Estimating the drivers of species distributions with opportunistic data using mediation analysis

David B. Huberman,1,† Brian J. Reich,1 Krishna Pacifici,2 and Jaime A. Collazo3

1Department of Statistics, North Carolina State University, Campus Box 8203, Raleigh, North Carolina 27695 USA
2Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, North Carolina 27695 USA
3Department of Applied Ecology, U.S. Geological Survey, North Carolina Cooperative Fish and Wildlife Research Unit, North Carolina State University, Raleigh, North Carolina 27695 USA

Citation: Huberman, D. B., B. J. Reich, K. Pacifici, and J. A. Collazo. 2020. Estimating the drivers of species distributions with opportunistic data using mediation analysis. Ecosphere 11(16):e03165. 10.1002/ecs2.3165

Abstract. Ecological occupancy modeling has historically relied on high-quality, low-quantity designed-survey data for estimation and prediction. In recent years, there has been a large increase in the amount of high-quantity, unknown-quality opportunistic data. This has motivated research on how best to combine these two data sources in order to optimize inference. Existing methods can be infeasible for large datasets or require opportunistic data to be located where designed-survey data exist. These methods map species occupancies, motivating a need to properly evaluate covariate effects (e.g., land cover proportion) on their distributions. We describe a spatial estimation method for supplementarily including additional opportunistic data using mediation analysis concepts. The opportunistic data mediate the effect of the covariate on the designed-survey data response, decomposing it into a direct and indirect effect. A component of the indirect effect can then be quickly estimated via regressing the mediator on the covariate, while the other components are estimated through a spatial occupancy model. The regression step allows for use of large quantities of opportunistic data that can be collected in locations with no designed-survey data available. Simulation results suggest that the mediated method produces an improvement in relative MSE when the data are of reasonable quality. However, when the simulated opportunistic data are poorly correlated with the true spatial process, the standard, unmediated method is still preferable. A spatiotemporal extension of the method is also developed for analyzing the effect of deciduous forest land cover on red-eyed vireo distribution in the southeastern United States and find that including the opportunistic data do not lead to a substantial improvement. Opportunistic data quality remains an important consideration when employing this method, as with other data integration methods.

Key words: mediation analysis; occupancy modeling; opportunistic data; spatial statistics.

Received 6 June 2019; revised 16 January 2020; accepted 2 March 2020. Corresponding Editor: Lucas Neal Joppa.

Copyright © 2020 The Authors. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
† E-mail: dbhuberm@ncsu.edu

INTRODUCTION

A primary ecological species distribution analysis tool is occupancy modeling, a method which estimates species occupancy from presence/absence data while accounting for imperfect species detection (MacKenzie et al. 2002). Occupancy modeling and analysis traditionally assume designed-survey data collected by a trained surveyor team. Occupancy modeling data analysis is predicated on the surveyors following the given ecological design, although methodological extensions have been developed to accommodate assumption violations like false-positive species presence recordings (Miller et al. 2011). This method of data collection can require significant effort, but typically results in a high-quality dataset that correlates well with the underlying
spatial process as long as a rigorous probabilistic sample design is followed. Recent technological advances have led to a massive increase in the availability and abundance of citizen-collected opportunistic data. The opportunistic data collection process is significantly less resource-intensive than traditional ecological designed-survey data collection, resulting in a spatiotemporal data coverage that is likely unattainable through the traditional designs. However, the increased coverage of opportunistic data can come at the price of reduced or uncertain data quality (Delauney et al. 2008, Jiguet 2009, Dickinson et al. 2010).

Effective integration of opportunistic data with existing methods for designed-survey data is essential for unlocking previously unavailable insights on species distributions. Multiple integration methods have been developed which assume opportunistic data to be in a presence-only format where absences are not recorded (Dorazio 2014, Renner et al. 2015).

Integration methods have also been developed for opportunistic data sources that can reasonably be considered presence-absence data (Paciﬁci et al. 2017, Miller et al. 2019). Paciﬁci et al. (2017) proposed several occupancy modeling techniques for data integration depending on the presumed quality of the data. In a scenario where the opportunistic data are believed to be of comparable quality to the designed-survey data, Paciﬁci et al. (2017) used a shared model where both data sources directly and equally influence occupancy estimation. If the opportunistic data are expected to be less correlated with occupancy than the designed-survey data, Paciﬁci et al. (2017) suggested a correlation model or a covariate model. In the correlation model, the opportunistic data influence occupancy estimation indirectly through a shared covariance matrix with the designed-survey data. In the covariate model, the opportunistic data are inserted into the design matrix as covariates.

In this paper, we build on the covariate model described in Paciﬁci et al. (2017) and propose a method for improving covariate effect estimation with mediation (or path) analysis concepts. In mediation analysis, the relationship between a predictor and outcome variable is theorized to be partially or fully mediated by an intervening variable. The total effect of the predictor on the outcome can be decomposed into a direct effect of the predictor on the outcome and an indirect effect of the predictor on the mediating variable which itself affects the outcome. In our context, we consider the total effect of the ecological predictor of interest on the designed-survey outcome and allow for the opportunistic data to function as the mediator. We exploit the total effect decomposition into direct and indirect effects to incorporate additional opportunistic data. The direct and indirect effects are directly estimated and combined in place of the total effect estimate.

Two key features of our method are its relative speed and ability to incorporate large quantities of opportunistic data. One of the indirect effect components can be estimated by linearly regressing the opportunistic data on the predictor variable. This is computationally convenient compared to a spatial analysis that would require the inversion of a non-diagonal covariance matrix, explained more in Methods (Gelfand et al. 2010). This also allows for the inclusion of additional opportunistic data from locations where no designed-survey data are available.

After a basic review of mediation analysis, we expand on the details of our method and its desirable features. Following this, we highlight the utility of our method with a simulation study and then apply it to real-world designed-survey and opportunistic datasets.

**METHODS**

**Mediation analysis review**

Mediation (or path) analysis is an extension of regression that attempts to quantify the causal relationship between a group of variables. For notational simplicity, we will only consider one independent variable; however, multiple covariates could be included if desired. In this framework, the expected effect of increasing independent variable $X$ by one unit on outcome variable $Y_1$ is referred to as the total effect (Hayes 2009). This effect might be direct, or it might occur indirectly through a mediator variable (Hayes 2009). Formally, the direct effect represents the change in $Y_1$ when $X$ is increased by one unit, holding the mediator value $Y_0$ fixed (Alwin and Hauser 1975). The indirect effect represents the change in $Y_1$ when $X$ is held fixed.
and $Y_0$ is adjusted by how much it is expected to change if $X$ had been increased by one unit (Alwin and Hauser 1975).

For concreteness, suppose we are interested in the effect of silvicultural practices on songbird distribution. We could straightforwardly regress the presence/absence of the songbird on different forest management regimes to get an estimate of the effect of silviculture. Alternatively, we might also consider an intervening variable such as forest type which mediates the relationship between forest management and songbird distribution. In other words, perhaps some of the total effect of silviculture on songbird distribution is due to its direct causal relationship, while the rest of the total effect is due to an indirect relationship where silviculture affects forest type which in turn affects songbird distribution. The decomposition of the total effect into direct and indirect effects is the key concept leveraged in our method.

In a linear regression context, the total effect can be decomposed into the sum of direct and indirect effects (Alwin and Hauser 1975). The regression equation for capturing the total effect of $X$ on $Y_1$ is:

$$Y_1 = X\beta + \varepsilon_1$$

(1)

where $X = (1 \ X)$, $\beta = (\beta_0 \ \beta_1)^T$, and $\varepsilon_1$ represents a mean-zero normal error.

We take the expectation of $Y_1$ and obtain $E[Y_1] = X\beta$, where $\beta_1$ represents the total effect of $X$ on $Y_1$. $\delta$ represents the effect of the mediating variable $Y_0$ on $Y_1$, given the effect of $X$. The regression equations representing the total effect decomposition through the mediator are:

$$Y_0 = X\eta + \varepsilon_2$$

(2)

$$Y_1|Y_0 = X\tilde{\beta} + Y_0\delta + \varepsilon_3$$

(3)

where $\eta = (\eta_0 \ \eta_1)^T$, $\tilde{\beta} = (\tilde{\beta}_0 \ \tilde{\beta}_1)^T$, and $\varepsilon_2$ and $\varepsilon_3$ are independent mean-zero normal errors. Using these equations, the expectation of $Y_1$ can be calculated as:

$$E[Y_1] = E[E[Y_1|Y_0]] = E[X\tilde{\beta} + Y_0\delta] = X\tilde{\beta} + X\eta\delta$$

(4)

Thus, the total effect $\beta$ can be equivalently written as $\beta = \tilde{\beta} + \eta\delta$, where $\beta_1 + \eta_1\delta$ represents the decomposed total effect of $X$ on $Y_1$. $\tilde{\beta}$ and $\eta\delta$ represent the direct and indirect effects, respectively (Alwin and Hauser 1975). Kenny (2018) notes that the equivalence of $\beta$ and $\tilde{\beta} + \eta\delta$ is guaranteed in a multiple regression context if the same cases are used across the analysis (as well as the same covariates). Kenny (2018) also states that the total effect and its decomposition may only be approximately equivalent in certain regression settings such as multilevel modeling or structural equation modeling with latent variables. For these scenarios, Kenny (2018) argues for computing $\tilde{\beta}$, $\eta$, and $\delta$ to estimate $\beta = \tilde{\beta} + \eta\delta$ over computing $\beta$ directly.

Occupancy models are often constructed hierarchically to accommodate the assumed latent occurrence process (Royle & Dorazio 2008). Moreover, opportunistic data will likely be available at significantly more locations than designed-survey data are. In many cases, it would be a suboptimal use of available information to restrict these data to only locations where designed-survey data are available. For these reasons, the implication from Kenny (2018) is that estimating $\beta = \tilde{\beta} + \eta\delta$ through estimating $\tilde{\beta}$, $\eta$, and $\delta$ might be more useful for occupancy modeling inference than estimating $\beta$ in an unmediated fashion as commonly done.

**Mediation for spatial occupancy covariate modeling**

Let $s$ represent a spatial location. For this problem, $X(s)$, $Y_1(s)$, and $Y_0(s)$ are functions of spatial location $s$ that represent a habitat variable of interest, designed-survey species binary/count data, and opportunistic species binary/count data, respectively. $X(s)$, $Y_1(s)$, and $Y_0(s)$ are the spatially analogous versions of $X$, $Y_1$, and $Y_0$ introduced in *Methods: Mediation analysis review*, with $X(s)$ being a $1 \times p$ vector of spatial covariates. For simplicity, we assume $p = 2$, an intercept and predictor variable of interest. We note that the method outlined in this paper assumes $Y_1(s)$ is the number of positive observations in a fixed number of trials and $Y_0(s)$ follows a log-normal distribution. Opportunistic data are likely to be recorded as positive count data, which is incompatible with our normality assumption. However, a log transform of the opportunistic data would have valid support for a normal distribution (from $-\infty$ to $\infty$) and be
able to retain the linear relationship. Thus, our resolution is to assume the log-transformed opportunistic data are normally distributed. The mediation method could be adapted and applied for other types of data as long as the linear relationships of these variables are retained.

Unmediated method.—This method estimates the coefficient in Eq. 1 for the effect of \(X(s)\) on \(Y_1(s)\). We define a spatial probit occupancy model analogous to Eq. 1. A probit model is selected to reflect the assumed presence/absence format of \(Y_1(s)\). A location must contain both \(Y_1(s)\) and \(X(s)\) information to be eligible for analysis with the unmediated method’s spatial probit occupancy model.

The designed-survey data \(Y_1(s) \sim \text{Binomial}(N(s), pO(s))\) are conditional on the number of samples \(N(s)\), detection probability \(p\) and \(O(s)\), the true occurrence state process. To model the true occurrence state, Johnson et al. (2013) suggest introducing \(Z(s)\), a latent, continuous, and normally distributed analog of \(O(s)\) so that

\[
O(s) = \begin{cases} 
1 & Z(s) > 0 \\
0 & Z(s) \leq 0 
\end{cases} . 
\] (5)

Note that the mediated independent effect calculations done in Methods: Mediation analysis review assume a linear relationship between \(Y_1(s)\) and \(X(s)\). However, \(Y_1(s)\) is binomially distributed, making a direct linear link with \(X(s)\) impossible. Instead, this method considers the latent variable \(Z(s)\) in the place of \(Y_1(s)\) to maintain the structure in Eq. 1.

We then model \(Z(s)\) as a conditional Gaussian process with \(X(s)\beta\) and spatial random effect \(\theta(s)\) entered into the mean function and \(1 - \lambda\) as the variance. Recall from Methods: Mediation analysis review that \(\beta = (\beta_0, \beta_1)\) is a 2 by 1 coefficient vector with \(\beta_0\) and \(\beta_1\) representing the intercept and coefficient, respectively. \(\theta(s)\) is modeled as a Gaussian process with \(E[\theta(s)] = 0\) and \(\text{Cov}([\theta(s_1), \theta(s_2)]) = \lambda f(h)\) where \(h = ||s_1 - s_2||, \lambda \in (0, 1)\), and \(f\) is an exponential correlation function.

In summary, the unmediated model is.

\[
Y_1(s) | O(s), p \sim \text{Binomial}(N(s), pO(s)) ,
\]

where \(O(s) = \mathbb{1}(Z(s) > 0)\) 

\[
Z(s) | \theta(s), \lambda \sim \text{Normal}(\theta(s) + X(s)\beta, 1 - \lambda) .
\] (7)

The variance in the conditional \(Z(s) | \theta(s)\) distribution is set as \(1 - \lambda\) so that the variance of \(Z(s)\) marginal of \(\theta(s)\) is one and thus \(\beta\) and \(\lambda\) are identified. To understand this, first recall that a normal variable can be standardized by subtracting its mean and dividing by its standard deviation so that the transformed variable follows the standard normal distribution. Then, the probability that a site is occupied (i.e., \(Z(s) > 0\)) is:

\[
\text{Pr}(Z(s) > 0) = \text{Pr} \left( \frac{Z(s) - E[Z(s)]}{\sqrt{\text{Var}[Z(s)]}} > \frac{-E[Z(s)]}{\sigma} \right) = \Phi \left( \frac{X(s)\beta}{\sigma} \right)
\] (8)

where \(\sigma\) is the standard deviation of \(Z(s)\), and \(\Phi\) is the standard normal cumulative density function. The standard deviation will change the interpretation of \(\beta\) if \(\sigma \neq 1\), so we define the unmediated model to avoid this.

Mediated method.—The mediated estimate is obtained in two separate steps, as we need to estimate coefficients from both Eqs. 2 and 3. \(Y_0(s)\) is taken to be the logarithm of the observed event rate (i.e., the number of observations divided by the observation effort). Let \(\eta = (\eta_0, \eta_1)\) be the coefficient vector capturing the effect of \(X(s)\) on \(Y_0(s)\) with \(\eta_0\) and \(\eta_1\) representing the intercept and coefficient of interest, respectively. For the coefficient in Eq. 2, we model \(Y_0(s) | X(s)\) as

\[
Y_0(s) | X(s) \sim \text{Normal}(X(s)\eta, \sigma_0^2) .
\] (9)

The linear effect of \(X(s)\) on \(Y_0(s)\) is quantified by \(\eta\) which can be estimated using ordinary least-squares (OLS) estimation. The linear relationship between \(Y_0(s)\) and \(X(s)\) ensures that \(\beta\) can be decomposed using \(\eta\) similarly to the decomposition described in Eq. 4. In this step, we ignore spatial correlation in \(Y_0(s) | X(s)\) for estimating \(\eta\). This is justified as we assume the opportunistic dataset is large, so the least-squares estimator of \(\eta\) is unbiased and consistent (under regularity conditions) even in the presence of spatial dependence. We would expect most ecological opportunistic datasets to be sufficiently large for use with our method, as less resource-intensive data collection is a key feature of the data type.
The spatial probit occupancy model for the mediated method is similar to the model in Methods: Mediation for spatial occupancy covariate modeling: Unmediated method with adjustments made to incorporate \( Y_0(s) \). Recall from Methods: Mediation analysis review that \( \hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1) \) is a 2 by 1 coefficient vector with \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) representing the intercept and coefficient, respectively. Then, the model is:

\[
Y_1(s)|O(s), p \sim \text{Binomial}(N(s), pO(s)), \text{where } O(s) = \mathbb{1}(Z(s) > 0),
\]

\[
Z(s)|Y_0(s), 0(s), \lambda \sim \text{Normal} \left( 0(s) + X(s)\hat{\beta} + Y_0(s)\delta, 1 - \lambda \right).
\]

(10)

As in Methods: Mediation for spatial occupancy covariate modeling: Unmediated method, \( 0(s) \) is a mean-zero Gaussian process with \( \text{Cov}[0(s_1), 0(s_2)] = \lambda f(h) \) where \( h = ||s_1 - s_2|| \) and \( f \) is an exponential correlation function.

The direct and indirect effects for the mediated method are given by the \( \hat{\beta} \) and \( \hat{\eta} \) vectors, respectively. Summing these two effects gives the total effect \( \hat{\beta} + \hat{\eta} \). This vector contains the mediated intercept and coefficient effects.

As mentioned in the Introduction, the mediated method allows for the usage of more location data than the unmediated method. The spatial probit occupancy model for the mediated method requires locations with \( Y_1(s), Y_0(s) \), and \( X(s) \) information, similar to the data requirements for the unmediated method’s spatial probit occupancy model. However, the mediated method regression step can also incorporate locations which only have \( Y_0(s) \) and \( X(s) \) information. The regression component thus allows for the use of locations that were not eligible for analysis in the unmediated method because they did not contain designed-survey information. Designed-survey data may be more difficult to obtain at many locations than opportunistic and covariate information. This allows the mediated method to potentially analyze significantly more data than the unmediated method. Furthermore, as explained in the Introduction, the linear regression step is a more convenient way of bringing in additional locations with opportunistic and covariate information than fitting another spatial model. Recall that \( p \) gives the number of predictor variables and \( n \) gives the number of observations. The computational complexity for the linear regression step is \( O(p^2n) \), compared to \( O(n^3) \) for a standard non-diagonal covariance matrix inversion. Consequently, the linear regression step is comparatively quicker as long as \( p \) is less than \( n \). For many applications, we would expect this to hold true.

Recall in Eq. 8 we explained that \( \text{Var}[Z(s)] \) needs to be a scalar to avoid identifiability issues. It can be shown for the mediated method that \( \text{Var}[Z(s)] = 1 + \delta^2 \sigma_0^2 \) using the law of total variance. Therefore, we must apply a correction to the intercept and coefficient estimates to remove the influence of \( \delta^2 \sigma_0^2 \) on \( \text{Var}[Z(s)] \). Let \( \sigma_0^2 \) be the residual standard error estimate of \( \sigma_0^2 \) calculated in the linear regression step in Eq. 9. Then, for the mediated method, we take Eq. 8 with

\[
\sigma = \sqrt{1 + \delta^2 \sigma_0^2}
\]

and obtain:

\[
\text{Pr}(Z(s) > 0) = \Phi \left( \frac{\hat{\beta}_0 + \hat{\eta}_0 \delta + X(s)(\hat{\beta}_1 + \hat{\eta}_1 \delta)}{\sqrt{1 + \delta^2 \sigma_0^2}} \right) = \Phi \left( \frac{\hat{\beta}_0^* + X(s)\hat{\beta}_1^*}{\sqrt{1 + \delta^2 \sigma_0^2}} \right)
\]

(12)

where \( \hat{\beta}_0^* \) and \( \hat{\beta}_1^* \) represent the corrected intercept and coefficient estimates for the mediated method, respectively. These estimates can be corrected after the models have been evaluated and the parameter estimates have been obtained.

**Simulation**

We conduct a simulation study to compare the performance of the unmediated and mediated methods from Methods: Mediation for spatial occupancy covariate modeling. Data analysis in this project was conducted using a 2014 MacBook Pro with a 2.2 GHz Intel Core i7 processor and 16 GB 1600 MHz DDR3 Memory. The statistical software used was RStudio v1.1447 with R v3.5.0.

**Settings.**—We spatially generated 1000 data points on an evenly spaced 40 \( \times \) 25 grid with grid spacing one unit. Habitat data \( X(s) \) are generated for these locations as realizations of a mean-zero Gaussian process with a variance equal to one and an exponential correlation

![Image](image-url)
function with range equal to one unit so that correlation between orthogonally adjacent observations is $e^{-1} \approx 0.368$. Then, designed-survey and opportunistic data are generated as:

$$Y_1(s) | O(s), p \sim \text{Binomial}(N(s), p O(s)),$$

where $O(s) = \mathbb{I}(Z(s) > 0)$

$$Y_0(s) | \theta(s) \sim \text{Normal}(a + b \theta(s), \sigma^2).$$

$$Z(s) | \theta(s), \lambda \sim \text{Normal}(\theta(s) + X(s) \beta, 1 - \lambda).$$

$$\theta \sim \text{MVN}(0, \lambda \Sigma(\kappa)).$$

where $\theta = (\theta(s_1) \ldots \theta(s_n))$ is the vector of random spatial effects, $a$ and $b$ are constants, $\sigma^2 > 0$ is the conditional variance for the mediating variable, and all other terms are as previously defined. Recall from Methods: Mediation for spatial occupancy covariate modeling: Mediated method that $Y_0(s)$ is the logarithm of a count variable. The log transform is necessary to obtain the appropriate support bounds for a normal distribution. $Y_0(s)$ is normally distributed to maintain the linear relationship between the variables needed for our mediation decomposition. The rest of the data generation process follows the model described in Methods: Mediation for spatial occupancy covariate modeling: Unmediated method. For all data in the simulation, we set $\beta_0 = 0$, $\beta_1 = 0.5$, $N(s) = 5$, $\lambda = 0.5$, $\nu = 0.5$, $a = 0$, and $b = 1$. Let $\kappa$ the spatial range parameter in the exponential correlation function, and recall that $p$ represents the detection probability. We randomly generate $X(s)$ and $Y_1(s)$ data for each grid location for all possible combinations of the simulation settings.

1. $\rho \in \{0, 0.8\},$
2. $p \in \{0.25, 0.75\},$ and
3. $\kappa \in \{0.5, 1.5\}.$

The parameter $\rho$ controls the dependence between the true latent process and the auxiliary data, with $\rho = 0$ giving $\text{cor}(Y_0(s), \theta(s)) \approx 0$ which makes $Y_0(s)$ irrelevant, and $\rho = 0.8$ giving $\text{cor}(Y_0(s), \theta(s)) \approx 0.8$. Fig. 1 gives an example spatial representation of data generated with the model and settings $p = 0.75$ and $\kappa = 0.5$. We repeat this data generation 300 times for each simulation setting. For each generated dataset, a random subset of either $n = 50$ or $n = 200$ grid locations are selected and the designed-survey data from the unselected sites are removed to create sixteen simulation settings. The opportunistic data are retained for all sites.

Bayesian MCMC techniques are implemented on the subsetted data to evaluate the spatial occupancy models for both methods and obtain the posterior distributions of the parameters. The additional linear regression step in the mediated method is evaluated through OLS estimation, using all 1000 data locations. The coefficient estimate for the mediated method is corrected as in Methods: Mediation for spatial occupancy covariate modeling: Mediated method. For details on the prior distributions, posterior updates, and other model details, refer to Appendix S1. Code for the simulation and case study can be found at https://github.com/dbhuberm/MediatedOccupancyAnalysis.

The average effect estimate mean squared error (MSE), coverage, relative MSE, and relative MSE standard error are calculated for both methods using all 300 datasets for each setting. The MSE is calculated as:

$$\frac{\sum_{i=1}^{300} (\hat{\beta}_i - \bar{\beta})^2}{n}$$

where $i$ indexes the 300 datasets used for the given setting, $\bar{\beta}$ represents the true effect of $X(s)$ on $Y_1(s)$, and $\hat{\beta}$ represents the posterior-mean estimator for the unmediated or mediated method. Coverage is calculated as the percentage of simulations for which the true effect is included in the equal-tailed 95% posterior interval. The relative average MSE is calculated as the mediated average MSE divided by the unmediated average MSE (multiplied by a factor of 100). A relative average MSE below 100 indicates the mediated average MSE is smaller than the unmediated average MSE. The associated standard error is calculated using bootstrap sampling.

**Case study**

To provide an example of this method’s usage in ecological statistics, we consider the species distribution of the red-eyed vireo (Vireo olivaceus) across the continental United States from 2008 to 2013. The objective is to assess the relationship between deciduous forest land cover change and red-eyed vireo occupancy.
We chose the red-eyed vireo because the species exhibits a number of attributes shared by many North American landbirds. For example, they are migratory, insectivorous, and use a wide range of plant communities. The red-eyed vireo breeds in the eastern and northern United States and much of southern Canada and winters in South America. It is a common and widespread species that uses a wide variety of forest habitats to nest and glean insects from foliage. Counts of red-eyed vireos have been increasing by 0.8% annually since 1966 and by 1.0% annually since 2003, with some localized declines in Florida, Texas, and the Pacific Northwest (Sauer et al. 2015).

Data sources.—The designed-survey dataset used in our case study is presence–absence data from the North American Breeding Bird Survey (BBS) (Sauer et al. 2015). BBS data arises from observers who sample a set of pre-specified North American routes with 50 stops. At each stop on a given route, the observers survey the surrounding area. For this case study, the surveys at each stop within a route are considered independent. Each route is associated with a unique latitude and longitude, so that the presence/absence surveys from all 50 stops are considered to be surveys from the same location. For each survey route, a naive occupancy proportion

---

**Fig. 1.** Plots of the $\theta(s)$ values, $Z(s)$ values, $Y_0(s)$ values when $\text{cor}(Y_0(s), \theta(s)) \approx 0.8$, and $Y_1(s)$ counts at each location for data generated with $\kappa = 1.5$ and $p = 0.75$. 

---
between 0 and 1 can be calculated by dividing the number of individuals recorded by the number of stops.

The opportunistic dataset used in this study is from the Cornell Lab of Ornithology’s eBird database (eBird) (Sullivan et al. 2009). Citizen scientists record and submit records from birds that they have seen or heard electronically. Here, we only used recordings that came from a checklist provided by eBirds. Therefore, we can consider these data as presence/absence because if no positive detection was recorded for a species, then it is safe to assume it was not detected. We could not make that determination without the extra user-provided information that all species on the full checklist were surveyed for, due to the possibility that the user did not survey the location for that species. In addition, we only used eBird data where the user indicated their effort (in hours). We standardized the eBird data by effort to give the average number of bird presence recordings per unit effort (see Eq. 15). Only stationary count eBird data from March through July were included to filter for data quality.

The land cover covariate data were obtained using the National Land Cover Database (NLCD) created by the Multi-Resolution Land Characteristics consortium (MRLC) (Vogelmann et al. 2001). The raster state datasets used in this paper were downloaded from https://viewerm.nationalmap.gov/basic/. The NLCD contains high-resolution land cover image data across the United States of America for years 2006 and 2011. Each pixel represents a coordinate with land cover type information. The site land cover data for years 2008 through 2010 and 2012 through 2013 was obtained through linear interpolation, using the 2006 and 2011 site land cover data as endpoints. All grid sites contained land cover information.

We generate a 1 by 1 degree rectangular grid over the mainland United States spanning from −125 to −67 degrees longitude and 24.5 to 49.5 degrees latitude, restricted to grid locations containing land area. We further restrict the longitude and latitude range of this grid to match the longitude and latitude range of our BBS red-eyed vireo data, resulting in a grid spanning from −99 to −67 degrees longitude and 24.5 to 49.5 degrees latitude. This restriction was made because we focused on the Eastern half of the United States.

Breeding Bird Survey, eBird, and covariate data are assigned to their respective grid cells and aggregated for analysis. The number of BBS species recordings and number of BBS survey stops are aggregated from all BBS routes within the given grid cell. The aggregated BBS variable is calculated as the proportion of species recordings to survey stops. For eBird, the species recordings and effort hours are aggregated within the given grid cell and the aggregated eBird variable is calculated as:

\[ Y_0(s) = \log\left(\frac{C_0(s)}{N_0(s)} + 1\right) \]  

where \( C_0(s) \) and \( N_0(s) \) represent the sum of reported eBird species presence recordings and the sum of reported eBird effort hours at grid site \( s \), respectively.

The aggregated land cover proportion represents the percentage of the given grid cell covered by deciduous forests. Due to data size issues for most states, the land cover raster data were inputted for each state individually and pre-aggregated at a small scale (squares of 50 horizontal and vertical pixels) in order to be feasible to convert to non-raster data for analysis. The raster data for Connecticut, District of Columbia, Delaware, Maryland, Massachusetts, New Hampshire, New Jersey, Rhode Island, and Vermont were small enough to be converted with no pre-aggregation. The data for each state are projected to latitude and longitude coordinates using the World Geodetic System (WGS) 84 projection. The latitudes and longitude extents for each state are determined to ensure overlap of locations from different state datasets is negligible. Finally, the land cover information is mapped to the specified grid. The aggregated land cover proportion is calculated as the number of pixels in the given grid cell representing deciduous forest land cover divided by the number of land cover pixels in the given grid cell overall.

Settings.—There were 463 grid locations for each year with at least one red-eyed vireo occupancy data source, resulting in 2778 total grid locations over 6 yr. There was BBS information on red-eyed vireo occupancy in 1884 or 68% of these grid locations, compared to eBird red-eyed vireo occupancy data in 2597 or 93% of the grid locations. The size of the grid was selected to limit the
number of grid locations which contained BBS but not eBird information. Our method assumes that all locations with BBS information also contain eBird information, so this information needs to be interpolated if not available. For a 1 by 1 degree grid, 60 of the locations with BBS information did not also have eBird information, or around 2% of the total grid locations. We chose this grid size as a balance between number of observations and keeping this percentage small, because the number of locations with BBS but no eBird information only increased at higher resolutions. eBird proportion values for these 2% of grid locations with incomplete information were interpolated through simple Kriging, with Kriged eBird proportion values below 0 truncated to 0 (Cressie 1990).

Fig. 2 displays the aggregated naive (and cleaned) BBS red-eyed vireo occupancy, aggregated naive eBird red-eyed vireo occupancy, and aggregated deciduous forest land cover proportions data for 2008, respectively.

The mediated and unmediated method introduced in this paper was modified to account for possible temporal correlation across seasons (years). The spatial probit occupancy model was modified to be a spatiotemporal probit occupancy model as:

\[ Y_{1t} | O_t, \theta \sim \text{Binomial}(N_t, O_t \theta) \]  
where \( O_t = 1(Z_t > 0) \)  

\[ Z_t | \theta_t, \beta \sim \text{MVN}(\theta_t + X_t \beta, (1 - \lambda)I_n) \]  

\[ \theta_1 \sim \text{MVN}(0, \lambda \Sigma(\kappa)) \]  

\[ \theta_t | \theta_{t-1} \sim \text{MVN}(\rho \theta_{t-1}, (1 - \rho^2)\lambda \Sigma(\kappa)). \]  

where \( Y_{0t} \delta \) is additively inserted into the conditional mean for \( Z_t \) when using the mediated method, as in the single season version. For the mediated method, the linear regression step introduced in Eq. 9 is conducted on eBird and land cover data from all seasons simultaneously. The corrected intercept and coefficient estimates described in Eq. 12 are calculated using these all season parameter estimates. The formulas for these two steps remain unchanged.

As in the simulation, the spatiotemporal models are evaluated using MCMC methods. For \( \rho \), a random sample is drawn from 1000 evenly spaced candidates between \(-1 \) and \( 1 \). The candidate probabilities are weighted by the associated standardized likelihood values. The other priors are similar to the priors given in Appendix S1 but generalized for a multi-season model.
RESULTS

Simulation

Table 1 compares model performance in terms of average MSE and coverage across settings when $p = 0.8$ and thus $\text{cor}(Y_0(s), \theta(s)) \approx 0.8$. Across all of these simulation settings, the mediated method produces relative MSEs below 100. All but two of these settings are more than a standard error below 100, showing the utility of the mediated method in estimating the effect of $X(s)$ when the mediating variable $Y_0(s)$ is strongly correlated with the true spatial process. Both methods result in around 95% coverage for all of the settings, and there do not appear to be any trends in coverage percentage related to the simulation setting parameters.

Altering $p$, $\kappa$, or $n$ all appear to influence the relative average MSEs. Generally, the relative average MSE decreases as $n$ is decreased from 200 to 50, indicating that the mediated method's value increases when there are comparatively fewer locations with $Y_1(s)$ information. Holding $p$ and $n$ constant, setting $\kappa = 1.5$ decreases the relative average MSE compared to setting $\kappa = 0.5$. This is likely due to the increased utility of extra spatial information. Lastly, decreasing $p$ from 0.75 to 0.25 appears to improve the relative average MSE, though this did not occur when $\kappa = 0.5$ and $n = 200$.

Table 2 compares model performance in terms of average MSE and coverage across settings when $p = 0$ and thus $\text{cor}(Y_0(s), \theta(s)) \approx 0$. As expected, the mediated method does not improve compared to the unmediated method when the auxiliary data are uncorrelated with the true process. The relative average MSEs increase compared to the strong correlation case.

---

Table 1. The average MSE, 95% coverage, and relative MSE results for the unmediated and mediated methods on data generated to have a correlation of $\approx 0.8$ on average between $Y_0(s)$ and $\theta(s)$.

| Settings | Unmediated | Mediated | Relative Avg MSE |
|----------|------------|----------|-----------------|
|          | Avg MSE    | Coverage (%) | Avg MSE | Coverage (%) | Estimate (%) | SE (%) |
| $p$ | $\kappa$ | $n$ |          | Avg MSE | Coverage (%) |          | Avg MSE | Coverage (%) | Estimate (%) | SE (%) |
| 0.25 | 0.5 | 200 | 0.025 | 95.7 | 0.023 | 92.7 | 94.0 | 13.0 |
| 0.25 | 0.5 | 50 | 0.120 | 95.3 | 0.090 | 95.0 | 75.5 | 8.6 |
| 0.25 | 1.5 | 200 | 0.027 | 93.7 | 0.021 | 95.0 | 80.3 | 9.4 |
| 0.25 | 1.5 | 50 | 0.121 | 93.0 | 0.084 | 93.3 | 69.2 | 9.6 |
| 0.75 | 0.5 | 200 | 0.011 | 94.7 | 0.010 | 94.7 | 92 | 6.3 |
| 0.75 | 0.5 | 50 | 0.048 | 95.7 | 0.045 | 95.7 | 93.0 | 7.3 |
| 0.75 | 1.5 | 200 | 0.010 | 96.0 | 0.009 | 97.3 | 87.1 | 6.4 |
| 0.75 | 1.5 | 50 | 0.047 | 95.7 | 0.040 | 94.7 | 85.2 | 8.1 |

Notes: $p$ and $\kappa$ refer the detection and range parameter values used to generate the data, respectively. $n$ refers to the number of subsetted locations of the original 1000 that retain $Y_1(s)$ data.

Table 2. The average MSE, 95% coverage, and relative MSE results for the unmediated and mediated methods on data generated to have a correlation of $\approx 0$ on average between $Y_0(s)$ and $\theta(s)$.

| Settings | Unmediated | Mediated | Relative Avg MSE |
|----------|------------|----------|-----------------|
|          | Avg MSE    | Coverage (%) | Avg MSE | Coverage (%) | Estimate (%) | SE (%) |
| $p$ | $\kappa$ | $n$ |          | Avg MSE | Coverage (%) |          | Avg MSE | Coverage (%) | Estimate (%) | SE (%) |
| 0.25 | 0.5 | 200 | 0.025 | 95.7 | 0.027 | 96.3 | 110.7 | 9.2 |
| 0.25 | 0.5 | 50 | 0.120 | 95.3 | 0.118 | 94.7 | 98.4 | 9.6 |
| 0.25 | 1.5 | 200 | 0.027 | 93.7 | 0.026 | 93.7 | 98.2 | 6.1 |
| 0.25 | 1.5 | 50 | 0.121 | 93.0 | 0.121 | 93.0 | 99.8 | 9.1 |
| 0.75 | 0.5 | 200 | 0.011 | 94.7 | 0.011 | 94.7 | 99.5 | 3.0 |
| 0.75 | 0.5 | 50 | 0.048 | 95.7 | 0.049 | 95.7 | 102.3 | 4.2 |
| 0.75 | 1.5 | 200 | 0.010 | 96.0 | 0.010 | 97.7 | 100.3 | 3.3 |
| 0.75 | 1.5 | 50 | 0.047 | 95.7 | 0.046 | 94.7 | 96.3 | 4.1 |

Notes: $p$ and $\kappa$ refer the detection and range parameter values used to generate the data, respectively. $n$ refers to the number of subsetted locations of the original 1000 that retain $Y_1(s)$ data.
for all of these settings. All relative average MSE estimates are within one standard error of 0.10 except the setting with $p = 0.25$, $\kappa = 0.5$, and $n = 200$, which is at least one standard error above. However, these results confirm that the mediated method still produces valid statistical inference in this case (i.e., coverage near 95%). As with most other opportunistic data fusion methods, the quality of the Data S1 should be considered before deciding to use this method.

**Case study**

Table 3 gives the posterior mean and standard deviation for each parameter under the unmediated and mediated methods. The eBird red-eyed vireo coefficient in the mediated method was only 0.18 (with a standard deviation of 0.84), suggesting a weak relationship with the BBS red-eyed vireo data. The unmediated method estimate of the effect of deciduous forest land cover on red-eyed vireo occupancy is 2.34 (with a standard deviation of 0.30).

For the mediated method, the linear regression coefficient estimate was $\hat{\beta}_1 \approx 0.01$ while the variance estimate was $\sigma^2_0 \approx 0.0002$. Thus, the mediated method effect of deciduous forest land cover on red-eyed vireo occupancy is calculated as:

$$\hat{\beta} + \hat{\eta}_1 \delta \approx \frac{2.29 + (0.01)(0.18)}{\sqrt{1 + \delta^2 \sigma^2_0}} \approx 2.29$$

Table 3. The posterior mean and standard deviation (SD) for each parameter under the unmediated and mediated methods for the red-eyed vireo analysis.

| Parameter                        | Unmediated Method | Mediated Method |
|----------------------------------|-------------------|-----------------|
|                                 | Mean   | SD   | Mean   | SD   |
| Partial Sill                    | 0.84   | 0.15 | 0.86   | 0.07 |
| Range                           | 0.88   | 0.25 | 0.68   | 0.21 |
| Detection                       | 0.25   | <0.01| 0.25   | <0.01|
| Temporal Correlation            | 0.96   | 0.02 | 0.97   | 0.02 |
| Unmediated/Mediated Intercept   | 1.06   | 0.12 | 1.20   | 0.16 |
| Unmediated/Mediated Land cover Coefficient | 2.34   | 0.30 | 2.29   | 0.65 |
| Mediated eBird Coefficient      | N/A    | N/A  | 0.18   | 0.84 |

with a posterior standard deviation of 0.65. These are identical to the estimates of $\hat{\beta}_1$ and SD ($\hat{\beta}_1$) given for the mediated method in Table 3. We expect that opportunistic data more highly correlated with the designed-survey data would have highlighted the mediated method’s capacity for insight. On the other hand, this result demonstrates the method’s conservativeness when using lower correlated opportunistic data. The mediated method point estimate of the deciduous forest land cover effect on red-eyed vireo occupancy is similar, with the main penalty for including lower correlated opportunistic data being a larger standard error. In cases where the opportunistic data have a low correlation with the designed-survey data, the supplementary incorporation of opportunistic data ensures the mediated estimate will not deviate too sharply from the unmediated estimate. If the opportunistic dataset is known to be well correlated with the designed-survey dataset and computational speed is not a concern, other more aggressive methods might be considered. Even then, many other methods would only allow for the incorporation of opportunistic data where designed-survey data already exists, so this method might still be preferable.

**Discussion**

The mediated method proposed in this paper attempts to improve upon estimates of a predictor variable’s effect on species occupancy from designed-survey data. This method utilizes a secondary data source, likely opportunistic data, in a supporting capacity as a mediating variable. The simulation study showed that when the opportunistic data source is well correlated with the designed data source, our mediated method produces smaller MSEs and higher coverage relative to the unmediated method. The mediated and unmediated methods performed similarly when the opportunistic data source was not well correlated with the designed data source in the simulation study.

The mediated method did not provide significant improvement when applied to red-eyed vireo occupancy data compared to the unmediated method. Both methods indicated that deciduous forest land cover had an effect on red-eyed
vireo occupancy. The two methods did not produce notably different estimates; however, the unmediated method obtained higher precision. This difference in precision is unsurprising. Every parameter estimator has some amount of associated variation (or uncertainty). The unmediated method estimates a single parameter, whereas the mediated method estimates a linear combination of multiple parameters. The mediated method estimator involves more sources of variation and consequently has a higher level of associated uncertainty. The mediated method result suggests the supplementary inclusion of the opportunistic data source buffers against poor estimation when the data are not well correlated. Furthermore, the mediated method is computationally convenient and constructed to allow for opportunistic data usage beyond the locations with designed-survey data. Therefore, we consider it a low risk, flexible, and relatively expeditious occupancy modeling option for multiple data sources.

Additionally, we introduce a spatiotemporal version of the method for the case study with an order 1 autoregressive (AR1) correlation structure. The data were highly temporally correlated, with the autocorrelation parameter being estimated close to one. Temporal variation was not made a focus in this paper, but it could be worth more thoroughly examining the effectiveness of the mediated method in a spatiotemporal context with various temporal correlation strengths.

Our proposed method can theoretically be applied to grid data or point data. However, the spatial probit occupancy model step in the mediated method requires that any location with designed-survey data must also have opportunistic and covariate data. This is less likely if data are collected for points rather than a grid. We would not expect that these three data sources to be sampled at the exact same locations, in which case we would need to map them to a common grid to perform the mediated analysis. It is possible that the designed-survey and covariate information are obtained from the same data source at the same locations, in which case unmediated analysis on point data is feasible. Otherwise, the unmediated method would also require data gridding.

Regarding the grid resolution selection, we expect that the highest possible grid resolution is preferable as this results in the most data points. At high grid resolutions, it is likelier that some grid locations have BBS but not eBird information, in which case the eBird information must be interpolated. We would like to minimize the amount of interpolated data, a preference in conflict with using the highest possible grid resolution. An alternative approach could include reconciling the difference in spatial resolution (change of support; Pacifici et al. (2019)) before conducting mediation analysis, but this would require fitting a much more complicated model. For our case study, we selected a grid size that resulted in around 2% of locations requiring eBird interpolation. In general, we anticipate the theoretically superior spatiotemporal coverage of opportunistic data will allow for the selection of relatively high grid resolutions with minimal or no need to interpolate opportunistic data values. Still, it might be worthwhile to analyze results from different grid resolutions and come up with a rule of thumb for grid selection that balances these two concerns.

Acknowledgments

We thank the Cornell Lab of Ornithology for making available the eBird data and the US Geological Survey’s Patuxent Wildlife Research Center (along with cooperating researchers) for making available the North American Breeding Bird Survey data on the red-eyed vireo. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Literature Cited

Alwin, D. F., and R. M. Hauser. 1975. The decomposition of effects in path analysis. American Sociological Review 40:37–47.

Cressie, N. 1990. The origins of kriging. Mathematical Geology 22:239–252.

Delaney, D. G., C. D. Sperling, C. S. Adams, and B. Leung. 2008. Marine invasive species: validation of citizen science and implications for national monitoring networks. Biological Invasions 10:117–128.

Dickinson, J. L., B. Zuckerberg, and D. N. Bonter. 2010. Citizen science as an ecological research tool: challenges and benefits. Annual Review of Ecology, Evolution, and Systematics 41:149–172.

Dorazio, R. M. 2014. Accounting for imperfect detection and survey bias in statistical analysis of
presence-only data. Global Ecology and Biogeography 23:1472–1484.

Gelfand, A. E., P. Diggle, P. Guttorp, and M. Fuentes. 2010. Handbook of spatial statistics. CRC Press, Boca Raton, Florida, USA.

Hayes, A. F. 2009. Beyond Baron and Kenny: statistical mediation analysis in the new millennium. Communication Monographs 76:408–420.

Jiguet, F. 2009. Method learning caused a first-time observer effect in a newly started breeding bird survey. Bird Study 56:253–258.

Johnson, D. S., P. B. Conn, M. B. Hooten, J. C. Ray, and B. A. Pond. 2013. Spatial occupancy models for large data sets. Ecology 94:801–808.

Kenny, D. A. 2018. “Mediation”. http://www.davidakenny.net/cm/mediate.htm

MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. Andrew Royle, and C. A. Langtimm. 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology 83:2248–2255.

Miller, D. A., J. D. Nichols, B. T. McClintock, E. H. C. Grant, L. L. Bailey, and L. A. Weir. 2011. Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. Ecology 92:1422–1428.

Miller, D. A., K. Pacifi, J. S. Sanderlin, and B. J. Reich. 2019. The recent past and promising future for data integration methods to estimate species distributions. Methods in Ecology and Evolution 10:22–37.

Pacifi, K., B. J. Reich, D. A. Miller, B. Gardner, G. Stauffer, S. Singh, A. McKerrow, and J. A. Collazo. 2017. Integrating multiple data sources in species distribution modeling: a framework for data fusion. Ecology 98:840–850.

Pacifi, K., B. J. Reich, D. A. Miller, and B. S. Pease. 2019. Resolving misaligned spatial data with integrated species distribution models. Ecology 100: e02709.

Renner, I. W., J. Ethlith, A. Baddeley, W. Fithian, T. Hastie, S. J. Phillips, G. Popovic, and D. I. Warton. 2015. Point process models for presence-only analysis. Methods in Ecology and Evolution 6:366–379.

Royle, J. A., and R. M. Dorazio. 2008. Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities. Academic Press, Amsterdam, the Netherlands. https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/j.1365-2664.2009.01724.x

Sauer, J., J. Hines, J. Fallon, K. Pardieck, D. Jr Ziolkowski, and W. Link. 2015. The North American breeding bird survey, results and analysis 1966–2013. Version 01.30.2015. US Geological Survey Patuxent Wildlife Research Center, Laurel, Maryland, USA. https://www.mbr-pwrc.usgs.gov/bbs/bbs.html

Sullivan, B. L., C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling. 2009. eBird: a citizen-based bird observation network in the biological sciences. Biological Conservation 142:2282–2292.

Vogelmann, J. E., S. M. Howard, L. Yang, C. R. Larson, B. K. Wylie, and N. Van Driel. 2001. Completion of the 1990s National Land Cover Data Set for the conterminous United States from Landsat Thematic Mapper data and ancillary data sources. Photogrammetric Engineering and Remote Sensing 67:26.

**Supporting Information**

Additional Supporting Information may be found online at: http://onlinelibrary.wiley.com/doi/10.1002/ecs2.3165/full