Ecological Archives A021-053-A1

Ephraim M. Hanks, Mevin B. Hooten, and Fred A. Baker. 2011. Reconciling multiple data sources to improve accuracy of large-scale prediction of forest disease incidence. *Ecological Applications* 21:1173–1188.

Appendix A: Simulation Study

We have conducted a simulation study to verify the ability of the approach we have presented here to produce meaningful results. The main goal of this study is to show the appropriateness of the stands surveyed in the intensive survey as the basis for predictions across the range of the DNR inventory. The stands in the intensive survey are not a random sample of all stands in the DNR inventory; rather, the survey was conducted before the DNR inventory was released to the public. Thus, the intensive survey can be thought of as a sample of convenience. For predictions based on this survey to be valid, we must show that the stands surveyed are representative of the stands in the DNR, and that inference made from this data is not likely to be compromised by the sampling design.

We give particular attention to the discrepancy between the proportion of stands in the DNR inventory that are labeled as having mistletoe present (11%) and the corresponding proportion of stands in the USU intensive survey that the DNR classified as having mistletoe present (17%). This discrepancy is larger than might be expected by random chance. If we were to randomly select 196 stands from the entire DNR dataset, the probability of observing a sample with 17% (or more extreme values) of the 196 stands infested is 0.025. Such a random sample is possible, but not likely. Thus, we focus our simulation study on the effect that this sample might have on inference.

We simulate a mistletoe process across 10,000 stands. At each stand, we draw two covariate values from standard normal distributions. Together with an intercept covariate (equal to 1 for all stands), these are concatenated into a covariate matrix X. Probit regression coefficients, $\boldsymbol{\beta}$, corresponding to the covariates in X are specified, and the "true" mistletoe presence at each stand is simulated by:

$$y_i \sim Bern(\theta_i) \quad , \quad \Phi^{-1}(\theta_i) = \mathbf{x}_i^T \boldsymbol{\beta}.$$

Four other covariate values are drawn for each stand, again from standard normal distributions. Two of these covariates relate to the probability of true detection (ϕ_i), and the remaining two are related to the probability of false detection (ψ_i). The mistletoe status as recorded by the less accurate survey ($\mathbf{w}$) is simulated by:

$$w_i \sim \begin{cases} Bern(\Phi(\mathbf{x}_{\phi_i}^T \boldsymbol{\beta}_\phi)) & , \ y_i = 1 \\ Bern(\Phi(\mathbf{x}_{\psi_i}^T \boldsymbol{\beta}_\psi)) & , \ y_i = 0 \end{cases}$$

The values for $\boldsymbol{\beta}$, $\boldsymbol{\beta}_\phi$, and $\boldsymbol{\beta}_\psi$ were chosen to give a similar situation as we have in the example presented in this paper. That is, they were chosen in such a way that we have mistletoe present in approximately 60% of all stands in the more accurate survey, but only 11% of all stands in the less accurate survey, with "false positive" and "false negative" rates similar to those seen in our study.

To simulate a non-random sampling design, 196 stands were sampled from the 10,000 stands with the constraint that a specified proportion of the stands were labeled as infested with mistletoe in the less accurate survey. Each of these samples were then used to fit the BDR model and predict mistletoe presence across the 10,000 stands in the simulation, using only the covariates and the less accurate survey data. Keating and Cherry (2004) discussed how non-random sampling designs can lead to biased results in binary regression models, especially in the intercept terms. The table below shows the inference that the BDR approach gives for the intercept terms for the three probit regressions in the model. Rows in bold are simulations in which the true parameter value did not fall within the symmetric 95% credible interval of the posterior distribution.

Percent	Intercept for $\boldsymbol{\beta}$			Intercept for $\boldsymbol{\beta}_\phi$			Intercept for $\boldsymbol{\beta}_\psi$		
$w = 1$	β_0	95% CI		β_{ϕ_0}	95% CI		β_{ψ_0}	95% CI	
5%	0.50	0.45	1.11	**-3.00**	**-22.69**	**-4.06**	-4.00	-4.48	-1.72
10%	0.50	0.29	0.83	-3.00	-6.27	-2.38	-4.00	-19.69	-3.79
15%	0.50	0.21	0.80	-3.00	-3.40	-1.58	-4.00	-11.21	-3.09
20%	0.50	0.46	1.11	-3.00	-3.20	-1.60	-4.00	-4.22	-1.52
26%	0.50	0.28	0.91	-3.00	-3.94	-1.69	-4.00	-4.24	-1.75
31%	0.50	0.21	0.78	-3.00	-4.63	-2.09	-4.00	-9.64	-2.97
36%	0.50	0.20	0.78	-3.00	-3.35	-1.51	-4.00	-5.89	-2.25
41%	0.50	0.17	0.72	-3.00	-3.06	-1.32	-4.00	-5.14	-1.81
46%	0.50	0.19	0.72	-3.00	-4.44	-1.66	-4.00	-4.81	-1.63
51%	0.50	0.29	0.82	-3.00	-3.88	-1.61	-4.00	-5.32	-1.72
56%	0.50	0.08	0.58	**-3.00**	**-2.20**	**-0.74**	-4.00	-5.52	-1.87
61%	0.50	0.33	0.90	**-3.00**	**-2.64**	**-1.14**	-4.00	-5.77	-1.47

From these results, we can see that for extremely non-representative samples ($w = 1$ in at least 56% of the 196 stands), inference about the probit regression parameters is not trustworthy. Larger percentages of stands (greater than 61%) with $w = 1$ also resulted in biased inference on the intercepts. For situations similar to the intensive survey presented in this paper ($w = 1$ in about 17% of the 196 stands), all “true” parameter values are within the symetric 95% credible intervals of the posterior distributions and there is little or no bias in the predictions of mistletoe presence in stands not observed in the more accurate survey. Thus, the results of this simulation study provide evidence that our findings concerning mistletoe are valid.

References

[1] Keating, K.A., and S. Cherry. 2004. Use and interpretation of logistic regression in habitat-selection studies. Journal of Wildlife Management. 68(4):774-789.