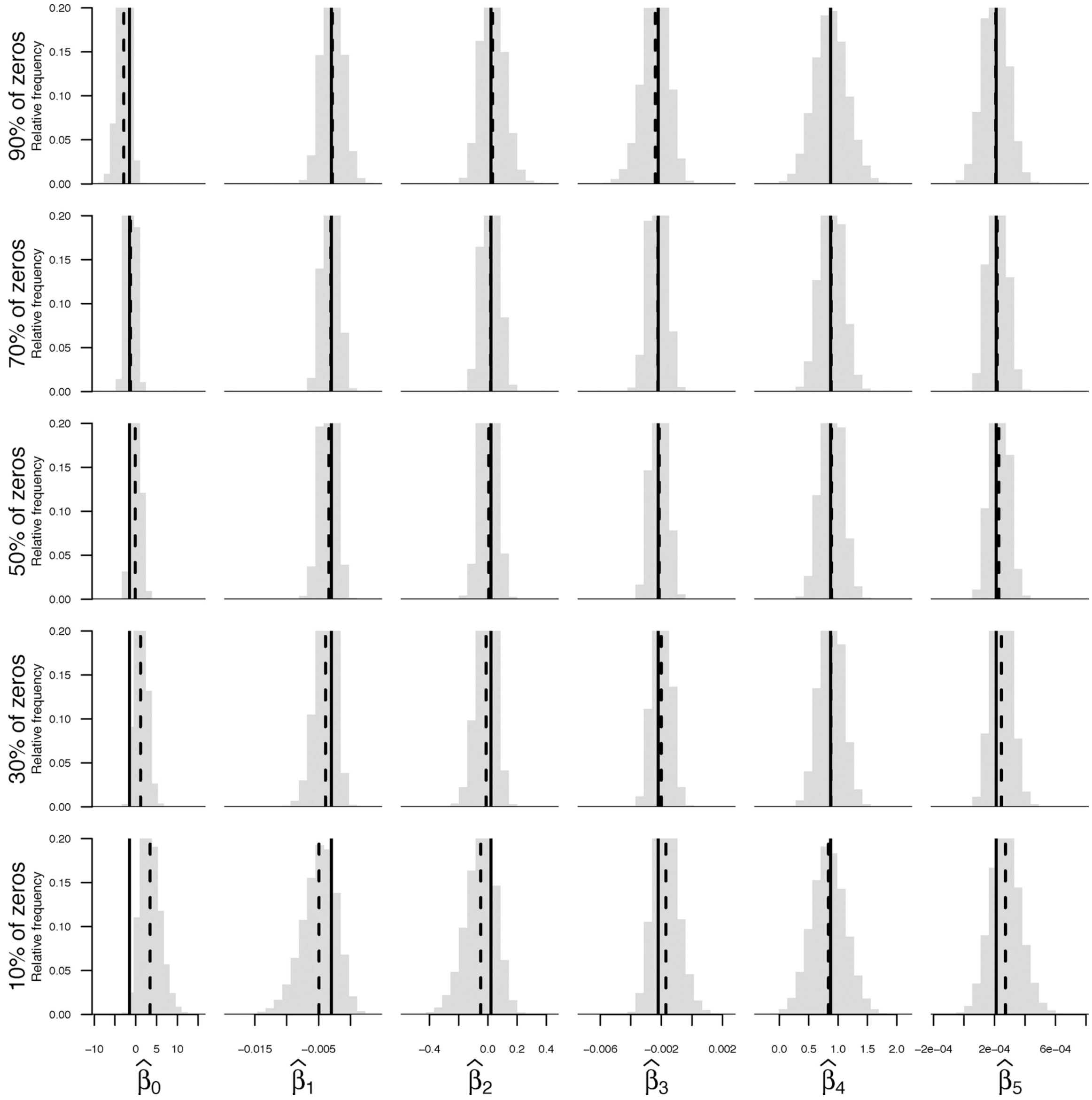


**Fig. 1.** Empirical distribution of estimated parameters for the forest mortality model (Eq. (1)) given different scenarios of zeros in the data. The vertical solid line and the vertical dashed line, within each histogram, represents the parameter value and the Monte Carlo expected value of the estimated parameter, respectively.

specificity or sensitivity are used as reference. Hence, specificity increases with higher number of zeros, but sensitivity decreases as number of zeros increases (Table 1).

## 4. Discussion

In this paper we demonstrated that the common unbalanced proportion of zeros and ones found in ecological data affects the statistical properties of logistic regression models being fitted. Because the variances of the estimated parameters are affected by the proportion of 0/1 data, all the statistical inference (e.g., hypothesis testing) of the fitted model will be affected. Thus, if we are investigating the driver variables of a ecological phenomenon, such as species distribution across a geographic area, we could be erroneously determining them, because the statistical significance of each parameter of the model is based upon their respective variance estimator. Therefore, the practice of balancing data must be carried out with caution, as well as fully considering its implications for model performance. Some authors have argued that there is no major effect in having unbalanced binary data (except for the bias in the intercept parameter, Maddala, 1992), but our results indicate that all statistical properties of the MLE parameters are affected. Although all the parameters estimates are biased, the magnitud of bias will diminish as soon as our sample mimic the proportion of zeros that are found in the population (see the crossed lines in Fig. 2a). Notice that it has previously been stated that all the parameters would be biased for small samples sizes (Schaefer, 1983; King and Zeng, 2001), but that was not necessarily the case in the present study (where $n$ = 1000).

We also claim that the prediction capabilities of the logistic regression model are affected, as also was found by McPherson et al. (2004) and Maggini et al. (2006), but using slightly different statistical models. Thus, a given ecological binary phenomenon could be erroneously predicted to occur (or not) if the fitted model suffers from statistical issues derived from using unbalanced data. This is especially critical for predicting habitat suitability for endangered species (and its conservation) or for predicting the distribution range of exotic invasive species and their subsequent control plans. In either case it can result in allocating efforts and resources in an inefficient manner. In this study we encourage researchers to carefully examine the nature of the data they have available and the 0/1 proportion of it before fitting the

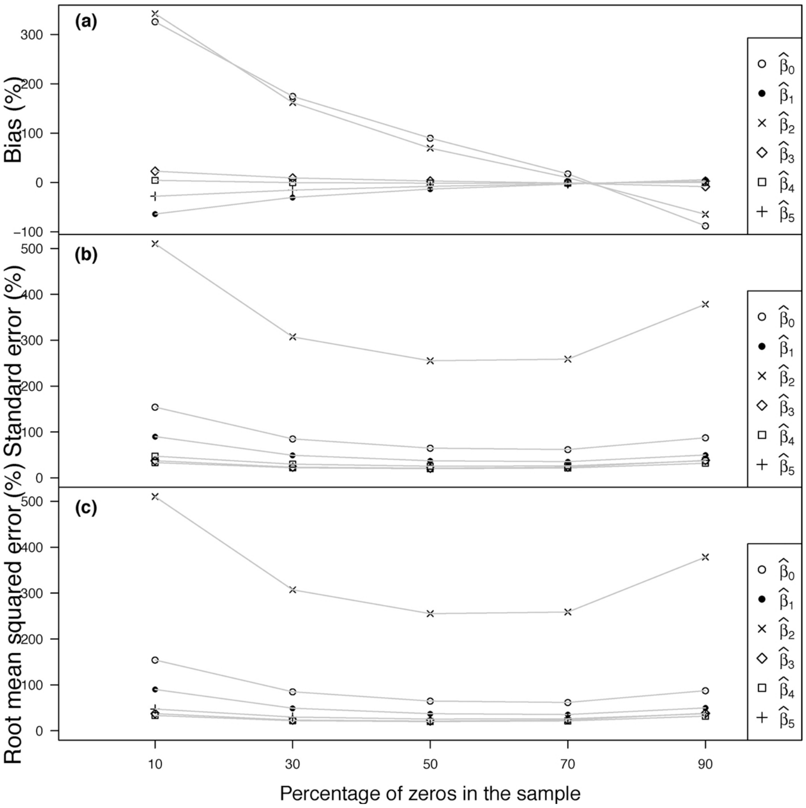


**Fig. 2.** Statistical properties of the estimated parameters for the forest mortality model (Eq. (1)), given different scenarios of zeros in the data. (a) Bias, (b) standard error, and (c) root mean squared error are shown as a percentage of the real parameter value.

**Table 1**
Prediction indexes of the logistic regression model depending upon unbalanced data scenarios. Each value is the empirical expected value of the respective index.

| | Proportions of zeros in the sample | | | | |
|---|---|---|---|---|---|
| | 10% | 30% | 50% | 70% | 90% |
| **Commission** | | | | | |
| Error (%) | 0.41 | 4.44 | 13.09 | 25.12 | 9.99 |
| Accuracy (%) | 89.59 | 65.56 | 36.91 | 4.88 | 0.01 |
| **Omission** | | | | | |
| Error (%) | 9.54 | 21.33 | 21.52 | 4.54 | 0.01 |
| Accuracy (%) | 0.46 | 8.67 | 28.48 | 65.46 | 89.99 |
| Sensitivity (%) | 99.54 | 93.65 | 73.81 | 16.28 | 0.02 |
| Specificity (%) | 4.58 | 28.91 | 56.95 | 93.52 | 99.91 |

logistic regression model, as this can greatly improve both the statistical properties of the estimated parameters of the model and the prediction capabilities applied to ecological phenomena.

We did not focus on analyzing alternatives for overcoming the effects of unbalanced data nor on finding the best fitted model. Some studies dealing with models in ecology have shown the necessity of effectively correct biased analyses for better interpretation and prediction capabilities (Lajeunesse, 2015; Ruffault et al., 2014), but we focused on pointing out the effects of unbalanced data when fitting logistic regression models. The bias in the intercept of a logistic model could be diminished when using the correction given by Manski and Lerman (1977), but this type of correction is better suited for disciplines where higher proportion of ones in the population is more common to find or sample (e.g., social, economy, and political sciences), than for ecological populations.

The statistical inference of fitted logistic regression models is affected by the unbalanced nature of ecological data. Our results show that the largest standard error and root mean squared error of the estimated parameters are found when having extreme proportion of zeros (or ones) in the data. More importantly, for the first time in the literature, as far as we are aware of, we described that the variability of the maximum likelihood estimated (MLE) parameters decreases when having a balanced sample. This finding may suggest that balancing data is an appropriate practice, if statistical inference (e.g., hypothesis testing), is what the researcher is concerned about. Hence, by using unbalanced data, we might conclude that a predictor variable is statistically significant when in fact it is not, or otherwise. Furthermore, this finding refutes what King and Zeng (2001) had claimed regarding that the addition of ones into the data, would decrease the variance of the MLE parameters.

We also found that unbalanced data heavily affects the prediction capabilities of a logistic regression model. Our study reflects that the occurrence of the event is better predicted when having larger proportions of ones in the data. On the other hand, non-occurrence of the event is better predicted when having larger proportions of zeros in the data. This trend is expected, because the model is fitted by ML, where the parameters estimates are those that maximize the likelihood of the data at hand, therefore we should predict them concordantly (Schabenberger and Pierce, 2002). Also, if we take into account the trade-off of building a model that predicts occurrence and non-occurrence as best as possible, the balanced data scenario with a 50% of zeros and ones offers a suitable way to proceed (Table 1). Overall, balancing the data seems to be an appropriate practice to improve some statistical properties and prediction capabilities of the fitted model. Regardless of balancing or not balancing the data before fitting a logistic regression model, we recommend to use the remaining sample (i.e., not used for fitting the model) for validation purposes and behavior analyses.

## 5. Recommendations

Given that the proportion of 0/1 data affects the variance of the estimated parameters of the fitted logistic regression model, the selection of the statistically significant predictor variables to conduct the analyses may also being influenced, ultimately leading to a wrong conclusion. From our study, we have provided evidence that using a balanced data scenario (i.e., 50% of zeros and ones) will yield smaller variances for the maximum likelihood estimates of parameters, therefore offering less uncertainty in the estimation process, and ultimately in identifying the driver variables for modelling presence/absence response variables. This finding is extremely relevant in ecological applications, as an important amount of studies are currently dealing with niche modeling and species distribution based on presence/absence data, especially within the climate change context. Thus, we recommend that when modelling binary response variables, researchers can safely use balanced datasets for fitting candidate models, in order to choose the best model given the variables used for the analysis. Giving our results, by performing these procedure, the analysis itself will gain more certainty because the researcher could better distinguish between the effects of the predictor variables being included in the model or whether is the ecological phenomenon really important (McPherson et al., 2004). Another issue to take into account is the drastic reduction of the sample for balancing purposes. We recommend to use the remaining data (i.e., the one not used for fitting purposes) for assessing the prediction capabilities of the models (using the indices and plots recommend by Jones et al., 2010), and assessing the models behavior by plotting the prediction of the response variable as a function of the predictor variable(s).

Regarding interpreting the predicted outcomes, we recommend not extrapolating the model results into areas where predictor variables were not measured. This implies that the presence/absence of a given organisms could be altered within certain ranges. In the case when extrapolation is indeed necessary, researchers should differentiate their predictions from those areas where no data were collected using, for instance, color-codded results to distinguish them from prediction results by the model using real data. This situation will be specially advantageous for modelling any ecological phenomena that is a function of spatially-recorded predictor variable(s).

## 6. Concluding remarks

The proportion of zeros and ones in a dataset affects the statistical inference and prediction capabilities of a fitted logistic regression model. Not only the accuracy of the estimated parameters is affected by unbalanced data, but also their precision. More importantly, the statistical inference (e.g., hypothesis testing) is influenced by the proportion of zeros and ones in the data. In addition, the prediction capabilities of the fitted logistic regression model are affected as well, therefore the model performance would greatly depend on the proportion of 0/1 data. Overall, the 0/1 proportion might affect the conclusions that can arise from the fitted model and its further application. Since unbalanced data in ecology are fairly common, this can have great implications in model building of several ecological phenomena being modelled by scientists.

## Acknowledgements

This study was supported by the Chilean research grant Fondecyt No. 1151495. AFR is supported by a Postdoctoral Scholarship from Vicerrectoría de Investigación y Postgrado, Universidad de La Frontera, Temuco, Chile.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ecolind.2017.10.030.

## References

Affleck, D.L.R., 2006. Poisson mixture models for regression analysis of stand level mortality. Can. J. For. Res. 36, 2994–3006.

Alberini, A., 1995. Testing willingness-to-pay models of discrete-choice contingent valuation survey data. Land Econ. 71 (1), 83–95.

Arana, J.E., Leon, C.J., 2005. Flexible mixture distribution modeling of dichotomous choice contingent valuation with heterogenity. J. Environ. Econ. Manage. 50 (1), 170–188.

Bastin, L., Thomas, C.D., 1999. The distribution of plant species in urban vegetation fragments. Landsc. Ecol. 14 (5), 493–507.

Bell, C.D., Roberts, R.K., English, B.C., Park, W.M., 1994. A logit analysis of participation in Tennessee's forest stewardship program. J. Agric. Appl. Econ. 26 (2), 463–472.

Bigler, C., Kulakowski, D., Veblen, T.T., 2005. Multiple disturbance interactions and drought influence fire severity in Rocky mountain subalpine forests. Ecology 86 (11), 3018–3029.

Bradstock, R.A., Hammill, K.A., Collins, L., Price, O., 2010. Effects of weather, fuel and terrain on fire severity in topographically diverse landscapes of south-eastern Australia. Landsc. Ecol. 25 (4), 607–619.

Brook, B.W., Bowman, D.M.J., 2006. Postcards from the past: charting the landscape-scale conversion of tropical Australian savanna to closed forest during the 20th century. Landsc. Ecol. 21, 1253–1266.

Cailleret, M., Bigler, C., Bugmann, H., Camarero, J.J., Cufar, K., Davi, H., Meszaros, I., Minunno, F., Peltoniemi, M., Robert, E.M.R., Suarez, M.L., Tognetti, R., Martinez-Vilalta, J., 2016. Towards a common methodology for developing logistic tree mortality models based on ring-width data. Ecol. Appl. 26 (6), 1827–1841.

Chao, K.-J., Phillips, O.L., Monteagudo, A., Torres-Lezama, A., Vásquez, R., 2009. How do trees die? Mode of death in northern Amazonia. J. Veg. Sci. 20, 260–268.

Davies, S.J., 2001. Tree mortality and growth in 11 sympatric Macaranga species in Borneo. Ecology 82 (4), 920–932.

Dickson, B.G., Prather, J.W., Xu, Y.G., Hampton, H.M., Aumack, E.N., Sisk, T.D., 2006. Mapping the probability of large fire occurrence in Northern Arizona. Landsc. Ecol. 21 (2), 747–761.

Eastman, J.R., 2006. Idrisi 15 andes, guide to GIS and Image Processing. Clark University, Worcester, MA, USA.

Echeverria, C., Coomes, D.A., Newton, M.H.A.C., 2008. Spatially explicit models to analyze forest loss and fragmentation between 1976 and 2020 in southern Chile. Ecol. Model. 212, 439–449.

Firth, D., 1993. Bias reduction of maximum likelihood estimates. Biometrika 80, 27–38.

Gregoire, T.G., Salas, C., 2009. Ratio estimation with measurement error in the auxiliary variate. Biometrics 65 (2), 590–598.

Gregoire, T.G., Schabenberger, O., 1999. Sampling-skewed biological populations: behavior of confidence intervals for the population total. Ecology 80 (3), 1056–1065.

Hastie, T., Fithian, W., 2013. Inference from presence-only data; the ongoing controversy. Ecography 36, 864–867.

Hu, X., Wu, C., Hong, W., Qiu, R., Li, J., Hong, T., 2014. Forest cover change and its drivers in the upstream area of the Minjiang River, China. Ecol. Indic. 46, 121–128.

Jones, C.C., Acker, S.A., Halpern, C.B., 2010. Combining local- and large-scale models to predict the distributions of invasive plant species. Ecol. Appl. 20 (2), 311–326.

King, G., Zeng, L., 2001. Logistic regression in rare events data. Polit. Anal. 9 (2), 137–163.

Komori, O., Eguchi, S., Ikeda, S., Okamura, H., Ichinokawa, M., Nakayama, S., 2016. An asymmetric logistic regression model for ecological data. Methods Ecol. Evol. 7, 249–260.

Kumar, R., Nandy, S., Agarwal, R., Kushwaha, S.P.S., 2014. Forest cover dynamics analysis and prediction modeling using logistic regression model. Ecol. Indic. 45, 444–455.

Lajeunesse, M.J., 2015. Bias and correction for the log response ratio in ecological meta-analysis. Ecology 96 (8), 2056–2063.

Lander, T.A., Bebber, D.P., Choy, C.T., Harris, S.A., Boshier, D.H., 2011. The circe principle explains how resource-rich land can waylay pollinators in fragmented landscapes. Curr. Biol. 21, 1302–1307.

Leyk, S., Zimmermann, N.E., 2007. Improving land change detection based on uncertain survey maps using fuzzy sets. Landsc. Ecol. 22 (2), 257–272.

Lindsey, J.K., 1997. Applying Generalized Linear Models. Springer, New York, USA, pp. 256.

Lloret, F., Calvo, E., Pons, X., Diaz-Delgado, R., 2002. Wildfires and landscape patterns in the eastern Iberian peninsula. Landsc. Ecol. 17 (8), 745–759.

Maddala, G.S., 1992. Introduction to Econometrics, 2nd ed. Macmillan Publishing Company, New York, NY, USA, pp. 631.

Maggini, R., Lehmann, A., Zimmermann, N.E., Guisan, A., 2006. Improving generalized regression analysis for the spatial prediction of forest communities. J. Biogeogr. 33 (10), 1729–1749.

Manski, C.F., Lerman, S.R., 1977. The estimation of choice probabilities from choice based samples. Econometrica 45 (8), 1977–1988.

McPherson, J.M., Jetz, W., Rogers, D.J., 2004. The effects of species range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? J. Appl. Ecol. 41 (5), 811–823.

Mermoz, M., Kitzberger, T., Veblen, T.T., 2005. Landscape influences on occurrence and spread of wildfires in Patagonian forests and shrublands. Ecology 86 (10), 2705–2715.

Palma, C., Cui, W., Martell, D., Robak, D., Weintraub, A., 2007. Assessing the impact of stand-level harvests on the flammability of forest landscapes. Int. J. Wildl. Fire 16 (5), 584–592.

Phillips, S.J., Elith, J., 2013. On estimating probability of presence from use-availability or presence-background data. Ecology 94 (6), 1409–1419.

Qi, Y., Wu, J., 1996. Effects of changing spatial resolution on the results of landscape pattern analysis using spatial autocorrelation indices. Landsc. Ecol. 11 (1), 39–49.

R Development Core Team, 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.Rproject.org.

Rueda, X., 2010. Understanding deforestation in the southern Yucatán: insights from a sub-regional, multi-temporal analysis. Reg. Environ. Change 10 (3), 175–189.

Ruffault, J., Martin-StPaul, N.K., Duffet, C., Goge, F., Mouillot, F., 2014. Projecting future drought in mediterranean forests: bias correction of climate models matters!. Theor. Appl. Climatol. 117 (1–2), 113–122.

Salas, C., Gregoire, T.G., 2010. Statistical analysis of ratio estimators and their estimators of variances when the auxiliary variate is measured with error. Eur. J. For. Res. 129 (5), 847–861.

Schabenberger, O., Pierce, F.J., 2002. Contemporary Statistical Models for the Plant and Soil Sciences. CRC Press, Boca Raton, FL, USA, pp. 738.

Schaefer, R.L., 1983. Bias correction in maximum likelihood logistic regression model. Stat. Med. 2, 71–78.

Schulz, J.J., Cayuela, L., Rey-Benayas, J.M., Schröder, B., 2011. Factors influencing vegetation cover change in Mediterranean Central Chile (1975–2008). Appl. Veg. Sci. 14, 571–582.

Scott, A.J., Wild, C.J., 1986. Fitting logistic models under case–control or choice based sampling. J. R. Stat. Soc. B 78 (2), 170–182.

Seto, K.C., Kaufmann, R.F., 2005. Using logit models to classify land cover and land-cover change from Landsat Thematic Mapper. Int. J. Rem. Sens. 25 (3), 563–577.

Vega-García, C., Chuvieco, E., 2006. Applying local measures of spatial heterogeneity to Landsat-TM images for predicting wildfire occurrence in Mediterranean landscapes. Landsc. Ecol. 21, 596–605.

Vega-García, C., Woodard, P., Titus, S., Adamowicz, W., Lee, B., 1995. A logit model for predicting the daily occurrence of human caused forest fires. Int. J. Wildl. Fire 5 (2), 101–111.

Vega-García, C., Woodard, P.M., Lee, B.S., Adamowicz, W.L., Titus, S.J., 1999. Dos modelos para la predicción de incendios forestales en Whitecourt Forest, Canadá. Investigación Agraria: Sistemas y Recursos Forestales 8 (1), 5–24.

Warton, D.I., Hui, F.K.C., 2011. The arcsine is asinine: the analysis of proportions in ecology. Ecology 92 (1), 3–10.

Wilson, K., Newton, A., Echeverría, C., Weston, C., Burgman, M., 2005. A vulnerability analysis of the temperate forests of south central Chile. Biol. Conserv. 122, 9–21.

Wu, J., Gao, W., Tueller, P.T., 1997. Effects of changing spatial scale on the results of statistical analysis with landscape data: a case study. Geogr. Inf. Sci. 3 (1-2), 30–41.

Wunder, J., Reineking, B., Bigler, C., Bugmann, H., 2008. Predicting tree mortality from growth data: how virtual ecologists can help real ecologists. J. Ecol. 96 (1), 174–187.

Xie, Y., Manski, C.F., 1989. The logit model and response-based samples. Sociol. Methods Res. 17 (3), 283–302.

Young, D.J.N., Stevens, J.T., Earles, J.M., Moore, J., Ellis, A., Jirka, A.L., Latimer, A.M., 2017. Long-term climate and competition explain forest mortality patterns under extreme drought. Ecol. Lett. 20 (1), 78–86.

Zuur, A.F., Ieno, E.N., Elphick, C.S., 2010. A protocol for data exploration to avoid common statistical problems. Methods Ecol. Evol. 1 (1), 3–14.