Resolving misaligned spatial data with integrated distribution models

Krishna Pacifici, Brian J. Reich, David A.W. Miller, and Brent Pease

Appendix S1

Models for spatial intensity function

Below we discuss three models and describe their balance between modeling flexibility and ease of evaluating or approximating the integral in (Eq 1 in main document) . Arguably the most flexible model is to allow $\lambda(\mathbf{s})$ to be a log-Gaussian process (Gelfand et al., 2010). This poses a computation challenge because (Eq 1 in main document) becomes a stochastic integral. One remedy is to approximate the integral by summing over the process evaluated at many locations in $\mathcal{B}$. However, this requires evaluation of a Gaussian model at a large number of spatial locations, which is a notorious computational bottleneck (Gelfand et al., 2010).

Another approach (Gelfand et al., 2010) is to assume that the continuous process is a function of L known basis function $Z_1(\mathbf{s}), ..., Z_L(\mathbf{s})$, so that $\log[\lambda(\mathbf{s})] = \sum_{l=1}^{L} Z_l(\mathbf{s})\beta_l$. For example, the basis functions could be spatial covariates such as land use or climate variables, or a flexible basis expansion such as spline, wavelet or Fourier functions (Hefley et al., 2017). The integral is then

$$\tilde{\lambda}(\mathcal{B}) = \int_{\mathcal{B}} \exp\left[\sum_{l=1}^{L} Z_l(\mathbf{s})\beta_l\right] d\mathbf{s}, \qquad \text{(Eq. S1)}$$

which can be approximated with numerical integration. Increasing L gives an arbitrarily flexible model, but at the expense of increasing the number of parameters to be estimated (i.e., the L basis coefficients β_l) and computational burden of the numerical integration.

A third approach is to specify the model for a fine grid of regions $\mathcal{B}_1, ..., \mathcal{B}_n$ that partition the spatial domain of interest, and assume the Poisson intensity is roughly constant within each grid

cell. That is, to approximate $\lambda(\mathbf{s}) = \lambda_i/A_i$ for all $\mathbf{s} \in \mathcal{B}_i$, where A_i is the area of $\mathcal{B}_i$ so that $\lambda_i = \tilde{\lambda}(\mathcal{B}_i)$ is the integrated intensity in region i. Under this model the aggregated intensity for an arbitrary region $\mathcal{G}$ is

$$\tilde{\lambda}(\mathcal{G}) = \sum_{i=1}^{n} w_i(\mathcal{G})\lambda_i \qquad \text{(Eq. S2)}$$

where $w_i(\mathcal{G}) = G_i/A_i$ and G_i is the area of intersection between $\mathcal{G}$ and $\mathcal{B}_i$. If the analysis includes spatial covariates then these too must be roughly constant in the fine grid, giving the log-linear model

$$\log(\lambda_i) = \mathbf{X}_i^T\boldsymbol{\beta} + \theta_i,$$

where $\mathbf{X}_i$ is vector of covariates in region i and θ_i is the spatial residual. Of course, this approach is only reasonable if the fine grid of n regions is sufficiently resolved to capture the important features of the process.

For example, Figure S1a plots simulated λ_i (gray scale) on a fine grid of n regions $\mathcal{B}_1, ..., \mathcal{B}_n$ (rectangles separated by dashed lines) that partition the spatial domain. In this hypothetical example, the data are defined on a coarse-resolution grid denoted $\mathcal{G}_1, ..., \mathcal{G}_m$ with $m < n$, and each coarse resolution grid cell is the combination of several fine resolution grid cells. Let $\mathcal{S}_j$ be the indices of the fine-resolution cells that comprise $\mathcal{G}_j$, i.e.,

$$\mathcal{G}_j = \bigcup_{i \in \mathcal{S}_j} \mathcal{B}_i.$$

The Poisson intensity for $\mathcal{G}_j$ is then

$$\tilde{\lambda}(\mathcal{G}_j) = \int_{\mathcal{G}_j} \lambda(\mathbf{s}) d\mathbf{s} = \sum_{i \in \mathcal{S}_j} \lambda_i.$$

Each coarse grid cell $\mathcal{G}_1, ..., \mathcal{G}_m$ (rectangles separated by bold solid lines) contains 9 fine scale grid cells and Figure S1b shows the aggregated Poisson intensity for this illustrative example.

References

Gelfand, A. E., Diggle, P., Guttorp, P. and Fuentes, M. (2010) *Handbook of Spatial Statistics*. CRC press.

Hefley, T. J., Broms, K. M., Brost, B. M., Buderman, F. E., Kay, S. L., Scharf, H. R., Tipton, J. R., Williams, P. J. and Hooten, M. B. (2017) The basis function approach for modeling autocorrelation in ecological data. *Ecology*, **98**, 632–646.

Figure S1: **Hypothetical spatial resolutions**: The plot on the left shows a simulated dataset shaded according to its value at fine resolution grid cells. The plot on the right shades the coarse resolution cells shaded by the average of the fine resolution cells within each larger grid cell.

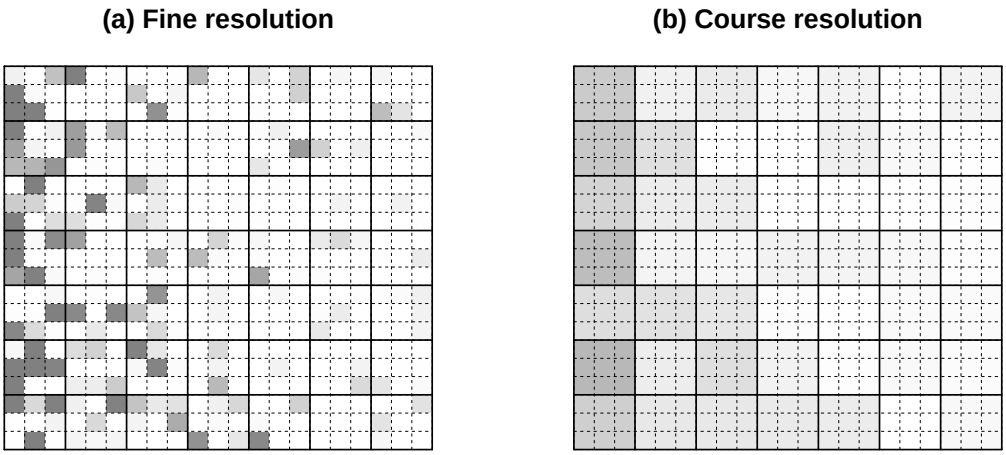

Figure S2: **An example of the types of spatial misalignment from combining two sources of data (Y_1 and Y_2)**: Examples include (clockwise starting in upper left) integrating point level data with areal data to make predictions to point level data, integrating point level data with areal data to make predictions to areal data, integrating point level data with areal data to make predictions to areal data at different spatial scales, and integrating point level data with areal data to make predictions to point level data at different spatial scales

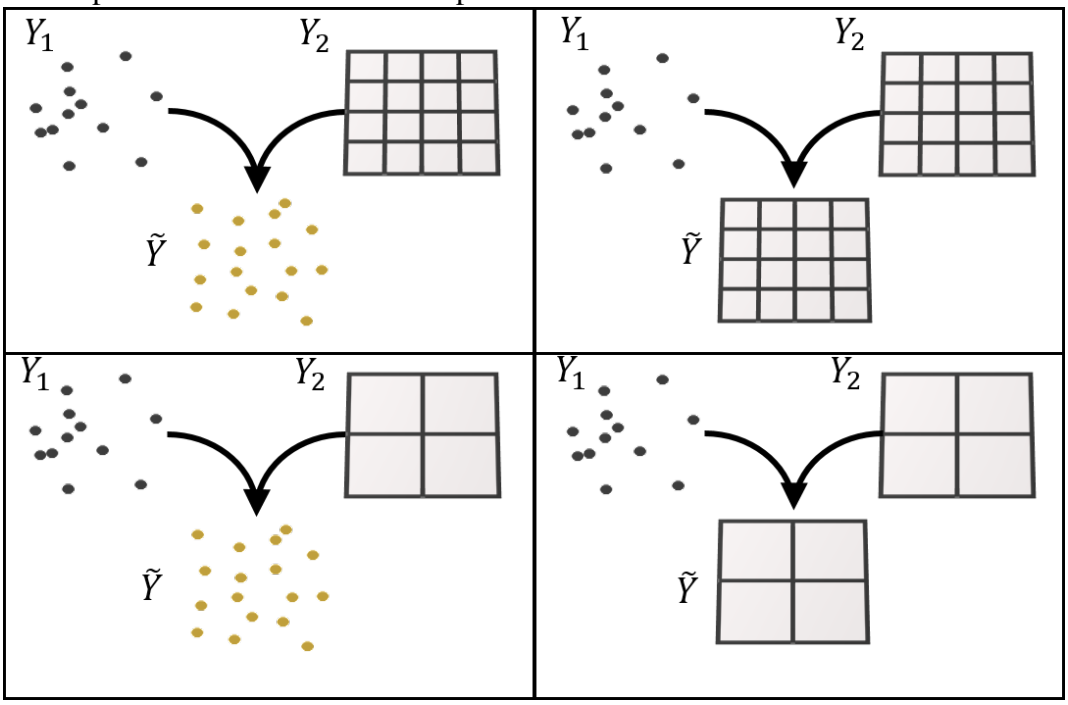

Figure S3: **Plot of one simulated example**: Panel (a) maps the latent occupancy indicator Z_i for each fine-resolution cell; Panel (b) shows the fine-resolution data $Y_{1i} \in \{0, 1, ..., 5\}$; Panel (c) shows the aggregated data source Y_{2j} for each 4×4 coarse-resolution grid cell; Panel (d) shows the second data source but presuming each observation corresponds to one fine-resolution grid cell rather than a sum over grid cells. An example of this is the equivalent of having a data set where observations are accounted to the county level with no knowledge of where the data was collected within the county, but are assigned to occur at the centroid of the county.

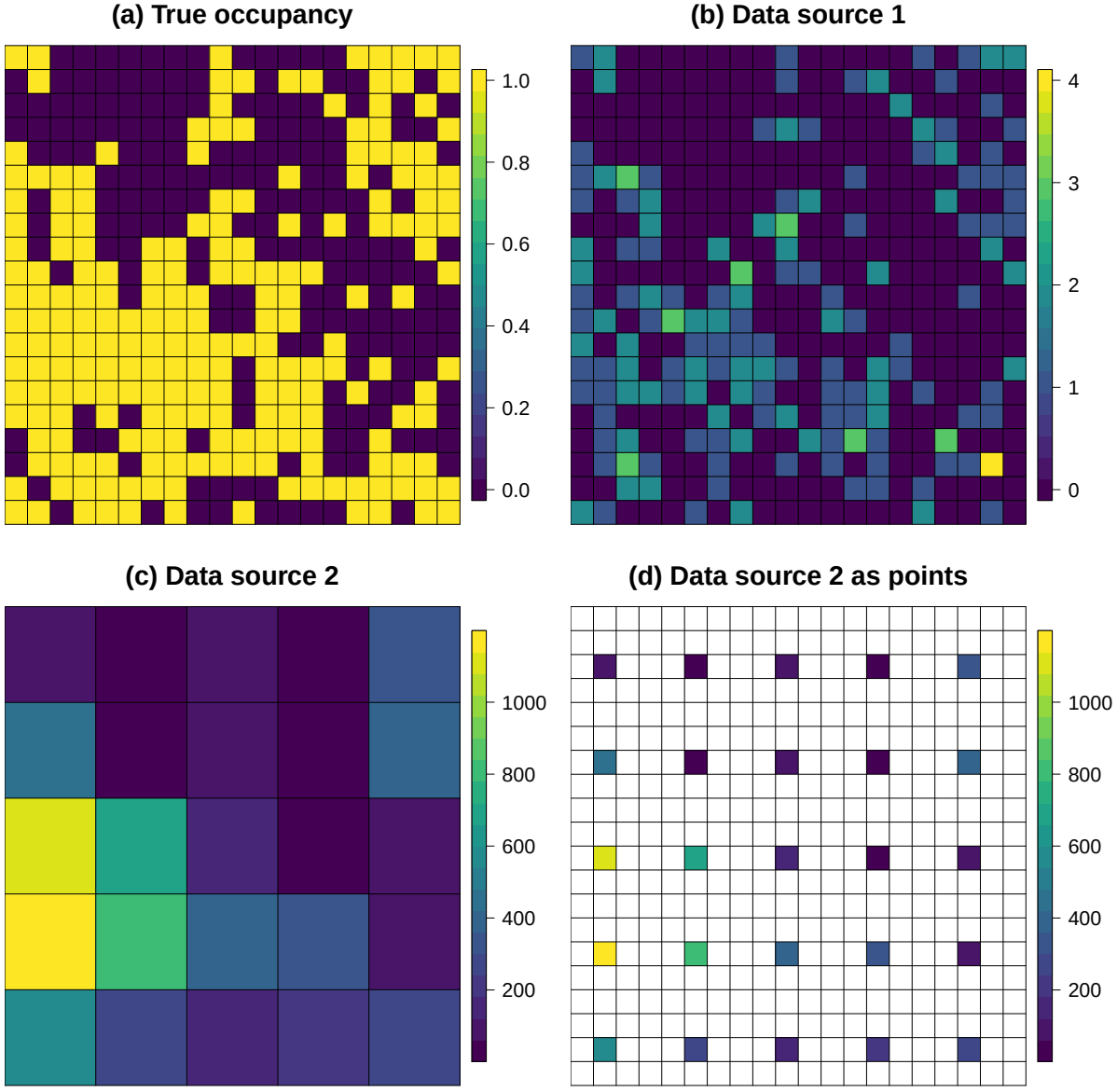

Figure S4: **Occupancy estimates from full data set, covariate model, shared no cos, and shared at grid level 1**: The plot in the upper left panel shows the distribution of occupancy across sites for the entire data set. The plot in the upper right panel shows the results from the covariate model, shared no cos, and shared models.

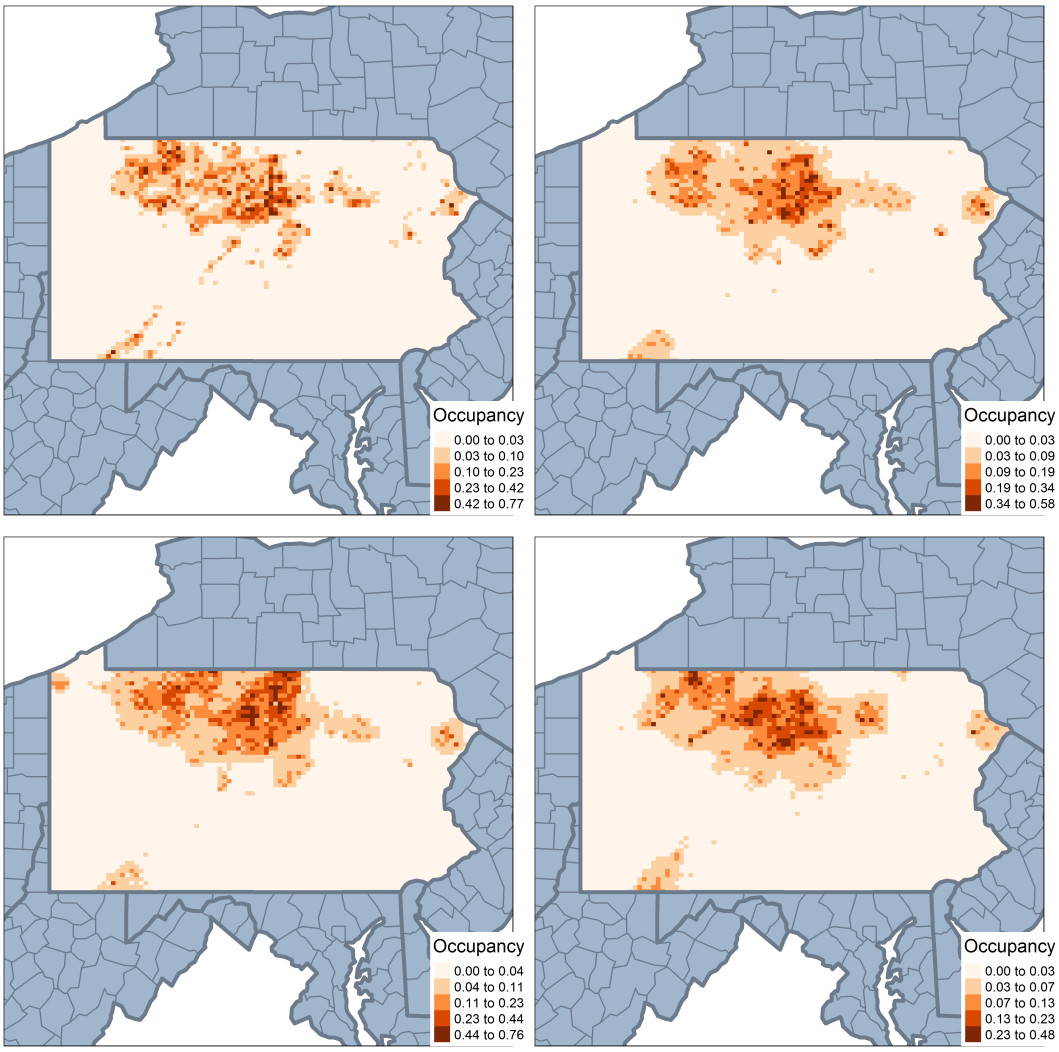

Figure S5: **Occupancy estimates from covariate model with covariates summarized in two different ways at grid level 4**: The plot in the upper left panel shows the results from the covariate model with the second data source averaged across all smaller sites, with uncertainty on upper right panel. The lower left panel shows results from the same covariate model with the covariates summarized differently.