Estimating species occupancy through time with the `autoOcc` R package.

# Introduction

# Explanation of the method

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 we explain it here to provide intuition into how the autologistic term, *θ*, is used. Thus, for *i* in 1,…,*I* sites and *t* in 1,…,*T* primary sampling periods, let *zi,t* be the occupancy status of a species at site *i* and time *t* and *ψi,t* be the occupancy probability. During the first timestep there is no information about the occupancy status of the species before sampling began. Thus, when t=1the latent state model is

|  |  |  |
| --- | --- | --- |
|  | *, t = 1* | Eq. 1 |

where is a vector of regression coefficients (including the model intercept) and their associated covariates, which are indexed by *i* and *t* because they 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 which estimate initial occupancy in the first timestep. However, autologistic occupancy models use the parameters that describe a species occupancy in the first timestep across all seasons instead of separately 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 timestep.

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

Thus, when the species is present during the previous timestep 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.

For the data model let *p* be the conditional probability of detecting the species during *j* in 1,…,*J* secondary sampling periods given the species presence. Further, let yi,t,j represent the detection / non-detection data for site *i*, primary sampling period *t*, and secondary sampling period *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 then

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

where is a vector of regression coefficients and their associated covariates that may vary across sites, primary, and secondary sampling periods. Given Eq. 1 through Eq. 3, the autologistic parameterization adds one new parameter to the model, *θ*, to account for temporal dependence in species occupancy from one timestep to the next.

Instead of describing this model with an underlying latent variable, which is mostly done in a Bayesian analysis, it is possible to construct the likelihood of observing each datapoint with matrix notation like the dynamic occupancy model developed by MacKenzie et al. (2003). By doing so, the model can be fitted via maximum likelihood. To make it easier to follow along with how the autologistic model is written in matrix notation, let’s assume a site has been sampled across four primary sampling periods with three repeat surveys to generate the detection history [110], [000], [101], [100]. Thus, the species was detected at least once on the first, third, and fourth primary sampling periods and was not detected at all during the second. For this section we need two occupancy probabilities. Thus, let *ψt* represent the occupancy probability at time *t* during either the first sampling period or when the species was not present at time t-1 while *ωt* is the occupancy probability if the species was present in the previous timestep. Following MacKenzie et al. (2003), and dropping the site subscript for simplicity, the probability of observing the first sampling period 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 primary sampling period the species was not detected. That means one of two things occurred. Either the species was not present at *t* = 2 or the species was present but not detected at *t* = 2. Thus, the probability of observing the second sampling period detection history is

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

For *t* = 3, we are uncertain if the species was present in the previous timestep. As such, we use *ψ3* and *ω3* to generate the probability of this survey, multiplying both by the necessary detection probabilities:

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

Note that Eq. 6 could be further simplified by factoring out the detection probabilities from either side of the addition. Finally, in 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 would then be 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, MacKenzie et al. (2003) demonstrate how it is far more convenient to use matrix notation. I will show how that is done here. 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 over time from time *t* - 1 to *t*.

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

Note that rows of ***ϕ****t* denote the occupancy state at *t – 1* while columns are the occupancy state at *t*. For example, *ω*t is the probability the species was present at *t* - 1 and *t*, while 1 – *ψt* is the probability the species was not present at *t* – 1 and *t*. Finally, let *p****y,t*** be a column vector that contains the probability you would observe detection history **y**t,1:J on sampling period *t*, conditional on the occupancy state. As such, *p****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 instead imagine *p****y,t*** as a matrix of column vectors with a number of columns equal to the total number of possible detection histories and **y**t,1:J indexes the appropriate column. Some examples of these column vectors are

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

Psi\_theta 1 – psi\_theta

Psi 1 - psi

## Deriving expected occupancy estimates from this model

# 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