Example data set and analysis.

To illustrate concepts discussed in the paper, we provide an example of an integrated model and associated code. Specifically, we focus on illustrating correlation and covariate approaches for combining data, the use of spatial random-effects in models, and the explicit estimation and incorporation of effort into models using non-standardized data.

We choose two data sets where a joint-likelihood model is difficult to fit. Thus, other methods that share information while not requiring a shared set of parameters are ideal. The data are for the black-throated blue warbler collected from 2005-2009 in Pennsylvania, USA. Rather than using data for the whole state, we subset the data to create a data set for which models can be fit more quickly, allowing for more efficient exploration and potential modification by the reader. The dataset includes all observations between 40 and 41.5 degrees of latitude and -78 and -76.5 degrees of longitude. This region includes wide variation in relative abundance for the species.

The first data set, which was collected under a standardized data collection protocol are point count surveys conducted as part of the Pennsylvania Breeding Bird Atlas. More than 34,000 point counts were completed over the 5-year study period, of which 5,165 occur within the domain used here. Point counts occurred during peak breeding period each year and were finished within 4 hours of sunrise each morning. Counts lasted 375 seconds and were divided into 5 equal 75-second intervals. During each interval, observers recorded whether a black-throated blue warbler was heard within 150-m of the point count location. The observation, Y_i, is the number of intervals for which the warbler was heard at site i.

The second data set did not follow a standard data collection protocol, but instead were observations reported by citizen scientists to eBird. We limited observations to the breeding season. In total, 146 black-throated warblers were recorded in 1606 lists reported to eBird during the study period. Uncertainty in spatial locations of reports occurs because of how locations are reported and because observers may travel while collecting observations. To minimize spatial error, we aggregated observations into grid cells that were 1/16 degree of longitude by 1/24 degree of latitude. For each cell we total the number of birds counted in the j^{th} cell, W_j, as well as three measures of effort: the total number of lists reported *eLists*, the total number of km traveled *eKm*, and the number of hours spent observing *eHrs*.

In this example, we chose to our state variable to the probability a point count location was occupied by the warbler for our standardized data set. For the non-standardized data, we build models where either the intensity of observations within a grid cell or the probability the cell is occupied was modeled when fitting models based on a correlation structure. Alternatively, when using the covariate approach, we used the number of birds counted standardized by effort as a covariate. For all models we include two covariates, the proportion of forest cover and the average elevation measured at the larger grid cell scale.

As stated above, we always fit the first data set using an occupancy model where the proportion of point count locations occupied by the warbler was the estimated state variable. In addition, all models include a CAR spatial random effect fit to spatial variation at the scale of the individual

grid cells. This was integrated with the eBird data using three different model structures. These were:

1. We fit a correlation model where the state variable for the eBird data was the intensity of observations, scaled to account for effort. Effort was treated as a random variable to be estimated as a function of the number of lists, distance traveled, and hours spent observing.
2. Again, we used a correlation approach, but eBird data was now estimated where occurrence at the grid cell level was the response variable and effort again was treated as a random variable.
3. eBird data is treated as a covariate used to predict the model used for the point count data. We scale eBird counts by the number of lists as an ad hoc approach to account for effort.

All models can be fit using WinBUGS called using R. In the following section we describe the model and give examples of the bugs code. Appendix B includes the complete set of data and code needed to fit the models.

Model 1.

The hierarchical model for the data is as follows. For point count observations we assume

$$Y_i \sim \text{Binomial}(z_i * p, 5).$$

The binomial probability is a function of whether the site was occupied, z_i, and the probability of detecting the warbler during any interval if the site was occupied, p. The observation model for eBird follows a count model structure, where:

$$W_j \sim \text{Poisson}(\lambda_j * E_j).$$

The observation intensity per unit of effort is given by λ_j while E_j is the estimated effort for grid cell j. Effort is estimated as:

$$E_j = eLists_j + \beta_1 * eKm_j + \beta_2 * eHrs_j$$

where β_1 and β_2 are constrained to be positive.

We join the two state variables through a correlation structure that also account for spatial autocorrelation. We do this using a multivariate conditional autoregressive model (MVCAR). For each grid cell a spatial random effect is estimated for each of our state variables (θ_{j1} and θ_{j2}). The MVCAR model allows the spatial random effects to be correlated, thus allowing for information transfer between the two data sets up to the degree to which patterns are correlated.

Finally, a linear model is used to link the spatial random-effects and covariates for forest cover and elevation to each of the state variables.

$$z_i \sim \text{Bernouli}(\psi_i)$$

$$\text{logit}(\psi_i) = \alpha_{01} + \alpha_1 * forest + \alpha_2 * elevation + \theta_{j1}$$

$$\log(\lambda_j) = \alpha_{02} + \theta_{j2}$$

The bugs code for the model, including priors, is as follows:

```
model {
# MV CAR prior for the spatial random effects
# MVCAR prior
    S[1:2, 1:ncell] ~ mv.car(adj[], weights[], num[], omega[ , ])
# Other priors
    for (k in 1:2) {
       alpha[k] ~ dflat()
    }
# Precision matrix of MVCAR
    omega[1:2, 1:2] ~ dwish(R[ , ], 2)
# Covariance matrix of MVCAR
    sigma2[1:2, 1:2] <- inverse(omega[ , ])
# conditional SD
    sigma[1] <- sqrt(sigma2[1, 1])
    sigma[2] <- sqrt(sigma2[2, 2])
# within-area conditional correlation
    corr <- sigma2[1, 2] / (sigma[1] * sigma[2])
### BBA model
### priors
    p ~ dunif(0,1)
    a.forest ~ dnorm(0,0.01)
    a.elev ~ dnorm(0,0.01)
    b.effort[1] ~ dunif(0,10)
    b.effort[2] ~ dunif(0,10)
### data model
    for (i in 1:nsite){
       z[i] ~ dbern(psi[cell[i]])
       muy[i] <- z[i]*p
       Y[i] ~ dbin(muy[i],5)
    }
### eBird Model
### Data Model
    for (j in 1:ncell){
       logit(psi[j]) <- S[1,j] + alpha[1] + a.forest*forest[j] + a.elev*elev[j]
       log(lambda[j]) <- S[2,j] + alpha[2]
       E[j] <- effort[j,1] + b.effort[1] *effort[j,2] + b.effort[2] *effort[j,3]
       muP[j] <- E[j]*lambda[j]
       W[j] ~ dpois(muP[j])
    }
}
```

Model 2

The second model is like the first, with the only change being to the state variable estimated using the non-standard data. Now instead of intensity, we estimate the probability a grid cell is occupied. Let W_j now be whether any warblers were observed in the grid cell, where

$$W_j \sim \text{Bernouli}(P^*_j)$$

$$P^*_j = \delta_j * (1-(1-p)\text{^}E_j)$$

$$\text{logit}(\delta_j) = \alpha_{02} + \theta_{j2}$$

The model code is as follows:

```
model {
# MV CAR prior for the spatial random effects
# MVCAR prior
    S[1:2, 1:ncell] ~ mv.car(adj[], weights[], num[], omega[ , ])
# Other priors
    for (k in 1:2) {
       alpha[k] ~ dflat()
    }
# Precision matrix of MVCAR
    omega[1:2, 1:2] ~ dwish(R[ , ], 2)
# Covariance matrix of MVCAR
    sigma2[1:2, 1:2] <- inverse(omega[ , ])
# conditional SD
    sigma[1] <- sqrt(sigma2[1, 1])
    sigma[2] <- sqrt(sigma2[2, 2])
# within-area conditional correlation
    corr <- sigma2[1, 2] / (sigma[1] * sigma[2])
### BBA model
### priors
    p[1] ~ dunif(0,1)
    p[2] ~ dunif(0,1)
    a.forest ~ dnorm(0,0.01)
    a.elev ~ dnorm(0,0.01)
    b.effort[1] ~ dunif(0,1000)
    b.effort[2] ~ dunif(0,1000)
### data model
    for (i in 1:nsite){
       z[i] ~ dbern(psi[cell[i]])
       muy[i] <- z[i]*p[1]
       Y[i] ~ dbin(muy[i],5)
    }
### eBird Model
### Data Model
    for (j in 1:ncell){
       logit(psi[j]) <- S[1,j] + alpha[1] + a.forest*forest[j] + a.elev*elev[j]
       logit(delta[j]) <- S[2,j] + alpha[2]
       E[j] <- effort[j,1] + b.effort[1] *effort[j,2] + b.effort[2] *effort[j,3]
       muP[j] <- delta[j] * (1 - pow((1-p[2]),E[j]))
       W[j] ~ dbern(muP[j])
    }
}
```

Model 3

The third model uses the same data sets, but rather than building a separate distribution model using eBird data, the eBird data is used as an explanatory covariate for the point count data set. As in the other models we assume a binomial for the point count data, where

$$Y_i \sim \text{Binomial}(z_i * p, 5).$$

The binomial probability is a function of whether the site was occupied, z_i, and the probability of detecting the warbler during any interval if the site was occupied, p. As in the first model, observations for eBird will be summarized counts for each grid cell, W_j.

We again include a CAR model for spatial random variation, but no longer use a multivariate since a single state variable is being used. The CAR random effects for each grid cell are denoted by θ_j.

We again use a generalized linear model with a logistic link to describe the probability a point count location will be occupied. The model now includes an intercept, spatial random-effect, our two environmental covariates – forest cover and elevation, and two covariates related to the eBird data.

$$\text{logit}(\psi_i) = \alpha_{01} + \alpha_1 * forest + \alpha_2 * elevation + \theta_j + \beta_{count} * W_j + \beta_{effort} * eLists_j$$

We constrain β_{count} to be positive and β_{effort} to be negative. The result is to allow counts to be positive predictors of occupancy but to discount this by the amount of effort expended in the grid cell. We are no longer able to estimate effort as a random variable, and instead choose a single metric for effort in this formulation.

The bugs code for this model is:

```
model {
# CAR prior for the spatial random effects
# CAR prior
    spacesigma ~ dunif(0,5)
    spacetau <- 1/(spacesigma*spacesigma)
    S[1:ncell] ~ car.normal(adj[],weights[],num[],spacetau)
    alpha.l ~ dunif(0,1)
    alpha <-  log(alpha.l) - log(1-alpha.l)
### BBA model
### priors
    p ~ dunif(0,1)
    b.forest ~ dnorm(0,0.01)
    b.elev ~ dnorm(0,0.01)
    b.eCount ~ dunif(0,10)
    b.eList ~ dunif(-10,0)
### data model
    for (i in 1:nsite){
      z[i] ~ dbern(psi[cell[i]])
      muy[i] <- z[i]*p
      Y[i] ~ dbin(muy[i],5)
    }
    for (j in 1:ncell){
      logit(psi[j]) <- S[j] + alpha + b.forest*forest[j] + b.elev*elev[j] + b.eCount*W[j] +
                           b.eList*effort[j,1]
    }
}
```