Wiley StatsRef:
Statistics Reference Online

# Simultaneous Autoregressive (SAR) Model

By *Mevin B. Hooten*[1], *Jay M. Ver Hoef*[2], *and Ephraim M. Hanks*[3]

**Keywords:** *autocorrelation, autoregressive model, dependence, lattice data, spatial statistics*

**Abstract:** Simultaneous autoregressive (SAR) models are useful for accommodating various forms of dependence among data that have discrete support in a space of interest. These models are often specified hierarchically as mixed-effects regression models with first-moment structure controlled by a conventional linear regression term and second-moment structure induced by correlated random effects. In their general form, SAR models resemble conditional autoregressive (CAR) models, and can be made equivalent but are often parameterized differently. Importantly, SAR models can be specified by simultaneously regressing a discrete spatial process on itself. Thus, they allow one to construct statistical models for processes with directional graphical properties that pertain to data generating mechanisms. Most commonly SAR models have been used to account for structure among data with areal spatial support in applications involving ecology, epidemiology, sociology, and environmental science.

## 1 Introduction

In the statistical analysis of autocorrelated data, autoregressive (AR; *see* **Autoregressive – Moving Average (ARMA) Models**) models are among the most commonly used in spatial and temporal statistics for their beneficial properties and ease of specification. AR models account for dependence in a data set and/or process by explicitly accommodating the relationships among observations in a space of interest. The space of interest is often the support of the data or process in the spatial or temporal domain. Thus, statistical AR models are customized for specific types of support, either continuous or discrete.

In the spatial context, an example of a continuous process (*see* **Geostatistics**) is an ambient temperature field that exists everywhere in a study area (even though it is only measured at a finite set of locations). By contrast, a discrete spatial process exists at a countable set of points or regions. An example of a discrete spatial process is disease incidence in predefined geographic regions. AR models for processes that have discrete spatial support are often referred to as lattice or areal data models[1]. These AR models are specified

[1]U.S. Geological Survey and Colorado State University, Fort Collins, CO, USA,
[2]National Oceanic and Atmospheric Administration, Seattle, WA, USA,
[3]The Pennsylvania State University, State College, PA, USA,



DOI: 10.1002/9781118445112.stat08208

from two different perspectives and are referred to as conditional autoregressive (CAR; *see* **Conditional Autoregressive (CAR) Model**) models and simultaneous autoregressive (SAR) models[2]. SAR models are the focus of this exposition.

We introduce SAR models in a classical mixed model framework. Consider a data set $\mathbf{y} \equiv (y_1, \ldots, y_n)'$ that arises from a data generating process comprised of both first-moment (mean) and second-moment (covariance) structure. Assuming linear relationships with a set of covariates $\mathbf{x}_i$ for $i = 1, \ldots, n$, additivity of components, and separability of the first- and second-moment structure, we can write a statistical AR model jointly as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{z} + \boldsymbol{\varepsilon} \tag{1}$$

where the $n \times p$ matrix $\mathbf{X}$ consists of the $p \times 1$ covariate vectors $\mathbf{x}_i'$, for $i = 1, \ldots, n$, on the rows and the coefficient vector is $\boldsymbol{\beta} \equiv (\beta_0, \ldots, \beta_{p-1})'$. In Equation (1), the first-moment structure is deterministic and accommodated by the linear predictor $\mathbf{X}\boldsymbol{\beta}$, which results in a regression model. If we specify $\varepsilon$ as an unstructured error vector with mean $E(\boldsymbol{\varepsilon}) \equiv \mathbf{0}$ and covariance as $\text{Cov}(\boldsymbol{\varepsilon}) \equiv \sigma_\varepsilon^2\mathbf{I}$ as in traditional linear regression models (*see* **Linear Regression**), this leaves the $n \times 1$ random vector $\mathbf{z}$ in Equation (1) to accommodate the AR structure in $\mathbf{y}$ and its presence results in a mixed model (i.e., containing both fixed and random effects). The difference between AR models and conventional mixed effects regression models is in how we specify the stochasticity for $\mathbf{z}$. In AR models, we typically rely on Gaussian assumptions for $\mathbf{z}$ and specify the mean as $E(\mathbf{z}) \equiv \mathbf{0}$ and covariance as $\text{Cov}(\mathbf{z}) \equiv \boldsymbol{\Sigma}_z$. While CAR and SAR models both assume this form of second-moment dependence, the covariance for the SAR model is generally specified as

$$\boldsymbol{\Sigma}_z \equiv \sigma_z^2(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Omega}(\mathbf{I} - \mathbf{B}')^{-1} \tag{2}$$

where the variance component $\sigma_z^2$ controls the variation in the correlated random effect $\mathbf{z}$, the $n \times n$ diagonal matrix $\boldsymbol{\Omega}$ has positive values as diagonal elements, and the $n \times n$ matrix $\mathbf{B}$ accounts for the second-moment dependence in the process[3]. Commonly, the matrix $\boldsymbol{\Omega}$ is set to be the identity matrix $\mathbf{I}$ which results in the simplified form of SAR covariance

$$\boldsymbol{\Sigma}_z \equiv \sigma_z^2((\mathbf{I} - \mathbf{B})(\mathbf{I} - \mathbf{B}'))^{-1} \tag{3}$$

The mixed model specification in Equation (1) includes a structured random effect ($\mathbf{z}$) and an unstructured error term ($\boldsymbol{\varepsilon}$). If we omit the structured random effect $\mathbf{z}$, the model in Equation (1) reduces to the conventional linear regression model with only fixed effects. If we omit the unstructured errors $\boldsymbol{\varepsilon}$, then we are left with a SAR model that contains only structured second-order dependence in addition to the first-moment effects. In mixed model specifications, normality is often assumed for $\mathbf{z}$ and $\boldsymbol{\varepsilon}$ because it facilitates implementation and inference.

## 2 Simultaneous Dependence

The dependence induced in the process by the covariance model in Equation (3) may not be immediately intuitive. However, some have found it helpful to consider a constructive approach for building a SAR covariance matrix. One approach for building a SAR model involves regressing the random effects in $\mathbf{z}$ on themselves simultaneously (hence the name). Using this constructive approach to formulating the SAR model, we have

$$\mathbf{z} = \mathbf{B}\mathbf{z} + \boldsymbol{\eta} \tag{4}$$

where $\boldsymbol{\eta}$ has mean $E(\boldsymbol{\eta}) \equiv \mathbf{0}$ and covariance $\text{Cov}(\boldsymbol{\eta}) \equiv \sigma^z\mathbf{I}$. Subtracting $\mathbf{B}\mathbf{z}$ from both sides of Equation (4) results in $\mathbf{z} - \mathbf{B}\mathbf{z} = \boldsymbol{\eta}$, which, after some algebra, results in $(\mathbf{I} - \mathbf{B})\mathbf{z} = \boldsymbol{\eta}$. This implies that $\mathbf{z} = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\eta}$



*Wiley StatsRef: Statistics Reference Online*, © 2014–2019 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08208

if the matrix $\mathbf{I} - \mathbf{B}$ is invertible (Refs 1, 4 which may not always be the case). Thus, because $\boldsymbol{\eta}$ has mean $E(\boldsymbol{\eta}) \equiv \mathbf{0}$ and covariance $\text{Cov}(\boldsymbol{\eta}) \equiv \sigma_z^2\mathbf{I}$, $\mathbf{z}$ also has mean $E(\mathbf{z}) = \mathbf{0}$, but has covariance $\text{Cov}(\mathbf{z}) = \sigma_z^2(\mathbf{I} - \mathbf{B})^{-1}((\mathbf{I} - \mathbf{B})^{-1})'$ which can be rewritten as the covariance in the SAR model Equation (3).

## 3 Parameterizing SAR Dependence

The SAR covariance matrix is often written generically as $\boldsymbol{\Sigma}_z \equiv \sigma_z^2((\mathbf{I} - \mathbf{B})(\mathbf{I} - \mathbf{B}'))^{-1}$ as in Equation (3). However, if fully parameterized, the SAR dependence matrix $\mathbf{B}$ contains $n^2$ elements and increases with the size of the data set (or cardinality of the discrete support). Thus, to reduce dimensionality of the inverse statistical problem, the dependence matrix is often parameterized as $\mathbf{B} = \rho\mathbf{W}$, where $\rho$ controls the amount of dependence and the $n \times n$ matrix $\mathbf{W}$ is defined *a priori* in a way that accounts for hypothesized relationships among areal units in the support. The matrix $\mathbf{W}$ is commonly referred to as a neighborhood or proximity matrix and is binary with zeros on the diagonal and ones on the off-diagonal elements that indicate a possible relationship among areal units. For example, in a scenario with 5 areal units consecutively indexed and symmetric relationships among only consecutive indices, $\mathbf{W}$ is the $5 \times 5$ matrix

$$\mathbf{W} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \tag{5}$$

Note that the proximity matrix in Equation (5) contains all zeros on the diagonal implying that there are no self relationships for individual areal units in this example.

Proximity matrices associated with SAR models need not be symmetric (although $\mathbf{W}$ happens to be symmetric in Equation (5)). This is in contrast with those used in CAR models and results from the form of covariance used in Equation (3). In the SAR covariance that is based on a proximity matrix, we have the term $(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}')$ and, while $(\mathbf{I} - \rho\mathbf{W})$ may not be symmetric, the product will be symmetric. This highlights another fundamental difference between SAR and CAR models. That is, in SAR models, we allow for dependence using a square root of the precision matrix associated with the model. By contrast, CAR models allow for dependence through the precision matrix (*see* **Precision Matrix**). Both types of areal data models differ from those used in geostatistics for continuous spatial processes that are based on direct formulations involving the correlation (rather than precision).

## 4 Forms of Proximity

Parameterizing the SAR dependence matrix as $\mathbf{B} = \rho\mathbf{W}$ reduces the number of unknowns from $n^2$ to 1 when $\mathbf{W}$ is considered fixed and known. Considering $\mathbf{W}$ to be known implies strong assumptions about the potential relationships among areal units. However, if there is no evidence for such dependence in the process based on the data, $\rho$ and/or $\sigma_z^2$ will be estimated near zero, attenuating the random effect $\mathbf{z}$. Also, the predictive ability of various specifications for $\mathbf{W}$ can be compared by scoring a set of competing models (e.g., Ref. 5).

We can specify the simplest type of proximity using $\mathbf{W}$ like that shown in Equation (5) where neighboring areal units are easily recognized such as in time series. However, in contiguous arrangements of areal units in higher dimensions, we may choose to parameterize $\mathbf{W}$ such that elements are equal to 1 if they



DOI: 10.1002/9781118445112.stat08208

share a boundary. In more complex arrangements of the discrete spatial support for the process, we may specify neighbors based on proximity calculated using an auxiliary data source. For example, if $d_{ij}$, for $i = 1, \ldots, n$ and $j = 1, \ldots, n$, represents a distance between areal units $i$ and $j$, then we could assign all other areal units within a certain distance threshold as neighbors. Alternatively, we could specify $\mathbf{W}$ such that the closest $q$ (for $q < n$) nearest areal units to the focal region are considered neighbors.

One benefit of using a binary proximity matrix in a SAR model is to take advantage of sparsity. For example, consider the proximity matrix $\mathbf{W}$ in Equation (5) which only has 8 non-zero elements out of a total of $n^2 = 25$. Thus, when fitting SAR models with sparse precision matrices to large data sets, we can use computer algorithms that are based on sparse matrix storage and calculations to reduce the computational time[6]. Another important aspect of both SAR and CAR models relative to geostatistical models is that we can reduce the primary computational burden to fit the model in Equation (1) to data. Because both SAR and CAR models are specified in terms of their precision, we do not need to invert their associated covariance matrix $\mathbf{\Sigma}_z$ when evaluating the Gaussian likelihood associated with Equation (1).

Binary proximity matrices may not adequately characterize the dependence in a given process. In that case, it may be useful to specify $\mathbf{W}$ such that it has continuous-valued elements. For example, in the previously discussed case where auxiliary information provides knowledge of pairwise distances $d_{ij}$, we could specify the elements of $\mathbf{W}$ as

$$w_{ij} = \exp\left(-\frac{d_{ij}}{\phi}\right) \tag{6}$$

where $\phi$ is a parameter controlling the range of dependence that can either be treated as fixed and known or estimated with $\rho$ and $\sigma_z^2$. This proximity matrix formulation is very similar to that used in geostatistical models for spatial processes. The type of process we are studying affects how we calculate the distances $d_{ij}$. If the areal units exist in a geographic space, $d_{ij}$ could be defined as the Euclidean distance between centroids for the areal units. Alternatively, in situations with contiguous areal units, the proximity $w_{ij}$ could be a function of the length of shared boundary between units.

The possibility of accommodating asymmetry in the SAR proximity matrix can be viewed as an advantage in formulating SAR models versus CAR models. For example, if the relationships among areal units can be viewed jointly as a graph[7], then a SAR model allows us to account for mechanisms giving rise to a directed graph (*see* **Multivariate Directed Graphs**). That is, suppose we wish to use a SAR model to accommodate dependence among species in a community. In this case, each areal unit in a SAR model represents a species and the proximity $w_{ij}$ characterizes dependence among species $i$ and $j$. If $w_{ij} = 1$ and $w_{ji} = 0$, then the model accommodates the situation where species $j$ influences species $i$, but not vice versa. Hanks[8] also demonstrated how natural dynamic mechanisms can lead to dependence structure in limiting processes that can be accommodated by SAR models.

## 5 Row Standardization

When the SAR dependence matrix is specified as $\mathbf{B} \equiv \rho\mathbf{W}$, the precision matrix $(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}')$ will only be proper when the dependence parameter $\rho$ is between the reciprocals of the smallest and largest eigenvalues (*see* **Eigenvalues and Eigenvectors in Statistics**) of $\mathbf{W}$. For our example SAR proximity matrix in Equation (5), the smallest and largest eigenvalues are $-1.732$ and $1.732$, respectively. Thus, for the associated SAR precision matrix to be valid, then $-0.577 < \rho < 0.577$ is sufficient for $(\mathbf{I} - \rho\mathbf{W})^{-1}$ to exist. Standardizing the rows of the proximity matrix $\mathbf{W}$ so that they sum to one has several benefits (e.g., Ref. 2). One such benefit is that, when the rows of the proximity matrix sum to one, it simplifies the condition for the dependence parameter such that $-1 < \rho < 1$ is sufficient for the SAR precision matrix to be valid.



*Wiley StatsRef: Statistics Reference Online*, © 2014–2019 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08208

However, when the rows of the proximity matrix sum to one, it has become common to set the dependence parameter $\rho = 1$ to reduce the overall number of parameters in the model because $\rho$ is often estimated to be near 1 regardless. SAR models for which $\rho = 1$ are referred to as intrinsic simultaneous autoregressive models (ISAR; or ICAR when used with CAR models; *see* **Conditional Autoregressive (CAR) Model**). In ISAR models, the covariance matrix does not exist, but the resulting form of the distribution can be used as an improper prior in a Bayesian analysis. Alternatively, additional constraints can be placed on $\mathbf{z}$ to ensure that the distribution is proper[9].

In our example proximity matrix from Equation (5), standardizing so that each row sums to one results in the new matrix

$$\widetilde{\mathbf{W}} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \tag{7}$$

The constructive perspective of the SAR model in Equation (4) sheds some light on the new interpretation of this row-standardized proximity matrix. The constructive SAR model in Equation (4) implies that each element of $\mathbf{z}$ is a weighted average (rather than a sum) of the other elements of $\mathbf{z}$ if we use the proximity matrix $\widetilde{\mathbf{W}}$ in Equation (7). When the process is averaged over more neighboring areal units, its variance will decrease implying that more central and well connected areal units will have lower variance than those on the margins of the graph that are less well connected. This concept matches our understanding of geostatistical (*see* **Geostatistics**) processes that have larger uncertainty when certain locations are farther from other measured locations in the process. Thus, row-standardization of the proximity matrix in SAR (and CAR) models is generally appropriate[2].

## 6 Equivalence of SAR and CAR Models

Ver Hoef *et al.*[10] showed that any SAR model can be written as a CAR model and vice versa. For comparison, a general CAR model takes the same form as the mixed model in Equation (1), but has the covariance matrix for $\mathbf{z}$ expressed as $\mathbf{\Sigma}_z \equiv \sigma_z^2(\mathbf{I} - \mathbf{C})^{-1}\mathbf{M}$, where $\mathbf{C}$ serves as the CAR dependence matrix and $\mathbf{M}$ is a diagonal matrix with elements $m_{ii}$ proportional to the conditional variance of $z_i$ given all its neighbors. CAR dependence matrices are often specified as $\mathbf{C} \equiv \rho\mathbf{W}$ like in SAR models. However, Wall[11] showed that correlation increases more rapidly with the dependence parameter $\rho$ in SAR models than in CAR models specified in this way. Intuitively, that makes sense because SAR models are parameterized such that the dependence is a function of square root of the precision matrix rather than the precision matrix directly as in CAR models.

## 7 Example: Avian Diversity

To demonstrate the use of a SAR model, we apply a reduced form of the model with second-order structure arising only from the spatially structured random effects to model avian diversity in Colorado, USA. We take a Bayesian (*see* **Bayes' Theorem and Updating of Belief**) approach to fitting the SAR model to the natural log of avian species richness in Colorado counties $\mathbf{y} \equiv (y_1, \ldots, y_n)'$ $(n = 64)$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{z} \tag{8}$$



DOI: 10.1002/9781118445112.stat08208

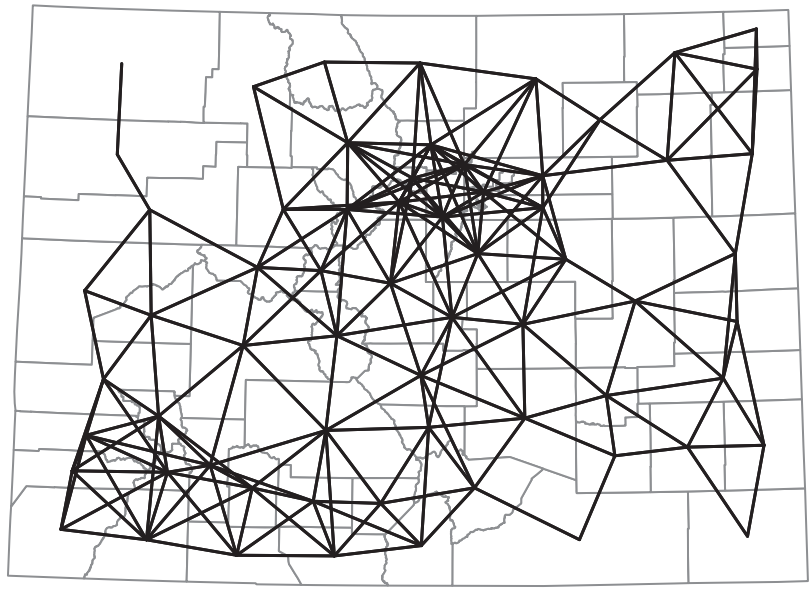

**Figure 1.** SAR network graph (graph edges shown in black) associated with our binary proximity matrix **W** based on a threshold set at 15% of the maximum distance between county (shown in gray) centroids in Colorado.

where $\mathbf{z} \sim \mathrm{N}(\mathbf{0}, \sigma_z^2((\mathbf{I} - \rho\widetilde{\mathbf{W}})(\mathbf{I} - \rho\widetilde{\mathbf{W}}'))^{-1})$. In our model, we used a row-standardized proximity matrix $\widetilde{\mathbf{W}}$ that was based on a binary proximity matrix $\mathbf{W}$ with elements equal to one only when the distances between Colorado county centroids do not exceed 15% of the maximum distance between any two counties. The threshold equal to 15% of the maximum distance ensured that every county in Colorado had at least one neighbor (Figure 1).

To complete our Bayesian SAR model, we specify priors for the regression coefficients $\boldsymbol{\beta}$, SAR variance $\sigma_z^2$, and SAR spatial dependence parameter $\rho$, such that

$$\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, 1000 \cdot \mathbf{I}) \tag{9}$$

$$\sigma_z^2 \sim \mathrm{IG}(0.001, 0.001) \tag{10}$$

$$\rho \sim \mathrm{Unif}(0, 1) \tag{11}$$

Note that the prior support for $\rho$ ($0 < \rho < 1$) results in a valid SAR covariance matrix but also restricts the model such that it can only account for positive spatial dependence.

We fit the Bayesian SAR model to the log species richness data in Figure 2a based on the spatially referenced covariates in Figure 2b–d using a Markov chain Monte Carlo (*see* **Bayesian Analysis and Markov Chain Monte Carlo Simulation**) algorithm and 100 000 iterations. Figure 3 shows marginal posterior histograms resulting from our model fit to the log species richness data. Based on our analysis, we can see that there is weak evidence for a negative effect of minimum elevation (Figure 3b), a strong effect of log human population density (Figure 3c), and some evidence for a positive effect of total county area on log avian species richness (Figure 3d). Furthermore, there was evidence for substantial learning about the SAR dependence parameter $\rho$ beyond our prior distribution (Figure 3f) which implies that the spatial random effect is needed in the model to account for dependence beyond what is explained by the first-moment effects.



*Wiley StatsRef: Statistics Reference Online*, © 2014–2019 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08208

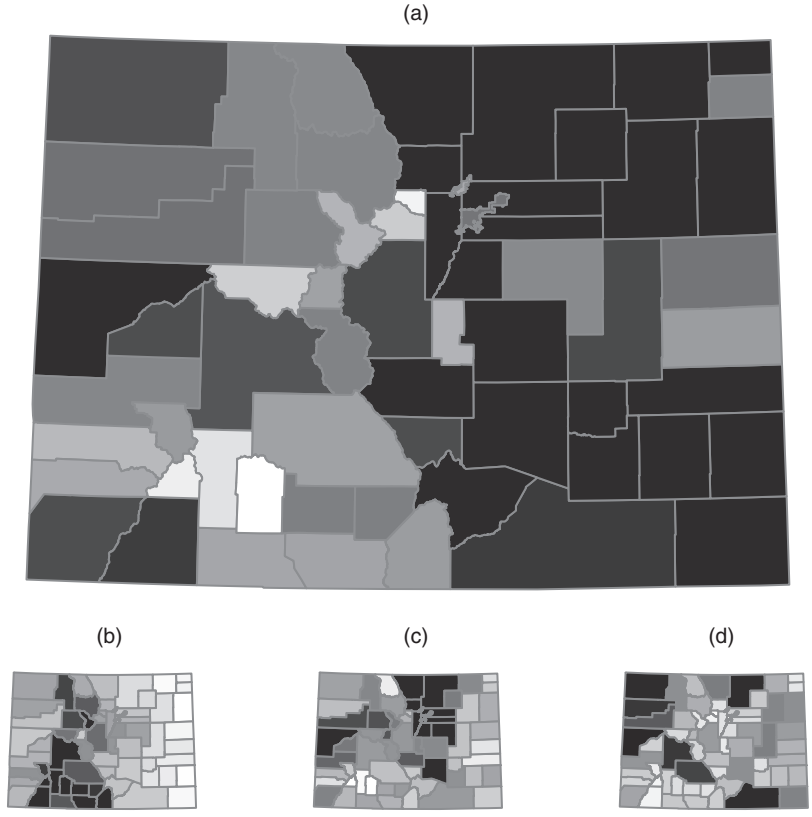


**Figure 2.** (a) Log species richness in Colorado counties, and standardized covariates: (b) minimum elevation in county, (c) log human population density per county, and (d) total area of county. Darker shades correspond to greater values in each map.

For comparison, we also fit a conventional linear regression model assuming independent errors (i.e., the model shown in Equation (8) with $\rho = 0$). We compared the predictive ability of the SAR model and conventional regression model using the deviance information criterion (DIC; smaller values indicate better predictive ability; (Refs 12, 13; *see* **Deviance Information Criterion (DIC)**). We calculated DIC for the conventional regression model as −64.6 and for the SAR model as −489.3 which indicated that the SAR model is the better predictive model for our data set and chosen covariates (because it has the smaller DIC value).

## 8 Extensions and Related Concepts

### 8.1 Confounding

In general, linear mixed models as in Equation (1) are useful for modeling data that have discrete support and a combination of first-moment and second-moment structure. Often such models are used



DOI: 10.1002/9781118445112.stat08208

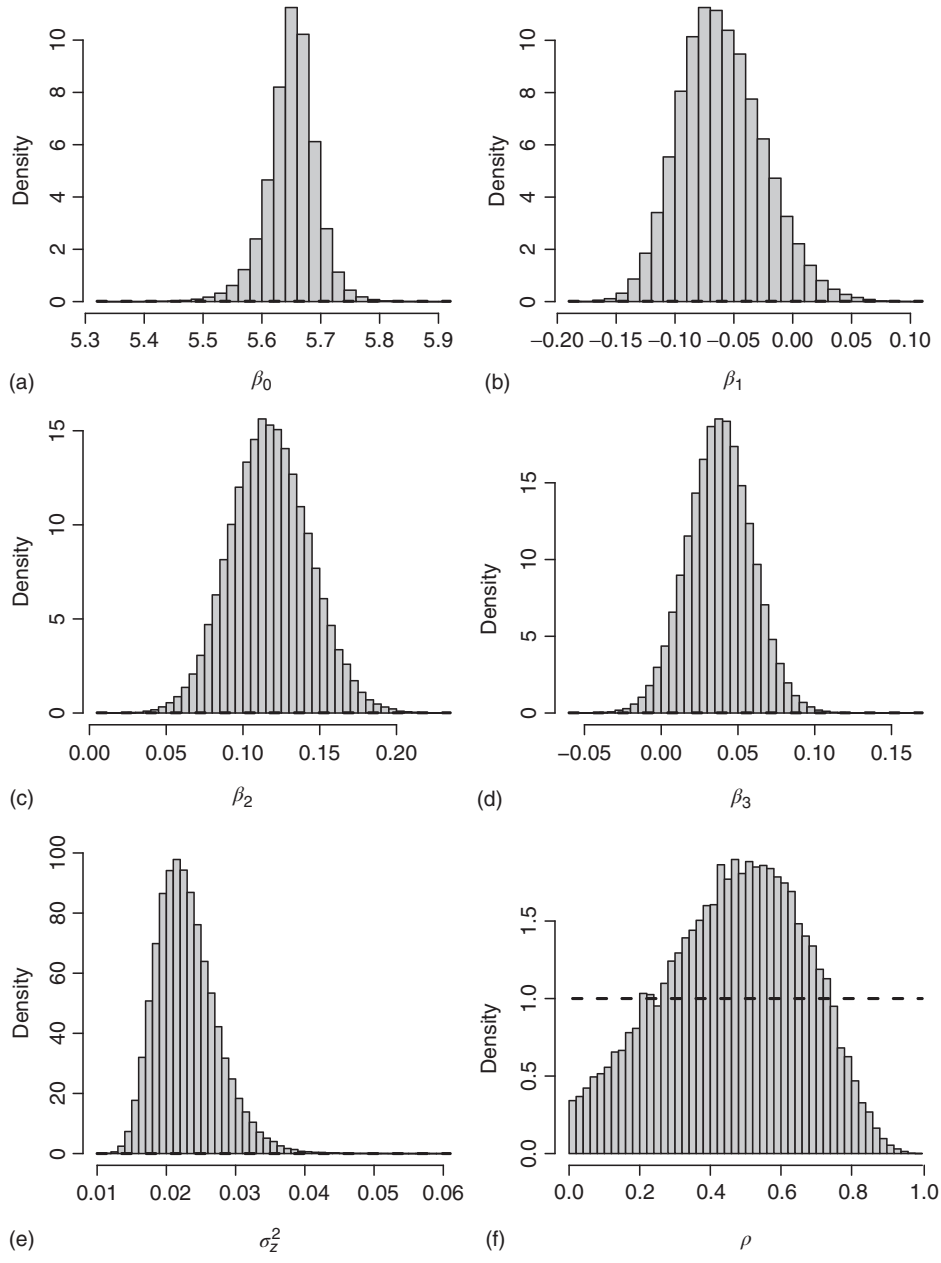


**Figure 3.** Marginal posterior histograms for (a) the intercept $\beta_0$, (b) the coefficient associated with minimum elevation ($\beta_1$), (c) the coefficient associated with log human population ($\beta_2$), (d) the coefficient associated with total area ($\beta_3$), (e) the SAR variance ($\sigma_z^2$), and (f) the SAR dependence parameter ($\rho$). Priors shown as dashed lines.



*Wiley StatsRef: Statistics Reference Online*, © 2014–2019 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08208

when a subset of predictor variables are measured and appear in the design matrix $\mathbf{X}$ yet there remains dependence that goes unaccounted for when considering only the first-moment fixed effects. An analysis of the residuals from conventional linear regression to assess additional dependence (e.g., using a Moran's I statistic[14]; (**Spatial Autocorrelation Coefficient, Moran's**) may indicate that we need to account for it in the model so that the assumptions are valid. However, in such cases, there may exist confounding among the first- and second-moment effects[15, 16]. To see this, consider the alternative form for the random effects, written as a linear combination of basis vectors $\mathbf{z} = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\eta}$. Using this formulation, the linear mixed model in Equation (1) is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\eta} + \boldsymbol{\epsilon} \tag{12}$$

and it becomes clear that both $\mathbf{X}$ and $(\mathbf{I} - \mathbf{B})^{-1}$ act as covariates in the model. If the columns of $\mathbf{X}$ and $(\mathbf{I} - \mathbf{B})^{-1}$ are correlated, our inference about the regression coefficients $\boldsymbol{\beta}$ may differ in the models with and without random effects. While remedial approaches such as restricted regression (e.g., Refs 16, 17) exist for constraining such models so that estimates for $\boldsymbol{\beta}$ are similar in models with and without random effects, determining when to use restricted regression is not always clear[18].

## 8.2 GLMMs

The linear mixed modeling approach we presented in Equation (1) may be appropriate for continuous-valued response data. However, in other fields such as ecology and epidemiology, binary or count response variables are common. Fortunately, SAR models can easily be applied in a generalized linear mixed model (GLMM; *see* **Generalized Linear Mixed Models**) framework. For example, for binary observations $\mathbf{y} \equiv (y_1, \ldots, y_n)'$ and associated success probabilities $\mathbf{p} \equiv (p_1, \ldots, p_n)'$, we could specify a GLMM with SAR random effects as $y_i \sim \text{Bern}(p_i)$ where $g(\mathbf{p}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{z}$. In this case, the autocorrelated random effects $\mathbf{z}$ are modeled as in Equation (3) and $g$ represents a link function of choice (e.g., logit or probit). In fact, for the Bayesian version of this type of GLMM, a latent variable representation can be used to facilitate computation when fitting the model to data (e.g., Refs 19, 20). Similar GLMM specifications based on other types of data (e.g., counts) follow naturally.

## 8.3 Network Models

Statistical models for networks (i.e., graphs) can assume a SAR form where the covariance is expressed as in Equation (3) with $\mathbf{B} = \rho\mathbf{W}$ and the proximity matrix $\mathbf{W}$ is not known *a priori*. The goal in such analyses is to infer the network structure. Thus, we seek to estimate the elements of $\mathbf{W}$, which we can do by assuming additional structure about the edges of the graph. One approach to parameterizing the edge weights $w_{ij}$ is to link them to an underlying latent point process (e.g., Ref. 21; *see* **Exponential Random Graph Models**). For example, we could define $w_{ij} = \exp(-||\mathbf{s}_i - \mathbf{s}_j||)$ where $\mathbf{s}_i$, for $i = 1, \ldots, n$, are points in a potentially multidimensional latent space $\mathcal{S}$. The simplest probability model for the points $\mathbf{s}_i$ is a standard multivariate normal distribution. If the latent space is two-dimensional, there will be $2n + 2$ unknowns to be estimated in the SAR covariance matrix $\boldsymbol{\Sigma}_z$ (i.e., all $\mathbf{s}_i$ for $i = 1, \ldots, n$, $\rho$, and $\sigma_z^2$) which is $2n$ more than the model where the proximity matrix $\mathbf{W}$ is assumed to be known, but much less than the $n^2$ that would be unknown if we allowed the proximity matrix $\mathbf{W}$ to be fully parameterized.



DOI: 10.1002/9781118445112.stat08208

## Acknowledgements

The authors thank Erin Peterson, Marie-Josee Fortin, and Noel Cressie for numerous helpful discussions and previous research in this area. This research was funded by NSF DMS 1614392. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

## Related Articles

**Conditional Autoregressive (CAR) Model**; **Geostatistics**.

## References

[1] Cressie, N.A.C. (1993) *Statistics for Spatial Data*, Revised edn, John Wiley and Sons, Inc., New York.
[2] Ver Hoef, J.M., Peterson, E.E., Hooten, M.B., et al. (2018) Spatial autoregressive models for statistical inference from ecological data. *Ecol. Monogr.*, **88**, 36–59.
[3] Whittle, P. (1954) On stationary processes in the plane. *Biometrika*, **41**, 434–449.
[4] Waller, L.A. and Gotway, C.A. (2004) *Applied Spatial Statistics for Public Health Data*, John Wiley & Sons, Somerset, NJ.
[5] Cassemiro, F.A., Diniz-Filho, J.A.F., Rangel, T.F.L., and Bini, L.M. (2007) Spatial autocorrelation, model selection and hypothesis testing in geographical ecology: Implications for testing metabolic theory in New World amphibians. *Neotrop. Biol. Conserv.*, **2**, 119–126.
[6] Rue, H. (2001) Fast sampling of Gaussian Markov random fields. *J. Roy. Stat. Soc.Ser. B*, **63**, 325–338.
[7] Lauritzen, S.L. (1996) *Graphical Models*, Oxford University Press, Oxford, UK.
[8] Hanks, E.M. (2017) Modeling spatial covariance using the limiting distribution of spatio-temporal random walks. *J. Am. Stat. Assoc.*, **112**, 497–507.
[9] Ver Hoef, J.M., Hanks, E.M., and Hooten, M.B. (2018a) On the relationship between conditional (CAR) and simultaneous (SAR) autoregressive models. *Spat. Stat.*, **25**, 68–85.
[10] Rue, H. and Held, L. (2005) *Gaussian Markov Random Fields: Theory and Applications*, Chapman and Hall/CRC, Boca Raton, FL.
[11] Wall, M.M. (2004) A close look at the spatial structure implied by the CAR and SAR models. *J. Stat. Plan. Inference*, **121**, 311–324.
[12] Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. (2002) Bayesian measures of model complexity and fit. *J. Roy. Stat. Soc. B*, **64**, 583–639.
[13] Hooten, M.B. and Hobbs, N.T. (2015) A guide to Bayesian model selection for ecologists. *Ecol. Monogr.*, **85**, 3–28.
[14] Moran, P.A. (1948) The interpretation of statistical maps. *J. Roy. Stat. Soc. Ser. B*, **10**, 243–251.
[15] Reich, B.J., Hodges, J.S., and Zadnik, V. (2006) Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, **62**, 1197–1206.
[16] Hodges, J.S. and Reich, B.J. (2010) Adding spatially-correlated errors can mess up the fixed effect you love. *Am. Stat.*, **64**, 325–334.
[17] Hefley, T.J., Hooten, M.B., Hanks, E.M., et al. (2017) The Bayesian group lasso for confounded spatial data. *J. Agricult. Biol. Environ. Stat.*, **22**, 42–59.
[18] Hanks, E.M., Schliep, E., Hooten, M.B., and Hoeting, J.A. (2015) Restricted spatial regression in practice: Geostatistical models, confounding, and robustness under model misspecification. *Environmetrics*, **26**, 243–254.
[19] Hooten, M.B., Larsen, D.R., and Wikle, C.K. (2003) Predicting the spatial distribution of ground flora on large domains using a hierarchical Bayesian model. *Landsc. Ecol.*, **18**, 487–502.
[20] Johnson, D.S., Conn, P.B., Hooten, M.B., et al. (2013) Spatial occupancy models for large data sets. *Ecology*, **94**, 801–808.
[21] Hoff, P.D., Raftery, A.E., and Handcock, M.S. (2002) Latent space approaches to social network analysis. *J. Am. Stat.Assoc.* **97**, 1090–1098.




DOI: 10.1002/9781118445112.stat08208