



---

## Spatial Generalized Linear Models in R Using **spmodel**

Michael Dumelle       Jay M. Ver Hoef       Matt Higham   
United States      Alaska Fisheries  
Environmental Protection Agency      Science Center      St. Lawrence University

---

### Abstract

Non-Gaussian data are common in practice and include binary, count, skewed, and proportion data types. Often, non-Gaussian data are modeled using a generalized linear model (GLM). GLMs typically assume that observations are independent of one another. This is an impractical assumption for spatial data, as nearby observations tend to be more similar than distant ones. The **spmodel** package in R provides a suite of tools for fitting spatial generalized linear models (SPGLMs) to non-Gaussian data and making spatial predictions (i.e., Kriging). SPGLMs for point-referenced (x- and y-coordinates) support are fit using the `spglm()` function, while SPGLMs for areal (lattice, polygon) support are fit using the `spgautor()` function. Both `spglm()` and `spgautor()` maximize a novel Laplace likelihood which marginalizes over the model's fixed effects and latent mean while formally incorporating spatial covariance among observations. The inputs and outputs of `spglm()` and `spgautor()` closely resemble the `glm()` function from base R, easing the transition from GLMs to SPGLMs. **spmodel** provides and builds upon several commonly used helper functions for model building like `summary()`, `plot()`, and `fitted()`, among others. Spatial predictions of the latent mean at unobserved locations are obtained using `predict()` or `augment()`. **spmodel** accommodates myriad advanced modeling features like geometric anisotropy, nonspatial random effects, analysis of variance, and more. Throughout, we use **spmodel** to fit SPGLMs to moose presence and counts in Alaska, United States (US), skewed conductivity data in the Southwestern US, harbor seal abundance trends in Alaska, US, and voter turnout rates in Texas, US.

*Keywords:* autoregressive model, geostatistical model, Poisson regression, link function, logistic regression, overdispersion, spatial covariance, spatial dependence.

---

## 1. Introduction

In practice, non-Gaussian data (e.g., binary, count, skewed, and proportion data) are ubiquitous. Non-Gaussian data that belong to an exponential family can be naturally modeled using a generalized linear model (GLM) regression framework (Nelder and Wedderburn 1972; McCullagh and Nelder 1989). In a GLM, an  $n \times 1$  response variable  $\mathbf{y}$  belongs to a statistical distribution (e.g., binomial, Poisson) with some mean and variance. Often, the analysis goal is to study the impact of a linear function of several explanatory variables on the mean of  $\mathbf{y}$  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:

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

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., if a probability, between zero and one), the link function generally makes  $\mathbf{w}$  unconstrained. Common link functions include the log odds (i.e., logit) link for binary and proportion data and the log link for count and skewed data. Equation 1 can also be written in terms of the inverse link function,  $f^{-1}(\cdot)$ :

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

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 an 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 side 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. For spatial data, nearby observations tend to be more similar than distant observations (Tobler 1970), which leads 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 the elements of  $\mathbf{w}$  exhibit both spatial and nonspatial variability. 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  $E(\boldsymbol{\tau}) = \mathbf{0}$  and  $Cov(\boldsymbol{\tau}) = \sigma_{\tau}^2 \mathbf{R}$ , where  $E(\cdot)$  and  $Cov(\cdot)$  denote expectation and covariance, respectively. The variance parameter  $\sigma_{\tau}^2$  controls the magnitude of spatial covariance and is often called a partial sill. The matrix  $\mathbf{R}$  is an  $n \times n$  spatial correlation matrix that depends on a range parameter controlling the distance-decay rate of the spatial correlation. One example of a spatial covariance matrix is the “exponential,” which is given by

$$Cov(\boldsymbol{\tau}) = \sigma_{\tau}^2 \mathbf{R}_{exp} = \sigma_{\tau}^2 \exp(-\mathbf{H}/\phi), \quad (2)$$

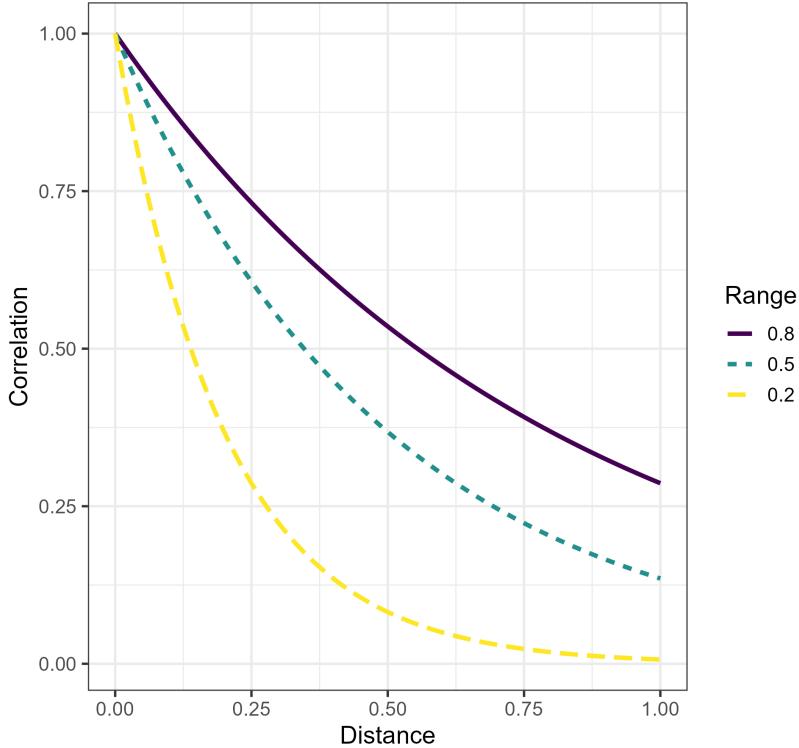


Figure 1: An exponential spatial correlation function with varying range parameters.

where  $\mathbf{H}$  is a matrix of pairwise distances among the elements of  $\mathbf{y}$  and  $\phi$  is the range parameter. From Equation 2, as the distance between two elements of  $\mathbf{y}$  increases, the spatial covariance decreases, which reflects intuition. Moreover, as the range parameter,  $\phi$ , 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  $E(\boldsymbol{\epsilon}) = \mathbf{0}$  and  $Cov(\boldsymbol{\tau}) = \sigma_{\epsilon}^2 \mathbf{I}$ , where  $\mathbf{I}$  is an  $n \times n$  identity matrix. The variance parameter  $\sigma_{\epsilon}^2$  controls the magnitude of nonspatial variability (i.e., fine-scale variation) and is often called a nugget. Often in spatial statistics, quantities are explicitly referenced with respect to  $\mathbf{s}$ , a vector of spatial coordinates indexing the observation (Cressie 1993). For example,  $\mathbf{y}$  and  $\mathbf{X}$  may instead be written  $\mathbf{y}(\mathbf{s})$  and  $\mathbf{X}(\mathbf{s})$ , respectively. We acknowledge the utility of this nomenclature but drop the explicit dependence on  $\mathbf{s}$  for simplicity of notation moving forward.

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

$$f(\boldsymbol{\mu} | \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\epsilon}) \equiv \mathbf{w} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\tau} + \boldsymbol{\epsilon}. \quad (3)$$

Assuming independence among  $\boldsymbol{\tau}$  and  $\boldsymbol{\epsilon}$ , it follows that

$$Cov(\boldsymbol{\tau} + \boldsymbol{\epsilon}) = Cov(\boldsymbol{\tau}) + Cov(\boldsymbol{\epsilon}) = \sigma_{\tau}^2 \mathbf{R} + \sigma_{\epsilon}^2 \mathbf{I}.$$

Henceforth, we refer to  $\sigma_{\tau}^2$  as  $\sigma_{de}^2$  (for spatially dependent error variance) and  $\sigma_{\epsilon}^2$  as  $\sigma_{ie}^2$  (for independent error variance). The parameters  $\sigma_{de}^2$ ,  $\sigma_{ie}^2$ , and  $\phi$ , in addition to 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), **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 (e.g., **R-INLA**) work with the precision matrix instead of the covariance matrix so that computationally expensive matrix inversion is not required. Working with precision matrices, however, can be more restrictive and less intuitive than working with covariance matrices. 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), **mgev** (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 Gaussian Markov random fields to fit a wide variety of SPGLMs, while **tinyVAST** (Thorson, Anderson, Goddard, and Rooper 2025) extends some of these models to multivariate or (dynamic) structural equation models.

Building from Evangelou, Zhu, and Smith (2011) and Bonat and Ribeiro Jr (2016), Ver Hoef, Blagg, Dumelle, Dixon, Zimmerman, and Conn (2024) proposed a novel approach for fitting SPGLMs that leverages the Laplace approximation while marginalizing over both the latent  $\mathbf{w}$  and the fixed effects ( $\beta$ ). This approach performed well in a variety of simulation settings, generally having appropriate confidence interval coverage for the fixed effects and prediction interval coverage for  $\mathbf{w}$  at new locations. It also 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. The novel Laplace approach is particularly attractive for two reasons. First, it is general enough that it 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 released a full set of modeling tools for SPGLMs fit using the novel Laplace approach described by 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 support (e.g., x- and y-coordinates representing point locations in a field; these models are sometimes called “geostatistical” models), while the **spgautor()** function fits SPGLMs for areal support (e.g., polygon boundaries representing

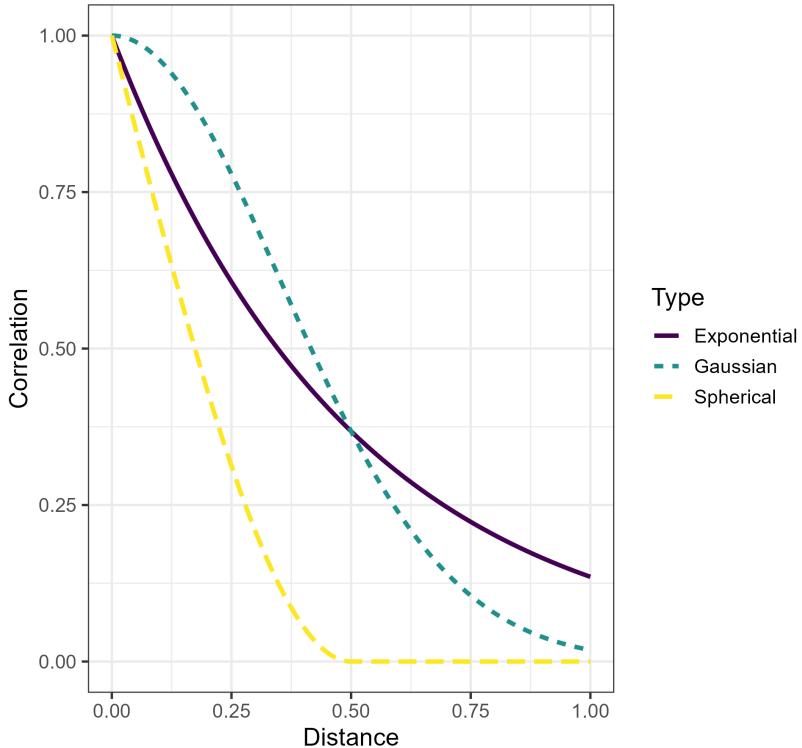


Figure 2: Exponential, Gaussian, and spherical spatial correlation functions all with range parameters equal to 0.5.

102 geographic subsets of a region; these models are sometimes called “autoregressive” models).  
 103 For both point-referenced and areal supports, **spmodel** supports the binomial distribution for  
 104 binary data, Poisson and negative binomial distributions for count data, Gamma and inverse  
 105 Gaussian distributions for skewed data, and the beta distribution for proportion data. There  
 106 are 20 different spatial covariance structures available including the exponential, Gaussian,  
 107 and spherical for point-referenced support (Figure 2) and the conditional autoregressive, and  
 108 simultaneous autoregressive structures for areal support **spmodel** provides tools for commonly  
 109 used model summaries, visualizations, and diagnostics (e.g., fitted values) using standard R  
 110 helper functions like `summary()`, `plot()`, and `fitted()`, among others. **spmodel** also pro-  
 111 vides tools to predict `w` at new locations and quantify uncertainty in those prediction using  
 112 `predict()` and `augment()`. This core functionality, combined with several advanced fea-  
 113 tures we describe throughout the manuscript, enables **spmodel** to introduce novel, important  
 114 SPGLM modeling tools previously missing from the existing R ecosystem.  
 115 Of the existing R packages for SPGLMs, **spmodel** (version 0.11.1) is arguably most similar  
 116 to **sdmTMB** (version 0.7.4) in terms of scope and feel. Both packages use similar syntax  
 117 as `glm()`, accommodate flexible `formula` arguments (e.g., offsets, splines), handle spatial  
 118 covariance that decays at different rates in different directions (i.e., geometric anisotropy),  
 119 incorporate nonspatial random effects, support other R packages for modeling like **broom**  
 120 (Robinson, Hayes, and Couch 2021; Kuhn and Silge 2022), **emmeans** (Lenth 2024), and **car**  
 121 (Fox and Weisberg 2019), and have tools for model summaries, prediction, and simulating  
 122 data. There are some notable differences between the two packages, however. **sdmTMB** sup-

123 ports several additional GLM distributions like the Tweedie, supports Hurdle models, and  
124 can incorporate prior information through Bayesian applications. **sdmTMB** also provides  
125 tools for working with temporal data and spatiotemporal data and provides enhanced vi-  
126 sualizations of the model's marginal effects. **sdmTMB** does require a preprocessing step of  
127 constructing a mesh prior to model fitting (using the stochastic partial differential equation  
128 approach), and the density of the mesh can affect model results and computational com-  
129 plexity. On the other hand, **spmodel** does not require the construction of a mesh prior to  
130 model fitting. **spmodel** supports 20 different spatial covariances and models them directly,  
131 rather than using a precision matrix approximation to the Matérn spatial covariance as in  
132 **sdmTMB**. **spmodel** can model data directly using neighborhood distance and autoregressive  
133 models, rather than relying on the polygon centroid (as in **sdmTMB**), which may not be  
134 within the polygon's boundaries. **spmodel** provides experimental design tools (e.g., analysis  
135 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  
136 has analytic solutions for fixed effect and prediction standard errors. Other similarities and  
137 differences do exist between **sdmTMB** and **spmodel**, and both packages continue to evolve.  
138 Overall, we believe that these packages are complementary and enhance the suite of SPGLM  
139 tools accessible to practitioners.

141 The rest of this article is organized as follows. In Section 2, we provide some background for  
142 the SPGLM fitting and prediction routines in **spmodel**. In Section 3, we provide an overview  
143 of core SPGLM functionality in **spmodel** by modeling moose presence in Alaska, United  
144 States (US). In Section 4, we model moose counts in Alaska, US; skewed lake conductivity  
145 in the Southwestern US; harbor seal abundance trend behavior in Alaska, US; and voter  
146 turnout rates in Texas, US. And in Section 5, we end with a discussion synthesizing **spmodel**'s  
147 contributions to SPGLMs in R.

## 2. The spatial generalized linear model and marginalization

148 The novel Laplace approach implemented in **spmodel** formally maximizes a hierarchical GLM  
149 likelihood ([Lee and Nelder 1996](#); [Wood 2017](#)), making likelihood-based statistics for model  
150 comparison like AIC ([Akaike 1974](#)), AICc ([Hoeting, Davis, Merton, and Thompson 2006](#)), BIC  
151 ([Schwarz 1978](#)), deviance ([McCullagh and Nelder 1989](#)), and likelihood ratio tests available.  
152 These types of statistics are not directly available for quasi-likelihood ([Wedderburn 1974](#);  
153 [Breslow and Clayton 1993](#)) or pseudo-likelihood approaches ([Wolfinger and O'Connell 1993](#)),  
154 which only specify the first two moments of a distribution. Next, we describe a brief overview  
155 of the approach and how it can be used for several primary data analysis tasks ([Tredennick,](#)  
156 [Hooker, Ellner, and Adler 2021](#)) like model comparison, parameter estimation, inference,  
157 model diagnostics, and prediction.

### 158 2.1. Formulating the hierarchical likelihood

159 We can write the SPGLM likelihood hierarchically as

$$[\mathbf{y}|\mathbf{X}, \varphi, \boldsymbol{\theta}] = \int_{\mathbf{w}} \int_{\boldsymbol{\beta}} [\mathbf{y}|f^{-1}(\mathbf{w}), \varphi] [\mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\theta}] d\boldsymbol{\beta} d\mathbf{w}, \quad (4)$$

160 where  $[\mathbf{y}|f^{-1}(\mathbf{w}), \varphi]$  is the density for the appropriate response distribution of  $\mathbf{y}$  (e.g., bi-  
161 nomial, Poisson) given the latent  $\mathbf{w}$  and dispersion parameter ( $\varphi$ ), and  $[\mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\theta}]$  is the

<sup>162</sup> multivariate Gaussian density for  $\mathbf{w}$  given the explanatory variables ( $\mathbf{X}$ ), fixed effects ( $\boldsymbol{\beta}$ ),  
<sup>163</sup> and spatial covariance parameters ( $\boldsymbol{\theta}$ ). The elements of  $[\mathbf{y}|f^{-1}(\mathbf{w}), \varphi]$  are conditionally in-  
<sup>164</sup> dependent (given  $\mathbf{w}$ ), but the elements of  $[\mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\theta}]$  share spatial covariance. Following  
<sup>165</sup> Harville (1974), we can integrate  $\boldsymbol{\beta}$  out of Equation 4, which yields

$$[\mathbf{y}|\mathbf{X}, \varphi, \boldsymbol{\theta}] = \int_{\mathbf{w}} [\mathbf{y}|f^{-1}(\mathbf{w}), \varphi][\mathbf{w}|\mathbf{X}, \boldsymbol{\theta}]d\mathbf{w}, \quad (5)$$

<sup>166</sup> where  $[\mathbf{w}|\mathbf{X}, \boldsymbol{\theta}]$  is the restricted (i.e., residual) multivariate Gaussian density (Patterson and  
<sup>167</sup> Thompson 1971) for  $\mathbf{w}$  given the explanatory variables and covariance parameters. The  
<sup>168</sup> restricted multivariate Gaussian density is given by

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

<sup>169</sup> where  $\tilde{\boldsymbol{\beta}} = (\mathbf{X}^{\top}\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^{\top}\Sigma^{-1}\mathbf{w}$ ,  $\Sigma$  denotes the covariance matrix (of  $\mathbf{w}$ ), and  $|\cdot|$  denotes  
<sup>170</sup> the determinant. Equation 5 can synonymously be written after profiling the overall variance  
<sup>171</sup> out of  $\Sigma$ , which reduces the dimension of  $\boldsymbol{\theta}$  by one for optimization (Wolfinger, Tobias, and  
<sup>172</sup> Sall 1994). Next, let

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

<sup>173</sup> and rewrite Equation 5 as

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

<sup>174</sup> A second-order Taylor series expansion of  $\ell_{\mathbf{w}}$  around a point  $\mathbf{w}^*$  yields

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

<sup>175</sup> where  $\mathbf{g}$  and  $\mathbf{G}$  are the gradient and Hessian, respectively, of  $\ell_{\mathbf{w}}$  with respect to  $\mathbf{w}$ . If  $\mathbf{w}^*$  is  
<sup>176</sup> a value for which  $\mathbf{g} = \mathbf{0}$ ,

$$[\mathbf{y}|\mathbf{X}, \varphi, \boldsymbol{\theta}] \approx \exp(\ell_{\mathbf{w}^*}) \int_{\mathbf{w}} \exp(-\frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^{\top}(-\mathbf{G})(\mathbf{w} - \mathbf{w}^*))d\mathbf{w}. \quad (6)$$

<sup>177</sup> The integral in Equation 6 can be solved by leveraging properties of the normalizing constant  
<sup>178</sup> of a multivariate Gaussian distribution. Thus, rewriting  $\exp(\ell_{\mathbf{w}^*})$  yields

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

<sup>179</sup> which can be directly evaluated. This result suggests a doubly iterative optimization over 1)  
<sup>180</sup>  $\boldsymbol{\theta}$  and  $\varphi$  and 2) the latent  $\mathbf{w}$  (to find each set of  $\mathbf{w}^*$ ), which ultimately yields the marginal  
<sup>181</sup> restricted maximum likelihood estimators  $\hat{\varphi}$  and  $\hat{\boldsymbol{\theta}}$  and their respective values of  $\mathbf{w}^*$ , which  
<sup>182</sup> we call  $\hat{\mathbf{w}}$ . Ver Hoef *et al.* (2024) provide further details, which includes explicit forms of  $\mathbf{g}$   
<sup>183</sup> and  $\mathbf{G}$  for various response distributions.

## <sup>184</sup> 2.2. Estimating fixed effects

We can estimate the fixed effects 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  $\beta$  is given by

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

where  $\mathbf{B} = (\mathbf{X}^\top \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \Sigma^{-1}$ . While we do not  $\mathbf{w}$ , we do estimate it via  $\hat{\mathbf{w}}$ , and thus it is reasonable to define  $\hat{\beta} = \mathbf{B}\hat{\mathbf{w}}$ . To derive properties of  $\hat{\beta}$  like expectation and variance, we must derive these properties for  $\hat{\mathbf{w}}$  by conditioning on  $\mathbf{w}$  as if it were observed and leveraging the laws of total expectation and variance. Based on asymptotic properties of (restricted) maximum likelihood estimators (Cressie and Lahiri 1993), we may assume that given  $\mathbf{w}$ ,  $\hat{\mathbf{w}}$  has mean  $\mathbf{w}$  and variance equal to  $-\mathbf{H}^{-1}$ , the negative inverse Hessian (i.e., the inverse observed information matrix). Thus it follows that  $E(\hat{\mathbf{w}})$  is given by

$$E(\hat{\mathbf{w}}) = E(E(\hat{\mathbf{w}}|\mathbf{w})) = E(\mathbf{w}) = \mathbf{X}\beta$$

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

$$\begin{aligned} \text{Var}(\hat{\mathbf{w}}) &= E(\text{Var}(\hat{\mathbf{w}}|\mathbf{w})) + \text{Var}(E(\hat{\mathbf{w}}|\mathbf{w})) \\ &= E(-\mathbf{H}^{-1}) + \text{Var}(\mathbf{w}) \\ &= -\mathbf{H}^{-1} + \Sigma \end{aligned}$$

Putting this all together, it follows that  $\hat{\beta}$  is unbiased for  $\beta$ :

$$E(\hat{\beta}) = E(\mathbf{B}\hat{\mathbf{w}}) = \mathbf{B}E(\hat{\mathbf{w}}) = (\mathbf{X}^\top \Sigma^{-1} \mathbf{X})^{-1} (\mathbf{X}^\top \Sigma^{-1} \mathbf{X})\beta = \beta.$$

Moreover, it follows that

$$\begin{aligned} \text{Var}(\hat{\beta}) &= \text{Var}(\mathbf{B}\hat{\mathbf{w}}) \\ &= \mathbf{B}\text{Var}(\hat{\mathbf{w}})\mathbf{B}^\top \\ &= \mathbf{B}(-\mathbf{H}^{-1} + \Sigma)\mathbf{B}^\top \\ &= \mathbf{B}(-\mathbf{H})^{-1}\mathbf{B}^\top + \mathbf{B}\Sigma\mathbf{B}^\top \\ &= \mathbf{B}(-\mathbf{H})^{-1}\mathbf{B}^\top + (\mathbf{X}^\top \Sigma^{-1} \mathbf{X})^{-1}. \end{aligned}$$

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

These results are important because they justify analytic (i.e., closed-form) solutions for  $\hat{\beta}$  and its associated variance. Analytic solutions are useful because they bypass the need for sampling-based strategies to evaluate the mean and variance of  $\hat{\beta}$ , a common technique for other approaches to SPGLMs like Bayesian MCMC that can be computationally intensive.

### 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 extreme response value relative to

209 expectation. The response GLM residuals simply compare the observation to its fitted latent  
 210 mean:

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

211 Because observations often have a unique support in a GLM (e.g., only two possible response  
 212 values for binary data) and the variance of an observation generally depends on its mean,  
 213 response residuals lack some utility. Deviance residuals are a function of response residuals  
 214 that are appropriately scaled to behave more like response residuals in a standard linear  
 215 model. Deviance residuals are given by

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

216 where  $\mathbf{d}$  is a vector of individual deviances. The sum of the squared deviance residuals equals  
 217 the sum of the elements of  $\mathbf{d}$ , known as the deviance of the model fit. The deviance of the  
 218 model fit quantifies twice the difference in log likelihoods between the a saturated model that  
 219 fits every observation perfectly (i.e.,  $y_i = f^{-1}(\hat{w}_i)$  for all  $i$ ) and the fitted model (Myers,  
 220 Montgomery, Vining, and Robinson 2012). Deviance is often used as a fit statistic; lower  
 221 values of deviance imply a better model fit (compared to the observed data). Pearson and  
 222 standardized residuals are other types of GLM residuals that involve a scaling of the response  
 223 residuals. The Pearson residuals scale  $\mathbf{r}_r$  by the square root of  $\mathbf{V}$ , a diagonal matrix with  
 224  $i$ th diagonal element equal to the variance of the response distribution evaluated at  $f^{-1}(\hat{w}_i)$   
 225 (Faraway 2016);  $\mathbf{V}$  is sometimes called the GLM weight matrix. The standardized residuals  
 226 scale the deviance residuals by  $\frac{1}{\sqrt{(1-\mathbf{L}_{ii})}}$ , where  $\mathbf{L}_{ii}$  is the  $i$ th diagonal element of the leverage  
 227 matrix, which we discuss next.

228 An observation has high leverage if its combination of explanatory variables is far away from  
 229 other observations. In a linear model, the leverage (i.e., hat) values are the diagonal of the  
 230 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  
 231 given by

$$\mathbf{L} = \mathbf{V}^{1/2} \mathbf{X} (\mathbf{X}^\top \mathbf{V} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{V}^{1/2}.$$

232 The larger the value of  $\mathbf{L}_{ii}$ , the more severe the leverage from the  $i$ th observation.  
 233 An observation is influential if it has a sizable impact on model fit. Influence is measured  
 234 using Cook's distance (Cook 1979; Cook and Weisberg 1982), which is given for a GLM by

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

235 where  $\mathbf{r}_s^2$  are the standardized residuals and  $\text{diag}(\mathbf{L})$  indicates the diagonal elements of the  
 236 leverage matrix. The larger the value of  $c_i$ , the more severe the influence from the  $i$ th obser-  
 237 vation. Montgomery *et al.* (2021) provide guidance for interpreting these types of statistics,  
 238 including cutoffs to consider when identifying extreme residual, leverage, or influence values.  
 239 In a linear model, the  $R^2$  (R-squared) statistic quantifies the proportion of variability in the  
 240 data captured by the explanatory variables. It is calculated as one minus the ratio of the error  
 241 sum of squares to the total sum of squares (Rencher and Schaalje 2008). In a GLM, there  
 242 are many ways to define a statistic that emulates the aforementioned meaning of  $R^2$  from the  
 243 linear model (Smith and McKenna 2013). This statistic is called a pseudo R-squared ( $PR^2$ ).

<sup>244</sup> One  $PR^2$  for GLMs simply replaces the sums of squares ratio from the linear model with the  
<sup>245</sup> deviance ratio:

$$PR^2 = 1 - \frac{\text{deviance}_{\text{error}}}{\text{deviance}_{\text{total}}},$$

<sup>246</sup> where  $\text{deviance}_{\text{error}}$  is the deviance of the fitted model (sometimes called the error or residual  
<sup>247</sup> deviance) and  $\text{deviance}_{\text{total}}$  is the deviance of the intercept-only model (sometimes called the  
<sup>248</sup> total or null deviance). In practice,  $\text{deviance}_{\text{total}}$  is derived by computing  $\hat{\mathbf{w}}$  when  $\mathbf{X} \equiv \mathbf{1}$   
<sup>249</sup> (a column of all ones), given  $\hat{\boldsymbol{\theta}}$  and  $\hat{\varphi}$  from the fitted model. Like  $R^2$ ,  $PR^2$  can be adjusted  
<sup>250</sup> to account for the numbers of parameters estimated in a model. Because the  $\text{deviance}_{\text{total}}$   
<sup>251</sup> denominator changes across fitted models (as the values of  $\hat{\boldsymbol{\theta}}$  and  $\hat{\varphi}$  change), this statistic  
<sup>252</sup> should not be used as a model comparison tool. Rather, it should be used as an informative  
<sup>253</sup> diagnostic tool that is unique to each model fit and describes how much variability from that  
<sup>254</sup> model is attributable to the explanatory variables.

## <sup>255</sup> 2.4. Predicting at new locations

<sup>256</sup> We may also predict values of the latent mean (on the link scale) at new locations by leveraging  
<sup>257</sup> the spatial covariance between observed locations and new locations (spatial prediction is  
<sup>258</sup> also called Kriging; see Cressie (1990)). Like in Section 2.2, suppose first that we observed  
<sup>259</sup>  $\mathbf{w}$  and we want to make predictions at  $\mathbf{u}$ , a vector of latent means at the new locations that  
<sup>260</sup> follows the same SPGLM from Equation 3 and has design matrix,  $\mathbf{X}_u$ . The vector  $(\mathbf{w}, \mathbf{u})^\top$   
<sup>261</sup> has expectation  $(\mathbf{X}\boldsymbol{\beta}, \mathbf{X}_u\boldsymbol{\beta})^\top$  and covariance matrix  $\begin{bmatrix} \Sigma & \Sigma_{\mathbf{w}\mathbf{u}} \\ \Sigma_{\mathbf{u}\mathbf{w}} & \Sigma_{\mathbf{u}\mathbf{u}} \end{bmatrix}$ , where  $\Sigma = \text{Var}(\mathbf{w}, \mathbf{w})$ ,  
<sup>262</sup>  $\Sigma_{\mathbf{w}\mathbf{u}} = \text{Var}(\mathbf{w}, \mathbf{u})$ ,  $\Sigma_{\mathbf{u}\mathbf{w}} = \Sigma_{\mathbf{w}\mathbf{u}}^\top$  and  $\Sigma_{\mathbf{u},\mathbf{u}} = \text{Var}(\mathbf{u}, \mathbf{u})$ . Thus we may derive the conditional  
<sup>263</sup> distribution of  $\mathbf{u}|\mathbf{w}$ , which has the following properties:

$$\begin{aligned} E(\mathbf{u}|\mathbf{w}) &= \mathbf{X}_u\boldsymbol{\beta} + \Sigma_{\mathbf{u},\mathbf{w}}\Sigma^{-1}(\mathbf{w} - \mathbf{X}\boldsymbol{\beta}) \\ \text{Var}(\mathbf{u}|\mathbf{w}) &= \Sigma_{\mathbf{u},\mathbf{u}} - \Sigma_{\mathbf{u},\mathbf{w}}\Sigma^{-1}\Sigma_{\mathbf{w},\mathbf{u}} \end{aligned}$$

<sup>264</sup> Recall, however, that we do not actually observe  $\mathbf{w}$  and instead compute  $\hat{\mathbf{w}}$ ; so, to predict  
<sup>265</sup>  $\mathbf{u}$  and quantify its uncertainty, we must again leverage the laws of total expectation and  
<sup>266</sup> variance. Ver Hoef *et al.* (2024) show that  $\hat{\mathbf{u}}$  and its associated variance are given by:

$$\begin{aligned} \hat{\mathbf{u}} &= \mathbf{X}_u\hat{\boldsymbol{\beta}} + \Sigma_{\mathbf{u},\mathbf{w}}\Sigma^{-1}(\hat{\mathbf{w}} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ \text{Var}(\hat{\mathbf{u}}) &= \Sigma_{\mathbf{u},\mathbf{u}} - \Sigma_{\mathbf{u},\mathbf{w}}\Sigma^{-1}\Sigma_{\mathbf{w},\mathbf{u}} + \mathbf{K}(\mathbf{X}^\top\Sigma^{-1}\mathbf{X})^{-1}\mathbf{K}^\top + \boldsymbol{\Lambda}(-\mathbf{H})^{-1}\boldsymbol{\Lambda}^\top, \end{aligned}$$

<sup>267</sup> where  $\mathbf{K} = \mathbf{X}_u - \Sigma_{\mathbf{u},\mathbf{w}}\Sigma^{-1}\mathbf{X}$  and  $\boldsymbol{\Lambda} = \mathbf{X}_u\mathbf{B} + \Sigma_{\mathbf{u},\mathbf{w}}\Sigma^{-1}(\mathbf{1} - \mathbf{X}\mathbf{B})$  for a vector of ones,  $\mathbf{1}$ . As  
<sup>268</sup> with  $\hat{\boldsymbol{\beta}}$ , these covariance matrices are evaluated at  $\hat{\boldsymbol{\theta}}$  in practice.

## 3. Modeling moose presence in Alaska, USA

<sup>269</sup> The `moose` data in **spmodel** contain information on moose (*Alces Alces*) presence in the Togiak  
<sup>270</sup> region of Alaska, USA. `moose` is an `sf` object, a special data frame that is supplemented with  
<sup>271</sup> spatial information using the `sf` package in R (Pebesma 2018). After loading **spmodel**, the  
<sup>272</sup> first few rows of `moose` look like:

```
R> library("spmodel")

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)
```

273 There are five columns in `moose`: `elev`, the numeric site elevation (meters); `strat` a stratification variable for sampling with two levels ("L" and "M") categorized by landscape metrics  
 274 at each site; `count`, the number of moose at each site; `presence`, a factor that indicates  
 275 whether at least one moose was observed at each site (0 implies no moose; 1 implies at least  
 276 one moose); and `geometry`, the NAD83/Alaska Albers (EPSG: 3338) projected coordinate of  
 277 each site. These data have point-referenced support because each observation occurs at point  
 278 coordinates represented by a POINT geometry. Moose are most prevalent in the southwestern  
 279 and eastern parts of the Togiak region (Figure 3).

281 The `moose_preds` data in `spmodel` is an `sf` object with point locations at which moose  
 282 presence predictions are desired. Like `moose`, `moose_preds` contains `elev` and `strat` for each  
 283 site:

```
R> head(moose_preds)

Simple feature collection with 6 features and 2 fields
Geometry type: POINT
Dimension:     XY
Bounding box:  xmin: 291839.8 ymin: 1436192 xmax: 401239.6 ymax: 1512103
Projected CRS: NAD83 / Alaska Albers
# A tibble: 6 x 3
  elev strat      geometry
  <dbl> <chr>    <POINT [m]>
1 143. L    (401239.6 1436192)
2 324. L    (352640.6 1490695)
3 158. L    (360954.9 1491590)
4 221. M    (291839.8 1466091)
5 209. M    (310991.9 1441630)
6 218. L    (304473.8 1512103)
```

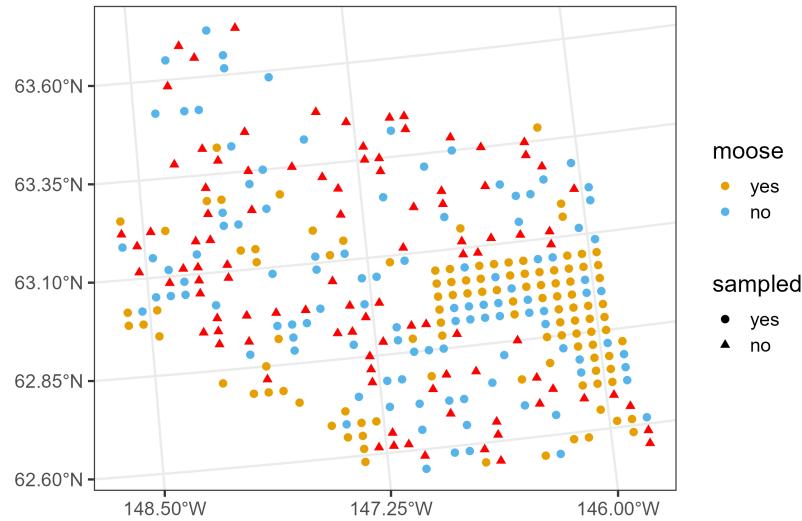


Figure 3: Moose presence in Alaska. Circles represent moose presence or absence (based on color) and triangles represent locations at which moose presence probability predictions are desired.

### 284 3.1. Model Fitting

285 SPGLMs in **spmodel** for point-referenced support are fit using the **spglm()** function. The  
 286 **spglm()** function requires four arguments: **formula**, the relationship between the response  
 287 and explanatory variables; **family**, the response distribution assumed for the response vari-  
 288 able; **data**, the data frame that contains the variables in **formula**, and **spcov\_type**, the type  
 289 of spatial covariance. The **formula**, **family**, and **data** arguments are familiar because they  
 290 are the three required arguments to **glm()** for nonspatial GLMs. So the transition from **glm()**  
 291 to **spglm()** simply requires one additional argument: **spcov\_type**. When **data** is not an **sf**  
 292 object, **spglm()** also requires the **xcoord** and **ycoord** arguments, which indicate the columns  
 293 in **data** that represent the projected x- and y-coordinates, respectively.

294 We use **spglm()** to fit a SPGLM (i.e., here, a spatial logistic regression model) quantifying  
 295 the effect of elevation and strata on moose presence:

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

296 The **summary()** function returns a model summary with relevant information like the function  
 297 call, deviance residuals, a coefficients table of fixed effects, the pseudo R-squared, spatial  
 298 covariance parameters, 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

```

- 299 The model provides some evidence that elevation is positively associated with the log odds  
 300 of moose presence ( $p$  value  $\approx 0.087$ ), after controlling for strata. The model also provides  
 301 strong evidence that moose have a higher log odds of presence in the "M" strata compared to  
 302 the "L" strata ( $p$  value  $< 0.001$ ), after controlling for elevation.  
 303 The fixed effects coefficients table from `summary()` is often of primary scientific interest, but  
 304 it is not immediately usable when printed directly to the R console. The `tidy()` function  
 305 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   -5.38e+0   0.447
2 elev        0.00604   0.00353     1.71  0.0869   -8.73e-4   0.0129
3 stratM      1.44       0.421     3.42  0.000622  6.15e-1    2.26

```

306 **3.2. Model Comparison**

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

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

313 Using `spglm()` instead of `glm()` ensures that **spmodel** helper functions are available and that  
 314 each of the `spglm()` models uses the same hierarchical likelihood (Equation~4):

```
R> glance(spbin)

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

```
R> glance(bin)

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

315 The likelihood-based statistics AIC, AICc, BIC, and deviance are much lower for the SPGLM,  
 316 indicating a better fit relative to the GLM.

317 Instead of relying on likelihood-based statistics, models can be compared using a cross-  
 318 validation procedure (James, Witten, Hastie, and Tibshirani 2013). The `loocv()` function  
 319 performs leave-one-out cross validation, comparing the predicted mean (on the response scale)  
 320 to the observed response variable for each hold-out observation, recomputing estimates of  $\beta$   
 321 in each iteration. Performing leave-one-out cross validation tends to be more computationally  
 322 efficient than fitting the model, as leave-one-out cross validation requires only one set of prod-  
 323 ucts involving the inverse covariance matrix (a primary computational burden), while fitting  
 324 traditional models requires these products for each optimization iteration. After performing  
 325 leave-one-out cross validation, statistics like bias, mean-squared-prediction error (MSPE), and  
 326 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
```

327 Both models have negligible bias, but the SPGLM has much lower MSPE and RMSPE than  
 328 the GLM, indicating the SPGLM predictions are far more efficient. Both the likelihood-based  
 329 leave-one-out cross validation metrics prefer the SPGLM to the GLM.

330 We may also want to compare the fit of two SPGLMs with different spatial covariance struc-  
 331 tures. For example, we may want to compare the fit of an SPGLM with the "exponential"  
 332 spatial covariance to the fit of a SPGLM with the "gaussian" spatial covariance:

```
R> spbin2 <- update(spbin, spcov_type = "gaussian")
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  674.  680.  680.  690. -337.    198.
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 -0.000261 0.146 0.382
```

333 The SPGLM with the "exponential" spatial covariance (`spbin`) has a slightly lower (better)  
 334 deviance but slightly higher (worse) AIC, AICc, and BIC than the SPGLM with the  
 335 "gaussian" spatial covariance (`spbin2`). Both SPGLMs have similar leave-one-out cross  
 336 validation metrics, though the SPGLM with the "gaussian" spatial covariance has slightly  
 337 lower (better) RMSPE. For practical purposes, these models fit similarly.  
 338 Frequently in spatial statistics, the difference in model fit between the best spatial model  
 339 and worst spatial model is much smaller than the difference in model fit between the worst  
 340 spatial model and the nonspatial model, implying that accounting for some form of spatial  
 341 covariance is very beneficial. Two spatial covariance functions to consider starting with are the  
 342 exponential and Gaussian, which have quite different origin behaviors (Figure 2), something  
 343 Stein (1999) argues is important to characterize accurately.

### 344 3.3. Model Diagnostics

345 `spmodel` provides a suite of tools for model diagnostics. One is `augment()`, which augments  
 346 the data used in the model with several model diagnostics (introduced in Section 2.3):

```
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
```

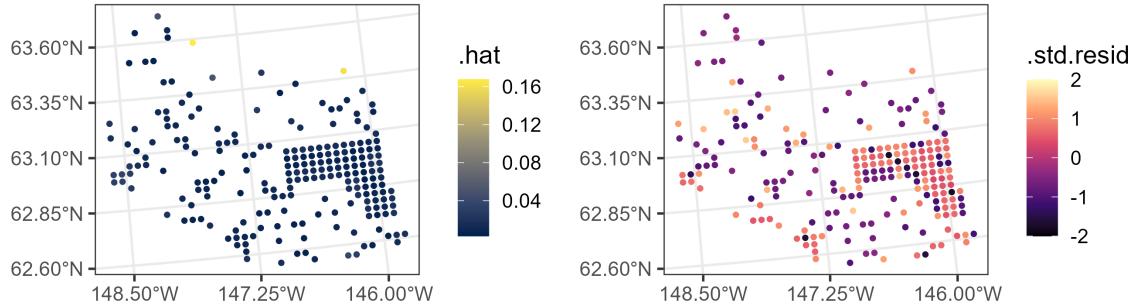


Figure 4: Moose presence model diagnostics, including leverage values (left) and standardized residuals (right).

```

presence elev strat .fitted .resid      .hat   .cooksdi .std.resid
* <fct>    <dbl> <chr> <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]>
```

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

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

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

358 The `varcomp()` function partitions model variability into several different components, helping  
 359 to elucidate the model's structure:

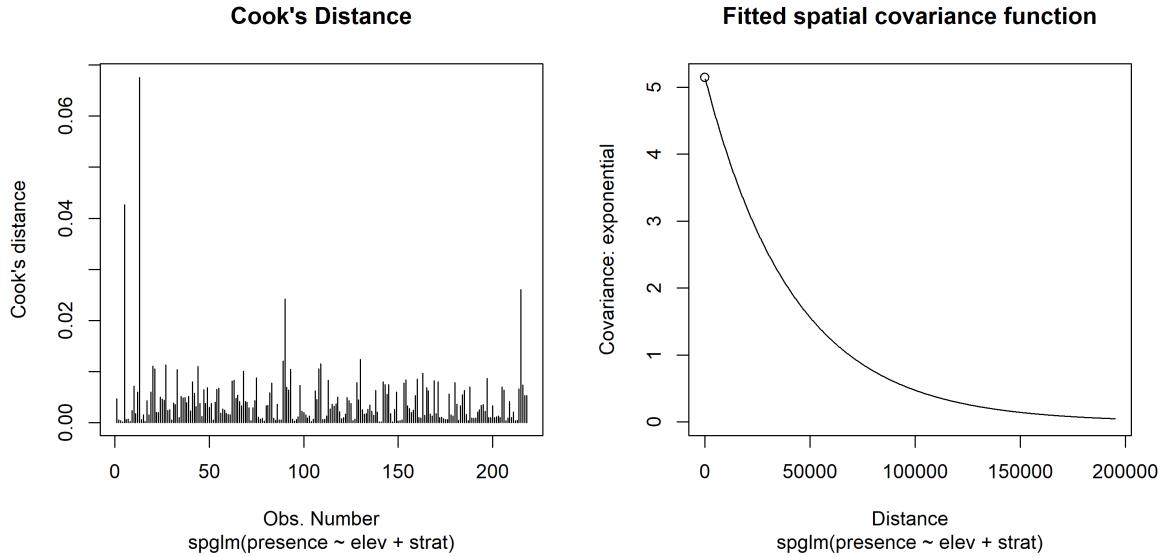


Figure 5: Moose presence model diagnostics, including Cook's distance (left) and the fitted spatial covariance as a function of distance (right).

```
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
```

360 The pseudo R-squared ( $PR^2$ ) is reported in the first row. The remaining variability ( $1 - PR^2$ )  
 361 is allocated proportionally to `de` and `ie` according to  $\sigma_{de}^2$  and  $\sigma_{ie}^2$ . This variability partitioning  
 362 is a useful tool that helps quantify how much the explanatory variables, residual spatial  
 363 variance, and residual nonspatial variance contribute to model fit; as with  $PR^2$ , it should not  
 364 be used for model comparison, but rather as a helpful model diagnostic.

### 365 3.4. Prediction

366 We can predict the probability of moose presence at the locations in `moose_preds` using  
 367 `predict()`:

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

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

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

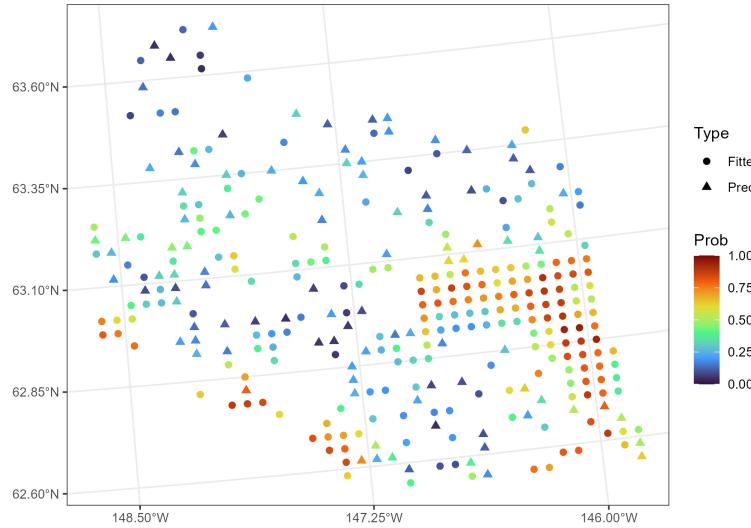


Figure 6: Moose presence probability fitted values ( $f^{-1}(\hat{\mathbf{w}})$ ) and predictions. Fitted values are represented by circles and predictions by triangles.

```
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

370 Moose are far more common in the eastern and southwestern portion of the domain (Figure 6).  
 371 Prediction intervals for the probability of moose presence (on the link scale) are returned by  
 372 supplying `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

373 We can alternatively use `augment()` to augment the prediction data with predictions. Argu-  
 374 ments 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
```

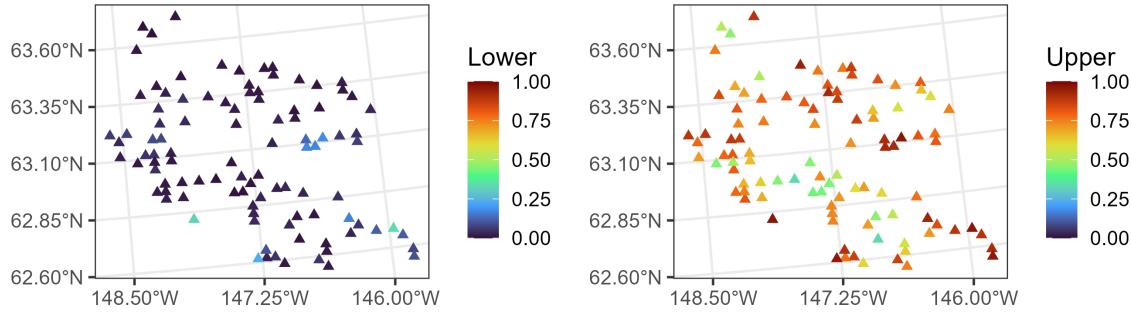


Figure 7: Moose presence 95% prediction interval lower bounds (left) and upper bounds (right).

```
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
```

<sup>375</sup> By using `augment()` when `newdata` is an `sf` object, predictions and their corresponding  
<sup>376</sup> uncertainties are readily available for spatial mapping (Figure 7).

## 4. Additional applications

<sup>377</sup> Throughout the remainder of this section, we briefly highlight some additional **spmodel** ca-  
<sup>378</sup> pabilities for SPCGLMs. In Section 4.1, we fit Poisson and negative binomial models with and  
<sup>379</sup> without geometric anisotropy for the point-referenced moose count data. In Section 4.2, we  
<sup>380</sup> fit a Gamma model to the point-referenced lake conductivity data, showing how to fit a model  
<sup>381</sup> with a partition factor, perform a spatial analysis of variance (ANOVA), and estimate con-  
<sup>382</sup> trasts for models with interactions. In Section 4.3, we fit a binomial model to the areal harbor  
<sup>383</sup> seal trend data with a nonspatial random effect. Finally in Section 4.4, we fit beta models to  
<sup>384</sup> Texas voter turnout data, which can be modeled using point-referenced or areal support, and  
<sup>385</sup> use maximum likelihood to compare two models with different explanatory variables. Table 1  
<sup>386</sup> outlines, for each application, the section number, data set, family (i.e., response distribu-

Section	Data	Family	Geometry	Additional Features
4.1	Moose Counts	Poisson NBinomial	Point	Geometric Anisotropy
4.2	Lake Conductivity	Gamma	Point	Partition Factor ANOVA Contrasts
4.3	Harbor Seals	Binomial	Areal	Nonspatial Random Effects
4.4	Texas Voter Turnout	Beta	Point Areal	Likelihood-Ratio Test

Table 1: Section number, data set, family, geometry type, and additional features for each application.

387     tion), geometry type (point-referenced or areal support), and additional **spmodel** features  
 388     highlighted.

#### 389     **4.1. Modeling moose counts in Alaska, USA**

390     In addition to moose presence, moose counts are also recorded in `moose` (Figure 8). The  
 391     Poisson and negative binomial response distributions can be used to model SPGLMs for count  
 392     data. The Poisson distribution mean is equal to its variance, while the negative binomial has  
 393     an extra parameter to accommodate overdispersion (where the variance is larger than the  
 394     mean). Using a spherical spatial covariance function, we may fit both a Poisson and negative  
 395     binomial SPGLM changing the `family` argument:

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

396     Because the Poisson and negative binomial distributions have the same response support  
 397     (nonnegative integers), we can compare them using AIC, AICc, or BIC:

```
R> BIC(sppois, spnb)
```

df	BIC
sppois	3 1344.574
spnb	4 1343.105

398     Implicit in our spatial covariance functions thus far has been an assumption of geometric  
 399     isotropy. A spatial covariance function is geometrically isotropic if it decays with distance  
 400     at the same rate in all directions (Figure 9; left). A spatial covariance is geometrically  
 401     anisotropic if it decays with distance at different rates in different directions (Figure 9; right).

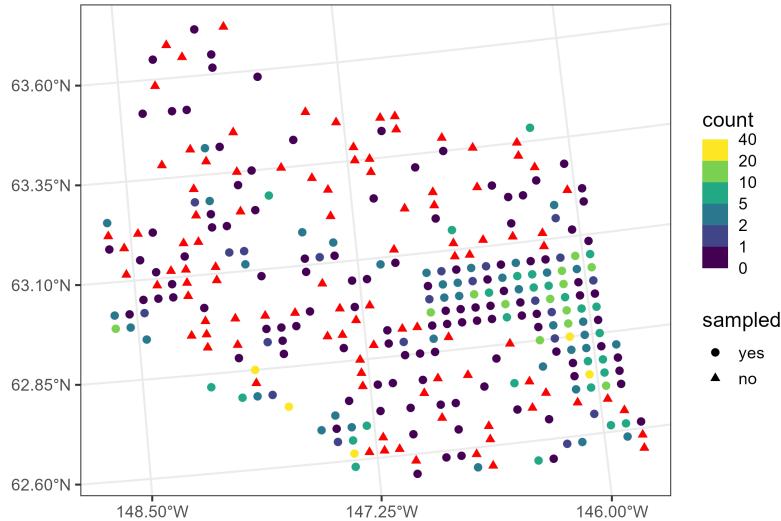


Figure 8: Moose counts in Alaska. Circles represent moose counts (based on color) and triangles represent locations at which mean count predictions are desired.

402 Geometric anisotropy is formally incorporated by rotating and scaling original coordinates,  
403 yielding transformed coordinates that are geometrically isotropic:

$$\begin{bmatrix} x^* \\ y^* \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1/\omega \end{bmatrix} \begin{bmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

404 The parameters  $\omega$  and  $\alpha$  controls the scaling and rotation, respectively, of the major and  
405 minor axes of a level curve of equal spatial covariance (Figure 9). Using these transformed  
406 coordinates, the partial sill ( $\sigma_{de}^2$ ), nugget ( $\sigma_{ie}^2$ ), and range ( $\phi$ ) parameters are estimated. We  
407 accommodate geometric anisotropy by supplying `anisotropy`:

```
R> sppois_anis <- update(sppois, anisotropy = TRUE)
R> spnb_anis <- update(spnb, anisotropy = TRUE)
```

408 According to BIC, the spatial negative binomial model with geometric anisotropy performs  
409 best:

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

	df	BIC
sppois	3	1344.574
spnb	4	1343.105
sppois_anis	5	1341.143
spnb_anis	6	1339.714

410 The `plot()` function can be used to visualize the anisotropy (Figure 9):

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

411 From Figure 9, the spatial covariance is strongest in a northwest-southeast direction (2.729  
412 radians) and weakest in the northeast-southwest direction (0.413 radians), which is intuitive  
413 given the similar patterns in moose counts from Figure 8.

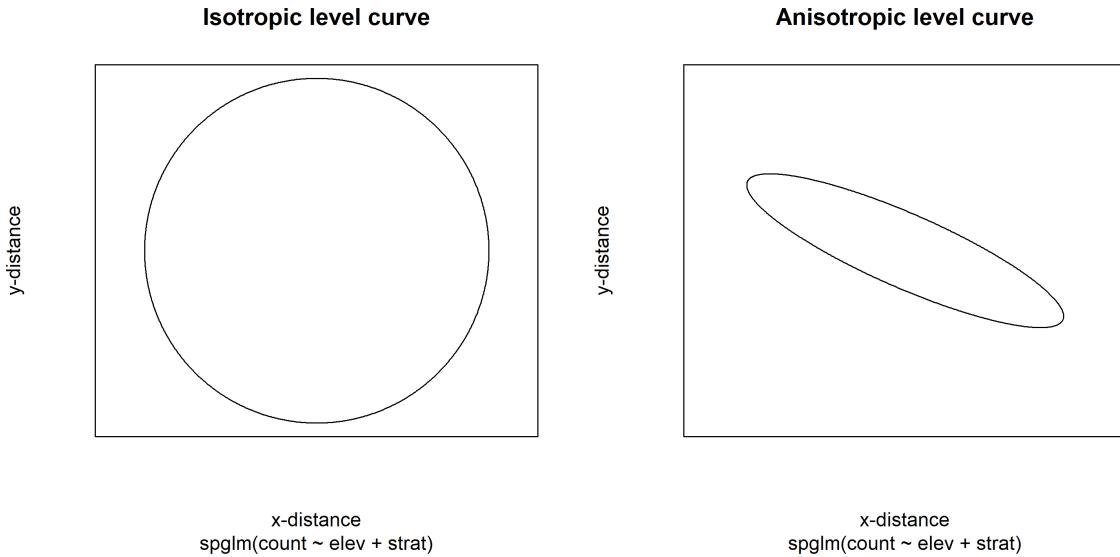


Figure 9: Level curves of equal spatial covariance for the negative binomial moose count models. The ellipse is centered at zero distance in the x-direction and y-direction, and points along the ellipse have equal levels of spatial covariance. In the isotropic level curve (left), spatial covariance decays equally in all directions. In the anisotropic level curve (right), spatial covariance decays fastest in the northeast-southwest direction (0.413 radians) and slowest in the northwest-southeast direction (2.729 radians).

#### 414 4.2. Modeling lake conductivity in Southwest, USA

415 The `lake` data in `spmodel` contains climate and chemical data for several lakes in four south-  
416 western states in the United States: Arizona, Colorado, Nevada, and Utah. We desire an  
417 SPGLM that characterizes the effect of temperature, state, and lake origin (whether the lake  
418 is naturally occurring or human made) on lake conductivity. Conductivity is a measure of  
419 dissolved ions (measured here in water), which is important for various physical, chemical,  
420 and biological processes. Chemical data are often heavily right-skewed, so we model them  
421 using an SPGLM assuming a Gamma distribution for the response. The `log_cond` variable  
422 in `lake` is the logarithm of conductivity, which we dynamically exponentiate within `formula`  
423 so that it is on the original scale:

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

- 424 We model conductivity as a function of temperature, state, and lake origin, and we allow  
 425 the effect of temperature to vary by state (`temp:state` interaction). The `year` partition  
 426 factor (specified via `partition_factor`) restricts spatial covariance to be nonzero only for  
 427 observations sampled during the same year. Data were collected in 2012 and 2017, so this  
 428 partition factor assumes independence between observations in 2012 and 2017. While we used  
 429 the partition factor here illustratively, more generally, the utility of partition factors can be  
 430 highly context dependent.
- 431 When categorical variables have more than two levels, the default reference group contrasts  
 432 are not well-suited to assess the variable's overall significance:

```
R> summary(spgam)
```

```
Call:
spglm(formula = exp(log_cond) ~ temp * state + origin, family = "Gamma",
      data = lake, spcov_type = "cauchy", partition_factor = ~year)

Deviance Residuals:
    Min      1Q   Median      3Q      Max
-1.35762 -0.20796 -0.03706  0.17869  1.10616

Coefficients (fixed):
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.59325   0.50058  7.178 7.06e-13 ***
temp        0.15182   0.03006  5.051 4.39e-07 ***
stateCO     -0.03214   0.56098 -0.057  0.95432
stateNV      0.75664   0.66851  1.132  0.25771
stateUT     -0.19696   0.55916 -0.352  0.72466
originNATURAL 0.08313   0.21988  0.378  0.70538
temp:stateCO 0.13679   0.04808  2.845  0.00444 **
temp:stateNV 0.01882   0.05820  0.323  0.74645
temp:stateUT 0.20015   0.04846  4.131 3.62e-05 ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Pseudo R-squared: 0.7061

Coefficients (cauchy spatial covariance):
      de      ie      range      extra
2.069e-02 2.952e-01 4.119e+06 5.645e-01

Coefficients (Dispersion for Gamma family):
dispersion
      3.761
```

- 433 A more effective approach is to use an analysis of variance (ANOVA), which is well-suited to  
 434 assess the overall significance of each variable:

```
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
```

- 435 The main effect for temperature and the temperature by state interaction are highly significant  
 436 (*p* value < 0.001), while the main effects for state and lake origin are not significant.  
 437 Variance inflation factors assess the degree to which standard errors  $\hat{\beta}$  are inflated due to  
 438 covariance among the columns of **X**. Generalized variance inflation factors can capture the  
 439 variance inflation for subsets of **X** that may include categorical variables with more than two  
 440 levels (Fox and Monette 1992):

```
R> library("car")
```

```
R> vif(spgam)
```

	GVIF	Df	GVIF <sup>(1/(2*Df))</sup>
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

- 441 The GVIF<sup>1/2df</sup> values for **temp**, **state**, and **temp:state** are just greater than two, which  
 442 suggests moderate multicollinearity for these terms – unsurprising given the **temp:state**  
 443 interaction in the model. The GVIF<sup>1/2df</sup> for **origin** is close to one, which suggests little to  
 444 no multicollinearity for this term.  
 445 Because of the interaction between **temp** and **state**, contrasts that assess mean differences  
 446 among states should condition upon a specific temperature value. By default, **emmeans** uses  
 447 the mean temperature value (here, 7.63) to assess contrasts:

```
R> library("emmeans")
```

```
R> pairs(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

- 448 Again, because of the interaction between `temp` and `state`, we should assess temperature  
 449 trends separately for each state:

```
R> 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

450 **4.3. Modeling harbor seal trends in Alaska, USA**

- 451 The `seal` data in `spmodel` contains harbor seal abundance trends for two different harbor  
 452 seal stocks (genetically distinct populations). While the `moose` and `lake` data have point-  
 453 referenced support, the `seal` data have areal support. Each polygon in the `seal` data repre-  
 454 sents a distinct harbor seal haulout region (Figure 10). A haulout region is an area of coastal  
 455 rocks that harbor seals go to rest, molt, and give birth.

- 456 For each polygon, a Poisson regression was used to quantify the mean trend in abundance  
 457 over approximately 30 years (Ver Hoef, Peterson, Hooten, Hanks, and Fortin 2018). If the  
 458 logarithm of mean abundance trends (`log_trend`) is negative (positive), it means abundance  
 459 is decreasing (increasing). We use a binomial SPGLM to quantify the likelihood that mean  
 460 abundance trends are decreasing:

```
R> is_decreasing <- seal$log_trend < 0
R> spbin <- spgautor(
+   formula = is_decreasing ~ 1,
+   family = binomial,
+   data = seal,
+   spcov_type = "car",
```

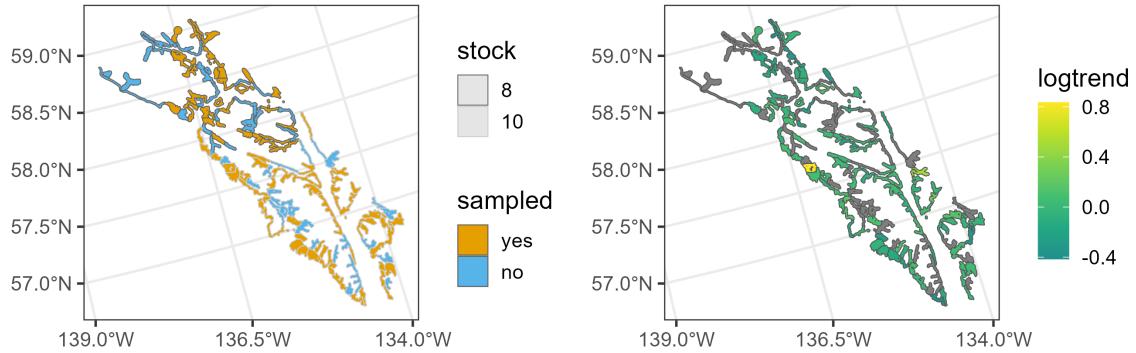


Figure 10: Seal trend distribution in Alaska. Observed and missing seal polygons by stock (left) and observed log seal trends (right).

```
+   random = ~ stock
+ )
```

To model spatial dependence, we used a conditional autoregressive function. Conditional and simultaneous autoregressive functions characterize spatial distance through neighborhood relationships (rather than Euclidean distance) and have `spcov_type` values of "car" and "sar", respectively. By default, Queen's distance is used to determine whether two sites are neighbors, though custom neighborhood matrices can be passed via `W`. Row standardization is also assumed by default; this can be changed via `row_st`. Using `random`, we also specified a nonspatial random effect for seal stock, which implies seals belonging to the same stock share extra covariance. The `random` argument uses similar syntax as `lme4` (Bates, Mächler, Bolker, and Walker 2015) and `nlme` (Pinheiro and Bates 2006) to specify nonspatial random effects.

Tidying the model reveals the estimates and confidence intervals on the log odds scale:

```
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   -0.979    1.66
```

Back-transforming the confidence interval to the probability scale yields:

```
R> emmeans(spbin, ~ 1, type = "response")

1          prob      SE  df asympt.LCL asympt.UCL
overall 0.584 0.164 Inf      0.273      0.84

Degrees-of-freedom method: asymptotic
Confidence level used: 0.95
Intervals are back-transformed from the logit scale
```

<sup>472</sup> The **SE** column is the standard error on the response scale obtained from the delta method  
<sup>473</sup> (Oehlert 1992; Ver Hoef 2012).

<sup>474</sup> In contrast to point-referenced data, prediction locations for areal data must be specified  
<sup>475</sup> at the time of model fitting, as they affect the spatial covariance function's neighborhood  
<sup>476</sup> structure. Prediction locations whose response values have an **NA** (i.e., missing) value are  
<sup>477</sup> converted into a **newdata** object that is stored in the model output. For example, rows one  
<sup>478</sup> and nine are locations without seal trends, meaning they are not used in model fitting but  
<sup>479</sup> are desired for prediction:

```
R> seal
```

```
Simple feature collection with 149 features and 2 fields
Geometry type: POLYGON
Dimension:     XY
Bounding box:  xmin: 913618.8 ymin: 855730.2 xmax: 1221859 ymax: 1145054
Projected CRS: NAD83 / Alaska Albers
# A tibble: 149 x 3
  log_trend stock                                     geometry
  * <dbl> <fct>                                     <POLYGON [m]>
1 NA     8   ((1035002 1054710, 1035002 1054542, 1035002 105354~
2 -0.282 8   ((1037002 1039492, 1037006 1039490, 1037017 103949~
3 -0.00121 8   ((1070158 1030216, 1070185 1030207, 1070187 103020~
4  0.0354 8   ((1054906 1034826, 1054931 1034821, 1054936 103482~
5 -0.0160 8   ((1025142 1056940, 1025184 1056889, 1025222 105683~
6  0.0872 8   ((1026035 1044623, 1026037 1044605, 1026072 104461~
7 -0.266 8   ((1100345 1060709, 1100287 1060706, 1100228 106070~
8  0.0743 8   ((1030247 1029637, 1030248 1029637, 1030265 102964~
9 NA     8   ((1043093 1020553, 1043097 1020550, 1043101 102055~
10 -0.00961 8  ((1116002 1024542, 1116002 1023542, 1116002 102254~
# i 139 more rows
```

<sup>480</sup> Then, **predict()** can be called without having to specify **newdata**:

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

      fit      lwr      upr
1  0.6807677 0.3863736 0.8783808
9  0.5945680 0.2467634 0.8678078
13 0.6189055 0.2974432 0.8616799
15 0.6040102 0.2921802 0.8493132
18 0.6375700 0.3356282 0.8596641
```

<sup>481</sup> We could have alternatively used a (geostatistical) SPGLM via **spglm()**. When areal data are  
<sup>482</sup> used with **spglm()**, the centroids of each polygon are used as the point-referenced coordinates.

#### <sup>483</sup> 4.4. Modeling voter turnout in Texas, USA

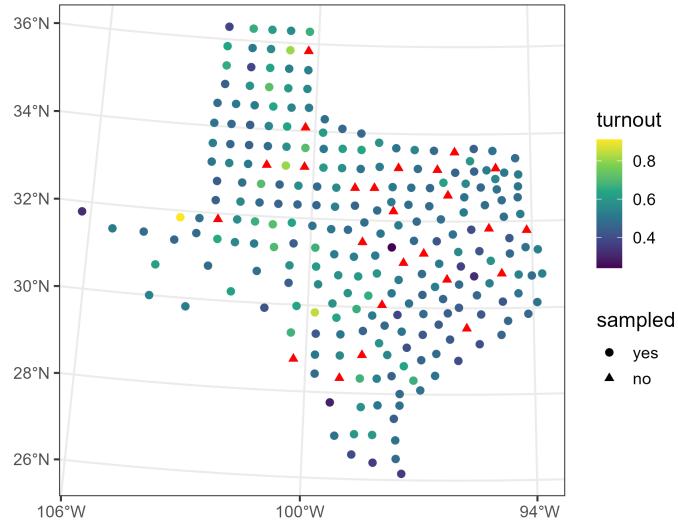


Figure 11: Proportion of voter turnout in Texas for the 1980 presidential election. Circles represent voter turnout (based on color) and triangles represent locations at which voter turnout predictions are desired.

484 The `texas` data in `spmodel` contains voter turnout data for Texas counties in the 1980 United  
 485 States Presidential Election (Bivand, Nowosad, and Lovelace 2024). The data have point-  
 486 referenced support, with polygon centroids representing the spatial location of each county  
 487 (Figure 11). Beta regression is a GLM used to model rate and proportion data in the (0,  
 488 1) interval (Ferrari and Cribari-Neto 2004; Cribari-Neto and Zeileis 2010). We model voter  
 489 turnout rates as a function of mean log income of county residents using a geostatistical  
 490 SPGLM assuming a beta distributed response variable:

```
R> spbeta_geo <- spglm(
+   formula = turnout ~ log_income,
+   family = "beta",
+   data = texas,
+   spcov_type = "matern"
+ )
```

491 Alternatively, we could fit an autoregressive SPGLM by assuming the data have areal support  
 492 and constructing a neighborhood matrix that treats counties as neighbors if the distance  
 493 between their centroids is less than `cutoff`:

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

494 According to AIC, the geostatistical SPGLM:

```
R> AIC(spbeta_geo, spbeta_auto)
```

df	AIC
spbeta_geo	5 -44.53113
spbeta_auto	3 -22.46104

495 The default estimation method in **spmodel** for SPGLMs is restricted maximum likelihood  
496 (REML), while maximum likelihood (ML) can also be used. A benefit of REML is  
497 that it can yield unbiased estimates of covariance parameters (Cressie and Lahiri 1993). A  
498 drawback of REML is that likelihood-based statistics are often considered invalid when the  
499 models have different explanatory variable or fixed effect structures (Wolfinger 1993), though  
500 Gurka (2006) provides some evidence to the contrary. In contrast to REML estimators, ML  
501 estimators are generally biased for covariance parameters, though in practice this bias tends  
502 to be small. Moreover, when using ML, likelihood-based comparisons are valid for models  
503 with different explanatory variable or fixed effect structures. Using ML, we can evaluate the  
504 significance of log income on voter turnout using a likelihood ratio test (for nested models):

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

#### Likelihood Ratio Test

Response: turnout

Df	Chi2	Pr(>Chi2)
----	------	-----------

spbeta_reduced_ml vs spbeta_full_ml	1	23.155	1.494e-06	***
-------------------------------------	---	--------	-----------	-----

---

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'. '	0.1	' '	1
----------------	---	-------	-------	------	------	-----	------	------	-----	-----	---

505 The likelihood ratio test provides strong evidence that log income is significantly related  
506 to voter turnout ( $p$  value  $< 0.001$ ). Alternatively, we could have instead used a different  
507 likelihood-based statistic like AIC:

```
R> AIC(spbeta_full_ml, spbeta_reduced_ml)
```

df	AIC
spbeta_full_ml	7 -31.25900
spbeta_reduced_ml	6 -10.10354

508 The AIC also prefers the full model, suggesting that log income is important for predicting  
509 voter turnout.

## 5. Discussion

SPGLMs are fit in **spmodel** using a novel application of the Laplace approximation that simultaneously marginalizes over the latent (i.e., unobserved) random effects and the fixed effects. **spmodel**'s `spglm()` (for point-referenced support) and `spgautor()` (for areal support) fit SPGLMs that are similar in structure and syntax as base R's `glm()` function, easing the transition for practitioners from GLMs to SPGLMs. The `spglm()` and `spgautor()` functions support six response distributions for binary, count, and skewed data and 20 spatial covariance functions. **spmodel** has a suite of tools for data visualization, inference, model diagnostics, and prediction, providing a framework that can be used for all stages of a data analysis. There are many additional **spmodel** features that are not covered here, including fitting multiple models simultaneously, fixing spatial covariance and dispersion parameters at known values, fitting models to large non-Gaussian data having thousands of observations via spatial indexing (Ver Hoef, Dumelle, Higham, Peterson, and Isaak 2023), incorporating spatial dependence in machine learning (e.g., random forests; Breiman (2001)), simulating spatially dependent data (e.g., `spbinom()`, `sprpois()`, etc.), and more. Further details are provided by <https://CRAN.R-project.org/package=spmodel> and links therein.

## Data and code availability

The results in this manuscript were obtained using R 4.4.0 with the **spmodel** 0.11.1 package. Figures were created using the `ggplot2` 3.5.1 package (Wickham 2016) and base R. All writing and code associated with this manuscript is available for viewing and download on GitHub at <https://github.com/USEPA/spmodel.glm.manuscript>. All data used are part of the **spmodel** R package available for download from CRAN at <https://CRAN.R-project.org/package=spmodel>.

## Acknowledgments

We would like to genuinely thank the associate editor, anonymous reviewers, and editorial staff for significant support and feedback that greatly improved the manuscript. The views expressed in this article are those of the author(s) and do not necessarily represent the views or policies of the U.S. government, U.S. Environmental Protection Agency or the National Oceanic and Atmospheric Administration. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

## 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](https://doi.org/10.1101/2022.03.24.485545).

- 543 Bachl FE, Lindgren F, Borchers DL, Illian JB (2019). “inlabru: An R Package for Bayesian  
 544 Spatial Modelling from ecological survey data.” *Methods in Ecology and Evolution*, **10**,  
 545 760–766. [doi:10.1111/2041-210X.13168](https://doi.org/10.1111/2041-210X.13168).
- 546 Bates D, Mächler M, Bolker B, Walker S (2015). “Fitting Linear Mixed-Effects Models Using  
 547 lme4.” *Journal of Statistical Software*, **67**(1), 1–48. [doi:10.18637/jss.v067.i01](https://doi.org/10.18637/jss.v067.i01).
- 548 Bivand R, Nowosad J, Lovelace R (2024). *spData: Datasets for Spatial Analysis*. R package  
 549 version 2.3.1, URL <https://CRAN.R-project.org/package=spData>.
- 550 Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White JSS (2009).  
 551 “Generalized Linear Mixed Models: A Practical Guide for Ecology and Evolution.” *Trends  
 552 in Ecology & Evolution*, **24**(3), 127–135.
- 553 Bonat WH, Ribeiro Jr PJ (2016). “Practical Likelihood Analysis for Spatial Generalized  
 554 Linear Mixed Models.” *Environmetrics*, **27**(2), 83–89.
- 555 Breiman L (2001). “Random Forests.” *Machine Learning*, **45**, 5–32.
- 556 Breslow NE, Clayton DG (1993). “Approximate Inference in Generalized Linear Mixed Mod-  
 557 els.” *Journal of the American Statistical Association*, **88**(421), 9–25.
- 558 Brooks ME, Kristensen K, van Benthem KJ, Magnusson A, Berg CW, Nielsen A, Skaug  
 559 HJ, Maechler M, Bolker BM (2017). “glmmTMB Balances Speed and Flexibility Among  
 560 Packages for Zero-Inflated Generalized Linear Mixed Modeling.” *The R Journal*, **9**(2),  
 561 378–400. [doi:10.32614/RJ-2017-066](https://doi.org/10.32614/RJ-2017-066).
- 562 Bürkner PC (2017). “brms: An R package for Bayesian Multilevel Models Using Stan.”  
 563 *Journal of Statistical Software*, **80**, 1–28.
- 564 Chambers JM, Hastie TJ (eds.) (1992). *Statistical Models in S*. Chapman & Hall, London.
- 565 Cook RD (1979). “Influential Observations in Linear Regression.” *Journal of the American  
 566 Statistical Association*, **74**(365), 169–174.
- 567 Cook RD, Weisberg S (1982). *Residuals and Influence in Regression*. New York: Chapman  
 568 and Hall.
- 569 Cressie N (1990). “The Origins of Kriging.” *Mathematical Geology*, **22**(3), 239–252.
- 570 Cressie N (1993). *Statistics for Spatial Data*. John Wiley & Sons.
- 571 Cressie N, Lahiri SN (1993). “The Asymptotic Distribution of REML Estimators.” *Journal  
 572 of multivariate analysis*, **45**(2), 217–233.
- 573 Cribari-Neto F, Zeileis A (2010). “Beta Regression in R.” *Journal of statistical software*,  
 574 **34**(1), 1–24.
- 575 Doser JW, Finley AO, Kéry M, Zipkin EF (2022). “spOccupancy: An R Package for Single-  
 576 Species, Multi-Species, and Integrated Spatial Occupancy Models.” *Methods in Ecology and  
 577 Evolution*, **13**(8), 1670–1678.

- 578 Doser JW, Finley AO, Kéry M, Zipkin EF (2024). “spAbundance: An R package for Single-  
579 Species and Multi-Species Spatially Explicit Abundance Models.” *Methods in Ecology and  
580 Evolution*, **15**(6), 1024–1033.
- 581 Dumelle M, Higham M, Ver Hoef JM (2023). “spmodel: Spatial Statistical Modeling and  
582 Prediction in R.” *PLOS ONE*, **18**(3), e0282524.
- 583 Evangelou E, Zhu Z, Smith RL (2011). “Estimation and Prediction for Spatial Generalized  
584 Linear Mixed Models Using High Order Laplace Approximation.” *Journal of Statistical  
585 Planning and Inference*, **141**(11), 3564–3577.
- 586 Faraway JJ (2016). *Extending the Linear Model with R: Generalized Linear, Mixed Effects  
587 and Nonparametric Regression Models*. CRC press.
- 588 Ferrari S, Cribari-Neto F (2004). “Beta Regression for Modelling Rates and Proportions.”  
589 *Journal of Applied Statistics*, **31**(7), 799–815.
- 590 Finley AO, Banerjee S, Carlin BP (2007). “spBayes: An R Package for Univariate and  
591 Multivariate Hierarchical Point-Referenced Spatial Models.” *Journal of Statistical Software*,  
592 **19**(4), 1–24. URL <https://www.jstatsoft.org/article/view/v019i04>.
- 593 Finley AO, Datta A, Banerjee S (2022). “spNNGP R Package for Nearest Neighbor Gaussian  
594 Process Models.” *Journal of Statistical Software*, **103**(5), 1–40. doi:[10.18637/jss.v103.i05](https://doi.org/10.18637/jss.v103.i05).
- 595 596 Fox J, Monette G (1992). “Generalized Collinearity Diagnostics.” *Journal of the American  
597 Statistical Association*, **87**(417), 178–183.
- 598 Fox J, Weisberg S (2019). *An R Companion to Applied Regression*. Third edition. Sage,  
599 Thousand Oaks CA. URL <https://www.john-fox.ca/Companion/>.
- 600 Gurka MJ (2006). “Selecting the Best Linear Mixed Model Under REML.” *The American  
601 Statistician*, **60**(1), 19–26.
- 602 Harville DA (1974). “Bayesian inference for variance components using only error contrasts.”  
603 *Biometrika*, **61**(2), 383–385.
- 604 Hoeting JA, Davis RA, Merton AA, Thompson SE (2006). “Model Selection for Geostatistical  
605 Models.” *Ecological Applications*, **16**(1), 87–98.
- 606 Hughes J, Cui X (2020). *ngspatial: Fitting the Centered Autologistic and Sparse Spatial  
607 Generalized Linear Mixed Models for Areal Data*. Frederick, MD. R package version 1.2-2.
- 608 James G, Witten D, Hastie T, Tibshirani R (2013). *An Introduction to Statistical Learning*.  
609 Springer-Verlag.
- 610 Kuhn M, Silge J (2022). *Tidy Modeling with R*. O'Reilly Media, Inc.
- 611 Lee D (2013). “CARBayes: An R Package for Bayesian Spatial Modeling with Conditional  
612 Autoregressive Priors.” *Journal of Statistical Software*, **55**(13), 1–24.
- 613 Lee Y, Nelder JA (1996). “Hierarchical Generalized Linear Models.” *Journal of the Royal  
614 Statistical Society: Series B (Methodological)*, **58**(4), 619–656.

- 615 Lenth RV (2024). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package  
 616 version 1.10.3, URL <https://CRAN.R-project.org/package=emmeans>.
- 617 Lindgren F, Rue H (2015). “Bayesian Spatial Modelling with R-INLA.” *Journal of Statistical  
 618 Software*, **63**, 1–25.
- 619 McCullagh P, Nelder JA (1989). *Generalized Linear Models, Second Edition*. Chapman and  
 620 Hall Ltd.
- 621 Montgomery DC, Peck EA, Vining GG (2021). *Introduction to Linear Regression Analysis*.  
 622 John Wiley & Sons.
- 623 Myers RH, Montgomery DC, Vining GG, Robinson TJ (2012). *Generalized Linear Models:  
 624 With Applications in Engineering and the Sciences*. John Wiley & Sons.
- 625 Nelder JA, Wedderburn RW (1972). “Generalized Linear Models.” *Journal of the Royal  
 626 Statistical Society A*, **135**(3), 370–384.
- 627 Oehlert GW (1992). “A Note on the Delta Method.” *The American Statistician*, **46**(1), 27–29.
- 628 Patterson D, Thompson R (1971). “Recovery of Inter-Block Information when Block Sizes  
 629 are Unequal.” *Biometrika*, **58**(3), 545–554.
- 630 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](https://doi.org/10.32614/RJ-2018-009). URL [https://doi.org/  
 631 10.32614/RJ-2018-009](https://doi.org/10.32614/RJ-2018-009).
- 633 Pinheiro J, Bates D (2006). *Mixed-Effects Models in S and S-PLUS*. Springer-Verlag Science  
 634 & Business Media.
- 635 R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foun-  
 636 dation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- 637 Rencher AC, Schaalje GB (2008). *Linear Models in Statistics*. John Wiley & Sons.
- 638 Robinson D, Hayes A, Couch S (2021). *broom: Convert Statistical Objects into Tidy Tibbles*.  
 639 R package version 0.7.6, URL <https://CRAN.R-project.org/package=broom>.
- 640 Ronnegard L, Shen X, Alam M (2010). “hglm: A Package for Fitting Hierarchical Generalized  
 641 Linear Models.” *The R Journal*, **2**(2), 20–28.
- 642 Rousset F, Ferdy JB (2014). “Testing Environmental and Genetic Effects in the Presence  
 643 of Spatial Autocorrelation.” *Ecography*, **37**(8), 781–790. URL [https://dx.doi.org/10.  
 644 1111/ecog.00566](https://dx.doi.org/10.1111/ecog.00566).
- 645 Sainsbury-Dale M, Zammit-Mangion A, Cressie N (2024). “Modeling Big, Heterogeneous,  
 646 Non-Gaussian Spatial and Spatio-Temporal Data Using FRK.” *Journal of Statistical Soft-  
 647 ware*, **108**, 1–39.
- 648 Schwarz G (1978). “Estimating the Dimension of a Model.” *The Annals of Statistics*, pp.  
 649 461–464.

- 650 Smith TJ, McKenna CM (2013). “A Comparison of Logistic Regression Pseudo R<sup>2</sup> Indices.”  
651 *General Linear Model Journal*, **39**(2), 17–26.
- 652 Stein ML (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer-Verlag  
653 Science & Business Media.
- 654 Thorson JT, Anderson SC, Goddard P, Rooper CN (2025). “tinyVAST: R Package with an  
655 Expressive Interface to Specify Lagged and Simultaneous Effects in Multivariate Spatio-  
656 Temporal Models.” *Global Ecology and Biogeography*, **34**(4), e70035. doi:10.1111/geb.  
657 70035. URL <https://doi.org/10.1111/geb.70035>.
- 658 Tobler WR (1970). “A Computer Movie Simulating Urban Growth in the Detroit Region.”  
659 *Economic Geography*, **46**(sup1), 234–240.
- 660 Tredennick AT, Hooker G, Ellner SP, Adler PB (2021). “A Practical Guide to Selecting  
661 Models for Exploration, Inference, and Prediction in Ecology.” *Ecology*, **102**(6), e03336.
- 662 Ver Hoef JM (2012). “Who Invented the Delta Method?” *The American Statistician*, **66**(2),  
663 124–127.
- 664 Ver Hoef JM, Blagg E, Dumelle M, Dixon PM, Zimmerman DL, Conn PB (2024). “Marginal  
665 Inference for Hierarchical Generalized Linear Mixed Models with Patterned Covariance  
666 Matrices Using the Laplace Approximation.” *Environmetrics*, **35**(7), e2872. doi:10.1002/  
667 env.2872.
- 668 Ver Hoef JM, Dumelle M, Higham M, Peterson EE, Isaak DJ (2023). “Indexing and Parti-  
669 tioning the Spatial Linear Model for Large Data Sets.” *PLOS ONE*, **18**(11), e0291906.
- 670 Ver Hoef JM, Peterson EE, Hooten MB, Hanks EM, Fortin MJ (2018). “Spatial Autoregressive  
671 Models for Statistical Inference From Ecological Data.” *Ecological Monographs*, **88**(1), 36–  
672 59.
- 673 Wedderburn RW (1974). “Quasi-Likelihood Functions, Generalized Linear Models, and the  
674 Gauss—Newton Method.” *Biometrika*, **61**(3), 439–447.
- 675 Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York.  
676 ISBN 978-3-319-24277-4. URL <https://ggplot2.tidyverse.org>.
- 677 Wolfinger R (1993). “Covariance structure selection in general mixed models.” *Communica-  
678 tions in statistics-Simulation and computation*, **22**(4), 1079–1106.
- 679 Wolfinger R, O’connell M (1993). “Generalized Linear Mixed Models: A Pseudo-Likelihood  
680 Approach.” *Journal of Statistical Computation and Simulation*, **48**(3-4), 233–243.
- 681 Wolfinger R, Tobias R, Sall J (1994). “Computing Gaussian Likelihoods and their Derivatives  
682 for General Linear Mixed Models.” *SIAM Journal on Scientific Computing*, **15**(6), 1294–  
683 1310.
- 684 Wood SN (2017). *Generalized Additive Models: An Introduction with R*. CRC press.
- 685 Zimmerman DL, Ver Hoef JM (2024). *Spatial Linear Models for Environmental Data*. CRC  
686 Press.

687 **Affiliation:**

688 Michael Dumelle  
689 United States  
690 Environmental Protection Agency  
691 200 SW 35th St  
692 Corvallis, OR, 97330  
693 E-mail: [Dumelle.Michael@epa.gov](mailto:Dumelle.Michael@epa.gov)

694