England Biodiversity Indicators 2020

This documents supports

4b: Status of UK priority species – Distribution
10: Status of Pollinating Insects

Technical background document: Deriving Indicators from Occupancy Models

Prepared[1] by Nick Isaac, Gary Powney, Jack Hatfield, Tom August, Charlotte Outhwaite, Stephen Freeman; Biological Records Centre, Centre for Ecology and Hydrology

For further information on the England Biodiversity Indicators visit
https://www.gov.uk/government/statistics/england-biodiversity-indicators

[1] NB some text re-used from 2013 technical document prepared by the Species Indicator Initiative working group and we wish to acknowledge the input from the authors of that original document.

Introduction

Recent studies have highlighted the value of Bayesian occupancy models for estimating species' status and trends from unstructured biological records (van Strien *et al*. 2013; Isaac *et al*., 2014). Based on these developments, it is now possible to produce biodiversity indicators from biological records, using occupancy models. However, the process of fitting occupancy models and generating indicators raises a number of technical, statistical and conceptual challenges. This document describes these challenges, and how they have been addressed for the 2019 biodiversity indicator set. The procedure described here represents the methodology for the C4b indicator: the status of priority species: distribution http://jncc.defra.gov.uk/page-6850. Additionally, the occupancy model described in section 2 below was used to produce the occupancy trends for D1c indicator: pollinating insects http://jncc.defra.gov.uk/page-6851.

Overview

The workflow for producing the indicators has 6 steps:

1. Data preparation
2. Estimating occupancy for each species
3. Filtering out reliable occupancy estimates
4. Creating the composite indicator
5. Assessing the indicator line
6. Species-specific trends

The entire procedure is handled by a pair of R packages developed at the Biological Records Centre (BRC). Steps 1–2 are implemented using an R package known as Sparta (https://github.com/BiologicalRecordsCentre/sparta); Steps 3–6 are implemented in an R package known as BRCIndicators (https://github.com/BiologicalRecordsCentre/BRCindicators). Each step is described in detail in the sections below.

Data preparation

In theory, any data on the National Biodiversity Network Gateway (NBN Atlas from April 2017) would be suitable for analysis in our occupancy model framework. In reality, the source data used is restricted to the verified records provided to the Biological Records Centre by national recording schemes and societies.

Biological records are first gathered into data-sets comprising species records that are typically co-recorded as an assemblage. For example, the pollinating insects' indicator is based on 2 data-sets: one comprising bee records from the Bees, Wasps & Ants Recording Scheme, and another comprising hoverfly records from the Hoverfly Recording Scheme.

The occupancy model used operates with a spatial resolution of 1km^2 and a temporal resolution of one day. Thus, all records with coarser resolution (in space or time) are excluded at this point. The remaining records are then converted into a set of 'site visits', defined as unique combinations of date and 1 km^2. The number of species recorded on each visit, known as the list length, is then calculated. Note that the list length includes all species in the assemblage, not merely the subset for which an indicator is to be calculated.

Estimating occupancy for each species

A Bayesian occupancy-detection model is fitted for each species. The name 'occupancy-detection' reflects the use of 2 hierarchically coupled sub-models: one model's occupancy (i.e. presence versus absence of each site-year combination), and the other model's detection (i.e. recorded versus not-recorded on each visit). Since true occupancy is

unknown, this form of occupancy-detection model is of a class of statistical models known as 'hidden process' or 'state space' models. The approach is based on theory mathematically similar to that underlying long-established methods of capture-recapture analysis, where replicated visits within a season are used to estimate the probabilities of detection and occupancy (MacKenzie 2006; van Strien *et al*., 2013). The 'season' (also known as the closure period) in our models is the year (i.e. occupancy for each year, using all the recording visits that took place during that year, is estimated).

For each species, the vector of observations at a site is made up of '1's (the species was recorded) and '0's (the species was not recorded, but other species in the assemblage were recorded – a pseudo-absence). The number of observations is equal to the number of 'site visits' defined above.

The specific implementation employed for the biodiversity indicators is an adaptation of the model tested by Isaac *et al.* (2014), in which sites visited in a single year were excluded. The 'list length', defined above, is used as an estimate of sampling effort and accounted for in the detection model. Generally detectability of the average species is expected to be to be lower on shorter lists, therefore visits were grouped into one of 3 categories based on the number of species recorded as follows: 1) a single species recorded, 2) short-day lists, 2 or 3 species recorded (*DT2*), and 3) visits with greater than 3 species recorded (*DT3*) (van Strien *et al.* 2013). This provided a categorical formulation of list length. See Box 1 for further details. Adaptations to this model involved the use of a random walk prior on the year effect of the state model which enables the estimation of occupancy for low-recording intensity data (Outhwaite *et al.*, 2018).

Box 1: Technical description of the Occupancy-Detection model used in the creation of biodiversity indicators.

State model: $z_{it} \sim \text{Bernoulli}(\psi_{it})$; $\text{logit}(\psi_{it}) = b_t + u_i$
Observation model: $y_{itv}|z_{it} \sim \text{Bernoulli}(z_{it} * p_{itv})$; $\text{logit}(p_{itv}) = a_t + \delta_1.DT2_{itv} + \delta_2.DT3_{itv}$

z_{it} = True (unknown) occupancy of site *i* in year *t*. Can be a 1 or 0 (present or absent).
ψ_{it} = The probability that site *i* is occupied in year *t*
b_t = Year effect (categorical)
u_i = Site effect (categorical)
y_{itv} = Observations (detected/not detected) at site (i) at year (t) on visit (v)
p_{itv} = The probability of detection at site *i* in year *t* on visit *v*, conditional on Z_{it} that is the species true presence or absence.
a_t = Year level random effect (normally distributed)
δ_1 = The effect of list length category 2 (*DT2*), relative to category 1.
δ_2 = The effect of list length category 3 (*DT*3), relative to category 1.

Models were fitted in Bayesian state space using the program JAGS (Just Another Gibbs Sampler: Plummer 2003) and run for a minimum of 20,000 iterations with 3 Markov chains, a burn-in of 10,000 and a thinning rate of 3. This means that each index acquires a vector of (20,000 – 10,000) * 3/3 = 10,000 elements representing random draws from its posterior distribution. Uninformative priors were used on most parameters: this means that our models have no expectation of whether any occupancy estimate will be closer to zero or one. Specifically, prior distributions on the logit (occupancy) estimates were set to have a mean of zero and standard deviation of 30. The prior on the year effect of the state model was represented by a random walk where estimates of the year effect of interest were a combination of the previous year plus or minus some variation around this (Outhwaite *et al.*, 2018).

The principal output from each species' model is the estimated occupancy for each year (i.e. the proportion of sites occupied). Since the model is fitted in a Bayesian framework, the annual occupancy estimates are expressed as a distribution, from which point estimates can be derived, rather than as classical maximum likelihood estimates. This distribution expresses all the forms of uncertainty that are captured by the model (including uncertainty about the data collection process), and forms the basis of indicator generation in subsequent steps. Specifically, the median of the distribution is taken as an estimate of the species' annual index value, and the standard deviation of the distribution defines the uncertainty of that index. This uncertainty is critical for subsequent steps in the construction of the indicator.

Filtering out reliable occupancy estimates

Not all model outputs can be considered reliable. A set of rules based on method exploration and testing were therefore employed to determine which index values were suitable for inclusion in a composite indicator. Species with less than 50 total records were excluded from the composite indicator, effectively removing the rare species for which trends in distribution could not be reliability estimated. To further improve reliability, species with a gap in records greater than 10 consecutive years were excluded. In addition, to ensure a reasonable time series was available, species required at least 10 years of occupancy estimates for inclusion.

The largest change in methodology compared to previous versions of the composite indicator was the clipping of the species time-series. Firstly, all species had their occupancy estimates restricted to the time frame of the available record data for their group. For example, all lichen species estimates are restricted to the period 1970 to 2015 as no lichen records are yet available for 2016. In addition, each individual species' occupancy estimates were clipped to after the first detection for the species in question. This change to the methodology was added to take account for the spatial expansion of many of the recording schemes and sampling of new sites.

Creating a composite indicator

Most species-based biodiversity indicators calculate the composite index as the geometric mean of indices for those species that contribute data in that year, relative to a value of 100 in the starting year.

$$I_t = 100.(\Pi\{x_{1,t}, x_{2,t} .. x_{n,t}\})^{1/n} / I_1 \qquad \text{Equation 1}$$

Where I_t is the value of the indicator in year t and $x_{1,t}$ is the abundance index for species 1 in year t. Under this approach, the proportional change in the indicator from one year to the next, Δ_t, is mathematically equivalent to the geometric mean growth rate from years t-1 to t.

$$\Delta_t = I_t / I_{(t\text{-}1)} = (\Pi\{\lambda_{1,t}, \lambda_{2,t} .. \lambda_{n,t}\})^{1/n} \qquad \text{Equation 2}$$

$$\lambda_{i,t} = x_{i,t} / x_{i,(t\text{-}1)} \qquad \text{Equation 3}$$

Where $\lambda_{i,t}$ is the growth rate for species i from year t-1 to t, and n is the number of species with index values in both year t-1 and t.

The geometric mean is appropriate for indices based on abundance data, which is bounded at zero but unbounded above. However, occupancy estimates are bounded at both zero and one (a species cannot occupy more than 100% of available sites). To retain this property in the indicator, the arithmetic mean of the change in log odds was selected as an appropriate statistic, thus:

$$\Delta_t = I_t/I_{(t-1)} = \Sigma\{\gamma_{1,t}, \gamma_{2,t} .. \gamma_{n,t}\}/n \qquad \text{Equation 4}$$

$$\gamma_{,t} = \log(p_{i,t}/(1-p_{i,t})) - \log(p_{i,(t-1)}/(1-p_{i,(t-1)})) \qquad \text{Equation 5}$$

Where $p_{i,t}$ is the proportion of occupied sites for species i in year *t,* and $\gamma_{i,t}$ is the log of the growth rate in the odds of the average site being occupied by species *i* between years *t*-1 and *t*. Following convention, the headline indicator is set to start at 100 with a lower bound of zero:

$$I_t = 100 * \exp(\Sigma\{\Delta_1, \Delta_2 .. \Delta_t\}) \qquad \text{Equation 6}$$

Reformulating the composite indicator in terms of growth rates has 2 distinct advantages over the conventional approach to constructing indicators. First, it means that the categorisation of species as 'increasing' or 'decreasing' can be made from the same set of data (the growth rates) as the construction of the headline indicator. Second, it provides an elegant solution to the problem of species that join the indicator after the first year (i.e. where the first year is unreliable): other indicators typically adopt a complicated rescaling approach to ensure that species entering the indicator after the first year do not bias the overall assessment. It also makes a simple and robust, though untestable, assumption about species that drop out of the indicator prior to the final year: specifically it assumes that their fluctuations are the same, in aggregate, as those of the species that remain in the indicator. By contrast, the abundance-based priority species indicator C4a assumes that species dropping out of the indicator remain at constant abundance when no data are available (Eaton *et al*., 2015).

Assessing the indicator line

As noted above, the index values are not classical point estimates, but rather derived from a posterior distribution of values. This distribution makes it possible to incorporate uncertainty in the annual occupancy estimates formally around the indicator line. Since each γ in equation 4 has a distribution of 10,000 values, it is trivial to calculate Δ_t and I_t as a distribution, the quantiles of which measure the credible intervals of the indicator. Since this approach leads to full propagation of uncertainty from the results of each species' model, the magnitude of uncertainty around the indicator line is expected to be larger than for other indicators, where the index values are assumed to be known without error and the uncertainty is estimated via bootstrapping.

In Bayesian statistics, the magnitude of uncertainty around parameter estimates is generally referred to as the 'credible intervals', as opposed to the 'confidence intervals' derived from classical frequentist statistics. For many applications, the Bayesian credible intervals and Frequentist confidence intervals are very similar. However, the interpretation is quite different, reflecting differences in the underlying philosophy of the 2 statistical paradigms. In Frequentist statistics, the 95% confidence intervals suggest that, if an experiment were repeated many times, the true value of the parameter (e.g. the index in the most recent year) would fall within the intervals 95% of the time. In other words, the uncertainty is an expression about the data collection process, and the parameter is assumed to be fixed. By contrast, the Bayesian approach treat the data (i.e. the species' data) as fixed and expresses uncertainty in terms of the parameter being estimated (whilst accounting for uncertainty in the data). The credible intervals around a Bayesian indicator reflect the probability that the indicator value lies within those intervals.

The 90% credible intervals were chosen for making the short- and long-term assessment of a trend in the indicator line. Thus, if the upper limit of the 90% credible interval falls below 100, this gives at least 95% probability (not 90%), based on the posterior distribution, of the

index having declined. Thus, the long-term assessment is a simple test of whether the value 100 lies inside or outside the 90% credible intervals for the focal year. Similarly, the short-term assessment tests whether the median value in year *t*-5 lies within the 90% credible intervals for the focal year, *t.*

Species-specific trends

Since γ is the log of the growth rate (on the odds scale), the mean growth rate across years for each species, Λ, can be calculated as follows:

$$\Lambda_i = \exp(\Sigma\{\gamma_{i,1} , \gamma_{i,2} .. \gamma_{i,t}\} / t) \qquad \text{Equation 7}$$

Species were grouped into 5 categories: 'strong increase', 'increase', 'no change', 'decrease' and 'strong decrease', based on the value of Λ for that species over both the long term (all years) and the short term (the most recent 5 years). The thresholds defined in the wild bird indicators have been adopted. Specifically, a strong increase is defined as a mean increase of at least 2.81% per annum, which is equivalent to a doubling over 25 years; a strong decrease is defined as a mean decrease of at least 2.73% per annum, which is equivalent to a halving over 25 years. This assessment is based on the best estimate of $\gamma_{i,t}$ (i.e. it uses the mean value of the distribution and does incorporate the uncertainty).

References

Brooks, S.P., & Gelman, A. (1998) General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, **7**, 434–455.

Eaton, M.A., Burns, F., Isaac, N.J.B., Gregory, R.D., August, T.A., Barlow, K.E., Brereton, T., Brooks, D.R., Al Fulaij, N., Haysom, K.A., Noble, D.G., Outhwaite, C., Powney, G.D., Procter, D. & Williams, J. (2015). The priority species indicator: measuring the trends in threatened species in the UK. *Biodiversity*, **16**, 108–119.

Isaac, N.J.B., van Strien, A.J., August, T.A., de Zeeuw, M.P., & Roy, D.B. (2014) Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution*, **5**, 1052–1060.

Kéry, M., & Schaub, M. (2011) *Bayesian Population Analysis using WinBUGS: a hierarchical perspective*. Academic Press, Amsterdam

King, R., Morgan, B.J.T., Gimenez, O. & Brooks, S.P. (2010) *Bayesian analysis for population ecology*. CRC Press, Boca Raton, USA.

MacKenzie, D.I. (2006). *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence*. Academic Press. Burlington, USA.

Outhwaite, C.L., Chandler, R.E., Powney, G.D., Collen, B., Gregory, R.D. & Isaac, N.J.B. (2018). Prior specification in Bayesian occupancy modelling improves analysis of species occurrence data. *Ecological Indicators*, **93**, 333–343.

Plummer, M. (2003) *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. http://mcmc-jags.sourceforge.net/.

van Strien, A.J., van Swaay, C.A.M., & Termaat, T. (2013) Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. *Journal of Applied Ecology*, **50**, 1450–1458.