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 time to collect information on the presence or absence of the species of interest at those sites. We assume that while the occupancy status of the species may change at sites over time, there are intervals where the occupancy status does not change (i.e., the closure assumption). Thus, if each site is sampled over *T* sampling periods, we assume that the occupancy status can change between sampling periods but not within. 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 accidentally detected).

There are a few other assumptions autologistic occupancy models make. First, we assume all sampled sites are independent. Second, 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. As such, we assume temporal dependence in adjacent sampling periods (e.g., *t – 1* to *t*) but assume independence over larger time frames (e.g., *t – 1* to *t + 1*). This temporal 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, 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, I would argue that 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. There are two ways this can be done, the first solution is generalized while the second solution is simpler. For the first solution, note that our transition probability matrix ***ϕ****t* does not contain any absorbing states (i.e., a state that once entered, cannot be left). As such, ***ϕ****t*  is an irreducible stochastic matrix. Irreducible stochastic matrices have several unique properties but the important one here is that they can be used to estimate a stationary probability vector, ***π***, that describes the long-term distribution of ***ϕ****t* (i.e., the expected proportion of sites occupied and not occupied). To calculate the expected occupancy of our species we can solve the equation *π* ***ϕ****t = π*, where ∑*π = 1*. This can be done by calculating the first left eigenvector of ***ϕ****t* and then normalizing the values (i.e., dividing each element by the sum of all elements; Fidino et al. 2019, Rivera et al. 2022). The normalized left eigenvector is *π* is of length 2, the first element is the expected occupancy probability, and the second element is the complement of the first (i.e., the expected probability sites are not occupied).

While the general solution above works for irreducible stochastic matrices of any size, it is far simpler to derive expected occupancy estimates with the second solution, which we borrow 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 probabilities we estimate from the autologistic occupancy model, we arrive at the following formula:

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

Which provides the same result as the first element of *π* (i.e., the expected occupancy). 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

## Underlying model assumptions

## Sample size considerations

# Worked examples

Bit of text. First about data coming being provided within autoOcc. Camera trap data. Second, a renalaysis of Stillman et al. (2023), that had point data data across FIRE STUFF. Unique because they only visited sampling locations once per year, but split that single visit into multiple secondary sampling periods.

## Virginia opossum throughout Chicago, IL

## Black-backed woodpecker throughout California

# Caveats

# Additional resources

# Conclusion

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.