SPATIAL MODELS FOR DISTANCE SAMPLING DATA:
RECENT DEVELOPMENTS AND FUTURE DIRECTIONS

APPENDIX B: CALCULATION OF VARIANCE IN DENSITY SURFACE MODELS

DAVID L. MILLER, M. LOUISE BURT, ERIC A. REXSTAD AND LEN THOMAS

1. INTRODUCTION

This appendix gives a brief mathematical explanation of the method proposed in Williams et al. (2011) for the propagation of uncertainty from the detection function to the DSM, as well as how to calculate the variance of a non-linear function of a GAM (e.g. when calculating the variance of the predicted abundance).

2. VARIANCE PROPAGATION

The formulation for a "count method" density surface model (DSM) is:

$$\mathbb{E}(n_j) = \exp\left[\log\left(p_j(\boldsymbol{\theta})A_j\right) + \sum_{k=1}^{K} f_k(z_{jk})\right],$$

where we model the expected number of animals per segment (n_j). The f_ks are smooth functions of the covariates and β_0 is an intercept term. A_j is the covered area and the probability of detection is given by ($\hat{p}_j$) and is estimated from the detection function.

Writing $\hat{p}_j$ explicitly as a function of the estimated detection function parameters $\hat{\boldsymbol{\theta}}$ and loging both sides yields:

$$\begin{aligned}\log\left[\mathbb{E}(n_j)\right] &= \log\left[p_j(\hat{\boldsymbol{\theta}})A_j\right] + \sum_{k=1}^{K} f_k(z_{jk}),\\ &= \log\left(A_j\right) + \log\left[p_j(\hat{\boldsymbol{\theta}})\right] + \sum_{k=1}^{K} f_k(z_{jk}).\end{aligned}$$

At this point we add another term to the model. This new term is the derivative of $\log\left[\hat{p}(\hat{\boldsymbol{\theta}})\right]$ multiplied by $\gamma = (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$.

$$\log\left[\mathbb{E}(n_j)\right] = \log\left(A_j\right) + \log\left[p_j(\hat{\boldsymbol{\theta}})\right] + \frac{d\log p(\boldsymbol{\theta})}{d\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\gamma + \sum_{k=1}^{K} f_k(z_{jk}). \tag{1}$$

This term has basically no effect on the model, since, using the definition of a finite difference:

$$\frac{d\log p(\boldsymbol{\theta})}{d\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \left\{\log\left[p(\hat{\boldsymbol{\theta}}+\delta)\right] - \log\left[p(\hat{\boldsymbol{\theta}})\right]\right\}\delta^{-1},$$

(if we assume that γ is small enough such that $\gamma \approx \delta$).

We may then write (1) as:

$$\log\left[\mathbb{E}(n_j)\right] = \log\left(A_j\right) + \log\left[p_j(\hat{\boldsymbol{\theta}})\right] + \frac{d\log p(\boldsymbol{\theta})}{d\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\gamma + \sum_{k=1}^{K} f_k(z_{jk}),$$

$$= \log\left(A_j\right) + \log\left[p_j(\hat{\boldsymbol{\theta}})\right] + \left\{\log\left[p(\hat{\boldsymbol{\theta}}+\delta)\right] - \log\left[p(\hat{\boldsymbol{\theta}})\right]\right\}\delta^{-1}\gamma + \sum_{k=1}^{K} f_k(z_{jk}).$$

Assuming that $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ are "close" we can say that $\gamma \approx \delta$, so:

$$\log\left[\mathbb{E}(n_j)\right] \approx \log\left(A_j\right) + \log\left[p_j(\hat{\boldsymbol{\theta}})\right] + \log\left[p(\hat{\boldsymbol{\theta}} + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}))\right] - \log\left[p(\hat{\boldsymbol{\theta}})\right] + \sum_{k=1}^{K} f_k(z_{jk})$$

$$\approx \log\left(A_j\right) + \log\left[p_j(\hat{\boldsymbol{\theta}})\right] + \log\left[p(\boldsymbol{\theta})\right] - \log\left[p(\hat{\boldsymbol{\theta}})\right] + \sum_{k=1}^{K} f_k(z_{jk})$$

$$\approx \log\left(A_j\right) + + \log\left[p(\boldsymbol{\theta})\right] + \sum_{k=1}^{K} f_k(z_{jk}).$$

So this extra term does not have a large effect on the resulting GAM. It does however have an effect on the variance estimates derived from the model. In practice, we can look at the difference between the model coefficients in the model with an without the extra term to check that there has been no large change in the model.

3. Calculating the variance of the abundance

To find the variance of the predicted abundance we are finding the variance of a function of the linear predictor (in the case of the abundance, this is simply the sum). We begin by revising some basic GAM theory before moving on to the specific case of DSMs.

When the identity link is used, finding the variance of some function of the model is relatively easy. The `lpmatrix` (Wood, 2006, page 245) is used, that is the matrix $\boldsymbol{X}_p$ such that:

$$\hat{\boldsymbol{\eta}}_p = \boldsymbol{X}_p\hat{\boldsymbol{\beta}}$$

i.e. $\boldsymbol{X}_p$ maps the model parameters ($\hat{\boldsymbol{\beta}}$) to the linear predictor ($\hat{\boldsymbol{\eta}}_p$). We can then use $\boldsymbol{X}_p$ to find the covariance matrix for the linear predictor if we can estimate the parameter covariance matrix ($\boldsymbol{V}_\beta$):

$$\boldsymbol{V}_{\hat{\eta}_p} = X_p\boldsymbol{V}_{\hat{\beta}}X_p^{\mathrm{T}}.$$

Only linear functions of the linear predictor can be calculated using this method but this just consists of changing the pre- and post-multiplying matrices. When the link function is not the identity calculations are not so straightforward, we now illustrate two ways of obtaining variance estimates when using a non-identity link function.

3.1. Calculation by simulation.

First note that the distribution of the parameters (given the data) is multivariate normal with mean as the parameter estimates and the covariance matrix of the parameters. (i.e. $\boldsymbol{\beta} \sim N(\hat{\boldsymbol{\beta}}, \boldsymbol{V}_{\hat{\boldsymbol{\beta}}})$).

The following algorithm is suggested by (Wood, 2006, page 246):

(1) For $b = 1, \ldots, N_b$ do the following:

(a) Simulate from $\boldsymbol{\beta} \sim N(\hat{\boldsymbol{\beta}}, \boldsymbol{V}_{\hat{\boldsymbol{\beta}}})$, to obtain $\boldsymbol{\beta}_b$.
(b) Calculate $\hat{\eta}_b = \exp(\boldsymbol{X}_p \boldsymbol{\beta}_b)$ (e.g. if we are using the log-link)
(c) Sum over the survey area

(2) Calculate the appropriate summary statistics, e.g. median, 95% quantiles etc over b.

In practice N_b does not have to be particularly large, Marra et al. (2011) achieve good results with $N_b = 100$.

3.2. **Calculation by the delta method.** Simulation may well be unnecessary and it may well be easier and more efficient to use a sandwich estimator:

$$\left(\frac{\partial \log_e \eta}{\partial \eta}\Big|_{\eta=\hat{\eta}} \circledast \boldsymbol{X}_p \right) \boldsymbol{V}_p \left(\frac{\partial \log_e \eta}{\partial \eta}\Big|_{\eta=\hat{\eta}} \circledast \boldsymbol{X}_p \right)^{\mathsf{T}},$$

where $\frac{\partial \log_e \eta}{\partial \eta}\Big|_{\eta=\hat{\eta}}$ is the vector of first derivatives of the link evaluated at the values of the linear predictor and $\circledast$ denotes R-style matrix-vector multiplication. A sandwich estimator inflates the variance based on the uncertainty in the linear predictor.

REFERENCES

Marra, G., D. L. Miller, and L. Zanin (2011, November). Modelling the spatiotemporal distribution of the incidence of resident foreign population. *Statistica Neerlandica 66*(2), 133–160.

Williams, R., S. L. Hedley, T. A. Branch, M. V. Bravington, A. N. Zerbini, and K. P. Findlay (2011). Chilean blue whales as a case study to illustrate methods to estimate abundance and evaluate conservation status of rare species. *Conservation Biology 25*(3), 526–535.

Wood, S. N. (2006). *Generalized Additive Models: An introduction with R* . Chapman & Hall/CRC.