8.8.1 Seal trend data

Models for the harbor seal data are based on neighborhood relationships (Figure 1.3) which leads to linear models based on spatial weights (Chapter 7). In general, these types of models are amenable to fast computing because they are specified through the inverse of the covariance matrix, e.g., $\boldsymbol{\Sigma}^{-1} = \mathbf{K}_{\text{CAR}}^{-1}(\mathbf{I} - \rho_{\text{CAR}}\mathbf{W})$ (eq. 7.7 using the parsimonious models in Section 7.2). It is the inverse of the covariance matrix that is used in the likelihoods (e.g, eqns 8.1 and 8.2), and this can be computationally demanding for large data sets if the models are specified through the covariance matrix, such as the geostatistical models (Chapter 6). Both geostatistical and spatial-weights models also require $|\boldsymbol{\Sigma}^{-1}|$ in the likelihood, and there are special algorithms for sparse matrices such as those for the spatial-weights models. Hence, the spatial-weights models, merely requiring an inverse of a diagonal matrix and a determinant of a sparse matrix, would seem to have a computational advantage over the geostatistical models.

However, one interesting aspect of the harbor seal example is that it contains missing data. Ultimately, we will want to predict at these missing locations, so, for the spatial-weights models, we must include the missing locations in the neighborhood structure. Recall that for the likelhood evaluations and optimization, we need the inverse of the covariance matrix for *only* the observed data, resulting in a situation where the computational advantages of the spatial-weights models can vanish. To make this clear, let us order the locations so that all of the observed locations are first, followed by the unobserved locations, and then the covariance matrix and inverse of the covariance matrix are

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{oo} & \boldsymbol{\Sigma}_{ou} \\ \boldsymbol{\Sigma}_{uo} & \boldsymbol{\Sigma}_{uu} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}^{oo} & \boldsymbol{\Sigma}^{ou} \\ \boldsymbol{\Sigma}^{uo} & \boldsymbol{\Sigma}^{uu} \end{bmatrix}, \tag{R.1}$$

where the subscripts and superscripts o and u are for the observed and unobserved locations, respectively. The most straight-forward idea for the spatial-weights models is to obtain the covariance matrix $\boldsymbol{\Sigma}_{oo}$ by first taking $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}^{-1})^{-1}$, and then computing $\boldsymbol{\Sigma}_{oo}^{-1}$ because, unfortunately, $\boldsymbol{\Sigma}_{oo}^{-1} \neq \boldsymbol{\Sigma}^{oo}$, and, also $\boldsymbol{\Sigma}_{oo}^{-1} \neq (\boldsymbol{\Sigma}^{oo})^{-1}$. However, there is a faster way than taking the two inverses $(\boldsymbol{\Sigma}^{-1})^{-1}$ and $\boldsymbol{\Sigma}_{oo}^{-1}$, by recalling from Section 5.6 that if we already have $\boldsymbol{\Sigma}^{-1}$, then we can obtain

$$\boldsymbol{\Sigma}_{oo}^{-1} = \boldsymbol{\Sigma}^{oo} - \boldsymbol{\Sigma}^{ou}(\boldsymbol{\Sigma}^{uu})^{-1}\boldsymbol{\Sigma}^{uo},$$

which requires a single numeric inverse $(\boldsymbol{\Sigma}^{uu})^{-1}$ that has dimensions less than those of the full $(\boldsymbol{\Sigma}^{-1})^{-1}$, and will be very beneficial if the dimensions of $\boldsymbol{\Sigma}_{uu}$ are (much) less than $\boldsymbol{\Sigma}_{oo}$. If the dimensions of $\boldsymbol{\Sigma}_{uu}$ are (much) larger than $\boldsymbol{\Sigma}_{oo}$, then the spatial-weights models, in terms of the computational demand due to matrix inverses, are more costly than geostatistical models. For the harbor seal example, there were 306 observed locations and 157 missing locations, so we only needed a 157×157 inverse.

We want to fit models by maximizing the log-likelihood for a variety of spatial-weights models. Figure 1.3 shows first, second, and fourth-order neighborhood relationships. If $\mathbf{W}_1$

is a symmetric spatial-weights matrix as described in Chapter 7, then a matrix that includes "neighbors of neighbors" is

$$\mathbf{W}_2 = \mathcal{I}(\mathcal{I}(\mathbf{W}_1\mathbf{W}_1 > 0) + \mathbf{W}_1 > 0) - \mathbf{I}$$

where $\mathcal{I}(\cdot)$ is the indicator function, equal to 1 if its argument is true, otherwise it is zero, and $\mathbf{I}$ is the identity matrix to ensure that the diagonal of $\mathbf{W}_2$ is all zeros. In a similar fashion, we can create a fourth-order neighborhood matrix, $\mathbf{W}_4$ from $\mathbf{W}_2$. Also, recall that for these data we have a single explanatory variable, which is a categorical (factor) variable for one of five different genetic stocks (Figure 1.1).

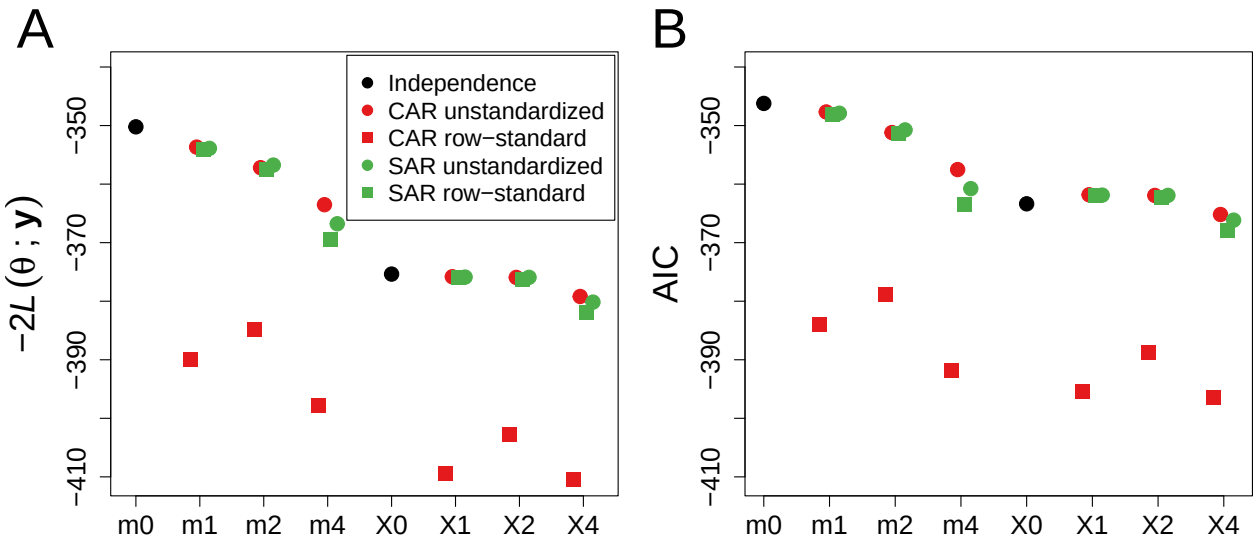

Figure 1: Log-likelihood and AIC for a variety of spatial-weights models. A. Minus two times the log-likelihood at the ML estimate. B. AIC for the same models as given to the left.

For the harbor seal example, we want to consider a variety of models that include whether or not a spatially-autocorrelated error structure is even necessary, and if so, whether it should be a CAR or SAR model. In both cases, we also can consider the first, second, and fourth-order neighborhood models as described above, and whether or not row-standardization was applied to those binary matrices. Figure 1 shows negative twice the log-likelihood, $-2L(\boldsymbol{\theta};\mathbf{y})$ (eq. 8.1), and AIC for all of these models. Generally, we see that models including the explanatory variable fit better than those without it. Interestingly, for the models with stock as an explanatory variable, the fits for CAR and SAR unstandardized, and SAR standardized, for first and second order neighbors were almost identical to the independence model, and so AIC favors the independence model because the spatial models have one more parameter (here, the spatial covariance can be thought of as $\boldsymbol{\Sigma} = \sigma^2\mathbf{R}_\rho$, where σ^2 is an overall variance parameter and $\mathbf{R}_\rho$ is a CAR or SAR covariance matrix that depends

on the single parameter ρ. The most dramatic feature of Figure 1 is the much improved fit when using row-standardization with the CAR models. The best overall models suggested by AIC were the first and fourth-order, row-standardized CAR models that had stock as an explanatory variable.

Because there are only 2 covariance parameters, it is very instructive to look at the restricted log-likelihood for the variance and range parameters. Consider the model

$$(y) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{R.2}$$

where $\mathbf{X}$ contains indicator variables for genetic stock membership, and $\boldsymbol{\varepsilon}$ has a CAR covariance matrix

$$\text{var}(\boldsymbol{\varepsilon}) = \sigma^2(\mathbf{D} - \rho\mathbf{W}_4)^{-1}$$

where $\mathbf{D}$ is a diagonal matrix with elements $\mathbf{W}_4\mathbf{1}$, $\mathbf{1}$ is a vector of all ones, and recall that $\mathbf{W}_4$ is the fourth-order neighborhood matrix. Note that this is an equivalent way to write a CAR model with row-standardization (Section 7.3). The resulting log-likelihood surface is shown in Figure 2, and the restricted maximum likelihood estimate is the value at the maximum of this surface, shown as a black circle, where $\hat{\rho} = 0.761$ and $\hat{\sigma}^2 = 0.267$. There are several ways to make inference on the covariance parameters, including profile likelihood and Fisher's Information matrix.

Consider the REML log-likelihood function in Section 8.2, which here is denoted as $L_{-i,R}(\theta_i; \hat{\boldsymbol{\theta}}_{-i}, \mathbf{y})$, where the ith component of $\boldsymbol{\theta}$ has been held constant at θ_i, and the function has been maximized for all other parameters, whose values are denoted as $\hat{\boldsymbol{\theta}}_{-i}$. For our model, $\boldsymbol{\theta} = (\rho, \sigma^2)$. Note that $\hat{\boldsymbol{\theta}}_{-i}$ changes with each i, but we suppress any notation to indicate such dependence. Then a profile likelihood plot for the ith component of $\boldsymbol{\theta}$ is one that plots $2L_{-i,R}(\theta_i; \hat{\boldsymbol{\theta}}_{-i}, \hat{\boldsymbol{\beta}}, \mathbf{y})$ for various values of θ_i, and these are seen in Figure 2B for σ^2 and Figure 2C for ρ. A $1-\alpha$ level confidence interval for θ_i can be obtained by finding all values of θ_i in the profile likelihood where $2L_{-i,R}(\theta_i; \hat{\boldsymbol{\theta}}_{-i}, \mathbf{y})$ are greater than $\hat{\theta}_i - \chi^2(1-\alpha, 1)$, where $\chi^2(x; \nu)$ is a quantile function, $0 \le x \le 1$, of a chi-squared distribution on ν degrees of freedom. Using $\alpha = 0.05$ leads to a 95% confidence interval and the well-known value of $\chi^2(0.95, 1) = 3.841$, and $\hat{\theta}_i - 3.841$ are shown by the dashed lines in Figure 2B,C. The confidence interval is the solid horizontal line, above which all values of $L_{-i,R}(\theta_i; \hat{\boldsymbol{\theta}}_{-i}, \hat{\boldsymbol{\beta}}, \mathbf{y})$ are greater than $\hat{\theta}_i - 3.841$. For ρ, the 95% confidence interval is from 0.347 to 0.941, and for σ^2, it is from 0.238 to 0.324.

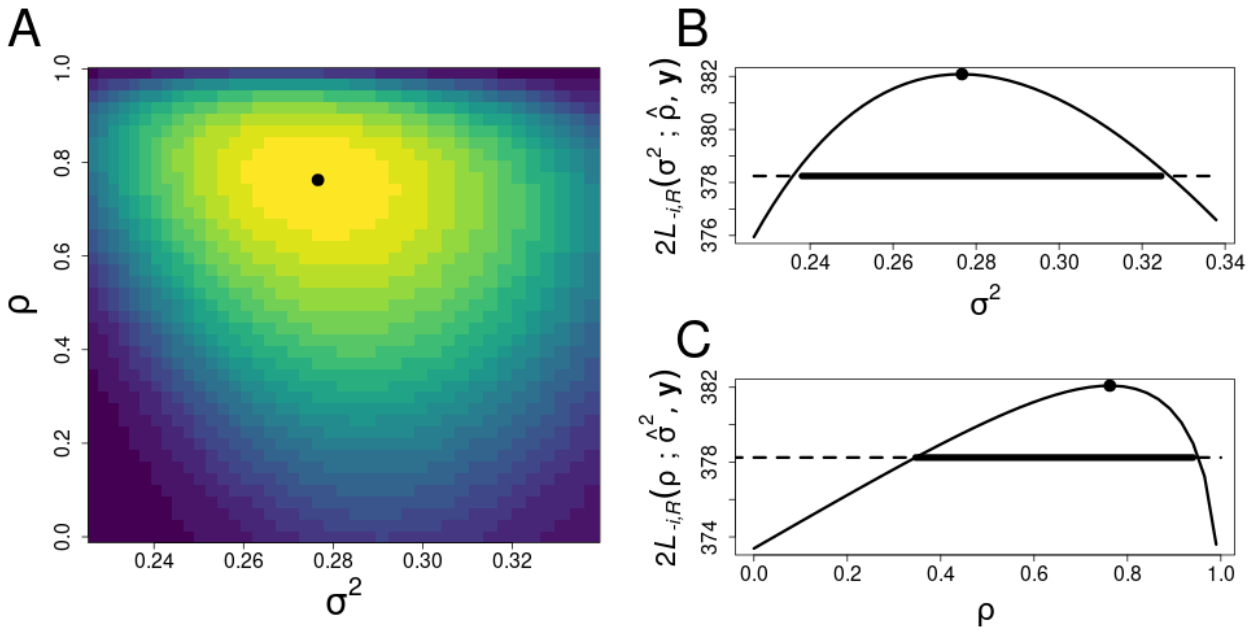

Figure 2: Restricted log-likelihood. A. The restricted log-likelihood surface for ρ and σ^2. More yellow colors are higher values, and bluer colors are lower values. The solid black circle is the maximum of the surface, yeilding the MLE. B. Profile likelihood confidence interval for σ^2. The curve is 2 times the restricted log-likelihood optimized for all parameters except σ^2, which is held constant at the value given by the x-axis. The solid black circle is the MLE, $\hat{\sigma}^2$, and the dashed line is $\hat{\sigma}^2 - \chi^2(1-\alpha, 1)$ for $\alpha = 0.05$. The horizontal black line forms the 95% confidence interval. C. Profile likelihood confidence interval for ρ in the same way as described above.

A second approach to finding a confidence interval is based on large sample asymptotics, where the curvature of the likelihood in Figure 2 at the REMLE (or we could also use MLE) is approximated by the observed Fisher's Information matrix, which can be used to construct confidence intervals as suggested in Section 8.4. After fitting the model, we compute observed Fisher's Information,

$$\boldsymbol{J}_a(\boldsymbol{\theta})_{i,j} = \frac{1}{2}\,\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_i}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_j}\right),$$

by using $\boldsymbol{\Sigma}^{-1} = (\mathbf{D} - \hat{\rho}\mathbf{W})/\hat{\sigma}^2$ and noting that

$$\begin{aligned}\frac{\partial \boldsymbol{\Sigma}}{\partial \sigma^2} &= (\mathbf{D} - \rho\mathbf{W}_4)^{-1},\\ \frac{\partial \boldsymbol{\Sigma}}{\partial \rho} &= \sigma^2(\mathbf{D} - \rho\mathbf{W}_4)^{-1}\mathbf{W}_4(\mathbf{D} - \rho\mathbf{W}_4)^{-1},\end{aligned}$$

where σ^2 and ρ are replaced by $\hat{\sigma}^2$ and $\hat{\rho}$, respectively.

An estimator of the asymptotic standard errors of $\hat{\boldsymbol{\theta}}$ are the square roots of the diagonal of $[\boldsymbol{J}_a(\boldsymbol{\theta})]^{-1}$. Formally, let $\mathbf{j}_i$ be a vector of all zeros, except there a single one at the ith element. Then a $(1-\alpha)100\%$ confidence interval for $\hat{\theta}_i$ is

$$\hat{\theta}_i \pm z_{1-\alpha/2}\sqrt{\mathbf{j}_i^T[\boldsymbol{J}_a(\boldsymbol{\theta})]^{-1}\mathbf{j}_i},$$

where $z_{1-\alpha/2}$ is a quantile of a standard normal distribution at $1-\alpha/2$. For example, if $\alpha = 0.05$, then $z_{1-\alpha/2}$ is the familiar 1.96. In our example, the standard errors for ρ and σ^2 were 0.101 and 0.018, respectively, so the 95% confidence interval for ρ was from 0.563 to 0.959 and for σ^2 it was from 0.240 to 0.312, which can be compared to those obtained from profile likelihood.

Another approach is to compute the Hessian matrix of the log-likelihood. We used `spmodel` for this as it allowed us to evaluate the log-likelihood for any given value of $\boldsymbol{\theta}$. We used the `R` package `numDeriv` to compute the numeric Hessian, $\mathbf{H}$, at the REMLE, and then an estimator of Fisher's Information is,

$$\boldsymbol{J}_n(\boldsymbol{\theta}) = -\mathbf{H},$$

and an estimator of the asymptotic standard errors follow in exactly the same way as they were developed when using $\boldsymbol{J}_a(\boldsymbol{\theta})$. In our example, the standard errors for ρ and σ^2 were 0.145 and 0.023, respectively, so the 95% confidence interval for ρ was from 0.476 to 1.046 and for σ^2 it was from 0.231 to 0.321, which can be compared to those obtained from the analytical computation of Fisher's Information and profile likelihood. Here, the confidence interval for ρ extends beyond 1 at the upper bound, showing a disadvantage of the large-sample asymptotic approach that relies on normality.

Nevertheless, it is interesting to look at the asymptotic correlation among the parameters as revealed by the observed Fisher's Information matrix. Let $\mathbf{S}_a$ be a diagonal matrix containing reciprocals of the square roots from the diagonal of $[\boldsymbol{J}_a(\boldsymbol{\theta})]^{-1}$, and similarly obtain $\mathbf{S}_n$ from $[\boldsymbol{J}_n(\boldsymbol{\theta})]^{-1}$. Then estimators of the asymptotic correlation matrix are

$$\mathbf{S}_a[\boldsymbol{J}_a(\boldsymbol{\theta})]^{-1}\mathbf{S}_a = \begin{pmatrix} 1.000 & 0.163 \\ 0.163 & 1.000 \end{pmatrix} \text{ and } \mathbf{S}_n[\boldsymbol{J}_n(\boldsymbol{\theta})]^{-1}\mathbf{S}_n = \begin{pmatrix} 1.000 & -0.189 \\ -0.189 & 1.000 \end{pmatrix},$$

which show little correlation between ρ and σ^2, and this is also evident in the shape of the log-likelihood surface (Figure 2A).

One of the features of the spatial-weights models is that they are nonstationary, and it is interesting to investigate this property for actual data. First, consider two models as in (R.2), using the fourth-order neighborhood structure, one where the $\mathbf{W}$ is standardized, i.e., $\text{var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}_{rs} = \sigma^2_{rs}(\mathbf{I} - \rho_{rs}\overline{\mathbf{W}}_4)\mathbf{K}_{rs}$, and the other where it is not, i.e., $\text{var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}_{un} = \sigma^2_{un}(\mathbf{I} - \rho_{un}\mathbf{W}_4)\mathbf{K}_{un}$. The diagonal elements of $\boldsymbol{\Sigma}_{rs}$, plotted as a function of the number of neighbors, is shown in Figure 3A. Notice that the marginal variance goes down with increasing numbers of neighbors. This is reasonable because, recall from eq. 7.6, the c_{ij}

weights are 1/(number of neighbors) when they are row standardized, so we are *averaging* over neighboring values, and the variance of an average goes down with larger sample sizes. On the other hand, the marginal variances of $\mathbf{\Sigma}_{un}$ are shown in Figure 3B, where $c_{ij} = 1$ in (7.6) so we are *summing* over neighboring values, and the variance of a sum goes up with larger sample sizes. Moreover, even for a fixed number of neighbors, the variance is not constant because (7.6) specifies a conditional expectation and the marginal variances involve the inverse of the matrix specified by these weights. In summary, both of these models exhibit nonstationary variances, in contrast to geostatistical models, where variances are constant regardless of distance.

Similarly, we can look at autocorrelation as a function of the order of the neighbor. Let $\mathcal{N}_i^{[1]}$ be the set of all neighbors of site i, $\mathcal{N}_i^{[2]}$ be the set of all neighbors of $\mathcal{N}_i^{[1]}$, exclusive of any already in $\mathcal{N}_i^{[1]}$, $\mathcal{N}_i^{[3]}$ be the set of all neighbors of $\mathcal{N}_i^{[2]}$, exclusive of any already in $\mathcal{N}_i^{[1]}$ or $\mathcal{N}_i^{[2]}$, etc., up to sixth-order neighbors. Then pairwise autocorrelations taken from $\mathbf{\Sigma}_{rs}$, as a function of neighbor order, are plotted as violin plots in Figure 3C. In general autocorrelation decreases as a function of neighbor order, but there is wide variation for a given neighbor order. It is also interesting to see that for the fourth-order neighbor model that was fit, there is a sudden decrease in autocorrelation, on average, after order four. Nonetheless, Figure 3C also shows that autocorrelation exists in $\mathbf{\Sigma}_{rs}$ beyond fourth order, even though $\overline{\mathbf{W}}_4$ contains all zeros beyond fourth order. Like variances, this is due to the fact that $\mathbf{\Sigma}_{rs}$ is obtained through an inverse involving $\overline{\mathbf{W}}_4$, rather than $\overline{\mathbf{W}}_4$ directly.

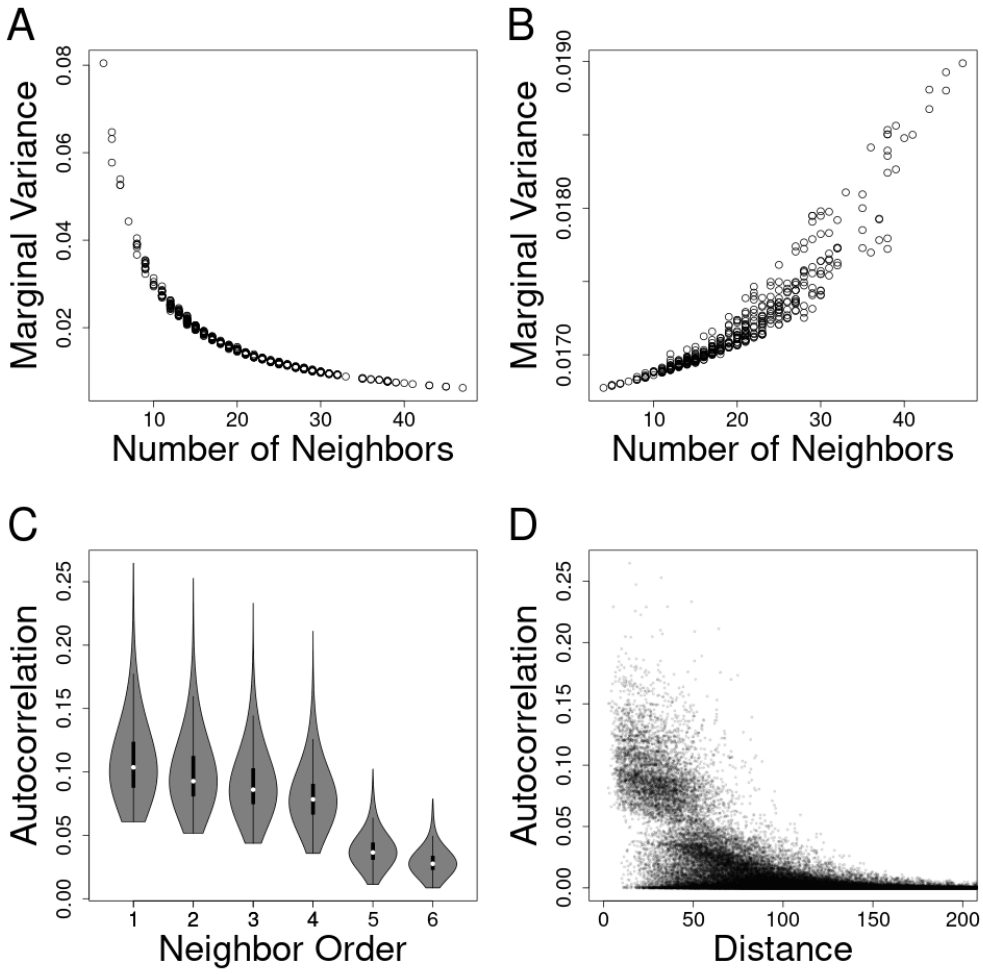

Figure 3: Nonstationarity in spatial-weights models. A. Marginal variance as a function of number of neighbors for a row-standardized CAR model. B. Marginal variance as a function of number of neighbors a binary weights CAR model. C. Violin plots of autocorrelation as a function of neighbor order for row-standardized cAR model. D. Scatterplot of all pairwise correlations as a function of centroid distance for row-standardized CAR model.

Although distance is not often well-defined for the spatial-weights models, such as when sites are polygons as in this harbor seals example, one unique definition of distance is obtained by associating a centroid with each polygon, and then taking Euclidean distances among centroids. Plotting autocorrelation as a function of centroid distance shows that, in general, autocorrelation decreases with distance (Figure 3D), just as it did for neighbor order (3C). Again, compare this situation to an isotropic geostatistical model, where autocorrelation is fixed for any given distance between two locations.

Before moving to estimation of fixed effects, we consider a few more models. One very handy feature of the `spmodel` software is that it allows the estimation of a separate

variance parameter for sites that are isolated, as discussed in Section 7.4. By default, the software uses the geometry in `sf` objects and determines neighbors as any polygons that share a common boundary. Fitting a CAR model in this way, including the genetic stock explanatory variable, and ordering the data so that the isolated polygons are listed first, the ML estimated covariance matrix for the row-standardized model was

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} 0.043 \times \mathbf{I} & \mathbf{0} \\ \mathbf{0} & 0.032 \times (\mathbf{I} - 0.264 \times \overline{\mathbf{W}})^{-1}\overline{\mathbf{K}} \end{pmatrix},$$

and $-2L(\boldsymbol{\theta}, \mathbf{y}) = -418.99$. With five fixed effects and three covariance parameters, AIC = -402.99, and, even though it has an extra covariance parameter, it is the best model when compared to all of those in Figure 1. Geostatistical models are stationary, and the centroids can be used as distance between polygons. Using distance in this way, the best geostatistical model with the explanatory variable, using MLE, was a circular model, which had $-2L(\boldsymbol{\theta}, \mathbf{y}) = -390.66$, and, again with three covariance parameters, AIC was -374.66, which was considerably worse than the row-standardized CAR model with islands. Finally, the variance-stabilizing weights of Section 7.3 were used in a CAR model (without islands) and with the explanatory variable, yielding $-2L(\boldsymbol{\theta}, \mathbf{y}) = -399.93$, and, with two covariance parameters, AIC was -385.93.

Any choice of a covariance model, and fitting it, leads to the next topic – inference about the fixed effects – which was broadly covered in Chapter 5. Let $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ be estimated by ML using a row-standardized CAR model with islands, as described in the previous paragraph, with a mean structure that includes an overall intercept and genetic stock membership as a categorical explanatory variable, as this appeared to be the best model according to AIC. Then an empirical *GLS* estimate of stock effect $\tilde{\tilde{\boldsymbol{\beta}}} = [\mathbf{X}^T[\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})]^{-1}\mathbf{X}]^{-1}\mathbf{X}^T(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}))^{-1}\mathbf{y}$ is given in Table 1 in typical `R` format, where $\tilde{\tilde{\boldsymbol{\beta}}}$ is given in the column labeled as "Estimate."

Table 1: Coefficient table of fixed effects estimated by ML using a row-standardized CAR model with islands, as given by the `spmodel` package in `R`. The first column is the estimated coefficients of $\tilde{\tilde{\boldsymbol{\beta}}}$, the second column is the estimated standard error, the third column is a computed z-value where column 1 is divided by column two, and the fourth column is the estimated probability of obtaining the z-value if the null hypothesis that the effect was zero was true.

Coefficients	Estimate	Std. Error	z-value	Pr(> |z|)
(Intercept)	0.0069	0.0127	0.547	0.58435
stock Dixon/Cape Decision	0.0449	0.0200	2.249	0.02453
stock Glacier Bay/Icy Strait	-0.0785	0.0265	-2.970	0.00298
stock Lynn Canal/Stephens	-0.0334	0.0242	-1.381	0.16743
stock Sitka/Chatham	0.0129	0.0211	0.613	0.53999

Standard errors for $\tilde{\hat{\beta}}_i$, the ith element of $\tilde{\hat{\boldsymbol{\beta}}}$, are estimated as $\tilde{\hat{\sigma}}_i = \sqrt{[\mathbf{X}^T[\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})]^{-1}\mathbf{X}]^{-1}_{i,i}}$. As discussed in Section 5.7, distributional properties of the empirical generalized least squares are complicated due to the fact that, from (8.3), $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}) = \tilde{\hat{\sigma}}^2\mathbf{R}(\hat{\boldsymbol{\theta}}_{-1})$. Nevertheless, assuming that $\tilde{\hat{\boldsymbol{\beta}}}$ is approximately normal, a z-value can be formed as $\tilde{\hat{\beta}}_i/\tilde{\hat{\sigma}}_i$, the third column in Table 1. Let z have a standard normal distribution with cumulative distribution function $\boldsymbol{\Phi}(z)$. Then an approximate α-level test against the null hypothesis that $\tilde{\hat{\beta}}_i = 0$ is given by the probability of obtaining the computed z-value, or larger, $\text{Prob}(|\tilde{\hat{\beta}}_i/\tilde{\hat{\sigma}}_{i,i}| > z)$, so $\alpha = 2 \times (1 - \boldsymbol{\Phi}(|\tilde{\hat{\beta}}_i/\tilde{\hat{\sigma}}_i|))$, which is the fourth column in Table 1.

It is also possible to pivot on the estimated z-value to obtain confidence intervals using the inverse cumulative distribution function, also called a quantile function, $\boldsymbol{\Phi}^{-1}(p)$. Let $z_{1-\alpha/2} = \boldsymbol{\Phi}^{-1}(1-\alpha/2)$, then a $(1-\alpha/2)\times 100\%$ confidence interval for $\tilde{\hat{\beta}}_i$ is $\tilde{\hat{\beta}}_i \pm z_{1-\alpha/2}\tilde{\hat{\sigma}}_i$. For example, recall that the stock effect for "Clarence Strait" has been absorbed into the overall intercept, so that the effect "Dixon/Cape Decision" is an estimate of the *difference* between Clarence Strait and Dixon/Cape Decision. A 95% confidence interval on this difference is $0.449 \pm 1.96 \times 0.02 = (0.0057, 0.0841)$. If an estimate of Dixon/Cape Decision is desired, then let $\boldsymbol{\ell}$ be the vector $(1, 1, 0, 0, 0)$, and then the estimator of the Dixon/Cape Decision effect is $\boldsymbol{\ell}^T\tilde{\hat{\boldsymbol{\beta}}}$, which for our example is 0.0518, with standard error $\sqrt{\boldsymbol{\ell}^T[\mathbf{X}^T(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}))^{-1}\mathbf{X}]^{-1}\boldsymbol{\ell}}$, which for our example is 0.0155. Then a 95% confidence interval for the log-trend of the Dixon/Cape Decision stock is 0.0215 to 0.0821, and we can be 95% certain that the stock is growing from somewhere between approximately 2% to 8% per year. In addition to estimates, we are often interested in specific contrasts. Figure 1.1 shows stocks labelled 8 and 9 as the two northern stocks, while stocks 10, 11, and 12 are southerns stocks. We would like to estimate the difference in trends between the northern and southern stocks. Stock 8 is Glacier Bay/Icy Strait, and stock 9 is Lynn Canal/Stephens, so the contrast of interest is $[\beta_1 + (\beta_2 + \beta_1) + (\beta_5 + \beta_1)]/3 - [(\beta_3 + \beta_1) + (\beta_4 + \beta_1)]/2 = \boldsymbol{\ell}^T\tilde{\hat{\boldsymbol{\beta}}}$, where $\boldsymbol{\ell} = (0, 1/3, -1/2, -1/2, 1/3)$. The estimate of this contrast is 0.0752, with standard error of 0.0177, and a 95% confidence interval difference in log-trend of the difference between the northern and southern stocks ranges from 0.0404 to 0.1101, and we can be 95% certain that the southern stocks trends are somewhere between approximately 4% to 11% higher per year than the northern stocks.

In addition to estimating individual effects, their linear combination, or a contrast, interest often centers on whether the stock effect as a whole should be included in the model. Chapter 5 developed an F-test for a joint linear hypothesis $\mathbf{L}\hat{\boldsymbol{\beta}} = \mathbf{0}$, where $\mathbf{0}$ is a vector of all zeros, in the case where $\mathbf{R}$ is known. The numerator of that F-test is distributed as a chi-squared random variable, which the `spmodel` package uses in the *EGLS* situation as an approximation to the distribution of $\mathbf{L}\tilde{\hat{\boldsymbol{\beta}}}$. Specifically, let $\mathbf{L}$ be the matrix $(\mathbf{0}_4|\mathbf{I}_4)$ where $\mathbf{0}_4$ is a vector of four zeros and $\mathbf{I}_4$ is the 4×4 identity matrix. Then, for our example,

$$X = (\mathbf{L}\tilde{\hat{\boldsymbol{\beta}}})^T[\mathbf{L}\mathbf{X}^T(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}))^{-1}\mathbf{X}\mathbf{L}^T]^{-1}\mathbf{L}\tilde{\hat{\boldsymbol{\beta}}} = 23.0895$$

and if $F_{\chi^2}(x, \nu)$ is the cumulative distribution function for a chi-squared random variable with ν degrees of freedom, then $= 1 - F_{\chi^2}(23.0895, 4) = 0.000122$ is the probability of obtaining

the observed value, or larger, if the null hypothesis were true. These results are obtained with the `anova` function in `spmodel`.

An alternative way to determine the effect of genetic stock, after accounting for the overall mean effect, is by a likelihood ratio test, which is very general. Recall that $-2L(\boldsymbol{\theta}; y, \mathbf{X}_{\text{stock}}) = -418.99$ for the model with stock as an explanatory variable, and allowing for isolated sites and a row-standardized CAR model. Using a mean-only model with the same covariance structure results in $-2L(\boldsymbol{\theta}; y, \mathbf{X}_{\text{ones}}) = -398.74$. The difference in these two log-likelihoods is 20.2427, and under large-sample aymptotics this has a chi-squared distribution, so $1 - F_{\chi^2}(20.2427, 4) = 0.000447$ and the probability of obtaining the observed value, or larger, if the mean-only model were true. These results can also be obtained with the `anova` function in `spmodel`.

9.10.1 Prediction of trends in harbor seal abundance

We continue with our analysis from Section 8.8.1 of the trends in harbor seal abundance in southeast Alaska. In Chapter 8, we made inferences on covariance parameters and fixed effects. Here, our inferential goal will be to make prediction, with prediction intervals, for sites with missing data, and also to investigate leave-one-out cross-validation (LOOCV) as a way to smooth autoregressive models.

Recall from the Chapter 8 harbor seal example that we discussed some computational issues when there are missing data for the spatial-weights models. From (R.1) we were able to obtain the inverse of the covariance matrix for the observed sites, which we called $\boldsymbol{\Sigma}_{oo}^{-1}$, from the inverse covariance matrix for all sites more efficiently than having to invert the inverse covariance matrix for all sites. For estimation of covariance parameters, and fixed effects, this was enough. However, for prediction, we also need $\boldsymbol{\Sigma}_{ou}$ and $\boldsymbol{\Sigma}_{uu}$ from the full covariance matrix (not its inverse), because, in the notation of Chapter 9, we need $\mathbf{R}^{-1}\mathbf{R}_{\mathbf{yu}}$ and $\mathbf{R}_{\mathbf{uu}}$ (see Section 9.1). It is possible to obtain $\mathbf{R}^{-1}\mathbf{R}_{\mathbf{yu}}$ by noting from (R.1) that $\boldsymbol{\Sigma}_{oo}^{-1}\boldsymbol{\Sigma}_{ou} = -\boldsymbol{\Sigma}^{ou}(\boldsymbol{\Sigma}^{uu})^{-1}$ and we may have already computed $(\boldsymbol{\Sigma}^{uu})^{-1}$ for likelihood evaluations and estimation of covariance parameters and fixed effects. However, for $\boldsymbol{\Sigma}_{uu}$, it appears that we need another inverse on the order of the $\boldsymbol{\Sigma}_{oo}$ by noting that $\boldsymbol{\Sigma}_{uu} = (\boldsymbol{\Sigma}^{uu})^{-1} + (\boldsymbol{\Sigma}^{uu})^{-1}\boldsymbol{\Sigma}^{uo}\boldsymbol{\Sigma}_{oo}^{-1}\boldsymbol{\Sigma}^{ou}(\boldsymbol{\Sigma}^{uu})^{-1}$. It is simpler to take $(\boldsymbol{\Sigma}^{-1})^{-1}$, but this only needs to be done once for prediction, while $\boldsymbol{\Sigma}_{oo}^{-1}$ needs repeated evaluations during likelihood optimization.

—— start LOOCV and n-fold ———

[Dale, the following is the same description of LOOCV that I wrote for the wet sulfate data. I think that this description, along with n-fold cross-validation, could be a topic in Chapter 9 apart from the examples, as I will likely refer to them for all of the examples. I will leave this here for now. Please feel free to re-organize.]

LOOCV eliminates one datum at a time, using all of the rest of the data to predict the one that was removed. There is a fast and a slow way to do this. The slow way is to remove a datum and then re-estimate all of the parameters using ML or REML estimation each time. However, with the removal of but a single datum, the parameter estimates change

very little. A fast way to achieve LOOCV is based on holding all parameters at their values as estimated by using all of the data, and then using results from partitioned matrices so that we only have to invert the covariance matrix once (which can be saved from the ML or REML estimation, so there is in fact no additional matrix inverses are required). Recall from Section 5.6 that, if a matrix is partitioned as,

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \text{ and } \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}^{11} & \boldsymbol{\Sigma}^{12} \\ \boldsymbol{\Sigma}^{21} & \boldsymbol{\Sigma}^{22} \end{bmatrix},$$

then

$$\boldsymbol{\Sigma}_{11}^{-1} = \boldsymbol{\Sigma}^{11} - \boldsymbol{\Sigma}^{12}(\boldsymbol{\Sigma}^{22})^{-1}\boldsymbol{\Sigma}^{21}.$$

Moreover, let us order the data such that the datum to be removed is last, so that $\boldsymbol{\Sigma}^{22}$ is a scalar, then the inverse of $\boldsymbol{\Sigma}^{22}$ is trivial and $\boldsymbol{\Sigma}_{11}^{-1}$ can be computed rapidly. The main computational expense of kriging predictions rely on the inverse covariance matrix for the observed data, but in LOOCV that is given by $\boldsymbol{\Sigma}_{11}^{-1}$, which is computed rapidly without any further matrix inverses if we already have $\boldsymbol{\Sigma}^{-1}$. The only other quantity from $\boldsymbol{\Sigma}$ needed for prediction is the vector $\boldsymbol{\Sigma}_{12}$. Conceptually, we just re-order the data, one at a time, putting the one to be removed last in the covariance matrices above, and that allows the predictions to be computed quickly.

Let $\hat{y}_i = \bar{u}$ from (9.2) be the ith predicted value using LOOCV where the ith datum has been removed, and let $\hat{v}_i = \sqrt{\text{var}(\bar{u} - u)}$ from (9.3) be the i prediction standard error. Then we will consider two metrics to assess model performance. One is the root-mean-squared prediction error (RMSPE), which we computed as

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}.$$

Models with lower RMSPE have better predictive performance. The other metric is the 90% prediction interval coverage, PIC90, which we computed as

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}(\hat{y}_i - z_{1-\alpha/2}v_i \le y_i \le \hat{y}_i + z_{1-\alpha/2}v_i),$$

where $\mathcal{I}(\cdot)$ is an indicator function, equal to 1 if its argument is true, and 0 otherwise, and $z_{1-\alpha/2}$ is a standard normal value below which contains $1 - \alpha/2$ of the probability density. We chose $\alpha = 0.1$ for PIC90, resulting in the familiar $z_{0.95} = 1.645$.

One problem with LOOCV is that it may be overly optimistic in assessing actual prediction. A simple thought experiment reveals why. Suppose 99 data locations were clustered very closely together, and another was separated from the cluster. Then, for LOOCV, as we removed each datum from the cluster, we would have many nearby locations and get very precise predictions with small prediction standard errors, and they would swamp RMSPE and PIC90 if our overall goal was to predict in a region that was substantially larger than that enclosing the cluster of locations. The same is essentially true if all locations were

pairs of locations that were very close to each other, but the pairs were scattered. Still, under normal sampling scenarios, it can be a good way to evaluate models as they are all operating under the same sampling scheme.

A second way to use cross-validation is called n-fold cross-validation. Here we divide the data into n groups, and remove one whole group to be predicted – this is often called the *test* dataset. The remaining data are called the *training* dataset, and are used to fit a model and make predictions at the locations of the test dataset. Then, the predictions at the locations for the test dataset can be compared to the actual values that were removed. In fact, RMSPE and PIC90 can be computed for n-fold cross-validation in exactly the same way as for LOOCV, and to distinguish them, we use RMSPE_{Lo} and PIC90_{Lo} for LOOCV, and RMSPE_{Nf} and PIC90_{Nf} for n-fold cross-validation. With n-fold cross-validation we do not have to fit the model as many times, so we completely re-fit the model for each group that is removed. Groups are often created randomly, and we will do it this way too. If we want to get the best feel how well a model will interpolate, that includes even a bit of extrapolation at the edges, it is desirable to have just a few groups. This may be more pessimistic about model performance than the real data because we are decreasing our sample sizes substantially. Nevertheless, to examine both extremes, where LOOCV is overly optimistic, we used 3-fold cross-validation, which is overly pessimistic. Because we created 3 groups randomly, we would like to ensure that our results do not depend too much on any particular randomized grouping. Hence, we do 3-fold cross-validation 10 times, and average the results for RMSPE (by first averaging the mean-squared prediction error, and then taking the square root) and PIC90.

—— end LOOCV and n-fold ————

Based on our evaluation of likelihoods in Section 8.8.1, CAR models with row-standardization appeared much better than SAR models and models without row-standardization. It also appeared that the model with isolated sites, at the cost of an extra parameter, was better than any of those where all sites were forced to be connected. Does this translate to predictive performance? We investigated using LOOCV. All models discussed will contain the stock effect unless otherwise noted and are estimated with REML. The CAR row-standardized covariance with islands (three covariance parameters) was still best, based on LOOCV RMSPE, with a value of 0.01739. This was followed by the same model without row standardization with RMSPE = 0.01751. A CAR covariance model without islands, based on a first-order neighbor structure, was next with RMSPE = 0.01765, but the same model with a fourth-order neighbor structure did very poorly with RMSPE = 0.02775, even though this was the best model in Figure 1B according to AIC. This reinforces the idea that a variety of model checks should be performed before settling on any one model, and it can be difficult to come to a decision when different criteria identify different "best" models. A model without any spatial autocorrelation, assuming uncorrelated random errors, had RMSPE = 0.01777. Finally, we tried a model based purely on spatial autocorrelation, without any fixed effects except a constant mean, and a CAR row-standardized covariance with islands, that resulted in RMSPE = 0.01800.

Based on these results, we feel confident in proceeding with the CAR row-standardized covariance with islands. An important inference concerning these data was about the effect of genetic stock on trend, which was explored in Chapter 8.8.1, but now we would also like to predict values for missing data, and make smoothed maps of the existing data. To explain smoothing, Figure **??**A shows a histogram of the observed values and the LOOCV values. It is possible for predictions to be more extreme than observed values, but here, with a covariate effect of group means, and fairly weak autocorrelation, predictions "shrink" away from extremes. LOOCV predictions are a combination of the estimated stock mean plus and a weighted average of local residuals, which causes the predictions to be much less variable than the original data. Predictions of the missing data are presented in Figure **??**B. For the independence model, the predictions can take on only one of five possible values, which are the five estimated stock means, shown by the darkest black bars. Predictions for the pure mean model, without the stock effect, are shown, as well as those with the stock effect. When spatial autocorrelation is included in the model with stock effect, the predictions can take on values that deviate from the stock means because they include weights from the residuals of neighboring sites in the same way that they did for LOOCV. Broadly speaking, the spreads of the predictions are roughly similar for the mean-only model and the one that includes the stock effect, and they are very similar to the spread of the LOOCV predictions. LOOCV, combined with prediction of missing data, is one way to smooth maps, but there are many others. Here, we will use LOOCV.

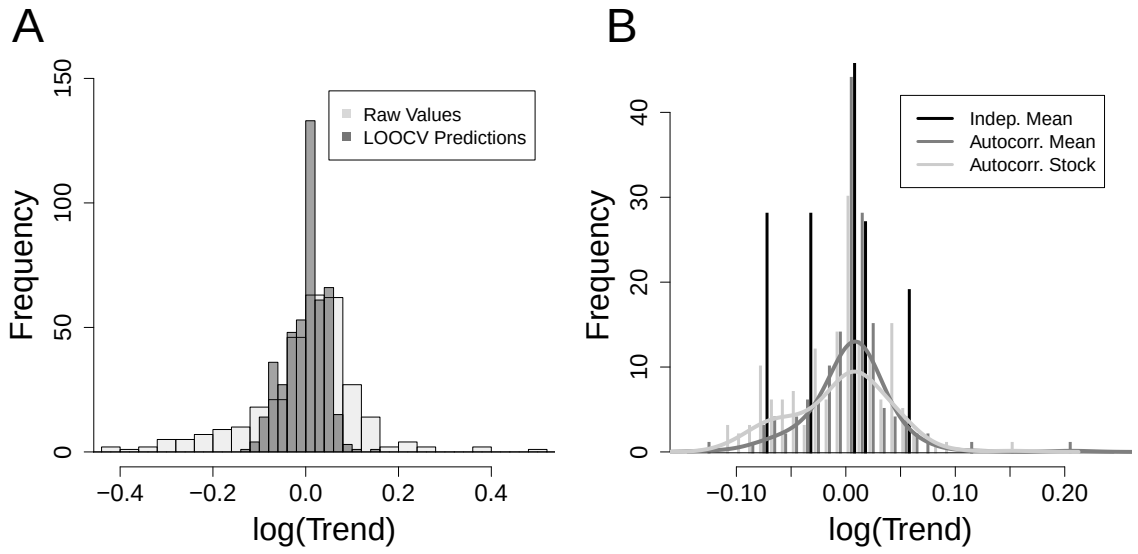

Figure 4: Histograms of raw data and predictions. A. Darker shaded histogram is laid over the histogram of the raw data in a lighter shade. B. Histograms of predictions for three different models. The darkest shade is a stock effect model assuming independent errors. The middle shade has autocorrelated errors, but a single mean effect. The lightest shade has stock effects in the model, along with autocorrelated errors.

Figure 5A show predictions for all 159 site with missing values. These are shown as colored circles, and are superimposed on the colored polygons of raw values. Notice that the range of values, given by the legend, are the same as the range for the raw values in Figure 4. The prediction standard errors (Figure 5B) are highest near the edges where sites have fewer neighbors, as we expect from the row-standardized model. On occasion, high standard errors occur for interior sites because it is possible for them to have few neighbors. The predictions of missing data, along with LOOCV values for locations with raw data, are given in Figure 5C. Notice (from the legend) that the range of values is much narrower than Figure 5A, and this qualifies as a smoothed map. The effect of stock mean is fairly evident in Figure 5C. As in Figure 5B, the prediction standard errors (Figure 5D) are highest near the edges where sites have fewer neighbors. Figure 5E show predictions of missing data, along with LOOCV, for for the same covariance model as previously, but without the stock effects (a common mean). Notice that this is also a smoothed map with slightly more range in values than the stock-effect model, and the effect of stock is not as evident. The standard error map (Figure 5F) looks very similar to the other standard error maps.

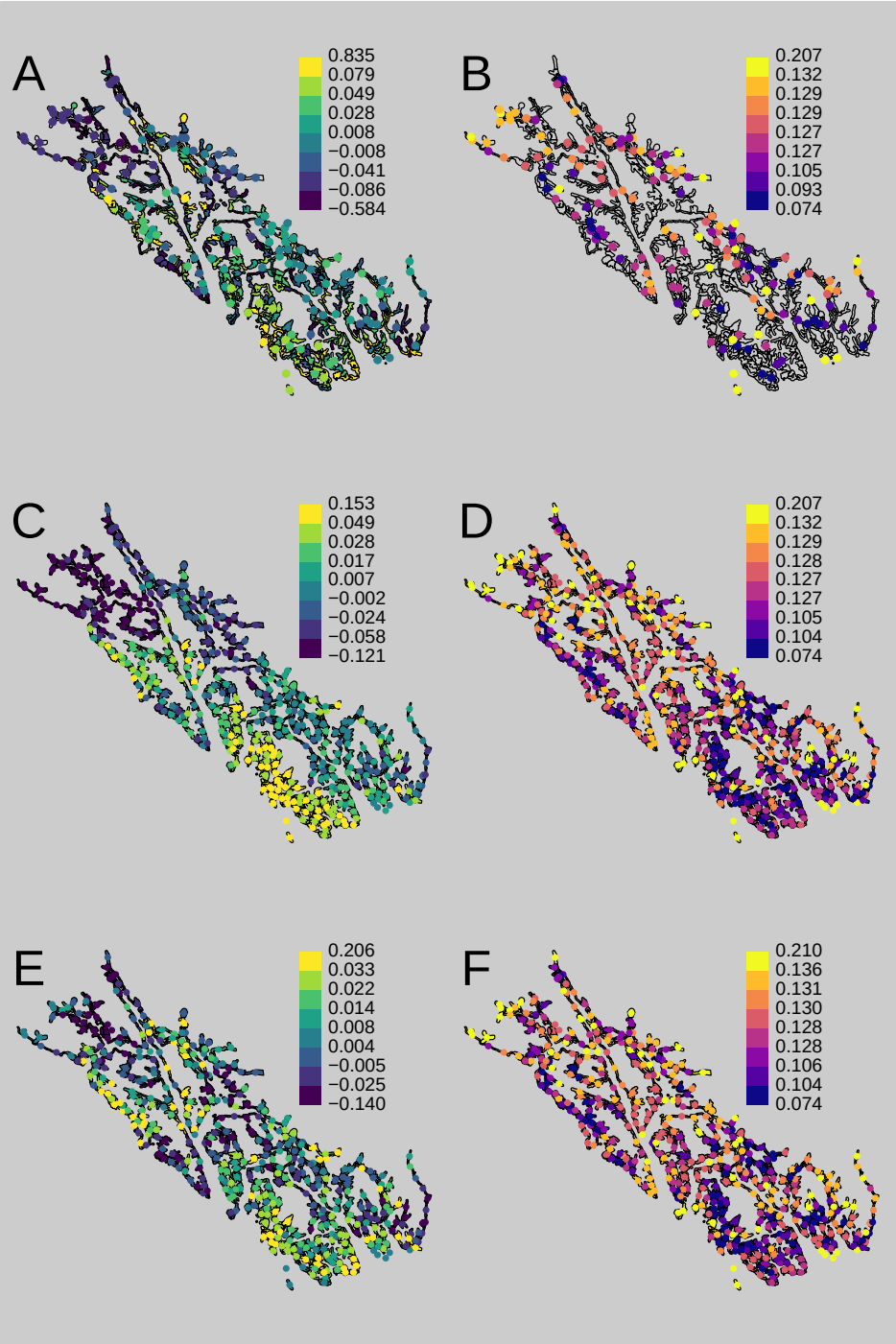

Figure 5: Prediction maps (left column) and prediction standard errors (right column) for a variety of models. A-B Model includes stock effect, and a row-standardized CAR model with islands (no neighbors, resulting in an extra parameter). C-D LOOCV, rather than raw values, for the same model, E-F LOOCV and predictions for the same covariance model as above but with a common mean model as the single fixed effect.