Avoiding extrapolation bias when using statistical models to make ecological prediction

Paul B. Conn[1*], Devin S. Johnson[1], and Peter L. Boveng[1]

[1]*National Marine Mammal Laboratory, Alaska Fisheries Science Center, NOAA National Marine Fisheries Service, Seattle, Washington 98115 U.S.A.*

Appendix S1: Model formulation and Gibbs sampling algorithms for certain classes and extensions of the generalized linear model

We write all statistical models for ecological prediction in the form

$$Y_i \sim f_Y(g^{-1}(\mu_i)), \tag{1}$$

where $\mu_i = \theta_i + \epsilon_i$, f_Y denotes a probability density or mass function (e.g. Bernoulli, Poisson), g gives a link function, θ_i is a linear predictor, and $\epsilon_i \sim \text{Normal}(0, \tau_\epsilon)$ is iid Gaussian error with precision parameter τ_ϵ. The specification in Eq. 1 is thus doubly stochastic, in the sense that we assume error associated with f_Y as well as in the location parameter μ_i. This setup can be useful computationally, and can also be used to approximate singly stochastic system by setting τ_ϵ to a large value. For instance, models for count data often assume a Poisson error structure with a log link function to ensure that the Poisson intensity parameter is greater than zero. In this case, we would specify

$$Y_i \sim \text{Poisson}(\exp(\theta_i + \epsilon_i)),$$

a configuration known as a log-Gaussian Cox process. We now describe how different classes of statistical models can be developed depending on how one structures the linear predictor, θ_i.

1 Models

1.1 Generalized linear models (GLM)

Generalized linear models (McCullagh and Nelder 1989) are one of the simplest (and most often used) statistical models used by ecologists to make spatial predictions. In GLMs, the linear predictor is a simple linear function of gathered covariates (including possible quadratic terms of these covariates. Statisticians often describe this relationship by $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$, which is written in matrix notation. In particular, $\boldsymbol{\theta}$ gives a vector of linear predictor values (one for each data point being analyzed), $\boldsymbol{\beta}$ gives a vector of regression coefficients, and $\mathbf{X}$ is a design matrix which includes all explanatory variables (and often a column vector of ones to represent an intercept).

1.2 Generalized additive models (GAM)

Although one can allow nonlinear relationships between response variables and regressors by including polynomial terms in the design matrix of a GLM, these need to be pre-selected by the analyst and it is often unclear how many such terms one should include. Generalized additive models (GAMs; Hastie and Tibshirani 1999, Wood 2006) build upon generalized linear models, but instead allow smooth relationships between the dependent and independent variables using flexible functions such as splines. Such models have been employed in a number of spatial prediction scenarios, including transect sampling models for animal abundance (Hedley and Buckland 2004) and SDMs (Guisan et al. 2002). For instance, animal density or presence can be modeled as a smooth, unknown function of a habitat covariate.

There are a number of ways smooth relationships can be modeled; in order to formulate a GAM in the notation of Eq. 1 we employ a knot-based kernel smoother with a radial basis function. The basic notion is to place k_j knots, which we will denote by $\boldsymbol{\omega}_j = \{\omega_{1j}, \omega_{2j}, \ldots, \omega_{k_j j}\}$, throughout the range of a given regressor j, where the number of knots controls the level of smoothing. For each regressor of interest, we can then calculate a matrix $\mathbf{K}_j$ of dimension (n, k_j), with entries $\mathcal{N}(X_{ij}; \omega_j, \tau_{\omega,j})$, where $\mathcal{N}(x; \mu, \tau)$ denotes a Gaussian distribution with mean μ and precision τ. We can then introduce additional regression coefficients $\boldsymbol{\alpha}_j$ to help model the smooth

relationship, and incorporate these into the linear predictor. In the case of one smooth term, we then have in Eq. 1:

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{K}_j\boldsymbol{\alpha}_j \tag{2}$$

When fitting such a model, one needs to select the number and location of knots, as well as the precision parameter $\tau_{\omega,j}$ for the Gaussian basis kernels. This could be done using some notion of optimality or prediction error (as with cross validation). In practice, a simple recommendation is to set $\tau_{\omega,j}$ less than the range of the modeled covariate, and greater than the typical difference between observed values of the covariate.

Our main point in writing the GAM as in Eq. 2 is to emphasize the structural similarity between GLMs and certain classes of GAMs. In particular, we can rewrite Eq. 2 as $\boldsymbol{\theta} = \mathbf{X}_{aug}\boldsymbol{\beta}_{aug}$, where $\mathbf{X}_{aug} = [\mathbf{X} \quad \mathbf{K}_j]$ (i.e., concatenating $\mathbf{X}$ and $\mathbf{K}_j$ horizontally), and $\boldsymbol{\beta}_{aug} = [\boldsymbol{\beta} \quad \boldsymbol{\alpha}_j]'$ (the subscript *aug* denotes augmentation). We shall exploit this structure when examining extrapolation bias.

1.3 Introducing spatial and/or temporal autocorrelation: spatio-temporal regression models (STRMs)

The previous two modeling frameworks (GLMs and GAMs) do not acknowledge spatial autocorrelation above and beyond that induced by modeled covariates. However, it is common for residuals from GLM and GAM model fits to include spatial autocorrelation, which violates their common assumption of independently distributed error (Legendre 1993, Lichstein et al. 2002). Parameter estimates from GLMs and GAMs that display residual autocorrelation should be interpreted with caution, as they will tend to have overstated precision and may even be biased. In such situations, analysts often employ spatial regression models which explicitly account for spatial autocorrelation above and beyond that explained by modeled covariates.

There are a variety of ways spatial autocorrelation can be included in regression models, depending on (i) the spatial support (i.e., continuous vs. discrete), and (ii) the particular mechanism used to impart correlation. Here, we shall focus on areal models for discrete spatial support, as our impression is that these are more commonly employed in ecological studies. Such

models require that data are aggregated at the level of some sample unit (plots, grid cells, etc.). Spatio-temporal regression models (STRMs) for areal data are often specified in a similar fashion to GLMs and GAMs:

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta}, \tag{3}$$

where $\boldsymbol{\eta}$ represent spatially autocorrelated effects. In this treatment we shall consider two approaches for inducing spatial autocorrelation in $\boldsymbol{\eta}$: process convolution (PC; Higdon 1998) and restricted spatial regression (RSR; Reich et al. 2006, Hodges and Reich 2010, Hughes and Haran 2013).

The PC implementation works by placing k knots throughout the spatial domain being analyzed, and like our GAM formulation, uses distances from the center point of each sample unit to each of these knot locations to induce spatial structure. Using similar notation to our GAM formulation, we denote these knots as $\boldsymbol{\omega} = \{\boldsymbol{\omega_1}, \boldsymbol{\omega}_2, \ldots, \boldsymbol{\omega}_k\}$; the $\boldsymbol{\omega}_m$ are bivariate in this instance as space is assumed to be two dimensional. We then construct a matrix $\mathbf{K}$ of dimension (n, k) with entries $\mathcal{BVN}(\mathbf{s}_i; \boldsymbol{\omega}_m, \tau_\omega)$, where $\mathcal{BVN}(\mathbf{x}; \mu, \tau)$ denotes a bivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and precision τ. As with our GAM model we introduce additional regression coefficients $\boldsymbol{\alpha}$ to estimate the level of spatial smoothing conditional on the assumed knot structure and reparameterize the STRM as

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{K}\boldsymbol{\alpha} \tag{4}$$

This model is structurally very similar to the GAM model (indeed it may be interpreted as a GAM with a bivariate smooth on spatial location). Similarly, we can rearrange terms to write it similar to the GLM, such that $\boldsymbol{\theta} = \mathbf{X}_{aug}\boldsymbol{\beta}_{aug}$, where in this case $\mathbf{X}_{aug} = [\mathbf{X} \quad \mathbf{K}]$, and $\boldsymbol{\beta}_{aug} = [\boldsymbol{\beta} \quad \boldsymbol{\alpha}]'$.

The RSR approach to spatial regression uses a reduced-rank version of the popular intrinsic conditionally autoregressive (ICAR; Besag and Kooperberg 1995, **?**) model for spatial random effects, reparameterized so that basis vectors are orthogonal to the main effects of interest. This approach has generated substantial recent interest, as fixed effects retain primacy in explaining

variation in the ecological process of interest and problems with spatial confounding between fixed and random effects are eliminated. As such, spatial random effects are only used to account for residual autocorrelation (Reich et al. 2006, Hodges and Reich 2010) and the decision to incorporate spatial autocorrelation has little effect on the point estimates of fixed effects. In addition, reduced dimension spatial models such as RSR lighten computational burden while still accounting for course-scale spatial autocorrelation (see e.g. Latimer et al. 2009, Wikle 2010, Hughes and Haran 2013). It turns out that RSR models can also be written in the form of Eq. 4, using a different choice of $\mathbf{K}$ (Hughes and Haran 2013), constructed as follows:

1. Define the residual projection matrix $\mathbf{P}^{\perp} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$
2. Calculate the Moran operator matrix $\mathbf{\Omega} = J\mathbf{P}^{\perp}\mathcal{W}\mathbf{P}^{\perp}/\mathbf{1}'\mathcal{W}\mathbf{1}$
3. Define $\mathbf{K}$ as an $(n \times m)$ matrix, where the columns of $\mathbf{K}$ are composed of the eigenvectors associated with the largest m eigenvalues of $\mathbf{\Omega}$. Hughes and Haran (2013) used simulation to explore a range of such values and concluded $m = 50 - 100$ should suffice for most applications.

Here, $\mathbf{I}$ an identity matrix, $\mathbf{1}$ is a column vector of ones, and $\mathcal{W}$ represents an association matrix describing the spatial neighborhood structure of sampling units. For instance, for a first order neighborhood structure, $\mathcal{W}$ would include a 1 for all rows i and columns j where sampling unit i and j are neighbors (see **?**, for alternative association matrices).

Literature Cited

Besag, J., and C. Kooperberg. 1995. On conditional and intrinsic autoregressions. Biometrika **82**:733–746.

Guisan, A., T. C. Edwards Jr., and T. Hastie. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling **157**:89–100.

Hastie, T. J., and R. J. Tibshirani. 1999. Generalized Additive Models. Chapman & Hall/CRC, Boca Raton, Florida.

Hedley, S., and S. Buckland. 2004. Spatial models for line transect sampling. Journal of Agricultural, Biological, and Environmental Statistics **9**:181–199.

Higdon, D. 1998. A process-convolution approach to modelling temperatures in the North Atlantic Ocean. Environmental and Ecological Statistics **5**:173–190.

Hodges, J., and B. Reich. 2010. Adding spatially-correlated errors can mess up the fixed effects you love. American Statistician **64**:325–334.

Hughes, J., and M. Haran. 2013. Dimension reduction and alleviation of confounding for spatial generalized mixed models. Journal of the Royal Statistical Society B **75**:139–159.

Latimer, A. M., S. Banerjee, H. Sang, E. S. Moshner, and J. A. Silander Jr. 2009. Hierarichical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northern United States. Ecology Letters **12**:144–154.

Legendre, P. 1993. Spatial autocorrelation: trouble or new paradigm? Ecology **74**:1659–1673.

Lichstein, J., T. Simons, S. Shriner, and K. E. Franzreb. 2002. Spatial autocorrelation and autoregressive models in ecology. Ecological Monographs **72**:445–463.

McCullagh, P., and J. A. Nelder. 1989. Generalized Linear Models. Chapman and Hall, New York.

Reich, B., J. Hodges, and V. Zadnik. 2006. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. Biometrics **62**:1197–1206.

Wikle, C. K., 2010. Low rank representations for spatial processes. Pages 89–106 *in* A. Gelfand, P. Diggle, M. Fuentes, and P. Guttorp, editors. Handbook of Spatial Statistics. Chapman & Hall.

Wood, S. N. 2006. Generalized additive models. Chapman & Hall/CRC, Boca Raton, Florida.