Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

Spatial Generalized Linear Models in R Using spmodel

Michael Dumelle
United States Environmental Protection Agency

Jay M. Ver Hoef
Alaska Fisheries Science Center

Matt Higham
St. Lawrence University

Abstract

Generalized linear models (GLMs) describe a non-normal response variable that may be binary, count, skewed, or a proportion. Typically, observations in a GLM are assumed independent of one another. For spatial data, this independence assumption is impractical, as nearby locations tend to be more similar than locations far apart. The **spmodel** R package provides tools to fit GLMs that incorporate spatial correlation (i.e., spatial generalized linear models, or SPGLMs). SPGLMs are fit in **spmodel** using a novel application of the Laplace approximation via `spglm()` for point-referenced data or `spgautor()` for areal (i.e., lattice), data. `spglm()` and `spgautor()` closely resemble glm from base R but include arguments that control the spatial correlation structure. **spmodel** has many helper functions for model inspection and diagnostics, some of which leverage other R packages like broom and emmeans. **spmodel** has tools to make predictions of the latent spatial-mean process at unobserved locations. **spmodel** also provides many advanced features like accommodating geometric anisotropy and nonspatial random effects, simulating spatially autocorrelated data, and more. Here we use **spmodel** to illustrate the modeling of binary, count, skewed and proportion response variables from several point-referenced and areal data sets.

Keywords: autoregressive model, geostatistical model, spatial covariance, spatial correlation.

1. Introduction

In practice, non-Gaussian data are ubiquitous. Non-Gaussian that belong to an exponential family data can be naturally modeled using a generalized linear model (GLM) regression framework (Nelder and Wedderburn 1972; McCullagh and Nelder 1989; Myers, Montgomery, Vining, and Robinson 2012; Faraway 2016). In a GLM, an $n \times 1$ response variable $\mathbf{y}$ belongs to a a statistical distribution (e.g., Poisson, Binomial) with some mean and variance. Often, the analysis goal is to study the impact of a linear function of several explanatory variables on y through a GLM. In this context, the latent (i.e., unobserved) mean of $\mathbf{y}$, $\boldsymbol{\mu}$, is linked to these explanatory variables via a link function:

$$f(\boldsymbol{\mu}|\mathbf{X}, \boldsymbol{\beta}) \equiv \mathbf{w} = \mathbf{X}\boldsymbol{\beta}, \tag{1}$$

where for a sample size n, $f(\cdot)$ is a link function that connects $\boldsymbol{\mu}$ to $\mathbf{w}$, $\mathbf{X}$ is the $n \times p$ design matrix of explanatory variables, and $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed effects. While the mean is typically constrained in some way (e.g., between zero and one if a probability), the link function generally makes $\mathbf{w}$ is unconstrained. Common link functions inlude the log odds (i.e., logit) link for binary and proprtion data and the log link count and skewed data. Equation 1 can also be written in terms of the inverse link function, $f^{-1}(\cdot)$:

$$\boldsymbol{\mu}|\mathbf{X}, \boldsymbol{\beta} \equiv f^{-1}(\mathbf{w}) = f^{-1}(\mathbf{X}\boldsymbol{\beta}), \tag{2}$$

The GLM fixed effects ($\boldsymbol{\beta}$) are typically estimated via maximum likelihood (Chambers and Hastie 1992). It is often convenient to compute the maximum likelihood estimates using the iteratively reweighted least squares (IRWLS) algorithm (Wood 2017), which is the approach used by the `glm()` function in the R programming language (R Core Team 2024). GLMs add an additional layer of complexity compared to linear regression models, as the left-hand size of Equation 1 is a function of the mean of $\mathbf{y}$ rather than $\mathbf{y}$ itself (as in linear regression models).

The standard GLM assumes the elements of $\mathbf{y}$ are independent. This independence assumption is typically impractical for spatial data. In spatial data, nearby observations tend to be more similar than distant observations (Tobler 1970), leading to positive spatial covariance among observations. The consequences of ignoring spatial covariance in statistical models for spatial data can be severe and include imprecise parameter estimates as well as misleading standard errors that inflate Type-I error rates and decrease power (Zimmerman and Ver Hoef 2024).

An approach for handling spatial data using a GLM is to assume $\mathbf{w}$ has spatial covariance. This is achieved by adding to Equation 1 two random effects, $\boldsymbol{\tau}$ and $\boldsymbol{\epsilon}$. The random effect $\boldsymbol{\tau}$ is an $n \times 1$ column vector of spatially dependent random errors. We assume that $\text{E}(\boldsymbol{\tau}) = \mathbf{0}$ and $\text{Cov}(\boldsymbol{\tau}) = \sigma^2_\tau \mathbf{R}$, where $\text{E}(\cdot)$ and $\text{Cov}(\cdot)$ denote expectation and covariance, respectively. The variance parameter σ^2_τ controls the magnitude of spatial covariance and is often called a partial sill, while the matrix $\mathbf{R}$ is an $n \times n$ spatial correlation matrix that depends on a range parameter controls the distance-decay rate of the spatial correlation. One example of a spatial covariance matrix is the "exponential", which is given by

$$\text{Cov}(\boldsymbol{\tau}) = \sigma^2_{de} \exp(-\mathbf{H}/\phi), \tag{3}$$

where $\mathbf{H}$ is a matrix of pairwise distances among the elements of $\mathbf{y}$ and ϕ is a range parameter. From Equation 3, as the distance between two elements of $\mathbf{y}$ increases, the spatial

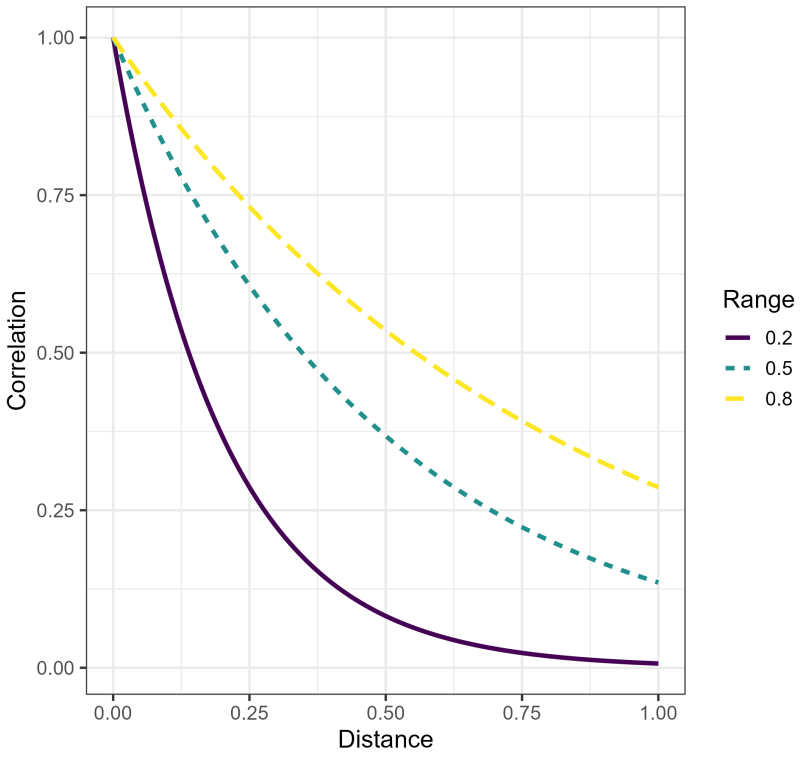

Figure 1: An exponential spatial correlation function with varying range parameters.

covariance decreases, which reflects intuition. Moreover, as the range parameter, ϕ, increases, the strength of spatial dependence increases (Figure 1). The random effect $\boldsymbol{\epsilon}$ is an $n \times 1$ column vector of independent random errors. We assume that $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\tau}) = \sigma^2_\epsilon \mathbf{I}$, where $\mathbf{I}$ is an $n \times n$ identity matrix. The variance parameter σ^2_ϵ controls the magnitude of nonspatial variability (i.e., fine-scale variation) and is often called a nugget.

Through inclusion of $\boldsymbol{\tau}$ and $\boldsymbol{\epsilon}$, the spatial GLM (SPGLM) can be written as

$$f(\boldsymbol{\mu}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\epsilon}) \equiv \mathbf{w} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\tau} + \boldsymbol{\epsilon}. \tag{4}$$

Often in spatial statistics, quantities are explicitly referenced with respect to $\mathbf{s}$, a vector of coordinates indexing the observation (Cressie 1993). For example, $\mathbf{y}$ and $\mathbf{X}$ may instead be written $\mathbf{y}(\mathbf{s})$ and $\mathbf{X}$, respectively. We acknowledge the utility of this nomenclature but drop the explicit dependence on $\mathbf{s}$ for simplicity of notation. Assuming independence among $\boldsymbol{\tau}$ and $\boldsymbol{\epsilon}$, it follows that

$$\mathrm{Cov}(\boldsymbol{\tau} + \boldsymbol{\epsilon}) = \mathrm{Cov}(\boldsymbol{\tau}) + \mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2_\tau \mathbf{R} + \sigma^2_\epsilon \mathbf{I}. \tag{5}$$

To better align with intuition, we henceforth σ^2_τ as σ^2_{de} (for spatial error variance) and σ^2_ϵ as σ^2_{ie} (for independent error variance). The parameters σ^2_{de}, σ^2_{ie}, the range parameter ϕ in $\mathbf{R}$, and any other parameters in $\mathbf{R}$ compose $\boldsymbol{\theta}$, the covariance parameter vector.

Fitting and using SPGLMs is challenging both conceptually and computationally (Bolker, Brooks, Clark, Geange, Poulsen, Stevens, and White 2009). Recently, however, there have been numerous, significant advances in R software that have made these models more accessible to practitioners. The **brms** (Bürkner 2017), **carBayes** (Lee 2013), **ngspatial** (Hughes and

Cui 2020), **R-INLA** (Lindgren and Rue 2015) and **inlabru** (Bachl, Lindgren, Borchers, and Illian 2019), **spBayes** (Finley, Banerjee, and Carlin 2007), **spOccupancy** (Doser, Finley, Kéry, and Zipkin 2022), **spAbundance** (Doser, Finley, Kéry, and Zipkin 2024), and **spNNGP** (Finley, Datta, and Banerjee 2022) packages take a Bayesian approach, either directly sampling from posterior distributions of parameters (e.g., using MCMC) or approximating them. A benefit of Bayesian approaches is that prior information can be incorporated and uncertainty quantification of parameter estimates is straightforward. However, Bayesian approaches, especially those using MCMC, can be computationally expensive. In order to reduce computation time, many of these packages work with the precision matrix instead of the covariance matrix so that computationally expensive matrix inversion is not required. For example, **R-INLA** uses the precision matrix and tends to be very fast. Working with precision matrices, however, can be more restrictive and less intuitive than working directly with the covariance matrix. The FRK (Sainsbury-Dale, Zammit-Mangion, and Cressie 2024), **glmmTMB** (Brooks, Kristensen, van Benthem, Magnusson, Berg, Nielsen, Skaug, Maechler, and Bolker 2017), **hglm** (Ronnegard, Shen, and Alam 2010), **mgcv** (Wood 2017), and **spaMM** (Rousset and Ferdy 2014) packages directly use Laplace, quasi-likelihood, or reduced-rank approaches to estimate parameters. These direct approaches tend to be computationally efficient, as they don't rely on MCMC sampling. In contrast to the Bayesian approach, a drawback of these direct approaches is that prior information cannot be formally incorporated and covariance parameter uncertainty is more challenging to quantify. The **sdmTMB** (Anderson, Ward, English, Barnett, and Thorson 2024) package combines elements of **R-INLA**, **glmmTMB**, and properties of Gaussian Markov random fields to fit a wide variety of SPGLMs, and **tinyVAST** (Thorson, Anderson, Goddard, and Rooper 2025) extends some of these models to multivariate or (dynamic) structural equation models.

Ver Hoef, Blagg, Dumelle, Dixon, Zimmerman, and Conn (2024) proposed a novel approach to fitting SPGLMs that leverages the Laplace approximation while marginalizing over both the latent $\mathbf{w}$ and the fixed effects ($\boldsymbol{\beta}$) and accommodating spatial covariance. Ver Hoef *et al.* (2024) showed that this approach performed efficiently in a variety of simulation settings, generally having appropriate confidence interval coverage for the fixed effects and prediction interval coverage for new $\mathbf{w}$. The approach performed similarly to the Bayesian SPGLM approach in **spBayes** and the automatic differentiation SPGLM approach in **glmmTMB** but was much faster. At small sample sizes, the approach outperformed the approximate Bayesian SPGLM approach in **R-INLA** and had similar computational times. For moderate sample sizes, it performed similarly to **R-INLA**, though **R-INLA** was faster. This novel approach is particularly attractive for two reasons. First, it is general enough that can be applied to any covariance structure (not just spatial). Second, after estimating the covariance parameters, analytical solutions exist for the fixed effects (and their standard errors) as well as predictions of the latent $\mathbf{w}$ at new locations (and their standard errors). The **spmodel** R package (Dumelle, Higham, and Ver Hoef 2023) recently provided full support for the methods in Ver Hoef *et al.* (2024) applied to binary, count, skewed, and proportion data for over 20 different spatial covariance types.

The **spmodel** R package (Dumelle *et al.* 2023) recently provided a full set of modeling tools for SPGLMs fit using the methods described in Ver Hoef *et al.* (2024). These modeling tools are approachable and mirror the familiar `glm()` syntax from base-R, making the transition from GLMs to SPGLMs relatively seamless. The `spglm()` function fits SPGLMs for point-referenced data (e.g., x-coordinates and y-coordinates representing point locations in a

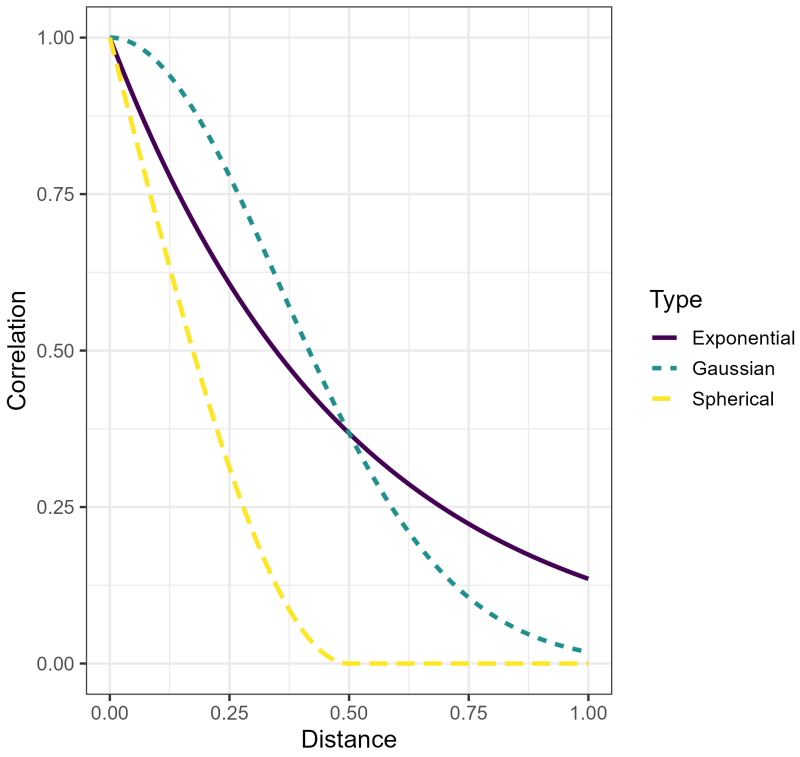

Figure 2: Exponential, Gaussian, and spherical spatial correlation functions all with range parameters equal to 0.5.

field), while the `spgautor()` function fits SPGLMs for areal data (e.g., polygon boundaries representing geographic subsets of a region). **spmodel** supports the binomial distribution for binary data, Poisson and negative binomial distributions for count data, Gamma and inverse Gaussian distributions for skewed data, and the beta distribution for proportion data. There are 20 different spatial covariance structures available including the exponential, Gaussian, and spherical for point-referenced data (Figure 2) and the conditional autoregressive, and simultaneous autoregressive structures for areal data. **spmodel** provides tools for commonly used model summaries, visualizations, and diagnostics (e.g., Cook's distance) using standard R helper functions like `summary()`, `plot()`, and `cooks.distance()`. **spmodel** also provides tools to predict $\mathbf{w}$ at new locations and quantify uncertainty in those prediction using `predict()`. This core functionality, combined with several advanced features we describe throughout the manuscript, enable **spmodel** to provide some novel and important capabilities previously missing from the existing SPGLM ecosystem in R.

spmodel (version 0.11.0) is arguably most similar to **sdmTMB** (version 0.7.4) in terms of scope and feel. Both packages use similar syntax as `glm()`, accommodate flexible `formula` arguments (e.g., offsets, splines), handle spatial covariance that decays at different rates in different rates (i.e., geometric anisotropy), incorporate nonspatial random effects, support other R packages for modeling like **broom** (Robinson, Hayes, and Couch 2021; Kuhn and Silge 2022), **emmeans** (Lenth 2024), and **car** (Fox and Weisberg 2019), and have tools for model summaries, prediction, and simulating data. There are some notable differences between the two packages, however. **sdmTMB** supports several additional GLM distributions like the

Tweedie, supports Hurdle models, and can incorporate prior information through Bayesian applications. **sdmTMB** also provides tools for working with temporal data and enhanced visualizations of marginal effects. **sdmTMB** does require a preprocessing step of constructing a mesh for the stochastic partial differential equation approach, and the density of the mesh can affect model results and computational complexity. **spmodel** does not require the construction of a mesh prior to modeling. **spmodel** supports 20 different spatial covariances and models them directly, rather than using a precision matrix approximation to the Matérn spatial covariance as in **sdmTMB**. **spmodel** also provides experimental design tools (e.g., analysis of variance, contrasts), supports **sf** objects in modeling and prediction functions (Pebesma 2018), has several specialized model diagnostics like leverage values and Cook's distances, and has analytic solutions for prediction standard errors. Other similarities and differences do exist between **sdmTMB** and **spmodel**, and both packages continue to evolve. Overall, we believe that these packages are complementary and enhance the suite of SPGLM tools accessible to practitioners.

The rest of this article is organized as follows. In Section 2, we provide some background for the SPGLM fitting and prediction routines in **spmodel**. In Section 3, we provide several applications of **spmodel** to spatial binary, count, skewed and proportion data with both point-referenced and areal supports. And in Section **??**, we end with a discussion synthesizing **spmodel**'s contributions to the analysis of SPGLMs in R.

2. The spatial generalized linear model and marginalizatoin

spmodel implements the novel methods described in Ver Hoef *et al.* (2024) to fit SPGLMs, which leverages the Laplace approximation and marginalizes over both the latent $\mathbf{w}$ and the fixed effects while accommodating spatial covariance. A beneficial aspect of this approach is that it formally maximizes a hierarchical GLM likelihood (Lee and Nelder 1996; Wood 2017). This makes likelihood-based statistics for model comparison like AIC (Akaike 1974), AICc (Hoeting, Davis, Merton, and Thompson 2006), BIC (Schwarz 1978), deviance (McCullagh and Nelder 1989), and likelihood ratio tests available. These types of statistics are not available for quasi-likelihood (Wedderburn 1974; Breslow and Clayton 1993) or pseudo-likelihood approaches (Wolfinger and O'connell 1993), which only specify the first two moments of a distribution. Ver Hoef *et al.* (2024) provides thorough details regarding the method and contextualizes its development which built upon similar methods (Evangelou, Zhu, and Smith 2011, Bonat and Ribeiro Jr (2016)). Next, we describe a brief overview of the approach and how it can be used for parameter estimation, inference, and prediction.

2.1. Formulating the hierarchical likelihood

We can write the SPGLM likelihood hierarchically as

$$[\mathbf{y}|\mathbf{X}, \varphi, \boldsymbol{\theta}] = \int_{\mathbf{w}} \int_{\boldsymbol{\beta}} [\mathbf{y}|f^{-1}(\mathbf{w}), \varphi][\mathbf{w}|\mathbf{X}, \boldsymbol{\theta}] d\boldsymbol{\beta} d\mathbf{w}, \tag{6}$$

where $[\mathbf{y}|f^{-1}(\mathbf{w}), \varphi]$ is the density for the appropriate response distribution of $\mathbf{y}$ (e.g., binomial, Poisson) given the latent $\mathbf{w}$ and dispersion parameter (φ), and $[\mathbf{w}|\mathbf{X}, \boldsymbol{\theta}]$ is the multivariate Gaussian density for $\mathbf{w}$ given the explanatory variables ($\mathbf{X}$), fixed effects ($\boldsymbol{\beta}$), and spatial covariance parameters ($\boldsymbol{\theta}$). The elements of $[\mathbf{y}|f^{-1}(\mathbf{w}), \varphi]$ are conditionally indepen-

dent (given $\mathbf{w}$), but the elements of $[\mathbf{w}|\mathbf{X}, \boldsymbol{\theta}]$ share spatial covariance. Following Harville (1977), we can integrate $\boldsymbol{\beta}$ out of Equation 4, which yields

$$[\mathbf{y}|\mathbf{X}, \varphi, \boldsymbol{\theta}] = \int_{\mathbf{w}} [\mathbf{y}|f^{-1}(\mathbf{w}), \varphi][\mathbf{w}|\mathbf{X}, \boldsymbol{\theta}]d\mathbf{w}, \tag{7}$$

where $[\mathbf{w}|\mathbf{X}, \boldsymbol{\theta}]$ is the restricted (i.e., residual) multivariate Gaussian density (Patterson and Thompson 1971) for $\mathbf{w}$ given the explanatory variables and covariance parameters. Equation 7 can be synonymous written after profiling the overall variance out of $\boldsymbol{\Sigma}$, which reduces the dimension of $\boldsymbol{\theta}$ by one for optimization (Wolfinger, Tobias, and Sall 1994). The restricted multivariate Gaussian density is given by

$$[\mathbf{w}|\mathbf{X}, \boldsymbol{\theta}] = \frac{\exp(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})^T)}{(2\pi)^{(n-p)/2}|\boldsymbol{\Sigma}|^{1/2}|\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}|^{1/2}}, \tag{8}$$

where $\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{w}$ and $|\cdot|$ denotes the determinant. Next, let

$$\ell_{\mathbf{w}} = \log([\mathbf{y}|f^{-1}(\mathbf{w}), \varphi][\mathbf{w}|\mathbf{X}, \boldsymbol{\theta}]) \tag{9}$$

and rewrite Equation 7 as

$$[\mathbf{y}|\mathbf{X}, \varphi, \boldsymbol{\theta}] = \int_{\mathbf{w}} \exp(\ell_{\mathbf{w}})d\mathbf{w}. \tag{10}$$

A second-order Taylor series expansion of $\ell_{\mathbf{w}}$ around $\hat{\mathbf{w}}$ yields

$$[\mathbf{y}|\mathbf{X}, \varphi, \boldsymbol{\theta}] \approx \int_{\mathbf{w}} \exp(\ell_{\hat{\mathbf{w}}} + \mathbf{g}^T(\mathbf{w} - \hat{\mathbf{w}}) + \frac{1}{2}(\mathbf{w} - \hat{\mathbf{w}})^T\mathbf{G}(\mathbf{w} - \hat{\mathbf{w}}))d\mathbf{w}, \tag{11}$$

where $\mathbf{g}$ and $\mathbf{G}$ are the gradient and Hessian, respectively, of $\ell_{\mathbf{w}}$ with respect to $\mathbf{w}$. If $\hat{\mathbf{w}}$ is a value for which $\mathbf{g} = \mathbf{0}$,

$$[\mathbf{y}|\mathbf{X}, \varphi, \boldsymbol{\theta}] \approx \exp(\ell_{\hat{\mathbf{w}}}) \int_{\mathbf{w}} \exp(-\frac{1}{2}(\mathbf{w} - \hat{\mathbf{w}})^T(-\mathbf{G})(\mathbf{w} - \hat{\mathbf{w}}))d\mathbf{w}. \tag{12}$$

The integral in Equation 12 can be solved by leveraging properties of the normalizing constant of a multivariate Gaussian distribution. Thus, rewriting $\exp(\ell_{\hat{\mathbf{w}}})$ yields

$$[\mathbf{y}|\mathbf{X}, \varphi, \boldsymbol{\theta}] \approx [\mathbf{y}|f^{-1}(\hat{\mathbf{w}}), \varphi][\hat{\mathbf{w}}|\mathbf{X}, \boldsymbol{\theta}](2\pi)^{n/2}|-\mathbf{G}_{\hat{\mathbf{w}}}|^{-1/2}. \tag{13}$$

Maximizing the natural logarithm of Equation 13 requires a doubly iterative process over $\boldsymbol{\theta}$ and φ as well as $\mathbf{w}$, eventually yielding the the marginal restricted maximum likelihood estimators $\hat{\varphi}$ and $\hat{\boldsymbol{\theta}}$ and their corresponding values of $\hat{\mathbf{w}}$. Maximizing this log likelihood is a computationally expensive operation that involves repeatedly evaluating $\boldsymbol{\Sigma}^{-1}$, $\mathbf{g}$, and $\mathbf{G}$; see Ver Hoef *et al.* (2024) for more details and forms of $\mathbf{g}$ and $\mathbf{G}$ for various response distributions.

2.2. Estimating fixed effects

Though the fixed effects are integrated out of the likelihood, we can still estimate them using generalized least squares (GLS) principles, a common practice for linear models estimated using restricted maximum likelihood methods. Had we observed $\mathbf{w}$, a GLS estimator for $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\Sigma}^{-1}\mathbf{w} = \mathbf{B}\mathbf{w}, \tag{14}$$

where $\mathbf{B} = (\mathbf{X}^\top\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\Sigma}^{-1}$. However, we only observe $\hat{\mathbf{w}}$, so it is reasonable to define $\hat{\boldsymbol{\beta}} = \mathbf{B}\hat{\mathbf{w}}$. Thus, to derive properties of $\hat{\boldsymbol{\beta}}$ like expectation and variance, we must derive these properties for $\hat{\mathbf{w}}$. To do so, we must condition on $\mathbf{w}$ as if it were observed and invoke properties of the laws of total expectation and variance. Because $\hat{\mathbf{w}}$ was optimized via the likelihood, we assume that given $\mathbf{w}$, $\hat{\mathbf{w}}$ has mean $\mathbf{w}$ and variance approximately equal to $-\mathbf{H}^{-1}$ (the inverse Hessian). It follows that $\mathrm{E}(\hat{\mathbf{w}})$ is given by

$$\mathrm{E}(\hat{\mathbf{w}}) = \mathrm{E}(\mathrm{E}(\hat{\mathbf{w}}|\mathbf{w})) = \mathrm{E}(\mathbf{w}) = \mathbf{X}\boldsymbol{\beta} \tag{15}$$

and $\mathrm{Var}(\hat{\mathbf{w}})$ is given by

$$\begin{aligned}
\mathrm{Var}(\hat{\mathbf{w}}) &= \mathrm{E}(\mathrm{Var}(\hat{\mathbf{w}}|\mathbf{w})) + \mathrm{Var}(\mathrm{E}(\hat{\mathbf{w}}|\mathbf{w})) && (16)\\
&= \mathrm{E}(-\mathbf{H}^{-1}) + \mathrm{Var}(\mathbf{w}) && (17)\\
&= -\mathbf{H}^{-1} + \boldsymbol{\Sigma} && (18)
\end{aligned}$$

Putting this all together, it follows that

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \mathrm{E}(\mathbf{B}\hat{\mathbf{w}}) = \mathbf{B}\mathrm{E}(\hat{\mathbf{w}}) = (\mathbf{X}^\top\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{X}^\top\boldsymbol{\Sigma}^{-1}\mathbf{X})\boldsymbol{\beta} = \boldsymbol{\beta} \tag{19}$$

and

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \mathrm{Var}(\mathbf{B}\hat{\mathbf{w}}) && (20)\\
&= \mathbf{B}\mathrm{Var}(\hat{\mathbf{w}})\mathbf{B}^\top && (21)\\
&= \mathbf{B}(-\mathbf{H}^{-1} + \boldsymbol{\Sigma})\mathbf{B}^\top && (22)\\
&= \mathbf{B} - \mathbf{H}^{-1}\mathbf{B}^\top + \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^\top && (23)\\
&= \mathbf{B} - \mathbf{H}^{-1}\mathbf{B}^\top + (\mathbf{X}^\top\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1} && (24)
\end{aligned}$$

In practice, $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ is estimated by evaluating $\boldsymbol{\Sigma}$ at $\hat{\boldsymbol{\theta}}$, the estimated covariance parameter vector.

These results are important because they justify closed-form solutions for $\hat{\boldsymbol{\beta}}$ and its associated variance. Closed-form solutions are useful because they bypass the need for computationally expensive sampling-based strategies to evaluate the mean and variance of $\hat{\boldsymbol{\beta}}$ – a common technique for other approaches to SPGLMs like Bayesian MCMC.

2.3. Inspecting model diagnostics

Inspecting model diagnostics is an important step of the modeling process that can yield valuable insights into model behavior and unusual observations. Montgomery, Peck, and Vining (2021) contextualize three components of unusual observations: outliers, leverage, and influence. An observation is an outlier if it has an unusual response value relative to expectation. The response GLM residuals simply compare the observation to its fitted latent mean:

$$\mathbf{r}_r = \mathbf{y} - f^{-1}(\hat{\mathbf{w}}) \tag{25}$$

Because observations often have a unique support in a GLM (e.g., only two possible response values for binary data) and the variance of an observation generally depends on its mean, response residuals lack some utility. Deviance residuals are a function of response residuals

that are appropriately scaled to behave more like response residuals in a standard linear model. Deviance residuals are given by

$$\mathbf{r}_d = sign(\mathbf{r}_r)\sqrt{\mathbf{d}}, \tag{26}$$

where $\mathbf{d}$ is a vector of individual deviances. The sum of the squared deviance residuals equals the sum of $\mathbf{d}$. The sum of $\mathbf{d}$ is the deviance of the model fit, which quantifies twice the difference in log likelihoods between the a saturated model that fits every observation perfectly (i.e., $\mathbf{y} = f^{-1}(\hat{\mathbf{w}}_i)$ for all i) and the fitted model. Deviance is often used as a fit statistic; lower values of deviance imply a better model fit. Pearson and standardized residuals are other types of GLM residuals that involve some scaling of the response residuals; the Pearson residuals scale $\mathbf{r}_r$ by the squrae root of $\mathbf{V}$, while the standardized residuals scale the deviance residuals by $\frac{1}{\sqrt{(1-\mathbf{L}_{ii})}}$, where $\mathbf{L}_{ii}$ is the ith diagonal element of the leverage matrix, which we discuss next. An observation has high leverage if its combination of explanatory variables is far away from other observations. In a linear model, the leverage values are the diagonal of the leverage (i.e., projection, hat) matrix, $\mathbf{L} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$. In a GLM, the leverage matrix is given by

$$\mathbf{L} = \mathbf{V}^{1/2}\mathbf{X}(\mathbf{X}^\top\mathbf{V}\mathbf{X})\mathbf{X}^\top\mathbf{V}^{1/2}, \tag{27}$$

where $\mathbf{V}$ is a diagonal matrix with ith diagonal element equal to the variance of the response distribution evaluated at $f^{-1}(\mathbf{w}_i)$ (Faraway 2016). The matrix $\mathbf{V}$ is sometimes called the GLM weight matrix. The larger the ith diagonal element of the hat matrix, the more severe the leverage from the ith observation. An observation is influential if it has a sizeable impact on model fit. Influence is measured using Cook's distance (Cook 1979; Cook and Weisberg 1982), which is given for a GLM by

$$\mathbf{c} = \mathbf{r}_s^2 \frac{diag(\mathbf{L})}{\text{tr}(\mathbf{L})(\mathbf{1} - diag(\mathbf{L}))}, \tag{28}$$

where $\mathbf{r}_s^2$ are the standardized residuals and $diag(\mathbf{L})$ indicates the diagonal elements of the leverage matrix. The larger the ith diagonal element of the hat matrix, the more severe the influence from the ith observation. Montgomery *et al.* (2021) provide guidance for interpreting these types of statistics, including cutoffs to consider when identifying unusual residual, leverage, or influence values.

In a linear model, the R^2 (R-squared) statistic quantifies the proportion of variability in the data captured by the explanatory variables and is calculated as one minus the ratio of the error sum of squares to the total sum of squares (Rencher and Schaalje 2008). In a GLM, there are many ways to define such a statistic (Smith and McKenna 2013). One such approach is to use one minus the deviance ratio:

$$PR^2 = 1 - \frac{deviance_{fit}}{deviance_{null}}, \tag{29}$$

where $deviance_{fit}$ is the deviance of the fitted model (sometimes called the residual deviance) and $deviance_{null}$ is the deviance of the model taking $\mathbf{X} \equiv \mathbf{1}$, a column of all ones (i.e., an intercept-only model). In practice, $deviance_{null}$ is derived by computing $\hat{\mathbf{w}}$ when $\mathbf{X} \equiv \mathbf{1}$ given $\hat{\boldsymbol{\theta}}$ and $\hat{\varphi}$ from the fitted model. Like the standard R^2, this statistic attempts to capture variability (i.e., deviance) attributable to the explanatory variables. Because the $deviance_{null}$ denominator changes across fitted models (as the values of $\hat{\boldsymbol{\theta}}$ and $\hat{\varphi}$ change), this statistic

should not be used as a model comparison tool. Instead, it should be used as an informative diagnostic tool unique to each model fit.

2.4. Predicting at new locations

We may also predict values of the latent mean (on the link scale) at new locations by leveraging the spatial covariance between observed locations and new locations (spatial prediction is also called Kriging; see Cressie (1990)). Again suppose that we observed $\mathbf{w}$ and we want to make predictions at $\mathbf{u}$, a vector of latent means at the new locations that follows the same SPGLM from Equation~4 with fixed effects design matrix, $\mathbf{X_u}$. The vector $(\mathbf{w}, \mathbf{u})^\top$ has the following properties:

$$\mathrm{E}(\mathbf{w}, \mathbf{u})^\top = (\mathrm{E}(\mathbf{w}), \mathrm{E}(\mathbf{u}))^\top = (\mathbf{X}\boldsymbol{\beta}, \mathbf{X_u}\boldsymbol{\beta})^\top \tag{30}$$

$$\mathrm{Var}(\mathbf{w}, \mathbf{u})^\top = \begin{bmatrix} \mathrm{Var}(\mathbf{w}, \mathbf{w}) & \mathrm{Var}(\mathbf{w}, \mathbf{u}) \\ \mathrm{Var}(\mathbf{u}, \mathbf{w}) & \mathrm{Var}(\mathbf{u}, \mathbf{u}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma}_{\mathbf{wu}} \\ \boldsymbol{\Sigma}_{\mathbf{uw}} & \boldsymbol{\Sigma}_{\mathbf{uu}} \end{bmatrix} \tag{31}$$

Because we have observed $\mathbf{w}$, we may derive the conditional distribution of $\mathbf{u}|\mathbf{w}$, which has the following properties:

$$\mathrm{E}(\mathbf{w}|\mathbf{u}) = \mathbf{X_u}\boldsymbol{\beta} + \boldsymbol{\Sigma}_{\mathbf{u},\mathbf{w}}\boldsymbol{\Sigma}^{-1}(\mathbf{w} - \mathbf{X}\boldsymbol{\beta}) \tag{32}$$

$$\mathrm{E}(\mathbf{w}|\mathbf{u}) = \boldsymbol{\Sigma}_{\mathbf{u},\mathbf{u}} - \boldsymbol{\Sigma}_{\mathbf{u},\mathbf{w}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\mathbf{w},\mathbf{u}} \tag{33}$$

Ver Hoef *et al.* (2024) show how these equations are adjusted to reflect uncertainty in both $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{w}}$ while leveraging the laws of total expectation and variance yet again. They derive the predictor of $\mathbf{u}$, $\hat{\mathbf{u}}$, and its associated variance, given by:

$$\hat{\mathbf{u}} = \mathbf{X_u}\hat{\boldsymbol{\beta}} + \boldsymbol{\Sigma}_{\mathbf{u},\mathbf{w}}\boldsymbol{\Sigma}^{-1}(\hat{\mathbf{w}} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{34}$$

$$\mathrm{Var}(\hat{\mathbf{u}}) = \boldsymbol{\Sigma}_{\mathbf{u},\mathbf{u}} - \boldsymbol{\Sigma}_{\mathbf{u},\mathbf{w}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\mathbf{w},\mathbf{u}} + \mathbf{K}(\mathbf{X}^\top\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{K}^\top + \boldsymbol{\Lambda}(-\mathbf{H})^{-1}\boldsymbol{\Lambda}^\top, \tag{35}$$

where $\mathbf{K} = \mathbf{X_u} - \boldsymbol{\Sigma}_{\mathbf{u},\mathbf{w}}\boldsymbol{\Sigma}^{-1}\mathbf{X}$ and $\boldsymbol{\Lambda} = \mathbf{X_u}\mathbf{B} + \boldsymbol{\Sigma}_{\mathbf{u},\mathbf{w}}\boldsymbol{\Sigma}^{-1}(\mathbf{1} - \mathbf{X}\mathbf{B})$ for a vector of ones, $\mathbf{1}$.

As with $\hat{\boldsymbol{\beta}}$, in practice these covariance matrices are evaluated at $\hat{\boldsymbol{\theta}}$. Moreover, these closed-form solutions provided enhance computational efficiency and clarity of the predictor's behavior.

3. Modeling moose presence in Alaska, USA

The `moose` data in **spmodel** contain information on moose (Alces Alces) presence in the Togiak region of Alaska, USA. `moose`is an `sf` object, a special data frame that is supplemented with spatial information using the **sf** package in R (Pebesma 2018). The first few rows of `moose` look like:

```
R> head(moose)

Simple feature collection with 6 features and 4 fields
Geometry type: POINT
Dimension:     XY
```

```
Bounding box:  xmin: 281896.4 ymin: 1518398 xmax: 311325.3 ymax: 1541016
Projected CRS: NAD83 / Alaska Albers
# A tibble: 6 x 5
   elev strat count presence           geometry
  <dbl> <chr> <dbl> <fct>           <POINT [m]>
1  469. L         0 0        (293542.6 1541016)
2  362. L         0 0        (298313.1 1533972)
3  173. M         0 0        (281896.4 1532516)
4  280. L         0 0        (298651.3 1530264)
5  620. L         0 0        (311325.3 1527705)
6  164. M         0 0        (291421.5 1518398)
```

There are five columns: `elev`, the numeric site elevation (meters); `strat` a stratification variable for sampling with two levels, `"L"` and `"M"`, which are categorized by landscape metrics at each site; `count`, the number of moose at each site; `presence`, a factor that indicates whether at least one moose was observed at each site (`0` implies no moose; `1` implies at least one moose); and `geometry`, the NAD83/Alaska Albers (EPSG: 3338) projected coordinate of each site (these data are point-referenced because each observation occurs at point coordinates and are represented by a `POINT` geometry. The `moose_preds` data in **spmodel** contain spatial locations at which predictions of moose presence are desired (and is also point-referenced). `moose_preds` is also an `sf` object with measurements for `elev` and `strat` and the same projection system. Figure 3 shows the `presence` variable in `moose` as well as the spatial locations of both `moose` and `moose_preds`. Moose are most commonly present in the southwestern and eastern parts of the domain and least commonly present in the northwest (Figure 3). Next we show how to use **spmodel** to study the effect of elevation and strata on moose presence while accounting for spatial covariance and to make predictions of moose presence at new locations.

3.1. Model Fitting

SPGLMs in **spmodel** are fit using the `spglm()` function. The `spglm()` function requires four arguments: `formula`, the relationship between the response and explanatory variables; `family`, the response distribution assumed for the repsonse variable; `data`, the data frame that contains the variables in `formula`, and `spcov_type`, the type of spatial covariance. These first three arguments are the three required arguments to `glm()` for nonspatial GLMs. So, the transition from `glm()` to `spglm()` simply requires one additional argument: `spcov_type`. When `data` is not an `sf` object, `spglm()` also requires the `xcoord` and `ycoord` arguments, which indicate the columns in `data` that represent the x- and y-coordinates, respectively (it is assumed these coordinates are already projected).

We use `spglm()` to fit a spatial logistic regression model quantifying the effect of elevation and strata on moose presence:

```
R> spbin <- spglm(
+   formula = presence ~ elev + strat,
+   family = binomial,
+   data = moose,
```

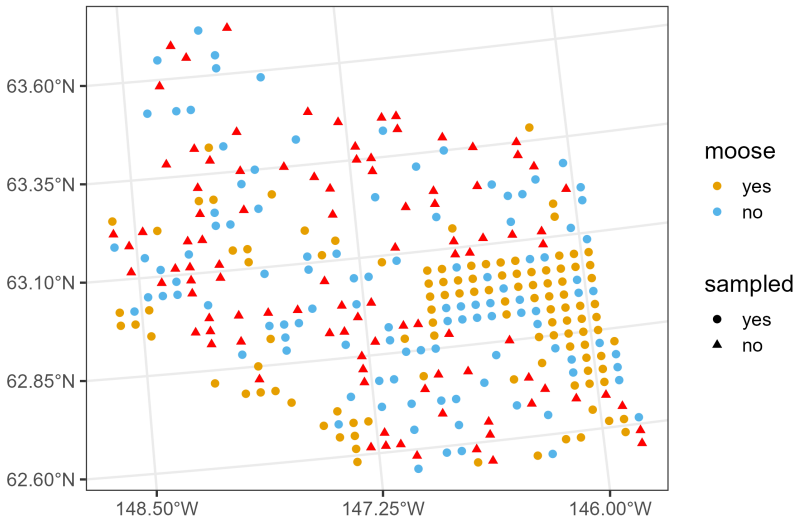

Figure 3: Moose presence in Alaska. Circles represent moose presence or absence (based on color) and triangles represent locations at which moose presence probability predictions are desired.

```
+   spcov_type = "exponential"
+ )
```

The `summary()` function returns a model summary that returns relevant information like the function call, deviance residuals, a coefficients table of fixed effects, the pseudo R-squared, spatial covariance parameter coefficient estimates, and the GLM dispersion parameter (fixed at one in logistic regression):

```
R> summary(spbin)


Call:
spglm(formula = presence ~ elev + strat, family = binomial, data = moose,
    spcov_type = "exponential")

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-1.7535 -0.8005  0.3484  0.7893  1.5797

Coefficients (fixed):
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.465713   1.486212  -1.659 0.097104 .
elev         0.006036   0.003525   1.712 0.086861 .
stratM       1.439273   0.420591   3.422 0.000622 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Pseudo R-squared: 0.06275

Coefficients (exponential spatial covariance):
       de        ie     range
5.145e+00 1.294e-03 4.199e+04

Coefficients (Dispersion for binomial family):
dispersion
         1
```

Based on this model, there is some evidence that elevation is associated with higher probabilities of moose presence (p-value ≈ 0.087) but strong evidence that moose are more prevalent in the `"M"` strata than the `"L"` strata (p-value < 0.001). The fixed effects coefficients table from `summary()` is often of primary practical interest, but it is not easily usable when printed directly to the R console. The `tidy()` function tidies this table, turning it into a data frame (i.e., a tibble) with standard column names:

```
R> tidy(spbin, conf.int = TRUE)

# A tibble: 3 x 7
  term        estimate std.error statistic  p.value  conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>     <dbl>     <dbl>
1 (Intercept) -2.47      1.49        -1.66 0.0971   -5.38        0.447
2 elev         0.00604   0.00353      1.71 0.0869   -0.000873    0.0129
3 stratM       1.44      0.421        3.42 0.000622  0.615       2.26
```

3.2. Model Comparison

The strength of spatial covarinace in the data affects how beneficial a SPGLM is relative to a GLM. When the spatial covariance is strong, the SPGLM should notably outperform the GLM. When the spatial covariance is weak, the SPGLM and GLM should perform similarly. We can quantify the benefits of incorporating spatial covariance for a particular data set by comparing the fit of a SPGLM to a GLM. We can fit a GLM in `spmodel` by specifying `spcov_type = "none"`:

```
R> bin <- spglm(
+   formula = presence ~ elev + strat,
+   family = binomial,
+   data = moose,
+   spcov_type = "none"
+ )
```

While the `spglm()` approach evaluates the HGLMM likelihood with $\sigma^2_{de} = 0$ and $\sigma^2_{ie} \approx 0$ instead of just the GLM likelihood, the parameter estimates and their standard errors are the same:

```
R> bin_glm <- glm(
+   formula = presence ~ elev + strat,
+   family = binomial,
+   data = moose,
+ )
R> round(coef(bin), digits = 4)

(Intercept)        elev      stratM
    -0.4247     -0.0003      0.8070

R> round(coef(bin_glm), digits = 4)

(Intercept)        elev      stratM
    -0.4247     -0.0003      0.8070

R> round(sqrt(diag(vcov(bin))), digits = 4)

(Intercept)        elev      stratM
     0.4208      0.0019      0.2906

R> round(sqrt(diag(vcov(bin_glm))), digits = 4)

(Intercept)        elev      stratM
     0.4208      0.0019      0.2906
```

However, using `spglm()` instead of `glm()` ensures that **spmodel** helper functions are available and that each of the `spglm()` models uses the same likelihood:

```
R> glance(spbin)

# A tibble: 1 x 10
      n     p  npar value   AIC  AICc   BIC logLik deviance pseudo.r.squared
  <int> <dbl> <int> <dbl> <dbl> <dbl> <dbl>  <dbl>    <dbl>            <dbl>
1   218     3     3  676.  682.  683.  693.  -338.     176.           0.0627

R> glance(bin)

# A tibble: 1 x 10
      n     p  npar value   AIC  AICc   BIC logLik deviance pseudo.r.squared
  <int> <dbl> <int> <dbl> <dbl> <dbl> <dbl>  <dbl>    <dbl>            <dbl>
1   218     3     0  708.  708.  708.  708.  -354.     294.           0.0280
```

The likelihood-based statistics AIC, AICc, BIC, and deviance are much lower for the SPGLM, indicating a better fit relative to the GLM. We may also perform a likelihood ratio test (LRT) between the two models, as the GLM is a special case of the SPGLM (i.e., is nested within the SPGLM):

```
R> tidy(anova(spbin, bin))

# A tibble: 1 x 5
  full  reduced    df statistic     p.value
  <chr> <chr>   <int>     <dbl>       <dbl>
1 spbin bin         3      31.5 0.000000652
```

The LRT test statistic is ≈ 31.5 with three degrees of freedom (the difference in covariance parameters), yielding a p-value < 0.001 that indicates preference for the full model (SPGLM) relative to the reduced model (GLM).

An alternative approach to model comparison is to use a cross-validation procedure (James, Witten, Hastie, and Tibshirani 2013). The `loocv()` function performs leave-one-out cross validation, comparing the predicted mean (on the response scale) to the observed response variable for each hold-out observation, recomputing estimates of $\boldsymbol{\beta}$ each time. Then, statistics like bias, mean-squared-prediction error (MSPE), and the square root of MSPE (RMSPE) can be used to evaluate models:

```
R> loocv(spbin)

# A tibble: 1 x 3
       bias  MSPE RMSPE
      <dbl> <dbl> <dbl>
1 0.0000206 0.156 0.394

R> loocv(bin)

# A tibble: 1 x 3
      bias  MSPE RMSPE
     <dbl> <dbl> <dbl>
1 -1.23e-9 0.240 0.490
```

Both models have negligible bias, but the SPGLM has much lower MSPE and RMSPE than the GLM, indicating the SPGLM predictions are far more efficient. Three separate metrics (likelihood-based statistics, likelihood-ratio test, and leave-one-out cross validation) prefer the SPGLM to the GLM.

We can compare two SPGLMs with different spatial covariance functions using likelihood-based statistics and leave-one-out cross validation, but we can't use the LRT because generally, the spatial covariance functions aren't nested:

```
R> spbin2 <- update(spbin, spcov_type = "spherical")
R> glances(spbin, spbin2)

# A tibble: 2 x 11
  model      n     p  npar value   AIC  AICc   BIC logLik deviance
  <chr>  <int> <dbl> <int> <dbl> <dbl> <dbl> <dbl>  <dbl>    <dbl>
1 spbin2   218     3     3  675.  681.  681.  691.  -338.     180.
2 spbin    218     3     3  676.  682.  683.  693.  -338.     176.
# i 1 more variable: pseudo.r.squared <dbl>
```

```
R> loocv(spbin)

# A tibble: 1 x 3
       bias  MSPE RMSPE
      <dbl> <dbl> <dbl>
1 0.0000206 0.156 0.394

R> loocv(spbin2)

# A tibble: 1 x 3
      bias  MSPE RMSPE
     <dbl> <dbl> <dbl>
1 0.000121 0.155 0.394
```

The `"exponential"` spatial covariance (`spbin`) has a slightly lower deviance but slightly higher AIC, AICc, and BIC than the `"spherical"` spatial covariance (`spbin2`). Both spatial covariance functions nearly identical leave-one-out cross validation metrics. For practical purposes, these models fit quite similarly.

3.3. Model Diagnostics

`spmodel` provides a suite of tools for model diagnostics. The `augment()` function augments the model data with diagnostics:

```
R> augment(spbin)

Simple feature collection with 218 features and 8 fields
Geometry type: POINT
Dimension:     XY
Bounding box:  xmin: 269085 ymin: 1416151 xmax: 419057.4 ymax: 1541016
Projected CRS: NAD83 / Alaska Albers
# A tibble: 218 x 9
   presence  elev strat .fitted .resid    .hat  .cooksd .std.resid
 * <fct>    <dbl> <chr>   <dbl>  <dbl>   <dbl>    <dbl>      <dbl>
 1 0         469. L       -1.95 -0.516 0.0476  0.00465      -0.528
 2 0         362. L       -2.70 -0.361 0.0123  0.000548     -0.363
 3 0         173. M       -1.96 -0.514 0.00455 0.000405     -0.516
 4 0         280. L       -3.15 -0.290 0.00413 0.000117     -0.291
 5 0         620. L       -1.19 -0.728 0.168   0.0427       -0.798
 6 0         164. M       -1.71 -0.576 0.00534 0.000598     -0.578
 7 0         164. M       -1.60 -0.606 0.00576 0.000714     -0.608
 8 0         186. L       -2.50 -0.397 0.00439 0.000233     -0.398
 9 0         362. L       -1.88 -0.532 0.0239  0.00237      -0.539
10 0         430. L       -1.54 -0.623 0.0497  0.00713      -0.639
# i 208 more rows
# i 1 more variable: geometry <POINT [m]>
```

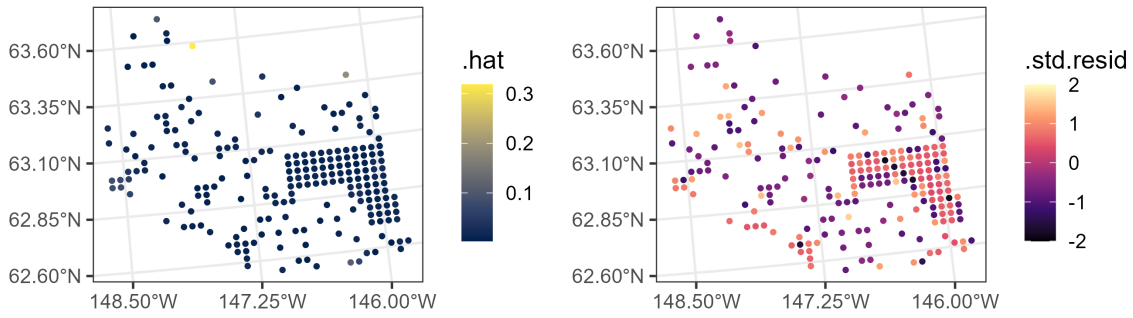

Figure 4: Spatial logistic regression model diagnostics from [augment].fct. The leverage (i.e., hat) values (left) and standardized residuals (right).

The fitted values (`.fitted`) can be returned on either the link ($\hat{\mathbf{w}}$) or response ($f^{-1}\hat{\mathbf{w}}$) scale and the residuals (`.resid`) can deviance, pearson, or response residuals. The defaults are the link scale and deviance residuals, respectively. Also returned by `augment()` are the leverage (`.hat`), Cook's distance (`.cooksd`), and standardized residuals `.std.resid`. A benefit of using `augment()` when `data` is an `sf` object is that the output is also an `sf` object, which makes it straightforward to create spatial diagnostic plots (Figure 4). Standard R helpers (e.g., `fitted()`, `residuals()`) are available to alternatively extract model diagnostics from the model object.

The `plot()` function can also be used to return similar diagnostics as from `lm()` and `glm()` with additional tools for spatial covariance. For example, we can inspect Cook's distance values and the empirical spatial covariance as a function (Figure 5) with

```
R> plot(spbin, which = c(4, 7))
```

The `varcomp()` function partitions model variability into several different components:

```
R> varcomp(spbin)
```

```
# A tibble: 3 x 2
  varcomp            proportion
  <chr>                   <dbl>
1 Covariates (PR-sq)   0.0627
2 de                   0.937
3 ie                   0.000236
```

The pseudo R-squared (PR^2), the proportion of variability attributable to the explanatory variables, is reported in the first row. The remaining variability ($1 - PR^2$) is allocated proportionally to `de` and `ie` according to σ^2_{de} and σ^2_{ie}. This variability partitioning is a useful that helps quantify how much the explanatory variables, residual spatial variance, and residual nonspatial variance contribute to model fit, but as with PR^2, should not be used as a model comparison tool.

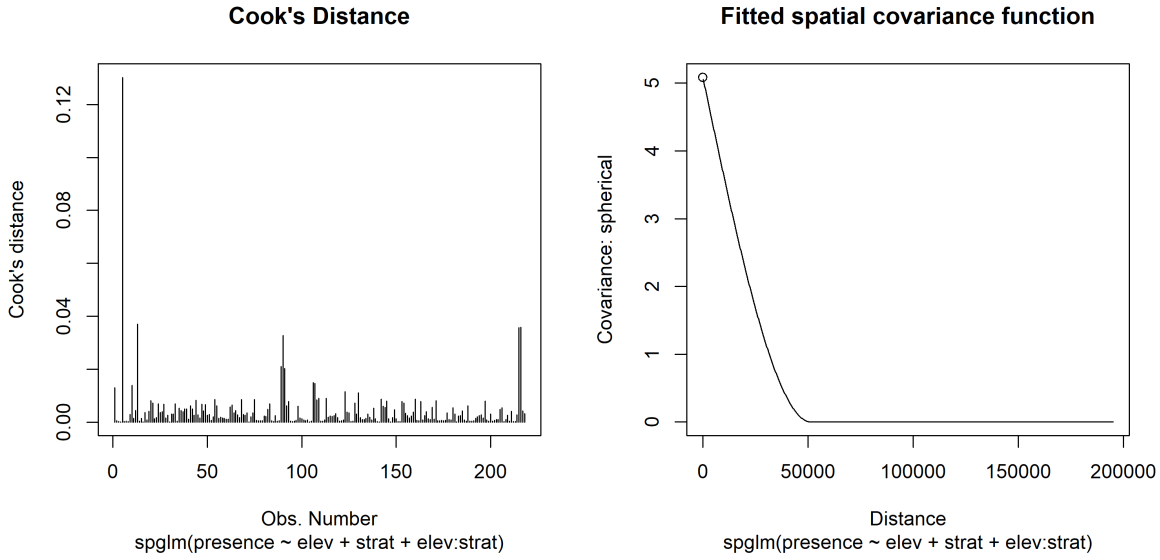

Figure 5: Spatial logistic regression model diagnostics from [plot].fct. The Cook's distance values (left) and the fitted spatial covariance as a function of distance (right).

3.4. Prediction

We can predict the probability of moose presence using `predict()`:

```
R> predict(spbin, newdata = moose_preds)[1:5]

          1           2           3           4           5
 0.06664165 -0.79069107 -1.60387940 -0.83159357  1.38183928
```

By default, predictions are returned on the link scale, but this can be changed to the response scale via `type`:

```
R> predict(spbin, newdata = moose_preds, type = "response")[1:5]

        1         2         3         4         5
0.5166542 0.3120203 0.1674401 0.3033082 0.7992862
```

Predictions on the response scale are visualized alongside the fitted values ($f^{-1}\hat{\mathbf{w}}$) in Figure 6. Prediction intervals are returned via `interval`

```
R> predict(spbin, newdata = moose_preds, interval = "prediction")[1:5, ]

          fit        lwr       upr
1  0.06664165 -2.0374370 2.1707203
2 -0.79069107 -3.4758514 1.8944692
3 -1.60387940 -4.0953329 0.8875741
4 -0.83159357 -3.0704818 1.4072947
5  1.38183928 -0.7692107 3.5328893
```

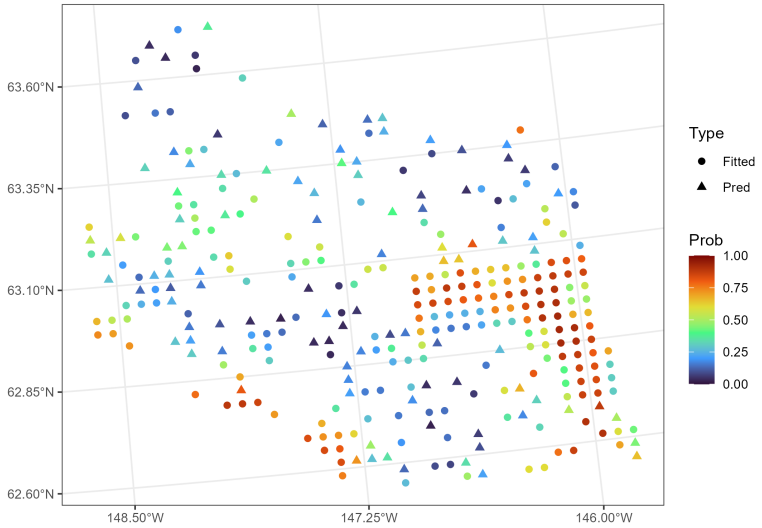

Figure 6: Moose presence probability fitted values and predictions. Fitted values are represeneted by circles and predictions by triangles.

We can alternatively use `augment()` to augment the prediction data with predictions. Arguments to `predict()` can also be passed to `augment()`:

```
R> augment(spbin, newdata = moose_preds, interval = "prediction")
```

```
Simple feature collection with 100 features and 5 fields
Geometry type: POINT
Dimension:     XY
Bounding box:  xmin: 269386.2 ymin: 1418453 xmax: 419976.2 ymax: 1541763
Projected CRS: NAD83 / Alaska Albers
# A tibble: 100 x 6
    elev strat .fitted .lower  .upper           geometry
 * <dbl> <chr>   <dbl>  <dbl>   <dbl>        <POINT [m]>
 1  143. L      0.0666 -2.04   2.17   (401239.6 1436192)
 2  324. L     -0.791  -3.48   1.89   (352640.6 1490695)
 3  158. L     -1.60   -4.10   0.888  (360954.9 1491590)
 4  221. M     -0.832  -3.07   1.41   (291839.8 1466091)
 5  209. M      1.38   -0.769  3.53   (310991.9 1441630)
 6  218. L     -2.59   -5.20   0.0177 (304473.8 1512103)
 7  127. L     -2.73   -5.24  -0.220  (339011.1 1459318)
 8  122. L     -2.32   -4.74   0.0920 (342827.3 1463452)
 9  191  L     -1.17   -4.01   1.66   (284453.8 1502837)
10  105. L     -0.905  -3.05   1.24   (391343.9 1483791)
# i 90 more rows
```

By using `augment()` when `newdata` is an `sf` object, predictions and their corresponding uncertainties are readily available for spatial mapping (Figure 7).

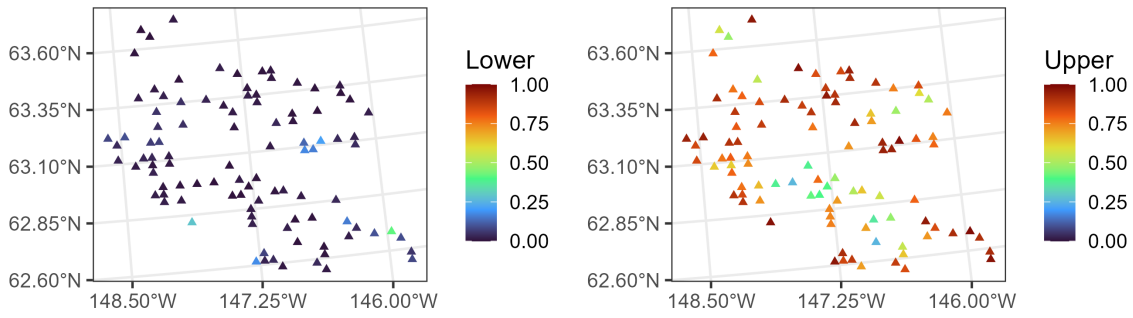

Figure 7: Moose presence probability prediction intervals. 95% prediction interval lower bound (left) and 95% prediction interval upper bound (right).

4. Additional applications

Throughout the remainder of this section, we briefly highlight some additional **spmodel** capabilities for SPGLMs. In Section 4.1, we fit Poisson and negative binomial models with and without geometric anisotropy for the point-referenced moose count data. In Section 4.2, we fit a binomial model to the areal seal trend data with a nonspatial random effect. In Section 4.3, we fit beta models to Texas voter turnout data. We explore how the Texas data can be treated as point-referenced or areal and compare the two spatial covariance structures empirically. We also highlight how the maximum likelihood estimation method is useful when comparing two models with different explanatory variables. Finally, in Section 4.4, we fit a Gamma model to the point-referenced lake conductivity data. We show how to perform a spatial analysis of variance (ANOVA) and leverage modeling functions from other R packages like **emmeans** and **car**.

4.1. Modeling moose counts in Alaska, USA

```
R> sppois <- spglm(
+   formula = count ~ elev + strat,
+   family = poisson,
+   data = moose,
+   spcov_type = "gaussian"
+ )
R> spnb <- update(sppois, family = nbinomial)
R> sppois_anis <- update(sppois, anisotropy = TRUE)
R> spnb_anis <- update(sppois_anis, family = nbinomial)

R> BIC(sppois, spnb, sppois_anis, spnb_anis)

            df      BIC
sppois       3 1343.892
spnb         4 1340.706
```

```
sppois_anis  5 1337.610
spnb_anis    6 1335.051

R> plot(spnb, which = 8)
R> plot(spnb_anis, which = 8)
```

4.2. Modeling harbor seal trends in Alaska, USA

```
R> spbin <- spgautor(
+   formula = log_trend > 0 ~ 1,
+   family = binomial,
+   data = seal,
+   spcov_type = "car",
+   random = ~ stock
+ )
R> tidy(spbin, conf.int = TRUE)

# A tibble: 1 x 7
  term        estimate std.error statistic p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
1 (Intercept)   -0.340     0.673    -0.506   0.613    -1.66     0.979
```

4.3. Modeling voter turnout in Texas, USA

```
R> spbeta_geo <- spglm(
+   formula = turnout ~ log_income,
+   family = "beta",
+   data = texas,
+   spcov_type = "matern"
+ )
R>
R> spbeta_auto <- spgautor(
+   formula = turnout ~ log_income,
+   family = "beta",
+   data = texas,
+   spcov_type = "car",
+   cutoff = 1e5
+ )

R> AIC(spbeta_geo, spbeta_auto)

            df       AIC
spbeta_geo   5 -44.53113
spbeta_auto  3 -22.46104
```

```
R> tidy(spbeta_geo)

# A tibble: 2 x 5
  term        estimate std.error statistic    p.value
  <chr>          <dbl>     <dbl>     <dbl>      <dbl>
1 (Intercept)   -5.20      1.11      -4.70 0.00000260
2 log_income     0.579     0.122      4.76 0.00000194

R> spbeta_full_ml <- update(spbeta_geo, estmethod = "ml")
R> spbeta_red_ml <- update(spbeta_full_ml, formula = turnout ~ 1)
R> tidy(anova(spbeta_full_ml, spbeta_red_ml))

# A tibble: 1 x 5
  full          reduced           df statistic    p.value
  <chr>         <chr>          <dbl>     <dbl>      <dbl>
1 spbeta_red_ml spbeta_full_ml     1      23.2 0.00000149
```

4.4. Modeling lake conductivity in Southwest, USA

```
R> spgam <- spglm(
+   formula = exp(log_cond) ~ temp * state + origin,
+   family = "Gamma",
+   data = lake,
+   spcov_type = "cauchy",
+   partition_factor = ~ year
+ )

R> anova(spgam)

Analysis of Variance Table

Response: exp(log_cond)
            Df    Chi2 Pr(>Chi2)
(Intercept)  1 51.5270 7.062e-13 ***
temp         1 25.5146 4.390e-07 ***
state        3  3.0747 0.3802528
origin       1  0.1429 0.7053819
temp:state   3 19.7668 0.0001897 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R> car::vif(spgam)

                GVIF Df GVIF^(1/(2*Df))
temp         4.691914  1        2.166083
state      127.082397  3        2.242234
origin       1.264940  1        1.124695
temp:state  76.387383  3        2.059856
```

```
R> pairs(emmeans::emmeans(spgam, ~ state | temp))

temp = 7.63:
 contrast estimate    SE  df z.ratio p.value
 AZ - CO    -1.012 0.337 Inf  -3.004  0.0142
 AZ - NV    -0.900 0.348 Inf  -2.584  0.0480
 AZ - UT    -1.331 0.326 Inf  -4.082  0.0003
 CO - NV     0.112 0.258 Inf   0.434  0.9727
 CO - UT    -0.319 0.223 Inf  -1.427  0.4822
 NV - UT    -0.431 0.244 Inf  -1.763  0.2915

Results are averaged over the levels of: origin
Degrees-of-freedom method: asymptotic
Results are given on the log (not the response) scale.
P value adjustment: tukey method for comparing a family of 4 estimates

R> emmeans::emtrends(spgam, ~ state, var = "temp")

 state temp.trend     SE  df asymp.LCL asymp.UCL
 AZ         0.152 0.0301 Inf    0.0929     0.211
 CO         0.289 0.0370 Inf    0.2161     0.361
 NV         0.171 0.0504 Inf    0.0718     0.270
 UT         0.352 0.0372 Inf    0.2791     0.425

Results are averaged over the levels of: origin
Degrees-of-freedom method: asymptotic
Results are given on the exp (not the response) scale.
Confidence level used: 0.95
```

5. Discussion

References

Akaike H (1974). "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control*, **19**(6), 716–723.

Anderson SC, Ward EJ, English PA, Barnett LAK, Thorson JT (2024). "sdmTMB: an R package for fast, flexible, and user-friendly generalized linear mixed effects models with spatial and spatiotemporal random fields." *bioRxiv*, **2022.03.24.485545**. doi:10.1101/2022.03.24.485545.

Bachl FE, Lindgren F, Borchers DL, Illian JB (2019). "inlabru: an R package for Bayesian spatial modelling from ecological survey data." *Methods in Ecology and Evolution*, **10**, 760–766. doi:10.1111/2041-210X.13168.

Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White JSS (2009). "Generalized Linear Mixed Models: A Practical Guide for Ecology and Evolution." *Trends in Ecology & Evolution*, **24**(3), 127–135.

Bonat WH, Ribeiro Jr PJ (2016). "Practical likelihood analysis for spatial generalized linear mixed models." *Environmetrics*, **27**(2), 83–89.

Breslow NE, Clayton DG (1993). "Approximate Inference in Generalized Linear Mixed Models." *Journal of the American Statistical Association*, **88**(421), 9–25.

Brooks ME, Kristensen K, van Benthem KJ, Magnusson A, Berg CW, Nielsen A, Skaug HJ, Maechler M, Bolker BM (2017). "glmmTMB Balances Speed and Flexibility Among Packages for Zero-Inflated Generalized Linear Mixed Modeling." *The R Journal*, **9**(2), 378–400. doi:10.32614/RJ-2017-066.

Bürkner PC (2017). "brms: An R package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software*, **80**, 1–28.

Chambers JM, Hastie TJ (eds.) (1992). *Statistical Models in S*. Chapman & Hall, London.

Cook RD (1979). "Influential Observations in Linear Regression." *Journal of the American Statistical Association*, **74**(365), 169–174.

Cook RD, Weisberg S (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall.

Cressie N (1990). "The origins of kriging." *Mathematical geology*, **22**(3), 239–252.

Cressie N (1993). *Statistics for Spatial Data*. John Wiley & Sons.

Doser JW, Finley AO, Kéry M, Zipkin EF (2022). "spOccupancy: An R package for single-species, multi-species, and integrated spatial occupancy models." *Methods in Ecology and Evolution*, **13**(8), 1670–1678.

Doser JW, Finley AO, Kéry M, Zipkin EF (2024). "spAbundance: An R package for single-species and multi-species spatially explicit abundance models." *Methods in Ecology and Evolution*, **15**(6), 1024–1033.

Dumelle M, Higham M, Ver Hoef JM (2023). "spmodel: Spatial Statistical Modeling and Prediction in R." *PLOS ONE*, **18**(3), e0282524.

Evangelou E, Zhu Z, Smith RL (2011). "Estimation and prediction for spatial generalized linear mixed models using high order Laplace approximation." *Journal of Statistical Planning and Inference*, **141**(11), 3564–3577.

Faraway JJ (2016). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. CRC press.

Finley AO, Banerjee S, Carlin BP (2007). "spBayes: An R Package for Univariate and Multivariate Hierarchical Point-Referenced Spatial Models." *Journal of Statistical Software*, **19**(4), 1–24. URL https://www.jstatsoft.org/article/view/v019i04.

Finley AO, Datta A, Banerjee S (2022). "spNNGP R Package for Nearest Neighbor Gaussian Process Models." *Journal of Statistical Software*, **103**(5), 1–40. doi:10.18637/jss.v103.i05.

Fox J, Weisberg S (2019). *An R Companion to Applied Regression*. Third edition. Sage, Thousand Oaks CA. URL https://www.john-fox.ca/Companion/.

Harville DA (1977). "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems." *Journal of the American Statistical Association*, **72**(358), 320–338.

Hoeting JA, Davis RA, Merton AA, Thompson SE (2006). "Model Selection for Geostatistical Models." *Ecological Applications*, **16**(1), 87–98.

Hughes J, Cui X (2020). *ngspatial: Fitting the Centered Autologistic and Sparse Spatial Generalized Linear Mixed Models for Areal Data*. Frederick, MD. R package version 1.2-2.

James G, Witten D, Hastie T, Tibshirani R (2013). *An Introduction to Statistical Learning*. Springer.

Kuhn M, Silge J (2022). *Tidy Modeling with R*. O'Reilly Media, Inc.

Lee D (2013). "CARBayes: An R Package for Bayesian Spatial Modeling with Conditional Autoregressive Priors." *Journal of Statistical Software*, **55**(13), 1–24.

Lee Y, Nelder JA (1996). "Hierarchical Generalized Linear Models." *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**(4), 619–656.

Lenth RV (2024). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.10.3, URL https://CRAN.R-project.org/package=emmeans.

Lindgren F, Rue H (2015). "Bayesian Spatial Modelling with R-INLA." *Journal of Statistical Software*, **63**, 1–25.

McCullagh P, Nelder JA (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall Ltd.

Montgomery DC, Peck EA, Vining GG (2021). *Introduction to Linear Regression Analysis*. John Wiley & Sons.

Myers RH, Montgomery DC, Vining GG, Robinson TJ (2012). *Generalized Linear Models: With Applications in Engineering and the Sciences*. John Wiley & Sons.

Nelder JA, Wedderburn RW (1972). "Generalized Linear Models." *Journal of the Royal Statistical Society: Series A (General)*, **135**(3), 370–384.

Patterson D, Thompson R (1971). "Recovery of Inter-Block Information when Block Sizes are Unequal." *Biometrika*, **58**(3), 545–554.

Pebesma E (2018). "Simple Features for R: Standardized Support for Spatial Vector Data." *The R Journal*, **10**(1), 439–446. doi:10.32614/RJ-2018-009. URL https://doi.org/10.32614/RJ-2018-009.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rencher AC, Schaalje GB (2008). *Linear models in statistics*. John Wiley & Sons.

Robinson D, Hayes A, Couch S (2021). *broom: Convert Statistical Objects into Tidy Tibbles*. R package version 0.7.6, URL https://CRAN.R-project.org/package=broom.

Ronnegard L, Shen X, Alam M (2010). "hglm: A Package for Fitting Hierarchical Generalized Linear Models." *The R Journal*, **2**(2), 20–28.

Rousset F, Ferdy JB (2014). "Testing Environmental and Genetic Effects in the Presence of Spatial Autocorrelation." *Ecography*, **37**(8), 781–790. URL https://dx.doi.org/10.1111/ecog.00566.

Sainsbury-Dale M, Zammit-Mangion A, Cressie N (2024). "Modeling Big, Heterogeneous, Non-Gaussian Spatial and Spatio-Temporal Data Using FRK." *Journal of Statistical Software*, **108**, 1–39.

Schwarz G (1978). "Estimating the Dimension of a Model." *The Annals of Statistics*, pp. 461–464.

Smith TJ, McKenna CM (2013). "A comparison of logistic regression pseudo R2 indices." *General Linear Model Journal*, **39**(2), 17–26.

Thorson JT, Anderson SC, Goddard P, Rooper CN (2025). "tinyVAST: R package with an expressive interface to specify lagged and simultaneous effects in multivariate spatio-temporal models." *Global Ecology and Biogeography*, **34**(4), e70035. doi:10.1111/geb.70035. URL https://doi.org/10.1111/geb.70035.

Tobler WR (1970). "A Computer Movie Simulating Urban Growth in the Detroit Region." *Economic Geography*, **46**(sup1), 234–240.

Ver Hoef JM, Blagg E, Dumelle M, Dixon PM, Zimmerman DL, Conn PB (2024). "Marginal Inference for Hierarchical Generalized Linear Mixed Models with Patterned Covariance Matrices Using the Laplace Approximation." *Environmetrics*, **35**(7), e2872. doi:10.1002/env.2872.

Wedderburn RW (1974). "Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss—Newton Method." *Biometrika*, **61**(3), 439–447.

Wolfinger R, O'connell M (1993). "Generalized Linear Mixed Models: A Pseudo-Likelihood Approach." *Journal of Statistical Computation and Simulation*, **48**(3-4), 233–243.

Wolfinger R, Tobias R, Sall J (1994). "Computing Gaussian Likelihoods and their Derivatives for General Linear Mixed Models." *SIAM Journal on Scientific Computing*, **15**(6), 1294–1310.

Wood SN (2017). *Generalized Additive Models: An Introduction with R*. CRC press.

Zimmerman DL, Ver Hoef JM (2024). *Spatial Linear Models for Environmental Data*. CRC Press.

Affiliation:

Michael Dumelle
United States
Environmental Protection Agency
200 SW 35th St
Corvallis, OR, 97330
E-mail: Dumelle.Michael@epa.gov

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd