

JULY 2005

Analysis of Presence/Absence Data when Absence is Uncertain (False Zeroes): An Example for the Northern Flying Squirrel using SAS®

J. D. Steventon
B.C. Ministry of Forests
Northern Interior Forest Region
Bag 6000 Smithers, BC V0J 2N0
Doug.Steventon@gov.bc.ca

W. A. Bergerud
B.C. Ministry of Forests
Research Branch
P.O. Box 9519 Stn Prov Govt
Victoria, BC V8W 9C2
Wendy.Bergerud@gov.bc.ca

P. K. Ott
B.C. Ministry of Forests
Research Branch
P.O. Box 9519 Stn Prov Govt
Victoria, BC V8W 9C2
Peter.Ott@gov.bc.ca

Introduction

Predicting species presence/absence, or other binomial responses from survey data, is common in wildlife and forestry research. In situations where a recorded absence may in fact represent a failure to detect what is actually there (a false zero) erroneous inferences may result with naïve application of procedures such as logistic regression (MacKenzie and Kendall 2002; Tyre et al. 2003; Wintle et al. 2004).

An example where this can be expected is in live-trapping studies of wildlife occurrence. With false zeroes, not only can the proportion of sites truly occupied be underestimated, but the relationship with explanatory variables (the predictive model, usually the aim of the exercise) may also be biased. Furthermore, the chances of a false zero may not be constant but rather a function of time or sample site characteristics, making comparisons among samples questionable.

Using SAS®¹ v9.1 we implemented the Zero-Inflated Binomial (ZIB) method of Tyre et al. (2003) that incorporates estimation of false zero

rates. In addition we illustrate the use of Akaike's Information Criterion (AIC) to weight alternative models, and the use of the area under the curve (AUC) of the Receiver Operating Characteristic curve (ROC) for assessing model fit and prediction accuracy (Cumming 2000; Boyce et al. 2002). Our purpose is to show how to implement these methods using SAS. We strongly recommend consulting the references for more thorough discussions of the theory, appropriate application, and limitations of the methods.

ZIB models consider the probability of observing an animal as the product of two independent processes at the time of survey: 1) the probability that the animal is present at the site (probability of occupancy), and 2) the probability of detecting an animal when it is there (detection probability). The method can apply, however, to any binomial data where one of the classifications (0 or 1) is subject to classification error.

We illustrate the methods with an example dataset from a pilot live-capture study of northern flying squirrels (*Glaucomus sabrinus*) (J.D. Steventon, unpublished data). The

¹ SAS Institute, Cary, N.C., USA.

objective was to test whether detection probability varied by date, moving from summer into fall, and if the probability of occupancy by squirrels was affected by a habitat-density score² for a 16-ha square around each of the 135 trap-sites. Each site was live-trapped (i.e., surveyed) for either 5 or 7 nights.

We had four competing models to assess:

- A. squirrel captures are a function of both date affecting detection probability, and habitat density affecting occupancy probability;
- B. squirrel captures are a function of date affecting detection probability, while occupancy probability is constant;
- C. squirrel detection probability is constant, while occupancy is affected by habitat density; and
- D. both detection probability and occupancy probability are constant.

The Zero-Inflated Binomial Procedure

ZIB considers the probability of observing an animal as the product of the true probability of the site being occupied (p), and the probability of detection (q) when in fact the site is occupied at the time of survey. The false zero probability (i.e., not observing an animal when it is present) is $1 - q$. Multiple surveys (preferably three or more) must be conducted at each site (Tyre et al. 2003) to estimate q , although the number of surveys (m) does not have to be equal for all sites. The number of observations of an animal for each site over m surveys is denoted as y , and the number of sites sampled by n .

To begin with, consider one site after m surveys have been conducted.

2 An area-weighted average Habitat Suitability Index score for the 16-ha square surrounding each site.

The probability of observing zero animals is:

$$\begin{aligned} P(y=0) &= P(\text{animal is present}) \times \\ &\quad P(\text{animal is not detected}|\text{animal is present}) + \\ &\quad P(\text{animal is not present}) \times \\ &\quad P(\text{animal is not detected}|\text{animal is not present}) \\ &= p(1-q)^m + (1-p)(1) \end{aligned} \quad (1)$$

Likewise, the probability of exactly y animals, where y is greater than 0 is:

$$\begin{aligned} P(y>0) &= P(\text{animal is present}) \times \\ &\quad P(\text{animal is detected}|\text{animal is present}) + \\ &\quad P(\text{animal is not present}) \times \\ &\quad P(\text{animal is detected}|\text{animal is not present}) \\ &= p \binom{m}{y} q^y (1-q)^{m-y} + (1-p)(0) \end{aligned} \quad (2)$$

Combining the above two probabilities leads to the likelihood for a single site (Tyre et al. 2003):

$$L(p, q | y, m) = \begin{cases} p \binom{m}{y} q^y (1-q)^{m-y}, & y > 0 \\ (1-p) + p(1-q)^m, & y = 0 \end{cases} \quad (3)$$

The above probability can also be expressed as a generalized Bernoulli distribution, making it more apparent on how to combine multiple (independent) sites into a full likelihood:

$$L(p, q | \{y_i, m_i, u_i\}) = \prod_{i=1}^n \left[(1-p) + p(1-q)^{m_i} \right]^{u_i} \times \left[p \binom{m_i}{y_i} q^{y_i} (1-q)^{m_i - y_i} \right]^{1-u_i} \quad (4)$$

where u_i is an indicator variable: $u_i = 1$ when $y_i = 0$ and $u_i = 0$ when $y_i > 0$.

The values of p and q need not be constant, and can be influenced by covariates. In most exercises we are most interested in modelling p , with q being a necessary nuisance. Covariates influence the log odds of p or q as:

$$\log \left(\frac{p_i}{1-p_i} \right) = \alpha_0 + \alpha_1 x_{1i} + \dots + \alpha_k x_{ki} \quad (5)$$

and

$$\log \left(\frac{q_i}{1-q_i} \right) = \beta_0 + \beta_1 x_{1i} + \dots + \beta_l x_{li} \quad \text{where } i = 1, 2, \dots, n \quad (6)$$

Here x_{ji} is the value of the j th covariate at site i , and α_j, β_j are coefficients that describe the influence of the covariates on the log odds. Some or all of that covariate may be common to both equation 5 and equation 6, or they may be entirely different (as in our squirrel example).

The reason for modelling the log odds, via the logit link function, is to linearize the predictive equation (see, for example, Section 2.3 of Bergerud 1996). However, the natural relationship between p (or q) and the covariates is nonlinear and involves the exponential function. The form of

this relationship is apparent by solving equations 5 (or 6) with respect to p (or q), and we use this back-transformed version in our SAS programming (e.g., lines 6 and 7 of the SAS code below) and in equations 7 and 8.

The following SAS code uses proc nlmixed to solve the equations using maximum likelihood³ (Table 1). We assume the SAS dataset “flyers” has already been created. Additional covariates for predicting p and q can be added in lines 2, 3, and 4. If p or q are to be constants, add them as parameters to line 2 while excluding lines 3, 4, 6, and 7 as appropriate.

```

1. proc nlmixed data=flyers
   df=131; ** change df=as
   appropriate4;
2. parms a0_est=0.3 a1_est=
   -0.5 b0_est=-8 b1_est=
   0.08;
3. expbitp=exp(a0_est +
   a1_est*x1); ** Model
   definition;
4. expbitq=exp(b0_est +
   b1_st*DateValue);
   ** Model definition;
5. combin=gamma(m+1)/(gamma
   (y+1)*gamma(m-y+1));
   ** Constant;
6. p_est=expbitp/(1 +
   expbitp); ** Probability
   of occupancy;
7. q_est=expbitq/(1 +
   expbitq); ** Probability
   of detection;
8. if y=0 then prob=
   (1 - p_est) + p_est *
   (1 - q_est)**m;
9. else prob=p_est * combin
   * (q_est**y)*(1 -
   q_est)**(m-y);
10. loglike=log(prob);
11. model y ~ general
   (loglike); /** output for
   use in ROC **/
12. CapExp=p_est*(1-(1-
   q_est)**m); ** Prob of
   observing at least one
   animal;
13. predict CapExp
   OUT=ROCDATA;
14. run;

```

- 3 Proc nlmixed allows the user to specify a customized log-likelihood function. In our case this is simply the log of equation 4; in general, it is easier numerically to optimize a log-likelihood than a raw likelihood.
- 4 Proc nlmixed often calculates the df incorrectly so this must be done manually. It should be the number of observations minus the number of parameters in the model.

TABLE 1 Parameter estimates from output of proc nlmixed for model A

Parameter	Estimate	Standard Error	DF	t Value	Pr > t	Alpha	Lower	Upper
a0_est	0.3964	1.1167	131	0.36	0.7232	0.05	-1.8121	2.6048
a1_est	0.2945	1.7147	131	0.17	0.8639	0.05	-3.0967	3.6857
b0_est	-9.5346	2.1987	131	-4.34	<0.0001	0.05	-13.8830	-5.1862
b1_est	0.08439	0.0248	131	3.41	0.0009	0.05	0.0354	0.1334

Note: a0_est is the intercept and a1_est is the coefficient for the covariate “habitat density” when modelling the log odds of p .
b0_est is the intercept and b1_est is the coefficient for the covariate “date value” when modelling the log odds of q .

The predicted probability of occupancy is:

$$p = e^{(0.396 - 0.295 x_1)} / (1 + e^{(0.396 - 0.295 x_1)}) \quad (7)$$

and the predicted probability of detection is:

$$q = e^{(-9.535 + 0.084 \text{DateValue})} / (1 + e^{(-9.535 + 0.084 \text{DateValue})}) \quad (8)$$

The estimated probability of capturing an animal in any given survey night is

$$pq \quad (9)$$

and the probability of capturing at least one animal after m trap-nights is

$$p[1 - (1 - q)^m] \quad (10)$$

Equation 10 will be used later when we examine ROC.

In this example (Figure 1), the predicted probability of catching at least one flying squirrel increased exponentially with date value going from summer into fall. The same probability decreased with habitat density, but the evidence supporting this prediction appears weak.

Applying model B, p_est was 0.64 (64% of sites occupied), vs. a “naïve” estimate (not accounting for detection probability) of 23 sites with captures

÷ 135 sites trapped for p_est of 0.17 (17% of sites occupied).

Model Selection and Weighting Using Akaike’s Information Criterion (AIC)

After fitting alternative models to the data, the next step is to compare the models, either to select the “best” model or to apply model weightings. Akaike’s Information Criterion (AIC) is rapidly becoming an established method of comparing models that accounts for the dangers of under- and over-fitting (Johnson and Omland 2003). It considers both the likelihood of the model and the number of parameters (Burnham and Anderson 2002). The output from proc nlmixed includes the AICc values (the small-sample corrected AIC), with a smaller AICc indicating a better model. Note that AIC is a relative score among models fit to the same data, and cannot be directly compared to AIC scores fit to other datasets. In our example all the models were nested, but this need not be the case for applying AIC.

The AIC weight to apply to each model is calculated as follows. First, take the difference in AIC scores for each model by subtracting its AIC score from the score of the best model (lowest AIC).

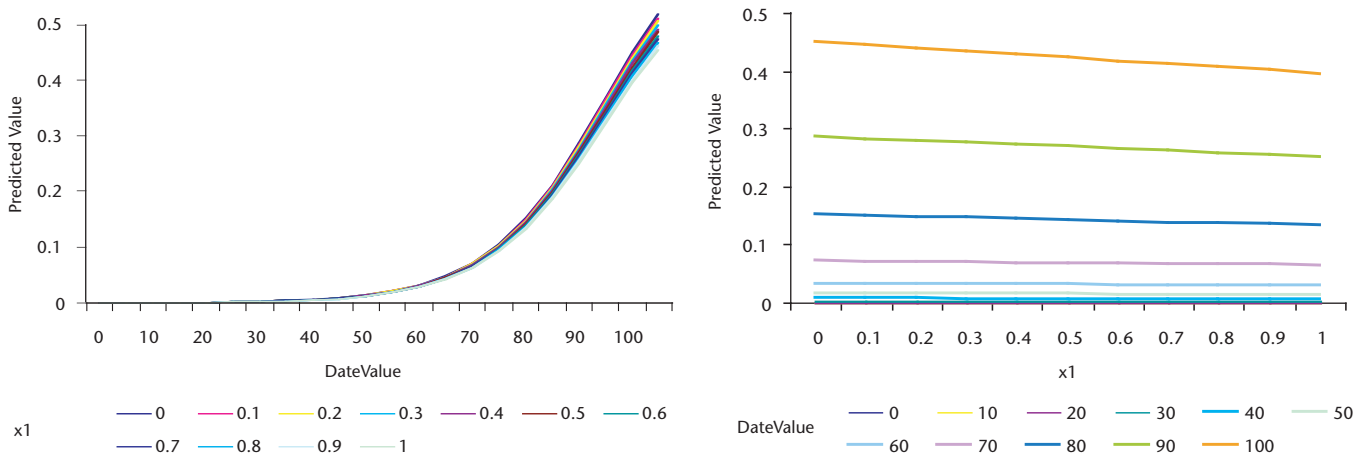


FIGURE 1 Predicted probability of capturing at least one squirrel (applying equation 10 with $m = 5$), as a function of DateValue and habitat density score (x_1). DateValue dominates the prediction.

The weight for each model is then calculated as (Anderson et al. 2000):

$$w_m = \frac{\exp\left(-\frac{1}{2}\Delta_m\right)}{\sum_{m=1}^M \exp\left(-\frac{1}{2}\Delta_m\right)} \quad (11)$$

where M is the total number of models being compared, and

$$\Delta_m = AICc_m - \min(AICc)$$

This weighting is then interpreted as the relative probability of each model being the “best” (i.e., best predictive ability) from among the models compared (Table 2).

Models C and D are very poor and should be rejected. The date-only model (B) is “best,” with a weight almost 3× that of the date/habitat-density model (A). However, it does not mean that any model is correct in an absolute sense, or that some model not considered is not better (Burnham and Anderson 2002).

In order to assess classification accuracy of the two models with potential (A and B), we then applied the Receiver Operating Characteristic curve (ROC) method (see following).

TABLE 2 Model comparison using AICc. Models are sorted in ascending order of AICc.

Model	AICc	Δ_m	W_m
B	131.3	0	0.74
A	133.4	2.1	0.26
C	160.2	28.9	<0.01
D	162.2	30.9	<0.01

Model A assumes that p is a function of habitat-density score, and that q is a function of date.

Model B assumes that p is constant, and that q is a function of date.

Model C assumes that q is constant, and that p is a function of habitat-density score.

Model D assumes that p and q are both constant.

Receiver Operating Characteristic Curve (ROC) and Area Under the Curve (AUC)

The ROC was developed in the 1950s and has been widely applied in the medical and engineering fields (McPherson et al. 2004). It is used to assess the fit or predictive accuracy of dichotomous models, and is now being used more frequently in the natural sciences (Boyce et al. 2002).

AUC is a robust measure of model performance/accuracy, relatively insensitive to prevalence (proportion of 0's or 1's), although less stable (higher

standard error) for prevalence less than 0.2 (McPherson et al. 2004). ROC and AUC do not require the choice of an arbitrary prediction probability threshold defining “capture” vs. “non-capture,” but rather summarize performance across the range of possible thresholds (Cumming 2000).

An AUC of 0.8, for example, means that in eight out of ten instances a site selected at random from those *with a capture* had a higher predicted probability of capture than a site randomly selected from those that *had no capture*. An AUC of 0.5 would indicate no ability of the model to discriminate among sites with captures vs. sites with no captures.

An ROC curve (Figure 2) is produced (e.g., Table 4 of Cumming 2000) by plotting the proportion of sites correctly classified as “captured” (termed Sensitivity) against the proportion of sites incorrectly classified as “captured” (termed $1 - \text{Specificity}$), for all possible thresholds of predicted capture probability (equation 10). Low thresholds are represented on the right side of the graph and high thresholds on the left. The steeper the curve rises and then flattens near 1, the better the model. The curve for a random model with no predictive ability would follow the diagonal from

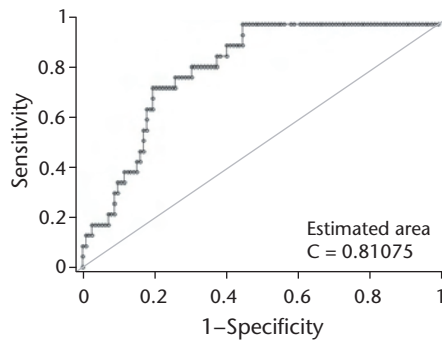


FIGURE 2 ROC curve for model A as output by SAS.

the x,y origin to 1,1.

ROC curves can be plotted quite readily by hand, in a spreadsheet, or with SAS. In our example the observations are first ordered in decreasing predicted probability of a capture from highest through lowest. Starting at the lower left of the graph with the first observation, the proportion of sites with captures that are above the predicted probability threshold (y-axis) is plotted against the proportion of sites with non-captures that are above that threshold (x-axis). This is repeated in sequence for each observation. For threshold values lower than the last observation, both y-axis and x-axis values will be 1 because all captures and non-captures have predicted probabilities greater than the threshold. (For a useful tutorial on ROC curves and how they are plotted, we recommend the web page <http://www.anaesthetist.com/mnm/stats/roc/>.)

In SAS, proc logistic has a handy option to produce the data necessary for plotting ROC (Figure 2) and calculating AUC. We implemented it in conjunction with our ZIB model using the coding below (SAS Institute 1999). A dataset created in proc nlmixed (shown earlier) provides the necessary data, and the fit statistic “c” in the proc logistic output is equivalent to AUC. Note that proc logistic does not produce the correct results for plotting ROC when sampling weights are used (Izrael et al. 2003), so be careful

how the dataset is structured. In our example, the dataset for proc logistic has a separate unweighted data record for each trapping site, consisting of the predicted capture probability (CapExp) and the response variable (SiteOc 1 or 0).

We obtained an AUC of 0.811 for model A and 0.815 for the date-only model (model B), again suggesting that the date-only model was a good predictor and that the more complex model provided little additional predictive value. We have not implemented it, but Izrael et al. (2000) present a bootstrap method to formally assess whether the predictive ability among models is significantly different. (The web tutorial at <http://www.anaesthetist.com/mnm/stats/roc/> also presents methods [which we have not verified] of calculating standard errors for ROC curves, for comparing curves, and for calculating sample sizes for AUC.)

```

1. Data ROC;
2. set ROCDATA;
3. if y > 0 then SiteOc=1;
4. if y=0 then SiteOc=0;
5. CapExp=Pred;
6. run;
7. ods html;
8. ods graphics on;
9. proc logistic descending
   data=ROC order=data;
10. model SiteOc
    (event='1')=CapExp /
    roceps=0 outroc=roc1
    (keep=_sensit_
    _1mspec_);
11. output out=preds
    prob=problog;
12. run;
13. ods graphics off;
14. ods html close;
15. symbol1 i=join v=none
    c=blue;
16. proc gplot data=roc1;
17. title 'ROC Curve';
18. plot _sensit_*_1mspec_
    =1/ vaxis=0 to 1 by .1
    cframe=ligr;
19. run;
20. quit;

```

The SAS output below from proc logistic is for squirrel model A, with the value “c” (bolded) equivalent to AUC:

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	79.9	Somers'D	0.622
Percent Discordant	17.8	Gamma	0.636
Percent Tied	2.3	Tau-a	0.177
Pairs	2576	C	0.811

Discussion

Accounting for false zeroes (i.e., accounting for detection probability) is necessary in presence/absence studies to avoid potentially serious errors of estimation and inference. Furthermore, AIC and AUC provide the ability to objectively compare alternative models and to assess classification performance. We show how to readily implement these procedures in SAS.

For our example dataset, flying squirrel detection probability was strongly affected by date moving from summer into fall, as was also reported for interior British Columbia by Ransome et al. (2004). This date effect is presumably a reflection of a change in squirrel foraging behaviour. The negative effect of habitat-density score on squirrel occurrence was unexpected, and may or may not be real based on the data at hand. We could, however, speculate on potential mechanisms such as greater overlap with predators, increased density due to packing effects, or a preference for edge.

Based on the pilot data, future sampling will be conducted later in the fall in order to take advantage of higher detection probability and thus greatly increased power (for a given trapping effort) to assess occupancy probability models. Tyre et al. (2003) suggested that, with detection probabilities < 0.5, an increased number of visits at each site is most efficient for

estimating p ; whereas with detection probabilities > 0.5 , increasing the number of sites is most efficient. For the squirrels, by simply delaying the trapping into the fall, detection probability can be increased and effort applied to an increased number of sites.

References

- Anderson, D.R., K.P. Burnham, and W.L. Thompson. 2000. Null hypothesis testing: problems, prevalence and an alternative. *Journal of Wildlife Management* 64:912–23.
- Bergerud, W.A. 1996. Introduction to logistic regression models: with worked forestry examples. Res. Br., B.C. Min. For., Victoria, B.C. Biometrics Info. Handb. 7. <<http://www.for.gov.bc.ca/hfd/pubs/Docs/Wp/wp26.htm>>.
- Boyce, M.S., P.R. Vernier, S.E. Nielson, and F.K.A. Schmiegelow. 2002. Evaluating resource selection functions. *Ecological Modelling* 157: 281–300.
- Burnham, K.P. and D.R. Anderson. 2002. Model selection and multi-model inference: a practical information-theoretic approach. Springer, New York, N.Y.
- Cumming, G.S. 2000. Using between-model comparisons to fine-tune linear models of species ranges. *Journal of Biogeography* 27: 441–55.
- Izrael, D., A.A. Battaglia, D.C. Hoaglin, and M.P. Battaglia. 2003. Use of the ROC curve and the bootstrap in comparing weighted logistic regression models/SAS Users Group International, Paper 248–27. SAS Institute [on line] <<http://www2.sas.com/proceedings/sugi27/p248-27.pdf>> – 178.5KB – ROC: 16.
- Johnson, J.B. and K.S. Omland. 2004. Model selection in ecology and evolution. *Trends in Ecology and Evolution* 19: 101–8.
- MacKenzie, D.I. and W.L. Kendall. 2002. How should detection probability be incorporated into estimates of relative abundance? *Ecology* 83: 2387–93.
- McPherson, J.A., W. Jetz, and D.J. Rogers. 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology* 41: 811–23.
- Ransome, D.B., P. Lindgren, D.S. Sullivan, and T.P. Sullivan. 2004. Long-term responses of ecosystem components to stand thinning in young lodgepole pine forest. I. Population dynamics of northern flying squirrels and red squirrels. *Forest Ecology and Management* 202: 355–67.
- SAS Institute. 1999. SAS online documentation, v8. LOGISTIC procedure. Cary, N.C. [on line] <<http://v8doc.sas.com/sashtml/>>.
- Tyre, A.J., B. Tenhumberg, S.A. Field, D. Niejalke, K. Parris, and H.P. Possingham. 2003. Improving precision and reducing bias in biological surveys: estimating false-negative error rates. *Ecological Applications* 13: 1790–801.
- Wintle, B.A., M.A. McCarthy, K.M. Parris, and M.A. Burgman. 2004. Precision and bias of methods for estimating point survey detection probabilities. *Ecological Applications* 14: 703–12.

Acknowledgements

Partial funding of this project was provided by the B.C. Ministry of Forests, Forest Science Program, Forest Innovation Account, via a grant to the Bulkley Valley Centre for Natural Resources Research and Management. Ardea Biological Consulting (Laurence Turney, Anne-Marie Roberts, Lis Rach, Sara Jaward) and Raven Ecological Services (Deb Wellwood) conducted the squirrel trapping. Houston Forest Products (Mellissa Todd) provided field support (Lauren MacDonald, Anne Macleod). Don Morgan assisted with live trapping and obtaining GIS files. Eric LoFroth and Walt Klenner loaned trapping equipment. Steven Smith provided helpful editorial review, and Paul Nystedt and Rick Scharf assisted with publication.

Citation: Steventon, J.D., W.A. Bergerud, and P.K. Ott. 2005. Analysis of presence/absence data when absence is uncertain (false zeroes): an example for the northern flying squirrel using SAS. Res. Br., B.C. Min. For. Range, Victoria, B.C. Exten. Note 74 <<http://www.for.gov.bc.ca/hfd/pubs/Docs/En/En74.htm>>.

The use of trade, firm, or corporation names in this publication is for the information and convenience of the reader. Such use does not constitute an official endorsement or approval by the Government of British Columbia of any product or service to the exclusion of others that may also be suitable. This Extension Note should be regarded as technical background only.