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 $\mathrm{E}(\boldsymbol{\tau}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\tau}) = \sigma^2_\tau \mathbf{R}$, where $\mathrm{E}(\cdot)$ and $\mathrm{Cov}(\cdot)$ 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

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

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

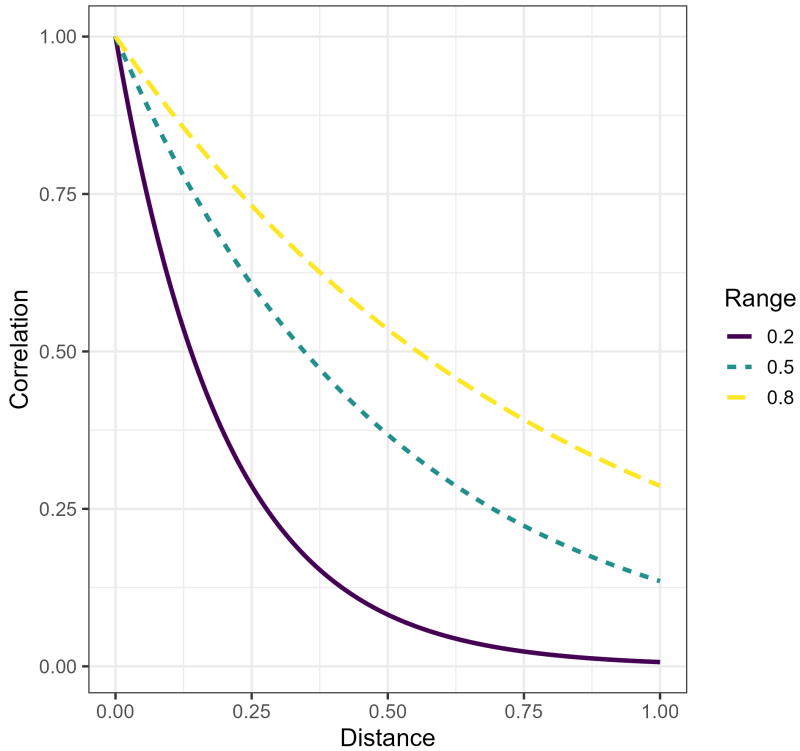

Figure 1: An exponential spatial correlation function with varying range parameters.

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

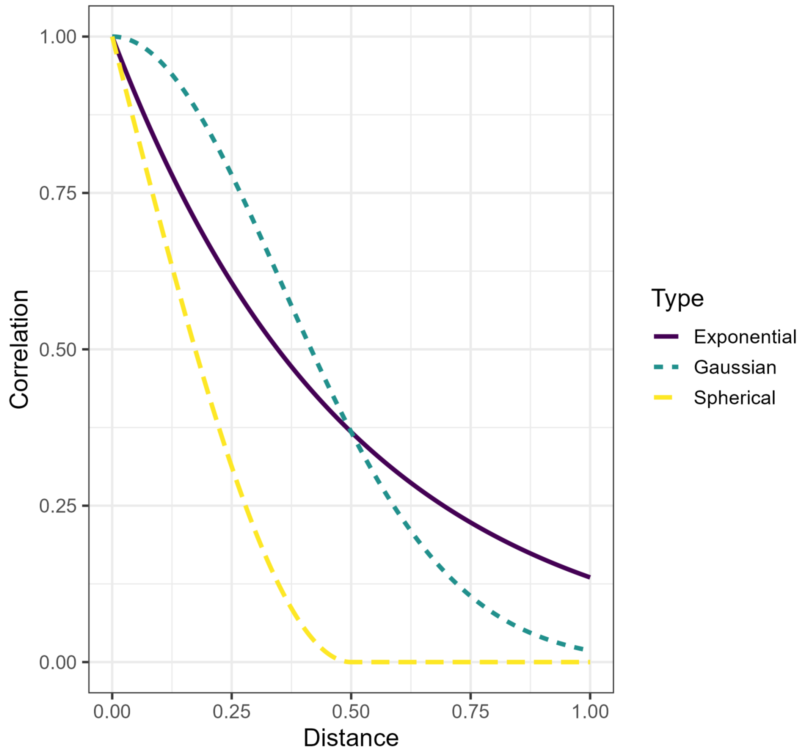

Figure 2: Exponential, Gaussian, and spherical spatial correlation functions all with range parameters equal to 0.5.

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) and **emmeans** (Lenth 2024), 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 ($\boldsymbol{\theta}$). 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

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

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. 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. 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, $\mathbf{X}_\mathbf{u}$. The vector $(\mathbf{w}, \mathbf{u})^\top$ has the following properties:

$$\mathrm{E}(\mathbf{w}, \mathbf{u})^\top = (\mathrm{E}(\mathbf{w}), \mathrm{E}(\mathbf{u}))^\top = (\mathbf{X}\boldsymbol{\beta}, \mathbf{X}_\mathbf{u}\boldsymbol{\beta})^\top \tag{25}$$

$$\mathrm{Var}(\mathbf{w}, \mathbf{u})^\top = \begin{bmatrix} \mathrm{Var}(\mathbf{w}, \mathbf{w}) & \mathrm{Var}(\mathbf{w}, \mathbf{u}) \\ \mathrm{Var}(\mathbf{u}, \mathbf{w}) & \mathrm{Var}(\mathbf{u}, \mathbf{u}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma}_{\mathbf{wu}} \\ \boldsymbol{\Sigma}_{\mathbf{uw}} & \boldsymbol{\Sigma}_{\mathbf{uu}} \end{bmatrix} \tag{26}$$

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

$$\mathrm{E}(\mathbf{w}|\mathbf{u}) = \mathbf{X}_\mathbf{u}\boldsymbol{\beta} + \boldsymbol{\Sigma}_{\mathbf{u},\mathbf{w}}\boldsymbol{\Sigma}^{-1}(\mathbf{w} - \mathbf{X}\boldsymbol{\beta}) \tag{27}$$

$$\mathrm{E}(\mathbf{w}|\mathbf{u}) = \boldsymbol{\Sigma}_{\mathbf{u},\mathbf{u}} - \boldsymbol{\Sigma}_{\mathbf{u},\mathbf{w}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\mathbf{w},\mathbf{u}} \tag{28}$$

Two more sources of uncertainty, $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{w}}$. Spatial prediction is often also called Kriging (Cressie 1990).

3. Modeling moose presence in Alaska, USA

3.1. Additional applications

3.2. Modeling moose counts in Alaska, USA

3.3. Modeling harbor seal trends in Alaska, USA

3.4. Modeling voter turnout in Texas, USA

3.5. Modeling lake conductivity in Southwest, USA

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.

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`.

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.

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.

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/.

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.

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