Estimating species occupancy across multiple sampling seasons with the autoOcc R package

# Introduction

# 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), where data is collected by surveying some number of sites over multiple sampling periods (e.g., years) to collect information on the presence or absence of the species of interest at those sites over time. We assume that while the occupancy status of the 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 a survey at a site. 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 does not false positives (i.e., the species is not present but was mistakenly detected).

Autologistic occupancy models three more assumptions in addition to those listed above,. First, 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 surveys. Second, while autologisistic occupancy models assume spatial independence, it does account for some temporal dependence within the data. More specifically, autologistic occupancy models include a first-order autoregressive term to account for whether the presence of a species in one time period influences the occupancy status in the following time period. However, we still assume independence over larger time frames (e.g., *t – 1* to *t + 1*). This second 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. Third, 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 this class of statistical model is with a latent binary variable that denotes whether the species of interest is present or not at a sampled location. While autoOcc does not use this parameterization to estimate the associated model parameters, this is the simplest 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 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* = 1the latent state model is

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

where is a vector of regression coefficients (including 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. 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.

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 presence. 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 (citation) or JAGS (citation). 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 how the autologistic model is written in matrix notation, 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 time *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 share parameters such that logit(*ηt*) = *β*0 and logit(*ωt*) = *β*0  + θ, both of which could be extended to accommodate coviates. 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 last season 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 *ψ3* 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 survey of season 3, not detected the second survey , and detected again on the third survey.“ Looking at Eq. 6, note that it could be further simplified by factoring out the detection probabilities from either side of the addition/ 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. Borrowing from MacKenzie et al. (2003), 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 to grab the correct detection probabilities. Two examples of these column vectors include

|  |  |  |
| --- | --- | --- |
|  |  | 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. As 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 and are demonstrated in the worked examples below.

## 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 those 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 re-write 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 to make predictions along spatiotemporal gradients.

# Things to consider before using this method

To demonstrate how much data should be collected before using autoOcc and illustrate how it performs to other techniques a researcher may use with similar data I conducted a simulation study to compare the accuracy and precision of autologistic occupancy models to dynamic occupancy models. I chose dynamic occupancy models for comparison instead of other techniques (e.g., a stacked design occupancy model) because dynamic occupancy models do not require random effects to estimate, making them a simpler choice. For each class of model, simulations varied in how common a species was, number of sites sampled, and 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 a given expected occupancy probability. After determining possible solutions, I chose one solution at random. Number of sites ranged from 30 to 100 locations sampled 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.

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 observational model the same environmental covariate was used, but the slope term associated with this covariate was set to 0.5. Finally, I set the detection intercept 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. 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 scenarios, I calculated the relative bias 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 model 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 though and was respectively 51 and 5.4 times smaller than the dynamic colonization and extinction intercepts. When considering different scenarios, 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 colonization slope terms increased 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 3.O). Averaged across scenarios, the autologistic model slope term was 2.67 and 1.56 times narrower than the dynamic model colonization and extinction slope terms, respectively (Figure 4). The largest difference between models was when the expected occupancy of the species was 0.5, especially at small sample sizes.

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 a smaller number of sites were sampled. 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 computer I coded up for these simulations in. As such, the 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 the analysis conducted by Stillman et al. (2023), who used an autologistic occupancy model to analyze survey data for black-backed woodpecker (*Picoides arcticus*) collected throughout the montane forests in California to determine how pyrodiversity, or the spatial and temporal variation in fire characteristics, affects woodpecker occupancy. Across both worked examples I compare the relative fit of different models using AIC (Anderson and Burnham, 2004). Furthermore, I used 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 influence 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.

Save for weekly temperature, all of the other data to conduct this analysis is included within autoOcc, and so our first objective would be to load the detection data and format it for analysis.

# load opossum detection / non-detection data

library(autoOcc)

data("opossum\_det\_hist")

If you looked at this dataset or checked it’s help file (?opossum\_det\_hist) you should notice that it has 6 self-explanatory columns: Site, Season, and Week\_1, Week\_2, Week\_3, and Week\_4. Furthermore, this dataset is in long format and sorted temporally by season and then alphabetically by site. To use this dataset in autoOcc, we first need to convert it to a site by season by survey three-dimensional array. This can be done with the function format\_y(), which requires you to specify which columns denote sites, seasons, and detection data.

# format the detection data as it is currently

# in long format.

opossum\_y <- format\_y(

x = opossum\_det\_hist,

site\_column = "Site",

time\_column = "Season",

history\_columns = "^Week"

)

The output of this function is the three-dimensional array we need for analysis. Following that, we need to collect the site-level covariates as well as the temporally varying temperature data. For the former, you can load the covariates that are included as data within autoOcc. To represent urban intensity I calculated the proportion of impervious cover within 1 km of each sampling location (NLCD 2016 citation). To represent neighborhood wealth I calculated the median per capita income within 1 km of each sampling location from the 2014-2018 American Community Survey (citation).

# load covariates

data("opossum\_covariates")

# subset only impervious and income

opossum\_covariates <- opossum\_covariates[,

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

]

While you will soon see how autoOcc can handle temporally varying covariates, for this analysis the latent state covariates only vary spatially. As such, all we should do with is scale these covariates so that they have unit variance and a mean of zero and provide them as a data.frame to autoOcc. The code below checks if a column in the dataset is numeric, and if so, scales it accordingly, and then returns a data.frame. With only two covariates, this can do done rather simply with the scale() function in R.

# scale covariates, and convert them back to numeric class.

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 covariates do vary temporally, then covariate data must be supplied as a named list object instead of a data.frame. Such a setup should hopefully be familiar to those who have used the unmarked package in the past for their occupancy analyses (unmarked citation).

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

# Caveats

# Additional resources

# Conclusion

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.

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.