# Appendix S1

## Spatial Autoregressive Models for Statistical Inference from Ecological Data

Jay M. Ver Hoef[1], Erin E. Peterson[2], Mevin B. Hooten[3], Ephraim M. Hanks[4], and
Marie-Josée Fortin[5]

---

[1]Marine Mammal Laboratory, NOAA-NMFS Alaska Fisheries Science Center
7600 Sand Point Way NE, Seattle, WA 98115
tel: (907) 456-1995      E-mail: jay.verhoef@noaa.gov
[2] ARC Centre for Excellence in Mathematical and Statistical Frontiers (ACEMS) and
the Institute for Future Environments, Queensland University of Technology
[3] U.S. Geological Survey, Colorado Cooperative Fish and Wildlife Research Unit,
Department of Fish, Wildlife, and Conservation Biology
Department of Statistics, Colorado State University
[4] Department of Statistics, The Pennsylvania State University
[5] Department of Ecology and Evolutionary Biology, University of Toronto

---

October 24, 2017

# Misconceptions and Errors in the Literature

The fact that CAR and SAR models are developed from the precision matrix, in contrast to geostatistical models being developed for the covariance matrix, has caused some confusion in the ecological literature. For example, in comparing geostatistical models to SAR models, Beguería and Pueyo (2009) stated "Semivariogram models account for spatial autocorrelation *at all possible distance lags*, and thus they do not require *a priori* specification of the window size and the covariance structure," (emphasis by the original authors). CAR and SAR models also account for spatial autocorrelation at all possible lags, as seen in Fig. 9c,d, of the main article. In a temporal analogy, the autoregressive AR1 time series models also account for autocorrelation at all possible lags, where the conditional specification $Z_{i+1} = \phi Z_i + \nu_i$, with $\nu_i$ an independent random shock and $|\phi| < 1$, implies that $\text{corr}(Z_i, Z_{i+t}) = \phi^t$ for all $t$. In fact, if we restrict $0 < \phi < 1$, then this can be reparameterized as $\text{corr}(Z_i, Z_{i+t}) = \exp(-t(-\log(\phi)))$, which is an exponential geostatistical model with range parameter $-\log(\phi)$. While there are interesting results in Beguería and Pueyo (2009), a restriction on the range of autocorrelation is not a reason that CAR/SAR models might perform poorly against a geostatistical model. The important concept is that the autoregressive specification is local in the precision matrix and not in the covariance matrix.

CAR models are often incorrectly characterized. For example, Keitt et al. (2002) characterized CAR models as: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{C}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\varepsilon}$, with a stated covariance matrix of $\sigma^2(\mathbf{I} - \rho\mathbf{C})^{-1}$, where $\mathbf{C}$ is symmetric. Their actual implementation may have been correct, and there are excellent and important results in Keitt et al. (2002); however, the construction they used leads to a SAR covariance matrix of $\sigma^2(\mathbf{I} - \rho\mathbf{C})^{-1}(\mathbf{I} - \rho\mathbf{C})^{-1}$ if $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$ and $\mathbf{C}$ is symmetric. Even to characterize a CAR model as $\sigma^2(\mathbf{I} - \rho\mathbf{C})^{-1}$ with symmetric $\mathbf{C}$ is overly restrictive, as we have demonstrated that an asymmetric $\mathbf{C}$ with the proper $\mathbf{M}$ will still satisfy Eq. 12, or alternatively that $\boldsymbol{\Sigma}^{-1} = (\mathbf{M}^{-1} - \mathbf{C})/\sigma^2$, where $\mathbf{C}$ is symmetric but $\mathbf{M}^{-1}$ is not necessarily constant on the diagonals. In fact, constraining a CAR model to $\sigma^2(\mathbf{I} - \rho\mathbf{C})^{-1}$ does not allow for row-standardized models. These mistakes are perpetuated in Dormann et al. (2007), and we have seen similar errors in describing CAR models as SAR models in other literature, presentations, and help sites on the internet.

Dormann et al. (2007) also claimed that any SAR model is a CAR model, which agrees with the literature (e.g., Cressie, 1993, p. 409), but then they show an incorrect proof (it is also incorrect in Haining (1990, p. 89), and likely beginning there), because they do not consider that $\mathbf{C}$ for a CAR model must have zeros along the diagonal. In fact, we demonstrate in Ver Hoef et al. (2017) that, despite statistical and ecological literature to the contrary, CAR models and SAR models can be written equivalently, and we provide details.

# Maximum Likelihood Estimation for CAR/SAR Models with Missing Data

We begin by finding analytical solutions when we can, and then substituting them into the likelihood to reduce the number of parameters as much as possible for the full covariance matrix. Assume a linear model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{y}$ is a vector of response variables, $\mathbf{X}$ is a design matrix of full rank, $\boldsymbol{\beta}$ is a vector of parameters, and the zero-mean random errors have a multivariate normal distribution, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a patterned covariance matrix; i.e., it has non-zero off-diagonal elements. Suppose that $\boldsymbol{\Sigma}$ has parameters $\{\theta, \boldsymbol{\rho}\}$ and can be written as $\boldsymbol{\Sigma} = \theta \mathbf{V}_{\boldsymbol{\rho}}$, where $\theta$ is an overall variance parameter and $\boldsymbol{\rho}$ are parameters that structure $\mathbf{V}_{\boldsymbol{\rho}}$ as a non-diagonal matrix, and we show the dependency as a subscript. Note that $\boldsymbol{\Sigma}^{-1} = \mathbf{V}_{\boldsymbol{\rho}}^{-1}/\theta$. Recall that the maximum likelihood estimate of $\boldsymbol{\beta}$ for any $\{\theta, \boldsymbol{\rho}\}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}_{\boldsymbol{\rho}}^{-1}\mathbf{X})^{-1}\mathbf{X}\mathbf{V}_{\boldsymbol{\rho}}^{-1}\mathbf{y}$. By substituting $\hat{\boldsymbol{\beta}}$ into the normal likelihood equations, $-2$ times the log-likelihood for a normal distribution is

$$\mathcal{L}(\theta, \boldsymbol{\rho}|\mathbf{y}) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + \log(|\boldsymbol{\Sigma}|) + n\log(2\pi),$$

where $n$ is the length of $\mathbf{y}$, but this can be written as,

$$\mathcal{L}(\theta, \boldsymbol{\rho}|\mathbf{y}) = \mathbf{r}_{\boldsymbol{\rho}}'\mathbf{V}_{\boldsymbol{\rho}}^{-1}\mathbf{r}_{\boldsymbol{\rho}}/\theta + n\log(\theta) + \log(|\mathbf{V}|) + n\log(2\pi) \qquad (S1)$$

where $\mathbf{r}_{\boldsymbol{\rho}} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ (notice that $\hat{\boldsymbol{\beta}}$ is a function of $\boldsymbol{\rho}$, so we show that dependency for $\mathbf{r}$ as well). Conditioning on $\boldsymbol{\rho}$ yields

$$\mathcal{L}(\theta|\boldsymbol{\rho}, \mathbf{y}) = \mathbf{r}_{\boldsymbol{\rho}}'\mathbf{V}_{\boldsymbol{\rho}}^{-1}\mathbf{r}_{\boldsymbol{\rho}}/\theta + n\log(\theta) + \text{terms not containing } \theta$$

and minimizing (analytically) for $\theta$ involves setting

$$\frac{\partial \mathcal{L}(\theta|\boldsymbol{\rho}, \mathbf{y})}{\partial \theta} = -\mathbf{r}_{\boldsymbol{\rho}}'\mathbf{V}_{\boldsymbol{\rho}}^{-1}\mathbf{r}_{\boldsymbol{\rho}}/\theta^2 + n/\theta$$

equal to zero, yielding the maximum likelihood estimate

$$\hat{\theta} = \mathbf{r}_{\boldsymbol{\rho}}'\mathbf{V}_{\boldsymbol{\rho}}^{-1}\mathbf{r}_{\boldsymbol{\rho}}/n. \qquad (S2)$$

Substituting Eq. S2 back into Eq. S1 yields the -2*log-likelihood as a function of $\boldsymbol{\rho}$ only,

$$\mathcal{L}(\boldsymbol{\rho}|\mathbf{y}) = n\log(\mathbf{r}_{\boldsymbol{\rho}}'\mathbf{V}_{\boldsymbol{\rho}}^{-1}\mathbf{r}_{\boldsymbol{\rho}}) + \log(|\mathbf{V}_{\boldsymbol{\rho}}|) + n(\log(2\pi) + 1 - \log(n)). \qquad (S3)$$

Equation Eq. S3 can be minimized numerically to yield the MLE $\hat{\rho}$, and then $\hat{\theta} = \mathbf{r}'_{\hat{\rho}} \mathbf{V}^{-1}_{\hat{\rho}} \mathbf{r}_{\hat{\rho}}/n$, and $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}_{\hat{\rho}}\mathbf{X})^{-1}\mathbf{X}\mathbf{V}^{-1}_{\hat{\rho}}\mathbf{y}$ yield analytical solutions for MLEs after obtaining (numerically) the MLE for $\boldsymbol{\rho}$.

We developed the inverse covariance matrix $\boldsymbol{\Sigma}^{-1}_A = \text{diag}(\mathbf{W1}) - \rho\mathbf{W}$, and here we use $\boldsymbol{\Sigma}_A$ to denote it is for *all* locations, those with observed data as well as those without. Without missing data, Eq. S3 can be evaluated quickly by factoring out an overall variance parameter from $\boldsymbol{\Sigma}^{-1}_A$ and using sparse matrix methods to quickly and efficiently evaluate $|\mathbf{V}_{\boldsymbol{\rho}}|$ by recalling that $|\mathbf{V}_{\boldsymbol{\rho}}|$ $= 1/|\mathbf{V}^{-1}_{\boldsymbol{\rho}}|$. However, when there are missing data, there is no guarantee that $\mathbf{V}_{\boldsymbol{\rho}}$ will be sparse. The obvious and direct approach is to first obtain $\boldsymbol{\Sigma}_A = (\boldsymbol{\Sigma}^{-1}_A)^{-1}$, and then obtain $\mathbf{V}_{\boldsymbol{\rho}} = \boldsymbol{\Sigma}[\mathbf{i}, \mathbf{i}]$, where $\mathbf{i}$ is a vector of indicators that subsets the rows and columns of $\boldsymbol{\Sigma}$ to only those for sampled locations. Then, a third step is a second inverse to find $\mathbf{V}^{-1}_{\boldsymbol{\rho}}$. However, this is computationally expensive. A faster way uses results from partitioned matrices and Schur complements. In general, let the square matrix $\boldsymbol{\Sigma}$ with dimensions $(m + n) \times (n + m)$ be partitioned into block submatrices,

$$
\underset{(m+n)\times(m+n)}{\boldsymbol{\Sigma}} = \left[ \begin{array}{cc} \underset{m\times m}{\mathbf{A}} & \underset{m\times n}{\mathbf{B}} \\ \underset{n\times m}{\mathbf{C}} & \underset{n\times n}{\mathbf{D}} \end{array} \right]
$$

with dimensions given below each matrix. Assume $\mathbf{A}$ and $\mathbf{D}$ are nonsingular. Then define the matrix function $\mathbf{S}(\boldsymbol{\Sigma}, \mathbf{A}) = \mathbf{D} - \mathbf{CA}^{-1}\mathbf{B}$ as the Schur complement of $\boldsymbol{\Sigma}$ with respect to $\mathbf{A}$. Likewise, there is a Schur complement with respect to $\mathbf{D}$ by reversing the roles of $\mathbf{A}$ and $\mathbf{D}$. Using Schur complements, it is well-known (e.g., Harville, 1997, p. 97) that an inverse for a partitioned matrix $\boldsymbol{\Sigma}$ is,

$$
\boldsymbol{\Sigma}^{-1} = \left[ \begin{array}{cc} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{BS}(\boldsymbol{\Sigma}, \mathbf{A})^{-1}\mathbf{CA}^{-1} & -\mathbf{A}^{-1}\mathbf{BS}(\boldsymbol{\Sigma}, \mathbf{A})^{-1} \\ -\mathbf{S}(\boldsymbol{\Sigma}, \mathbf{A})^{-1}\mathbf{CA}^{-1} & \mathbf{S}(\boldsymbol{\Sigma}, \mathbf{A})^{-1} \end{array} \right]
$$

Then, note that $\mathbf{A}^{-1} = \mathbf{S}(\boldsymbol{\Sigma}^{-1}, \mathbf{S}(\boldsymbol{\Sigma}, \mathbf{A})^{-1})$; that is, if we already have $\boldsymbol{\Sigma}^{-1}$, then $\mathbf{A}^{-1}$ is the Schur complement of $\boldsymbol{\Sigma}^{-1}$ with respect to the rows and columns that correspond to $\mathbf{D}$. Additionally, the largest matrix that we have to invert is $[\mathbf{S}(\boldsymbol{\Sigma}, \mathbf{A})^{-1}]^{-1}$, which is $n \times$n and has dimension less than $\boldsymbol{\Sigma}$, and only one inverse is required. Thus, if we let $\mathbf{A}$ correspond to the $m$ rows and columns of the observed locations, and $\mathbf{D}$ correspond to the $n$ rows and columns of the missing data, then this provides a quick and efficient way to obtain $\mathbf{V}^{-1}_{\boldsymbol{\rho}}$ from $\boldsymbol{\Sigma}^{-1}_A$, and the largest inverse required is of dimension $n \times n$, where $n$ is the number of sites with missing data.

# Prediction and Smoothing

In what follows, we provide the formulas used in creating Fig. 8. For universal kriging, the formulas can be found in Cressie and Wikle (2011, p. 148),

$$\hat{y}_i = \mathbf{x}'_i\hat{\boldsymbol{\beta}} + \mathbf{c}_i\boldsymbol{\Sigma}_{-i}^{-1}(\mathbf{y}_{-i} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

where $\hat{y}_i$ is the prediction for the $i$th node, $\mathbf{x}_i$ is a vector containing the covariate values for the $i$th node, $\mathbf{X}$ is the design matrix for the covariates (fixed effects), $\mathbf{c}_i$ is a vector containing the fitted covariance between the $i$th site and all *other* sites with observed data, $\boldsymbol{\Sigma}$ is the fitted covariance matrix among all observed data, $\mathbf{y}$ is a vector of observed values for the response variable, and $\hat{\boldsymbol{\beta}} = \mathbf{X}'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y}$ is the generalized least squares estimate of $\boldsymbol{\beta}$. The covariance values contained in $\mathbf{c}_i$ and $\boldsymbol{\Sigma}$ were obtained using maximum likelihood estimate for the parameters as detailed in Appendix B. We include the $-i$ subscript on $\boldsymbol{\Sigma}_{-i}^{-1}$ and $\mathbf{y}_{-i}$ to indicate that, when smoothing, we predict at the $i$th node by removing that datum from $\mathbf{y}$, and by removing its corresponding rows and columns in $\boldsymbol{\Sigma}$. If the value is missing, then prediction proceeds using all observed values. Hence, Fig. 8a contains the observed values plus the predicted values at nodes with missing values, while Fig. 8c contains predicted values at all nodes, where any observed value at a node was removed and predicted with the rest of the observed values. The prediction standard errors are given by,

$$\hat{\text{se}}(\hat{y}_i) = \sqrt{\mathbf{c}'_i\boldsymbol{\Sigma}_{-i}^{-1}\mathbf{c}_i + \mathbf{d}'_i(\mathbf{X}'_{-i}\boldsymbol{\Sigma}_{-i}^{-1}\mathbf{X}_{-i})^{-1}\mathbf{d}_i}$$

where $\mathbf{d}_i = \mathbf{x}'_i - \mathbf{X}'_{-i}\boldsymbol{\Sigma}_{-i}^{-1}\mathbf{c}_i$.

For the IAR model smoothing in Fig. 8d, we used the WinBUGS (Lunn et al., 2000) software, final version 1.4.3. The model code is very compact and given below:

```
model
  {
    for(i in 1:N) {
      trend[i] ~ dnorm(mu[i],prec)
      mu[i] <- beta0 + beta[stockid[i]] + b[i]
      }
    b[1:N] ~ car.normal(adj[], weights[], num[], tau)
    beta0 ~ dnorm(0,.001)
    beta[1] ~ dnorm(0,.001)
    beta[2] ~ dnorm(0,.001)
    beta[3] ~ dnorm(0,.001)
    beta[4] <- 0
    beta[5] ~ dnorm(0,.001)
    prec ~ dgamma(0.001,0.001)
    sigmaEps <- sqrt(1/prec)
    tau    ~ dgamma(0.5, 0.0005)
    sigmaZ <- sqrt(1 / tau)
  }
```

The means of the MCMC samples from the posterior distributions of `mu[i]` were used for the IAR smoothing for the $i$th location in Fig. 8d.

4

# References

Beguería, S. and Pueyo, Y. (2009), "A comparison of simultaneous autoregressive and generalized least squares models for dealing with spatial autocorrelation," *Global Ecology and Biogeography*, 18, 273–279.

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

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

Dormann, C. F., McPherson, J. M., Araújo, M. B., Bivand, R., Bolliger, J., Carl, G., Davies, R. G., Hirzel, A., Jetz, W., Kissling, W. D., Kühn, I., Ohlemüller, R., Peres-Neto, P. R., Reineking, B., Schröder, B., Schurr, F. M., and Wilson, R. (2007), "Methods to account for spatial autocorrelation in the analysis of species distributional data: a review," *Ecography*, 30, 609–628.

Haining, R. (1990), *Spatial Data Analysis in the Social and Environmental Sciences*, Cambridge, UK: Cambridge University Press.

Harville, D. A. (1997), *Matrix Algebra from a Statistician's Perspective*, New York, NY: Springer.

Keitt, T. H., Bjørnstad, O. N., Dixon, P. M., and Citron-Pousty, S. (2002), "Accounting for spatial pattern when modeling organism-environment interactions," *Ecography*, 25, 616–625.

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

Ver Hoef, J. M., Hanks, E. M., and Hooten, M. B. (2017), "On the relationship between conditional (CAR) and simultaneous (SAR) autoregressive models," *arXiv*, 1710.07000v1, 1–18.