Occupancy Modeling Species-Environment Relationships with Non-ignorable Survey Designs

Ecological Applications

Kathryn M. Irvine[1,4], Thomas J. Rodhouse [2], Wilson J. Wright[1], & Anthony R. Olsen[3]

[1] U.S. Geological Survey, Northern Rocky Mountain Science Center,
Bozeman, MT 59715, USA

[2] U.S. National Park Service, Upper Columbia Basin Network,
Bend, OR 97701, USA

[3] U.S. Environmental Protection Agency, Western Ecology Division,
Corvallis, OR 97333, USA

[4]E-mail: kirvine@usgs.gov

Appendix S2: ***Verification pseudo-likelihood estimation (P-MLE) appropriate for probability master sample designs***

The same 1000 simulated populations were used as described in the main text. The proposed implementation of the NABat master sample will result in a realized design that is an unequal probability design. For the proposed NABat design, the strata or subsets could be based on ownership as described for the Oregon example or specific spatial sub-domains of interest. For convenience to construct equal and unequal probability samples, we used stratified random sampling. However, with a truly unequal probability design as implemented in the mdcaty option within the grts function of spsurvey package in R (Kincaid et al. 2016), the number of sites within the specified categories is not guaranteed as with a stratified design. Strata membership for each sample unit was assigned based on dividing the mean elevation covariate into three groups of equal size that did not overlap $N_h = 887$ for $h = 1, \cdots, 3$ (i.e., a third of the sample units with smallest elevation formed one strata Fig. 1).

We explore one type of ignorable design, a self-weighting stratified design. In self-weighting stratified designs all strata had the same sample weights, $w_{i \text{ in } h} = \frac{N_h}{n_h} = c$ for all strata h where c is the sampling intensity. This equal probability design is similar to a simple random sample design with $c = N/n$. The constant c can be ignored when solving for the maximum or in our case using adjusted weights is ≈ 1 (P-MLE $=$ MLE and design is ignorable). We explored sampling intensities of 5%, 10%, 20% (which correspond to $n = 138, 278,$ or 555) with $n_h = \{46, 46, 46\}, \{93, 93, 93\}, \{185, 185, 185\}$ (Appendix S2: Table S1). We compared the equal probability design to one in which the selection of sample units was related to mean elevation within each areal unit (unequal probability design). In other words, sampling intensity varied within each of the three strata ($n_h = \{69, 46, 23\}, \{139, 93, 46\}, \{278, 185, 92\}$), but total sample size n was the same as in the equal probability designs (Appendix S2: Table S1).

As in the main paper, we fit the same four mean structures for occupancy: (1) the data generating model with both elevation and percent forest explanatory variables (denoted, Elev.+For.); (2) a model with only percent forest (denoted, For.); (3) a model with only elevation (denoted, Elev.); (4)

and an intercept only model assuming no heterogeneity in site-occupancy (denoted, Int.). We assumed constant p. When elevation was included in the mean structure for site-occupancy (models For.+Elev. or Elev.), the design becomes ignorable because selection of sample units was related to elevation for equal and unequal designs. However, if elevation was not included (models For. or Int.), the design is non-ignorable because it was not properly accounted for in the model.

Similar to the main simulation study, we compared estimates ($\hat{\beta}_{D|M_k}$) to fitting the same model but assuming a census was conducted ($n = 2660$), $\hat{\beta}_{census|M_k}$ based on maximizing Equation 1. Also, we compared the estimates for a given design and model $\hat{\beta}_{D|M_k}$ to the data generating values $\beta_{truth|M_{truth}}$. For each sampling design (Table S2) and fitted model, average 95% CIs (averages of the upper and lower bounds) and average of the point estimates ($\hat{\beta}_{D|M_k}$) were calculated across all simulated datasets as a summary. We also examined the same two different coverage properties for the census and data generating parameters.

Appendix S2: Table S1. Probability (equal and unequal) sampling designs and sampling intensities explored in simulation study. Each of 1000 populations was simulated with 8 or 4 revisits per season based on the site-occupancy model with forest cover and elevation and constant detection. Then each was sampled with the design specifications below for total sample size $n = \sum_{h=1}^{3} n_h$ and different sample sizes within a strata h, n_h.

Sampling Intensity	n		n_h		Sampling Design
5%	138	46	46	46	Equal
5%	138	69	46	23	Unequal
10%	278	93	93	93	Equal
10%	278	139	93	46	Unequal
20%	555	185	185	185	Equal
20%	555	278	185	92	Unequal

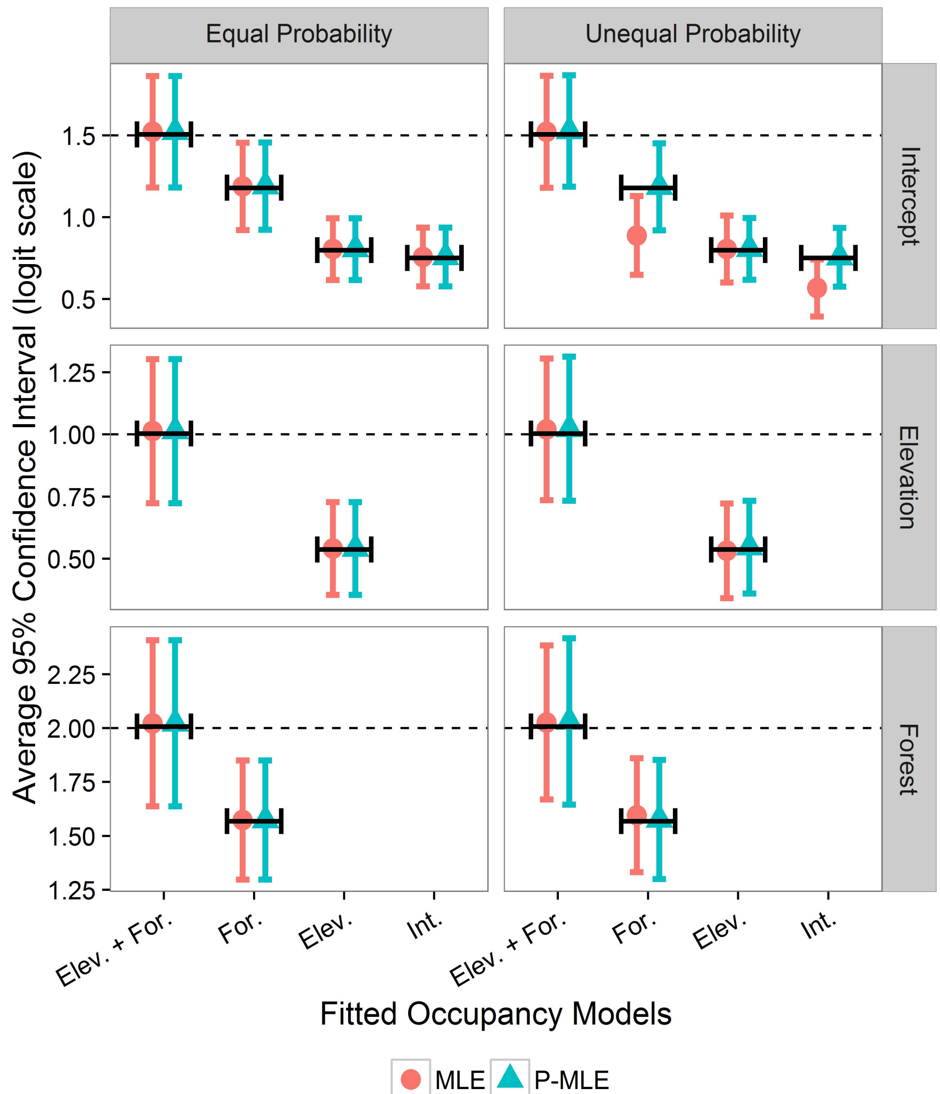
We show just the case of 20% sampling intensity with 8 revisits (Appendix S2: Fig. S1) because the other combinations displayed similar patterns. Although, as expected, decreasing sample size and number of revisits led to increased uncertainty in parameter estimates. In the case of equal probability sampling, which is an ignorable design for all fitted models, P-MLE and MLE were the same as expected because the adjusted weights $\tilde{w} \equiv 1$. Generally, unequal probability sampling produced substantially biased ML estimates for proportion of sites occupied (β_0) for models that ignored elevation, the variable used for stratification. This bias was mitigated by using P-ML estimation for the intercept β_0. However, including sample weights in estimation (P-MLE) does not alleviate model misspecification bias for non-data generating models. Coverage of the true parameter values was much lower than the desired 95%. P-MLEs only helped adjust for design-based bias arising from observing a sample of the population and not censusing every sample unit (e.g., all 2660 grid cells in Oregon).

These simulations support the use of comparing P-MLE and MLE confidence intervals as a way to diagnose a non-ignorable design for a given fitted model (as motivated in Appendix S1). P-MLE confidence intervals were similar to MLE confidence intervals when fitting For.+Elev. or Elev. model with unequal probability of site selection (ignorable designs) or all models with equal probability of selection. However, the intervals differed for unequal probability sampling and fitting For. or Int. because these models ignore that the sites were selected based on elevation (non-ignorable designs for these models). These results suggest that for a probability master sample with definable strata or subsets the P-MLE approach could be used for design unbiased inferences. Alternatively, a simpler approach could be to include a random effect or fixed effect for each unique subset; however, all the criticisms pointed out in the introduction by Pfeffermann (2007) should be considered and this assumes that the unique subsets can be defined prior to a combined analysis.

Literature Cited

Kincaid, T., T. R. Olsen, D. Stevens, C. Platt, D. White, & R. Remington. Aug. 19, 2016. *Spatial Survey Design and Analysis*. Version 3.3. URL: `https://cran.r-project.org/web/packages/spsurvey/index.html`.

Pfeffermann, D. 2007. Comment: Struggles with survey weighting and regression modeling. Statistical Science 22:179–183.



Appendix S2: Fig. S1 Equal probability and unequal probability sampling design impacts on site-occupancy parameters for four different sets of occupancy covariates. Fitted models varied site-level covariate structures: data generating model with elevation and percent forest (Elev. + For); Forest cover only (For.); elevation only (Elev.); and constant occupancy (Int.). The equal probability sample was a self-weighting stratified design. The unequal probability designs were created by varying sampling intensity among strata defined along the elevational gradient in Oregon (Fig. 1 gray scale). Occupancy estimates based on MLE or P-MLE with a total sample size equal to 555 with 8 revisits. The dashed lines display data generating values and line segments show model estimates assuming a census was taken of all 2660 Oregon sample units.