Species Density Models from Opportunistic Citizen Science Data

Jay M. Ver Hoef[1], **Devin Johnson**[1], **Robyn Angliss**[1], **Matt Higham**[2]

---

[1]Marine Mammal Laboratory, NOAA-NMFS Alaska Fisheries Science Center
7600 Sand Point Way NE, Seattle, WA 98115
E-mail: jay.verhoef@noaa.gov tel: (907) 347-5552
[2]Department of Statistics
St. Lawrence University, Canton, NY

Running Headline: SDMs from Opportunistic Citizen Science Data

---

June 25, 2021

ABSTRACT

- With the advent of technology for data-gathering and storage, opportunistic citizen-science data are proliferating. Species distribution models (SDMs) aim to use species occurrence or abundance for ecological insights, prediction, and management. We analyzed a massive opportunistic data set with over 100,000 records of incidental shipboard observations of marine mammals. Our overall goal was to create maps of species density from massive opportunistic data by using spatial regression for count data with an effort offset. We illustrate the method with two marine mammals in the Gulf of Alaska and Bering Sea.

- We counted the total number of animals in 11,424 hexagons based on presence-only data. To decrease bias, we first estimated a spatial density surface for ship-days, which was our proxy variable for effort. We used spatial considerations to create pseudo-absences, and left some hexagons as missing values. Next, we created SDMs that used modeled effort to create pseudo-absences, and included the effort surface as an offset in a second stage analysis of two example species, northern fur seals and Steller sea lions.

- For both effort and species counts, we used spatial count regression with random effects that had a multivariate normal distribution with a conditional autoregressive (CAR) covariance matrix, providing 2.5 million Markov chain Monte Carlo (MCMC) samples (1000 were retained) from the posterior distribution. We used a novel MCMC scheme that maintained sparse precision matrices for observed and missing data when batch sampling from the multivariate normal distribution. We also used a truncated normal distribution to stabilize estimates, and used a look-up table for sampling the autocorrelation parameter. These innovations allowed us to draw several million samples in just a few hours.

- From the posterior distributions of the SDMs, we computed two functions of interest. We normalized the SDMs and then applied an overall abundance estimate obtained from the literature to derive spatially explicit abundance estimates, especially within subsetted areas. We also created "certain hotspots" that scaled local abundance by standard deviation and using thresholds. Hexagons with values above a threshold were deemed as hotspots with enough evidence to be certain about them.

KEY WORDS: Conditional autoregressive, species distribution models, opportunistic data, marine mammals, density models

1 INTRODUCTION

Natural resource surveys conducted with a structured sampling design are the preferred method for most ecological assessments because they provide results with the highest levels of confidence for the species or system of interest. However, more recently, citizen science efforts are being incorporated into scientific research, including data collection, analyses, and interpretation (Miller-Rushing et al., 2012). Species distribution models (SDMs) (Elith and Leathwick, 2009; Guisan et al., 2013), which aim to use species occurrence or abundance for ecological insights and prediction, are as fundamental as the definition of ecology itself (Krebs, 1972). At the intersection of citizen science and SDMs are opportunistic data sets of species occurrence/abundance (Soroye et al., 2018) and analytical methods (Elith et al., 2006) that are rapidly growing (Renner et al., 2015; Lukyanenko et al., 2020). Our overall goal is to create maps of species density from massive opportunistic data sets by using spatial regression for count data with an effort offset.

Most citizen science data are obtained without a formal sampling design, and hence, according to Kelling et al. (2019), there is no "fully statistically defensible way of accounting for the biases inherent in the data collection." Statistical bias occurs when the expected value of a statistical technique is different from the true quantity that it is estimating. A sampling design controls the distribution of observation and measurement effort, which generally allows for unbiased estimates. For example, an unbiased estimator for a population total, τ, introduced by Horvitz and Thompson (1952), is,

$$\hat{\tau} = \sum_{i=1}^{n} y_i/\pi_i, \tag{1}$$

where $y_i, i = 1, 2, \ldots, N$, is a value from N population units, π_i is the probability that unit i was included in the sample, n ($< N$) is sample size, and τ is the sum of $\{y_i\}$ over all N sample units. The $\{\pi_i\}$ are probabilities, but it is interesting to view them as the "effort" to include an

observation, which is sufficient for unbiased estimation.

Also consider Poisson regression for rates. We often have counts that need to be put on an equal basis with other counts. For example, the number of diseases among a county population, the number of animals per geographic area, or the number of manufacturing errors per unit time. Let y_i be the count, then y_i/π_i is the rate (diseases per person, animals per area, or manufacturing errors per time), where here we use π_i to be county population, geographic area, and time, respectively. We assume that y_i is the result of a random variable, Y_i, and a reasonable model is $g(\mathrm{E}[Y_i]/\pi_i) = \mathbf{x}_i'\boldsymbol{\beta}$, where $\mathbf{x}_i$ is a vector of covariates associated with the ith observation, $\boldsymbol{\beta}$ is a vector of regression parameters, and $g(\cdot)$ is a link function that keeps counts positive. For most count distributions (e.g., Poisson, negative binomial, etc.), we model the log of the mean as linear, so $\log(\mathrm{E}[Y_i]/\pi_i) = \mathbf{x}_i'\boldsymbol{\beta}$, where the distribution provides the error terms. The value π_i is fixed, so we move the denominator of the rate to the right of the equal sign,

$$\log(\mathrm{E}[Y_i]) = \log(\pi_i) + \mathbf{x}_i'\boldsymbol{\beta}, \tag{2}$$

and $\log(\pi_i)$ is known as the "offset," which can be considered as another covariate with known regression coefficient equal to one.

In both Equations (1) and (2) we used notation π_i to make the connection that dividing by a probability, an area, or some proxy to effort, is needed; without $\{\pi_i\}$, we do not know how to properly weight $\{y_i\}$ to construct an unbiased estimator. Citizen science data, where observations are collected opportunistically, often lack information on effort, which is a main source of bias (Bird et al., 2014). Note that going from π_i in the denominator, to $\log(\pi_i)$ as an offset, Equation (2), requires that it be a constant (non-random) for each i.

There are now many review articles for SDMs (e.g., Austin, 2007; Elith and Leathwick,

2009; Hefley and Hooten, 2016; Robinson et al., 2017; Araújo et al., 2019). Data types might be counts, presence-absence, or presence-only (Hefley and Hooten, 2016). Models can be primarily focused on estimating relationships to covariates, similar to the basic ecological idea of a niche (Soberón, 2007; Elith and Leathwick, 2009), and/or focused on prediction in space (Austin, 2002; Elith and Leathwick, 2009). There is an uneasy relationship between models formed in the covariate (niche) space and coordinate (geographic) space (Randin et al., 2006), and here we will focus purely on geographic space.

One of the most popular SDM methods is MAXENT (Phillips et al., 2006), based on maximum entropy modeling, which uses covariates in a spatially-explicit grid, and models the presence/absence of species in the grid. This is equivalent to Poisson regression (Renner and Warton, 2013), and there are further connections to spatial point processes. An inhomogeneous spatial point process that is aggregated to plots with nonzero areas leads to Equation 2 (Warton and Shepherd, 2010), where the offset can allow for unequal areas. We will stay in the grid framework. It is also possible to add a spatial random error term (Guélat and Kéry, 2018), which we will feature in our model development below.

When citizen science data consists of presence-only, or only nonzero counts, researchers have often created pseudo-absences, or zeros in the data, in an attempt to model where individuals do not occur (Pearce and Boyce, 2006; Conn et al., 2015a). A simple approach is to create zeros at random from all plots other than those with observed values (Stockwell and Peterson, 2002), but better approaches correct for sampling bias (Phillips et al., 2009; Conn et al., 2017) and include case-control methods (Fithian and Hastie, 2014) and local background sampling (Daniel et al., 2020), among others. In what follows, we will propose new ideas for creating pseudo-absences based on spatial considerations.

1.1 Objectives

Our overall objective is to develop a new approach to species distribution models based exclusively on ideas motivated by spatial autocorrelation for count data from gridded plots. Our specific objectives are to 1) use spatial count regression to develop models for effort, 2) develop spatial count regression SDMs that include effort, 3) provide novel considerations for creating zeros based on spatial autocorrelation rather than covariates, 4) outline a novel Markov Chain Monte Carlo (MCMC) approach for these data and models, 5) combine a normalized SDM with a separate species abundance estimate to provide spatially explicit abundance estimates, and 6) show several useful outcomes that can be provided by computing on the posterior distribution of the SDMs.

1.2 Motivating Example

A Platforms of Opportunity (POP) data set of marine mammal sightings, predominantly from ships, including National Oceanic and Atmospheric Administration (NOAA), U.S. Coast Guard, Navy, fishing, research, and tourist vessels, was collected from 1958 - 2016, containing 109,465 records. Data collection and quality control were described in Himes Boor and Small (2012). Each record in the POP data set was a marine mammal sighting event, which was the observation of one or more individuals of a single species, and also included date, latitude and longitude, and estimated number of animals. Sightings were contributed to the database by individuals with training and experience that ranged from professional, experienced biologists with extensive knowledge of species identification to members of the public with little or no training.

Our study area consisted of the Bering Sea and Gulf of Alaska, situated between Alaska (USA) and Siberia (Russia) (Figure 1A). Additionally, we were especially interested in marine mammal density within the dashed area in the Gulf of Alaska. This is a Density Extent Area (DEA) where, within a much smaller area called the Temporary Maritime Activities Area

(TMAA), the U.S. Navy conducts activity simulations, including acoustic signals that may be harmful to marine mammals. Environmental impacts on marine mammals are required by law. We gridded the study area into $N = 11{,}424$ hexagons, with a close-up provided in Figure 1B. One goal was to provide marine mammal density surfaces within the whole DEA for further subsetting as needed.

Within the study area, we considered northern fur seal (*Callorhinus ursinus*) and Steller sea lion (*Eumetopias jubatus*) counts from the POP data set, as each illustrated different results. Because both species are seasonally migratory, we subsetted the data to the months from May through September, which contained most of the data. These data, from 1958 - 2000, were analyzed by Himes Boor and Small (2012) with a non-spatial model, and we borrowed their idea of a ship-day. We will use our data twice; once to estimate effort, and a second time to model species densities while accounting for effort. The basic idea is to use all of the data, for all species, as a variable for effort, before analyzing any particular species. However, any particular species is obviously included in the effort variable, causing undesirable dependence between the two variables. A ship-day was defined as the presence of one or more marine mammal observations from a single ship on a single day, and ship-days were counted by hexagon (Figure 2A). We used ship-days, rather than counting the total number of animals per hexagon, in order to uncorrelate, as much as possible, the modeling of effort (ship-days) from that of any single species (Himes Boor and Small, 2012). A ship-day is a presence, rather than a count, and these are summed when a hexagon is visited multiple times. Hence, we assumed that ship-days were proportional to effort, and essentially independent of total counts of any single species. Notice that some hexagons have missing data, which may be due to lack of observed animals, or because a ship never visited the polygon, and the effort data will require spatial interpolation and smoothing, just a we will do for species-specific counts.

The numbers of ship-days (Figure 2A) shows a clear sampling bias, and it is not surprising given shipping routes and other activities by various vessels. Many records are from NOAA ships, whose mission includes ocean charting and research on fish and marine mammals, explaining the dense effort in certain areas. The total counts per hexagon for northern fur seals and Steller sea lions are shown in Figures 2B and 2C, respectively. The problem is clear when comparing Figures 2A and Figure 2B or 2C. High counts of northern fur seals and Steller sea lions might be due to high effort, and if either were not counted in a hexagon, it may be due to lack of animals, or lack of effort. However, if we assume that ship-days are proportional to effort, then we can adjust for it.

2 METHODS

We illustrate the methods using the motivating example in Section 1.2, and use terminology like "ship-days" to be concrete, but it is just a surrogate for any variable on effort. Likewise, we will use "animals" as the count variable, but these could be any count variable.

2.1 Models

Consider a spatial count regression, where the latent spatial error is multivariate normal with a covariance matrix specified by a conditional autoregressive (CAR) structure (e.g., Ver Hoef et al., 2018). Let $\mathbf{c} = (C_1, C_2, \ldots, C_N)$ be a vector of count random variables for ship-days indexed by N total hexagons in Figure 1B. We assume

$$[\mathbf{c}|\alpha_0, \mathbf{r}] = \text{Poi}(\exp(\alpha_0 + \mathbf{r})) \tag{3}$$

where α_0 is an intercept term, $\mathbf{r}$ is a vector of spatially autocorrelated random effects, and $[a|b]$ indicates a conditional distribution where random variable(s) a depends on b, and $\text{Poi}(\boldsymbol{\eta})$ is a Poisson distribution with mean vector $\boldsymbol{\eta}$. Let $[\mathbf{r}|\sigma_e^2, \rho_e] = \text{MVN}_{CAR}(\mathbf{0}, \boldsymbol{\Sigma}_e)$ be spatial random effects where MVN_{CAR} is a zero-mean multivariate normal distribution with covariance matrix $\boldsymbol{\Sigma}_e = \sigma_e^2(\mathbf{I} - \rho_e \mathbf{W}_e)^{-1}\mathbf{M}_e$ structured as a conditional autoregressive (CAR) model (Besag, 1974). The matrix $\mathbf{M}$ is diagonal, and $\mathbf{W}$ is sparse, having non-zero values only for neighboring indexes. That is, in general (for non-edge hexagons), each row of $\mathbf{W}$ has 6 non-zero values (interior hexagons in Figure 1B have 6 neighboring cells). For the non-zero values, we used row-standardization (Ver Hoef et al., 2018). By considering flat priors for $-\infty < \alpha_0 < \infty$, $0 < \sigma_e^2 < \infty$, $-\infty < \nu_e < \infty$, and $0 < \rho_e < 1$, we obtain the posterior distribution $[\mathbf{r}, \alpha_0, \rho_e, \sigma_e^2|\mathbf{c}]$. The choice of flat priors was a pragmatic one, and we had no prior information on these parameters. If there is prior information, e.g., from previously fit models or expert information, the model can be easily modified to accommodate the priors.

In practice, we use Markov chain Monte Carlo (MCMC) methods to sample from the posterior distribution $[\mathbf{r}, \alpha_0, \rho_e, \sigma_e^2|\mathbf{c}]$, which also provides a sample from the posterior distribution of $[\mathbf{e}|\mathbf{c}]$, where $\mathbf{e} = \exp(\alpha_0 + \mathbf{r})$ is the modeled effort in ship-days back on the nominal scale. More details on our MCMC methods are given below, and in the Appendix.

Now, let the count of the species of interest (in our motivating example, that is either northern fur seals or Steller sea lions) be a vector of random variables $\mathbf{y} = (Y_1, Y_2, \ldots, Y_N)$ with the same indexes as $\mathbf{c}$ from the hexagons in Figure 1B. We assume

$$[\mathbf{y}|\beta_0, \mathbf{z}, \mathbf{e}, \nu_y] = \text{NB}(\exp(\beta_0 + \log(\mathbf{e}) + \mathbf{z}), \nu_y). \tag{4}$$

Here, $[\mathbf{z}|\sigma_z^2, \rho_z] = \text{MVN}_{CAR}(\mathbf{0}, \boldsymbol{\Sigma}_z)$ is a vector of spatial random effects, Z_i, with covariance

matrix $\mathbf{\Sigma}_z = \sigma_z^2(\mathbf{I} - \rho_z \mathbf{W}_z)^{-1}\mathbf{M}_z$ structured as a CAR model. In Equation (4) $\log(\mathbf{e})$ is an offset that adjusts for effort. Note that we have adopted the recommendation of Warton et al. (2013) by modeling the observer bias first ($\mathbf{e}$), and then condition on it to model the species distribution. Thus, even though we modeled $\mathbf{e}$ as random in Equation (3), by conditioning we are treating it as a constant, as in Equation (2).

As for Equation (3), we use Markov chain Monte Carlo (MCMC) methods to sample from the posterior distribution $[\mathbf{z}, \beta_0, \rho_z, \sigma_z^2, \nu_y | \mathbf{c}, \mathbf{e}]$. Note that we use a two-stage analysis, where we first model $\mathbf{e}$, condition on it, and then use MCMC sampling to integrate over it,

$$[\mathbf{z}, \beta_0, \rho_z, \sigma_z^2, \nu_z | \mathbf{y}, \mathbf{c}] = \int [\mathbf{z}, \beta_0, \rho_z, \sigma_z^2, \nu_z | \mathbf{y}, \mathbf{e}][\mathbf{e}|\mathbf{c}] d\mathbf{e}. \tag{5}$$

We approximate Equation (5) by sampling from the posterior distribution of $[\mathbf{e}|\mathbf{c}]$ while we are sampling from the posterior of $[\mathbf{z}, \beta_0, \rho_z, \sigma_z^2, \nu_z | \mathbf{y}, \mathbf{e}]$. This idea of composition sampling within MCMC has also been used in, e.g., Banerjee et al. (2008), Hooten et al. (2010), and Babcock et al. (2015). Intuitively, this passes along all uncertainty in $\mathbf{e}$. Because we draw from $\mathbf{e}$ independently, note that we are making the assumption that $\mathbf{e}$ is independent of $\mathbf{z}$, which be believe is reasonable, or at least approximately, and it is the only way to make progress without joint modeling. Again, we considered flat priors for $-\infty < \beta_0 < \infty$, $0 < \sigma_z^2 < \infty$, $0 < \nu_z^2 < \infty$, and $0 < \rho_z < 1$.

It would also be possible to use a negative binomial distribution (or other count distribution) in Equation (3), and a Poisson distribution in (4). Figure 2B shows that counts range from 0 to 5,146 for northern fur seals, and from 0 to 11,747 for Steller sea lions, and are highly overdispersed, even with a spatial random effect. We tried both distributions, and the inclusion of ν was far from ∞ for both species, so we present the negative binomial results. At $\nu = \infty$, the first two moments are equivalent for negative binomial and Poisson (Ver Hoef and

Boveng, 2007), so we should choose the simpler Poisson model. The posterior distribution for ν was very large when modeling ship-days, so a Poisson distribution was used.

2.2 Computing on the Posterior Distribution

One of the reasons that we chose MCMC sampling is that we want samples from the full joint distribution of all parameters, which will allow us to compute functions of interest from the whole spatial surface. Because we have MCMC samples of the whole surface, we also have MCMC samples of any quantity computed on that whole surface. Hence, we easily obtain standard errors and uncertainty quantification when computing on the posterior distribution of the spatial surface.

We give two examples where we compute on the joint posterior distribution of the spatial surface. The first is to compute the total number of animals in the DEA (described in Section 1.2). Let θ_i be the true expected number of animals in hexagon i. Then

$$\theta_i^* = \frac{\theta_i}{\sum_{i=1}^N \theta_i}$$

is "standardized relative abundance" (SRA). Note that if the total abundance is known, $T = \sum \theta_i$, then $T\theta_i^* = \theta_i$ allows the recovery of each hexagon's abundance from a total abundance estimate and SRA. From Equation (4), let $\mu_i = \exp(\beta_0 + Z_i)$, which is expected counts per unit effort in the ith hexagon. A reasonable assumption is that μ_i is proportional to θ_i; $\mu_i = \phi\theta_i$. We standardize $\{\mu_i\}$,

$$\mu_i^* = \frac{\mu_i}{\sum_{i=1}^N \mu_i} = \frac{\theta_i}{\sum_{i=1}^N \theta_i} = \theta_i^*, \tag{6}$$

which allows for recovery of each hexagon's abundance if we have a separate total abundance estimate, $\theta_i = T\theta_i^* = T\mu_i^*$. The MCMC sample from Equation (5) provides a sample from the posterior distribution of $[\boldsymbol{\mu}|\mathbf{y}]$, where $\boldsymbol{\mu} = \exp(\beta_0 + \mathbf{z})$, which is the expected number of animals

per hexagon for a single ship-day, back on the nominal scale. If we have a posterior distribution for T, then sampling from posteriors of both T and μ_i^* provides a sample of the posterior for the abundance in each hexagon, and MCMC allows a sample from their joint distribution as well.

Turning to the DEA, let $\mathcal{A} = \{1, 2, ..., N\}$ be the set of indexes for all hexagons, and let $\mathcal{M} \subset \mathcal{A}$ be the set of indexes for the DEA. Then $\tau = \sum_{i \in \mathcal{M}} \theta_i$ is the total number of animals in the DEA. Let $\mu_{i,k}^*$ be the kth MCMC sample for μ_i^*. Then the kth MCMC sample for the posterior distribution of abundance within the DEA is

$$\hat{\tau}_k = T \sum_{i \in \mathcal{M}} \mu_{i,k}^*, \quad \text{or} \quad \hat{\tau}_k = T_k \sum_{i \in \mathcal{M}} \mu_{i,k}^* \tag{7}$$

depending on whether T is fixed, or T_k is the kth MCMC sample for T. Inferences, such as mean, median, mode, standard deviation, credible intervals, etc., can be obtained from $\{\hat{\tau}_k; k = 1, \ldots, K\}$ for K MCMC samples. Thus, although modeling counts with an effort offset does not yield true abundance per hexagon, it can be a key piece of information if a total abundance estimate is available.

As a second example of computing on the joint posterior distribution, consider the idea of trying to obtain "certain hotspots" of animal abundance, which accounts for areas where we are certain abundance is above average, and discounts areas where abundance estimation may be high, but highly uncertain. To help visualize areas of higher abundance, it is generally desirable to perform some amount of smoothing. Let $\mathcal{N}_i \subset \mathcal{A}$ be the set of indexes in some neighborhood of hexagon i, including i. Neighborhoods could be those hexagons within a certain radius, or a fixed number of nearest neighbors, etc. Then a smoothed value at location i is

$$s_i = \frac{\sum_{j \in \mathcal{N}_i} \mu_j}{|\mathcal{N}_i|} \tag{8}$$

where $|\mathcal{N}_i|$ is the number of neighbors. Using the kth MCMC posterior sample of $\mu_{j,k}$ for the jth hexagon, we obtain the kth MCMC sample $s_{i,k}$ for s_i. Take the mean of the MCMC samples for each i; call it $\bar{s}_i$, and let the ordered values, from smallest to largest, be denoted $\bar{s}_{(1)}, \bar{s}_{(2)}, \ldots, \bar{s}_{(N)}$. Let q be a quantile of interest, say 0.95. Then $\{\bar{s}_{([qN])}, \bar{s}_{([qN]+1)}, \ldots, \bar{s}_{(N)}\}$ are the top 5% of sites with the highest estimated abundance, where $[a]$ rounds up. Let σ_i be the standard deviation among MCMC samples, $s_{i,k}$, for each i, and let $\sigma_{(1)}, \sigma_{(2)}, \ldots, \sigma_{(N)}$ be $\{\sigma_i\}$ in the same order as $\bar{s}_{(1)}, \bar{s}_{(2)}, \ldots, \bar{s}_{(N)}$. Then

$$\left\{ \frac{\bar{s}_{([qN])}}{\sigma_{([qN])}}, \frac{\bar{s}_{([qN]+1)}}{\sigma_{([qN]+1)}}, \ldots, \frac{\bar{s}_{(N)}}{\sigma_{(N)}} \right\} \tag{9}$$

can be viewed as standard normal values, which will be large when $\bar{s}_i$ is high and σ_i is low, and, if sufficiently large, we can be certain a site has above average abundance. Cut-off values can be computed by comparing each value in Equation (9) to a quantile in the standard normal distribution. A cutoff can be proposed, and only values above the cutoff are claimed to be "certain hotspots." For example, an α-level of 0.95 yields the familiar 1.96 as a cutoff value. However, correcting for 11,424 possible comparisons, and using the conservative Bonferronni adjustment for multiple comparisons, we obtain a cutoff value of 4.59. We declare any value in Equation (9) above 4.59 to be a certain hotspot. Of course, many other options exist for creating various thresholds of interest.

2.3 MCMC Overview

We used MCMC methods to obtain samples from the posterior distribution for all parameters and latent random effects. Here, we give the broad outline of our MCMC sampling scheme, which contained some innovations. More details are given in the Appendix. The models in

Equations (3) and (4) are substantially the same, so we write the problem generically as sampling from the hierarchical model,

$$[\mathbf{y}|\beta, \mathbf{z}_o, \nu][(\mathbf{z}'_o, \mathbf{z}'_m)'|\mathbf{W}, \mathbf{M}, \sigma, \rho][\nu][\sigma][\rho]$$

where $[\mathbf{y}|\beta, \mathbf{z}_o, \nu]$ is a count model with $\mathrm{E}[\mathbf{y}] = \exp(\beta + \mathbf{z}_o)$ and possibly an extra parameter ν, $[(\mathbf{z}'_o, \mathbf{z}'_m)'|\mathbf{W}, \mathbf{M}, \sigma, \rho]$ is a multivariate normal distribution with a CAR model covariance matrix $\mathbf{\Sigma} = \sigma^2(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{M}$, and $[\nu][\sigma][\rho]$ are prior distributions. Adding effort, using its posterior distribution, was described for Equation (5), and because it does not require MCMC updating (its regression coefficient is fixed at one), we ignore it here. Note that the dimension of $\mathbf{y}$ is not the same as $\mathbf{z}$. Therefore, it will be necessary to split $\mathbf{z} = (\mathbf{z}'_o, \mathbf{z}'_m)'$ into hexagons with observed data $\mathbf{z}_o$, and missing data $\mathbf{z}_m$. Moreover, there is an issue about what are zeros, and what are missing values. Originally, $[\mathbf{y}|\beta, \mathbf{z}_o, \nu]$ has no zeros. For ship-days (effort), there are only hexagons with at least one ship-day, and for animals, there are only hexagons where at least one animal was counted (Figure 2). So one problem is where to add zeros, which we mentioned in the Introduction. We discuss our specific approaches while giving examples. Secondly, even when zeros have been added, there will still be hexagons with missing data.

Our hierarchical model leads to the following posterior distribution, $[\beta, \mathbf{z}_o, \mathbf{z}_m, \sigma, \rho, \nu|\mathbf{y}, \mathbf{W}, \mathbf{M}]$, and we discuss Metropolis-Hastings sampling from the conditional distribution for each quantity in turn.

- $[\beta|\mathbf{z}_o, \mathbf{z}_m, \sigma, \rho, \nu, \mathbf{y}, \mathbf{W}, \mathbf{M}]$ We use a Metropolis step involving the ratio $[\mathbf{y}|\beta^*, \mathbf{z}_o, \nu]/[\mathbf{y}|\beta, \mathbf{z}_o, \nu]$, where β^* is a proposal from a symmetric distribution.

- $[\nu|\beta, \mathbf{z}_o, \mathbf{z}_m, \sigma, \rho, \mathbf{y}, \mathbf{W}, \mathbf{M}]$ We use a Metropolis step involving the ratio $[\mathbf{y}|\beta, \mathbf{z}_o, \nu^*]/[\mathbf{y}|\beta, \mathbf{z}_o, \nu]$, where ν^* is a proposal from a symmetric distribution.

- $[\mathbf{z}_o|\beta, \mathbf{z}_m, \sigma, \rho, \nu, \mathbf{y}, \mathbf{W}, \mathbf{M}]$ We use a Metropolis step involving the ratio

$$\frac{[\mathbf{y}|\beta, \mathbf{z}_o^*, \nu][\mathbf{z}_o^*|\mathbf{z}_m, \mathbf{W}, \mathbf{M}, \sigma, \rho]}{[\mathbf{y}|\beta, \mathbf{z}_o, \nu][\mathbf{z}_o|\mathbf{z}_m, \mathbf{W}, \mathbf{M}, \sigma, \rho]}, \tag{10}$$

where $\mathbf{z}_o^*$ is a batch proposal by adding small, independent normal increments to the current values of $\mathbf{z}_o$. It would be more typical to sample each $z_i|\mathbf{z}_{-i}, \ldots$ one-at-a-time, where $\mathbf{z}_{-i}$ contains all the rest of $\{z_j; j \neq i\}$, as this derives directly from the conditional definition of the CAR model. Although no matrix inverses are required, this is still quite slow, looping through all 11,424 hexagons for a single MCMC sample. The evaluation of $[\mathbf{y}|\beta, \mathbf{z}_o, \nu]$ is very fast because all y_i are assumed conditionally independent. Note that, in the multivariate normal distribution for all $\mathbf{z}$, $\mathbf{z}_o$ occurs only in $\exp(-\mathbf{z}\mathbf{\Sigma}^{-1}\mathbf{z}/2)$, where $\mathbf{\Sigma}^{-1} = \mathbf{M}^{-1}(\mathbf{I} - \rho\mathbf{W})/\sigma^2$, and $\mathbf{W}$ is a sparse matrix and the only inverse required is $\mathbf{M}^{-1}$, which is diagonal, making matrix computations fast, with less storage. In the Appendix, we show how that sparse structure can be maintained, even when splitting $\mathbf{z} = (\mathbf{z}_o, \mathbf{z}_m)$. A single batch update is then very fast, and although the independent increments need to be very small for acceptance, thousands can be proposed in the time it takes for a single MCMC loop when sampling one-at-a-time. Additionally, we assumed a truncated multivariate normal distribution, rather than the usual multivariate normal distribution, and we explain why next.

- $[\mathbf{z}_m|\beta, \mathbf{z}_o, \sigma, \rho, \nu, \mathbf{y}, \mathbf{W}, \mathbf{M}]$ All $\mathbf{z}_m$ are contained only in the multivariate CAR model, so we could use Gibbs sampling one-at-a-time, directly using the definition of a CAR model. However, batch sampling with Metropolis was faster, as described above. Additionally, we noticed that the MCMC sampler for missing values, especially, was unstable when there were many missing values, lots of zeros, and high overdispersion. The reason is that the

model can always fit the zeros better by making Z_i, corresponding to an observed $y_i = 0$, more and more negative in $\exp(\beta + \mathbf{z})$, and adjusting β downward so the larger values of $\mathbf{z}$ still fit the observed $y_i > 0$. This drives the overall variance up for $\mathbf{z}$. Missing data are not anchored by observed values, and the larger variance in $\mathbf{z}$ occasionally leads to extremely large (unrealistic) values after exponentiation, for some $\exp(\beta + Z_i)$ with missing values. Similar results were found by Conn et al. (2015b) and Higham (2019). We tried a narrow prior on σ^2, but this did not solve the problem, as it was sensitive to the prior, and individual Z_i could still get very large when exponentiated. Truncating the Z-values was a simple solution. It still allowed good fits to the data, it was relatively robust to the truncation limit, and it was easy and fast to make proposals from a truncated multivariate normal distribution. Thus, ultimately, we did not use a typical CAR multivariate normal distribution, but rather one that was truncated. Truncation was easy to implement in our framework because it only involved renormalizing the distribution, and the renormalization constant canceled in the Metropolis ratio for MCMC (Equation 10) when sampling $\mathbf{z}_o$, and in the ratio $[\mathbf{z}_m^*|\mathbf{z}_o, \mathbf{W}, \mathbf{M}, \sigma, \rho]/[\mathbf{z}_m|\mathbf{z}_o, \mathbf{W}, \mathbf{M}, \sigma, \rho]$ when sampling $\mathbf{z}_m$.

- $[\rho|\beta, \mathbf{z}, \sigma, \nu, \mathbf{y}, \mathbf{W}, \mathbf{M}]$. Sampling for $\mathbf{z}$ only involves the exponent in the (truncated) multivariate normal distribution, $\exp(-\mathbf{z}'\boldsymbol{\Sigma}^{-1}\mathbf{z})$, and for CAR models, no matrix inverses are necessary. However, ρ is contained in both $\boldsymbol{\Sigma}^{-1}$ and the determinant $|\boldsymbol{\Sigma}|$ of a multivariate normal distribution. Therefore, for each MCMC iteration, we needed to compute $|\boldsymbol{\Sigma}|$, which was very time consuming for an $11,424 \times 11,424$ matrix. Instead, we pre-computed $|\boldsymbol{\Sigma}|$ for $\text{logit}(\rho) = (-40, -39, \ldots, 39, 40)/5$ in $|\sigma^2(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{M}|$. Special methods exist for determinants of sparse matrices, and because of relations between determinants and inverses, the 81 determinants only took a few minutes, and were stored as a look-up table during MCMC sampling. Thus, ρ was only sampled on the grid of values,

$\text{logit}(\rho) = (-40, -39, \ldots, 39, 40)/5$, with pre-computed determinants, and the Metropolis ratio $[\mathbf{z}|\mathbf{W}, \mathbf{M}, \sigma, \rho^*]/[\mathbf{z}|\mathbf{W}, \mathbf{M}, \sigma, \rho]$ could be rapidly evaluated with proposal ρ^*. The proposed ρ^* was chosen by sampling the 3 nearest values on either side of the current value of ρ, all with equal probability, from the set with pre-computed determinants. If the current value of ρ was near either endpoint, so that it did not have 3 values above or below it, we used Hastings sampling due to the asymmetric proposal.

- $[\sigma|\beta, \mathbf{z}, \rho, \nu, \mathbf{y}, \mathbf{W}, \mathbf{M}]$ Typically, σ could be sampled with an inverse-chi-squared distribution (Gelman et al., 2013) because it only occurs in $[\mathbf{z}|\mathbf{W}, \mathbf{M}, \rho, \sigma]$. However, due to truncation changing the normalizing constant, we used a Metropolis step involving the ratio $[\mathbf{z}|\mathbf{W}, \mathbf{M}, \rho, \sigma^*]/[\mathbf{z}|\mathbf{W}, \mathbf{M}, \rho, \sigma]$, where $\log(\sigma^*)$ had a symmetric proposal distribution.

The major innovations in our MCMC scheme were splitting $\mathbf{z} = (\mathbf{z}'_o, \mathbf{z}'_m)$ and maintaining sparse matrices without the need to compute matrix inverses when batch sampling for the multivariate normal distribution, using a truncated normal distribution, and using a look-up table for sampling ρ. These innovations allowed us to draw several million samples in just a few hours. We first used a burn-in time, and tuned the sampler so that Metropolis acceptance rates were between 0.2 and 0.5. We then used 2.5 million samples for the final run, retaining only one in every 2,500, yielding 1,000 MCMC samples from the full posterior distribution. We only kept 1,000 MCMC samples because we needed to store all 11,424 Z-values for each sample. Note that 2.5 million samples were required because batch sampling of $\mathbf{z}_o$ and $\mathbf{z}_m$ required small steps. However, these draws were very fast, and there were only two draws per iteration compared to 11,242 draws for the standard one-location-at-a-time sampling for the CAR model. Thus there was high autocorrelation for MCMC iterations, so we used many iterations that still converged relatively quickly.

We evaluated MCMC convergence using effective sample size (ESS, Flegal et al., 2008; Gong and Flegal, 2016) and MCMC standard error. The minimum ESS among all $\mathbf{z}$ was > 32, which was deemed acceptable, as 30 is an often-used criteria for sufficient sample sizes when data are independent (likely originating in Student, 1908). Most ESS for $\mathbf{z}$ were much larger than 32.

3 EXAMPLES

We illustrate our methods by continuing the motivating examples in Section 1.2.

3.1 Results of Model Fits

We first modeled ship-days (Figure 2A) using Equation (3) with MCMC sampling. Prior to MCMC sampling, we needed to add zeros to the data (Figure 3A). To add zeros, we used spatial considerations. Upon close examination, some hexagons had no animals, of any kind, but were completely surrounded by other hexagons with at least one ship-day observation. It seems highly likely that a ship passed through those empty hexagons surrounded by hexagons with > 0 ship-days , but no animals were observed, so those remain as missing values. However, there are other areas of many connected hexagons without any animal observations, and it seems likely that no ships passed through those areas, so they receive zero effort (maroon color in Figure 3A). We made the decision to buffer any hexagon with an observed ship-day with missing values, as the ship had to travel to some adjacent hexagon. However, beyond that, all original missing values were turned to zero, yielding Figure 3A.

After burnin and tuning for the MCMC sampler, the mean of the 1000 retained samples (from 2.5 million MCMC samples) for $\mathbf{e} = \exp(\alpha_0 + \mathbf{r})$ is shown in Figure 3B. The mean of the posterior distribution for ρ was greater than 0.99, showing a high amount of autocorrelation. This resulted in a fairly smooth map of effort that matches what we expect when looking at the raw

data (Figure 2A). We also point out that use of zeros and missing values had the effect of smoothing. For $\mathbf{e} = \exp(\alpha_0 + \mathbf{r})$, no values are exactly zero, reflecting the possibility that a ship traveled there but no animals were seen. Buffered hexagons left as missing, but mostly surrounded by zeros, had mean posterior values near zero. However, hexagons that had missing values, but were surrounded by hexagons with at least one ship-day, had posterior means of nearly one ship-day, or more (depending on the counts in the neighboring hexagons) in the posterior distribution (Figure 3B). In summary, our assessment is that Figure 3B is a good reflection of effort.

Next, we modeled northern fur seals (Figure 2B) with MCMC sampling, using Equation (4) with the posterior of $\mathbf{e} = \exp(\alpha_0 + \mathbf{r})$ as an offset. By sampling from the posterior distribution of $\mathbf{e}$ during MCMC sampling, we obtain the desired posterior distribution (Equation 5). Prior to MCMC sampling, we needed to add zeros to the data (Figure 2B), which required different spatial considerations than we used for ship-days. Here, we decided that if the posterior mean of ship-days (effort) was greater than one, but no northern fur seals were seen, then we would assign a zero (maroon color in Figure 4A). Otherwise, they were left as missing values. This reflects the idea that with sufficient effort, animals that were present had the possibility of being seen, but were absent. In other words, when ship-day effort was modeled as >1, then a ship was likely present in a hexagon at least once, but because a count for a species was missing, we set it to zero. We are acting as if we actually had ship tracks, and if we knew a ship entered a hexagon but did not see anything, we would record a zero. Modeled effort is simply replacing actual ship tracks.

After burnin and tuning for the MCMC sampler, the mean of the 1000 retained samples (from 2.5 million MCMC samples) for $\exp(\beta_0 + \mathbf{z})$ is shown in Figure 4B. Note that we set $\log(\mathbf{e}) = \mathbf{0}$, so Figure 4B represents the expected count per ship-day. The mean of the posterior distribution for ρ was greater than 0.99, indicating a high amount of autocorrelation. The map of

395 northern fur seal distribution shows the known distribution of northern fur seals during the summer months, May - Sept., where northern fur seals are concentrated on the Pribilof Islands in the middle of the Bering Sea during pupping (Figure 4B). Our methods have adjusted for effort. For example, the area of maximum ship-days was located at the pass in the Aleutian Islands which separates the Bering Sea from the Gulf of Alaska (Figure 3B). This resulted in quite a few 400 northern fur seal sightings near this pass as well (Figure 4A). However, after correcting for effort, this pass does not have high fur seal concentrations (Figure 4B).

As a second example, we modeled Steller sea lions (Figure 2C), using Equation (4) with the posterior of $\mathbf{e} = \exp(\alpha_0 + \mathbf{r})$ as an offset. Again, we decided that if the posterior mean of ship-days was greater than 1, but no Steller sea lions were seen, that we would assign a zero to a 405 hexagon (maroon color in Figure 5A), otherwise it was left as a missing value. After burnin and tuning for the MCMC sampler, the mean of the 1000 retained samples for $\exp(\beta_0 + \mathbf{z})$ is shown in Figure 5B, representing the expected count per ship-day. The mean of the posterior distribution for ρ was greater than 0.99, indicating a high amount of autocorrelation. The map of Steller sea lion distribution shows their known distribution during the summer months, May - Sept., which is 410 primarily along the coast of the Gulf of Alaska and out into the Aleutian Islands (Figure 5B).

It is also important to understand uncertainty about the maps presented in Figures 3 through 5. First, we present the posterior standard deviation, hexagon by hexagon, for $\mathbf{r}$ in Equation (3). This is on the log scale in relation to the data. As we might expect, Figure 6A shows standard deviations are lower where we have larger sample sizes (more ship-days) and 415 higher around the edges, which is typical for spatial models. However, when we look at the posterior standard deviation of $\exp(\alpha_0 + \mathbf{r})$ (Figure 6B), we see the more typical pattern for count distributions, where the variance is positively related to the mean. We also show the posterior standard deviation, hexagon by hexagon, for $\mathbf{z}$ in Equation (4) for northern fur seals (Figure 6C)

and Steller sea lions (Figure 6E). Again, for the random effects, standard deviations are lowest where values were counted, and highest where data were missing, and around the edges. The posterior standard deviation of the expected counts, $\exp(\beta_0 + \mathbf{z})$, for northern fur seals (Figure 6D) and Steller sea lions (Figure 6F) show the typical pattern for count distributions, where the variance is positively related to the mean. Both types of maps are useful, where the maps of R-uncertainty show more certainty with more sampling, while the mean-uncertainty show higher uncertainty with higher means.

3.2 Results from Computing on Posterior Distributions

One of our primary goals was estimating at-sea abundance, by combining an existing abundance estimate with a standardized species distribution map (Equation 7). We obtained the most current abundance estimate for northern fur seals from the stock assessment report, which is 620,660 seals (Muto et al., 2019). No standard error was presented with the estimate, so we hold it fixed. The mode of the posterior density for each grid cell within the DEA provides a spatially-explicit map for northern fur seal (Figure 7A). For each MCMC iteration, we also summed the abundance estimates for all grid cells within the DEA, providing a total estimate for the DEA. A sample from the posterior distribution of total northern fur seals in the DEA is given as a histogram in Figure 7B. Similarly, we obtained the most current abundance estimate for Steller sea lions from the stock assessment report, which is 54,267 sea lions (Muto et al., 2019); no standard error was given, so we hold it fixed. The mode of the posterior density for each grid cell within the DEA for Steller sea lion is given in Figure 7C, and a sample from the posterior distribution of total Steller sea lions in the DEA is given as a histogram in Figure 7D.

Maps, using Equation (8) by smoothing over 50 nearest neighbors, are shown for northern fur seals in Figure 8A, and for Steller sea lions in Figure 8C. Standardization of the smoothed

values (Equation 9), using the highest 10 %, and using a Bonferonni-adjusted cut-off for an α-level of 0.95, yields "certain hotspots" for northern fur seals (Figure 8B) and Steller sea lions (Figure 8D). These match our prior experience about areas known to have high abundance for both species.

4 DISCUSSION AND CONCLUSIONS

We used spatial count regression to estimate SDMs for two marine mammals in the Gulf of Alaska and Bering Sea. We created a hexagonal grid and counted the number of animals per hexagon based on presence-only data collected as shipboard observations without a pre-specified sampling design. To decrease bias, we first estimated a spatial density surface for ship-days, which was our proxy variable for effort. We created zeros for some hexagons that were far from hexagons with observed animals, and created missing values for those hexagons adjacent to hexagons with observed animals. We retained an MCMC sample of 1000 spatial surfaces from 2.5 million iterations from the posterior distribution of ship-days by using spatial Poisson regression with random effects that had a multivariate normal distribution with a CAR covariance matrix.

Next, we created SDMs for two species. Here, we created zeros for hexagons that had a mean effort of at least one ship-day and no observed animals, and any remaining hexagons with no observed animals were treated as missing values. We included the effort surface as an offset in spatial negative-binomial regression with random effects that had a multivariate normal distribution with a CAR covariance matrix, and sampled from the posterior distribution of the effort surface while retaining 1000 samples from 2.5 million MCMC iterations from the posterior distribution of the SDMs for northern fur seals and Steller sea lions.

From the posterior distributions of the SDMs, we computed two functions of interest and high importance. We normalized the SDMs so that they summed to one, and then applied an

overall abundance estimate that we obtained from the literature to derive spatially explicit abundance estimates. These were then summed in a subset of the study area, the DEA, and the MCMC samples provided a histogram reflecting the posterior distribution of total abundance in the DEA. This provided a needed estimate in an area that lacked good information until now. The data also identified what we called "certain hotspots," first by smoothing the spatial posterior distributions, dividing each hexagon by the MCMC standard deviation, and then creating thresholds. Hexagons with values above a threshold were deemed as hotspots with enough evidence to say that we are certain about them. This provides managers with ways to protect critical areas for both species.

Analysis of citizen-science data requires more assumptions and decisions than statistically designed sampling, as often occurs when crucial information is missing, so we accumulate and discuss them here.

- Hierarchical models are highly parametric. We used a Poisson or negative binomial distribution for count data, and, on the log scale, a multivariate normal distribution with a CAR covariance matrix for random effects. Both of these were chosen in part for speed and tractability of the ensuing hierarchical model. These distributions have been combined often in both space (e.g., Wakefield, 2007; Mohebbi et al., 2014) and time (e.g., Zhu, 2011; Chen et al., 2016).

- We assumed that the ship-days variable was proportional to effort, following Himes Boor and Small (2012), and others have used a similar idea in a different context (e.g., Gomes et al., 2018). If effort is dominated by, say, total counts of the most abundant species, then the modeling of effort and an SDM for that species could be highly confounded, especially for the most abundant species. The use of ship-days breaks the direct dependence of effort

on species counts. That, along with conditioning and marginalization of the SDM on effort (Equation 5) decreases, as much as possible, the confounding between effort and the SDM. In other words, we do not attempt to model the joint distribution of effort and the SDM, but rather treat them sequentially.

- We had over 55 years of data but used only spatial locations in the analysis. Hence, for method development, we assumed that the spatial distribution was constant over years. In defense of this idea, we think that most spatial data sets in ecology are actually collected over a range of time values; it would be very difficult to collect field data at all locations simultaneously, with the exception being remotely sensed data and images. However, it is important to consider the implications. Spatial distributions likely vary from year to year, and this variation is lost when data are compressed over time. Moreover, some years have more data than others, so there are unequal year effects in the data. Some of these issues can be resolved with spatio-temporal models (Cressie and Wikle, 2011), but at a large computational expense. We developed a new MCMC algorithm just to handle large spatial data, and our aim is to extend these methods to spatio-temporal models next. It is an open research problem.

- We assumed that detection of animals was spatially constant. Absolute detection rate is not an issue for these models, because at the modeling stage, we are not trying to estimate true abundance, but rather a count that is proportional to true abundance (Equations (3) and (4)). Also, any proportional constant cancels from the ratio in Equation (6), which allows for estimation of actual density with a separate estimate of total abundance. If there is information on variable detection rates, such as habitat, different observers, group size, etc., then it could be included in the model. We did not have such information, and any

variation, so long as it was not spatially-patterned, simply got passed on to random components of the model, which then increased uncertainty. Habitat might create spatially-patterned detection, but for our data, the visual detection from ships on the ocean is essentially unchanging (or it is spatially unpredictable, such as sea condition, lighting, etc.), so habitat is a minor concern.

- We used hexagons, rather than working with the actual point data. It would be more natural to use spatial point process models (Warton and Shepherd, 2010; Renner and Warton, 2013; Renner et al., 2015), but the notion of a random surface leads to log-Gaussian Cox-process models (Møller et al., 1998), and these are difficult and time-consuming to fit (Teng et al., 2017). A CAR model as a random effect on a grid is a fast approximation for the spatial point process intensity surface (Rathbun and Cressie, 1994), and still provides an MCMC sample from the posterior distribution (Besag, 1994). A sensitivity analysis could be performed (Kéry and Royle, 2016, p. 415) on hexagon size, and ultimately, the coarsest-scale grid that meets objectives should be adopted as the fastest method. Here, we wanted to show that it is reasonably fast for tens of thousands of samples, so we chose a fine-scale grid.

- We did not include any covariates. This was a major departure from most SDMs, as we reviewed in the Introduction. We wanted to highlight spatial considerations, especially when adding zeros, and focus on the prospect of creating SDMs without the need for covariates. Of course, spatial count models for regression can easily allow for covariates, in addition to spatial random effects, to include the best of both methods, and we will focus on this for future research.

- We created missing values for the effort (ship-days) data by buffering hexagons with presence-only as missing data, and any hexagons farther away were set as zero. Obviously,

other buffering rules could be applied. Again, we suggest a sensitivity analysis to examine the effect of buffering on missing values and zeros.

- We created zeros for species data by using a threshold based on the posterior distribution for ship-days, setting a hexagon to zero if it had no species counts and a mean posterior value for ship-days greater than 1. All other hexagons that had no species counts were treated as missing values. Obviously, thresholds other than 1 ship-day could be tried, and again we recommend sensitivity analysis to explore the effect of various thresholds.

- We used truncation limits of ± 6 for the normal distribution. This still allowed for a wide range of mean count values (assuming the intercept term is zero), ranging from $\exp(-6) = 0.0025$ to $\exp(6) = 403$ animals per ship-day, which easily encompassed the mean values of our two species (Figures 4B, 5B). Larger values caused problems with MCMC convergence, and smaller values had little effect until truncation limits were less than approximately ± 4, when they were too restricted to fit the data well. Truncation can be tailored to the data at hand.

As seen from above, there are challenges when using presence-only citizen-science data. In addition to modeling effort, there are consequences due to model choice, time effects, detection, spatial scale, creating zeros, and computation. Virtually all models require some assumptions, and citizen science data often require more. Nevertheless, there is information in these data, and it is also a mistake to waste information. Our goal is to make appropriate assumptions, interpret cautiously, and, if available, compare to, or combine with, other information.

A full data analysis should include model diagnostics, and we recommend them here too (Conn et al., 2018). We have focused on method development, so we omit sensitivity and model-checking. However, we did perform the most basic model checks – do the results make

sense, and are they useful? The certain hotspots (Figure 8A) for northern fur seal are centered on the Pribilof Islands, which is the major breeding center for all northern fur seals in the eastern Bering Sea, so this reflects known distribution for this species. Interestingly, a second hotspot shows up just west and south of the eastern-most Aleutian Islands, indicating a concentration possibly consisting of animals from Russian rookeries, providing interesting new information. The main certain hotspots for Steller sea lions (Figure 8B), from left to right, are the Seguam, Bogoslov, and Shelikof critical habitat foraging areas identified for Stellar sea lions (Himes Boor and Small, 2012). Two additional hotspots show up farther east, near the outside of the Prince William Sound and the open ocean side of southern southeast Alaska. Both maps confirm known concentrations, but provide further insight on species distributions. The histograms for northern fur seal and Steller sea lions (Figure 7) are both biologically reasonable according to biologists familiar with the species and area.

The methods that we have presented offer some advantages over existing methods that are used for species distribution modeling. There are now many other such methods, too many to compare individually. Nonetheless, our method focuses on spatial autocorrelation for prediction, while most others focus on covariates. Thus, our method provides an option when covariates are not easily available, and this also led us to consider novel ways to create zeros when fitting models. We developed a fast way to use exact MCMC methods for a latent CAR model, while many other methods use approximations. We also showed how to normalize any relative density surface that, when combined with an overall abundance estimate, can provide a spatial probability density surface for estimating abundance in any small area. These advantages come with some disadvantages as well. To make progress, we relied on many assumptions that were discussed above, and while our computing algorithms are fast, full MCMC still requires considerable time to fit models and store output.

The methods presented here can extend easily to other citizen science data. Our data contained counts, but Equation (3) could be a Bernoulli distribution, binomial distribution, negative binomial distribution, etc., depending on the type of citizen science data, with only a small change in MCMC sampling. Likewise, covariates could be added to Equation (3) and (4). In order create an effort surface similar to our methods, a dataset would need to be collected on many species, and the idea of a ship-day would need to be modified to some variable from the whole dataset that is a good proxy for effort. Other literature modeling species occurrence that also accounts for effort includes van Strien et al. (2013) and Dennis et al. (2017).

We have improved on the methods used in Himes Boor and Small (2012) by using spatial considerations to provide complete maps with smoothing. With some reasonable assumptions, proper models, efficient computing techniques, and a set of analytical decisions, we were able to take presence-only data, along with an overall abundance estimate for each species, and provide spatially explicit abundance estimates, with uncertainties, in a remote part of the Gulf of Alaska with no designed survey effort. This presents a novel approach to presence-only data to answer important management questions that depend on spatially explicit information. We stress that surveys designed to provide unbiased population estimates are always preferred but our approach provides important information to natural resource managers when directed scientific survey effort is unavailable.

Acknowledgments

The project received financial support from the National Marine Fisheries Service, NOAA. The findings and conclusions in the paper of the author do not necessarily represent the views of the reviewers nor the National Marine Fisheries Service, NOAA. Any use of trade, product or firm names does not imply an endorsement by the U.S. Government.

Data Availability

An R package with the all data and analyses are provided at https://github.com/jayverhoef/POP, with instructions on downloading, installing, and use. This manuscript is included in the package, with the ability to reproduce all analyses and figures. If this manuscript is accepted, Zenodo will be used as a permanent repository with a DOI.

Authors' Contributions

JVH, MH, and DJ conceived the ideas and designed methodology; RA managed and provided the data; JVH and DJ analysed the data; JVH led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

References

Araújo, M. B., Anderson, R. P., Barbosa, A. M., Beale, C. M., Dormann, C. F., Early, R., Garcia, R. A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R. B., Zimmermann, N. E., and Rahbek, C. (2019), "Standards for distribution models in biodiversity assessments," *Science Advances*, 5, eaat4858.

Austin, M. (2007), "Species distribution models and ecological theory: A critical assessment and some possible new approaches," *Ecological Modelling*, 200, 1–19.

Austin, M. P. (2002), "Spatial prediction of species distribution: an interface between ecological theory and statistical modelling," *Ecological Modelling*, 157, 101–118.

Babcock, C., Finley, A. O., Bradford, J. B., Kolka, R., Birdsey, R., and Ryan, M. G. (2015),

"LiDAR based prediction of forest biomass using hierarchical models with spatially varying coefficients," *Remote Sensing of Environment*, 169, 113–127.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008), "Gaussian predictive process models for large spatial data sets," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 825–848.

Besag, J. (1974), "Spatial interaction and the statistical analysis of lattice systems (with discussion)," *Journal of the Royal Statistical Society, Series B*, 36, 192–236.

Besag, J. E. (1994), "Discussion to the paper: Representation of knowledge in complex systems. by Grenander, U. and M.I. Miller." *Journal of the Royal Statistical Society, Series B*, 56, 549–603.

Bird, T. J., Bates, A. E., Lefcheck, J. S., Hill, N. A., Thomson, R. J., Edgar, G. J., Stuart-Smith, R. D., Wotherspoon, S., Krkosek, M., Stuart-Smith, J. F., Pecl, G. T., Barrett, N., and Frusher, S. (2014), "Statistical solutions for error and bias in global citizen science datasets," *Biological Conservation*, 173, 144–154.

Chen, C. W., So, M. K., Li, J. C., and Sriboonchitta, S. (2016), "Autoregressive conditional negative binomial model applied to over-dispersed time series of counts," *Statistical Methodology*, 31, 73–90.

Conn, P. B., Johnson, D. S., and Boveng, P. L. (2015a), "On extrapolating past the range of observed data when making statistical predictions in ecology," *PLoS One*, 10, e0141416.

Conn, P. B., Johnson, D. S., Ver Hoef, J. M., Hooten, M. B., London, J. M., and Boveng, P. L. (2015b), "Using spatiotemporal statistical models to estimate animal abundance and infer ecological dynamics from survey counts," *Ecological Monographs*, 85, 235–252.

Conn, P. B., Johnson, D. S., Williams, P. J., Melin, S. R., and Hooten, M. B. (2018), "A guide to Bayesian model checking for ecologists," *Ecological Monographs*, 88, 526–542.

Conn, P. B., Thorson, J. T., and Johnson, D. S. (2017), "Confronting preferential sampling when analysing population distributions: diagnosis and model-based triage," *Methods in Ecology and Evolution*, 8, 1535–1546.

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

Daniel, J., Horrocks, J., and Umphrey, G. J. (2020), "Efficient modelling of presence-only species data via local background sampling," *Journal of Agricultural, Biological and Environmental Statistics*, 25, 90–111.

Dennis, E. B., Morgan, B. J., Freeman, S. N., Ridout, M. S., Brereton, T. M., Fox, R., Powney, G. D., and Roy, D. B. (2017), "Efficient occupancy model-fitting for extensive citizen-science data," *PLoS ONE*, 12, e0174433.

Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, Bette A., B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M. M., Peterson, A. T., Phillips, S. J., Richardson, K., Scachetti-Pereira, R., Schapire, R. E., Soberón, J., Williams, S., Wisz, M. S., and Zimmermann, N. E. (2006), "Novel methods improve prediction of species' distributions from occurrence data," *Ecography*, 29, 129–151.

Elith, J. and Leathwick, J. R. (2009), "Species distribution models: Ecological explanation and prediction across space and time," *Annual Review of Ecology, Evolution, and Systematics*, 40, 677–697.

Fithian, W. and Hastie, T. (2014), "Local case-control sampling: Efficient subsampling in imbalanced data sets," *Annals of Statistics*, 42, 1693–1724.

Flegal, J. M., Haran, M., and Jones, G. L. (2008), "Markov chain Monte Carlo: Can we trust the third significant figure?" *Statistical Science*, 23, 250–260.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013), *Bayesian Data Analysis*, CRC Press, Boca Raton, FL, USA.

Gomes, V. H., IJff, S. D., Raes, N., Amaral, I. L., Salomão, R. P., de Souza Coelho, L., de Almeida Matos, F. D., Castilho, C. V., de Andrade Lima Filho, D., and López, D. C. (2018), "Species Distribution Modelling: Contrasting presence-only models with plot abundance data," *Scientific Reports*, 8, 1–12.

Gong, L. and Flegal, J. M. (2016), "A practical sequential stopping rule for high-dimensional Markov chain Monte Carlo," *Journal of Computational and Graphical Statistics*, 25, 684–700.

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I., Regan, T. J., Brotons, L., McDonald-Madden, E., and Mantyka-Pringle, C. (2013), "Predicting species distributions for conservation decisions," *Ecology Letters*, 16, 1424–1435.

Guélat, J. and Kéry, M. (2018), "Effects of spatial autocorrelation and imperfect detection on species distribution models," *Methods in Ecology and Evolution*, 9, 1614–1625.

Hefley, T. J. and Hooten, M. B. (2016), "Hierarchical species distribution models," *Current Landscape Ecology Reports*, 1, 87–97.

Higham, M. (2019), "Spatial Prediction for Finite Populations with Ecological Applications," Ph.D. thesis, Oregon State University, Corvallis, OR.

Himes Boor, G. K. and Small, R. J. (2012), "Steller sea lion spatial-use patterns derived from a Bayesian model of opportunistic observations," *Marine Mammal Science*, 28, E375–E403.

Hooten, M. B., Johnson, D. S., Hanks, E. M., and Lowry, J. H. (2010), "Agent-based inference for animal movement and selection," *Journal of Agricultural, Biological and Environmental Statistics*, 15, 523–538.

Horvitz, D. G. and Thompson, D. J. (1952), "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, 47, 663–685.

Kelling, S., Johnston, A., Bonn, A., Fink, D., Ruiz-Gutierrez, V., Bonney, R., Fernandez, M., Hochachka, W. M., Julliard, R., Kraemer, R., and Guralnick, R. (2019), "Using semistructured surveys to improve citizen science data for monitoring biodiversity," *BioScience*, 69, 170–179.

Krebs, C. (1972), *Ecology: The Experimental Analysis of Distribution and Analysis*, New York, NY: Harper and Row.

Kéry, M. and Royle, J. A. (2016), *Applied Hierarchical Modeling in Ecology: Analysis of Distribution, Abundance and Species Richness in R and BUGS*, Amsterdam, The Netherlands: Academic Press.

Lukyanenko, R., Wiggins, A., and Rosser, H. K. (2020), "Citizen science: An information quality research frontier," *Information Systems Frontiers*, 22, 961–983.

Miller-Rushing, A., Primack, R., and Bonney, R. (2012), "The history of public participation in ecological research," *Frontiers in Ecology and the Environment*, 10, 285–290.

Mohebbi, M., Wolfe, R., and Forbes, A. (2014), "Disease mapping and regression with count data in the presence of overdispersion and spatial autocorrelation: A Bayesian model averaging approach," *International Journal of Environmental Research and Public Health*, 11, 883–902.

Muto, M. M., Helker, V. T., Angliss, R. P., Boveng, P. L., Breiwick, J. M., Cameron, M. F., Clapham, P., Dahle, S. P., Dahlheim, M. E., Fadely, B. S., Ferguson, M. C., Fritz, L. W., Hobbs, R. C., Ivashchenko, Y. V., Kennedy, A. S., London, J. M., Mizroch, S. A., Ream, R. R., Richmond, E. L., Shelden, K. E. W., Sweeney, K. L., Towell, R. G., Wade, P. R., Waite, J. M., and Zerbini, A. N. (2019), "Alaska Marine Mammal Stock Assessments, 2018," Tech. Rep. NOAA Technical Memorandum NMFS-AFSC-393, U.S. Department of Commerce.

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998), "Log Gaussian Cox processes," *Scandinavian Journal of Statistics*, 25, 451–482.

Pearce, J. L. and Boyce, M. S. (2006), "Modelling distribution and abundance with presence-only data," *Journal of Applied Ecology*, 43, 405–412.

Phillips, S. J., Anderson, R. P., and Schapire, R. E. (2006), "Maximum entropy modeling of species geographic distributions," *Ecological Modelling*, 190, 231–259.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., and Ferrier, S. (2009), "Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data," *Ecological Applications*, 19, 181–197.

Randin, C. F., Dirnböck, T., Dullinger, S., Zimmermann, N. E., Zappa, M., and Guisan, A. (2006), "Are niche-based species distribution models transferable in space?" *Journal of Biogeography*, 33, 1689–1703.

Rathbun, S. L. and Cressie, N. (1994), "A space-time survival point process for a longleaf pine forest in southern Georgia," *Journal of the American Statistical Association*, 89, 1164–1174.

Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., Popovic, G., and

Warton, D. I. (2015), "Point process models for presence-only analysis," *Methods in Ecology and Evolution*, 6, 366–379.

Renner, I. W. and Warton, D. I. (2013), "Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology," *Biometrics*, 69, 274–281.

Robinson, N. M., Nelson, W. A., Costello, M. J., Sutherland, J. E., and Lundquist, C. J. (2017), "A systematic review of marine-based species distribution models (SDMS) with recommendations for best practice," *Frontiers in Marine Science*, 4, 421 doi:10.3389/fmars.2017.00421.

Soberón, J. (2007), "Grinnellian and Eltonian niches and geographic distributions of species," *Ecology Letters*, 10, 1115–1123.

Soroye, P., Ahmed, N., and Kerr, J. T. (2018), "Opportunistic citizen science data transform understanding of species distributions, phenology, and diversity gradients for global change research," *Global Change Biology*, 24, 5281–5291.

Stockwell, D. R. B. and Peterson, A. T. (2002), "Controlling bias in biodiversity data," in *Predicting Species Occurrences: Issues of Scale and Accuracy*, (eds. J.M. Scott, P.J. Heglund, M.L. Morrison, J.B. Haufler, M.G. Raphael, W.A. Wall & F.B. Samson), Island Press, Washington, DC, pp. 537–546.

Student (1908), "Probable error of a correlation coefficient," *Biometrika*, 6, 302–310.

Teng, M., Nathoo, F., and Johnson, T. D. (2017), "Bayesian computation for log-Gaussian Cox processes: A comparative analysis of methods," *Journal of Statistical Computation and Simulation*, 87, 2227–2252.

van Strien, A. J., van Swaay, C. A., and Termaat, T. (2013), "Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models," *Journal of Applied Ecology*, 50, 1450–1458.

Ver Hoef, J. M. and Boveng, P. L. (2007), "Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data?" *Ecology*, 88, 2766–2772.

Ver Hoef, J. M., Peterson, E. E., Hooten, M. B., Hanks, E. M., and Fortin, M.-J. (2018), "Spatial autoregressive models for statistical inference from ecological data," *Ecological Monographs*, 88, 36–59.

Wakefield, J. (2007), "Disease mapping and spatial regression with count data," *Biostatistics*, 8, 158–183.

Warton, D. I., Renner, I. W., and Ramp, D. (2013), "Model-based control of observer bias for the analysis of presence-only data in ecology," *PLOS ONE*, 8, e79168.

Warton, D. I. and Shepherd, L. C. (2010), "Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology," *The Annals of Applied Statistics*, 4, 1383–1402.

Zhu, F. (2011), "A negative binomial integer-valued GARCH model," *Journal of Time Series Analysis*, 32, 54–67.

FIGURES

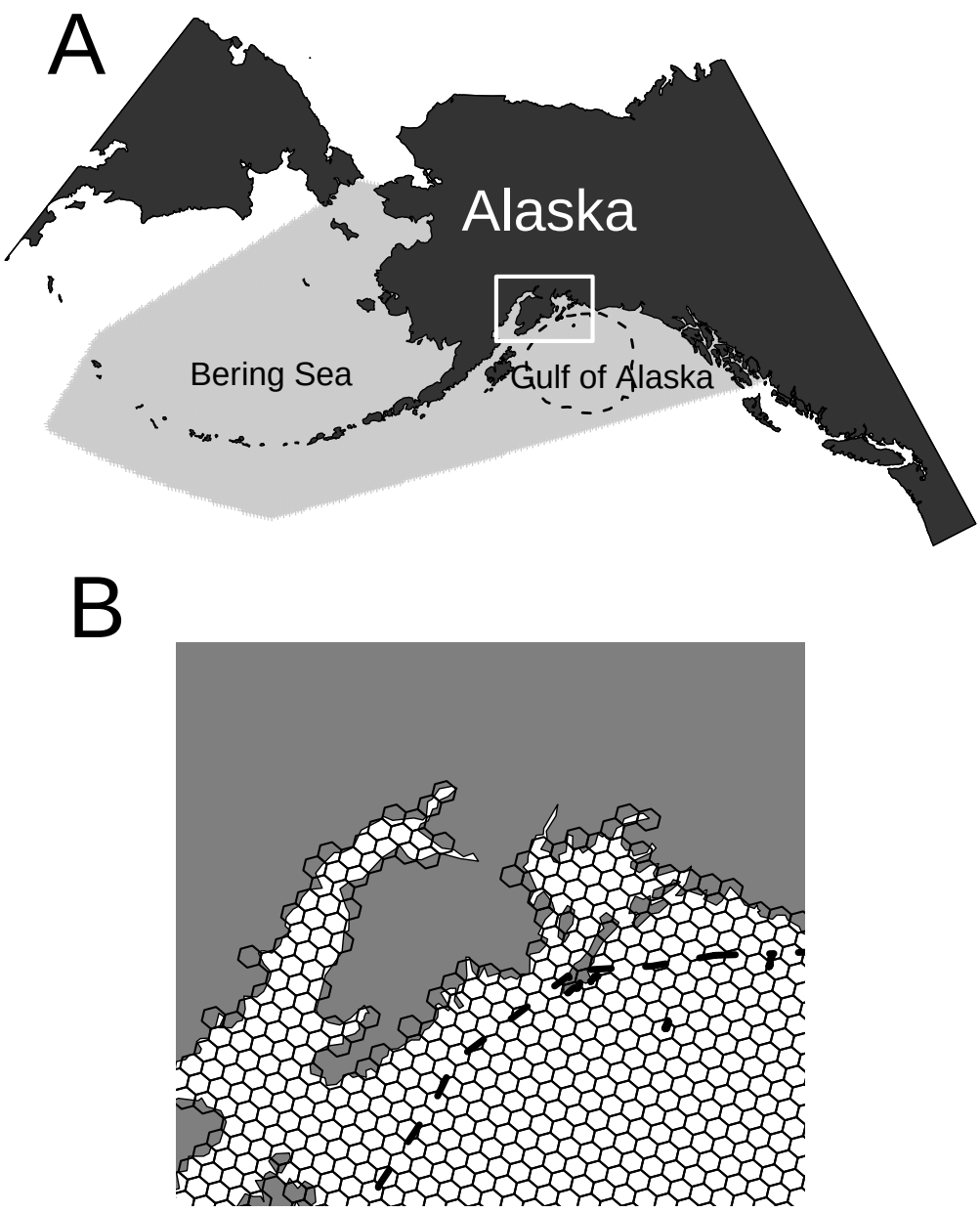

Figure 1: Study area. A. Data were taken from the Bering Sea and Gulf of Alaska, shown by the light gray shade. The study area was gridded with 11,424 hexagons, but resolution is insufficient to plot them all. The white rectangular inset allows for more detail. The polygon with a dashed black line in the Gulf of Alaska is a Density Extent Area (DEA) used by the U.S. Navy. B. A close-up of the white rectangular inset, showing hexagonal sample units, each of which was approximately 289 km^2.

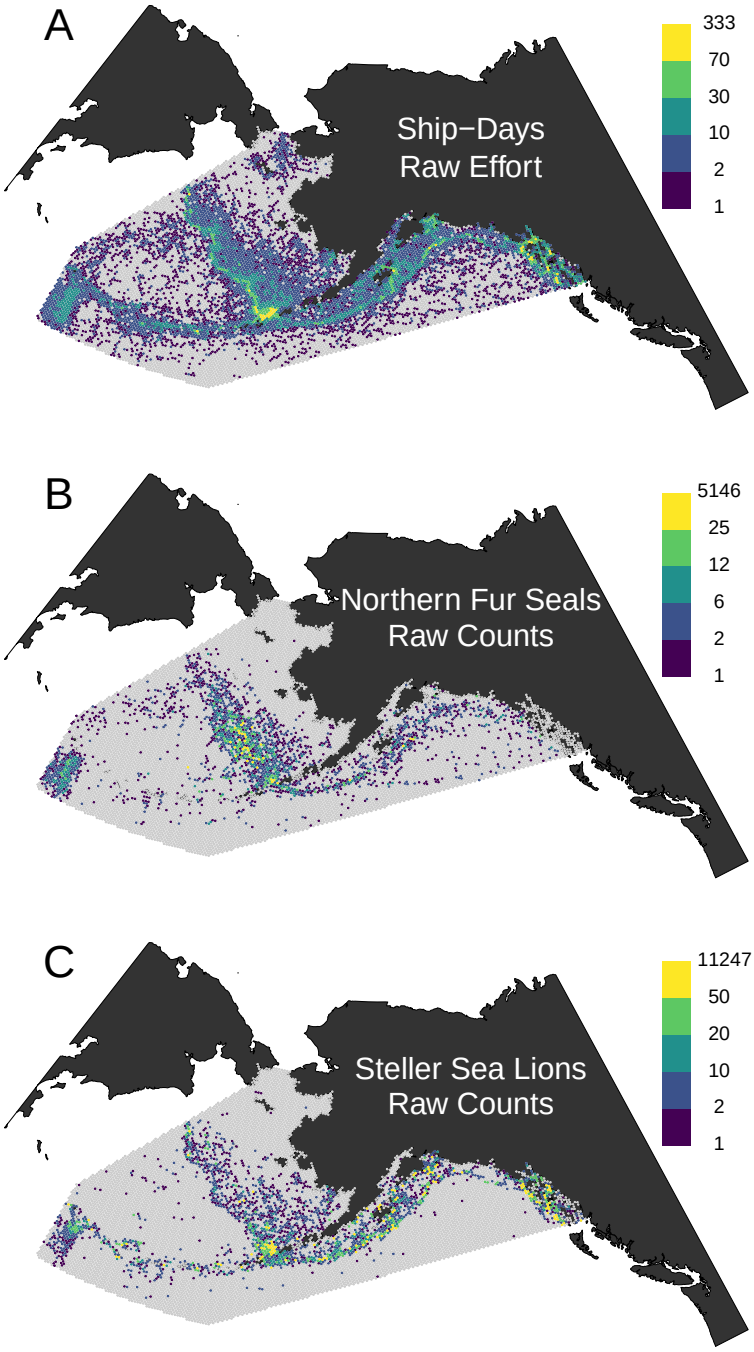

Figure 2: Raw data used for analyses. A. Ship-days in the study area. Hexagons without any ship days are contained in gray background. B. Northern fur seal counts in each hexagon, where zero counts are part of the gray background. C. Steller sea lion counts in each hexagon.

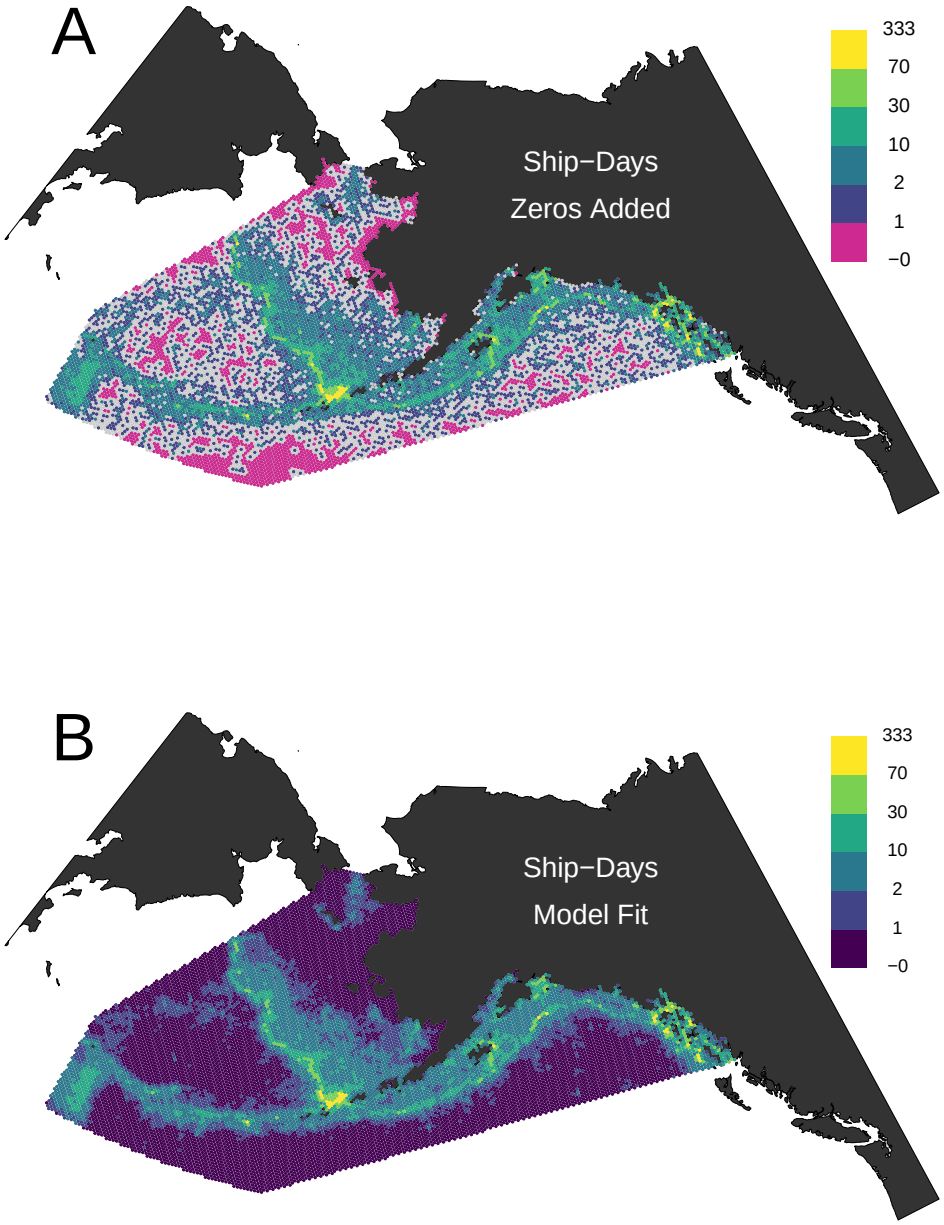

Figure 3: Effort data and model. A. Ship-days in the study area, where observed counts are the same as Figure 2A, except structural zeros have been added. Hexagons with a gray background were treated as missing data. B. Mode of the posterior distribution for ship-day for each hexagon using a negative binomial regression model with spatial random effects.

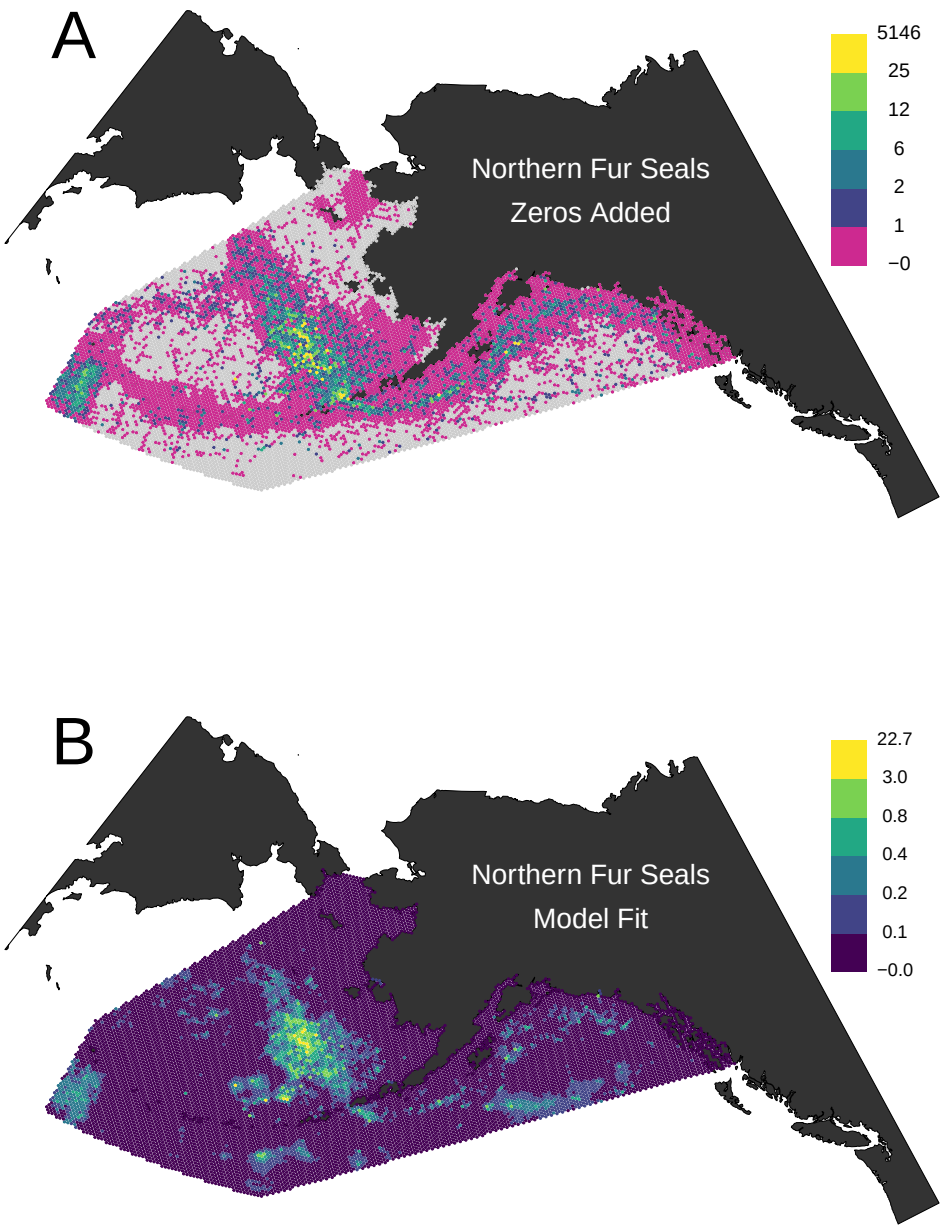

Figure 4: Northern fur seal data and model. A. Northern fur seal counts in the study area, where observed counts are the same as Figure 2B, except structural zeros have been added. Hexagons with a gray background were treated as missing data. B. Mode of the posterior distribution for northern fur seals per ship-day for each hexagon using a negative binomial regression model with spatial random effects.

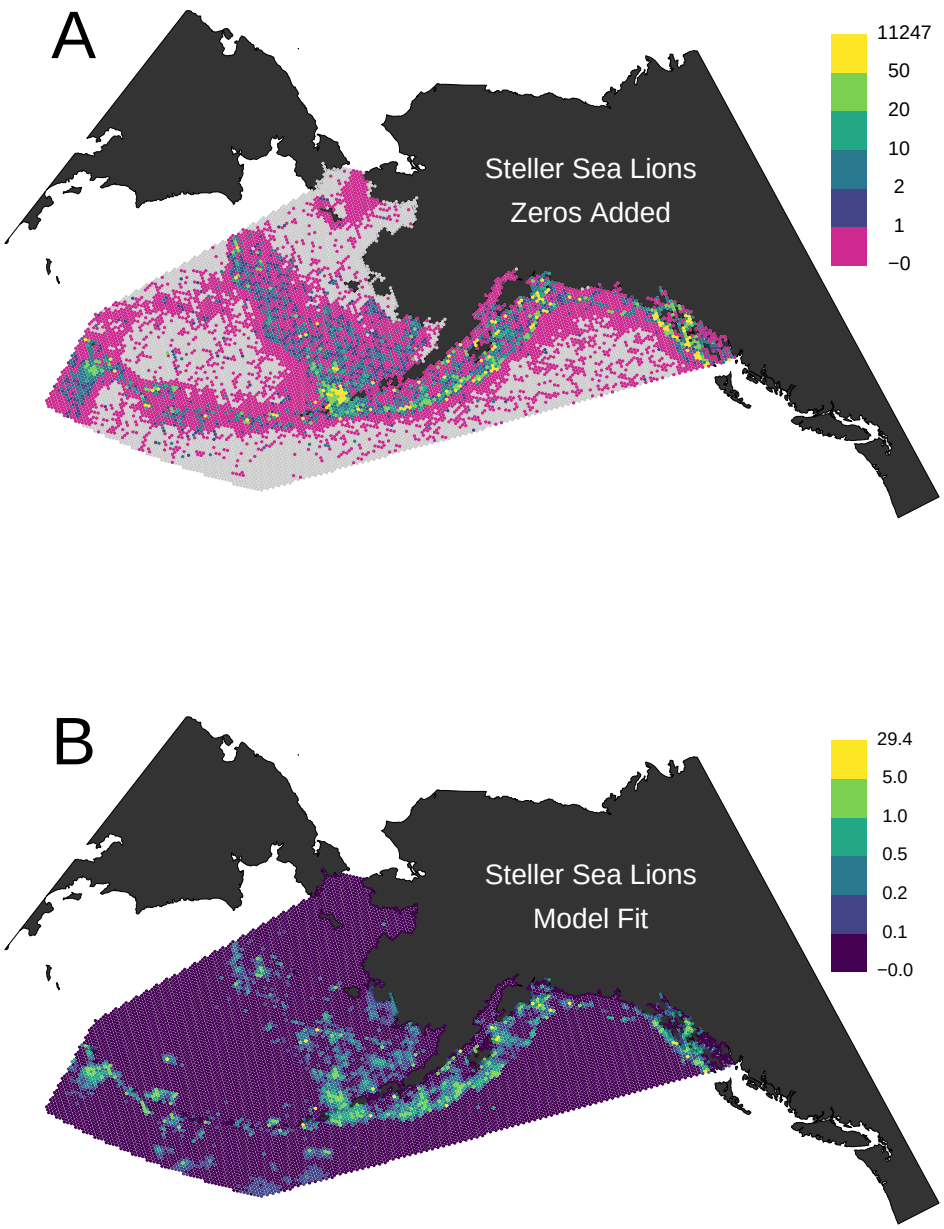

Figure 5: Steller sea lion data and model. A. Steller sea lion counts in the study area, where observed counts are the same as Figure 2C, except structural zeros have been added. Hexagons with a gray background were treated as missing data. B. Mode of the posterior distribution for Steller sea lions per ship-day for each hexagon using a negative binomial regression model with spatial random effects.

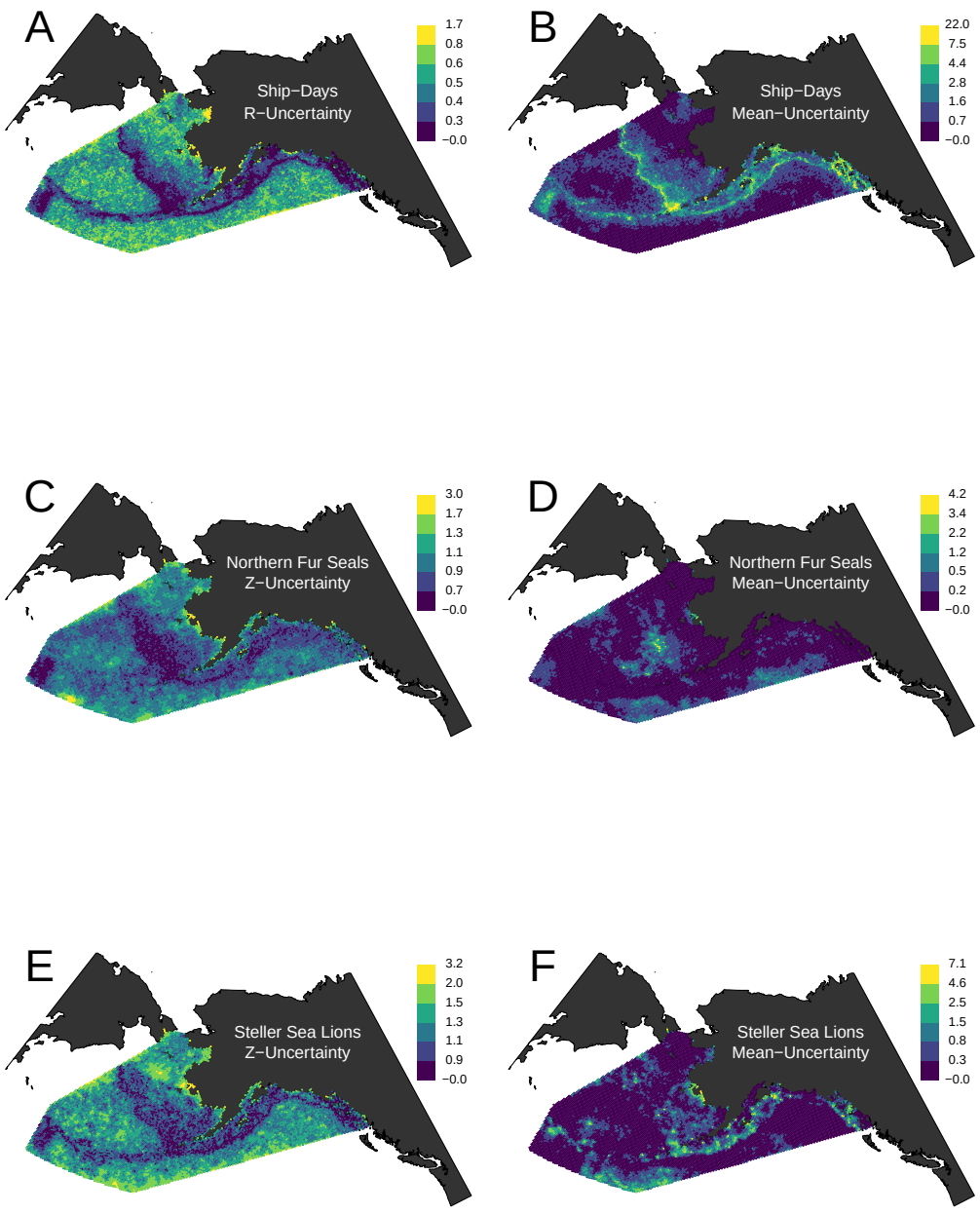

Figure 6: Model uncertainty. Standard deviation of each hexagon's posterior distribution for: A) R_i in $\mathbf{r}$ from Eq. (3), B) $\exp(\alpha_0 + R_i)$ from Eq. (3), C) Z_i in $\mathbf{z}$ from Eq. (4) for northern fur seals, D) $\exp(\beta_0 + Z_i)$ from Eq. (4) for northern fur seals, E) Z_i in $\mathbf{z}$ from Eq. (4) for Steller sea lions, D) $\exp(\beta_0 + Z_i)$ from Eq. (4) for Steller sea lions.

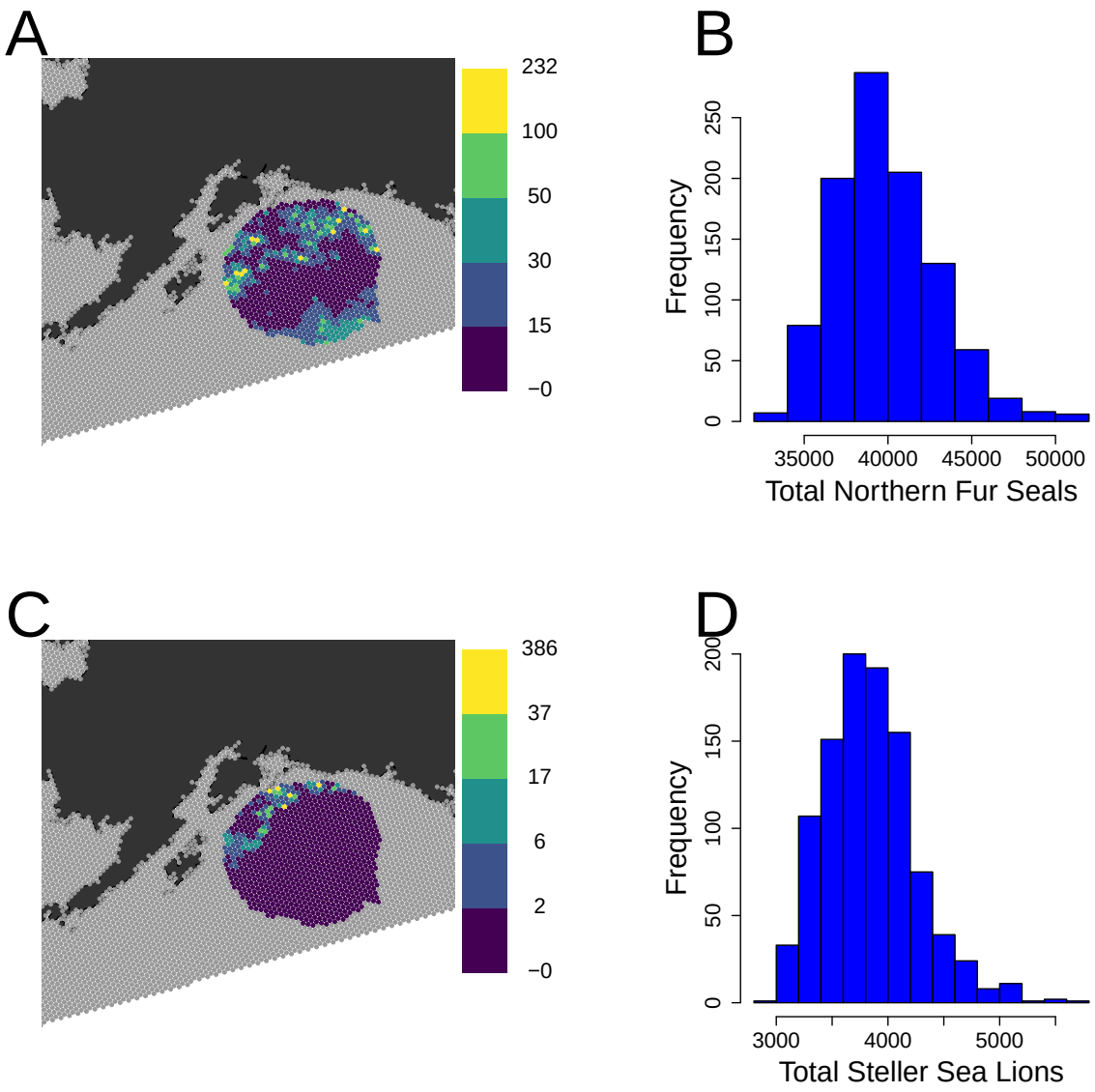

Figure 7: Population estimates per hexagon in the Gulf of Alaska Density Extent Area (DEA). A. Map of mode of the posterior distribution for northern fur seal abundance in DEA. B. Histogram of MCMC sample from posterior distribution for total abundance of northern fur seals in DEA. C. Map of mode of the posterior distribution for Steller sea lions in DEA. D. Histogram of MCMC sample from posterior distribution for total abundance of Steller sea lions in DEA.

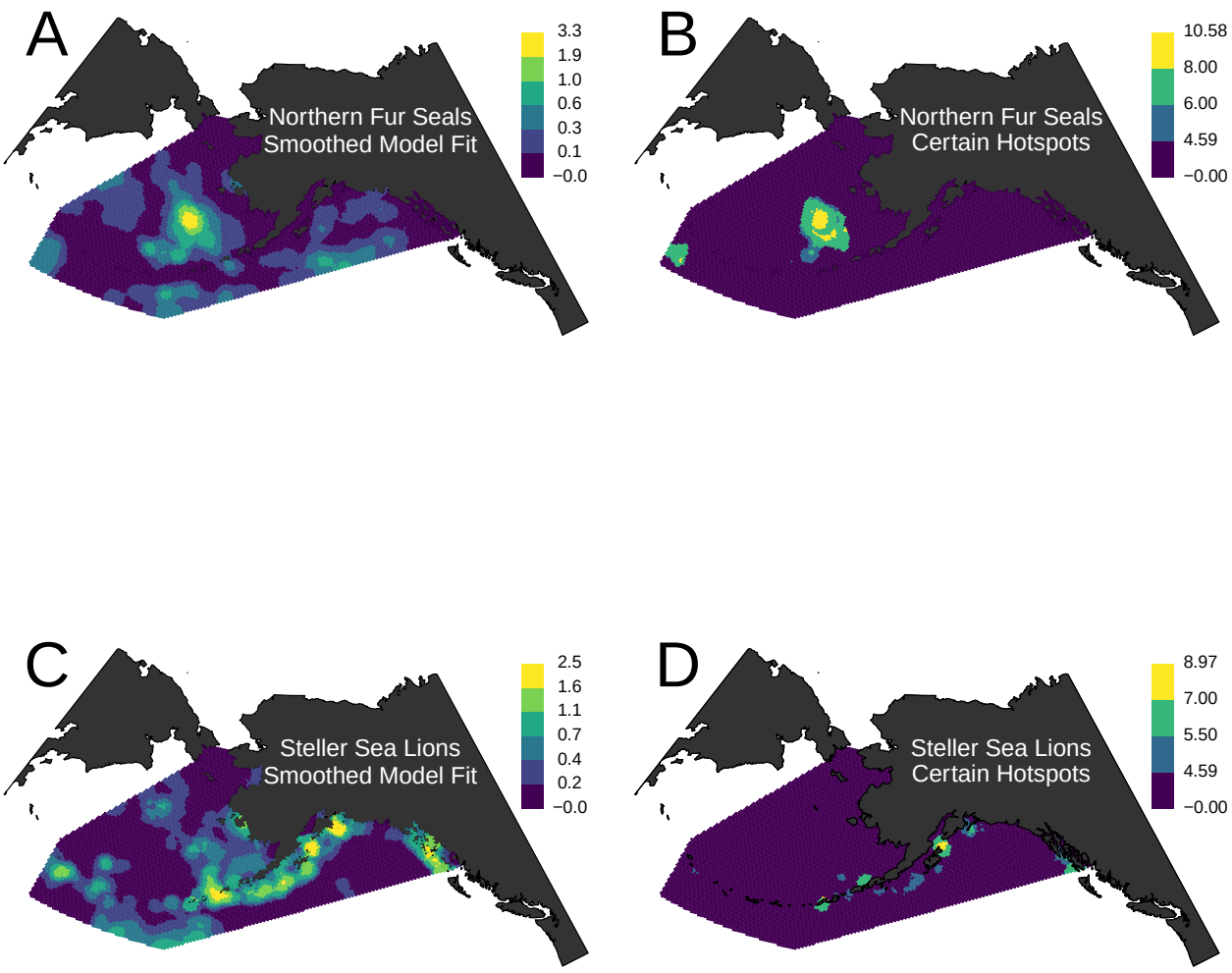

Figure 8: Smoothing and Hotspots. A. Mode of each hexagon's posterior distribution after smoothing by averaging over 50 nearest neighbors for northern fur seals. B. Certain hotspots created by combining smoothing, thresholds, and uncertainty for northern fur seals. C. Mode of each hexagon's posterior distribution after smoothing by averaging over 50 nearest neighbors for Steller sea lions. D. Certain hotspots created by combining smoothing, thresholds, and uncertainty for Steller sea lions.

APPENDIX: Details on MCMC Sampling Methods

4.1 Updating Spatial Random Effects Sequentially

It would be most natural to set up the Metropolis-Hastings (MH) to sample spatial random effects sequentially. First consider sites with observed counts. The MH acceptance probability for z_i at a sampled location i is

$$\min\left(1, \frac{[y_i|\beta_0, z_i^*, \nu][\mathbf{y}_{-i}|\beta_0, \mathbf{z}_{-i}, \nu][z_i^*|\mathbf{z}_{-i}, \sigma, \rho][\mathbf{z}_{-i}|\sigma, \rho][\beta_0][\sigma][\rho]q(z_i|z_i^*)}{[y_i|\beta_0, z_i, \nu][\mathbf{y}_{-i}|\beta_0, \mathbf{z}_{-i}, \nu][z_i|\mathbf{z}_{-i}, \sigma, \rho][\mathbf{z}_{-i}|\sigma, \rho][\beta_0][\sigma][\rho]q(z_i^*|z_i)}\right),$$

where z_i^* is the proposed new value of z_i, $\mathbf{z}_{-i}$ is the vector $\mathbf{z}$ minus the ith element, and $q(z_i|z_i^*)$ is the proposal distribution, which, if symmetric, can be eliminated from the acceptance probability, and recall that $[\mathbf{y}|\beta_0, \mathbf{z}, \nu] = [y_i|\beta_0, z_i, \nu][\mathbf{y}_{-i}|\beta_0, \mathbf{z}_{-i}, \nu]$ because of conditional independence. Of course, many parts of the above equation cancel, and, assuming a symmetric proposal distribution, yields the acceptance ratio

$$\min\left(1, \frac{[y_i|\beta_0, z_i^*, \nu][z_i^*|\mathbf{z}_{-i}, \sigma, \rho]}{[y_i|\beta_0, z_i, \nu][z_i|\mathbf{z}_{-i}, \sigma, \rho]}\right).$$

Due to the special nature of the way we use the CAR model with row-standardization, $[z_i^*|\mathbf{z}_{-i}, \sigma, \rho] = \mathrm{N}(\rho\bar{z}_i, \sigma^2/n_i)$, where $\bar{z}_i$ is the mean of z-values that are neighbors of site i, and n_i is the number of neighbors. This comes directly from the definition of a CAR model. No inversion of matrices are required for one-at-a-time updating of z's at sites with observed counts. For z's at sites with missing count values, we can do Gibbs sampling straight from the definition. Again, this is simple, and the only limitation is that looping one-at-a-time might be slow, especially in an interpreted language like R.

4.2 Block Updating for Spatial Random Effects

We used block updating, which was faster than sequential sampling. We split the sites into those with observed counts $\mathbf{z}_o$, and those with missing counts $\mathbf{z}_m$. First consider sampling for $\mathbf{z}_o$. The MH acceptance probability is very similar to above,

$$\min\left(1, \frac{[\mathbf{y}|\beta_0, \mathbf{z}_o^*, \nu][\mathbf{z}_o^*|\mathbf{z}_m, \sigma, \rho][\mathbf{z}_m|\sigma, \rho][\beta_0][\sigma][\rho]q(\mathbf{z}_o|\mathbf{z}_o^*)}{[\mathbf{y}|\beta_0, \mathbf{z}_o, \nu][\mathbf{z}_o|\mathbf{z}_m, \sigma, \rho][\mathbf{z}_m|\sigma, \rho][\beta_0][\sigma][\rho]q(\mathbf{z}_o^*|\mathbf{z}_o)}\right), \tag{A.1}$$

where $\mathbf{z}_o^*$ is a proposal for sites with observed counts, $[\mathbf{y}|\beta_0, \mathbf{z}_o, \nu]$ is simply the product of the count distribution model for all sites with counts, and $q(\mathbf{z}_o|\mathbf{z}_o^*)$ is the proposal distribution. Assuming the proposal distribution is symmetric, simplifying the MH acceptance probability becomes

$$\min\left(1, \frac{[\mathbf{y}|\beta_0, \mathbf{z}_o^*, \nu][\mathbf{z}_o^*|\mathbf{z}_m, \sigma, \rho]}{[\mathbf{y}|\beta_0, \mathbf{z}_o, \nu][\mathbf{z}_o|\mathbf{z}_m, \sigma, \rho]}\right). \tag{A.2}$$

It turns out that can we easily and rapidly draw an MH sample using Equation (A.2).

Note that

$$\frac{[\mathbf{y}|\beta_0, \mathbf{z}_o^*, \nu][\mathbf{z}_o^*|\mathbf{z}_m, \sigma, \rho]}{[\mathbf{y}|\beta_0, \mathbf{z}_o, \nu][\mathbf{z}_o|\mathbf{z}_m, \sigma, \rho]} = \frac{[\mathbf{y}|\beta_0, \mathbf{z}_o^*, \nu][\mathbf{z}_o^*, \mathbf{z}_m|\sigma, \rho]}{[\mathbf{y}|\beta_0, \mathbf{z}_o, \nu][\mathbf{z}_o, \mathbf{z}_m|\sigma, \rho]}.$$

So, for a block proposal, $\mathbf{z}_o^*$, we can form a MH probability acceptance ratio as

$$\min\left(1, \frac{[\mathbf{y}|\beta_0, \mathbf{z}_o^*] \exp\left(-\frac{1}{2}(\mathbf{z}_m, \mathbf{z}_o^*)\boldsymbol{\Sigma}^{-1}\begin{pmatrix}\mathbf{z}_m \\ \mathbf{z}_o^*\end{pmatrix}\right)}{[\mathbf{y}|\beta_0, \mathbf{z}_o] \exp\left(-\frac{1}{2}(\mathbf{z}_m, \mathbf{z}_o)\boldsymbol{\Sigma}^{-1}\begin{pmatrix}\mathbf{z}_m \\ \mathbf{z}_o\end{pmatrix}\right)}\right). \tag{A.3}$$

Recall that $\boldsymbol{\Sigma}^{-1}$ is given directly to us in the CAR model (that is, this is what we are actually modeling), and it is sparse. Equation (A.3) contains a simple, fast matrix multiplication because

we have the full, sparse $\mathbf{\Sigma}^{-1}$, and allows for a block update of all $\mathbf{z}_o$. By the same logic, but without the count distribution, we can do block updates of $\mathbf{z}_m$ as well,

$$\min\left(1, \frac{\exp\left(-\frac{1}{2}(\mathbf{z}_m^*, \mathbf{z}_o)\mathbf{\Sigma}^{-1}\begin{pmatrix}\mathbf{z}_m^* \\ \mathbf{z}_o\end{pmatrix}\right)}{\exp\left(-\frac{1}{2}(\mathbf{z}_m, \mathbf{z}_o)\mathbf{\Sigma}^{-1}\begin{pmatrix}\mathbf{z}_m \\ \mathbf{z}_o\end{pmatrix}\right)}\right) \tag{A.4}$$

where $\mathbf{z}_m^*$ is the proposal for $\mathbf{z}_m$.

4.3 The Problem of Many Zeros, Missing Data, and High Dispersion

We ran into a problem because there were many zeros, plus some very high counts, and missing values were unconstrained by observed counts. The $\mathbf{z}_o$ for sites with observed counts will naturally be sampled so that

$$\exp(\beta_0 + z_i)$$

will be near the observed $y_i > 0$. However, for $y_i = 0$, the "fit" will get better and better as z_i goes toward negative infinity. This drives up the variance of the latent CAR model while the intercept goes down, and when we predict z_i at sites without missing counts, we can obtain large values. When they are exponentiated, they become even larger, which leads to unrealistically large values on the exponentiated scale. We considered several solutions, including constraining the CAR variance, but that had limited success because the individual z_i still became very large. Instead we treated $\mathbf{z}$ as a truncated normal distribution to keep any value from becoming too large. So long as the truncation was not too limiting, it still allowed z_i to fit the 0's very well. The truncation helped to stabilize both the mean and CAR variance parameter as well. A

850

truncated multivariate normal distribution simply changes the normalizing constant, which cancels in Equations (A.3) and (A.4), so we only needed to make proposals that stayed within the truncation limits. We used uniform proposal distributions for each z_i, where the median of the uniform distribution was centered on z_i. If z_i was near a truncation value, so that the uniform distribution exceeded the truncation limit, then the uniform distribution was simply adjusted so that its endpoint was the truncation limit. The proposal distribution was symmetric, because both the current values and proposed values have the same probability density from the uniform distribution.

4.4 Using Sparse Matrices for Determinants, and Creating a Lookup Table

In the above sections, we showed how to sample $\mathbf{z}$ quickly using block updates. Because there are only 2 samplings, one for sites with observed counts, and one for sites with missing counts, it scales linearly with sample size, as only sparse matrix multiplications are involved. Updates for β_0 and σ proceed as usual. They are very fast because no inverses or determinants are necessary. However, ρ presents a new challenge because it is involved in the determinant $|\mathbf{R}_\rho|$,

$$\log[\rho|\ldots] \propto (1/2)\log(|\mathbf{R}_\rho^{-1}|) - \mathbf{z}'\mathbf{R}_\rho^{-1}\mathbf{z}/(2\sigma^2),$$

where $\mathbf{R}_\rho^{-1} = \mathbf{M}^{-1}(\mathbf{I} - \rho\mathbf{W})$ is sparse. Fortunately, determinants for sparse matrices are quite fast. For example, using the sparse matrix representation for a 11,424 $\times$ 11,424 CAR model, we were able to do 81 determinants in several minutes. These determinants can then be used as a lookup table when doing MCMC sampling. That is, rather than sampling from ρ continuously, we sample from a grid of values. The MCMC chain will mix over the discrete values, so very little is lost. The determinant is known from the lookup table, and we may as well store the sparse matrices

875

$\mathbf{R}_\rho^{-1} = \mathbf{M}^{-1}(\mathbf{I} - \rho\mathbf{W})$ for each ρ value as well, so they won't need to be computed during MCMC.

There are some issues with creating the proposals. Our strategy was to take the nearest 3 indexes on each side of the index of the current ρ value. For example, we used a grid with 81 ρ values, ordered from lowest to smallest. They were evenly spaced on the logit scale, so values were
880 compressed near 0 and 1 on the expit scale. Say the index for the current ρ value is 41. Then, we choose a proposal from indexes 38 to 40, and 42 to 44, with a discrete uniform probability of 1/6. As long as the current index is not near the ends, the proposal distribution will be symmetric. However, say the current value has index 81. Then, the proposal will come from indexes 78:80, with a discrete uniform probability of 1/3 for chosing an index. Suppose 78 is chosen. Then a
885 proposal from index 78 would come from indexes 75 to 77 and 79 to 81, with a discrete uniform probability of 1/6 for chosing index 81. Hence, in general, we used Hastings and included the proposal probabilities $q(\rho|\rho^*)$ and $q(\rho^*|\rho)$,

$$\min\left(1, \frac{(|\mathbf{R}_{\rho^*}^{-1}|)^{1/2}\exp[-\mathbf{y}'\mathbf{R}_{\rho^*}^{-1}\mathbf{y}/(2\sigma^2)]q(\rho|\rho^*)}{(|\mathbf{R}_{\rho}^{-1}|)^{1/2}\exp[-\mathbf{y}'\mathbf{R}_{\rho}^{-1}\mathbf{y}/(2\sigma^2)]q(\rho^*|\rho)}\right)$$

890

4.5 An Overall Sampling Strategy

When block samping $\mathbf{z}_o$ and $\mathbf{z}_m$, it seems unlikely that, when each vector has 1000's of elements, we can create autocorrelation by independently drawing each of the elements of $\mathbf{z}_o$ and $\mathbf{z}_m$. This does require literally millions of draws. So, a good strategy is to draw $\mathbf{y}_o$ and $\mathbf{y}_m$ many times
895 before advancing to other parameters. Both of these draws are very fast, simply requiring the proposal, and a sparse matrix multiplication. We sampled both $\mathbf{y}_o$ and $\mathbf{y}_m$ twenty times before advancing to other parameters.