Developing a Bayesian species occupancy/abundance indicator for the UK National Plant Monitoring Scheme

O.L. Pescott, CEH Wallingford (olipes@ceh.ac.uk), G.D. Powney (CEH Wallingford) & K.J. Walker (Botanical Society of Britain and Ireland)

Summary

- The National Plant Monitoring Scheme (NPMS) is a volunteer-based structured plant recording scheme. This report focuses on the development of a new statistical model for the species-level data generated by the NPMS. The aim is ultimately for this to contribute to a new indicator of UK habitat quality.
- NPMS surveyors collect data on plant abundance (percentage covers) from small plots targeted at specific habitats. They can participate at different levels, with the level of participation influencing the list of species sought in the field. Typically, surveyors record around 5 small plots in a 1 km square, with each plot being visited twice a year.
- NPMS data must be processed in order to accurately represent the information content of the plot surveys. Because surveyors use different lists of species depending on their level, in some cases we need to distinguish between true absences (species on a surveyor's target list but not reported) and unknown cases (species not on a surveyor's target list, meaning that absence from a list is not informative).
- We present a novel hierarchical statistical model for NPMS species-level data. This model seeks to make maximum use of the data collected, and integrates a standard occupancy modelling approach for plot detections with a Beta distribution model for a species' non-zero cover data.
- We evaluate the proposed model using a variety of different simulated datasets. The performance of the model is assessed in relation to the bias and variance shown relative to the actual parameters used in the data simulations.
- The simulations indicate that the model performs as expected under a "perfect" scenario. Smaller datasets induce various biases, many of which can be traced to the fact that, in our simulations, abundance and detectability are closely related. This biases the estimated mean of the underlying cover distribution upwards, and also impacts estimates of the intercept and regression coefficient in the detection sub-model. In real datasets this relationship would likely be less clear-cut, and we do not expect these biases to affect species' relative annual trend estimates.
- Finally, we apply the model to NPMS data collected between 2015 and 2018 for 86 grassland species. The model estimates ecologically sensible mean cover values for the species analysed. However, mean plot occupancies tended to centre on 0.5, suggesting that many species may not yet have sufficient data for mean occupancy to be well estimated.
- A novel combined abundance/occupancy indicator has been developed for NPMS data in a Bayesian framework. The simulation tests and applications to real data explored in this report indicate that the model performs well in ideal scenarios; biases in less data-rich scenarios can largely be explained by relationships between abundance and detectability. These are likely to be less clear-cut in real datasets, and future work will explore how additional covariates describing a species' detectability could be incorporated. Extending the model to create annual indices, and considering how these may be aggregated, will also be required for the future creation of a habitat quality indicator using NPMS data.

Table of contents

1. Introduction

The National Plant Monitoring Scheme (NPMS) is a volunteer-based structured plant recording scheme. This report focuses on the development of a new statistical model for the species-level data generated by the NPMS. The aim is ultimately for this model to contribute to a new indicator of UK habitat quality.

The National Plant Monitoring Scheme (NPMS; www.npms.org.uk) is a volunteer-based, habitat and plant monitoring scheme launched in 2015 (Walker et al., 2015). It was conceived as a national sample of high quality semi-natural habitats, but one that was straightforward enough in design to be appealing to volunteer botanical recorders. Comprehensive information on the founding motivations and design of the scheme is available in several reports and papers (Pescott et al., 2019, 2016, 2015, 2014; Walker et al., 2015, 2010), and we do not review all of this information here.

This report deals with the development of an analytical approach designed to produce species-level trend lines from proportional cover and occupancy data collected through the NPMS. The two main topics covered are data preparation (i.e. processing raw data collected by NPMS volunteers into a structure suitable for the proposed model), and the modelling approach itself. The data processing steps and model presented here do not represent the only possible approaches, but are an attempt to extract the maximum amount of information from NPMS data, based on knowledge of the design of the scheme. Depending on the amount of information on real world states and processes actually present in the dataset (rather than posited to be present in advance of any actual data inspection or analysis), the approach here may be simplified, or indeed expanded, in the future.

2. The NPMS sampling protocol and dataset

NPMS surveyors collect data on plant abundance (percentage covers) from small plots targeted at specific habitats. They can participate at different levels, with the level of participation influencing the list of species sought. Typically, surveyors record around 5 small plots in a 1 km square, with each plot being visited twice a year.

The core aim of the NPMS is to sample plant communities within habitats of conservation value using small plots. Volunteers are assigned a 1 km square within which such plots are established. Ideally these are visited twice a year, every year, although it is acknowledged that in reality the frequency of visits may be less than this, either because a square may be in a remote location, or because volunteers rotate 1 km squares between years to introduce additional novelty to their survey activities, and to reduce pressure on sensitive habitats. Figure 1 is an overview of the NPMS sampling process; much more detail can be found in Pescott et al. (2019).

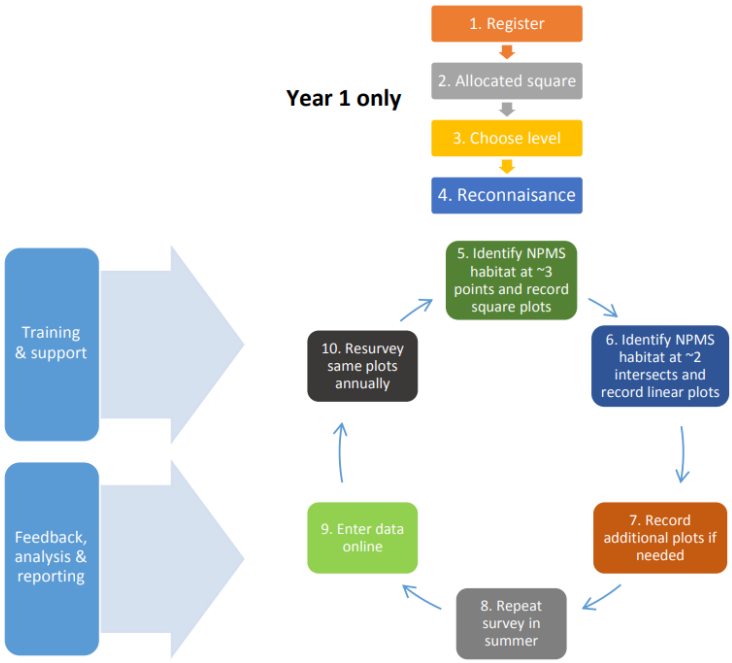


Figure 1. A schematic of the NPMS survey process from the volunteer perspective.

The key steps in Figure 1 are (2) Volunteer allocated square, (3) Choose level, (5) Identify NPMS habitat at ~3 points and record square plots, and (6) Identify NPMS habitat at ~2 intersects and record linear plots. This captures the process whereby volunteers chose an available 1 km square and attempt to set-up several small square and/or linear plots within that square (Walker et al., 2015). The process of choosing plots is initially governed by a structured process designed to minimise the various biases involved in giving surveyors a completely free choice (Pescott et al., 2019). The below figure demonstrates the structure of this plot selection process (Figure 2).

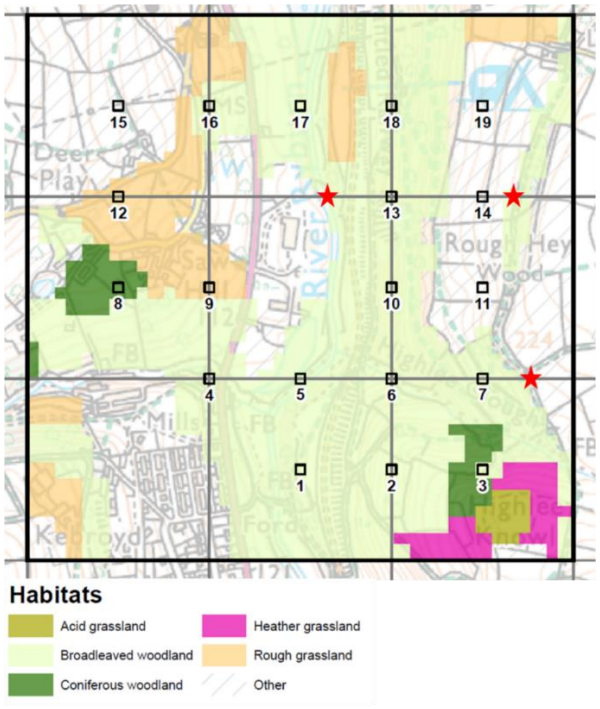


Figure 2. An example of the type of 1 km square map that a surveyor receives with their survey pack. The gridded numbered squares and gridlines are intended to reduce bias in plot placement. See the main text for more information.

The small numbered squares in Figure 2 are laid out according to a 5 x 5 grid, although squares are omitted if they intersect urban, suburban or improved grassland land cover types as defined in the CEH 2007 Land Cover Map (Morton et al., 2011; Pescott et al., 2019); these squares are the unbiased candidate locations for volunteer plot surveys. The four grid lines dissecting the 1 km square into ninths are an unbiased approach to indicating potential locations for volunteers to set up linear plots (the gridlines are unbiased with respect to the underlying land surface). The red starred locations along these gridlines (Fig. 2) indicate potential survey points along linear habitat features, such as hedgerows or arable field margins. Note however, that the NPMS also allows for the self-selection of plots in order to account for the eventuality that these pre-selected locations are all unavailable, or do not coincide with NPMS habitats.

Once selected, volunteers identify the habitat types within their plots. As a result of pre-launch consultations and field trials with volunteers, a two-tiered approach to habitat identification was developed (Pescott et al., 2019). The main reason for this was the lack of confidence of some volunteers when identifying the "fine-scale" habitat types developed for the NPMS. This two-tiered approach means that groups of associated fine-scale habitats are grouped into broad-scale categories, and volunteers then have the choice between recording at each of the two levels for any given plot. The broad- and fine-scale habitat types are shown in the table below (Table 1; Walker et al., 2015).

Table 1. Broad and fine-scale NPMS habitat categories with the associated numbers of wildflower and indicator species (at the broad scale). Reproduced from Walker et al. (2015).

Broad category	Fine-scale habitat(s) included	Wildflower	Indicator
Arable field margins	Arable field margins	15	30
Bog & wet heath	Blanket bog; raised bog; wet heath	31	53
Broadleaved woodland	Dry deciduous woodland; hedgerows of native species; wet woodland	49	75
Coast	Coastal saltmarsh; coastal sand-dunes; coastal vegetated shingle; machair; maritime cliff-tops and slopes	65	110
Freshwater	Nutrient-poor lakes and ponds; nutrient-rich lakes and ponds; rivers and streams	29	56
Heathland	Dry heathland; dry montane heathland	28	48
Lowland grassland	Dry acid grassland; dry calcareous grassland; neutral damp grassland; neutral pastures and meadows	62	98
Marsh & fen	Acid fens, flushes, mires and springs; base-rich fens, flushes, mires and springs	33	51
Upland grassland	Montane acid grassland; montane calcareous grassland	31	53
Native pinewood & juniper scrub	Conifer woods and juniper scrub	21	29
Rock outcrops, cliffs & screes	Inland rocks and scree; montane rocks and scree	34	52

All fine-scale NPMS habitats have associated positive and negative indicator species; positive indicators are taken to indicate a higher quality, or a more typical composition, of a habitat, whereas negative indicators are taken as signalling some type of decline in habitat quality. Counts of such indicators are shown in the above table (Table 1) for each broad-scale habitat category. As well as choosing the level at which habitats are discriminated (broad or fine), volunteers also choose their general level of participation in terms of the plant identification challenge. These levels are named Wildflower, Indicator and Inventory (these roughly coincide with beginner, improver and expert levels of identification ability, although even a "beginner" should be reasonably confident identifying the set of plants listed at that level). An Inventory-level recorder should be recording all species in a plot (regardless of whether flowering or not), although of course some level of error is inevitable in some habitats (e.g. recently mown or heavily grazed grassland), as is often the case for professional surveys (Morrison, 2016; Scott et al., 2008). Wildflower- and Indicator-level surveyors record a

specific set of indicator species linked to a habitat; Indicator-level recorders search from a longer list of species than those recording at the Wildflower level (see Table 1 above). The full list of indicator species, with their habitat affiliations, can be viewed at the NPMS website here; they are also illustrated in the NPMS species identification guide to be found here. The methodology behind the selection of the final set of habitat indicator species can be found in Pescott et al. (2019).

Whilst the preceding sampling methodology was designed to be as accommodating as possible from the surveyor point of view, as befitting an accessible and sustainable citizen science scheme (Pescott et al., 2019), surveyor-choice flexibility introduces a number of challenges from an analytical perspective. For example, depending on the choices that a surveyor makes with respect to habitat resolution and level of general participation (i.e. Wildflower, Indicator etc.), a particular species may be present in a plot, absent, or in some unknown state (because the species of interest was not listed for the particular habitat-surveyor level combination at which a plot was surveyed). The important process of manipulating the raw NPMS dataset in order to generate a maximally informative set of "states" for any given species across all samples relevant to a particular modelling exercise is described in the next section.

3. Processing the NPMS dataset

This section overviews the processing required for the raw data to accurately represent the information content of the NPMS plot surveys. Because surveyors use different lists of species depending on their level, in some cases we need to distinguish between true absences (species on a surveyor's target list but not reported) and unknown cases (species not on a surveyor's target list, so absence from a reported survey is not informative).

Scripts 1 and 2[1]

The first step in the processing of the raw NPMS data, as captured into the PostgreSQL database underlying the NPMS website, is to extract the required information into an appropriate format for modelling. At the current time two separate SQL queries are run from within R (script 1), retrieving the relevant information. These are defined as functions, and are run separately in script 2; note that to run this script a password for the PostgreSQL database is required. The first function in script 1 (getNpmsData_PlotsSamples) retrieves sample level information that will be required to later infer the underlying inferred status (i.e. present/absent/unknown) of a particular species in relation to surveyor sampling decisions. These data are returned in the following form (Table 2).

Table 2. Information concerning samples (i.e. plot visits) retrieved for processing. plot_id is the label of a small plot; monad is the 1 km grid reference; sample is the identifier for a sampling visit; surv_habitat is the surveyor-reported habitat for the sampling visit.

plot_id	monad	sample	title	surv_habitat
42360	TQ1168	884186	Wildflower survey	Neutral pastures and meadows
42361	TQ1168	884198	Wildflower survey	Neutral pastures and meadows
42363	TQ1168	884203	Wildflower survey	Neutral pastures and meadows
42364	TQ1168	884216	Wildflower survey	Nutrient-poor lakes and ponds
42367	TQ1168	884257	Wildflower survey	Hedgerows of native species
42420	TQ2814	888635	Indicator survey	Dry deciduous woodland

[1] Links to scripts can be found in the addendum at the end of this document.

The second function in script 1 (getNpmsData_SamplesSpecies) returns species-level information, along with fields linking these data back to the sample-level information retrieved above. These data are returned as shown in Table 3 below ("sample_id" links to the "sample" field in Table 2 above).

Table 3. Information about species records retrieved for processing. Id is the individual occurrence identifier; sample_id is the sampling visit identifier; date is the survey date; preferred_taxon is the recorded taxon name; tvk is a link to the UK Species Inventory; domin is the cover-abundance category according to the Domin scale.

id	sample_id	date	preferred_taxon	tvk	domin
4231103	2066461	30/09/2016	Digitalis purpurea	NBNSYS0000004094	2
5933608	3006272	17/06/2017	Rumex crispus/obtusifolius	NHMSYS0021123579	3
2409246	1122261	30/08/2015	Epilobium hirsutum	NBNSYS0000003536	5
4246241	2072787	11/08/2016	Calluna vulgaris	NBNSYS0000003902	8
5575360	2777671	11/08/2017	Mercurialis perennis	NBNSYS0000003721	3
4144731	2024256	30/05/2016	Mentha aquatica	NBNSYS0000004198	2
4161692	2033676	05/06/2016	Viola reichenbachiana/riviniana	NHMSYS0020083544	3
6036022	3073543	12/09/2017	Cirsium vulgare	NBNSYS0000004490	2

Script 3

These two datasets are subsequently processed together in script 3. This script processes the retrieved data in the following ways: First, it reads in a standardised list of NPMS indicator species, with a row for every fine- and broad-scale habitat combination for each taxon. This information is then summarised into separate fine and broad habitat lists; new columns are also created indicating whether a species is an indicator at both the Wildflower and Indicator levels for a particular habitat. This information serves as a lookup table when later processing the NPMS field data. Finally, a reduced table called "indsLookup" is produced which provides information about the status of every NPMS indicator species in relation to all possible habitat-surveyor level combinations. An example extract from this table is given below (Table 4).

Table 4. Look-up table created for interpreting presence/absence/unknown status of species records in relation to volunteer survey-level choices. indicaName is the recommended taxon name; indiciaPrefTvk is the link to the preferred name in the UK Species Inventory; 'combined' is a field that indicates the habitat × survey level for which the species is an indicator.

indiciaName	indiciaPrefTvk	combined
Alopecurus myosuroides	NHMSYS0000455779	Arable field margins, Indicator survey
Spartina anglica	NHMSYS0000463855	Coastal saltmarsh, Indicator survey
Deschampsia flexuosa	NBNSYS0000002623	Dry acid grassland, Indicator survey
Brachypodium pinnatum s.l.	NHMSYS0021123603	Dry calcareous grassland, Indicator survey
Deschampsia flexuosa	NBNSYS0000002623	Montane acid grassland, Indicator survey
Agrostis capillaris	NBNSYS0000002638	Montane acid grassland, Indicator survey
Anthoxanthum odoratum	NBNSYS0000002667	Montane acid grassland, Indicator survey

The next stage is the extraction of a relevant set of samples for the analysis of any particular species. At this point inferential considerations enter play, because the choice of samples used for any particular analysis brings with it implications for subsequent inference. That is to say, considerations about the population from which a particular sample is a sample are of importance. For example, if we are interested in developing trends for dry calcareous grassland, but only select plots that were classified as such in the field, then we may be missing samples from grassland that are marginal in terms of their affiliation to that habitat type. Such samples may improve over time (e.g. due to changed management or other restoration efforts), and we would like any indicator to capture these changes. In the first instance, a sensible level for indicator development seems to be the broad habitat, and this is the approach developed here.

The function getSamples joins the species and sample level information previously extracted together, and filters these according to a user-defined list of fine- and broad-scale habitats that together define the population of habitats that we think represent a sensible environmental space around our target community. Other, less closely affiliated habitats could also transition into our target community: arable could be restored into an unimproved or semi-improved grassland for example. However, rather than include all plots and samples in the calculation of the indicator for a particular species, the current proposal is that plots are only included when at least one sampling event for a plot has been labelled as one of the habitats relevant to a particular species. For example, a plot that had previously been classified as arable would only enter the set of plots used for the creation of a trend for a particular grassland species once a sample from that plot had been recorded as an affiliated grassland type. From this point on, all older samples recorded at this plot would be included in the species' trend; this would represent an improvement in the national stock of that habitat, as the historic samples would typically have zero cover for the grassland species until the point of restoration, whereupon a positive cover would be recorded.

The next function in script 3, spSamplePA, takes the list of samples previously extracted according to our set of habitats (the example in the script uses the following set of NPMS broad and fine grassland habitats: Neutral pastures and meadows, Dry acid grassland, Dry calcareous grassland, Neutral damp grassland, Lowland grassland), and a focal species (or other taxon grouping used by our scheme). The function evaluates the status of the focal species according to the habitat-surveyor level of every sample, creating a dataset consisting of presences, absences, and unknowns (coded "NA") on this basis. At the same time, the proportional cover information (collected according to the Domin scale) is unified across the species' dataset; this last step is required because cover information collected by surveyors at the Wildflower level is coded differently to that collected at the other two levels in the database. For any given habitat group-species combination for which this function is run, a data frame in the following format is produced (Table 5; the example is for Yarrow, *Achillea millefolium*, across the grassland samples described above).

Table 5. Processed species records information for *Achillea millefolium* in relation to the Lowland grassland broad habitat category. domin is the reported cover-abundance category according to the Domin scale; sample_id is a sample identifier; combination is the combination of habitat and survey level within which the species occurrence originated; date.x is the date of the survey; PAN is an indicator simplifying the Domin scale information to presence/absence/unknown; plot_id is a plot identifier; monad is the 1 km square grid reference; title is the survey level; surv_habitat is the reported habitat type.

domin	sample_id	combination	date.x	PAN	plot_id	monad	title	surv_habitat
0	1155416	Lowland grassland, Indicator survey	10/09/2015	0	144658	TM2226	Indicator survey	Lowland grassland
0	1193152	Neutral pastures and meadows, Inventory survey	01/07/2015	0	145706	SJ8850	Inventory survey	Neutral pastures and meadows
1	1243480	Lowland grassland, Wildflower survey	31/07/2015	1	146530	SK6092	Wildflower survey	Lowland grassland
7. 34-50%	1125910	Dry calcareous grassland, Indicator survey	19/05/2015	1	144454	TQ8353	Indicator survey	Dry calcareous grassland
NA	884186	Neutral pastures and meadows, Wildflower survey	27/04/2015	NA	42360	TQ1168	Wildflower survey	Neutral pastures and meadows

Note that, in the first column ('domin'), both zeros, presences (with abundances according to the Domin scale), and an NA are represented (the column 'PAN' represents these simply as presence/absence/NA data). The NA here relates to a sampling visit recorded in "Neutral pastures and meadows" by a Wildflower-level surveyor, a combination for which *Achillea* is not included as an indicator; this means that we do not know whether *Achillea* was present or not, because the surveyor was not asked to explicitly report on the presence of this species.

Script 4

This script simply serves as an example of the preceding process. The rest of this report describes the statistical model currently being developed for such processed NPMS species-level data.

4. Developing a statistical model for NPMS species data

Here we present a hierarchical statistical model for the NPMS data. This model seeks to make maximum use of the data collected, and integrates a standard occupancy modelling approach with a Beta distribution model for a species' non-zero cover data.

This section describes the statistical model that we have formulated for NPMS species-level data. The model is developed in a Bayesian framework using the JAGS language (Plummer, 2013) given the relative ease with which one can both specify complex hierarchical models and deal with missing data. See the addendum at the end of this report for a link to the GitHub repository containing the JAGS code representing this model. In addition, we also provide an illustrated representation of the NPMS survey processes that the following model seeks to represent in Appendix 1.

The model estimates both per species non-zero proportional cover and occupancy at the plot scale. This is informed by two pieces of information generated by NPMS surveyors: the recorded cover in a sampling visit to a plot, and the detection history of a species within a plot in a given year. Due to the NPMS methodology specifying that surveyors should aim to visit their plots twice a year, some information will often be available concerning the within-year detectability of a species (although note that not all surveyors will be able to follow this guidance, particularly in remoter areas of the UK). A link between the recorded proportional cover and a species' detectability is posited by the model (see McCarthy et al., 2013 for a discussion of this topic), in that a species' recorded abundance is used as a covariate in the detection sub-model. This means that the model should be able to adjust for the fact that species in plots at low abundance are, all other things being equal, more likely to be missed by surveyors. This, as with all occupancy models (Royle and Dorazio, 2008), means that we should be able to estimate true occupancy, as opposed to the confounded product of occupancy and detectability, as we have explicitly accounted for non-detections (false absences, where a present species is missed by a surveyor). The formulation used here thus attempts to account for imperfect detection using a standard occupancy model, but also uses information about the estimated distribution of proportional covers when present to make inferences concerning a species' true abundance in a plot.

Note that because we do not have replicated observations of a species' cover *within* sampling visits to a plot, we accept the reported plant species' proportional cover as an accurate estimate of the true proportional cover state at the time of survey (cf. Wright et al., 2017). If one has multiple estimates of a species' cover for a single sampling visit (e.g. if all plots were subject to independent recording by two or more surveyors during every visit) then these could also be included in the model, and the necessity of accepting a single report of proportional cover as "truth" would be removed. In that case, we would also be able to model the observation process for proportional cover, thus better separating observational error from the underlying true state (Wright et al., 2017).

The first part of our model treats only the non-zero proportion cover data collected by the NPMS surveyors. (See Fig. 3 for additional information). The true underlying (latent) proportional cover values are estimated based on the distribution across all non-zero proportional cover observations for the species being modelled. Thus the non-zero observations across the *i* sites and *j* years are assumed to be distributed to follow a Beta distribution with shape parameters a and b:

Eqn 1. $CposLatent_{i,j} \sim Beta(a, b)$

Eqn 2. $a = \mu \cdot \tau$

Eqn 3. $b = (1 - \mu) \cdot \tau$

However, the actual observation made by the surveyor during survey *s* of site *i* in year *j* is the interval-censored[2] observation $D_{i,j,s}$. $D_{i,j,s}$ is a random variable that can take the integer values {0,1,…,9,10}, these being the possible ordinal categories that a surveyor can score observed plant abundance at using the Domin scale[3]. Therefore, for any value of $D_{i,j,s}$ there are lower ($L_{i,j,s}$) and upper ($U_{i,j,s}$) bounds on the proportional cover scale defined by the associated Domin category, where both $L_{i,j,s}$ and $U_{i,j,s}$ lie in the interval {0,1} and $L_{i,j,s} \leq U_{i,j,s}$. If we recall that CposLatent is estimated for each site × year combination, but that there are in fact potentially two surveys for each site every year, then, for any survey cover score $D_{i,j,s}$, it is implied that,

Eqn 5. $\min(L_{i,j,s}) < \text{CposLatent}_{i,j} < \max(U_{i,j,s})$

That is to say, the Domin abundance category recorded by the surveyor implies that the underlying cover value is within the proportional cover bounds associated with that observation. The lower and upper bounds of $\text{CposLatent}_{i,j}$ are $\min(L_{i,j,s})$ and $\max(U_{i,j,s})$, i.e. the lowermost lower and uppermost upper cover boundaries for the species implied by the separate visits to a site during the year. As noted above (eqn 1), values of a and b are then estimated that best capture the distribution of the estimated "true cover" values of CposLatent across all sites and years.

The second part of the model uses the detection history of the species within a site × year combination to make inferences about true occupancy; that is, it is a standard site-occupancy model (Kéry and Royle, 2016).

Eqn 6. $z_{i,j} \sim \text{Bern}(\Psi_{i,j})$

Equation 6 indicates that the true state is Bernoulli-distributed random variable with $\Psi_{i,j}$ (psi) giving the estimated occupancy (presence/absence) probability for a species for any site × year combination. $x_{i,j,s}$ (eqn 7 below) indicates the observed state of a species (present/absent) during survey *s*.

Eqn 7. $x_{i,j,s} \sim \text{Bern}(\pi_{i,j,s})$

Eqn 8. $\pi_{i,j,s} = z_{i,j} \cdot \alpha_{i,j,s}$

Eqn 9. $\text{logit}(\alpha_{i,j,s}) = \gamma 0 + \gamma 1 * D_{i,j,s}$

Here, the observed presence/absence of a species during sampling visit *s* ($x_{i,j,s}$) arises from a Bernoulli distribution with per trial success probability $\pi_{i,j,s}$. In its turn, the per trial success $\pi_{i,j,s}$ is a function of the species true presence/absence at site *i* during year *j* and the detectability of the species during the visit $\alpha_{i,j,s}$, again, as in standard Bayesian occupancy models. The logit transformation of this detectability can be a function of a range of covariates (e.g. species type, such as graminoid[4] or non-graminoid, the time of year, or site-specific management). However, in the current example, detectability is determined by an intercept term (γ0, or gamma0) and by the regression coefficient for the recorded ordinal cover category[5] $D_{i,j,s}$, γ1 (gamma1). In the simulations described in Section 5, the recorded cover is the actual simulated plot cover value after it has been

[2] Censoring is when the true value of a variable is unknown; interval-censoring is when the true value is unknown, but is known to lie in a particular interval.

[3] Note that although 0 is included here, as it is a possible value for $D_{i,j,s}$, when the values of CposLatent are estimated plot visits where $D_{i,j,s} = 0$ are excluded because we only model non-zero covers as arising from the Beta distribution.

[4] The term 'graminoid' means grass-like, and includes grasses (Poaceae), sedges (Cyperaceae) and rushes (Juncaceae).

[5] Note that the prior distribution chosen for the intercept γ0 and the regression coefficient γ1 should take into account the fact that this is a logistic regression. For example, a prior that puts a significant amount of weight on zero is actually emphasising the value 0.5 on the probability scale (Northrup and Gerber, 2018). We use the formulation of Kéry & Royle (2016) to avoid this problem (i.e. we used a two stage prior, with a beta distribution, uniform between zero and one, subsequently logit transformed, for the intercept).

subject to a probabilistic decision as to whether it is detected or not (which is dependent on its true abundance); however, when we fit the model to real NPMS data below in section 7, this is the actual recorded Domin cover value (including unknown values, which are given a prior distribution).

Figure 3 provides an annotated Directed Acyclic Graph (DAG) of the model. DAGs normally map stochastic relationships using solid arrows and deterministic relationships using dashed arrows (Hobbs and Hooten, 2015). The state, z, a random variable, could also be influenced by covariates (e.g. climate), but the current model does not include this addition. Figure 3 indicates that the data $x_{i,j,s}$ and $D_{i,j,s}$ (the visit-level surveyor detection and recorded Domin-category cover respectively) are linked to the underlying true state z (the true presence/absence) in different ways: $x_{i,j,s}$ is a stochastic variable determined by the true state and detectability; whereas $D_{i,j,s}$ enables the estimation of μ and tau, but does not currently directly influence the estimation of z, except via its influence on detectability ($\alpha_{i,j,s}$). The recorded cover category at visit *s*, $D_{i,j,s}$ is a direct (deterministic) outcome of the estimated underlying cover in site *i*, year *j*, $CposLatent_{i,j}$, which is itself a stochastic outcome of an underlying Beta distribution (parameterised by its mean (μ) and precision (tau)).

In this example the global mean cover and precision are estimated across all years, although they could be indexed by year if an annual trend in a species' cover distribution was required. The DAG also details the derived variable, $C_{i,j}$, which is a combination of the estimated true state and the mean of the estimated cover distribution for the species modelled.

Eqn 10. $C_{i,j} = z_{i,j} \cdot Cpos_{i,j}$

Eqn 11. $Cpos_{i,j} \sim$ Beta (a, b)

Where Cpos is a new cover value estimated from the Beta distribution specified by the estimated parameters μ and tau. Although this does not have much practical value in the current set-up, if we extended our model to include covariates that influence the values that μ and tau can take, or information on spatial auto-correlation, then site × year estimates of Cpos, combined with $z_{i,j}$, could allow for site-specific estimates of the zero-inflated[6] cover of the species being modelled.

[6] A distribution is zero-inflated when it contains a large proportion of zeros that cannot be accounted for by the probability distribution otherwise used to describe it.

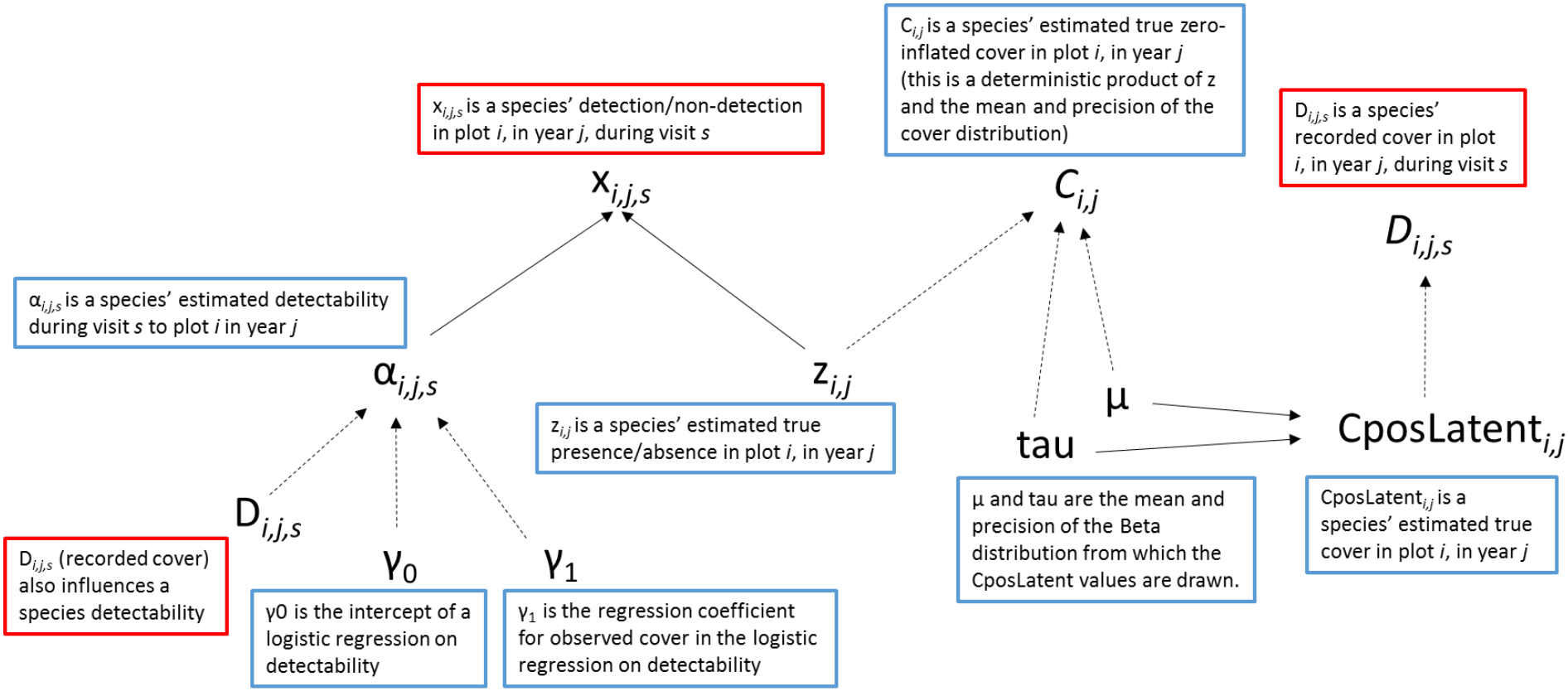


Figure 3. A Bayesian network or Directed Acyclic Graph showing the stochastic relationships (solid arrows) and deterministic relationships (dashed arrow) within the candidate model for NPMS species data. Note that μ and tau could be indexed by year to give annual mean estimates of a species' cover distribution. Red boxes are data, whereas blue boxes are parameters or states estimated by the model.

5. Testing the model using simulated data

Here we evaluate the proposed model using a variety of different simulated datasets. The performance of the model is assessed in relation to the bias and variance shown in relation to the underlying parameters used in the data simulations.

Novel models that are going to be applied to real world data should first be tested to see that they perform well in simulated examples: a model that fails to retrieve known, simulated, parameter values may not be suitable for use on noisier data from the real world. Model "performance" can be evaluated in many ways, although often the primary focus is on bias and variance, the two standard components of statistical error. That is to say, the estimates of parameter values should be centred around "truth", i.e. they should be unbiased, and estimates with lower variance are often to be preferred.[7]

The simulations presented in this section focus on whether the estimation of parameter values is accurate at large sample sizes, and then inspect the potential loss of accuracy and precision at smaller values more typical of the NPMS dataset (Table 6). A total of 11 scenarios are inspected. The simulations focus on whether the estimation of parameter values is accurate at large sample sizes, and then inspect the potential loss of accuracy and precision at smaller values likely to be more typical of the actual NPMS dataset. The simplest situation is that where all plots are occupied (i.e. all $\Psi_{i,j} = 1$), under this scenario any missing plot cover values are due to observer non-detections. We also include a "perfect detection" model (Table 6), where all occurrences of a species are recorded if present (i.e. there is not a probabilistic relationship introduced relating detection to abundance).

All model results focus on the estimation of μ, τ, γ0, γ1 and the estimated average annual occupancy, $(\sum_{j=1}^{J}\sum_{i=1}^{N} z_{i,j}/N)\,/J$, i.e. the average of the annual estimates of plot occupancy (in our simulation occupancy does not change across years). All models were run using JAGS and R2jags with 3 chains, with a 500 iteration burn-in, followed by a 500 iteration sample. Convergence was assessed using Rhat and visual inspection of traceplots; although these are not long MCMC chains, Rhats were always < 1.1 (Brooks and Gelman, 1998) and traceplots indicated stable, well-mixed chains. Note that during the actual model fitting the observed cover values used in the detection model were mean-centred, however, the intercept, gamma0, was transformed back to the original scale for plotting. The mean-centring was done to reduce the correlation between gamma0 and gamma1, this in turn should reduce the variance of their estimates. This type of mean-centring only affects the value of the intercept, not the slope, hence only gamma0 requires the back-transformation.

[7] Note, however, that this is not universally the case. The relative importance of bias and variance is context dependent; some statistical methods aiming for predictive (rather than, e.g., explanatory/inferential) accuracy trade off the tolerance of some bias against lower variance, with the net result of more accurate predictions on average (Shmueli, 2010).

Table 6. Parameters used for the simulation of data for models 0-3, and for the perfect detection scenario.

Variable (variable name in R code)	Notes (applicable to all tables)	Perfect detection	Model 0	Model 1	Model 2	Model 3
Number of plots	-	100	500	100	50	50
Number of within-year plot visits	-	10	10	10	5	2
Number of years	-	10	20	10	5	4
Occupancy (Ψ, psi)	-	1	1	1	1	1
Mean plot cover where present (mu)	Mean of beta distribution	0.5	0.5	0.5	0.5	0.5
Precision for plot cover where present (phi)	Precision of beta distribution	10	10	10	10	3
Detection model intercept (gamma0)	Logistic regression	NA	-2	-2	-2	-2
Detection model slope (gamma1)	Logistic regression	NA	3	3	3	3

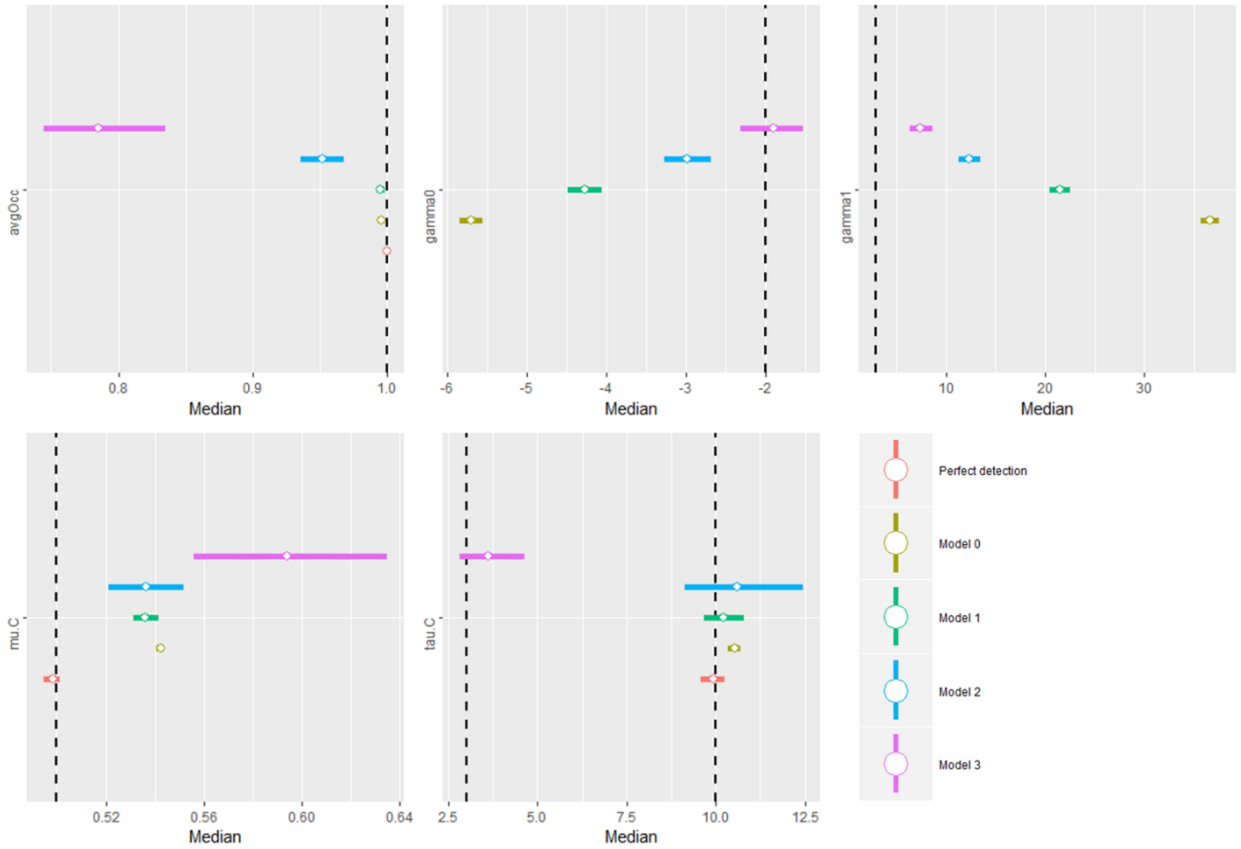


Figure 4. Estimates of parameters (gamma0, gamma1, mu and tau) and the average annual occupancy across years (avgOcc) for the simulated scenarios 0-3 and "perfect detection". Note that the "perfect detection" model (Table 6) does not have estimates of gamma0 and gamma1. White circles are the posterior 50th percentiles; coloured bars represent 95% credible intervals. Vertical black broken lines indicate the true value or values simulated.

Simulations 4–6 focus on the situation where the underlying true occupancy is gradually reduced (Table 7).

Table 7. Parameters used for the simulation of data for models 4-6.

Variable (variable name in R code)	Model 4	Model 5	Model 6
Number of plots	100	100	100
Number of within-year plot visits	5	5	5
Number of years	5	5	5
Occupancy (psi)	0.75	0.5	0.25
Mean plot cover where present (mu)	0.5	0.5	0.5
Precision for plot cover where present (phi)	3	3	3
Detection model intercept (gamma0)	-2	-2	-2
Detection model slope (gamma1)	3	3	3

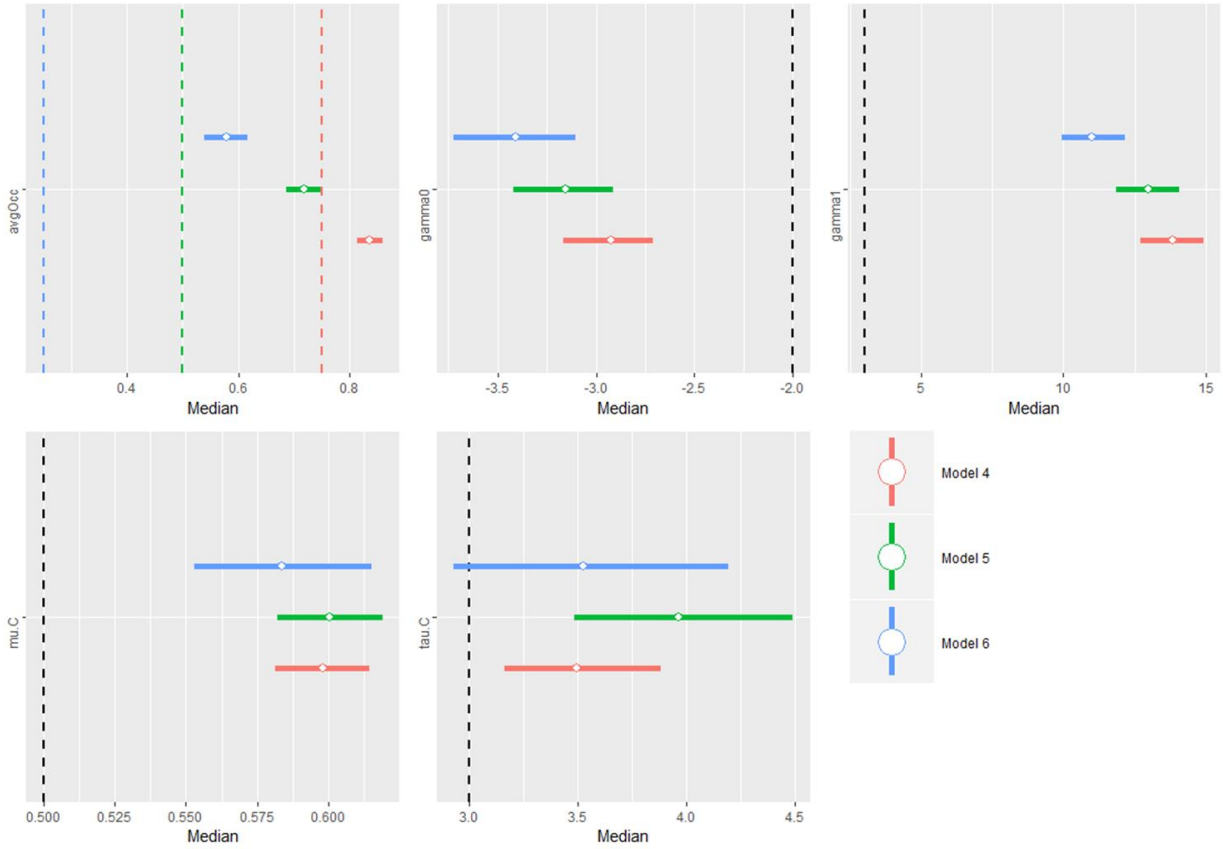


Figure 5. Estimates of parameters (gamma0, gamma1, mu and tau) and the average annual occupancy across years for the simulated scenarios 4-6. White circles are the posterior 50th percentiles; coloured bars represent 95% credible intervals. Vertical broken lines indicate the true value or values simulated.

Finally, models 7–9 examine the situation where the mean cover of a species when present is low, coupled with low occupancy and fewer within-year visits (Table 8).

Table 8. Parameters used for the simulation of data for models 7-9.

Variable (variable name in R code)	Model 7	Model 8	Model 9
Number of plots	100	100	100
Number of within-year plot visits	2	2	2
Number of years	5	5	5
Occupancy (psi)	0.25	0.25	0.25
Mean plot cover where present (mu)	0.25	0.10	0.025
Precision for plot cover where present (phi)	3	3	3
Detection model intercept (gamma0)	-2	-2	-2
Detection model slope (gamma1)	3	3	3

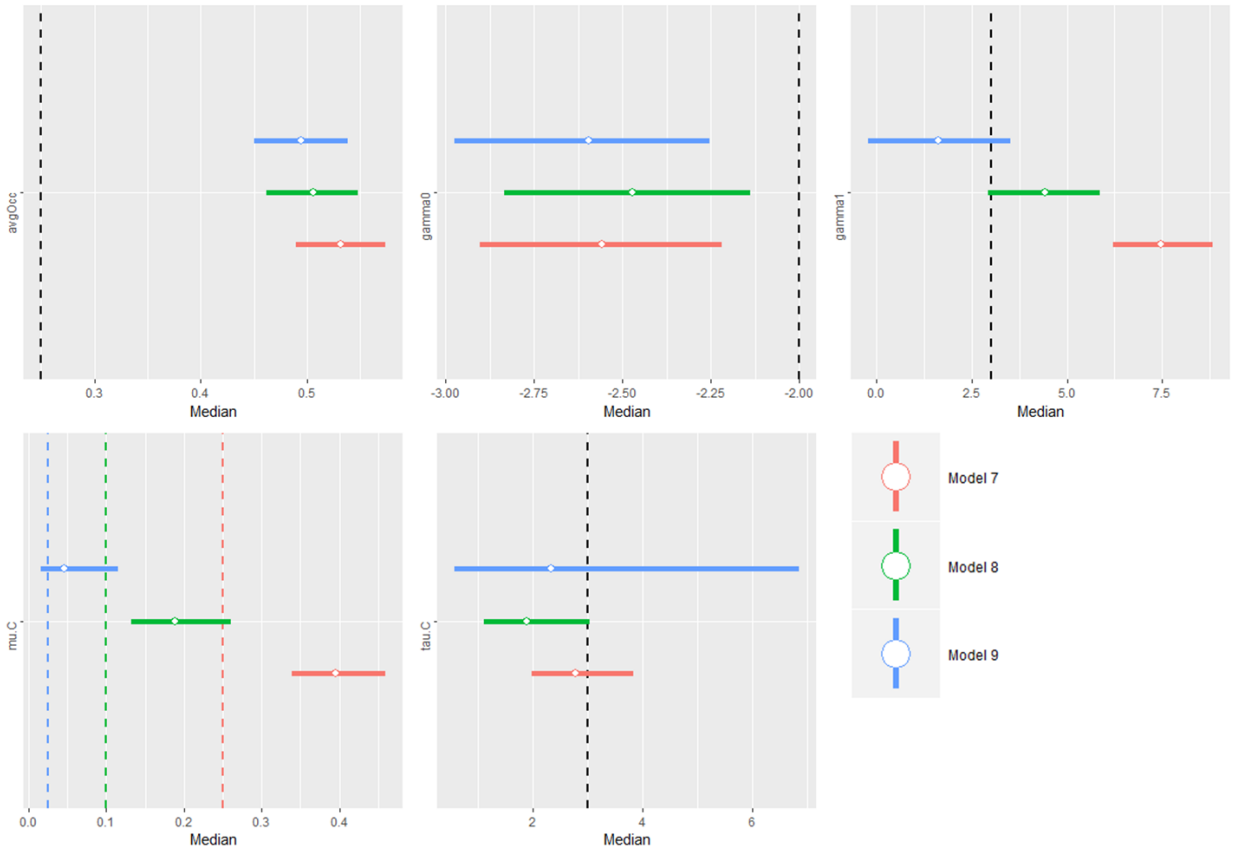


Figure 6. Estimates of parameters (gamma0, gamma1, mu and tau) and the average annual occupancy across years for the simulated scenarios 7-9. White circles are the posterior 50th percentiles; coloured bars represent 95% credible intervals. Vertical broken lines indicate the true value or values simulated.

6. Simulated data discussion

The results of the simulations indicate that the model performs as expected under "perfect" and large dataset scenarios. Smaller datasets induce various biases, many of which can be traced to the fact that, in our simulations, abundance and detectability were closely related. This biases the estimated mean of the underlying cover distribution upwards (because the distribution is left-truncated), and also impacts estimates of the intercept and regression coefficient in the detection sub-model. In a real dataset, this relationship would likely be less clear-cut, and we do not expect these biases to affect estimates of relative annual trends.

We have tested the model using simulated data under eleven scenarios. The results presented focus on whether the estimation of parameter values is accurate at large sample sizes, and then inspect the potential loss of accuracy and precision[8] at smaller values likely to be more typical of the actual NPMS dataset. The "perfect detection" model (Table 6) accurately estimated both the annual plot occupancy and the mean and precision of the Beta distribution underlying our simulated cover data (Figure 4), suggesting that there are no deficiencies with the specification of our model when the available data represent the true situation exactly. From this scenario, we preceded to a set (models 0-3; Table 6) where all plots were occupied (i.e. $\Psi_{i,j} = 1$ for all i,j); under these scenarios any missing plot cover values are due to observer non-detections. Models 0 and 1 were intended as "best case" scenarios, which are not likely to be representative of NPMS data (e.g. these models simulate ten visits to each plot within each year). Model 3 is closer to reality for a common species within the NPMS dataset (albeit with likely lower occupancy and mean cover values); model 2 was included as a stepping stone between models 0-1 and model 3. Models 4–9 retain the likely structure of the NPMS dataset in terms of visit frequencies, but look at situations in which either occupancy (models 4–6) or cover (models 7–9) are reduced.

Models 0-3

Figure 4 indicates that the "best case" models 0 and 1 perform reasonably well in retrieving the true underlying states: the average annual occupancy and the precision of the Beta distribution underlying the cover values are estimated more-or-less without bias. Reducing the amount of data and/or decreasing precision (tau, i.e. increasing the variance of the Beta distribution) results in increasing bias and variance, although for differing amounts for different parameters and states. μ, the mean of the Beta distribution underlying a species' cover values is, however, consistently overestimated (although not by large amounts) whenever detection is not perfect, irrespective of the number of plots, years and within-year visits (Fig. 4; models 0-3). This is considered to be due to the fact that detectability is a function of the observed cover: within our simulations observations are retained with probability anti-logit(gamma0 + gamma1 × original cover), therefore simulated observations with lower covers are more likely to be "overlooked" by our virtual surveyor. This effectively means that the information available to estimate the parameters of the Beta distribution underlying the observed covers is left-truncated, hence the mean is consistently biased upwards, as seen in Fig. 4 (mu plot).

To demonstrate this, we ran an additional version of model 2 (not plotted) where detection was only dependent on the intercept term, i.e. observations were subject to non-detection with probability anti-logit(gamma0), this model estimated mu as 0.49 (95% CI 0.46-0.52), demonstrating that when detection is independent of cover, the ability to estimate the mean of the Beta distribution without bias is retained (this is equivalent to thinning the distribution rather than truncating it). Removing

[8] Recall that accuracy and precision relate to bias and variance respectively.

the dependence of detection on abundance, however, introduces its own issues: if observations are detected with anti-logit(gamma0), recall that gamma0 is the intercept in the detectability logistic regression, then the value of gamma0 can be influential. Low values of gamma0, i.e. intrinsically low detectability, results in underestimates of occupancy. For example, when gamma0 is -2, detectability is 0.12, and annual occupancy is estimated at around 70% for model 2 (when the truth is 100%, cf. Fig. 4). This effect can be ameliorated by increasing the number of visits to a plot within a year, e.g. for model 2, increasing the number of plot visits to 20 means that annual occupancy is estimated at 95% (assuming again that detection is constant at 0.12 and true occupancy is 100%); this, however, is not a realistic scenario in terms of the actual NPMS survey protocol. Note that the issue of different combinations of detectability, within-season re-visits, occupancy etc. and their influence on model performance (in terms of variance and bias), are all discussed in detail by Welsh et al. (2013), with a rebuttal by Guillera-Arroita et al. (2014), response by Welsh et al. (2015), and a blog-based discussion (with numerous comments) by McGill (2014). The implications of these various issues are discussed below in the section "Simulation study conclusions".

The other striking conclusion from models 0-3 is the consistent underestimation of the intercept in the detectability regression (gamma0), and the consistent overestimation of the regression slope coefficient gamma1. This is likely to be due to the fact that, in our simulation, a probabilistic relationship exists between cover and detectability. As described above, the probability that an occurrence is not overlooked is anti-logit(gamma0 + gamma1 × original cover). Therefore, as for the estimation of mu, the result of left-truncating the cover distribution is to change the estimates of the intercept and slope in the detection logistic regression. This point is demonstrated below in Fig. 7, where a linear regression is fitted to a series of cover values and the corresponding series of detection probabilities are calculated as y = anti-logit(-2 × 3x), i.e. the values of gamma0 and gamma1 used in our simulations (e.g. see Table 6). The linear model fitted to the full data series is shown in red in Fig. 7a, Fig. 7b shows the linear model fitted to the same series truncated below 0.5. The linear model from the truncated series is also shown in blue in Fig. 7a, demonstrating the lower intercept and steeper slope that has resulted from this truncation.

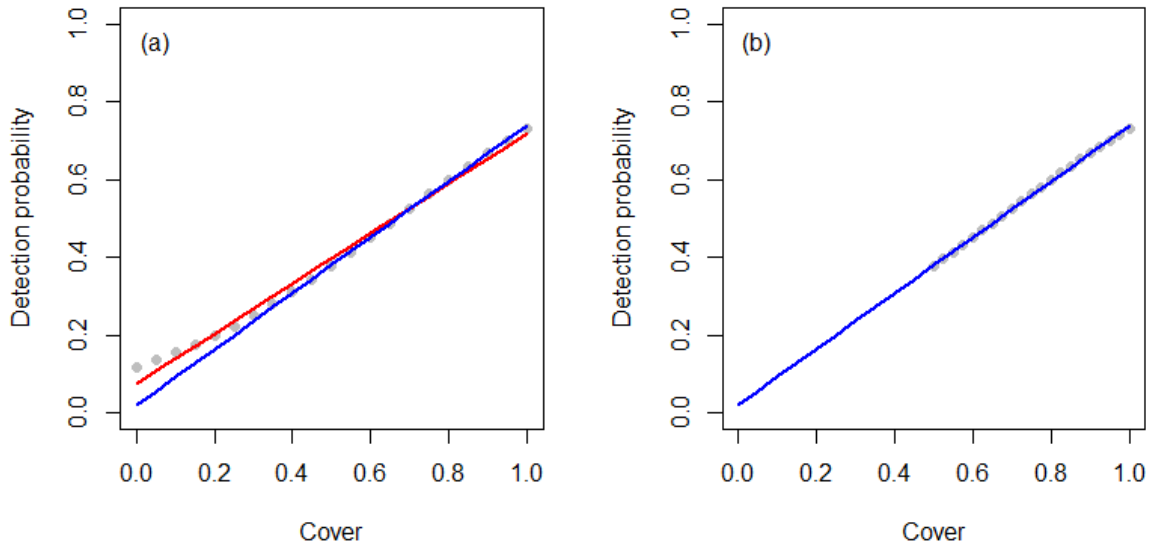


Fig. 7. (a) Two linear models fitted to simulated cover data and their implied detection probabilities, calculated as anti-logit(gamma0 + gamma1 × x), with gamma0 = -2 and gamma1 = 3. The red line indicates the model resulting from using the full range of x; the blue is from fitting a linear regression to the truncated data series (cover > 0.5). (b) The linear model fitted to the truncated data only. Simulated data points are shown in grey in both cases.

Somewhat counter-intuitively, Figure 4 indicates that the biases in gamma0 and gamma1 get worse as you accumulate data, at least for this set of scenarios; this is discussed further below. This issue of truncation also presents an additional problem, one of separation: once the data are truncated, there is complete, or near-complete, separation between the presence of "zero abundance" cover observations and higher detection. That is to say, the slope parameter becomes very hard to estimate accurately because of near-complete confounding between one of the predictors (cover) and the outcome variable (detectability). In our simulations we have dealt with this issue using an informative prior for gamma1, a normal distribution with mean = 0 and standard deviation = 1 (cf. Gelman et al., 2008)[9], this allows us to estimate a value for gamma1. The pattern of the biases in gamma0 and gamma1 (Fig. 4) across models 0-3, then, is likely to be a combination of the increasing amounts of available cover data coupled with the impacts of near-complete[10] separation (Menard, 2002). Note that in this situation it is not clear that the estimates of gamma1 are particularly meaningful: they represent the effect of combining an over-estimated slope coefficient with an informative prior dragging the estimate back towards zero.

Models 4-6

Models 4-6 explore the situation of decreasing true plot occupancy. In Figure 5 we see several consistent patterns: occupancy is overestimated, the intercept and slope of the detectability logistic regression are both under- and overestimated respectively. The mean of the cover distribution, mu, continues to be overestimated, and the precision of this distribution is also slightly overestimated; these biases in the parameters of the Beta distribution are considered to be for the same reasons as discussed above i.e. the left-truncation of the true cover distribution.

For occupancy, the degree of overestimation appears to be dependent on the underlying true occupancy, with the overestimates appearing worse for lower true plot occupancy; this also appears to be the case for gamma0, the intercept, although the estimates of gamma1 appear to improve with decreasing true occupancy. The effect of true occupancy on the model estimates is likely to be due to the fact that with decreasing true occupancy distinguishing true absences from non-detections becomes more difficult on average. This effect can be seen in the estimation of gamma0 – the true estimate of gamma0 becomes increasingly negatively biased as true occupancy declines, indicating that the model estimates that higher occupancy and lower detectability is more likely than the true situation of lower occupancy and higher baseline detectability. The fact that estimates of gamma1 appear to get worse with increasing occupancy may be due to a complex trade-off between the issue of distinguishing between the two occupancy/detectability scenarios just described, and the amount of (truncated) cover data available to the model as occupancy increases, the latter perhaps worsening the issue of near-complete separation previously discussed.

Models 7-9

Models 7-9 explore the situation where, for a constant level of occupancy ($\Psi = 0.25$), the true underlying cover decreases (Fig. 6). Models 7-9 also have a reduced number of within-year plot visits (2) compared to models 4-6. As for models 4-6, the average annual occupancy is over-estimated for all models to a similar extent to model 6 (which also had an underlying true occupancy of 0.25). The estimates of gamma0 and gamma1 are slightly more accurate than models 4-6. Increasing the amounts of data available to these models by increasing the number of within-year visits or the number of years of monitoring changed the estimates of gamma0 and gamma1 in relation to the true underlying cover (results not shown). The parameter estimates for model 9 (true cover = 0.025)

[9] See also the recommendations at https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

[10] Also known as "quasi-complete" separation in the literature.

were not greatly affected by increasing these variables, whereas changing these variables for model 7 (true cover = 0.25) resulted in estimates of gamma0 that were more negatively biased, and in estimates of gamma1 that were considerably larger (e.g. gamma1 estimate = 12.01 for model 7 with 10 years of monitoring and 5 within-year plot visits). This indicates that biases in these estimates are driven by the amount of cover data available to the model, as suggested above. Finally, the larger uncertainty associated with tau for model 9 is likely to be associated with the interval-censored nature of the cover data: when the Beta distribution has a mean of 0.025, most observations end up in the bottom cover interval used in our simulation (1e-16 < cover < 0.05), meaning that the precision of the true simulated distribution is challenging to assess.

Simulation study conclusions

The simulations suggest that, whilst the fundamental model proposed here is sound (it estimates parameters and states correctly when detection is perfect), the introduction of a probabilistic relationship between cover-abundance and detectability will introduce various biases into the model estimates. Biases in estimates of mu, the mean of the underlying Beta distribution from which positive covers are drawn, are easy to understand, in that the cover information available is left-truncated due to the fact that plots containing lower covers are more likely to go undetected. Biases in the average annual occupancy may be related to the more fundamental issues that occupancy models may have when only limited repeat visits within a time period are available (McGill, 2014; Welsh et al., 2013). However, these biases may also interact with the issue of quasi-complete separation that arises due to the left-truncation of the cover data.

In reality the relationship between cover and detectability will likely be weaker than that included in our simulation, reducing or even removing issues with near-complete separation in the detection sub-model. In addition, the inclusion of other relevant parameters in the detection model that are at least partially independent of cover (e.g. habitat management, the time of year) should improve the ability of the model to accurately adjust for a species' detectability. Investigating such additions to the model will be the focus of future work.

Is the proposed model fit for purpose?

The purpose of a statistical model is to capture important processes determining the data that we have observed, thus hopefully providing us with insight into the magnitude and direction of such processes. The model proposed here allows us to estimate both true plot occupancy and the distribution of recorded cover values of a species when it is present; it does this by combining a standard state-space model for occupancy, i.e. one incorporating an observation model, with a zero-inflated Beta distribution model for recorded covers. Although our simulations indicate that the model may be biased in some scenarios, these biases can be understood, and should not impact on the key aim of our model, which is ultimately to provide annual trends in species occupancy and cover values. We expect that, across a large, heterogeneous dataset such as the NPMS, biases will be relatively constant, and relative changes in these parameters over time will reflect true increases or declines; that is, we have no strong reason to suspect that such biases will change over time, meaning that our trends should index the true situation.

7. Applying the model to NPMS grassland data

Here we apply the model to NPMS data on 86 grassland species collected between 2015 and 2018. The model estimates ecologically sensible mean cover values for the species analysed, although mean plot occupancies tended to centre on 0.5, suggesting that many species do not have sufficient data for mean occupancy to be well-estimated at this early point in the survey.

In this section we describe the application of the model to the first four years of data collected within the NPMS. The model was applied to 86 species with available data that are listed as positive indicator species within the "Lowland grassland" NPMS broad habitat. Plots were only included in the modelling if there was at least one record of a species that had a known presence or absence, i.e. any plots consisting solely of 'NAs' were excluded. Two states were extracted from these model runs for plotting below: the estimated global mean cover of the Beta distribution underlying all years of a species' cover observations (μ; Fig. 8), and estimates of a species' plot occupancy for year 1 (2015; Fig. 9).

The estimates of the global mean of species' cover distributions made ecological sense (e.g. common grasses tended to have higher means), and commoner species had tighter credible intervals around the estimate of the mean (Fig. 8). Occupancy estimates tended to be centred on 50% (Fig. 9), except for a few commoner species, and it is believed that this is due to plots with no or little data for a species at the current time point receiving estimates that are dominated by the prior distribution used by the model for the occupancy parameter. This prior is designed to be uninformative (a uniform distribution between zero and one is used); however, the parameter plotted below (Fig. 9) is the estimated mean plot occupancy for a given year, and the mean of *n* random draws from a uniform distribution between zero and one is still 0.5. The centring of the species' occupancy estimates for 2015 around 0.5 indicates that, at this point, the prior used is likely to be dominating our conclusions about most species' annual plot occupancies. In the current results mean cover is likely to be more precisely estimated than occupancy for two reasons: (1) we are estimating mean cover for a species across all years, so more data is available than for the single year occupancy estimate plotted; and, (2), even within a year there is up to twice as much information available concerning a species mean cover compared to occupancy. This is because a maximum of two separate observations of a species' cover in a plot are made within a year, whereas a plot only contributes one data point towards the estimation of occupancy, and even this is less certain due to the issue of detectability.

Those species that most clearly deviate from 50% are those that are commonest in the UK landscape at small scales (e.g. *Rumex acetosa*, *Bellis perennis*, *Ranunculus repens* etc.) Note that it is of course possible that a species does have a true plot occupancy of 50%, although if this were the case for any species plotted below, we might expect tighter credible intervals around the estimate of the mean due to the likelihood of the data and the prior coinciding. At this point, recall that what is plotted in Fig. 9 is not the full predicted occupancy distribution for a species, but an estimate of the mean of that distribution with the associated uncertainty of that estimate. That is, the 95% credible interval presented is conceptually akin to a standard error in that it provides information on the certainty associated with the estimate of the mean. Plotting the full posterior occupancy distribution for a species would also be informative for investigating the degree to which the prior dominates the available data within any given year.

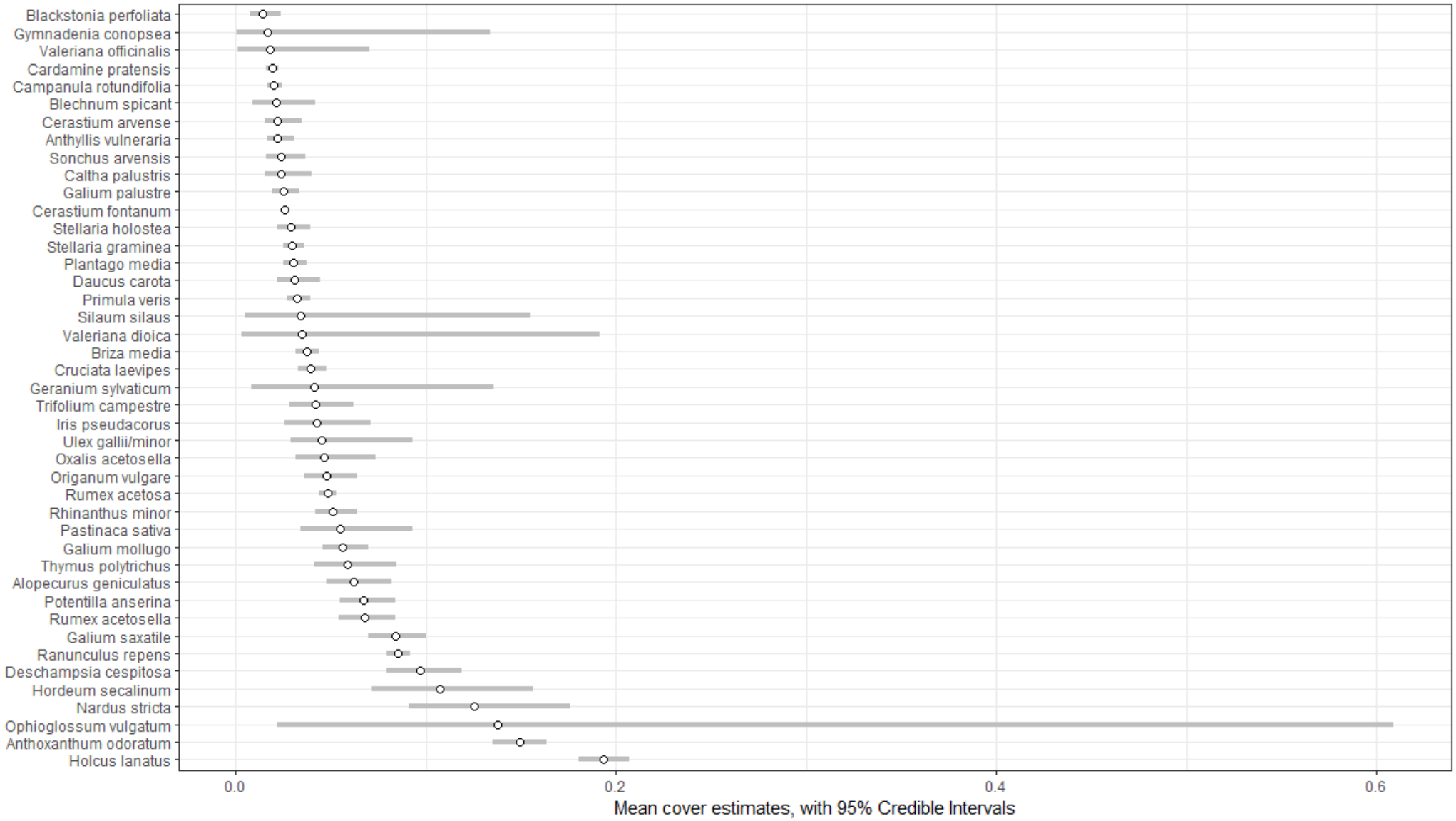


Fig. 8. Per species mean cover estimates (with 95% Credible Intervals) for 43 of the 86 Lowland grassland species modelled across all years; ordered by mean. Recall that this is the estimate of the mean of the cover distribution when the species is present (i.e. zero covers are excluded).

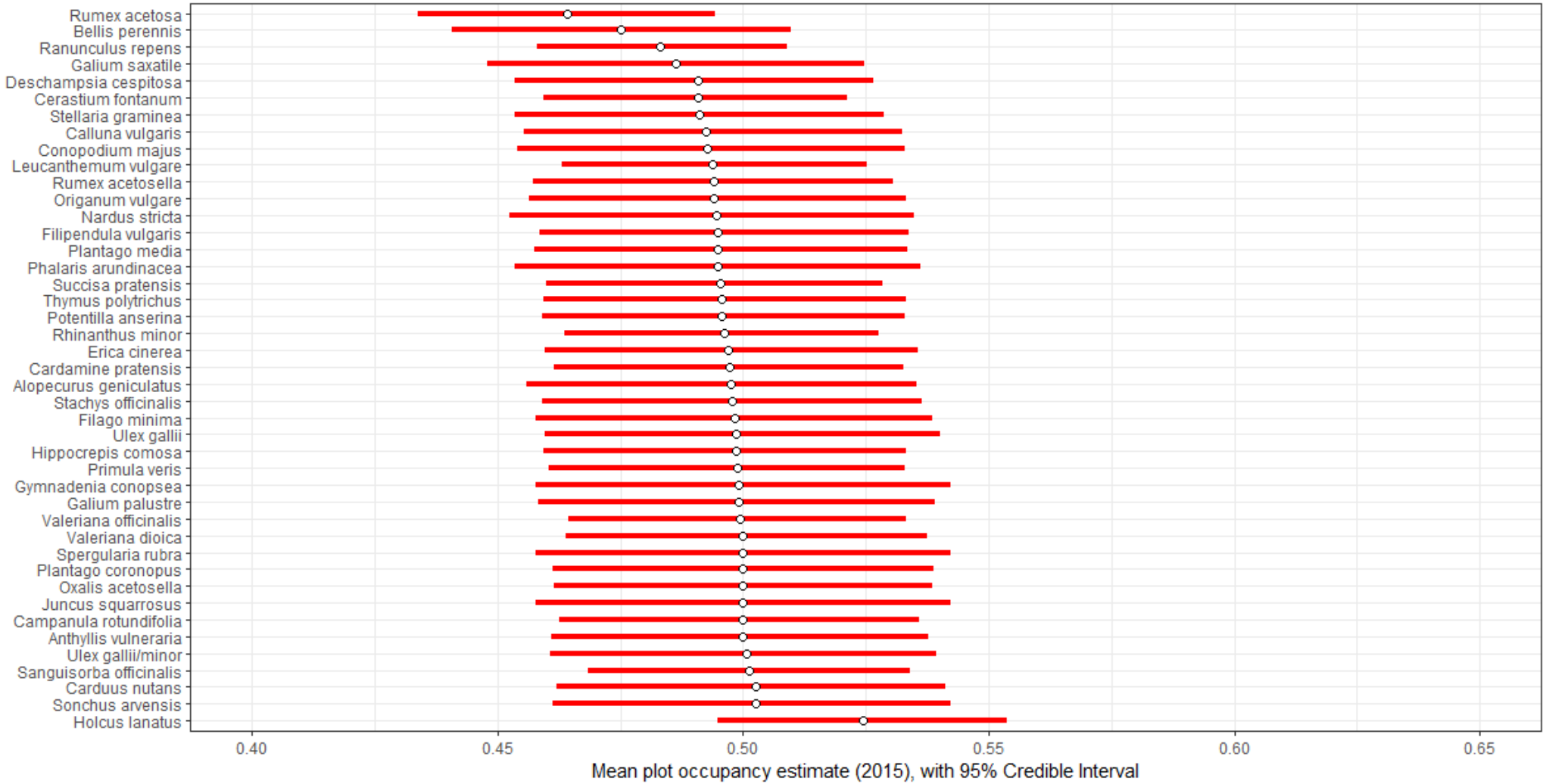


Fig. 9. Per species mean plot occupancy estimates for 2015 (with 95% Credible Intervals) for 43 of the 86 Lowland grassland species modelled; ordered by mean. Recall that plot occupancy can either be defined as the proportion of truly occupied plots, or as the probability that any given plot is occupied.

8. Summary, conclusions and future work

A novel combined abundance/occupancy indicator has been developed for NPMS data in a Bayesian framework. The simulation tests and applications to real data explored in this report indicate that the model performs well in ideal scenarios; biases in less data-rich scenarios can largely be explained by relationships between abundance and detectability. These are likely to be less clear-cut in real datasets, and future work will explore how additional covariates, thought to be of importance a priori, can be incorporated into future applications of the model to NPMS data. Extending the model to create annual indices, and considering how these may be aggregated, are considered to be the next steps required in the creation of a habitat quality indicator using NPMS data.

This report has outlined the structure of the NPMS dataset, and described the ways in which these data have been processed in order to extract the maximal amount of information for per species trend analyses. A hierarchical model, formulated within a Bayesian framework, has been developed which can: (i) deal with the missing data characteristic of the NPMS sampling scheme; (ii) deal with the interval-censored nature of the plant cover-abundance data collected; and which, (iii) integrates information on occupancy (accounting for detectability) and proportional cover into a single, zero-inflated model (cf. Wright et al., 2017). This model was described in a variety of ways, including through an annotated directed acyclic graph.

Data were simulated and processed in order to mimic the types of data likely to arise from the NPMS, as well other more data-rich scenarios, and a variety of such simulations were run in order to explore the strengths and weaknesses of the proposed model. A simulation in which species' occurrences were detected perfectly indicated that the model was capable of estimating simulated states and parameters without bias. Various additional scenarios in which the amount of data, the underlying true annual plot occupancy, and the true underlying distribution of plant covers (conditional on presence) were all varied revealed various characteristics of the current model. Foremost amongst these insights was the fact that if there is a strong relationship between abundance and detectability, then an upwards bias in estimates of the mean of the underlying cover distribution of species is inevitable. In addition, such a strong relationship may also result in near-complete separation in the detection sub-model (a logistic regression, as is standard in occupancy modelling). Additional issues with bias in estimating occupancy probably result from having few repeat visits within a time period (here a year), resulting in issues separating occupancy and detectability. This issue has been explored in detail elsewhere (Guillera-Arroita et al., 2014; McGill, 2014; Welsh et al., 2013), but could be further explored within the specific context of our model.

The model was also applied to Lowland grassland data (an NPMS broad habitat category) collected between 2015-18, and the results investigated graphically. These summaries made ecological sense where the mean of a species' underlying cover distribution was concerned, with species such as common grasses and ericaceous subshrubs (e.g. *Calluna vulgaris*, Heather) being estimated to have relatively high means with low uncertainty. Common forbs, which are widespread at larger scales but typically at low cover in plots, had lower means with low uncertainty (e.g. *Campanula rotundifolia*, Harebell, and *Cardamine pratensis*, Cuckooflower). More rarely encountered species had correspondingly higher uncertainty (e.g. *Gymnadenia conopsea*, Fragrant Orchid).

Future work

What changes to the model, then, are warranted by the work presented here? The detection model is a key issue, and, whilst we can learn something about the impacts of particular specifications

through simulations, ultimately (given that the truth is inaccessible) we must make a decision, or set of decisions, and hope that they capture key aspects of the processes at work in the real world. Assuming that detectability scales with abundance in some way is reasonable for the types of small plot surveys (25 m^2-100 m^2) that form the heart of the NPMS; this assumption has also been borne out by various bits of experimental and theoretical work (e.g. Dennett et al., 2018; McCarthy et al., 2013). Additional covariates could be added to this model to account for other likely drivers of detectability in our survey, e.g. plant functional group (graminoid or non-graminoid) or size is also likely to be a key driver, as is whether or not the habitat in the plot has been managed in some way (e.g. whether a grassland is reported as mown or grazed at the time of survey, information which is collected by NPMS volunteers). An alternative route would be to provide fixed estimates of detectability for species, these could be combined with other coefficients to be estimated from the data (such as the adjustments for habitat management noted above, or surveyor participation level), or simply used in isolation. Fixed detection probabilities per species have the advantage of reducing variance in the estimated states and parameters, but may induce bias if the chosen values are at odds with reality in some way. A related option might be the provision of a more informative prior for detectability, such that detection is assumed to be high unless the data indicate otherwise; we are not aware that this approach has been implemented in the ecological literature, although the ecological statistician Mark Brewer notes this strategy in one of the many comments on McGill (2014).

Another key extension of the work presented here is the estimation of annual mean covers for species (recall that above we have estimated a global mean cover across all four years of the survey, rather than estimates for each year). Such work would feed in the decisions required around whether occupancy, cover distributions, or both, would be best taken forward as habitat quality indicators. Finally, other challenges for the future include decisions around the visualisation and aggregation of trends. As we have seen from this report, both parameters relating to occupancy and proportional cover could be generated across species, decisions on which to present need to be made. Subsequently, it is envisaged that existing approaches to combining species trend lines could be employed (e.g. Isaac et al., 2015), this should be a relatively straightforward process once a model is decided upon.

References

Brooks, S.P., Gelman, A., 1998. General Methods for Monitoring Convergence of Iterative Simulations. Journal of Computational and Graphical Statistics 7, 434–455. https://doi.org/10.1080/10618600.1998.10474787

Dennett, J.M., Gould, A.J., Macdonald, S.E., Nielsen, S.E., 2018. Investigating detection success: lessons from trials using decoy rare plants. Plant Ecol 219, 577–589. https://doi.org/10.1007/s11258-018-0819-1

Gelman, A., Jakulin, A., Pittau, M.G., Su, Y.-S., 2008. A weakly informative default prior distribution for logistic and other regression models. Ann. Appl. Stat. 2, 1360–1383. https://doi.org/10.1214/08-AOAS191

Guillera-Arroita, G., Lahoz-Monfort, J.J., MacKenzie, D.I., Wintle, B.A., McCarthy, M.A., 2014. Ignoring Imperfect Detection in Biological Surveys Is Dangerous: A Response to 'Fitting and Interpreting Occupancy Models''.' PLOS ONE 9, e99571. https://doi.org/10.1371/journal.pone.0099571

Hobbs, N.T., Hooten, M.B., 2015. Bayesian Models: A Statistical Primer for Ecologists. Princeton University Press.

Isaac, N.J.B., Powney, G.D., Outhwaite, C.L., Freeman, S.N., 2015. Technical background document: Deriving Indicators from Occupancy Models. Biological Records Centre, Centre for Ecology & Hydrology, Wallingford, UK.

Kéry, M., Royle, J.A., 2016. Applied Hierarchical Modeling in Ecology: Analysis of distribution, abundance and species richness in R and BUGS. Volume 1: Prelude and Static Models. Academic Press.

McCarthy, M.A., Moore, J.L., Morris, W.K., Parris, K.M., Garrard, G.E., Vesk, P.A., Rumpff, L., Giljohann, K.M., Camac, J.S., Bau, S.S., Friend, T., Harrison, B., Yue, B., 2013. The influence of abundance on detectability. Oikos 122, 717–726. https://doi.org/10.1111/j.1600-0706.2012.20781.x

McGill, B.J., 2014. Detection probabilities, statistical machismo, and estimator theory. Dynamic Ecology. URL https://dynamicecology.wordpress.com/2014/09/15/detection-probabilities-statistical-machismo-and-estimator-theory/ (accessed 3.24.19).

Menard, S., 2002. Applied Logistic Regression Analysis, 2nd ed, Quantitative Applications in the Social Sciences. SAGE Publications, Inc., Thousand Oaks, California. https://doi.org/10.4135/9781412983433

Morrison, L.W., 2016. Observer error in vegetation surveys: a review. J Plant Ecol 9, 367–379. https://doi.org/10.1093/jpe/rtv077

Morton, D., Rowland, C.S., Wood, C.M., Meek, L., Marston, C., Smith, G., Wadsworth, R.A., Simpson, I.C., 2011. Final Report for LCM2007 - the new UK land cover map. Countryside Survey Technical Report No 11/07 (Publication - Report No. CEH Project Number: C03259). NERC/Centre for Ecology & Hydrology, Wallingford.

Northrup, J.M., Gerber, B.D., 2018. A comment on priors for Bayesian occupancy models. PLOS ONE 13, e0192819. https://doi.org/10.1371/journal.pone.0192819

Pescott, O.L., Jitlal, M., Redhead, J.W., Pocock, M.J.O., Roy, D.B., Walker, K.J., Harris, F., Southway, S.E., 2014. Design of a National Plant Monitoring Scheme (NPMS): report on development work between April 2014 and October 2014 (2nd contract variation) (Unpublished report to JNCC.). BSBI/CEH/Plantlife, Wallingford, UK.

Pescott, O.L., Jitlal, M., Smart, S.M., Walker, K.J., Roy, D.B., Freeman, S.N., 2016. A comparison of models for interval-censored plant cover data, with applications to monitoring schemes. PeerJ Preprints 4, e2532v1. https://doi.org/10.7287/peerj.preprints.2532v1

Pescott, O.L., Walker, K.J., Harris, F., New, H., Cheffings, C.M., Newton, N., Jitlal, M., Redhead, J., Smart, S.M., Roy, D.B., 2019. The design, launch and assessment of a new volunteer-based plant monitoring scheme for the United Kingdom. PLOS ONE 14, e0215891. https://doi.org/10.1371/journal.pone.0215891

Pescott, O.L., Walker, K.J., Pocock, M.J.O., Jitlal, M., Outhwaite, C.L., Cheffings, C.M., Harris, F., Roy, D.B., 2015. Ecological monitoring with citizen science: the design and implementation of schemes for recording plants in Britain and Ireland. Biol. J. Linn. Soc. 115, 505–521. https://doi.org/10.1111/bij.12581

Plummer, M., 2013. JAGS Version 3.4.0 User Manual. http://sourceforge.net/projects/mcmc-jags/files/Manuals/3.x/jags_ user_manual.pdf.

Royle, J.A., Dorazio, R.M., 2008. Hierarchical modelling and inference in ecology. Elsevier Academic Press, London, UK.

Scott, W.A., Smart, S.M., Clarke, R., 2008. Quality Assurance Report - Countryside Survey: QA and bias in vegetation recording. (Internal report.). NERC Centre for Ecology and Hydrology, Lancaster.

Shmueli, G., 2010. To explain or to predict? Statistical science 289–310.

Walker, K.J., Dines, T., Hutchinson, N., Freeman, S., 2010. Designing a new plant surveillance scheme for the UK (JNCC Report No. 440). JNCC, Peterborough.

Walker, K.J., Pescott, O.L., Harris, F., Cheffings, C., New, H., Bunch, N., Roy, D.B., 2015. Making plants count. British Wildlife 26, 243–250.

Welsh, A.H., Lindenmayer, D.B., Donnelly, C.F., 2015. Adjusting for One Issue while Ignoring Others Can Make Things Worse. PLOS ONE 10, e0120817. https://doi.org/10.1371/journal.pone.0120817

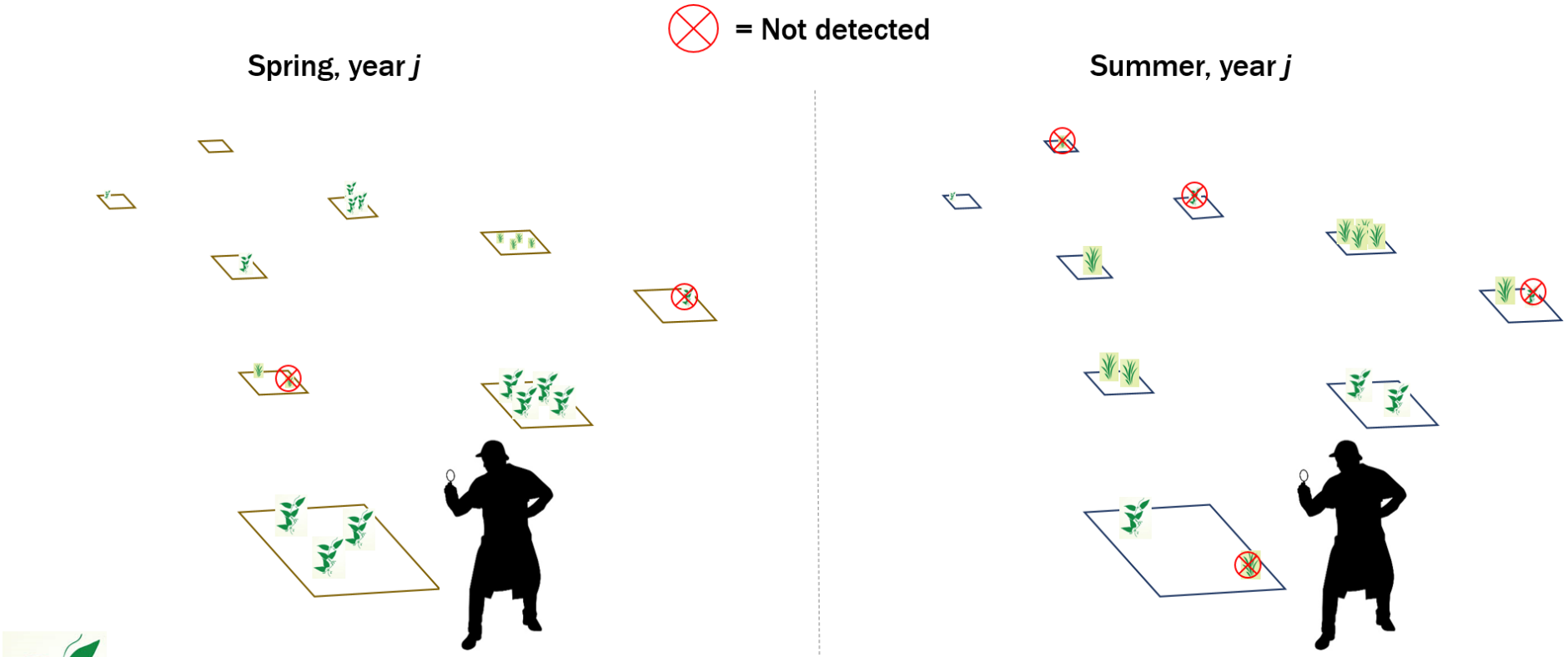
Welsh, A.H., Lindenmayer, D.B., Donnelly, C.F., 2013. Fitting and Interpreting Occupancy Models. PLoS One 8. https://doi.org/10.1371/journal.pone.0052015

Wright, W.J., Irvine, K.M., Warren, J.M., Barnett, J.K., 2017. Statistical design and analysis for plant cover studies with multiple sources of observation errors. Methods in Ecology and Evolution 8, 1832–1841. https://doi.org/10.1111/2041-210X.12825

Acknowledgements

This study was supported by funding from JNCC to CEH for project number NEC06730 National Plant Monitoring Scheme. This work was also supported by the Natural Environment Research Council award number NE/R016429/1 as part of the UK-SCAPE programme delivering National Capability. Thank you to Stephen Freeman (CEH), Charlotte Amos, James Williams, Julie Day, Alun Jones, Kirsi Peck (all JNCC), Christine Holleran, Karen Thomas (both Defra), and the Biodiversity Indicators Steering Group for comments that improved the work and report.

Appendix 1



Species	Spring survey data		Summer survey data		Naive annual occupancy	Naive non-zero μ annual cover	True annual occupancy	True non-zero μ annual cover
	Plot detections	Non-zero μ cover	Plot detections	Non-zero μ cover				
1	5/6	12.0%	3/6	6.7%	5	9.4%	6	13.4%
2	2/6	12.5%	4/6	10.0%	4	11.3%	6	11.7%

Figure A1. This figure attempts to represent the key observation processes that our statistical model seeks to represent for any given species within the NPMS. Within a given year, multiple small plots are surveyed at two different time points (visits). Given that individual occurrences of plants in plots have a chance of being overlooked, the naïve estimates of plot occupancy (i.e. the number of presences/number of plots considered) and cover-when-present are both likely to be inaccurately estimated. Occupancy modelling approaches allow us to use the information about a species' detectability to estimate the true annual occupancy, rather than being satisfied with the naïve (unadjusted) estimates. Note that, without multiple surveys of a plot within a visit, we cannot estimate the true cover-when-present – our model relies on the reported numbers only for estimates of this variable.

Addendum

All code and data referred to in this report, except for the actual database underlying the NPMS data-capturing website (www.npms.org.uk), are available at https://github.com/sacrevert/NPMStrends. Extracts of the raw NPMS data, with all relevant metadata, are deposited in the NERC CEH Environmental Information Data Centre on an annual basis. The corresponding raw biological records, i.e. species presence data only without the additional information required to reconstruct the sampling history of particular plots, are also available through the National Biodiversity Network.

Within the GitHub repository, the URLs for the scripts referred to in section 3 of this report are as follows:

Script 1: https://github.com/sacrevert/NPMStrends/blob/master/scripts/1_getDataFromIndiciaFuns.R
Script 2: https://github.com/sacrevert/NPMStrends/blob/master/scripts/2_getDataFromIndiciaExec.R
Script 3: https://github.com/sacrevert/NPMStrends/blob/master/scripts/3_processDataFuns.R
Script 4: https://github.com/sacrevert/NPMStrends/blob/master/scripts/4_extractData.R
Simulation script:
https://github.com/sacrevert/NPMStrends/blob/master/scripts/X2f_simData_JAGS_intervalCens_forSims.R