Ecological Modelling 387 (2018) 61–69

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Accounting for misidentification and heterogeneity in occupancy studies using hidden Markov models

Julie Louvrier[a,b,*], Thierry Chambert[a], Eric Marboutin[c], Olivier Gimenez[a]

[a] CEFE, Univ Montpellier, CNRS, Univ Paul Valéry Montpellier 3, EPHE, IRD, Montpellier, France
[b] Office National de la Chasse et de la Faune Sauvage, CNERA prédateurs et animaux déprédateurs, Parc Micropolis, 05000 Gap, France
[c] ONCFS, Gières, France

ARTICLE INFO

Keywords:
Occupancy models
Detection heterogeneity
Species imperfect detection
False-positives
Finite-mixture models

ABSTRACT

Occupancy models allow assessing species occurrence while accounting for imperfect detection. As with any statistical models, occupancy models rely on several assumptions amongst which (i) there should be no unmodelled heterogeneity in the detection probability and (ii) the species should not be detected when absent from a site, in other words there should be no false positives (e.g., due to misidentification). In the real world, these two assumptions are often violated. To date, models accounting simultaneously for both detection heterogeneity and false positives are yet to be developed. Here, we first show how occupancy models with false positives can be formulated as hidden Markov models (HMM). Second, benefiting from the HMM framework flexibility, we extend models with false positives to account for heterogeneity with finite mixtures. First, using simulations, we demonstrate that, as the level of heterogeneity increases, occupancy models accounting for both heterogeneity and misidentification perform better in terms of bias and precision than models accounting for misidentification only. Next, we illustrate the implementation of our new model to a real case study with grey wolves (*Canis lupus*) in France. We demonstrate that heterogeneity in wolf detection (false negatives) is mainly due to a heterogeneous sampling effort across space. In addition to providing a novel modeling formulation, this work illustrates the flexibility of HMM framework to formulate complex ecological models and relax important assumptions that are not always likely to hold. In particular, we show how to decompose the model structure in several simple components, in a way that provides much clearer ecological interpretation.

1. Introduction

Occupancy models (Mackenzie et al., 2006) are commonly used to infer species occurrence while accounting for imperfect detection (Bailey et al., 2014; Guillera-Arroita, 2017). These models rely on species detections and non-detections recorded during surveys repeated across time and across several spatial sampling units (sites). As with any statistical models, inferences made from occupancy analyses heavily rely on several assumptions that should be checked and validated (Mackenzie et al., 2003, 2006), although in reality this is rarely done (see however, Mackenzie et al., 2004; Warton et al., 2017).

Here, we focus on two important assumptions. First, there should be no unmodelled heterogeneity in species detection. In other words, all heterogeneity should be accounted for with covariates. If ignored, heterogeneity in detection leads to underestimating occupancy (Royle and Nichols, 2003; Royle, 2006). Detection heterogeneity can be due to a heterogeneous sampling effort in space (Louvrier et al., 2018), variation in animal abundance (Royle and Nichols, 2003) or variation in site characteristics (Mackenzie et al., 2011). Often, site-level covariates can be measured on the field and incorporated in occupancy models to account for detection heterogeneity. However, unexplained variation may remain or measuring the relevant covariates may simply be impossible in the field. When we suspect substantial unmodelled heterogeneity to occur, we should consider modeling it, either with continuous latent variables (through normally distributed site random effects, e.g. Gimenez et al., 2014). Modelling heterogeneity using normally distributed random effect has long been studied in the field of theoretical biology (e.g., Perc, 2011). Alternatively, modelling heterogeneity can be done using finite mixtures. In finite-mixture models, a latent variable is defined to assign sites to a mixture components (i.e.,

Abbreviations: MMO, model accounting for misidentification only; MMH, model accounting for misidentification and heterogeneity in the detection probabilities; MMS, model accounting for misidentification only and sampling effort as a covariate on the detection probability; MMHS, model accounting for misidentification and heterogeneity in the detection probabilities and sampling effort as a covariate on the detection probabilities

* Corresponding author at: Centre d'Écologie Fonctionnelle et Évolutive, CNRS, 1919 route de Mende, 34090 Montpellier, France.
E-mail address: julie.louvrier@cefe.cnrs.fr (J. Louvrier).

https://doi.org/10.1016/j.ecolmodel.2018.09.002
Received 4 January 2018; Received in revised form 6 August 2018; Accepted 3 September 2018

groups of heterogeneity) characterized by specific parameters (Royle, 2006; Pledger and Phillpot, 2008). While heterogeneity in detection probability using mixture models has been long studied in the capture-recapture (CR) literature (review in Gimenez et al., 2017), less attention has been given to this issue in occupancy models (Gimenez et al., 2014; Miller et al., 2015).

A second important assumption of occupancy models is that the species should not be detected when absent from a site (i.e. no false positives). False positives occur when the species of interest is detected at sites where it did not occur, usually as a result of misidentification (Miller et al., 2013). Several studies have underlined the importance of accounting for false positives on species distribution (Royle and Link, 2006; Miller et al., 2011, 2013; Chambert et al., 2015). Ignoring false positives and counting them as true positives causes important biases, such as overestimating occupancy and colonization probabilities, and underestimating extinction probability (Royle and Link, 2006; Mcclintock et al., 2010). Miller et al. (2011, 2013) developed static and dynamic occupancy models that accommodate both false negatives and false positives. As example of applications, these models have been used to estimate occurrence of amphibians (Miller et al., 2011), bats (Clement et al., 2014), and several large vertebrates in India (Pillay et al., 2014), as well as occurrence dynamics of wolves in Montana (Miller et al., 2013).

While several studies have accounted for heterogeneity in occupancy models with false positives by using site-level covariates (Mcclintock et al., 2010; Ferguson et al., 2015; Miller, 2015), methods that simultaneously account for both unmodelled heterogeneity through finite mixtures and false positives have yet to be developed. Here, we fill this gap and illustrate the use of hidden Markov modelling (HMM) framework as a powerful tool for further developments aiming at relaxing occupancy models' assumptions.

Standard occupancy models can be formulated as HMMs describing two time-series running in parallel. The first time-series captures the dynamics of the latent states with the state process following a Markovian sequence (e.g. site occupied vs. unoccupied); the other time series models the observation process consisting in detections conditional on the underlying but possibly unknown states (Gimenez et al., 2014). The originality of our approach is twofold. First, we show how occupancy models with false positives can be formulated as HMMs. Second, benefiting from the HMM framework flexibility, we extend models with false positives to account for unmodelled heterogeneity using a finite-mixture approach.

To assess the performance of our approach, we performed a simulation study comparing parameter bias and precision in a model accounting for misidentification and heterogeneity vs. a model accounting for misidentification only. To do so, we considered scenarios with an increasing level of heterogeneity in the probability of false positive detection. We also used a case study on the grey wolves' (*Canis lupus*) distribution in France to illustrate implementation of the method in a real-world scenario. Our objectives were (i) to investigate how detection heterogeneity, when ignored, affects the accuracy of occupancy estimation and (ii) assess the extent at which this heterogeneity might be explained by sampling effort variability across space.

2. Methods

In the statistical literature, there are three main problems of interest when using HMM (Rabiner, 1989). In what follows, we review each of these problems in the context of occupancy models. In the *evaluation* problem, we ask what the probability that the observations are generated by our model is – see Section 2.1. In the *decoding* problem, we ask what the most likely state sequence in the model that produced the observations is – see Section 2.5. In the *learning* problem, we ask how we should adjust the model parameters to maximize the likelihood – see Section 2.3.

2.1. HMM formulation of occupancy models with misidentification

Occupancy models can be viewed as HMM whereby the ecological states are considered as partially hidden states, i.e. imperfectly observed (Gimenez et al., 2014). Occupancy models incorporating false positives can also be framed within this approach. The HMM formulation allows flexibility in the model formulation. By decomposing the occupancy approach into simpler steps, the HMM formulation allows better understanding of the ecological and observation processes. To account for false positives, we followed the multi-season dynamic model formulation of Miller et al. (2013). For occupied sites, three observations can be made: (i) an unambiguous detection which is a true detection that has been validated, (ii) an ambiguous detection which is also a true detection but that could not be validated and (iii) no detection. At unoccupied sites, by definition, unambiguous detections cannot occur, thus, only two possible observations can be made: an ambiguous detection, which in this case is a false positive detection due to species misidentification, or no detection. The parameters of interest are ψ_1 the probability of initial occupancy, the probability of local extinction ε and of colonization γ, the probability of correctly detecting the species at an occupied site p_{11}, the probability to falsely detect the species at an unoccupied site p_{10}, and the probability b to classify a true-positive detection as unambiguous (Miller et al., 2011). The specification of a HMM is divided in three steps: the vector of initial state probabilities, the matrix of transition probabilities linking states between successive sampling occasions and the matrix of observation probabilities linking observations and states at a given occasion (Gimenez et al., 2014). We define $z_{i,k}$ the latent state of a site i during the primary occasion (e.g., season or year) k. At the first primary occasion, $k = 1$, a site can only be in one of two states ('unoccupied' or 'occupied'), with probabilities in the vector of initial state probabilities:

$$\Psi = \begin{matrix} \text{unoccupied} & \text{occupied} \\ [1-\psi_1 & \psi_1] \end{matrix}$$

Then, the states are distributed as a first-order Markov chain governed by a transition matrix of the form:

$$T = \begin{matrix} & \text{unoccupied at } k+1 & \text{occupied at } k+1 \\ \text{unoccupied at } k & 1-\gamma & \gamma \\ \text{occupied at } k & \varepsilon & 1-\varepsilon \end{matrix}$$

where rows describe states at occasion k in, and columns describe states at $k + 1$.

Next, we describe the observation process, which is conditional on occupancy states. We define $y_{i,j,k}$ the observation of a site i during the secondary occasion (e.g. visit or survey) j during the primary occasion k. For unoccupied sites, unambiguous detections ($y_{i,j,k} = 1$) do not occur while ambiguous detections ($y_{i,j,k} = 2$) or no detections ($y_{i,j,k} = 0$) may occur. For occupied sites, unambiguous detections, ambiguous detections and no detection can all occur. For the sake of clarity, it is more convenient to write the observation process as the product of two matrices. The first matrix summarizes the detection state process conditional on occupancy state (rows) 'unoccupied' and 'occupied' at k. Columns describe the following intermediate latent detection states: 'no detection', 'true positive detection' and 'false positive detection':

$$P = \begin{matrix} & \text{no detection} & \text{true positive detection} & \text{false positive detection} \\ \text{unoccupied} & 1-p_{10} & 0 & p_{10} \\ \text{occupied} & 1-p_{11} & p_{11} & 0 \end{matrix}$$

It is important to keep in mind that the true, underlying state (i.e., false or true positive) of the detections is unknown. At this stage of the modeling, we are still dealing with latent state, not with actual data. The second matrix then summarizes the classification of a true-positive detection as unambiguous or ambiguous, with probability b and 1-b, respectively. In this matrix, rows represent the above intermediate