Estimating species occupancy across multiple sampling seasons with temporal autologistic occupancy models via the autoOcc R package

# Abstract

To be written, not worried about it.

# Introduction

Dynamic occupancy models—which estimate a species distribution through time as a function of local colonization and extinction rates—are a powerful statistical tool. Such models have been used, for example, to quantify wolf (*Canis lupus*) recolonization throughout France (Louvrier et al. 2018), predict avian range dynamics (Kalle et al. 2017, Briscoe et al. 2021), and estimate pulses in colonization rates based on a species reproductive phenology (Fidino and Magle 2017). Dynamic occupancy models are also data hungry (Briscoe et al. 2021). Simulation studies, for example, propose that at least 120 unique sampling locations are required to accurately estimate colonization and extinction rates (Mckann et al., 2012). If true, this suggestion from Mckann et al. (2012) severely limits researchers interested in quantifying species distributions through time because most occupancy studies sample fewer than 120 locations (Kays et al. 2020). Making matters worse, wildlife survey data collected through time is common, especially as long-term research networks grow (Cove et al. 2021, Magle et al. 2020). Consequently, there is a need for an accessible tool that effectively quantifies species distributions through time, especially with reduced sample sizes.

One approach to quantify species distributions through time would be to fit a single-season occupancy model with a “stacked” design. Such a model treats each site and primary sampling period combination as a separate detection history and—because sites are sampled over multiple primary sampling periods—accounts for pseudoreplication via a site-level random intercept. A study with 50 sites and 4 primary sampling periods, for example, would have 200 detection histories if no data was missing and 50 random effect terms to estimate. Stacked design occupancy models are now relatively easy to fit because random intercepts can be incorporated into occupancy models via the unmarked R package (Fiske and Chandler 2011). This model does not, however, account for temporal dependence in occupancy between seasons and instead such patterns are soaked into the site-level random effect, which complicates interpretation. Furthermore, if there is insufficient data to fit a dynamic occupancy model because they have too many parameters, stacked design occupancy models may have the same problem, if not worse. As the site-level random effect requires many parameters to be estimated from your detection/non-detection data, “stacked” design occupancy models are not likely the simplest solution to address sample size issues with data collected over time.

Autologistic occupancy models are another approach that can account for spatial or temporal dependence in species distributions. Spatial autologistic occupancy models account for spatial autocorrelation such that the occupancy probability at a site can vary if nearby sites are also occupied. Temporal autologistic occupancy models, on the other hand, account for temporal autocorrelation such that the occupancy probability at a site can vary if that site was also occupied in the previous timestep. While I will specifically focus on temporal autologistic occupancy models here, those who are interested in spatial autologistic models can refer to Royle and Dorazio (2008). Temporal autologistic occupancy models (hereafter autologistic occupancy models) have been applied to numerous taxa. This model, for example, has been used to quantify the relationship between urban mammals and gentrification (Fidino et al. 2024), between bird diversity and variation in fire severity (Tingley et al. 2016), and between amphibian occupancy and climate change (Zipkin et al. 2012). Unlike dynamic occupancy models or single-season “stacked” design occupancy models, autologistic occupancy models are a useful option because they only require one additional parameter to quantify temporal dependence in occupancy. Yet, despite their utility, autologistic occupancy models are seldom used because they require researchers to write custom code to fit them and parameterize the model under a Bayesian framework, which severely limits the pool of researchers able to use this statistical method.

In this paper I introduce the autoOcc R package (<https://github.com/mfidino/autoOcc>), which can be used to fit autologistic occupancy models under a frequentist framework, compare the relative fit of models via AIC, and easily make model predictions. After explaining the basic sampling scheme and statistical formulation of this class of model I use simulations to show that autologistic occupancy models are accurate and precise under most scenarios, even when working with limited sample size. Furthermore, I show how autoOcc can be used in a standard analysis via two worked examples: estimating habitat associations of Virginia opossum (*Didelphis virginiana*, hereafter opossum) throughout Chicago, Illinois, USA and quantifying spatiotemporal patterns in black-backed woodpecker (*Picoides arcticus*) distributions as a function of fire severity throughout California’s montane forests (Stillman et al. 2023). I hope that by formally introducing autologistic occupancy models, their assumptions, and making them easy to fit in R, that this class of model will finally become a valid option for researchers collecting species detection/non-detection data through time.

# Explanation of the method

## Basic sampling scheme and model assumptions

The sampling protocol for autologistic occupancy models is identical to the multi-season protocol developed by MacKenzie et al. (2003) for dynamic occupancy models. Briefly, some number of sites are surveyed over multiple sampling periods (e.g., years) to collect species detection/non-detection data. We assume that while the occupancy status of a species may change at sites between our sampling periods, the occupancy status does not change within a sampling period (i.e., the closure assumption). During each sampling period, a researcher conducts multiple independent surveys at each site to generate a detection history, which are a collection of 1’s and 0’s that respectively indicate whether a species was or was not detected on an independent survey at the site of interest. As such, each site has a vector of detection data for each sampling season.

Autologistic occupancy models have four key assumptions in addition to the closure assumption. First, we assume that the techniques used to sample the species of interest can result in false negatives (i.e., a species is present but not detected) but not false positives (i.e., the species is not present but was mistakenly detected). Second, we assume all sampled sites are spatially independent. Thus, the presence of a species at a site does not influence species presence at other locations nor does the detection of a species at a site have an influence on detecting the species on other independent surveys. Third, autologistic occupancy models include a first-order autoregressive term to account for whether the presence of a species in one sampling period influences the occupancy status in the following sampling period. As such, this model assumes temporal dependence over adjacent sampling periods (e.g., *t-1* to *t*) but independence over larger time frames (e.g., *t – 1* to *t + 1*). This assumption is similar in spirit to dynamic occupancy models which condition on species presence in the previous timestep to estimate local colonization and extinction rates in the current timestep. Fourth, autologistic occupancy models assume the probability of occupancy and detection is either constant across sites or explained by covariates. In other words, there is no unmodeled site-specific heterogeneity. If such assumptions are violated then the resulting model may be over precise or estimators could be biased and, as a result, the inference made from the associated model could be wrong (Bailey et al. 2013).

## The model

The simplest way to describe autologistic occupancy models is with a latent binary variable that denotes whether the species of interest is present or not at a site. While autoOcc does not use this parameterization to estimate the associated model parameters, this is the best way to understand how the autologistic term, *θ*, is used in the model. Thus, for *i* in 1,…,*I* sites and *t* in 1,…,*T* primary sampling periods (hereafter seasons), let *zi,t* be the latent binary occupancy status of a species at site *i* and time *t* and let *ψi,t* be the occupancy probability. During the first season there is no information about the occupancy status of the species before sampling began. Thus, when *t* = 1 the latent state model is

|  |  |  |
| --- | --- | --- |
|  | *,* | Eq. 1 |

where is a vector of regression coefficients (, which includes the model intercept) and their associated covariates (), which is indexed by *i* and *t* because the covariates could vary across space, time, or both. Note that the first element of is a 1 to account for the model intercept. This setup for the first season of data is identical to dynamic occupancy models. However, autologistic occupancy models use the occupancy parameters across all seasons of data instead of explicitly estimating local colonization and extinction rates. For *t* > 1 we modify the logit-linear predictor by adding our autologistic term, *θ*, that is multiplied by a species occupancy status at the site of interest in the previous season.

|  |  |  |
| --- | --- | --- |
|  | *,* for  *t > 1* | Eq. 2 |

When the species is present during the previous season then and *θ* is added to the logit-linear predictor, otherwise it is excluded. The most common error I’ve seen with interpretation of *θ* is the assumption that *θ* informs autocorrelation in the occupancy status of a species even if they are absent in the previous timestep. This is false. If the species was absent at t-1, then *θ* = 0, which means that *θ* has no bearing on the occupancy status at time *t*. Instead, *θ* is fully conditional on species presence in the previous timestep. Positive *θ* values indicate that species presence in the previous timestep increases *ψi,t* in the current timestep, whereas negative *θ* values indicate the opposite. As such, when the autologistic *θ* term is added to the linear predictor it just increases or decreases the latent state model intercept, depending on its direction, given that the species is present at t-1.

For the data model let *pi,t,j* be the conditional probability of detecting the species during *j* in 1,…,*J* secondary sampling periods (hereafter surveys) given the species is present. Further, let *y*i,t,j represent the detection / non-detection data for site *i*, season *t*, and survey *j* which equals 1 if the species was detected, 0 if not, and NA if data was not collected. This level of the model is

|  |  |  |
| --- | --- | --- |
|  |  | Eq. 3 |

where is a vector of regression coefficients (**)** and their associated covariates () that can vary across sites, seasons, or surveys. Given Eq. 1 through Eq. 3, you can see that autologistic occupancy models only add one new parameter to the model, *θ*, to account for temporal dependence in species occupancy from one timestep to the next.

The latent variable approach described above is perhaps the easiest way to understand autologistic occupancy models and could be coded up as a Bayesian hierarchical model in either NIMBLE (de Valpine et al. 2017) or JAGS (Plummer 2003). However, if we drop the latent variable and write the model with matrix notation it is possible to estimate the associated parameters via maximum likelihood, which is what I did for autoOcc. To make it easier to follow along with this model formulation, let’s assume one site has been sampled across four seasons with three surveys per season to generate the detection history [110], [000], [101], [100]. Thus, the species was detected at least once in the first, third, and fourth seasons and was not detected at all during the second season. For this section we need to define two occupancy probabilities. Let *ηt* represent the occupancy probability at survey *t* during either the first season or when the species was not present at time t-1 while *ωt* is the occupancy probability if the species was present at *t - 1*. These two probabilities are nearly identical and as such share parameters. For example, an intercept only model for these probabilities could be written as logit(*ηt*) = *β*0 and logit(*ωt*) = *β*0  + θ. Following MacKenzie et al. (2003), and dropping the site subscript for simplicity, the probability of observing the first season detection history is

|  |  |  |
| --- | --- | --- |
|  |  | Eq. 4 |

A verbal description of Eq. 4 is “the species was present and detected on the first two surveys but not the last survey.” In the second season the species was not detected. That means one of two independent events occurred. Either the species was not present at *t* = 2 or the species was present but not detected. Because the species was present in the previous season we use *ω*2 instead of *η2*. Thus, the probability of observing the second season detection history is

|  |  |  |
| --- | --- | --- |
|  |  | Eq. 5 |

For *t* = 3, we are uncertain if the species was present in the previous season. As such, we use *η3* and *ω3* to generate the probability of this survey, multiplying both by the necessary detection probabilities:

|  |  |  |
| --- | --- | --- |
|  |  | Eq. 6 |

A verbal description of Eq. 6 is “the species either was or was not present in season 2 and was detected on the first and third survey of season 3, but not the second survey.” Looking at Eq. 6, note that it could be further simplified by factoring out the detection probabilities. Finally, for t = 4 we know that the species was present in the previous timestep, so we only need to use *ω4*:

|  |  |  |
| --- | --- | --- |
|  |  | Eq. 7 |

The probability of observing the entire detection history is the product of Eq. 4 – 7.

While it would be possible to write out the likelihood of each detection history as I did in Eq. 4 – 7, doing so would be difficult to generalize to any dataset. Fortunately, MacKenzie et al. (2003) describe a more general approach that uses matrix notation for their dynamic occupancy model, which I modified for autoOcc. Let *ϕ0* be a row vector for the first sampling period such that

|  |  |  |
| --- | --- | --- |
|  |  | Eq. 8 |

For the remaining seasons we need a 2 x 2 matrix of transition probabilities, ***ϕ****t*, that describes how a site may move from one state to the next from season *t* to *t + 1*.

|  |  |  |
| --- | --- | --- |
|  |  | Eq. 9 |

Note that rows of ***ϕ****t* denote the occupancy state at *t*, columns are the occupancy state at *t + 1*, and all rows sum to 1. For example, *ω*t is the probability the species was present at *t* and *t + 1*, while 1 –*ηt* is the probability the species was not present at *t* and *t + 1*. The elements that make up ***ϕ****t* in Eq. 9 is the primary difference between autologistic and dynamic occupancy models, wherein the latter explicitly estimates local colonization and extinction rates. Finally, let *δ****y,t*** be a column vector that contains the probability you would observe detection history **y**t,1:J on season *t*, conditional on the occupancy state. As such, *δ****y,t*** changes as a result of the observed data. While this is not how autoOcc handles this component of the model, it may help to imagine *δ****y,t*** as a matrix of column vectors with a number of columns equal to the total number of possible detection histories. When considered in this way, **y**t,1:J indexes the appropriate column vector that matches it’s detection history to grab the correct detection probabilities. Two examples of these column vectors are

|  |  |  |
| --- | --- | --- |
|  |  | Eq. 10 |

Note that the second element of each column vector takes either the value 0 if the species was detected at least once across surveys during a season or 1 if the species was not detected. With those three components you can calculate the probability of observing a given detection history as

|  |  |  |
| --- | --- | --- |
|  |  | Eq. 11 |

In Eq. 11, the D(***δ****y,t*) function indicates that the elements in the column vector ***δ****y,t* are placed along the main diagonal of a diagonal matrix. This transformation is done to ensure that all the elements of Eq. 11 are conformable and that the appropriate likelihood is calculated. See supplemental material 1 for a worked example of Eq. 11 with the four-season detection history we used for Eq. 4 – 7. Because Eq. 11 is the probability of one detection history, the model likelihood is

|  |  |  |
| --- | --- | --- |
|  |  | Eq. 12 |

As with the dynamic occupancy model outlined in MacKenzie et al. (2003), the autologistic occupancy model here can accommodate covariates via the logit link and handle missing surveys. These extensions have already been added to autoOcc.

## Deriving expected occupancy estimates from autologistic occupancy models

Expected occupancy estimates from autologistic occupancy models can be generated in a similar fashion to expected occupancy estimates from dynamic occupancy models. More specifically, dynamic occupancy models are used to estimate local colonization (γ) and extinction rates (ε), and the expected occupancy of these probabilities can be derived as γ / (γ + ε). By replacing those probabilities with the ones we estimate from the autologistic occupancy model, we arrive at the following formula:

|  |  |  |
| --- | --- | --- |
|  |  | Eq. 12 |

To demonstrate how this may look with parameters estimated via autoOcc and their associated covariates we could rewrite Eq. 12 as

|  |  |  |
| --- | --- | --- |
|  |  | Eq. 13 |

Where ilogit() represents the inverse logit link function. Eq. 13 makes it clear that to derive expected occupancy estimates from autologistic occupancy models we must estimate occupancy probabilities with and without the autologistic term, both of which can be done with covariates in autoOcc.

# Things to consider before using this method

To provide some guidance on how much data should be collected before using autoOcc — and illustrate how it performs to other statistical techniques a researcher may use — I conducted a simulation study to compare the accuracy and precision of autologistic occupancy models to dynamic occupancy models across a range of sample sizes. I chose dynamic occupancy models for comparison instead of other techniques (e.g., a “stacked” design single-season occupancy model with site-level random effects) because dynamic occupancy models do not require random effects to estimate, which can complicate model fitting and interpretation. For each class of model, simulations varied in how common the species was, the number of sites sampled, and the number of seasons sampled. More specifically, I used four scenarios where the expected occupancy of a species was either 0.2, 0.3, 0.4, and 0.5. For the autologistic occupancy model this was done by setting *ϑ* = 1 and using Eq.13 to determine what the model intercept should be to achieve the correct expected occupancy. For the dynamic occupancy model there were often multiple colonization and extinction rates that could generate these expected occupancy probabilities. After determining possible solutions for each expected occupancy scenario I chose one at random. The number of sites sampled ranged from 30 to 100 locations in intervals of 10 while seasons sampled ranged from 4 to 12 in intervals of 2. Thus, for the four expected occupancy, seven site, and six season scenarios there was a total of 160 different combinations to simulate for both the autologistic and dynamic occupancy models. Every combination was simulated and fitted 550 times and resulted in a total of 176,000 simulations and model fits.

For each simulation scenario I held some values constant. First, I included one environmental gradient for each logit-linear predictor (i.e., occupancy and detection for autologistic models and initial occupancy, colonization, extinction, and detection for dynamic models). Across both classes of model I set slope terms to 1 on the logit scale for all latent states except for the initial occupancy slope term of the dynamic model, which was set to 0 (i.e., the environmental covariate was not associated to initial occupancy). The initial occupancy intercept in the dynamic model was set to 1 (i.e., an initial occupancy probability of ~0.73). For the data model the same environmental covariate was used, but the slope term associated with this covariate was set to 0.5 across both models. Finally, I set the intercept of the data model to -0.9 on the logit scale, which resulted in a 0.29 average detection probability per survey. As I assumed 4 surveys per sampling period the overall probability of detecting the species at least once if they were present was roughly 0.75 (i.e., 1 – (1 – 0.29)4). With the addition of these slope terms, the autologistic and dynamic occupancy model respectively had five and eight parameters to estimate from their simulated datasets. After I fitted the simulated datasets across all scenario combinations I calculated the relative bias (root mean square error, hereafter RMSE) and precision (i.e., width of 95% confidence intervals) of latent state parameters.

Overall, autologistic occupancy models had less bias than dynamic occupancy models. Averaged across scenarios the relative bias of the autologistic latent state intercept was about 6.5 times less than the dynamic colonization intercept but was about 1.45 times greater than the dynamic extinction intercept (Figure 1). The autologistic term, θ, term had consistently less bias which was respectively 51.0 and 5.4 times smaller than the dynamic colonization and extinction intercepts. Across all parameters the dynamic model colonization intercept had the greatest bias when small sample sizes were used (Figure 1K, Fig 1O). Averaged across scenarios, the relative bias in the autologistic model slope term was 4.2 times smaller than the dynamic model colonization slope term and 1.6 times smaller than the extinction slope term (Figure 2). Furthermore, the relative bias in the colonization slope term increased with the expected occupancy of the species and, like the colonization intercept, was highest when small sample sizes were used (Figure 2.K).

Autologistic occupancy models were also more precise than dynamic occupancy models. Averaged across scenarios the average 95% CI width of the autologistic intercept was 1.86 times narrower than the dynamic model colonization intercept and 1.3 times narrower than the extinction intercept (Figure 3). However the autologistic term, θ, had greater uncertainty under some scenarios. The average 95% CI width for θ was roughly 1.09 times larger than the dynamic colonization intercept and 1.6 times larger than the dynamic extinction intercept. This difference was largely driven by the 0.4 and 0.5 expected occupancy scenarios, which had relatively wide 95% CI widths for both θ (Figure 3N) and the colonization intercept (Figure 3O). Averaged across scenarios, the autologistic model slope term confidence interval was 2.67 and 1.56 times narrower than the dynamic model colonization and extinction slope term confidence intervals, respectively (Figure 4). The largest difference between models was when the expected occupancy of the species was 0.5, especially at small sample sizes. This last result should not be surprising given that binomially distributed variables have the greatest variance when the probability of success is 0.5.

When considered together, autologistic occupancy models had less bias and more precision than dynamic occupancy models across a wider range of sample sizes. Notably, across all scenarios autologistic occupancy models had relatively low bias for all latent state parameters. This was not true for dynamic models, which especially struggled when the species was more common and when a smaller number of sites were sampled. As other researchers have suggested that dynamic occupancy models require at least 120 sites to be used reliably (McKann et al. 2012), autologistic occupancy models are clearly a valid option when dealing with small sample sizes. Increasing sample size did deliver a notable increase in precision for both models, especially with respect to θ in the autologistic model. Most importantly, autologistic model slope terms were far more precise than dynamic model slope terms, which showcases that autoOcc can be especially useful if the goal of a study is to evaluate the habitat associations of a species and a researcher is limited with the amount of data they may be able to collect.

In closing this section, I want to caution that the results of this simulation study cannot provide rigorous suggestions for how much data you need before even considering using autologistic occupancy models. The natural world is, after all, far more complex than the simulations I ran. As such, your appropriate sample size will vary depending on your research questions, logistical constraints, and the ecology of the species you plan to study. If I had to provide some recommendations to start with, I would focus on trying to increase precision because the relative bias was low across all scenarios of the autologistic model simulations. As such, people interested in using this class of model may be able to achieve high precision with a minimum of 60 sites sampled for 8 seasons and a moderate precision with 40 sites sampled for 6 seasons.

# Worked examples

To demonstrate how models can be fitted within autoOcc I have two worked examples coming from different taxa and data collection methods. For the first, I analyzed camera trap data collected throughout Chicago, Illinois to quantify if different social-ecological gradients are associated to the distribution of Virginia opossum (*Didelphis virginiana*). For the second, I recreated an analysis by Stillman et al. (2023) who used a Bayesian autologistic occupancy model with survey data to assess how pyrodiversity, or the spatial and temporal variation of fire characteristics, affects black-backed woodpecker occupancy throughout the montane forests of California. Across both worked examples I compare the relative fit of different models using AIC (Anderson and Burnham, 2004), and use a ΔAIC of 2 as a cutoff value to determine which models within a model set were competitive.

## Virginia opossum occupancy throughout Chicago, IL

The data for this example comes from 96 spatial locations across the greater Chicago metropolitan area (Chicago, IL, USA). In 2019, camera traps were deployed throughout urban greenspace for 28-day sampling seasons in January, April, July, and October for a total of four primary sampling periods (see Magle et al. 2019 for further sampling details). Daily detection histories were summarized to weekly detection histories for this analysis.

For this example, let’s assume we are interested in understanding how patterns of urban intensity and neighborhood wealth are associated with opossum occupancy, both of which may be associated with opossum occupancy in non-linear ways. Furthermore, as opossum activity changes throughout the year due to Chicago’s cold winters (Gallo et al. 2022), we also want to quantify the relationship between opossum detection probability and average weekly temperature.

For our analysis we need to assemble three components in R: 1) the opossum detection history, 2) occupancy covariates, and 3) detection covariates. The detection data for this example are already included within autoOcc so those can be loaded and set up for further analysis.

# load opossum detection / non-detection data

library(autoOcc)

data("opossum\_det\_hist")

# convert long format data to an array

opossum\_y <- format\_y(

x = opossum\_det\_hist,

site\_column = "Site",

time\_column = "Season",

history\_columns = "Week"

)

If you looked at this dataset or checked it’s help file (?opossum\_det\_hist) you will notice that this dataset has 6 columns that provide information on the name of the sampled location (Site), the season sampled (Season), and the detection data (Week\_1 through Week\_4). Furthermore, this dataset is in long format and is sorted along two columns, season and site, such that sites are sorted alphabetically within each of the four seasons of data. To use this dataset in autoOcc our detection history cannot be in long format and instead needs to be set up as a site by season by survey three-dimensional array, which is similar to other occupancy modeling packages like spOccupancy (Doser et al. 2022). The format\_y() function carries this out for you so long as you specify which columns denote sites, seasons, and detection data. For the history columns, format\_y() uses regular expressions to locate all columns that include whatever is included as an argument. As such, care must be taken to ensure you are selecting the appropriate columns in your dataset. If other columns included ‘Week’ then we could add more specificity to that argument by changing it to "^Week\_", which would look for columns that begin with the word “Week” followed by an underscore.

After setting up our detection history we need to prepare our occupancy covariates for analysis. To represent a gradient of urban intensity I used the proportion of impervious cover within 1 km of each sampling location which was queried from the National Land Cover Database (Dewitz, 2019). To represent neighborhood wealth I used the median per capita income within 1 km of each site from the 2014-2018 American Community Survey (U.S. Census Bureau 2012). These two covariates can be queried from the opossum\_covariates dataset located within autoOcc. After subsetting the data I scaled the covariates for our analysis by subtracting their mean and dividing by their standard deviation, which can help improve model convergence.

# load covariates

data("opossum\_covariates")

# subset only impervious and income

opossum\_covariates <- opossum\_covariates[,

grep("Impervious|Income", colnames(opossum\_covariates))

]

# make new data.frame, scale covariates, and convert back to numeric

occ\_cov\_list <- opossum\_covariates

occ\_cov\_list$Impervious <- as.numeric(scale(occ\_cov\_list$Impervious))

occ\_cov\_list$Income <- as.numeric(scale(occ\_cov\_list$Income))

If your occupancy covariates were temporal or spatiotemporal we would have to use a named list instead of a data.frame to store this information. In this case they are not so a data.frame is sufficient. My detection covariates, however, do vary temporally in this example. As a reminder, we wanted to quantify the relationship between opossum detection probability and average weekly temperature, which was summarized from daily temperatures provided by NCDC (NCDC 2023). As temperature varies across each week of sampling, these data need to be stored in a matrix with a number of rows equal to the number of sites and a number of columns equal to the number of surveys conducted over the entire study. In our example we have 96 sites, 4 seasons of data, and 4 weeks of sampling within each season. As such, the matrix for our weekly temperature covariate will have 96 rows and 16 columns such that the first four columns are associated to the four weeks of sampling in the first season and the last four columns are associated to the four weeks of sampling in the last season. As the temperature data I queried did not vary across space I replicated the same value along each column vector. Thus, assuming we have summarized our weather data down to 16 values, one for each week of sampling, the matrix for this detection covariate is

Temperature <- matrix(

rep(

as.numeric(

scale(

weather\_data\_vector

)

),

each = dim(opossum\_y)[1]

),

nrow = dim(opossum\_y)[1],

ncol = prod(

dim(opossum\_y)[2:3]

)

)

Where weather\_data\_vector is a numeric vector that contains the 16 temperature values that were queried. If we had a temporal covariate or spatiotemporal covariate that only varied by primary sampling period and not surveys we could instead store those data in a site by season matrix. Our temperature matrix, as well as any other covariates we may want to include on opossum detection probability, are stored within a named list. For example, if I wanted to control for our two social-ecological gradients on detection probability they can be included like so

det\_cov\_list <- list(

Temperature = Temperature,

Impervious = occ\_cov\_list$Impervious,

Income = occ\_cov\_list$Income

)

With these three pieces of data you can use autoOcc to fit a suite of models with the auto\_occ() function and then compare their relative fit with compare\_models(). For this analysis I fitted 10 models. Models understandably varied in which covariates were included but also whether those covariates had a quadratic term. For example, there are 8 possible models that could be fitted with two occupancy covariates, both with and without quadratic terms. The last two models included a temperature only model (i.e., intercept only for occupancy, but temperature and temperature2 on detection probability) and a null model. Every model except for the null model included temperature and temperature2 in the detection logit-linear predictor. As an example, the global model could be specified as

global\_quadratic <- auto\_occ(

~Temperature + I(Temperature^2) +

Impervious + I(Impervious^2) + Income + I(Income^2)

~Impervious + I(Impervious^2) + Income + I(Income^2),

y = opossum\_y,

det\_covs = det\_cov\_list,

occ\_covs = occ\_cov\_list

)

Note that the model formulas are the first argument of this function, and just as with the unmarked package, are written as a double right-hand side formula for detection and occupancy in that order. After fitting the remaining models and storing them in a list you can compare their relative fit with the compare\_models() function via AIC. There was only one competitive model which included a linear effect of income on opossum occupancy but a quadratic effect of impervious cover, and we can use the summary() function to look at the estimated parameters of this model.

# What the model list could look like after fitting the 10 models

model\_list <- list(

global\_quadratic = global\_quadratic,

global = global,

income\_quadratic = income\_quadratic,

income\_quad\_imperv = income\_quad\_imperv,

income = income,

imperv\_quadratic = imperv\_quadratic,

imperv\_quad\_income = imperv\_quad\_income,

imperv = imperv,

temperature = temp,

null = null

)

# compare models via AIC

aic\_results <- compare\_models(

model\_list,

digits = 2

)

# Look at first few models

head(aic\_results, 3)

model npar AIC delta AICwt cumltvWt

1 imperv\_quad\_income 11 1213.11 0.00 0.55 0.55

2 global\_quadratic 13 1215.27 2.16 0.19 0.74

3 imperv\_quadratic 9 1216.16 3.05 0.12 0.86

# Summarise best-fit model

summary(imperv\_quad\_income)

Call:

autoOcc::auto\_occ(formula = ~Temperature + I(Temperature^2) +

Impervious + I(Impervious^2) + Income ~ Impervious + I(Impervious^2) +

Income, y = opossum\_y, det\_covs = det\_cov\_list, occ\_covs = occ\_cov\_list)

optim convergence code: 0

optim iterations: 51

Occupancy estimates:

parameter Est SE lower upper p

1 psi - (Intercept) -0.324 0.200 -0.716 0.0674 1.05e-01

2 psi - Impervious -0.341 0.165 -0.664 -0.0177 3.87e-02

3 psi - I(Impervious^2) -0.274 0.131 -0.531 -0.0180 3.60e-02

4 psi - Income -0.279 0.164 -0.601 0.0436 9.02e-02

5 psi - theta 1.637 0.365 0.922 2.3517 7.16e-06

Note: psi - theta is the autologistic term

Detection estimates:

parameter Est SE lower upper p

6 rho - (Intercept) 0.3826 0.145 0.0984 0.6667 0.008315

7 rho - Temperature 0.3374 0.133 0.0761 0.5986 0.011373

8 rho - I(Temperature^2) -0.4384 0.116 -0.6661 -0.2106 0.000162

9 rho - Impervious -0.0536 0.145 -0.3386 0.2315 0.712669

10 rho - I(Impervious^2) -0.1291 0.110 -0.3446 0.0863 0.240025

11 rho - Income -0.2054 0.130 -0.4601 0.0493 0.113935

AIC: 1213.113

Making predictions for this model is relatively straightforward via the predict() function in autoOcc. For example, let’s assume we want to predict opossum occupancy across a gradient of impervious cover, which ranged from about 20% to 80% across our study. These predictions can be made in three steps. First, we generate a dataset with covariate values we’d like to make predictions with, holding other covariates at their mean. As we mean-centered our continuous covariates we can keep all other columns at their mean value of 0.

# A vector of impervious cover

imperv\_vec <- seq(20, 80, length.out = 300)

# The prediction data.frame

imperv\_dm <- data.frame(

matrix(

0,

ncol = ncol(imperv\_quad\_income@occcovs),

nrow = length(imperv\_vec)

)

)

# add column names

colnames(imperv\_dm) <- names(imperv\_quad\_income@occcovs)

Second, our impervious cover data needs to be scaled exactly as we did before. Thus, we need to center and scale the prediction impervious cover vector by the mean and standard deviation of the impervious cover data at our sampled sites, which are stored in the opossum\_covariates object.

# add in imperv and scale it in the same way as we did in the model

imperv\_dm$Impervious <- (

imperv\_vec - mean(opossum\_covariates$Impervious)

) / sd(opossum\_covariates$Impervious)

Finally, for our third step we can make the predictions. At a minimum, the predict() function in autoOcc requires you to provide the model you’d like to make predictions with and the type of prediction you’d like to make (“psi” for the latent state, “rho” for detection). In addition to this you can add a new dataset to make predictions with and input the confidence level you would like for you confidence intervals. Finally, because this function uses Monte Carlo simulations to approximate confidence intervals you can also specify a sufficiently large number of Monte Carlo simulations (defaults to 3000) and a seed to ensure you generate the same results.

imperv\_pred <- predict(

object = imperv\_quad\_income,

type = "psi",

newdata = imperv\_dm,

nsim = 3000,

seed = 453

)

These predictions can then be plotted out (Figure 5). In this example opossum occupancy was greatest when impervious cover was roughly 43% and decreased with increasing levels of neighborhood wealth. I also detected a non-linear relationship between opossum detection probability and weekly temperature such that opossum detection probability was highest when the weekly average temperature was around 15 °C (Figure 5).

## Black-backed woodpecker occupancy throughout the montane forests of California

For this second worked example I recreated the analysis conducted by Stillman et al. (2023), who used a Bayesian autologistic occupancy model to quantify how pyrodiversity affects black-backed woodpecker occupancy. For this study these authors had an impressive 10 years of woodpecker survey data at nearly 2,400 sites, though sites were not surveyed each year. For their sampling protocol, these authors used single-visit surveys that were divided into multiple independent surveys. Specifically, each survey consisted of playing black-backed woodpecker vocalizations for 30 seconds and then listening for responses for 1.5 minutes. A maximum of three playback surveys were conducted at each site, and playback surveys were stopped after the first detection. Finally, point counts were also conducted at roughly half of these sites after the playback surveys. For further information on sampling see Tingley et al. (2018, 2020).

For their paper, Stillman et al. (2023) compared three models: a static occupancy model, a temporal autologistic occupancy model, and a landscape-temporal autologistic occupancy. These last two models varied in which covariates were present within the model. While Stillman et al. (2023) also incorporated random effects into their models, autoOcc cannot. As such, random effects were not used for this worked example.

I fitted three models to these data. All models had the same detection logit-linear predictor, which included whether a survey was 2 or 3 minutes in length (2 min = 0, 3 min = 1), the ordinal day of a survey, and the survey type (point count = 0, playback = 1). As a departure from Stillman et al. (2023), I did not fit a static occupancy model as the first model and instead used an intercept-only autologistic model. The second model, which Stillman et al. (2023) described as their “temporal occupancy model,” accounted for the effects of years since fire on occupancy. This model included 5 occupancy covariates: the number of years since a fire had started at a site, the average burn severity of the last fire within 100m, the proportion pre-fire canopy cover within 100m, elevation, and latitude. This model included quadratic terms for elevation and fire age as well as a statistical interaction between elevation and latitude. The third, “landscape-temporal occupancy model,” had the covariates from the temporal occupancy model but also included other covariates to account for post-fire habitat dynamics. As such, in addition to the aforementioned covariates this model included a metric for diversity in burn severity, distance to a low burn severity forest patch, and the summed basal area of pre-fire live red fir and white fir within 100m. This landscape-temporal occupancy model included quadratic terms for elevation and fire age as well as statistical interactions between burn severity and fire age, elevation and fire age, and fir basal area and fire age.

Just like Stillman et al. (2023), I found that the landscape-temporal occupancy model had the best relative fit with the temporal occupancy model having a ΔAIC of 122.48 from the landscape-temporal model. Parameter estimates from the best-fit model were similar to the Bayesian analysis of Stillman et al. (2023) such that the autoOcc estimates fell within the 95% credible intervals of the Bayesian analysis for nearly every parameter (Figure 6). There was also less parametric uncertainty with the autoOcc analysis across every model parameter such that confidence intervals were always smaller than the Bayesian credible intervals in Stillman et al. (2023). This difference is likely because I omitted some random effects that Stillman et al. (2023) had used, which increased uncertainty in parameter estimates. Regardless, had this analysis originally been conducted with autoOcc, Stillman et al. (2023) would have found similar results and concluded that woodpecker occupancy was highest directly after a fire, especially if the site had a high burn severity (Figure 6). For the complete analysis in R, see supporting information.

# Caveats

One of the main strengths of autologistic occupancy models is that temporal dependence in site-level occupancy is accounted for with a single parameter. This makes the model simple and easier to apply to datasets that are smaller. However, this simple formulation may also be a weakness as it inherently assumes that *θ* does not vary across seasons or sites. Certainly, spatiotemporal covariates can be included within the model, but that does not explicitly quantify whether site-level persistence (i.e., the probability a site remains occupied if a species is present there in the previous timestep) varies across space or time. Autologistic occupancy models could be extended to address this specific issue, but at that point such a model is nearly equivalent to a dynamic occupancy model. Therefore, as with any analysis, it is important to consider what your research questions or management goals are to ensure that the model used is appropriate.

# Additional resources

In addition to this manuscript, autoOcc has an associated vignette that can be accessed so long as it is built when installing autoOcc. This can be opened up by running vignette("Overview") in R after loading the package. This vignette provides a description of autologistic occupancy models, how to add spatial, temporal, or spatiotemporal covariates to your autologistic occupancy models, and shows how all the functions within autoOcc can be used to run a complete analysis on the opossum dataset.

# Conclusion

Biodiversity monitoring surveys that collect data through time are increasingly common, yet the statistical tools to quantify habitat associations with such data have been somewhat limited when said data are subject to imperfect detection. The autoOcc package offers up perhaps the simplest approach, and through simulations I demonstrated how autologistic occupancy models can have relatively little bias even under small sample sizes. Through worked examples I provided a general outline for how an analysis can be conducted within the autoOcc ecosystem and then recreated the results of an already published study assessing black-backed woodpecker responses to fire severity (Stillman et al. 2023). Ultimately, I hope that all the information and results presented here, in combination with the autoOcc R package, will be help to researchers who are interested in understanding species occupancy patterns over time.

# Author contributions

Mason Fidino developed the idea for this Research Methods Guide, conducted the simulation study and worked examples, and wrote the manuscript.

# Acknowledgements

I am grateful to Andrew Stillman and Morgan Tingley for their incredibly helpful conversations while I was working on recreating the woodpecker analysis. I also thank Seth Magle and Brian Gerber for this comments on a previous draft of this manuscript.

Anderson, D., & Burnham, K. (2004). Model selection and multi-model inference. *Second. NY: Springer-Verlag*, *63*(2020), 10.

Bailey, L. L., MacKenzie, D. I., & Nichols, J. D. (2014). Advances and applications of occupancy models. *Methods in Ecology and Evolution*, *5*(12), 1269-1279.

Briscoe, N. J., Zurell, D., Elith, J., König, C., Fandos, G., Malchow, A. K., ... & Guillera‐Arroita, G. (2021). Can dynamic occupancy models improve predictions of species' range dynamics? A test using Swiss birds. *Global Change Biology*, *27*(18), 4269-4282.

Cove, M. V., Kays, R., Bontrager, H., Bresnan, C., Lasky, M., Frerichs, T., ... & Jordan, M. J. (2021). SNAPSHOT USA 2019: a coordinated national camera trap survey of the United States.

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., & Bodik, R. (2017). Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, *26*(2), 403-413.

Doser, J. W., Finley A. O., Kéry, M., & Zipkin E. F. (2022). spOccupancy: An R package for single-species, multi-species, and integrated spatial occupancy models Methods in Ecology and Evolution, 13, 1670-1678.

Dewitz, J., U.S. Geological Survey. National Land Cover Database (NLCD) 2019 Products, version 2.0. U.S. Geological Survey. https://doi.org/10.5066/P9KZCM54. Accessed 25 June 2022

Fidino, M. and Magle, S. B. (2017). Using Fourier series to estimate periodic patterns in dynamic occupancy models. *Ecosphere*, 8(9), e01944.

Fiske, I., & Chandler, R. (2011). Unmarked: an R package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of statistical software*, *43*, 1-23.

Gallo, T., Fidino, M., Gerber, B., Ahlers, A. A., Angstmann, J. L., Amaya, M., ... & Magle, S. B. (2022). Mammals adjust diel activity across gradients of urbanization. *Elife*, *11*, e74756.

Kalle, R., Ramesh, T., & Downs, C. T. (2018). When and where to move: Dynamic occupancy models explain the range dynamics of a food nomadic bird under climate and land cover change. *Global Change Biology*, *24*(1), e27-e39.

Kays, R., Arbogast, B. S., Baker‐Whatton, M., Beirne, C., Boone, H. M., Bowler, M., ... & Spironello, W. R. (2020). An empirical evaluation of camera trap study design: How many, how long and when?. *Methods in Ecology and Evolution*, *11*(6), 700-713.

Louvrier, J., Duchamp, C., Lauret, V., Marboutin, E., Cubaynes, S., Choquet, R., ... & Gimenez, O. (2018). Mapping and explaining wolf recolonization in France using dynamic occupancy models and opportunistic data. *Ecography*, *41*(4), 647-660.

National Climatic Data Center [NCDC]. 2023. National Oceanic and Atmospheric Administration National Climatic Data Center. <https://www.ncdc.noaa.gov/cdo-web/datatools/findstation> . Accessed 25 June 2023.

Magle, S. B., Fidino, M., Lehrer, E. W., Gallo, T., Mulligan, M. P., Ríos, M. J., ... & Drake, D. (2019). Advancing urban wildlife research through a multi‐city collaboration. *Frontiers in Ecology and the Environment*, *17*(4), 232-239.

Mckann, P. C., Gray, B. R., & Thogmartin, W. E. (2013). Small sample bias in dynamic occupancy models. *The Journal of Wildlife Management*, *77*(1), 172-180.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In Proceedings of the 3rd international workshop on distributed statistical computing. 124(125.10), 1–10.

Royle, J. A., & Dorazio, R. M. (2008). *Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities*. Elsevier.

Tingley, M. W., Ruiz-Gutiérrez, V., Wilkerson, R. L., Howell, C. A., & Siegel, R. B. (2016). Pyrodiversity promotes avian diversity over the decade following forest fire. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1840), 20161703.

Tingley, M. W., Stillman, A. N., Wilkerson, R. L., Howell, C. A., Sawyer, S. C., & Siegel, R. B. (2018). Cross‐scale occupancy dynamics of a postfire specialist in response to variation across a fire regime. *Journal of Animal Ecology*, *87*(5), 1484-1496.

Tingley, M. W., Stillman, A. N., Wilkerson, R. L., Sawyer, S. C., & Siegel, R. B. (2020). Black-backed woodpecker occupancy in burned and beetle-killed forests: disturbance agent matters. *Forest Ecology and Management*, *455*, 117694.

U.S. Census Bureau, 2014–2018 American community survey 3-year public use microdata samples (2021). https://www.census.gov/programs-surveys/acs/data.html. Accessed 25 June 2023.

Zipkin, E. F., Grant, E. H. C., & Fagan, W. F. (2012). Evaluating the predictive abilities of community occupancy models using AUC while accounting for imperfect detection. *Ecological Applications*, *22*(7), 1962-1972.