Marginal Inference for Hierarchical Generalized Linear Mixed Models with Patterned Covariance Matrices using the Laplace Approximation

Jay M. Ver Hoef[1], Eryn Blagg[2], Michael Dumelle[3], Philip Dixon[2], Dale Zimmerman[4], Paul Conn[1]

---

[1]National Marine Mammal Laboratory
NOAA-NMFS Alaska Fisheries Science Center
7600 Sand Point Way NE, Seattle, WA 98115
E-mail: jay.verhoef@noaa.gov

---

[2] Department of Statistics, Iowa State University, Ames, Iowa
[3] United States Environmental Protection Agency,
200 SW 35th St, Corvallis, Oregon
[4] Department of Statistics and Actuarial Science,
The University of Iowa, Iowa City, Iowa

March 2, 2023

Abstract

We take a fully parametric approach to creating covariance dependence for generalized linear mixed models through a hierarchical construction. We estimate covariance parameters and fixed effects marginally while integrating out over all latent random effects using the Laplace approximation. We use Newton-Raphson updates, which also leads to predictions for latent random effects. We provide complete marginal inference, from estimating covariance parameters and fixed effects, to making predictions for unobserved data, for any patterned covariance matrix in the hierarchical generalized linear mixed models framework. We show how the marginal likelihood can be developed for six commonly used distributions that are often used for binary, count, and positive continuous data. The methods are illustrated with simulations from known parameters, and their efficacy is shown through simulation experiments. Three examples are used to illustrate all six distributions with a variety of patterned covariance structures that include spatial models, time series models, and mixtures with typical random intercepts based on grouping.

1 INTRODUCTION

The classical linear model relies on a normal distribution that has continuous support on the real line, but many data are binary, counts, or positive continuous. Such data can be transformed to stabilize variances and create empirical distributions that are "near normal," allowing the use of classical linear models (e.g., Snedecor and Cochran, 1980, p. 288). For example, a square root transformation can be used for count data. However, Nelder and Wedderburn (1972) introduced the generalized linear model (GLM, McCullagh and Nelder, 1989) as a natural extension to linear models, such as the Poisson distribution for counts, the Bernoulli distribution for binary data, etc., which have become very popular and generally preferred to data transformations (e.g., Warton and Hui, 2011). GLMs can be extended by introducing latent random effects as a linear mixed model to create a class of generalized linear mixed models (GLMM, Breslow and Clayton, 1993). These latent random effects are usually assumed to be independent and identically distributed from a normal distribution (Zeger and Karim, 1991), however it is also possible for the latent random effects to be temporally autocorrelated (e.g., Stiratelli et al., 1984; Zeger et al., 1988), spatially autocorrelated (e.g, Clayton and Kaldor, 1987; Gotway and Stroup, 1997; Diggle et al., 1998), or both (Cressie and Wikle, 2011, p. 380). A unifying framework for this literature is through a hierarchical generalized linear mixed model (HGLMM, Lee and Nelder, 1996).

1.1 Hierarchical Generalized Linear Mixed Models

Most GLMs are motivated by the exponential family of distributions (e.g., Fisher, 1934; Lehmann and Casella, 2006). However, GLMs are ultimately modeled with quasi-likelihood through the first two moments, the variance and mean functions, and fitted through iteratively-weighted least-squares estimation (Wedderburn, 1974), which creates a very flexible class of models. In fact, some models have no true likelihood, such as the quasi-Poisson model.

We will take a fully parametric approach to create covariance dependence for GLMMs through a hierarchical construction. We will use the notation $[\mathbf{y}|\boldsymbol{\mu}]$ to denote any probability density function of the vector of random variables $\mathbf{y}$ conditional on a vector of parameters, or other fixed variables, $\boldsymbol{\mu}$. We can have a joint distribution on the left side of the conditional bar, and multiple parameter and fixed value vectors on the right, e.g., $[\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k|\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_\ell]$. For example, let $[\mathbf{y}|\boldsymbol{\mu}]$ be the product of independent Poisson distributions with mean parameters contained in the vector $\boldsymbol{\mu}$, so $\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu}$. The model for the data $\mathbf{y}$ can have more parameters than just the mean, in which case we write it $[\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\phi}]$. For an example with extra parameters for $\mathbf{y}$, consider the product of independent negative binomial distributions, which can be parameterized with a mean vector, and a common extra parameter that allows for overdispersion, which we would write as $[\mathbf{y}|\boldsymbol{\mu}, \phi]$, where ϕ is the overdispersion parameter.

For the hierarchical construction of the generalized linear mixed models that we consider in this article, we allow the mean to vary by other random variables $\mathbf{w}$, and we condition on these random variables, $[\mathbf{y}|g^{-1}(\mathbf{w}), \boldsymbol{\phi}]$, through the mean function $\mathrm{E}(\mathbf{y}) = g^{-1}(\mathbf{w})$. For the Poisson example, $g(\cdot)$ is often the log function, and in general $g(\cdot)$ is called the link function (McCullagh and Nelder, 1989). Link functions are monotonic so that $g^{-1}(\cdot)$ is one-to-one with $g(\cdot)$. Recall that the mean of a Poisson distribution must be positive, and if $g(\cdot)$ is the log function, then $g^{-1}(\cdot)$ is the exponential function so $\boldsymbol{\mu} = g^{-1}(\mathbf{w})$ is always positive, which allows $\mathbf{w}$ to be unconstrained on the real line.

We will only consider models where $\mathbf{w}$ is $n \times 1$ and has a multivariate normal distribution which is constructed through the linear mixed model,

$$\mathbf{w} = \mathbf{X}\boldsymbol{\beta} + \sum_{k=1}^{q} \mathbf{Z}_k \mathbf{r}_k + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{X}$ is an $n \times p$ full rank design matrix of explanatory variables, $\boldsymbol{\beta}$ is a $p \times 1$ parameter

vector of fixed effects, $\mathbf{Z}_k$ is a design matrix for the kth random effect $\mathbf{r}_k$, and $\boldsymbol{\epsilon}$ is independent error. We assume that $\mathrm{E}(\mathbf{r}_k) = \mathbf{0}$ for all k, $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$, $\mathrm{var}(\mathbf{r}_k) = \mathbf{V}_k$, $\mathrm{cov}(\mathbf{r}_j, \mathbf{r}_k) = \mathbf{0}$ when $j \neq k$, and $\mathrm{var}(\boldsymbol{\epsilon}) = \sigma_0^2\mathbf{I}$. We use the notation $[\mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \{\mathbf{Z}_k\}, \{\mathbf{V}_k\}, \boldsymbol{\theta}]$ to indicate the probability density function $\mathbf{w} \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$ where

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \sum_k \mathbf{Z}_k \mathbf{V}_k \mathbf{Z}_k' + \sigma_0^2 \mathbf{I}.$$

The covariance matrices $\{\mathbf{V}_k\}$ for $k = 1, \ldots, q$ can have additional covariance parameters beyond σ_0^2, all of which are contained in the vector $\boldsymbol{\theta}$. We will give more specific details on $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ later.

For the fully parametric, hierarchical models, a very general model can be constructed hierarchically as,

$$[\mathbf{y}, \mathbf{w}|\boldsymbol{\phi}, \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}] = [\mathbf{y}|g^{-1}(\mathbf{w}), \boldsymbol{\phi}][\mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}], \tag{2}$$

where Berliner (1996) called $[\mathbf{y}|g^{-1}(\mathbf{w}), \boldsymbol{\phi}]$ the *data model* and $[\mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}]$ the *process model*. As a concrete example, suppose that $[\mathbf{y}|\exp(\mathbf{w})]$ is Poisson, and $[\mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}]$ is multivariate normal, then the joint likelihood is

$$[\mathbf{y}, \mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}] = \left(\prod_{i=1}^{n} \frac{\exp(w_i)^{y_i} \exp(-\exp(w_i))}{y_i!} \right) \frac{\exp\left(-(\mathbf{w} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} (\mathbf{w} - \mathbf{X}\boldsymbol{\beta})\right)}{(2\pi)^{n/2} |\boldsymbol{\Sigma}_{\boldsymbol{\theta}}|^{1/2}},$$

and note the use of $\exp(w_i)$ for $\mathrm{E}(y_i)$.

1.2 Patterned Covariance Matrices

To construct a likelihood for (2) we will need parametric models for $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ in (1). There are few constraints here, and any valid covariance model for $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is possible. For example, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ may be constructed from typical mixed models where $\mathbf{Z}_k$ contains indicator variables for random intercepts, or explanatory variables for random slopes, and where $\mathbf{V}_k = \sigma_k^2 \mathbf{I}$, and

then $\mathbf{\Sigma}_{\boldsymbol{\theta}} = \sum_k \sigma_k^2 \mathbf{Z}_k \mathbf{Z}_k' + \sigma_0^2 \mathbf{I}$. We can also consider time series models (e.g., Hamilton, 1994). For example, for a first-order autoregressive (AR1) model with i and j being integers, let $q = 1$ and $\mathbf{Z}_1 = \mathbf{I}$, then $\mathbf{V}_1[i, j]$ has as its i, jth entry $\sigma_1^2 \rho^{|i-j|}/(1-\rho^2)$, where $0 < \sigma_1^2$ and $0 \leq \rho < 1$. Similarly, we can have geostatistical models (Chiles and Delfiner, 1999, e.g.,), such as the exponential autocovariance model, where $q = 1$, $\mathbf{Z}_1 = \mathbf{I}$, and the i, jth element of $\mathbf{V}_1[i, j]$ is $\sigma_1^2 \exp(-\delta_{i,j}/\rho)$ where $\delta_{i,j}$ is Euclidean distance between the ith and jth locations, $0 < \sigma_1^2$, and $0 < \rho$. Other spatial covariance types include the conditional autoregressive (CAR, Besag, 1974; Cressie, 1993) and simultaneous autoregressive models (SAR, Whittle, 1954; Ver Hoef et al., 2018), moving average models in time series (e.g., Hamilton, 1994) and spatial statistics (Haining, 1978), spatio-temporal models (Cressie and Wikle, 2011, p. 380), and models on non-Euclidean topologies such as a sphere (e.g., the earth, Huang et al., 2011; Gneiting, 2013) and networks such as roads (Ver Hoef, 2018) and streams (Ver Hoef and Peterson, 2010). Because a covariance matrix can be constructed by summing covariance matrices as variance components, mixtures of all models mentioned above can create a rich set of patterned covariance matrices for modeling dependent structures. In what follows, we develop inference based on any valid covariance matrix.

1.3 Inference for HGLMMs

The combination of the data model, $[\mathbf{y}|g^{-1}(\mathbf{w}), \boldsymbol{\phi}]$, where any distribution could be used that matches the type of data, and the process model, $[\mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \mathbf{\Sigma}_{\boldsymbol{\theta}}]$, that can allow for any patterned covariance matrix, provides a hierarchical construction (2) that is a very rich and flexible class of models. This class of models is not new.

There are two broad methods of analysis. The most obvious method is to take a Bayesian approach and compute the posterior distribution of all latent variables and parameters. Due to intractable integrals, this is usually achieved with Markov chain Monte Carlo

(MCMC) methods (Gelfand and Smith, 1990; Gilks et al., 1996), of which there are now many varieties. Bayesian hierarchical models in our context have been extremely popular, beginning with spatial statistics (e.g. Clayton and Kaldor, 1987), clustered data (e.g. Zeger and Karim, 1991), time series (e.g. Berliner, 1996), and longitudinal data (Kleinman and Ibrahim, 1998) among others, and with the introduction of the `WinBUGS` software (Lunn et al., 2000).

A second approach attempts to estimate covariance parameters and fixed effects marginally while integrating out over all latent random effects. This can also be done using MCMC methods as a numerical integrator (e.g., Zhang, 2002; Christensen, 2004) but a more popular and deterministic method uses a Laplace approximation (Tierney and Kadane, 1986). In particular, Rue et al. (2009) proposed integrated nested Laplace approximation (INLA) as approximate Bayesian inference when using Gaussian Markov random fields. Our development builds primarily on Evangelou et al. (2011) and Bonat and Ribeiro Jr (2016). We will point out differences from our development and that of Bonat and Ribeiro Jr (2016) in our Methods section.

Despite the relatively long history of this subject, there is no unified framework for the case where covariance matrices are patterned by random effects, spatial and/or temporal models. Our general goal is to provide complete marginal inference, from estimating covariance parameters and fixed effects, to making predictions for unobserved data, for any patterned covariance matrix in the HGLMM framework. In particular, our goals are to: 1) find marginal maximum likelihood and restricted maximum likelihood estimates for covariance parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, 2) predict the latent values of $\mathbf{w}$, 3) estimate fixed effects $\boldsymbol{\beta}$, and 4) make predictions of new values of the process that generated $\mathbf{w}$ at unsampled times or places.

The rest of this paper is organized as follows. In Section 2, we use the Laplace approximation to develop marginal maximum likelihood estimates for $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ using Newton-

Raphson updates, which also leads to predictions of $\mathbf{w}$. From the predictions of $\mathbf{w}$ we develop estimators of $\boldsymbol{\beta}$ with proper confidence intervals and prediction of new values of the process generating $\mathbf{w}$ with proper prediction intervals. In Section 3, we conduct simulations to illustrate all methods and validate the earlier development. Section 4 provides three separate examples to further illustrate the methods. We conclude with some discussion in Section 5.

2 Methods

When considering the hierarchical model formulation of the HGLMMs, we would like to marginalize the distribution $[\mathbf{w}, \mathbf{y}|\boldsymbol{\phi}, \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}] = [\mathbf{y}|g^{-1}(\mathbf{w}), \boldsymbol{\phi}][\mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}]$ over $\mathbf{w}$ and be free of $\boldsymbol{\beta}$ as well to obtain a distribution of the only the data and variance/covariance parameters. The Laplace method helps achieve that.

2.1 Laplace for HGLMMs

First, consider integrating over $\boldsymbol{\beta}$ as well as $\mathbf{w}$,

$$[\mathbf{y}|\boldsymbol{\phi}, \boldsymbol{\theta}] = \int_{\mathbf{w}\in\mathbb{R}^n} \int_{\boldsymbol{\beta}\in\mathbb{R}^p} [\mathbf{w}, \mathbf{y}|\boldsymbol{\phi}, \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}] d\boldsymbol{\beta} d\mathbf{w} = \int_{\mathbf{w}} [\mathbf{y}|g^{-1}(\mathbf{w}), \boldsymbol{\phi}] \int_{\boldsymbol{\beta}} [\mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}] d\boldsymbol{\beta} d\mathbf{w}.$$

When $[\mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}]$ is Gaussian, $\int_{\boldsymbol{\beta}} [\mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}] d\boldsymbol{\beta}$ is the likelihood for restricted maximum likelihood estimation (REML, see Appendix). Note that REML was originally derived as a set of $n - p$ independent linear combinations of the observations known as error contrasts (Patterson and Thompson, 1971, 1974), and there is little literature on its derivation from integration. Alternatively, consider $[\mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}]$ where $\boldsymbol{\beta}$ has been replaced by its conditional (on $\mathbf{w}$) maximum likelihood estimator, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{w}$. Then, both cases are free of $\boldsymbol{\beta}$,

$$[\mathbf{w}|\mathbf{X}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}] = \frac{1}{C_n} \exp[(\mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}})],$$

where for ML estimation $C_n = \sqrt{2\pi^{n/2}|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}|}$ and for REML estimation $C_n = \sqrt{2\pi^{(n-p)/2}|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}||\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X}|}$. Note that Bonat and Ribeiro Jr (2016) only considered the marginal likelihood integrated over $\mathbf{w}$, and did not consider the likelihood where $\boldsymbol{\beta}$ was also integrated out (as in REML estimation) or back-substituted (as in ML estimation).

To obtain the marginal distribution of the data and covariance parameters we need the integral,

$$[\mathbf{y}|\boldsymbol{\phi}, \mathbf{X}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}] = \int_{\mathbf{w}} [\mathbf{y}|g^{-1}(\mathbf{w}), \boldsymbol{\phi}][\mathbf{w}|\mathbf{X}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}]d\mathbf{w}.$$

Let us denote $\ell(\mathbf{w}, \cdot) = \log([\mathbf{y}|g^{-1}(\mathbf{w}), \boldsymbol{\phi}][\mathbf{w}|\mathbf{X}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}])$, and consider $\int e^{\ell(\mathbf{w},\cdot)}d\mathbf{w}$. Let $\mathbf{v}$ be the gradient vector with ith element

$$v_i(\boldsymbol{\phi}, \boldsymbol{\theta}) = \frac{\partial \ell(\mathbf{w}, \cdot)}{\partial w_i},$$

and let $\mathbf{H}$ be the Hessian matrix with i, jth element,

$$H_{i,j}(\boldsymbol{\phi}, \boldsymbol{\theta}) = \frac{\partial^2 \ell(\mathbf{w}, \cdot)}{\partial w_i \partial w_j},$$

where for both $v_i(\boldsymbol{\phi}, \boldsymbol{\theta})$ and $H_{i,j}(\boldsymbol{\phi}, \boldsymbol{\theta})$ we show dependence on parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. Using the multivariate Taylor series expansion around some point $\mathbf{a}$,

$$\int_{\mathbf{w}} e^{\ell(\mathbf{w},\cdot)}d\mathbf{w} \approx \int_{\mathbf{w}} e^{\ell(\mathbf{a},\cdot)+\mathbf{v}'(\mathbf{w}-\mathbf{a})+1/2(\mathbf{w}-\mathbf{a})'\mathbf{H}(\mathbf{w}-\mathbf{a})}d\mathbf{w}.$$

Now if $\mathbf{a}$ is a value for $\ell(\mathbf{a}, \cdot)$ such that $\mathbf{v} = \mathbf{0}$, then

$$\int_{\mathbf{w}} e^{\ell(\mathbf{w},\cdot)}d\mathbf{w} \approx e^{\ell(\mathbf{a},\cdot)} \int_{\mathbf{w}} e^{-1/2(\mathbf{w}-\mathbf{a})'(-\mathbf{H})(\mathbf{w}-\mathbf{a})}d\mathbf{w} = e^{\ell(\mathbf{a},\cdot)}(2\pi)^{n/2}|-\mathbf{H}_{\mathbf{a}}(\boldsymbol{\phi}, \boldsymbol{\theta})|^{-1/2}.$$

where $\mathbf{H}_{\mathbf{a}}(\boldsymbol{\phi}, \boldsymbol{\theta})$ indicates $\mathbf{H}$ evaluated at $\mathbf{a}$ and we again show its dependence on $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. The result on the most right-hand side is familiar from the normalizing constant of a

multivariate Gaussian distribution. Hence,

$$[\mathbf{y}|\boldsymbol{\phi}, \mathbf{X}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}] = \int_{\mathbf{w}} e^{\ell(\mathbf{w},\cdot)} d\mathbf{w} \approx [\mathbf{y}|g^{-1}(\mathbf{a}), \boldsymbol{\phi}][\mathbf{a}|\mathbf{X}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}](2\pi)^{n/2}|-\mathbf{H}_{\mathbf{a}}(\boldsymbol{\phi}, \boldsymbol{\theta})|^{-1/2}. \quad (3)$$

2.2 Marginal Maximum Likelihood for Covariance Parameters

From (3) a marginal maximum likelihood estimator for $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, given $\mathbf{a}$, is

$$\{\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}\} = \arg\max_{\boldsymbol{\phi},\boldsymbol{\theta}} \left(\log[\mathbf{y}|g^{-1}(\mathbf{a}), \boldsymbol{\phi}] + \log[\mathbf{a}|\mathbf{X}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}] - \frac{1}{2}\log\left(|-\mathbf{H}_{\mathbf{a}}(\boldsymbol{\phi}, \boldsymbol{\theta})|\right) \right), \quad (4)$$

where we drop terms that do not contain $\boldsymbol{\phi}$ or $\boldsymbol{\theta}$. Note that $\log[\mathbf{a}|\mathbf{X}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}]$ form exactly the same loglikelihood equations for ML or REML as in standard Gaussian models, but here they are evaluated at $\mathbf{a}$, where for ML

$$\log[\mathbf{a}|\mathbf{X}, \{\mathbf{Z}_k\}, \boldsymbol{\theta}] = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}| - \frac{1}{2}(\mathbf{a} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{a}})'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\mathbf{a} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{a}}),$$

and for REML

$$\log[\mathbf{a}|\mathbf{X}, \{\mathbf{Z}_k\}, \boldsymbol{\theta}] = -\frac{n-p}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}| - \frac{1}{2}\log|\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\mathbf{X}| - \frac{1}{2}(\mathbf{a} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{a}})'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\mathbf{a} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{a}}),$$

where in both cases $\hat{\boldsymbol{\beta}}_{\mathbf{a}} = (\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{a}$. The result (4) depends on finding $\mathbf{a}$ so that $\mathbf{v} = \mathbf{0}$. To achieve this, we use Newton-Raphson, conditional on $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, which we describe next.

Assuming conditional independence of $\mathbf{y}$ on $g^{-1}(\mathbf{w})$,

$$\log([\mathbf{y}|g^{-1}(\mathbf{w}),\boldsymbol{\phi}][\mathbf{w}|\mathbf{X},\boldsymbol{\Sigma_\theta}]) = \sum_{i=1}^{N}\log[y_i|w_i,\boldsymbol{\phi}] - \frac{1}{2}(\mathbf{w}-\mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\mathbf{w}-\mathbf{X}\hat{\boldsymbol{\beta}}) + C, \tag{5}$$

where C are terms that do not contain $\mathbf{w}$. Let $\mathbf{d}_{\boldsymbol{\phi}}$ be the vector with ith component,

$$d_i \equiv \frac{\partial \log[y_i|g^{-1}(w_i),\boldsymbol{\phi}]}{\partial w_i},$$

and note that

$$\frac{\partial[-\frac{1}{2}(\mathbf{w}-\mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\mathbf{w}-\mathbf{X}\hat{\boldsymbol{\beta}})]}{\partial \mathbf{w}} = -\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{w} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}},$$

so the gradient of (5) is

$$\mathbf{v} = \mathbf{d}_{\boldsymbol{\phi}} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{w} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{d}_{\boldsymbol{\phi}} - \mathbf{P}_{\boldsymbol{\theta}}\mathbf{w}, \tag{6}$$

where $\mathbf{P}_{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}$. For the Hessian, let $\mathbf{D}_{\boldsymbol{\phi}}$ be a diagonal matrix with ith component,

$$D_{i,i} \equiv \frac{\partial^2 \log[y_i|g^{-1}(w_i),\boldsymbol{\phi}]}{\partial w_i^2}, \tag{7}$$

where all off-diagonal elements are zero because all second partials are 0 when $i \neq j$ due to conditional independence. Then the Hessian of (5) is

$$\mathbf{H} = \mathbf{D}_{\boldsymbol{\phi}} - \mathbf{P}_{\boldsymbol{\theta}}. \tag{8}$$

Note the difference in (6) and (8) from Bonat and Ribeiro Jr (2016), where for their development $\mathbf{P}_{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}$, because we used $\hat{\boldsymbol{\beta}}$ in (5), which contains $\mathbf{w}$, whereas Bonat and Ribeiro Jr (2016) used $\boldsymbol{\beta}$.

A table of $\mathbf{d}_i$ and $\mathbf{D}_{i,i}$ for a few common distributions and link functions is given in

Table 1. In Table 1, we used alternative parameterizations for the negative binomial, gamma, and beta distributions so that $\mathrm{E}(y) = \mu$. We also reparameterize the inverse Gaussian distribution, and details for all distributions are given in the Appendix.

Table 1: Flexibility of the HGLMM, showing how different distributions can be matched with different patterned covariance matrices. We also show distributions, inverse link functions, and first and second partial derivative with respect to w_i for various parts of the loglikelihood.

$\log[\mathbf{y}\|g^{-1}(\mathbf{w}),\boldsymbol{\phi}]$		$-(1/2)\log\|-\mathbf{H}_{\mathbf{a}}(\boldsymbol{\phi},\boldsymbol{\theta})\|$		$+\log[\mathbf{a}\|\mathbf{X},\{\mathbf{Z}_k\},\boldsymbol{\theta}]$
Distribution	$\boldsymbol{\mu}=g^{-1}(\mathbf{w})$	$\mathbf{d}_i$	$\mathbf{D}_{i,i}$	$\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$-types
Binomial	$\boldsymbol{\mu}=\frac{\exp(\mathbf{w})}{1+\exp(\mathbf{w})}$	$y_i-\frac{n_i\exp(w_i)}{1+\exp(w_i)}$	$-\frac{n_i\exp(w_i)}{(1+\exp(w_i))^2}$	Random Effects
Poisson	$\boldsymbol{\mu}=\exp(\mathbf{w})$	$y_i-\exp(w_i)$	$-\exp(w_i)$	Geostatistical
Neg. Binomial	$\boldsymbol{\mu}=\exp(\mathbf{w})$	$\frac{\phi(y_i-e^{w_i})}{\phi+e^{w_i}}$	$-\frac{\phi e^{w_i}(\phi+y_i)}{(\phi+e^{w_i})^2}$	Spatial Areal
Gamma	$\boldsymbol{\mu}=\exp(\mathbf{w})$	$-\phi+y_i\phi e^{-w_i}$	$-y_i\phi e^{-w_i}$	Time Series
Inv. Gaussian	$\boldsymbol{\mu}=\exp(\mathbf{w})$	$\phi\left(\frac{y}{2e^{w_i}}-\frac{e^{w_i}}{2y}\right)+\frac{1}{2}$	$-\frac{\phi(e^{2w_i}+y_i^2)}{2ye^{w_i}}$	Spatio-temporal
Beta	$\boldsymbol{\mu}=\frac{\exp(\mathbf{w})}{1+\exp(\mathbf{w})}$	$\frac{-\phi e^{w_i}k_0(w_i\|\phi,y_i)}{(e^{w_i}+1)^2}$	$\frac{-\phi e^{2w_i}k_1(w_i\|\phi,y_i)}{(e^{w_i}+1)^4}$	

$k_0(w_i|\phi,y_i)=\psi^{(0)}\left(\frac{\phi e^{w_i}}{1+e^{w_i}}\right)-\psi^{(0)}\left(\frac{\phi}{1+e^{w_i}}\right)+\log\left(\frac{1}{y_i}-1\right)$
$k_1(w_i|\phi,y_i)=\phi\left(\psi^{(1)}\left(\frac{\phi e^{w_i}}{1+e^{w_i}}\right)+\psi^{(1)}\left(\frac{\phi}{1+e^{w_i}}\right)\right)-2\sinh(w_i)\left(k_0(w_i|\phi,y_i)+2\tanh^{-1}(1-2y_i)\right)$
$\psi^{(n)}(\cdot)$ is the nth derivative of the digamma function
sinh and tanh are the hyperbolic sine and tangent functions, respectively

Conditional on $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, a Newton-Raphson update is,

$$\mathbf{w}^{[k+1]}=\mathbf{w}^{[k]}-\mathbf{H}^{-1}\mathbf{v},$$

and upon convergence we set $\mathbf{a}=\mathbf{w}$ in (4) for any evaluation of the likelihood for given $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. Notice that this makes the marginal maximum likelihood doubly iterative, as we solve for $\mathbf{a}$ while optimizing for $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. It is possible to use other maximization routines,

such as the EM algorithm, but, generally, the Newton-Raphson algorithm converges rapidly (often around 10 iterations in our experience), and this was favored by Bonat and Ribeiro Jr (2016) also. However, on occasion, the stepsize needs to be adjusted so that $\mathbf{v}$ does not diverge. For example, it is easy and fast to check $\mathbf{v}^{[k+1]} = \mathbf{d}_{\boldsymbol{\phi}} - \mathbf{P}_{\boldsymbol{\theta}}\mathbf{w}^{[k+1]}$, and if $\mathbf{v}^{[k+1]}$ is "larger" than $\mathbf{v}$ by some criterion (e.g., largest or average element of $\mathbf{v}$), then take

$$\mathbf{w}^{[k+1]} = \mathbf{w}^{[k]} - \alpha\mathbf{H}^{-1}\mathbf{v},$$

where $0 < \alpha < 1$. In the simulations below, we check $\mathbf{v}^{[k+1]}$ in the manner described above, and set $\alpha = 0.1$ if the largest element of $\mathbf{v}^{[k+1]}$ is larger than the largest element of $\mathbf{v}$. The advantage of using Newton-Raphson is that it provides $\mathbf{H}$, which is useful for making adjustments to variances when estimating of fixed effects and making predictions, which we describe in the next section.

In summary, estimation of covariance parameters and $\mathbf{w}$ can be written in the following steps,

1. Get initial values for covariance parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. For example, for variance components. such as σ_0^2 and σ_1^2, apportion $\text{var}(g(\mathbf{y}))$ equally among each variance component. If there are many explanatory variables, a linear model can be fit to $g(\mathbf{y})$) and residual variance could be used.

2. Pick initial values for $\mathbf{w}$. For example, set $\mathbf{w} = g(\mathbf{y})$ or as the residuals from a linear model fit to $g(\mathbf{y})$.

3. Use Newton-Raphson to estimate $\mathbf{w} = \mathbf{a}$ for given $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ in (4).

4. Evaluate the loglikelihood in in (4) for $\mathbf{w}$, $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$.

5. Loop through steps 3 and 4 for different values of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ while optimizing for the

loglikelihood in step 4 until convergence.

2.3 Inference for Fixed Effects

In order to estimate $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ it was necessary to optimize the likelihood for $\mathbf{w}$, which we called $\mathbf{a}$, using Newton-Raphson, for each evaluation of the likelihood. Upon convergence in estimating $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, we also have optimized values for $\mathbf{w}$, and let us denote them as $\hat{\mathbf{w}} = \mathbf{a}$.

Bonat and Ribeiro Jr (2016) proposed profile likelihood for estimating $\boldsymbol{\beta}$ and obtaining confidence intervals, but their proposal will be very slow and does not extend well to cases with many coefficients in $\boldsymbol{\beta}$. Instead, an obvious estimator of $\boldsymbol{\beta}$ is to consider $\hat{\mathbf{w}}$ as if they were observed data, and then use the generalized least squares estimator $\hat{\boldsymbol{\beta}} = \mathbf{B}\hat{\mathbf{w}}$, where $\mathbf{B} = (\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}$. However, $\mathbf{w}$ contains predictions of unobserved, latent random variables, rather than observed values. We need to make some adjustments in order to estimate the variance of $\hat{\boldsymbol{\beta}}$. It is convenient to condition on $\mathbf{w}$ as if we had observed them, and then

$$\text{var}(\mathbf{B}\hat{\mathbf{w}}) = \text{E}_{\mathbf{w}}[\text{var}(\mathbf{B}\hat{\mathbf{w}}|\mathbf{w})] + \text{var}_{\mathbf{w}}[\text{E}(\mathbf{B}\hat{\mathbf{w}}|\mathbf{w})].$$

We will assume that $\hat{\mathbf{w}}$ is unbiased for $\mathbf{w}$, i.e., $E(\hat{\mathbf{w}}|\mathbf{w}) = \mathbf{w}$, so $\text{var}_{\mathbf{w}}[\text{E}(\mathbf{B}\hat{\mathbf{w}}|\mathbf{w})] = \mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\mathbf{B}'$, which simplifies to $\mathbf{C}_{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X})^{-1}$, the usual variance-covariance matrix of fixed effects when using generalized least squares. We will denote this as $\mathbf{C}_{\hat{\boldsymbol{\beta}}}$ when replacing $\boldsymbol{\theta}$ with its marginal estimate $\hat{\boldsymbol{\theta}}$ in $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}$. We can use the inverse of the observed Fisher-Information to obtain $\text{var}(\hat{\mathbf{w}}|\mathbf{w})$, which is $-\mathbf{H}_{\hat{\mathbf{w}}}(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})^{-1}$, where, as an approximation, we replace $\mathbf{w}$ with their predicted values $\hat{\mathbf{w}} = \mathbf{a}$. Notice that $-\mathbf{H}_{\hat{\mathbf{w}}}(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})^{-1}$ depends on $\mathbf{w}$ and $\boldsymbol{\phi}$ through the diagonal elements of $\mathbf{D}$ in (8), and it depends on parameters in $\boldsymbol{\theta}$ through $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$. Then an estimator of the covariance matrix of fixed effects is

$$\widehat{\text{var}}(\hat{\boldsymbol{\beta}}) = \mathbf{B}[-\mathbf{H}_{\hat{\mathbf{w}}}(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})^{-1}]\mathbf{B}' + \mathbf{C}_{\hat{\boldsymbol{\beta}}}. \tag{9}$$

2.4 Inference for Prediction

So far, we have estimated $\boldsymbol{\theta}$, $\boldsymbol{\phi}$, $\boldsymbol{\beta}$, and $\mathbf{w}$, and obtained estimated covariance matrices for $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{w}}$. Now let us consider the case of prediction for unsampled data, which may be in space, or time, or by design. We will denote unsampled $\{w_i\}$ with the vector $\mathbf{u}$. We can extend the linear model (1) as

$$\begin{pmatrix} \mathbf{w} \\ \mathbf{u} \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \mathbf{X_u} \end{pmatrix} \boldsymbol{\beta} + \sum_{k=1}^{q} \begin{pmatrix} \mathbf{Z}_k \\ \mathbf{Z}_{\mathbf{u},k} \end{pmatrix} \mathbf{r}_k + \begin{pmatrix} \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon}_\mathbf{u} \end{pmatrix}. \tag{10}$$

Our goal is prediction of $\mathbf{u}$. The universal kriging predictor is linear in the "data," which here is $\hat{\mathbf{w}}$, and can be written as $\hat{\mathbf{u}} = \boldsymbol{\Lambda}\hat{\mathbf{w}}$, where $\boldsymbol{\Lambda} = \mathbf{X}_\mathbf{u}'\mathbf{B} + \boldsymbol{\Sigma}_{\mathbf{wu}}'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} - \boldsymbol{\Sigma}_{\mathbf{wu}}'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X}\mathbf{B}$ and where $\boldsymbol{\Sigma}_{\mathbf{wu}}'$ is the covariance matrix between $\mathbf{w}$ and $\mathbf{u}$ (Cressie, 1993, p. 173). We want an estimator of the mean-squared-prediction errors, also called the prediction variance, which is $\text{var}(\hat{\mathbf{u}} - \mathbf{u}) = \text{var}(\boldsymbol{\Lambda}\hat{\mathbf{w}} - \mathbf{u})$. Now, if we had observed $\mathbf{w}$, rather than predicting $\hat{\mathbf{w}}$, then $\text{var}(\boldsymbol{\Lambda}\hat{\mathbf{w}} - \mathbf{u})$ is given by the usual mean-squared prediction errors, also known as the universal kriging equations in geostatistics,

$$\text{var}(\hat{\mathbf{u}} - \mathbf{u}) = \boldsymbol{\Sigma}_{\mathbf{uu}} - \boldsymbol{\Sigma}_{\mathbf{wu}}'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\Sigma}_{\mathbf{wu}} + \mathbf{K}\mathbf{C}_{\boldsymbol{\beta}}\mathbf{K}', \tag{11}$$

where $\mathbf{K} = \mathbf{X}_\mathbf{u} - \boldsymbol{\Sigma}_{\mathbf{wu}}'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X}$ (Cressie, 1993, p. 173), but again we need to make some adjustments because we are estimating $\hat{\mathbf{w}}$. Conditioning on $\mathbf{w}$ and $\mathbf{u}$, we have

$$\text{var}(\boldsymbol{\Lambda}\hat{\mathbf{w}} - \mathbf{u}) = \text{E}_{\mathbf{w},\mathbf{u}}[\text{var}(\boldsymbol{\Lambda}\hat{\mathbf{w}} - \mathbf{u}|\mathbf{w}, \mathbf{u})] + \text{var}_{\mathbf{w},\mathbf{u}}[\text{E}(\boldsymbol{\Lambda}\hat{\mathbf{w}} - \mathbf{u}|\mathbf{w}, \mathbf{u})].$$

As we did for estimating fixed effects, we will assume that $\hat{\mathbf{w}}$ is unbiased for $\mathbf{w}$, so $\text{E}(\boldsymbol{\Lambda}\hat{\mathbf{w}} - \mathbf{u}|\mathbf{w}, \mathbf{u}) = \boldsymbol{\Lambda}\mathbf{w} - \mathbf{u}$, and the variance of this will be the same as if we had observed $\mathbf{w}$, so $\text{var}_{\mathbf{w},\mathbf{u}}(\boldsymbol{\Lambda}\mathbf{w} - \mathbf{u})$ is given by (11). Conditionally, $\text{var}_{\hat{\mathbf{w}}}(\boldsymbol{\Lambda}\hat{\mathbf{w}} - \mathbf{u})$ does not depend on $\mathbf{u}$,

so $\mathrm{E}_{\mathbf{w},\mathbf{u}}[\mathrm{var}(\boldsymbol{\Lambda}\hat{\mathbf{w}} - \mathbf{u}|\mathbf{w},\mathbf{u})] = \mathrm{E}_{\mathbf{w}}(\boldsymbol{\Lambda}[-\mathbf{H}_{\mathbf{w}}(\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\theta}})^{-1}]\boldsymbol{\Lambda}')$, and, to take expectation, we simply replace $\mathbf{w}$ in $\mathbf{H}$ with its estimator $\hat{\mathbf{w}} = \mathbf{a}$. Putting them together, we obtain

$$\widehat{\mathrm{var}}(\boldsymbol{\Lambda}\hat{\mathbf{w}} - \mathbf{u}) = \boldsymbol{\Lambda}[-\mathbf{H}_{\hat{\mathbf{w}}}(\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\theta}})^{-1}]\boldsymbol{\Lambda}' + \boldsymbol{\Sigma}_{\mathbf{uu}} - \boldsymbol{\Sigma}'_{\mathbf{wu}}\boldsymbol{\Sigma}^{-1}_{\hat{\boldsymbol{\theta}}}\boldsymbol{\Sigma}_{\mathbf{wu}} + \mathbf{K}\mathbf{C}_{\hat{\boldsymbol{\beta}}}\mathbf{K}', \tag{12}$$

where $\boldsymbol{\Sigma}_{\mathbf{uu}}$ is the covariance matrix of $\mathbf{u}$. All covariance matrices depend on $\boldsymbol{\theta}$, which is replaced by its estimator $\hat{\boldsymbol{\theta}}$, and the fitted covariance function that was used to obtain $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is used for $\boldsymbol{\Sigma}_{\mathbf{wu}}$ and $\boldsymbol{\Sigma}_{\mathbf{uu}}$.

3 Simulations

We first illustrate our method with a simulation of spatial data so that we know all true values. We created a square grid of 20×20 locations equally spaced on a $(0,1) \times (0,1)$ domain. Let $\mathbf{X} = \mathbf{1}$ with a single overall mean parameter $\beta_0 = 2$. We generated $\mathbf{w}$ from an exponential autocovariance model, $\mathrm{cov}(w(\mathbf{s}_i), w(\mathbf{s}_j)) = \sigma_1^2 \exp(-\delta_{i,j}/\rho) + \sigma_0^2 \mathcal{I}(\delta_{i,j} = 0)$, where $\mathbf{s}_i$ are the spatial coordinates at location i, $\delta_{i,j}$ is Euclidean distance between the ith and jth locations, $\mathcal{I}(\cdot)$ is the indicator function, equal to one if its argument is true, otherwise it is zero, and we let $\sigma^2 = 1$, $\rho = 1$, and $\sigma_0^2 = 0.0001$. The 400 simulated $\mathbf{w}$ values are shown in Figure 1B. Conditional on the $\mathbf{w}$, at each spatial location we independently simulated a Poisson random variable with mean equal to $\exp(w_i)$, which are shown in Figure 1A.

Using the values in Figure 1A, we assumed an unknown mean and exponential covariance function. Optimizing the likelihood in (4) using REML for $\boldsymbol{\theta} = (\sigma_1^2, \rho, \sigma_0^2)$ we obtained the values $\hat{\sigma}_1^2 = 1.247$, $\hat{\rho} = 1.341$, and $\hat{\sigma}_0^2 = 1.392 \times 10^{-11}$. The likelihood surface for σ_1^2 and ρ is shown in Figure 1C. A pronounced ridge shows the positive association in the likelihood between σ_1^2 and ρ, which is typical for geostatistical models. The estimation of $\boldsymbol{\theta} = (\sigma_1^2, \rho, \sigma_0^2)$ also provided $\hat{\mathbf{w}}$, which are shown in Figure 1D, and it appears that we were able to recover

the spatial patterning of the true simulated $\mathbf{w}$ quite well.

Figure 1 here: Estimation for simulated data. A. Simulated count data using the model described in the text. B. The true simulated $\mathbf{w}$ values. C. The likelihood surface of the simulated data. The white circle shows the estimated value. D. The estimated $\hat{\mathbf{w}}$ values.

Of course, this is just one simulation. We did a computer simulation experiment in order to check for bias and confidence interval coverage. Here, we created 200 locations randomly on the unit square. We used the same autocovariance model as above, $\text{cov}(w(\mathbf{s}_i), w(\mathbf{s}_j)) = \sigma_1^2 \exp(-\delta_{i,j}/\rho) + \sigma_0^2 \mathcal{I}(\delta_{i,j} = 0)$ with $\sigma^2 = 1$, $\rho = 1$, and $\sigma_0^2 = 0.0001$. For the beta we took $phi = 10$, and for negative binomial, gamma, and inverse Gaussian we took $\phi = 1$.

For the mean structure, we used

$$\text{E}(w_i) = \beta_0 + \beta_1 x_i + \beta_2 \tau_i + \beta_3 (x{:}\tau)_i,$$

where x_i was randomly and independently simulated from $\text{N}(0, 1)$, τ_i was randomly and independently simulated as a Bernoulli variable with probability $p = 0.5$, and $(x{:}\tau)_i$ was in interaction between the normally-distributed and Bernoulli-distributed explanatory variables. We set $\boldsymbol{\beta} = (0.5, 0.5, -0.5, 0.5)$. We also created 100 prediction locations where we used a 10×10 square grid of equally-spaced prediction points throughout the unit square. Explanatory variables were also simulated at the prediction locations, and so 300 w_i values were simulated from $\text{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$. We then created the observed data as counts from a Poisson distribution conditional on the $\mathbf{w}$, where at each spatial location we independently simulated the Poisson random variable with mean equal to $\exp(w_i)$. We simulated 2000 data sets to assess bias and confidence/prediction interval coverage.

For each simulated data set, we first estimated the covariance parameters using (4), where $\log[\mathbf{a}|\mathbf{X}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}]$ were the REML equations, and we used an exponential autocovariance

model for fitting. We truncated the parameter space for σ_1^2 to be less than 10 times var$(g(y))$ and ρ to be less than 10 times maximum distance among all spatial locations. We did this because sometimes either the estimation of σ^2 or ρ, or both, tends to infinity as the ridge seen in Figure 1C can be very flat, and only their ratio is important for estimation and prediction (Zhang, 2004). This stabilized estimation over so many simulations.

Then, we used the estimated covariance parameters as plugin values for the autocovariance model to obtain $\mathbf{\Sigma}_{\hat{\boldsymbol{\theta}}}$, and along with the estimated $\hat{\mathbf{w}}$, we estimated fixed effects $\hat{\boldsymbol{\beta}} = \mathbf{B}\hat{\mathbf{w}}$. To estimate bias, we took the average of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ over all 2000 simulated data sets. We also formed 90% confidence intervals as $\hat{\boldsymbol{\beta}} \pm 1.645\widehat{\text{se}}(\hat{\boldsymbol{\beta}})$, where $\widehat{\text{se}}(\hat{\boldsymbol{\beta}})$ were the square roots of the diagonal elements of (9). Over the 2000 simulations, we computed the proportion of times that the confidence intervals contained the true values. If we are estimating the variances of $\hat{\boldsymbol{\beta}}$ well, the coverage should be close to 90%. We also computed the confidence interval coverage based on the naive unadjusted $\mathbf{C}_{\hat{\boldsymbol{\beta}}}$.

The results are shown in Table 2, where we see that there is very little bias in estimating any of the parameters in $\boldsymbol{\beta}$. The confidence interval coverage for β_0 is low, but estimating the overall intercept is difficult for normal spatial models as well. The confidence interval coverages for β_1 through β_3 are very close to 90% when using (9), but they are much too short when using only $\mathbf{C}_{\hat{\boldsymbol{\beta}}}$.

We also used the estimated covariance parameters in $\mathbf{\Sigma}_{\hat{\boldsymbol{\theta}}}$ and the estimated $\hat{\mathbf{w}}$ to make predictions at all 100 values for each simulated data set using $\hat{\mathbf{u}} = \mathbf{\Lambda}\hat{\mathbf{w}}$. To estimate bias, we took the average of $\hat{\mathbf{u}} - \mathbf{u}$ for each simulated data set, where recall that $\mathbf{u}$ contains 100 simulated values, and then averaged those across the 2000 simulated data sets. We also formed 90% prediction intervals as $\hat{\mathbf{u}} \pm 1.645\widehat{\text{se}}(\hat{\mathbf{u}})$, where $\widehat{\text{se}}(\hat{\mathbf{u}})$ were the square roots of the diagonal elements of (12). Over the 2000 simulations, we computed the proportion of times that the prediction intervals contained the true values, which should be about 90%. Table 2 gives the prediction results, which show little indication of bias, and coverage was very close

to 90% when using (12), but too short when using the naive (11).

Table 2: Bias and coverage for estimation of fixed effects $\boldsymbol{\beta}$ and for prediction of $\mathbf{u}$ at unobserved locations. Coverage is for 90% confidence and prediction intervals, and CI90_c used the corrected versions in (9) and (12), while CI90_u shows coverage for the uncorrected standard-error estimator based on $\mathbf{C}_{\boldsymbol{\beta}}$ and the uncorrected prediction standard errors using (11).

effect	bias	CI90_u	CI90_c
$\hat{\beta}_0$	0.031	0.728	0.751
$\hat{\beta}_1$	-0.005	0.367	0.902
$\hat{\beta}_2$	0.001	0.325	0.912
$\hat{\beta}_3$	0.000	0.335	0.917
$\hat{\mathbf{u}}$	0.037	0.701	0.899

In addition to the Poisson distribution, we did simulations for all 5 of the other distributions in Table 1. All methods appear to be unbiased, so we only show the corrected 90% confidence interval coverage in Table 3.

Table 3: Interval coverage for estimation of fixed effects $\boldsymbol{\beta}$ and for prediction of $\mathbf{u}$ at unobserved locations for the five distributions in Table 1 that were not covered in Table 2; binomial (bino), beta, negative binomial (nbin), gamma (gamm) and inverse Gaussian (iGau). Coverage is for 90% confidence and prediction intervals, using the corrected versions in (9) and (12).

effect	bino	beta	nbin	gamm	iGua
$\hat{\beta}_0$	0.712	0.923	0.730	0.839	0.708
$\hat{\beta}_1$	0.901	0.828	0.885	0.898	0.908
$\hat{\beta}_2$	0.914	0.837	0.898	0.894	0.888
$\hat{\beta}_3$	0.915	0.816	0.901	0.892	0.900
$\hat{\mathbf{u}}$	0.887	0.744	0.883	0.892	0.875

4 Examples

We demonstrate the methods with three example data sets that use all of the distributions in Table 1, combined with covariance matrices developed through spatial autoregressive models, time series models, geostatistical models, and variance components that include random effects.

4.1 1980 Presidential Turnout in Texas

This dataset contains the proportion of population over age 19 that cast votes in the 1980 presidential election in the United States. The proportions are for each of the 254 counties in Texas. The data for the whole of the United States were collected and reported in Pace and Barry (1997), and available in the `R` package `spData`. We created a subset of the data for Texas only. The response variable is reported as a proportion, but we also created a

binary variable by taking those proportions greater than 0.5 and creating a one, otherwise it was zero. This means our binary response variable measures whether the proportion of turnout was greater than 0.5 or not. To illustrate, we will fit the binomial distribution (as a Bernoulli distribution because all sample sizes are one) to the binary response variable, and the beta distribution to the proportional response variable.

There are three explanatory variables in the data set: 1) proportion of population with college degrees, 2) proportion of home ownership, and 3) income per capita, where for all three variables the values are with respect to the total population over age 19 that were eligible to vote. A scatterplot of the logit of the proportional response variable for all three explanatory variables is given in Figure 2. Note that in an attempt to linearize relationships, we cubed the explanatory variable for the proportion of home ownership and took the natural logarithm of per capita income. The linear model that we consider then is,

$$\mathbf{w} = \mathbf{X}\boldsymbol{\beta} + \mathbf{r}_1$$

where $\mathbf{X}$ contained a column of ones for an overall mean and 3 columns for the (transformed) explanatory variables.

Figure 2 here: Scatterplot of the logit of voter-turnout response variable by the three explanatory variables. Note the transformations of some explanatory variables, where proportion of home ownership was cubed, and natural logs were taken of per capital income.

For the spatial random effects $\mathbf{r}_1$, we fit two spatial autoregressive models to the data for the 254 counties, a conditional autoregressive (CAR) and a simultaneous autoregressive (SAR). These models rely on neighbor definitions, rather than distance directly. We defined a neighbor as any other county whose centroid was within 150 km. Using that definition, some counties had but a single neighbor, while the maximum number of neighbors was 38. Let $\mathbf{W}$ denote a matrix of binary values indicating neighbors, where the diagonal is all zeros

(a site is not a neighbor of itself). Let $\mathbf{W}_{rs}$ be a "row-standardized" version of $\mathbf{W}$, where each row in $\mathbf{W}$ is divided by its row sum, $w_{i,+} = \sum_j W_{i,j}$, and where $W_{i,j}$ is the i, jth element of $\mathbf{W}$. Then the covariance matrix for a CAR model is

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \sigma^2(\mathbf{I} - \rho\mathbf{W}_{rs})^{-1}\mathbf{M}_{rs},$$

where $\mathbf{M}_{rs}$ is a diagonal matrix with ith diagonal element $1/w_{i,+}$. The SAR covariance matrix is

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \sigma^2[(\mathbf{I} - \rho\mathbf{W}_{rs})(\mathbf{I} - \rho\mathbf{W}'_{rs})]^{-1}.$$

For the binary response variable and the three explanatory variables, using (4) with REML, for the CAR covariance matrix we estimated $\sigma^2 = 0.625$ and $\rho = 0.999$, while for the SAR covariance matrix we estimated $\sigma^2 = 0.265$ and $\rho = 0.964$. The minimized value of the -2 $\times$ the loglikelihood in (4) was 298.14 for the CAR model, while it was 279.15 for the SAR model, indicating the SAR model was a better choice. Table 4 gives fixed effects estimates. Note the large difference in standard errors using the naive approach based on $\mathbf{C}_{\hat{\boldsymbol{\beta}}}$ and the corrected version given in (9).

Table 4: Estimated fixed effects table for Texas turnout data using binary response variable. The estimates are given by Est., while s.e.$_1$ is the naive standard error using only $\mathbf{C}_{\boldsymbol{\beta}}$ from Section 2.3 and while s.e.$_2$ is the corrected standard error using (9). The t-val. is the estimated divided by the corrected standard error, and the p-val. is the probability of obtaining the t-value if the effect were truly zero, which a t-distribution with 254 - 4 = 250 degrees of freedom.

	SAR model					CAR model	
Effect	Est.	s.e.$_1$	s.e.$_2$	t-val.	p-val.	Est.	s.e.$_2$
Intercept	-5.725	0.502	2.397	2.388	0.0177	-2.898	1.975
College	4.439	0.701	3.694	1.202	0.2307	4.137	2.979
Home-owner	69.768	2.221	12.771	5.463	< 0.0001	56.635	10.318
Income	-0.171	0.260	1.414	0.121	0.9040	-0.860	1.151

For the proportion turnout response variable, the beta distribution in Table 1 was used, and for the CAR covariance matrix we estimated $\sigma^2 = 0.312$, $\rho = 0.999$, and $\phi = 48.7$, while for the SAR covariance matrix we estimated $\sigma^2 = 0.0126$, $\rho = 0.941$, and $\phi = 46.9$. The minimized value of -2 $\times$ loglikelihood in (4) was -554.6 for the CAR model, while it was -557.8 for the SAR model, indicating the SAR model was a better choice, just as for the binary data. Table 5 gives fixed effects estimates. As for the binary data, there is a large difference in standard errors using the naive approach based on $\mathbf{C}_{\hat{\boldsymbol{\beta}}}$ and the corrected version given in (9). The overall patterns of coefficient estimates and their precision are similar between SAR and CAR models in both Tables 4 and 5. In comparing Table 4 to Table 5, there appears to be more precision in the estimates in Table 5, especially regarding the significance of the per capita income variable. This is not surprising because when we transformed the proportional turnout data into binary data, we invariably lose information.

Table 5: Estimated fixed effects table for Texas turnout data using proportional response variable. The headings are the same as for Table 4.

	SAR model					CAR model	
Effect	Est.	s.e.$_1$	s.e.$_2$	t-val.	p-val.	Est.	s.e.$_2$
Intercept	-1.614	0.105	0.253	6.379	<0.0001	-1.427	0.340
College	0.407	0.154	0.407	1.000	0.3184	0.481	0.394
Home-owner	8.711	0.486	1.300	6.703	<0.0001	8.552	1.273
Income	0.470	0.057	0.142	3.317	0.0010	0.390	0.143

The predicted $\hat{\mathbf{w}}$ values are shown in Figure 3. A spatial visualization of the binary data shows some apparent clustering of 1's and 0's (Figure 3A). The predicted $\hat{\mathbf{w}}$ values for the binary data using a SAR model have the highest values in the northern Texas "panhandle" and in central Texas (Figure 3B), and the pattern is similar for the predicted $\hat{\mathbf{w}}$ values for the binary data when using a CAR model (Figure 3C). A logit transformation of the raw proportional turnout data are shown in Figure 3D and the predicted $\hat{\mathbf{w}}$ values for the SAR model (Figure 3E) appear to smooth the raw data, with a similar spatial pattern to the binary data (Figure 3B) and to the the predicted $\hat{\mathbf{w}}$ values using a CAR model (Figure 3F). Note that, in contast to classical linear models where the elements of $\mathbf{w}$ would sum to zero, for spatial models this is not true, and $\mathbf{w}$ interacts somewhat with the overall mean β_0. Hence, we do not use the same breakpoints in the color ramps in Figure 3, as it is the relative patterns among subfigures that is of greatest interest.

Figure 3 here: Raw data and predicted spatial random effects ($\mathbf{w}$) for the Texas turnout data. A) raw binary data, where open circles are zeros and solid circles are ones, B) predicted $\hat{\mathbf{w}}$ using SAR model for binary data, C) predicted $\hat{\mathbf{w}}$ using CAR model for binary data, D) logit-transformed proportional turnout data, E) predicted $\hat{\mathbf{w}}$ using SAR model for proportional turnout data, F) predicted $\hat{\mathbf{w}}$ using CAR model for proportional turnout data.

For the beta distribution with the SAR covariance that was used for the proportional turnout data ϕ was estimated to be 46.9. The density of $[y|\mu, \phi]$ will depend on μ, which is affected by the explanatory variables. In Figure 4 we show a histogram of the raw data, and also the fitted probability density function, $[y|\mu, 46.9]$, which is a beta distribution, for $\mu =$ 0.3, 0.5, and 0.8.

Figure 4 here: Histogram of proportional turnout and fitted probability density functions for a beta distribution with $\phi = 46.9$ at μ values of 0.3 (dashed line), 0.5 (solid line), and 0.8 (dotted line).

4.2 Harbor Seal Counts in Alaska

For over 30 years, aerial surveys of harbor seals throughout Alaska have been flown by the Marine Mammal Laboratory of the Alaska Fisheries Science Center, part of the US government NOAA Fisheries. These surveys, primarily during the late summer months when seals are molting, are the primary method for monitoring and estimating the abundance of harbor seals (Muto et al., 2022). Based on genetic sampling, all seals in Alaska have been divided into 13 different "stocks," or genetic populations. Abundance estimates are created for each stock, and here we will use the stock known as the Sitka/Chatham Strait population. This dataset consists of 716 observations in the years 1998, 2003, 2008 - 2011, and 2015. All known harbor seal haul-out sites were collected into 74 sample polygons that were completely surveyed, but some sample polygons were counted multiple times per year, whereas some sample polygons might be completely skipped in a year. For each aerial count, explanatory variables included time-from-low-tide and time-of-day.

We used Poisson and negative binomial models in Table 1. We considered a specific

linear model (1),

$$\mathbf{w} = \mathbf{X}\boldsymbol{\beta} + \mathbf{r}_1 + \mathbf{Z}_2\mathbf{r}_2 + \boldsymbol{\epsilon},$$

where $\mathbf{X}$ contained a column for an overall mean, 73 columns with indicators of a mean effect for each sample polygon (as deviations from the mean for the first polygon, which was absorbed into the overall mean), and the explanatory variable time-of-day was the fraction of day from solar noon (the time when the sun is at the zenith) and the time-from-low-tide was in hours from low tide (tide cycles last about 12 hours in this area). Both explanatory variables also had a squared term, so $\mathbf{X}$ had 78 columns. The random effect $\mathbf{r}$ was assumed to have a first-order autoregressive (AR1) time series model so $\mathrm{cov}(r_{1,i}, r_{1,j}) = \sigma_1^2\rho^{|i-j|}/(1-\rho^2)$, where i and j are integers. Moreover, we assumed that all polygons were independent of each other, so the autocovariance only occurred within polygons, giving the covariance matrix a block diagonal structure. Within site, we had repeated measures per year, so $\mathbf{Z}_2$ was a design matrix created as an interaction between site and year, and $\sigma_2^2\mathbf{Z}_2\mathbf{Z}_2'$ allowed for additional correlation among repeated samples per year. When using a model with a Poisson distribution, we allowed for further uncorrelated overdispersion by using $\boldsymbol{\epsilon}$ as given in (1), but for the negative binomial distribution, which allows for overdispersion, we did not use $\boldsymbol{\epsilon}$.

Table 6: Estimated fixed effects table for the harbor seal count data. The headings are the same as for Table 4, where here, for the p-val, we used a t-distribution with 716 - 78 = 638 degrees of freedom. We show $s.e_1$ and $s.e_2$ for both negative binomial and Poisson distributions. All other columns in the Table were the estimated values for the negative binomial distribution.

	Negative Binomial					Poisson	
Effect	Est.	$s.e._1$	$s.e._2$	t-val.	p-val.	$s.e._1$	$s.e._2$
time-from-low-tide	-0.081	0.00005	0.042	1.938	0.0530	0.041	0.043
(time-from-low-tide)^2	-0.064	0.00003	0.026	2.495	0.0128	0.025	0.026
hour-of-day	-0.273	0.00013	0.107	2.542	0.0113	0.104	0.107
(hour-of-day)^2	-0.747	0.00017	0.146	5.120	<0.0001	0.139	0.144

Using the Poisson distribution with the AR1 covariance model, using (4) with REML we estimated $\sigma_0^2 = 0.859$, $\sigma_1^2 = 0.660$, $\sigma_2^2 = 0.0003$ and $\rho = 0.940$, while using the negative binomial distribution with the AR1 covariance model, we estimated $\sigma_1^2 = 3.637$, $\sigma_2^2 = 0.0012$, $\rho = 0.997$, and $\phi = 1.529$. The minimized value of the likelihood in (4) was 8416.11 for the Poisson distribution model, while it was 8242.07 for the negative binomial distribution, indicating that the negative binomial distribution was a better choice. Table 6 gives fixed effects estimates for the negative binomial model, except that we include the naive and corrected variances for the Poisson distribution as well. It is especially interesting that when $\boldsymbol{\epsilon}$ was included in the $\mathbf{w}$ values for the Poisson distribution, the diagonal elements of $\mathbf{C}_{\boldsymbol{\beta}}$ were large and increased only slightly when adjusted by $\mathbf{B}(-\mathbf{H}_{\mathbf{a}}^{-1})\mathbf{B}'$ as given in (9). This is in contrast to the negative binomial distribution, whose diagonal elements of $\mathbf{C}_{\boldsymbol{\beta}}$, as reflected by $s.e._1$ in Table 6, are very small. Yet, $s.e._2$ is almost identical for the negative binomial and Poisson models, despite their apparently different covariance structures.

The fitted explanatory variables allow insight into seal behavior, and seals prefer to

haul out of the water around midday and at low tides. The model confirms that counts are highest at these times, as the coefficients in Table 6 are plotted in Figure 5. This shows changes on the log scale where all other explanatory variables are held fixed at zero, so Figure 5 is interpreted as the log of the proportional change in expected counts with unit changes in the explanatory variable.

Figure 5 here: Fitted effects of A) hour-of-day and B) time-from-low-tide on harbor seal counts. The fitted effect shows the log of the expected proportional change from the zero value of all covariates.

Our ultimate goal was to predict the abundance of seals at each site, which used the results in Section 2.4. Predicted w-values, after exponentiating, are shown in Figure 6, for 4 of the 74 different sites, which are labeled AC10, AC11, BC01, and BC02. For each year, we predicted $\hat{\mathbf{u}}$ in (10) with prediction intervals using (12), and then exponentiated both predictions and prediction intervals. We see that the errors are large, but this is not unreasonable given the relatively few data per site. Note that we borrowed strength across sites for estimating the autocorrelation parameter ρ, assuming all sites had the same amount of autocorrelation among $\mathbf{w}$. Although each site had a large prediction interval, all sites will be summed to get an overall estimate of abundance. When summed over 74 sites, the variance goes down to an acceptable level. Note that, in general, the predictions tend to shrink toward the overall mean for the site, but for site BC02 especially, the predictions are greater than the observed values. This can be explained because the predictions are standardized to optimal conditions for the explanatory variables (time-of-day and time-to-low-tide). Site BC02 was counted in suboptimal conditions on almost all occasions, which is entirely possible because a high tide may occur at solar noon. It is impossible to optimize explanatory variables through a sampling design unless you are willing to wait and only sample when a low tide occurs at near solar noon.

Figure 6 here: Predicted w-values for 4 of the 74 sites. Open circles are raw counts, and solid circles are predicted w-values connected by a solid line. The dashed line shows the prediction intervals.

4.3 Heavy Metal Concentrations in Moss

Cape Krusenstern National Park is in northwest Alaska, USA, and nearby is the Red Dog mine, where zinc, lead, cadmium and other heavy metals are mined. Trucks haul ore to the coast from the Red Dog Mine on a haul road that traverses Cape Krusenstern National Park. There is speculation that dust escapes into the environment from those trucks. Mosses obtain much of their nutrients from the air, so they are ideal biomonitors for heavy metals attached to airborne dust. In 2001 (Hasselbach et al., 2005) and again in 2006 (Neitlich et al., 2017), mosses were sampled for heavy metals, with the sampling being more dense near the road. Current annual growth of moss tissue was sampled, ground, homogenized, and then sent for laboratory analysis. Here, we just consider lead concentrations, although many other elements were analyzed. Potentially important explanatory variables that we include are distance-from-haul-road, side-of-the-road (north or south), and year of sample. There are 365 records in the data set, with 244 from 2001 and 121 from 2006.

Data on lead concentration have only positive values, and are often skewed, which led Hasselbach et al. (2005) and Neitlich et al. (2017) to use log transformations. We used the gamma and inverse Gaussian models given in Table 1. For the covariance structure, we considered a special case of the linear model (1),

$$\mathbf{w} = \mathbf{X}\boldsymbol{\beta} + \mathbf{r}_1 + \mathbf{Z}_2\mathbf{r}_2 + \mathbf{Z}_3\mathbf{r}_3 + \boldsymbol{\epsilon},$$

where $\mathbf{X}$ contained a column for an overall mean, an indicator column for year 2006 (2001

was absorbed into the overall mean), log of distance to road (in meters), and and indicator for south of the road (north was absorbed into the overall mean). We also considered an interaction between distance to road and the side of the road.

The random effect $\mathbf{r}_1$ was assumed to have a geostatistical autocovariance structure, known as the exponential model, where $\text{cov}(r_1(\mathbf{s}_i), r_1(\mathbf{s}_j)) = \sigma_1^2 \exp(-\delta_{i,j}/\rho)$, where $\mathbf{s}_i$ is a vector containing the spatial coordinates of the ith location and $\delta_{i,j}$ is the Euclidean distance between locations $\mathbf{s}_i$ and $\mathbf{s}_j$. The parameter σ_1^2 is often called the partial sill, and ρ is the range parameter, which controls the distance-decay rate of the autocovariance with distance. The variance of $\boldsymbol{\epsilon}$, σ_0^2, is often called the nugget effect. We assumed that years were independent of each other. Within year and location, at some sites, duplicate samples were obtained to account for microscale variation; that is grabbing one handful of moss versus reaching over and grabbing another (distance among them was assumed to be zero). Hence, $\mathbf{Z}_2$ is a design matrix with indicator variables for location; this causes increased autocorrelation for any samples from the same location. Some samples were ground into two replicate samples for laboratory analysis, as there can be some variation in machines that measure concentration or the way it is homogenized. Therefore $\mathbf{Z}_3$ is a design matrix that contains indicator variables for a duplicate nested within location; this causes repeated replicates to have higher autocorrelation due to having a common duplicate. In the resulting covariance matrix, σ_1^2 is often called the partial sill, σ_2^2 is a variance due to location, σ_3^2 is a variance due to duplication, and σ_0^2 is measurement error that is confounded with what is often called the "nugget effect" in geostatistics, which is variance at a distance too small to measure.

Using the gamma distribution with the exponential covariance model and maximum likelihood (rather than REML), from (4) we obtained -2 $\times$ loglikelihood equal to 2318.253, and from the fixed effects table it appeared that the main effect for side-of-road was not significant. We re-ran the model without that main effect, and obtained -2 $\times$ loglikelihood

from (4) equal to 2319.391. Using either AIC or a likelihood ratio test, we have evidence to drop the main effect for side-of-road. After doing so, the marginal estimates of the covariance parameters were $\sigma_1^2 = 0.1703$, $\sigma_2^2 = 0.0634$, $\sigma_3^2 = 0.0266$, $\sigma_0^2 = 0.0023$, $\rho = 9.033$, and $\phi = 2218$. We fit the same mean and covariance structure, but with the inverse Gaussian distribution, and obtained -2 $\times$ loglikelihood = 2319.354, which is almost identical to the value when using the gamma distribution. The covariance parameters for the inverse Gaussian model were estimated to be $\sigma_1^2 = 0.1699$, $\sigma_2^2 = 0.0650$, $\sigma_3^2 = 0.0265$, $\sigma_0^2 = 0.0028$, $\rho = 9.133$, and $\phi = 2382$. The variance components are almost identical to the gamma distribution. Interestingly, if we fit a normal spatial model to the log of the data we obtain $\sigma_1^2 = 0.1702$, $\sigma_2^2 = 0.0633$, $\sigma_3^2 = 0.0267$, $\sigma_0^2 = 0.0028$, $\rho = 9.055$, which are almost identical to both the gamma and inverse Gaussian distributions.

Table 7: Estimated fixed effects table for the moss lead data when using the gamma distribution. The headings are the same as for Table 4, where here, for the p-val, we used a t-distribution with 365 - 4 = 361 degrees of freedom.

Effect	Est.	s.e.$_1$	s.e.$_2$	t-val.	p-val.
Intercept	8.074	0.20148	0.20178	40.0169	<0.0001
Year	-0.442	0.22149	0.22157	1.9932	0.0470
Distance to Road	-0.578	0.01835	0.01839	31.4377	<0.0001
Distance-to-road:Southside	-0.112	0.01165	0.01166	9.6119	<0.0001

The fixed effects tables looks almost identical for all three models, so only the one for the gamma model is shown Table 7. Note that for this example, in contrast to the previous two examples, there is little difference in the standard errors of the fixed effects based on s.e.$_1$ and s.e.$_2$. This can happen when essentially all of the variation is captured by $\mathbf{w}$ and the contribution of $\log[\mathbf{y}|g^{-1}(\mathbf{w}), \boldsymbol{\phi}]$ is small in comparison. This contribution is controlled

in part by ϕ for the gamma and inverse Gaussian distributions, whose estimated values are very large. To visualize, consider Figure 7. The histogram of the raw data (Figure 7A) shows highly skewed data over a large range of values. However, the fitted gamma distribution when the mean is 40 (conditional on the w-values) is quite narrow and symmetric (Figure 7B). As the mean increases to 150 (Figure 7C) and 850 (Figure 7D), the gamma distribution get wider due to the fact that the variance of the gamma distribution grows as μ^2.

Figure 7 here: A) Histogram of lead concentration in moss, B) fitted probability density at $\mu = 40$ with $\phi = 1415$ for the gamma distribution (dashed line) and $\phi = 0.00000147$ for the inverse Gaussian distribution (solid line) C) $\mu = 150$, and D) $\mu = 850$.

Predictions of the w-values are similar when using either a gamma distribution or an inverse Gaussian distribution, and they are both very similar to simply taking a log transformation of the data. Predictions of $\mathbf{w}$ by years 2001 and 2006 when using the gamma distribution are shown in Figure 8. The prediction locations are divided into 3 groups, one which was closely spaced near the haul road, and then more coarsely spaced locations with each successive group as it got farther from the road. It is clear that predicted values are largest near the road (Figure 8), but also that the predicted values generally went down from 2001 to 2006 due to a change from 2001 to 2006 where coverings were used on the trucks hauling ore on the road. The prevailing winds are from the south, so it is also clear that predicted values are higher on the north side of the road. The prediction standard errors show the typical pattern in geostatistics, where the standard errors are smallest near a sample or in dense concentrations of samples (Figure 8).

Figure 8 here: Prediction and their standard errors for 2001 and 2006 at locations near the haul road through Cape Krusenstern National Park, Alaska. The green $\times$ symbols show sample locations.

---

5 Discussion and Conclusions

We have developed a very flexible framework for modeling binary, count, positive continuous, and other data types in a hierarchical generalized linear mixed model framework. Virtually any data type can be accommodated by the many distributions that are known in statistics, and these distributions can be matched to virtually any patterned covariance matrix, where a short list is given in Table 1. Our examples illustrate all of the distributions in Table 1, and, for covariance matrices, we used CAR and SAR spatial autoregressive models, we used AR1 time series models, and we used exponential geostatistical models. We also showed how complex covariance matrices can be created by mixing random effects with other covariance structures. Any covariance matrix is possible in the HGLMM framework, including spatio-temporal, covariances for data on a sphere, covariances derived for linear networks such as streams and roads, etc. Using the Laplace approximation, the resulting loglikelihood is composed of the loglikelihood of the data distribution, the ML or REML loglikelihood for normally-distributed data, and a determinant of a Hessian matrix (4).

We have developed marginal inference for three of the most common objectives in linear models. First, in order to estimate fixed effects and make predictions, we must estimate all covariance parameters, which is accomplished from (4). Then, it is necessary to adjust variances for the fact that $\mathbf{w}$ are latend in the model, and not observed, which is accomplished from (9) when estimating fixed effects and from (12) when predicting at unsampled locations.

The models can be computationally demanding as they require computing the determinant of the Hessian matrix and, in our implementation, its inverse as well. Optimizing the likelihood is doubly iterative as Newton-Raphson updates are used during likelihood optimization for covariance parameters, requiring $\mathbf{H}^{-1}$ for each update. While this may limit the size of data sets for our HGLMM framework, we would like to point out some time-saving

features. First, note from (8) that

$$\mathbf{H} = [\mathbf{D} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}] + [\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X}](\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X})^{-1}[\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}],$$

where we add brackets to show that this has Sherman-Morrison-Woodbury form $\mathbf{A} + \mathbf{BCB}'$ (Sherman and Morrison, 1949; Woodbury, 1950). If $\mathbf{A}$ is $n \times n$ but has a fast inverse, and $\mathbf{C}$ has small dimension, then the inverse $(\mathbf{A} + \mathbf{BCB}')^{-1}$ can be made much faster than a full $n \times n$ inverse based on how quickly $\mathbf{A}^{-1}$ occurs. For example, consider our second example on harbor seals with 716 records at 74 sample sites. We assumed a time series model within site, but independence among sites, giving a block diagonal structure to the covariance matrix. Thus, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}$ has a block-wise inverse, and $\mathbf{D}$ is diagonal, so $[\mathbf{D} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}]^{-1}$ can be inverted block-wise, which is much faster than a single inverse for the whole $n \times n$ matrix. Similarly, $\mathbf{Z}_k\mathbf{Z}_k$ is often block-diagonal, and multiple variance components can use Sherman-Morrison-Woodbury recursively (Dumelle et al., 2021).

An important consideration for these models is the interplay of the independent component $\boldsymbol{\epsilon}$ in (1) and $\boldsymbol{\phi}$ in $[\mathbf{y}|g^{-1}(\mathbf{w}), \boldsymbol{\phi}]$. The parameters $\boldsymbol{\phi}$ often control variance, and can be confounded with $\boldsymbol{\epsilon}$. For example, consider when $[\mathbf{y}|g^{-1}(\mathbf{w}), \boldsymbol{\phi}]$ is a normal distribution where $g^{-1}(\mathbf{w})$ is the identity function and ϕ has but one element – the variance parameter. Then ϕ and σ_0^2 will not be identifiable. More often $\boldsymbol{\phi}$ controls how variance is related to the mean, but we expect that there still can be some confounding. For any particular data set, this can be investigated though loglikelihood plots of σ_0^2 and ϕ, similar to Figure 1C, or with more experience on how these parameters interact for particular models.

The HGLMM framework in this paper can be contrasted to the mixed model extension of GLMs. The GLM framework is inspired by the regular exponential family of distributions, and these lead to what are called the "canonical" link functions. For example, the canonical link function for the gamma distribution is $-1/\mu$, but it is often changed to $w = g(\mu) =$

$1/\mu$. However, that implies that $\mu = g^{-1}(w) = 1/w$, but because w can be negative, it is possible for μ to have negative values. In a moment-based modeling framework using pseudo-likelihood with iteratively weighted least squares, this can be tolerated if the values stay fairly close to the parameter space, and it allows for a wide variety of link functions which provides a great amount of flexibility. However, in the HGLMM framework, which is fully parametric, the evaluation of the loglikelihood for $[\mathbf{y}|g^{-1}(\mathbf{w}), \boldsymbol{\phi}]$ is not possible if $g^{-1}(\mathbf{w})$ is outside of the parameter space for the mean. For HGLMMs, link and mean functions must be chosen to respect the parameter space.

We have given a broad outline of marginal inference under the HGLMM. There are many topics to explore that we have not mentioned. For example, we may want to make inference on predictions where $\mathbf{w}$ is back-transformed as $g^{-1}(\mathbf{w})$, and where the variability of $\mathbf{y}|\mathbf{w}$ is added. We may also want further functions of $g^{-1}(\mathbf{w})$ such as block averages. Likewise, we may want inferences on random effects (best linear unbiased predictions) of $\mathbf{r}_i$ in (1). Like most linear models, we can consider linear combinations of $\boldsymbol{\beta}$, or contrasts of $\boldsymbol{\beta}$ parameters, in making inference on fitted models, treatment effects, etc. We only covered the basic framework is this paper and there are many further research topics to develop.

Acknowledgments

The project received financial support from the National Marine Fisheries Service, NOAA and the U.S. Environmental Protection Agency (EPA). The findings and conclusions in the paper are those of the author(s) and do not necessarily represent the views of the reviewers nor the EPA or the National Marine Fisheries Service, NOAA. Any use of trade, product, or firm names does not imply an endorsement by the US Government.

Data and Software Availability

All data and code will be made available upon publication.

References

Berliner, L. M. (1996), "Hierarchical Bayesian time series models," in *Maximum Entropy and Bayesian Methods: Santa Fe, New Mexico, USA, 1995 Proceedings of the Fifteenth International Workshop on Maximum Entropy and Bayesian Methods*, Springer, pp. 15–22.

Besag, J. (1974), "Spatial interaction and the statistical analysis of lattice systems (with discussion)," *Journal of the Royal Statistical Society, Series B*, 36, 192–236.

Bonat, W. H. and Ribeiro Jr, P. J. (2016), "Practical likelihood analysis for spatial generalized linear mixed models," *Environmetrics*, 27, 83–89.

Breslow, N. E. and Clayton, D. G. (1993), "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Association*, 88, 9–25.

Chiles, J.-P. and Delfiner, P. (1999), *Geostatistics: Modeling Spatial Uncertainty*, New York: John Wiley & Sons.

Christensen, O. F. (2004), "Monte Carlo maximum likelihood in model-based geostatistics," *Journal of Computational and Graphical Statistics*, 13, 702–718.

Clayton, D. and Kaldor, J. (1987), "Empirical Bayes estimates of age-standardized relative risks for use in disease mapping," *Biometrics*, 43, 671–681.

Cressie, N. and Wikle, C. K. (2011), *Statistics for Spatio-Temporal Data*, Hoboken, NJ: John Wiley & Sons.

Cressie, N. A. C. (1993), *Statistics for Spatial Data, Revised Edition*, New York: John Wiley & Sons.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998), "Model-based geostatistics (with discussion)," *Journal of the Royal Statistical Society, Series C: Applied Statistics*, 47, 299–326.

Dumelle, M., Ver Hoef, J. M., Fuentes, C., and Gitelman, A. (2021), "A linear mixed model formulation for spatio-temporal random processes with computational advances for the product, sum, and product–sum covariance functions," *Spatial Statistics*, 43, 100510.

Evangelou, E., Zhu, Z., and Smith, R. L. (2011), "Estimation and prediction for spatial generalized linear mixed models using high order Laplace approximation," *Journal of Statistical Planning and Inference*, 141, 3564–3577.

Fisher, R. A. (1934), "Two new properties of mathematical likelihood," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 144, 285–307.

Gelfand, A. E. and Smith, A. F. M. (1990), "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, 85, 398–409.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996), "Introducing Markov Chain Monte Carlo," in *Markov Chain Monte Carlo in Practice*, Chapman and Hall, pp. 1–19.

Gneiting, T. (2013), "Strictly and non-strictly positive definite functions on spheres," *Bernoulli*, 19, 1327 – 1349.

Gotway, C. A. and Stroup, W. W. (1997), "A generalized linear model approach to spatial data analysis and prediction," *Journal of Agricultural, Biological, and Environmental Statistics*, 2, 157–178.

Haining, R. P. (1978), "The moving average model for spatial interaction," *Transactions of the Institute of British Geographers*, 3, 202–225.

Hamilton, J. D. (1994), *Time Series Analysis*, vol. 2, Princeton, NJ, USA: Princeton University Press.

Hasselbach, L., Ver Hoef, J. M., Ford, J., Neitlich, P., Crecelius, E., Berryman, S., Wolk, B., and Bohle, T. (2005), "Spatial patterns of cadmium and lead deposition on and adjacent to National Park Service lands in the vicinity of Red Dog Mine, Alaska," *Science of the Total Environment*, 348, 211–230.

Huang, C., Zhang, H., and Robeson, S. M. (2011), "On the validity of commonly used covariance and variogram functions on the sphere," *Mathematical Geosciences*, 43, 721–733.

Kleinman, K. P. and Ibrahim, J. G. (1998), "A semi-parametric Bayesian approach to generalized linear mixed models," *Statistics in Medicine*, 17, 2579–2596.

Lee, Y. and Nelder, J. A. (1996), "Hierarchical generalized linear models," *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 619–656.

Lehmann, E. L. and Casella, G. (2006), *Theory of point estimation*, Springer Science & Business Media.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000), "WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility," *Statistics and Computing*, 10, 325–337.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models, 2nd Edition*, Chapman & Hall Ltd.

Neitlich, P. N., Hoef, J. M. V., Berryman, S. D., Mines, A., Geiser, L. H., Hasselbach, L. M., and Shiel, A. E. (2017), "Trends in spatial patterns of heavy metal deposition on national park service lands along the Red Dog Mine haul road, Alaska, 2001–2006," *PLOS ONE*, 12, e0177936.

Nelder, J. A. and Wedderburn, R. W. M. (1972), "Generalized linear models," *Journal of the Royal Statistical Society, Series A: General*, 135, 370–384.

Pace, R. K. and Barry, R. (1997), "Quick computation of spatial autoregressive estimators," *Geographical analysis*, 29, 232–247.

Patterson, H. and Thompson, R. (1974), "Maximum likelihood estimation of components of variance," in *Proceedings of the 8th International Biometric Conference*, Biometric Society, Washington, DC, pp. 197–207.

Patterson, H. D. and Thompson, R. (1971), "Recovery of inter-block information when block sizes are unequal," *Biometrika*, 58, 545–554.

Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71, 319–392.

Sherman, J. and Morrison, W. J. (1949), "Adjustment of an inverse matrix corresponding to changes in the elements of a given column or a given row of the original matrix," *Annals of Mathematical Statistics*, 20, 621.

Snedecor, G. W. and Cochran, W. G. (1980), *Statistical Methods, Seventh Edition*, Iowa State University.

Stiratelli, R., Laird, N., and Ware, J. H. (1984), "Random-effects models for serial observations with binary response," *Biometrics*, 40, 961–971.

Tierney, L. and Kadane, J. B. (1986), "Accurate approximations for posterior moments and marginal densities," *Journal of the American Statistical Association*, 81, 82–86.

Ver Hoef, J. M. (2018), "Kriging models for linear networks and non-Euclidean distances: Cautions and solutions," *Methods in Ecology and Evolution*, 9, 1600–1613.

Ver Hoef, J. M. and Peterson, E. (2010), "A moving average approach for spatial statistical models of stream networks (with discussion)," *Journal of the American Statistical Association*, 105, 6–18.

Ver Hoef, J. M., Peterson, E. E., Hooten, M. B., Hanks, E. M., and Fortin, M.-J. (2018), "Spatial autoregressive models for statistical inference from ecological data," *Ecological Monographs*, 88, 36–59.

Warton, D. I. and Hui, F. K. (2011), "The arcsine is asinine: the analysis of proportions in ecology," *Ecology*, 92, 3–10.

Wedderburn, R. W. M. (1974), "Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method," *Biometrika*, 61, 439–447.

Whittle, P. (1954), "On stationary processes in the plane," *Biometrika*, 41, 434–449.

Woodbury, M. A. (1950), *Inverting modified matrices*, Statistical Research Group, Princeton N.J., published: Memorandum Report 42.

Zeger, S. L. and Karim, M. R. (1991), "Generalized linear models With random effects; A Gibbs sampling approach," *Journal of the American Statistical Association*, 86, 79–86, publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988), "Models for longitudinal data: A generalized estimating equation approach," *Biometrics*, 44, 1049–1060, publisher: [Wiley, International Biometric Society].

Zhang, H. (2002), "On estimation and prediction for spatial generalized linear mixed models," *Biometrics*, 58, 129–136.

— (2004), "Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics," *Journal of the American Statistical Association*, 99, 250–261.

FIGURES

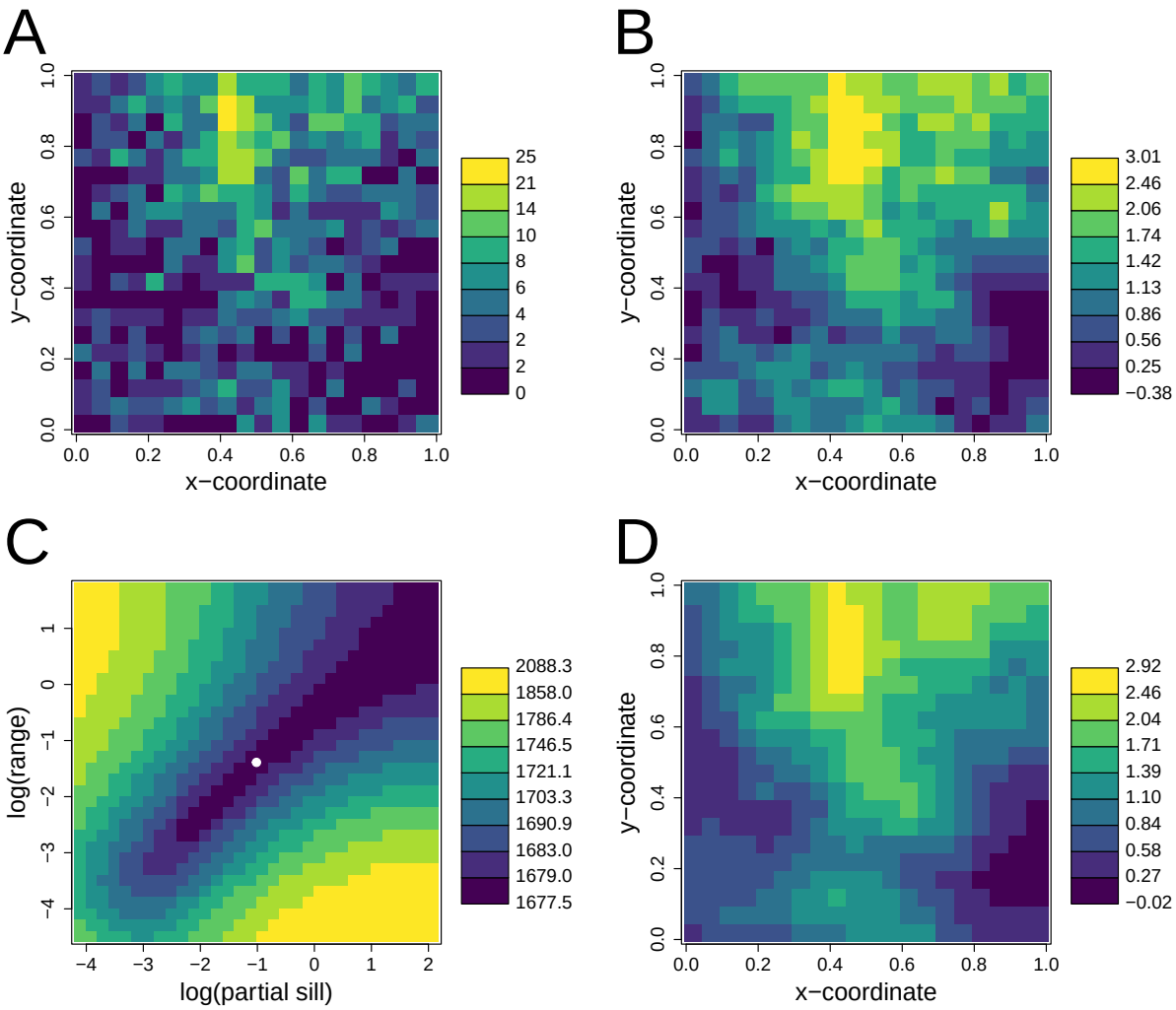

Figure 1: Estimation for simulated data. A. Simulated count data using the model described in the text. B. The true simulated $\mathbf{w}$ values. C. The likelihood surface of the simulated data. The white circle shows the estimated value. D. The estimated $\hat{\mathbf{w}}$ values.

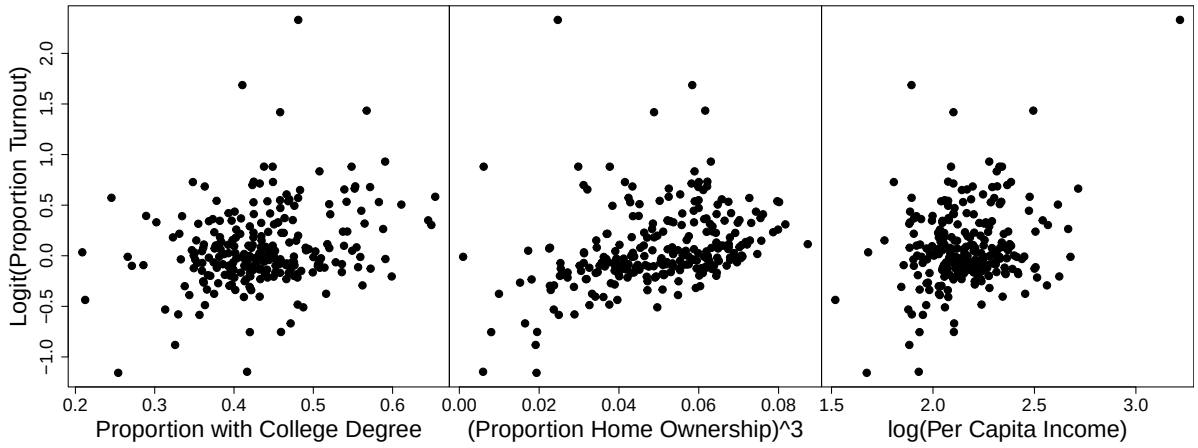

Figure 2: Scatterplot of the logit of voter-turnout response variable by the three explanatory variables. Note the transformations of some explanatory variables, where proportion of home ownership was cubed, and natural logs were taken of per capital income.

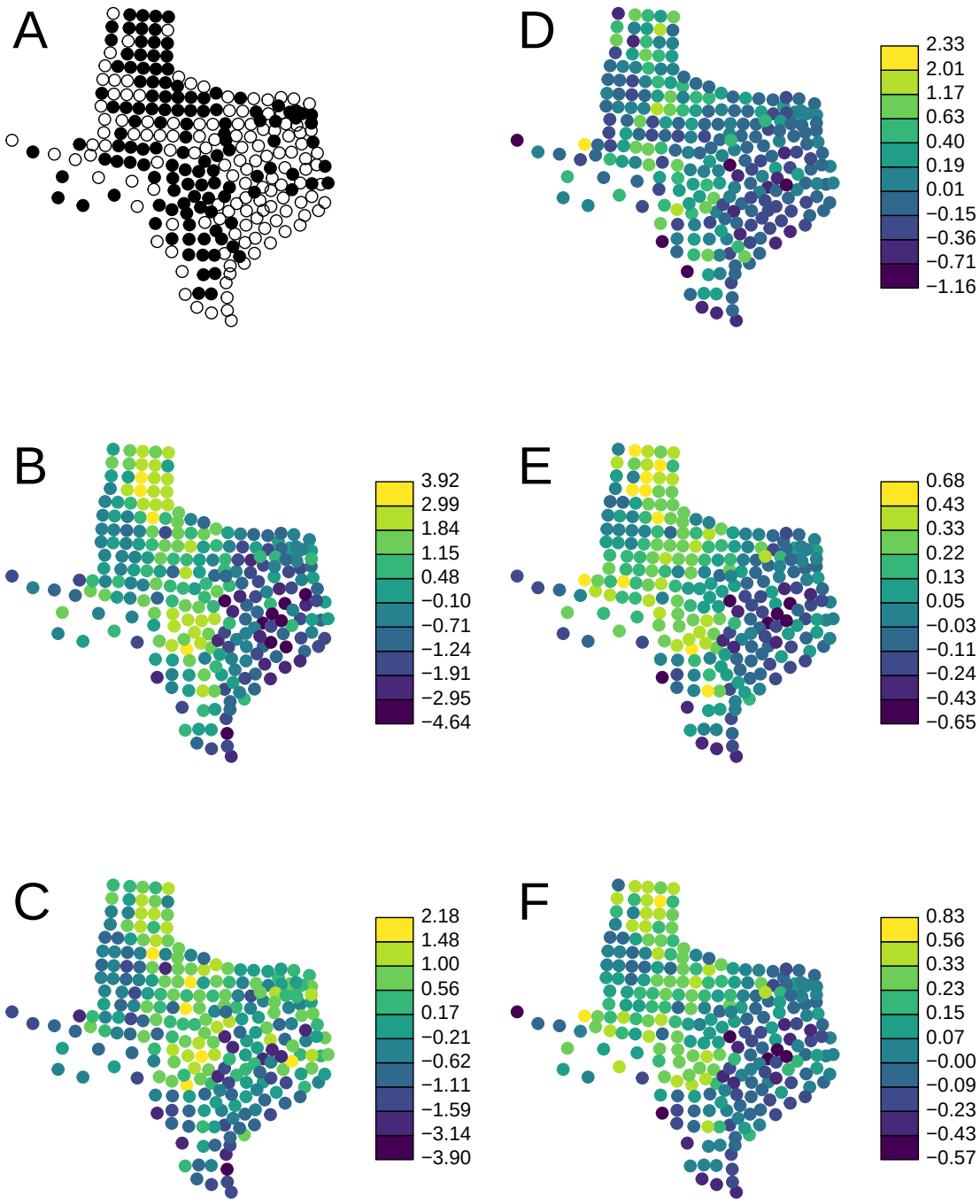

Figure 3: Raw data and predicted spatial random effects ($\mathbf{w}$) for the Texas turnout data. A) raw binary data, where open circles are zeros and solid circles are ones, B) predicted $\hat{\mathbf{w}}$ using SAR model for binary data, C) predicted $\hat{\mathbf{w}}$ using CAR model for binary data, D) logit-transformed proportional turnout data, E) predicted $\hat{\mathbf{w}}$ using SAR model for proportional turnout data, F) predicted $\hat{\mathbf{w}}$ using CAR model for proportional turnout data.

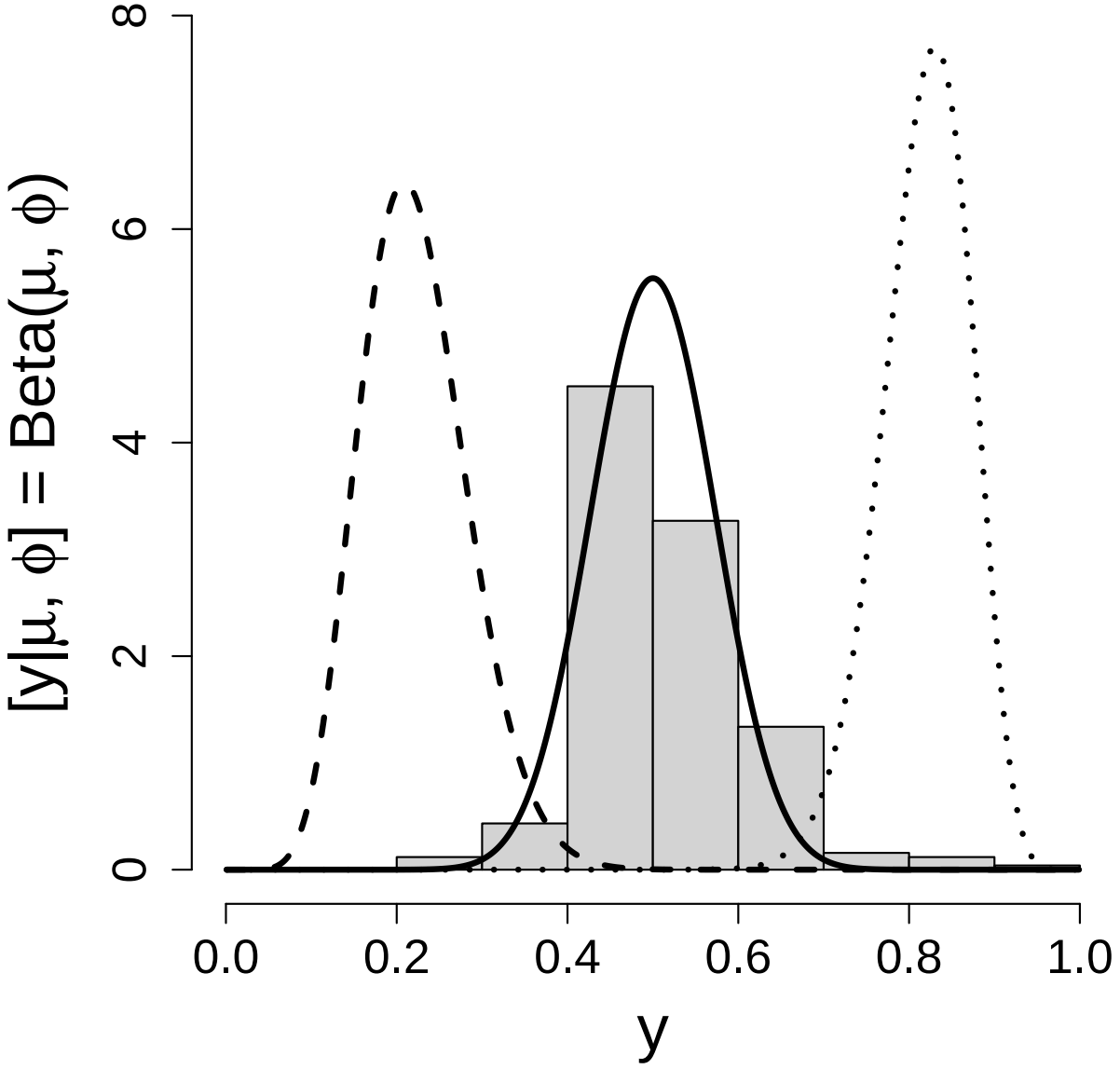

Figure 4: Histogram of proportional turnout and fitted probability density functions for a beta distribution with $\phi = 46.9$ at μ values of 0.3 (dashed line), 0.5 (solid line), and 0.8 (dotted line).

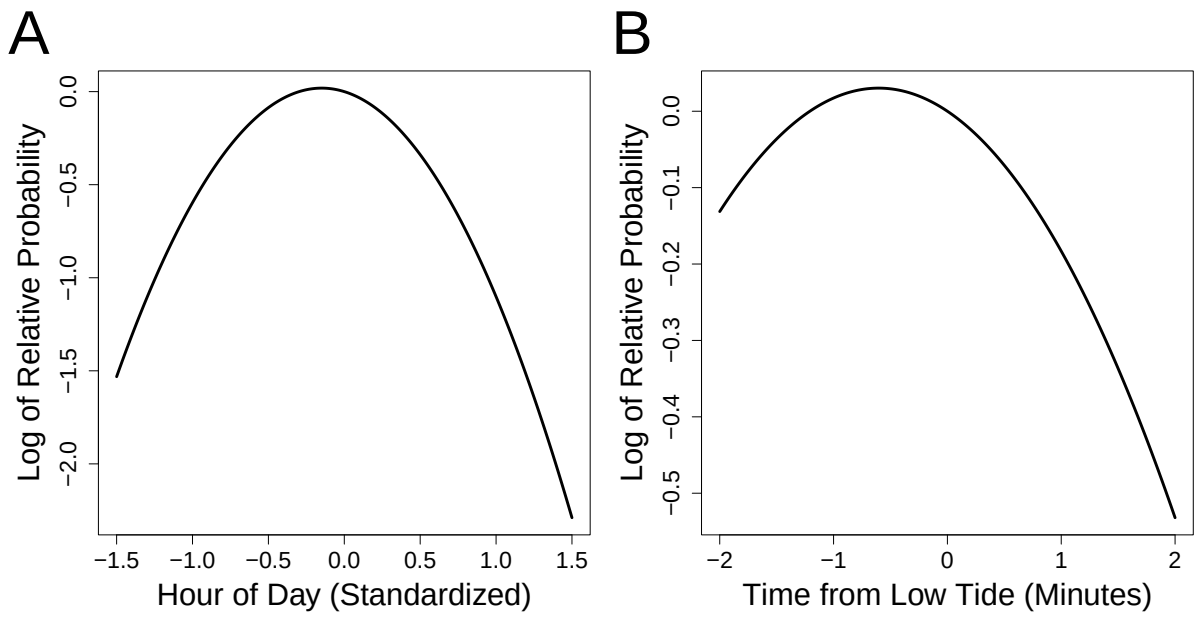

Figure 5: Fitted effects of A) hour-of-day and B) time-from-low-tide on harbor seal counts. The fitted effect shows the log of the expected proportional change from the zero value of all covariates.

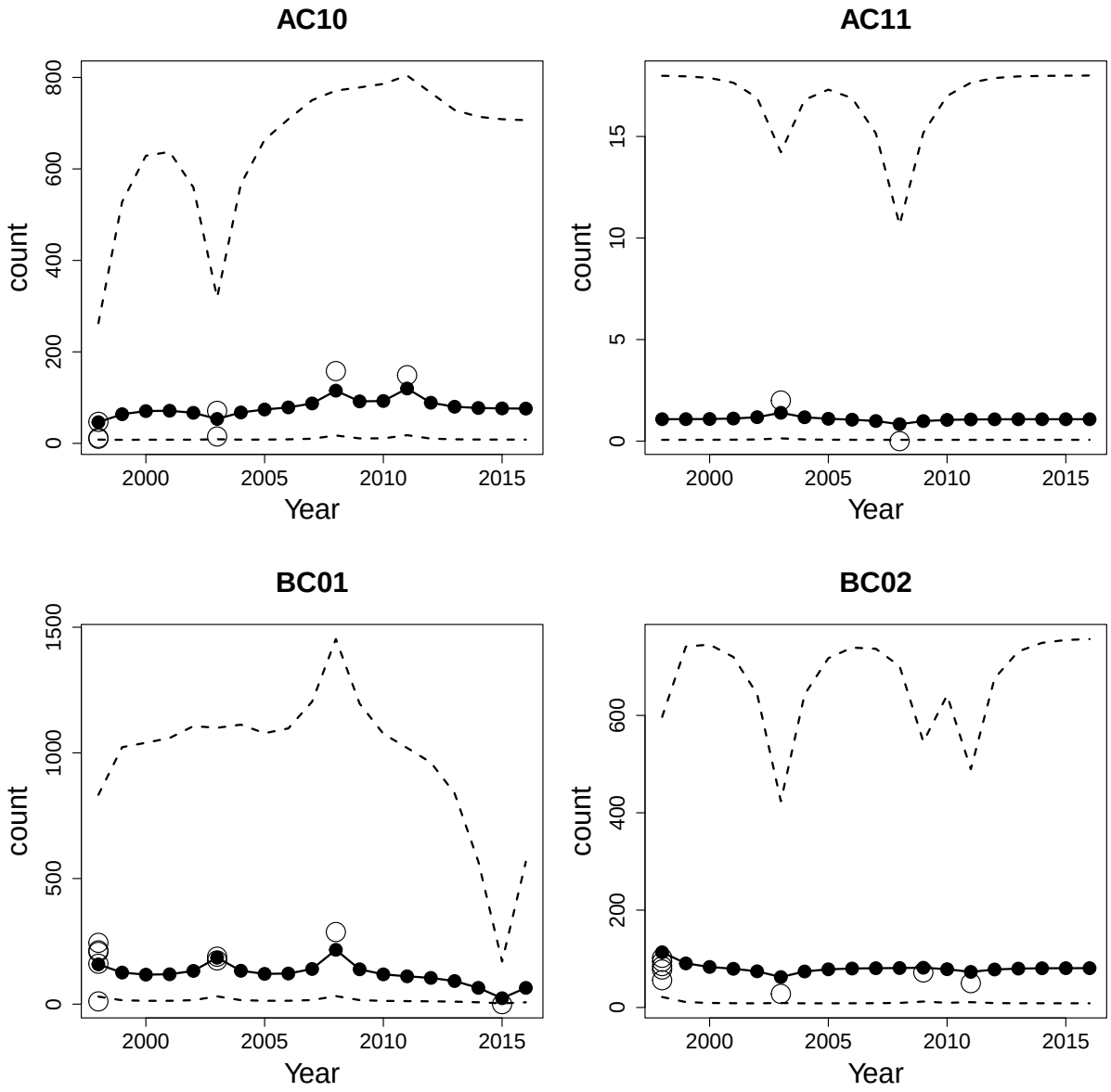

Figure 6: Predicted w-values for 4 of the 74 sites. Open circles are raw counts, and solid circles are predicted w-values connected by a solid line. The dashed line shows the prediction intervals.

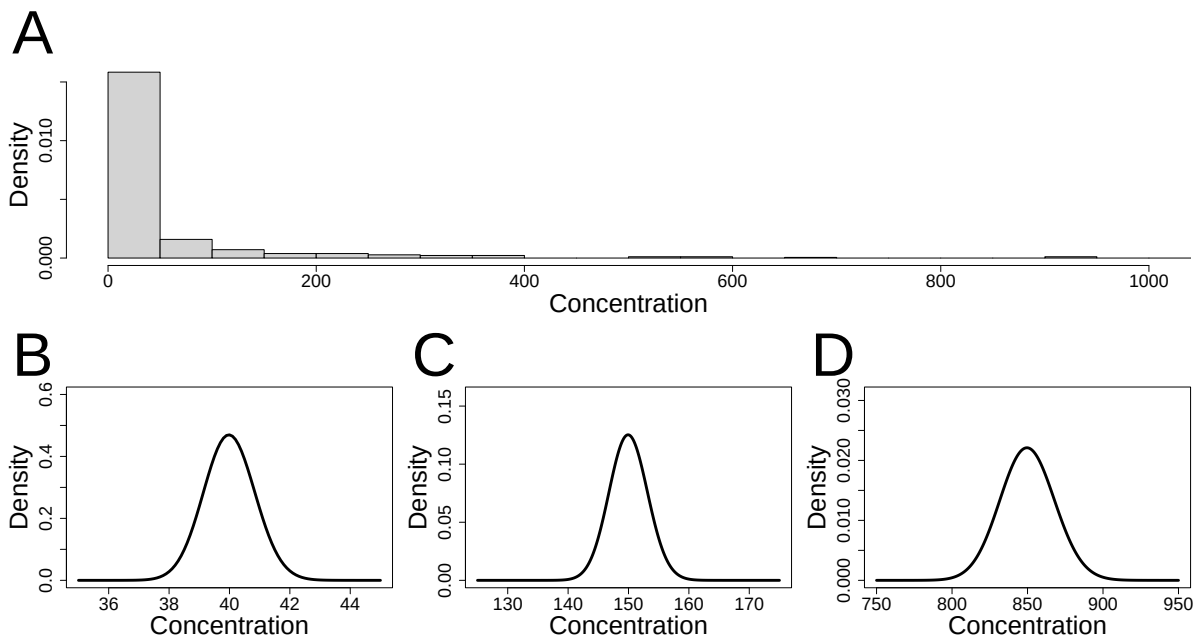

Figure 7: A) Histogram of lead concentration in moss, B) fitted probability density at $\mu = 40$ with $\phi = 2218$ for the gamma distribution (solid line) C) $\mu = 150$, and D) $\mu = 850$.

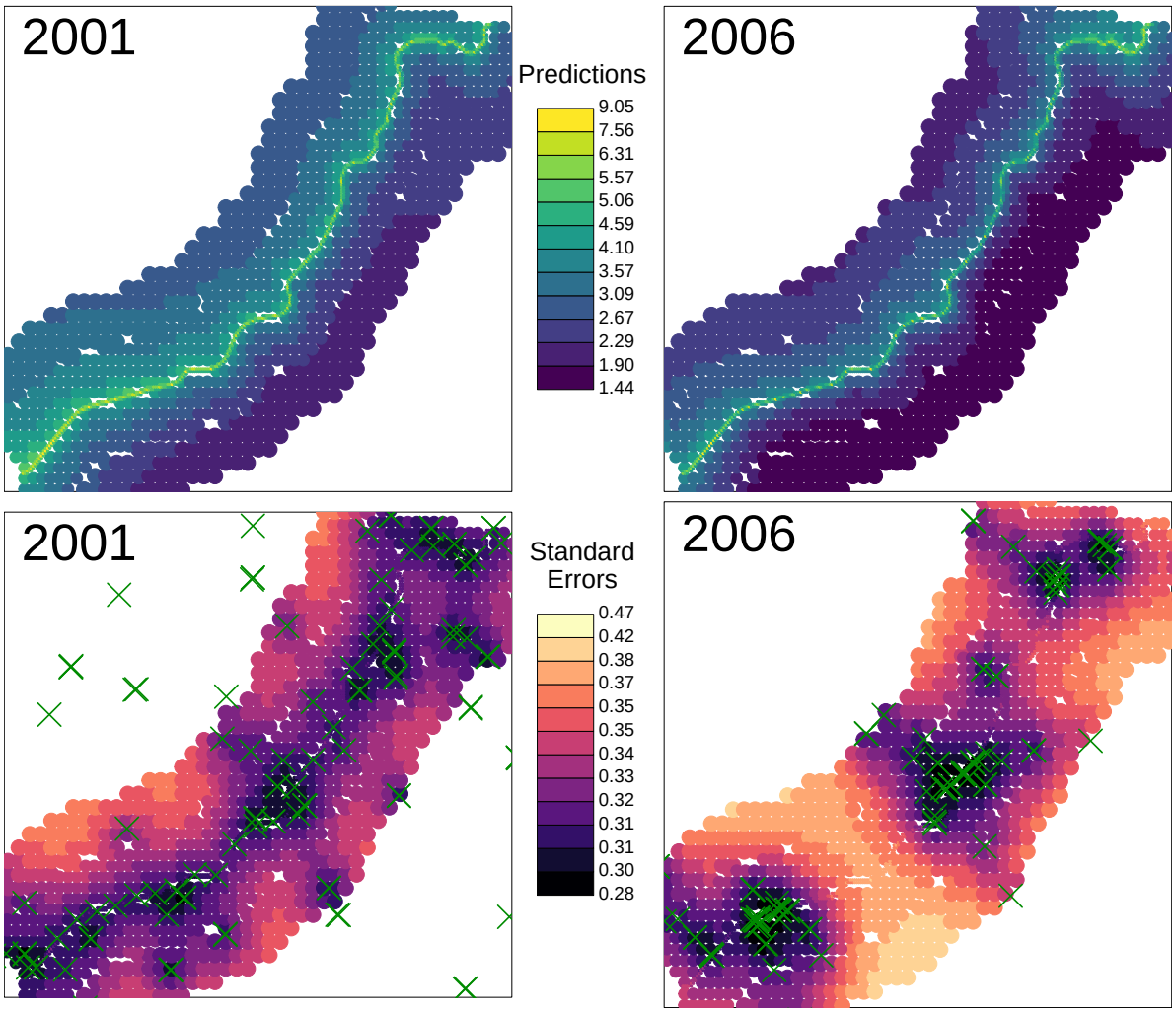

Figure 8: Prediction and their standard errors for 2001 and 2006 at locations near the haul road through Cape Krusenstern National Park, Alaska. The green × symbols show sample locations.

6 APPENDIX

6.1 Derivation of REML from Integration

Consider a multivariate normal distribution for a general linear model,

$$[\mathbf{y};\boldsymbol{\beta},\boldsymbol{\theta}] = \frac{\exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right)}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}}, \tag{A.1}$$

where $\mathbf{y}$ is an $n\times 1$ vector for the response variable, $\mathbf{X}$ is a $n\times p$ design matrix of explanatory variables, $\boldsymbol{\beta}$ is a $p\times 1$ vector of fixed effects, $\boldsymbol{\theta}$ contains covariance parameters contained in the $n\times n$ covariance matrix $\boldsymbol{\Sigma}$. It is possible to obtain REML equations by integrating out the fixed effects $\boldsymbol{\beta}$,

$$\int_{\mathbb{R}^p} f(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\theta})d\boldsymbol{\beta},$$

to obtain a likelihood that is a function of just the covariance parameters $\boldsymbol{\theta}$ and the data $\mathbf{y}$. In particular

$$-2\ln\left(\int_{\mathbb{R}^p} f(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\theta})d\boldsymbol{\beta}\right) = (n-p)\ln(2\pi) + \ln|\boldsymbol{\Sigma}| + \ln|\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}| + (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}),$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y}$.

Proof

Write (A.1) as

$$\begin{aligned}[\mathbf{y};\boldsymbol{\beta},\boldsymbol{\theta}] &= \frac{\exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta})\right)}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}},\\ &= \frac{\exp\left(-\frac{1}{2}[(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) + (\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta}) + C\right)}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}},\end{aligned}$$

where $C = 2(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}) = 0$. Factor out terms that do not contain $\boldsymbol{\beta}$,

$$\int_{\mathbb{R}^p} [\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\theta}] d\boldsymbol{\beta} = M \int_{\mathbb{R}^p} \exp\left(-\frac{1}{2}(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta})\right) d\boldsymbol{\beta},$$

where $M = \exp[-\frac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})]/[(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}]$. Notice that

$$\begin{aligned}
&\int_{\mathbb{R}^p} \exp\left(-\frac{1}{2}(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta})\right) d\boldsymbol{\beta}, \\
=&\int_{\mathbb{R}^p} \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right) d\boldsymbol{\beta}, \\
=&2\pi^{p/2}|(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})|^{-1/2},
\end{aligned}$$

by recalling that, for positive definite $\mathbf{A}_{m\times m}$ and any conformable $\mathbf{x} \neq \mathbf{0}$,

$$\int_{\mathbb{R}^m} \exp(-\mathbf{x}'\mathbf{A}\mathbf{x}/2) d\mathbf{x} = (2\pi)^{m/2}|\mathbf{A}|^{-1/2}.$$

Hence, we arrive at

$$[\mathbf{y}; \boldsymbol{\theta}] = \int_{\mathbb{R}^p} [\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\theta}] d\boldsymbol{\beta} = \frac{\exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\right)}{(2\pi)^{(n-p)/2}|\boldsymbol{\Sigma}|^{1/2}|\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}|^{1/2}},$$

and taking $-2\ln[\mathbf{y}; \boldsymbol{\theta}]$ we obtain the desired result.

6.2 Distribution Parameterizations

6.2.1 Negative Binomial Distribution

For the negative binomial, y_i is a non-negative integer with probability density function (PDF)

$$[y|\mu,\phi] = \frac{\Gamma(y+\phi)}{\Gamma(\phi)y!}\left(\frac{\mu}{\mu+\phi}\right)^{y}\left(\frac{\phi}{\mu+\phi}\right)^{\phi},$$

where $0 < \mu < 1$, $0 < \phi$, $\mathrm{E}(Y) = \mu$, $\mathrm{var}(Y) = \mu + \mu^2/\phi$, and $\Gamma(\cdot)$ is the gamma function.

6.2.2 Gamma Distribution

For the gamma distribution, y_i is positive with PDF

$$[y|\mu,\phi] = \frac{1}{\Gamma(\phi)}\left(\frac{\phi}{\mu}\right)^{\phi} y^{\phi-1}\exp\left(\frac{-y\phi}{\mu}\right),$$

where $0 < \mu$, $0 < \phi$, , $\mathrm{E}(Y) = \mu$, and $\mathrm{var}(Y) = \mu^2/\phi$.

6.2.3 Beta Distribution

For the beta distribution, $0 < y_i < 1$ with PDF

$$[y|\mu,\phi] = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},$$

where $0 < \mu < 1$, $0 < \phi$, $\mathrm{E}(Y) = \mu$, and $\mathrm{var}(Y) = \mu(1-\mu)/(1+\phi)$.

6.2.4 Inverse Gaussian Distribution

The inverse Gaussian distribution is usually written as,

$$[y;\mu,\lambda] = \sqrt{\frac{\lambda}{2\pi y^3}}\exp\left(-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right), \tag{A.2}$$

where $y > 0$, $\mu > 0$, and $\lambda > 0$. In this parameterization λ is a shape parameter, and $\mathrm{E}(Y) = \mu$ and $\mathrm{var}(Y) = \mu^3/\lambda$. In order to keep μ positive and w unconstrained in (1), we let $\boldsymbol{\mu} = \exp(\mathbf{w})$. However, under this construction, from (7), we obtain

$$D_{i,i} = \frac{(e^{w_i} - 2y_i)}{\phi e^{2w_i}},$$

and some $D_{i,i}$ can be positive whenever $e^{w_i} > 2y_i$, which can lead to $\mathbf{H}$ in (8) being singular. We propose an alternative parameterization. For inverse Gaussian models, λ is often scaled, and here we do so by taking $\phi = \lambda/\mu = \lambda/\exp(w)$, yielding a μ-scaled-λ inverse Gaussian model,

$$[y; \mu, \lambda] = \sqrt{\frac{\phi \exp(w)}{2\pi y^3}} \exp\left(-\frac{\phi(y - \exp(w))^2}{2\exp(w)y}\right), \tag{A.3}$$

where $\phi > 0$ and now $\mathrm{var}(Y) = \mu^2/\phi$. Under this parameterization, we have

$$D_{i,i} = -\frac{\phi(e^{2w_i} + y_i^2)}{2ye^{w_i}},$$

which is always negative, and so (8) is always well-behaved. Under this construction, we also have

$$d_i = \phi\left(\frac{y}{2e^{w_i}} - \frac{e^{w_i}}{2y}\right) + \frac{1}{2}.$$