Original Articles

Prior specification in Bayesian occupancy modelling improves analysis of species occurrence data

Charlotte L. Outhwaite[a,b,c,*], Richard E. Chandler[d], Gary D. Powney[a], Ben Collen[b], Richard D. Gregory[b,c], Nick J.B. Isaac[a,b]

[a] Centre for Ecology & Hydrology, Maclean Building, Benson Lane, Wallingford, Oxfordshire OX10 8BB, UK
[b] Centre for Biodiversity and Environment Research, University College London, Gower Street, London WC1E 6BT, UK
[c] RSPB Centre for Conservation Science, RSPB, The Lodge, Sandy, Bedfordshire SG19 2DL, UK
[d] Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK

ABSTRACT

Multi-species biodiversity indicators are increasingly used to assess progress towards the 2020 'Aichi' targets of the Convention on Biological Diversity. However, most multi-species indicators are biased towards a few well-studied taxa for which suitable abundance data are available. Consequently, many taxonomic groups are poorly represented in current measures of biodiversity change, particularly invertebrates. Alternative data sources, including opportunistic occurrence data, when analysed appropriately, can provide robust estimates of occurrence over time and increase the taxonomic coverage of such measures of population change. Occupancy modelling has been shown to produce robust estimates of species occurrence and trends through time. So far, this approach has concentrated on well-recorded taxa and performs poorly where recording intensity is low. Here, we show that the use of weakly informative priors in a Bayesian occupancy model framework greatly improves the precision of occurrence estimates associated with current model formulations when analysing low-intensity occurrence data, although estimated trends can be sensitive to the choice of prior when data are extremely sparse at either end of the recording period. Specifically, three variations of a Bayesian occupancy model, each with a different focus on information sharing among years, were compared using British ant data from the Bees, Wasps and Ants Recording Society and tested in a simulation experiment. Overall, the random walk model, which allows the sharing of information between the current and previous year, showed improved precision and low bias when estimating species occurrence and trends. The use of the model formulation described here will enable a greater range of datasets to be analysed, covering more taxa, which will significantly increase taxonomic representation of measures of biodiversity change.

1. Introduction

Targets to stem the loss of biodiversity have been in place globally since 2002 when the Convention on Biological Diversity (CBD) agreed the goal for signatory parties to "significantly reduce the rate of biodiversity loss by 2010". The recognised failure to meet this target was followed by the development of the 'Aichi' targets for 2020 (Convention on Biological Diversity, 2010). The new targets focussed on different facets of biodiversity loss, both direct and indirect, such as awareness of biodiversity, the causes of loss, sustainable land management, the pressures on biodiversity, and the benefits gained from it. To monitor progress towards these goals, a set of biodiversity indicators have been developed to track change in measures related, either directly or indirectly, to these elements (Butchart et al., 2010; Tittensor et al., 2014).

Biodiversity research has, therefore, increasingly focussed on the development of tools to produce robust measures of biodiversity change to accurately measure progress towards these targets (Buckland et al., 2005; Gregory et al., 2005).

Species based indicators are the primary means of monitoring change in the state of biodiversity over time. Several indicators of population change have been developed, at various scales and taxonomic coverage. The Living Planet Index (LPI) is a multi-species indicator that was developed to monitor change in vertebrate abundance at a global scale (Collen et al., 2009). Other examples include the wild bird indicator for Europe (Gregory et al., 2005) and the recent development of butterfly indicators for the UK and Europe (Brereton et al., 2010; Van Swaay et al., 2015). The lack of taxonomic representation is primarily due to dependence on the availability of abundance data from large-

scale structured monitoring schemes. In Europe and North America this is limited to birds, mammals, butterflies and moths: elsewhere such schemes are rare. Consequently, current biodiversity indicators are taxonomically biased towards these groups and their ability to act as surrogates of wider biodiversity has been questioned (Rodrigues and Brooks, 2007; Westgate et al., 2014) but rarely evaluated. The lack of taxonomic representativeness is an ongoing problem. If the goal is to understand how biodiversity is changing as a whole, it is important that all groups are represented where possible when such metrics are produced.

One way to achieve greater representation is to use occurrence data for those taxonomic groups that lack abundance data. Occurrence data are "presence-only" records of a species at a known time and location. By also using occurrence data to measure change in biodiversity, it is possible to broaden taxonomic coverage of biodiversity metrics and improve understanding of biodiversity change. For example, the "Dutch LPI" (van Strien et al., 2016) utilises distributional data on dragonflies, fish, mammals, amphibians and butterfly species alongside abundance data. The UK Priority Species Indicator (PSI) uses occurrence data to assess species status for whom abundance data are not available (Eaton et al., 2015; Outhwaite et al., 2015). These indicators have taken advantage of the development of occupancy modelling to incorporate species occurrence data into their assessments (Isaac et al., 2014; Kéry et al., 2010; van Strien et al., 2013).

To date, most applications of occupancy modelling to occurrence data has been limited to well-recorded taxa, such as butterflies (van Strien et al., 2013), dragonflies (Powney et al., 2015; Termaat et al., 2015; van Strien et al., 2010) and birds (Kamp et al., 2016; Kéry et al., 2010). An exception was the 2015 UK PSI, which used a Bayesian occupancy model framework to analyse a range of taxonomic groups (hymenoptera, bryophytes, carabids, odonata, fish, moths, orthoptera and soldierflies), few of which can be considered to be well-recorded (Isaac and Pocock, 2015). Without occurrence data as an alternative data source, no information would be available for the vast majority of the UK "priority species".

The specific occupancy model used for the PSI was that tested by Isaac et al. (2014): they reported high power for estimating species trends compared with alternative methods, and low type I error rates. Isaac et al. explored "high", "medium" and "low" levels of recording intensity, benchmarked against UK and Dutch occurrence datasets. However, occupancy model outputs were useable for only 20% of UK priority species, as most are in taxonomic groups with low levels of recording (Outhwaite et al., 2015). Although occurrence data is a vast resource, particularly in Europe and North America, the availability and coverage varies hugely (Meyer et al., 2015). To date, no research has examined the formulation of occupancy models for use with low recording intensity data. To make the most of the occurrence data available the modelling techniques used must be appropriate for the data available and produce outputs with a high precision where possible. The use of current methods on low-intensity data, as shown by the UK PSI, has revealed the need for improvement if they are to be more widely applicable.

One reason for the restricted applicability of current occupancy models is a lack of realistic year-on-year variation in their representation of species occupancy. For example, the model formulation of Isaac et al. (2014) specifies that the occupancy probability of a site is independent from one year to the next. In reality however, for many species the occupancy probabilities in successive years will tend to be similar, with the degree of similarity depending on the species' ecology. This insight can be exploited to constrain the results of an occupancy analysis in a principled manner, with the constraints providing crucial additional "information" that extends the applicability of such techniques to much sparser data sets than has previously been possible. In practice, the constraints are specified using carefully constructed prior distributions within a Bayesian framework. The use of informative and biologically plausible prior distributions can increase confidence in estimates produced from ecological studies (McCarthy and Masters, 2005), although care is required to ensure that the priors do not influence the results unduly. This approach has not been used previously in occupancy trend estimation.

Here, we use the occupancy model framework tested by Isaac et al. (2014) as a base against which to compare alternative specifications that differ in how information about the occupancy state is shared among years. The aim is to determine whether the additional information from sharing across years can advance current modelling practice to improve precision and reduce bias of trend estimates from datasets with low recording intensity. Specifically, we ask: (1) can alternative prior formulations in a Bayesian occupancy model framework improve the precision of species annual occurrence estimates? (2) Do these alternative formulations increase the precision and reduce the bias of species trend estimates compared to the original method tested by Isaac et al. (2014)? If a more appropriate formulation of this occupancy based method can be determined, it will extend the range of taxonomic groups to which occupancy models can be reasonably applied, thus contributing to broadening knowledge on biodiversity status.

2. Methods

Occupancy models are designed for the analysis of 'presence-absence' data from a collection of sites over time: an occupancy dataset for a particular species typically consists of a set of binary values $\{y_{itv}\}$ say, where y_{itv} takes the value 1 if the focal species was observed at visit v to site i in year t and 0 otherwise. These elements may be supplemented by other variables, such as sampling effort, or weather, that are potentially related to the probability of observing a species if it is present. To determine whether the use of occupancy models could be improved for the analysis of low recording intensity occurrence data we compared two variants of the Bayesian modelling framework tested by Isaac et al. (2014) to the original model formulation used by those authors (hereafter the 'base model'). We compared model variants using data for ants in Great Britain, for whom the data available is similar to the low recording intensity simulated by Isaac et al. (2014). We also tested the model variants in a simulation experiment to compare their performance with respect to the precision and bias of trend estimates.

2.1. The base model

The base model of Isaac et al. (2014) is split into two distinct sub models: a state model and an observation model. The "closure period" (the temporal precision of the state model) is one year; the observation model estimates the probability of detection based on repeat visits within years. The per visit detection history of each species is inferred from records of other species in the assemblage.

The state model, as defined by Eqs. (1) and (2), describes the true occupancy state, z_{it}, of site i in year t. This will be 1 if occupied or 0 if unoccupied. Let ψ_{it} denote the probability that the site is occupied. Then z_{it} has a Bernoulli distribution:

$$z_{it} \sim \text{Bernoulli}(\psi_{it}). \tag{1}$$

Then occupancy probability varies with site and year:

$$\text{logit}(\psi_{it}) = \log\left(\frac{\psi_{it}}{1-\psi_{it}}\right) = b_t + u_i, \tag{2}$$

where b_t and u_i are referred to as a 'year effect' and 'site effect' respectively (more details in Eqs. (5) and (6), below).

Next, the observation submodel describes how the data were generated. Let p_{itv} denote the probability that a species will be observed on a single visit, given that it is present at the site ($z_{it} = 1$). Then the observation parameter y_{itv} is itself a Bernoulli variable, with conditional distribution modelled as: