Check for updates

ECOGRAPHY

Research article

Estimating relative species abundance using fossil data identified to different taxonomic levels

Trond Reitan[1,2], Emanuela Di Martino[1] and Lee Hsiang Liow[1,2]

[1]Natural History Museum, University of Oslo, Oslo, Norway
[2]Centre for Planetary Habitability, Department of Geosciences, University of Oslo, Oslo, Norway

Correspondence: Lee Hsiang Liow (l.h.liow@nhm.uio.no)

Ecography
2023: e06866
doi: 10.1111/ecog.06866
Subject Editor: Kate Lyons
Editor-in-Chief: Miguel Araújo
Accepted 30 August 2023

Site-occupancy modelling is widely used in ecology for understanding species distribution, habitat-use and community changes but its application is still limited in paleoecology, where incomplete detection is also routine. Here, we make extensive expansions to an earlier multispecies occupancy model used to estimate the dynamics of relative species abundance in fossil communities. These expansions include incorporating counts of individuals at sites, explicitly allowing for the inclusion of specimens assignable to genus – but not species-level, a situation common in paleontology, and modelling regional presence/absence. We provide simulations to check the performance of this new model, as well as simulations to quantify the benefits of using individual count data versus subsample occupancy data, and model estimates versus face-value (raw) estimates, respectively. We also provide an empirical case study using occupancy data from a community of marine benthic colonial animals preserved in the Pleistocene of New Zealand. We find that the new model performs well, especially when it comes to recovering relative abundance dynamics and that it is well worth the effort to both collect individual count data and to include individuals unidentified to species-level in the site-occupancy modelling framework. This extended model can be widely applied in paleoecological settings and is necessary when both the average and uncertainty values of relative abundance dynamics need to be robustly estimated.

Keywords: Bryozoa, fossil communities, hierarchical modelling, multispecies site-occupancy models, preservation

Introduction

The abundance of a given species in its community is the consequence of population growth, which in turn is a consequence of survival and reproduction. The latter are influenced by competition, predation, disease, and intraspecific variability and environmental stochasticity. The relative abundance or dominance of different species in natural, contemporary communities are observed to shift on shorter time-scales, where such shifts can be directly attributed to environmental change, invasive species

© 2023 The Authors. Ecography published by John Wiley & Sons Ltd on behalf of Nordic Society Oikos
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

NORDIC SOCIETY OIKOS
www.ecography.org

and cyclical behaviour, among other factors. On longer timescales where observations are more challenging, however, the imprint of multiple processes not only obscure underlying mechanisms of such shifting dominance, but may also veil true differences in relative abundances. Yet, it is important to be able to reconstruct population dynamics deeper in time, using genetic evidence, biogeographic and/or paleoecological data to understand the past (Hoban et al. 2019, Dussex et al. 2021) and to use the past as baselines for anthropogenic change (Dillon et al. 2022).

Site-occupancy modelling uses information from repeated site visits to account for incomplete detection while estimating population and community parameters, including relative abundance. It is widely applied in many branches of ecology (MacKenzie et al. 2017) but not in paleoecology. We are aware of its application in only three previous papers (Liow 2013, Lawing et al. 2021, Reitan et al. 2022), although Foote 2016 approached the problem of occupancy while accounting for preservation without explicit replicates in sampling. Likewise, estimating relative abundance has long been a focus of paleoecological research (Kidwell 2001) but there has been little to no work on modelling incomplete detection when it comes to fossil community dynamics (Nichols and Pollock 1983 and references citing them for capture–recapture approaches for estimating origination and extinction probabilities). Incomplete detection in the fossil record can be in part attributed to non-biological factors, including varying sedimentation rates, storms, bioturbation, lateral transport, erosion and other processes that themselves tend to be temporally varying on longer time scales. A recent study used fossil data to estimate the dynamics of relative species abundance in a Pleistocene benthic community by developing a multispecies occupancy model that takes into consideration the features of fossil preservation (Reitan et al. 2022).

Reitan et al. (2022) were interested in how different species of marine invertebrates encrusting hard substrates change in their relative abundances over two million years. More specifically, they wanted to build a hierarchical model to estimate how several co-existing cheilostome bryozoan species waxed and waned over time across several geological formations within the Wanganui Basin of New Zealand. In the model they developed, which can also be applied to other paleoecological study systems, detection was in a one-to-one relationship with underlying abundance given site-occupancy.

This previous fossil multispecies occupancy model had features that are particularly suited to data commonly collected or are collectable in paleoecological settings. Like all site-occupancy models, (fossil) sites are re-sampled such that data from the replicate sampling allow us to tease apart site-occupancy and detection. The replicate sampling are subsamples within sites, which in the case of Reitan et al. (2022) were unique shells found within the sites, on which different species of encrusting cheilostome bryozoans were observed.

The current paper extends the Reitan et al. (2022) model by 1) using counts of individuals rather than only presence/absence of species on the subsample-level, 2) adding species-level random effects, 3) incorporating specimens assignable to genera but not species, 4) modelling regional presence/absence and 5) incorporating information when regional presence is known. Like the original model, these improvements are applicable to many paleoecological systems, in addition to the one presented in Reitan et al. (2022). To this end, we extend the dataset presented in Reitan et al. (2022), adding 18 species and 25 sites where observations were made. We provide simulations to explore how well the expanded model recovers parameters of interest, and the performance of model-estimated parameters based on individual counts or subsample-level presence/absence data versus 'face-value' information, i.e. raw estimates. We hypothesize that the estimation accuracy will increase when: 1) using site-occupancy modelling instead of face-value patterns (e.g. raw ratios); 2) using counts of individuals instead of only observation/non-observation of species at subsample-level and 3) including information on specimens only identified to genus level instead of ignoring this information. We end by discussing why it is important to explicitly model detection and present general recommendations for paleoecological work.

Material and methods

Data

The site-occupancy data are collected from a community of fossilized benthic, encrusting cheilostome bryozoans found in the Wanganui Basin of New Zealand (Carter and Naish 1998, Proust et al. 2005, Pillans 2017) previously presented in Reitan et al. (2022). There are now subsamples (= shells, typical substrates for bryozoans) for encrusting cheilostomes in 144 sites in transgressive system track (TST) shell beds from 10 geological formations, spanning about two million years. Such shellbeds reflect similar depositional conditions (facies). We tabulated the observed presence of any fossilized individuals of 21 focal cheilostome species on each shell (i.e. subsample) sampled from any given site, including the three previously analyzed in Reitan et al. (2022). With the exception of five species of *Microporella*, two of *Escharoides* and two of *Exochella*, each of these species are, as far as we know, sole representatives of their genera in the Wanganui Basin. This is important for later modelling considerations. As in the previous study, the superspecies represents all other encrusting bryozoan species in the community, excluding the 21 focal species. The observed presence of the superspecies gives information to improve parameter estimates (Model description). These observations constitute the occupancy dataset. For additional sources concerning regional occupancy (Extension 5) below), we draw on data collected for a separate study (Liow et al. 2016) as well as more recently collected material (provided in the folder 'RAMU-MSOM' available at https://github.com/trondreitan/RAMU-MSOM).

Original model: a brief recap

The objective of Reitan et al. (2022), was to estimate the temporal dynamics of relative species abundance. The data in

16000587, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ecog.06866 by University Of Oslo, Wiley Online Library on [02/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

that study had one row per site containing information about the number of subsamples having an observed presence of each species, i.e. subsample counts. A given species, s, has the potential of being observed in a given subsample if it is present in a given site, i. If a given site is not observed to contain the given species in any of its subsamples, it could mean either that 1) the site was truly devoid of that species or 2) that the species was present but not sampled (MacKenzie et al. 2002).

We denote the site-occupancy probability of a given species as Ψ and detection probability as p. More specifically, p is the probability that each subsample has at least one observation of the given species. The probability that a species is found on a given subsample is thus Ψp. The site-occupancy and detection probabilities can be specific to sites i within formations belonging to specific time-intervals. Here, formation, $f \in 1, \ldots, N_f$ where N_f is the number of formations, and species, $s \in 1, \ldots, S$, where S is the number of species (and the superspecies is indexed as S). Thus, we write $\Psi_{i,s}(\theta)$ and $p_{i,s}(\theta)$ for the site-occupancy and detection probabilities respectively, where θ is the set of parameters and random variables of the model. Since p is independent for each subsample, the binomial distribution can be used to summarize the chance of observing species, s, $y_{i,s}$ times out of T_i subsamples in site i. In other words, $y_{i,s}$, indicate here subsample counts. However, there may be variation in true abundance of a species from site to site, and hence variation in its detection probability, giving rise to overdispersion. Temporal variation within each formation, observational errors and local heterogeneity in preservation can further introduce extra variation, thus, we use a beta-binomial distribution. A binomial distribution is suitable for presence/absence information while the beta part of the distribution accounts for the extra variation beyond a standard binomial distribution. Since site-occupancy is not guaranteed, this further expands into a zero-inflated beta-binomial distribution. Here, the zero-inflation is for the numerous sites that have no observation due to true lack of occupancy of the species involved. We assume site-occupancy probability and the detection probability are each affected by a random factor ($\delta_{f(i),s}$ and $\varepsilon_{f(i),s}$, respectively) representing individual species dynamics in a given formation. Additional random factors representing dynamics common across species ($v_{f(i)}$ and $u_{f(i)}$ for site-occupancy and detection probabilities, respectively) encompass variation in preservation characteristics and hence detection probabilities in different geological formations.

To estimate species relative abundance, we assume that detection probability given occupancy, p, is linked to abundance-given-occupancy such $p = 1 - e^{-\lambda}$ via a Poisson model where λ is the mean number of detections. λ is associated with relative abundance dynamics via a log-link (i.e. the abundance-focused model in Reitan et al. 2022). We use a logistic link between site-occupancy probability and the accompanying random factors. When the likelihood of observing presence in $y_{i,s}$ out of T_i subsamples is a zero-inflated beta-binomial distribution that replies on formation-dependent and species + formation-dependent random factors, this is expressed as:

$$y_{i,s} \sim z\beta bin\left(T_i, p_{i,s}(\theta) = 1 - \exp\left(-\exp\left(\beta_s + u_{f(i)} + \varepsilon_{f(i),s}\right)\right)\right),$$
$$\kappa_s, \Psi_{i,s}(\theta) = I(s = S) + I(s < S)\,\mathrm{logit}^{-1}\left(\alpha_s + v_{f(i)} + \delta_{f(i),s}\right)\Big) \quad (1a)$$

$$u_f \sim N\left(0, \sigma_u^2\right), v_f \sim N\left(0, \sigma_v^2\right)$$
$$\delta_{f,s} \sim N\left(0, \sigma_{\delta,s}^2\right), \varepsilon_{f,s} \sim N\left(0, \sigma_{\varepsilon,s}^2\right) \quad (1b)$$

Here, κ_s is an overdispersion parameter (which we retrospectively found did not need the species-dependency we imposed on it). $I()$ is the indicator function which takes value one when the statement inside is true and 0 if false. S is the total number of species. α_s and β_s give average site-occupancy and detection probabilities for each species on their transformed scales (Reitan et al. 2022).

Using this set-up, relative abundance for a species is estimated as the product of occupancy probability and abundance-given-occupancy in that formation, compared to the sum of these products for all species in question in that formation, such that:

$$R_{f,s} = \frac{\Psi_{f,s}(\theta)\lambda_{f,s}(\theta)}{\sum_{s'=1}^{S} \Psi_{f,s'}(\theta)\lambda_{f,s'}(\theta)}. \quad (2)$$

We replaced the site index, i, with the formation index f, as both site-occupancy probability and abundance-given-occupancy only depend on species and formation here. Site-dependent variation is modelled through overdispersion.

We propose a set of modifications to the above model. Note that this model and its modifications assume that the probability of preservation is approximately equal in all species of interest (Reitan et al. 2022). Mathematical details of the new model follow after verbal descriptions of the extensions in the following section.

Model extensions

Extension 1: individual counts versus subsample count data per site

The original modelling was performed on the number of subsamples observed to have at least one individual of a given species (subsample counts). Some subsamples were observed to have tens of individuals of some species, while others just a few or none, reduction of the information to subsample counts constitutes a potentially huge loss of information.

Handling the data on the subsample level for individual counts is likely computationally unfeasible (Reitan et al. 2022), but we can move the analysis up to the site-level (arguments given in the Supporting information). Here, we use the negative binomial for an overdispersed version of the Poisson distribution for count data. We assume that the expected number of individuals at a site scales with the number of subsamples in the site.

16000587, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ecog.06866 by University Of Oslo, Wiley Online Library on [02/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Extension 2: species constants are replaced by random effects

In Reitan et al. (2022), data for only three focal species were available. Because only three species had to be modelled, they were each given two constant values (the time-averaged occupancy probability and abundance-given-occupancy). However, most communities are more species-rich, even when considering only common species, as done here. With more species, we turn these constants into random effects (such that the random factors represent time-dependent deviations from these constants) since the data are rich enough for inference on the distribution of species-dependent quantities. By adapting the distribution of these quantities to the data rather than giving each species its own prior distribution, the model is less sensitive to biases and uncertainty assumptions in the specification of priors.

Extension 3: individuals assignable to at least genus but not to species

Cheilostome bryozoans, like some other calcified marine taxa, can be assigned to their species with high confidence based on morphology (Jackson and Cheetham 1990). However, poor preservation due to taphonomic processes can reduce the possibility of assigning an individual to a lower taxonomic level (e.g. species or even genus), a situation common in paleoecology. However, if the individual can be identified to genus but not species-level, it still gives information for occupancy modelling. Imagine there are three species in a region, species A1, A2 and B, where B belong to a separate genus while A1 and A2 are in the same genus. Then, detecting 100 A1, 100 A2, 200 unidentified individuals belonging to genus A and 100 individuals to B, should suggest there were really 200 A1 and 200 A2 individuals and thus that the abundance of A1 relative to B was two to one rather than one to one. Note that all species belonging to a genus in the region should be included in the analysis; otherwise, the bias introduced by unidentified individuals cannot entirely be corrected. However, if the species left out are very rare, the bias introduced to the estimates will be negligible.

We thus need to multiply the estimated abundance-given-occupancy with the probability of non-identification to species-level, in order to get the apparent abundance-given-occupancy for the identified individuals. Note that this is only possible for individual count data, not subsample count data.

Extension 4: modelling regional occupancy

In some cases, there were no detections in any of the sites in a given formation for a species that is otherwise quite detectable in other formations. This suggests that it could be absent from the region at that time because that species had not arrived in the region yet; have been permanently or temporarily extirpated from the area; not have originated yet; or have gone globally extinct.

Because site-occupancy is required for site-detections, and regional occupancy (in a formation) is needed for any occupied sites, we now have a deeper hierarchy of explanations:

1. Species detected at a site: both site and regional occupancy are required.
2. Zero species detections at a given site, but some detection at other sites in the formation (regional occupancy): either 1) no detection though there is occupancy at the site (at unmeasured or non-preserved subsamples) or 2) absence at the given site (most parsimonious).
3. Zero detection in any of the sites in a formation: either 1) no detections though there is undetected occupancy at some sites and thus regional occupancy, 2) absence in all the sampled sites but presence at unmeasured sites, hence regional occupancy or 3) regional absence (most parsimonious).

Extension 5: external information concerning regional occupancy

In our dataset, and commonly so in other paleoecological datasets, some species that are quite detectable in some formations have no detections in others. Here, we could consider additional data sources (e.g. collected for other purposes or previously documented) external to the occupancy dataset to inform time-interval specific regional occupancy. If external data with certainty tells us that a certain species is in the region at a particular time, we can set regional occupancy to one for that species; where the external does not tell us that the species is present, we can allow for non-zero probability of regional absence.

Likelihood components

As described in Extension 1, we now use individual counts rather than subsample counts, so $y_{i,s}$ stands for individual count data henceforth. As mentioned in Extension 1, we use the negative binomial distribution to calculate the likelihood for the number of individuals of species s in a specific site given occupancy, $y_{i,s} \sim \text{negbinom1}(\mu_{i,s},\kappa)$, where $\mu_{i,s}$ is the expected value and κ is the overdispersion parameter. This is not the standard way of parametrizing the negative binomial distribution, so we designate it 'negbinom1' in Eq. 3–4 (compare with Eq. 7). We assume the same overdispersion for all species and formation as Reitan et al. (2022) suggested that overdispersion could not be distinguished among species. We also separate the expected value per subsample, $\lambda_{f(i),s}$, from T_i. The probability distribution of a single data point in an occupied site is then:

$$P_{\text{negbinom1}}\left(y_{i,s} \mid T_i\lambda_{f(i),s},\kappa\right) = \left(y_{i,s}+1/\kappa-1 y_{i,s}\right)\frac{\left(\lambda_{f(i),s}T_i\kappa\right)^{y_{i,s}}}{\left(1+\lambda_{f(i),s}T_i\kappa\right)^{y_{i,s}+1/\kappa}} \quad (3)$$

The expected value of this distribution is $\mu_{i,s}=\lambda_{f(i),s}T_i$ and the variance is $\lambda_{f(i),s}T_i(1+\kappa\lambda_{f(i),s}T_i)$. Thus, the closer the overdispersion is to zero, the closer the variance is to the expected value (as for the Poisson distribution).

16000587, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ecog.06866 by University Of Oslo, Wiley Online Library on [02/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

However, Eq. 3 assumes occupancy. If s does not occupy the site, the expected value will be zero and the only possible outcome is $y_{i,s}=0$. Let the independent probability of site-occupancy of each site belonging to a specific species s and formation $f(i)$ be designated $\Psi_{f(i),s}$. Then, the distribution of $y_{i,s}$ unconditioned on site-occupancy will be zero-inflated:

$$P_{\text{zero,negbinom1}}\left(y_{i,s}\,|\,T_i\lambda_{f(i),s},\kappa,\Psi_{f(i),s}\right)$$
$$=\left(1-\Psi_{f(i),s}\right)I\left(y_{i,s}=0\right)+\Psi_{f(i),s}\left(y_{i,s}+1/\kappa-1 y_{i,s}\right) \quad (4)$$
$$\frac{\left(\lambda_{f(i),s}T_i\kappa\right)^{y_{i,s}}}{\left(1+\lambda_{f(i),s}T_i\kappa\right)^{y_{i,s}+1/\kappa}}$$

Here, $I()$, is the indicator function, which is one if the statement inside the parenthesis is true, and zero, if false. We assume the superspecies occupies all sites.

A species can be absent from all sites in a region in the same formation, thus a non-independent lack of occupancy (Extension 4). We represent the presence/absence of s with a continuous variable $\omega_{f,s} \sim N(\mu=\Phi^{-1}(r), \sigma=1)$, but only for the species+formation combinations where we do not have external information that the species is present in the region (Extension 5). r represents the probability of regional presence for the set of species + formation combinations and $\Phi()$ is the cumulative distribution function of the standard normal distribution. We then define a binary variable,

$$\Omega_{f,s} = I(\omega_{f,s}>0 \text{ or } A_{f,s}=1), \quad (5)$$

which indicates whether the region is occupied, where $A_{f,s} \equiv I$ (external data sources tell that species s occupies formation f). Since $\omega_{f,s}$ is centered around $\Phi^{-1}(r)$, $\Omega_{f,s}=1$ with probability r whenever $A_{f,s}=0$. Since site-occupancy depends on regional occupancy, the expression $\Omega_{f,s}\Psi_{f,s}$ replaces $\Psi_{f,s}$ in the zero-inflation part of the likelihood component in Eq. 4. We then let r determine the distribution of $\omega_{f,s}$ for cases where $A_{f,s}=0$ and use likelihood r for the cases where $A_{f,s}=1$. Hence r will represent the probability for regional occupancy in total, rather than just regional occupancy for those cases where $A_{f,s}=0$. For each species-formation combination, the likelihood picks up a term

$$L_{f,s} \equiv I\left(A_{f,s}=1\right)r + I\left(A_{f,s}=0\right)f_N\left(\mu=\Phi^{-1}(r),\sigma=1\right), \quad (6)$$

where $f_{N()}$ is the probability density function of the normal distribution.

With unidentified-to-species-level individuals belonging to a genus, given that there are multiple species of that genus, (shortened as 'unidentified' and conversely as 'identified'), the probability of the combination of identified and unidentified individuals will be the product of the distribution of the identified individuals and the distribution of the unidentified individuals given the identified ones. The identified individuals are described by Eq. 4, though when taking into account the possibility of unidentified individuals, the expected value of identified individuals will be modified to $\gamma_{g,f,s}\lambda_{f,s}$ where $\gamma_{g,f}$ is the identification probability of an individual. The number of unidentified individuals, $U_{i,g}$, given the identified individuals, $I_{i,g}$, then follows the negative binomial distribution (see the Supporting information for details):

$$P\left(U_{i,g}|I_{i,g}\right)=\binom{U_{i,g}+I_{i,g}}{U_{i,g}}\gamma_{g,f(i)}^{I_{i,g}+1}\left(1-\gamma_{g,f(i)}\right)^{U_{i,g}} \quad (7)$$

Final likelihood expression

$$L=\left(\prod_{s=1}^{S}\prod_{f}^{F}L_{f,s}\prod_{i|f(i)=f}P_{zero,negbinom1}\right.$$
$$\left.\left(y_{i,s}\,|\,T_i\gamma_{g(s),f(i)}\lambda_{f(i),s},\kappa,\Omega_{f(i),s}\Psi_{f(i),s}\right)\right)$$
$$\left(\prod_{g\in UG}\prod_{s\in g}\prod_{i=1}^{\#\,sites}P\left(U_{i,s}|I_{i,g}\right)\right) \quad (8)$$

where UG is the set of genera that has unidentified individuals. Note that we now let the expected number of identified individuals for each species scale with identifiability probability of the genus it belongs to, $\gamma_{g,f(i)}$. We set $\gamma_{g,f}=1$ for each genus where there is no possibility for unidentified individuals.

The likelihood depends on the state of the random effects, both the common formation-dependent random effects for site-occupancy and abundance-given-occupancy respectively, v_f and u_f, as well as the species- and formation-dependent random effects for site-occupancy and abundance-given-occupancy respectively, $\delta_{f,s}$ and $\varepsilon_{f,s}$. The site-occupancy and abundance-given-occupancy component in the likelihood expression Eq. 9 are thus

$$\lambda_{f,s}(\theta)=\exp\left(\beta_s+u_{f(i)}+\varepsilon_{f(i),s}\right) \quad (9a)$$

$$\Psi_{i,s}(\theta)=I(s=S)+I(s<S)\,\text{logit}^{-1}\left(\alpha_s+v_{f(i)}+\delta_{f(i),s}\right) \quad (9b)$$

θ is the parameter set (random variables and top parameters, Fig. 1). Here, both abundance-given-occupancy and site-occupancy itself are decomposed into a species-dependent, a species + formation-dependent and a purely formation-dependent random variable, parallel to the original model Eq. 1a. The expression for relative abundance Eq. 2 is also retained.

Random effects

The random effects for species-dependent dynamics and common dynamics Eq. 1b are likewise retained in the new model.

16000587, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ecog.06866 by University Of Oslo, Wiley Online Library on [02/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

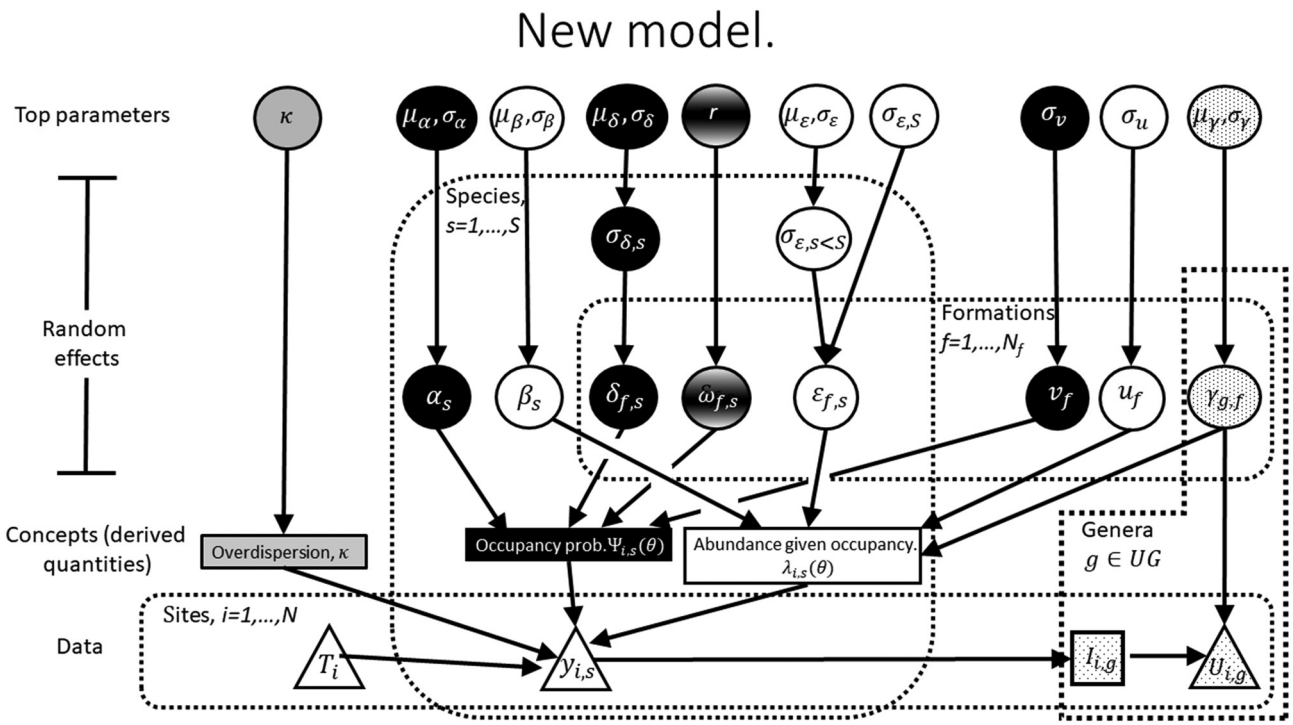

Figure 1. A schematic view of the new model. This overview shows the hierarchical relationships between data, the core components of the occupancy model, random effects and top parameters. The arrows show dependencies. Shapes with white background are associated with abundance-given-occupancy or base data (individual and subsample counts, excluding data associated with taxon identifiability probability). Shapes with black backgrounds are associated with occupancy (solid black for site-dependent occupancy and gradient black for regional occupancy). Shapes with grey backgrounds are associated with overdispersion. Lastly, shapes with dotted backgrounds are associated with taxon identifiability probabilities. Round shapes are parameters/random effects, rectangles are concepts expressed as functions and triangles are data. How the regional occupancy random effects, $\omega_{f,s}$, determines the regional occupancy states are not shown here (Eq. 5–6, 8). The functions $\lambda_{f,s}(\theta)$ and $\Psi_{i,s}(\theta)$ are expressed in Eq. 9. Note also $I_{i,g}$ is a sum of the species data, $y_{i,g}$, for each genus with unidentified colonies, shown as a separate entity because this aggregate is used in a separate part of the likelihood.

$$u_f \sim N\left(0,\sigma_u^2\right), v_f \sim N\left(0,\sigma_v^2\right)$$
$$\delta_{f,s} \sim N\left(0,\sigma_{\delta,s}^2\right), \varepsilon_{f,s} \sim N\left(0,\sigma_{\varepsilon,s}^2\right) \quad (10)$$

However, we also include new random effects for the species-dependent constants,

$$\alpha_s \sim N\left(\mu_\alpha,\sigma_\alpha^2\right), \beta_s \sim N\left(\mu_\beta,\sigma_\beta^2\right) \quad (11a)$$

$$\sigma_{\delta,s} \sim \log N\left(\mu_\delta,\sigma_\delta^2\right), \varepsilon_{\delta,s} \sim \log N\left(\mu_\varepsilon,\sigma_\varepsilon^2\right),$$
$$\text{for } s < S \text{ (superspecies exempted)} \quad (11b)$$

where the original species-dependent constants effects Eq. 11a and the size of the dynamics Eq. 11b are now both random factors. Note that the size of the superspecies dynamics for abundance-given-occupancy, $\sigma_{\varepsilon,S}$, is not part of this equation but is instead a top parameter. As the superspecies is an aggregate of many different species, it can be expected to be less dynamic than any single species. The information content of the superspecies is much greater than for any other species. We hence exclude it in Eq. 11 to avoid swamping of random effect parameters for species dynamics.

Since we have one identifiability probability for each combination of formation and genera with unidentified colonies, we let it be a random factor, just like the other components in our model that describes dynamics:

$$\text{logit}\left(\gamma_{g,f}\right) \sim N\left(\mu_\gamma,\sigma_\gamma^2\right). \quad (12)$$

Top parameters and prior distributions

With our current parametrization, the top parameters are

$$\theta_{top} = \left\{\mu_\alpha,\sigma_\alpha,\mu_\beta,\sigma_\beta,\mu_\delta,\sigma_\delta,\mu_\varepsilon,\sigma_\varepsilon,\sigma_u,\sigma_v,\sigma_{\varepsilon,S},r,\kappa,\mu_\gamma,\sigma_\gamma\right\}. \quad (13)$$

Note that this parameter set does not increase with an increasing number of species, so the number of top parameters is always 15. For comparison, the Reitan et al. (2022) model

16000587, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ecog.06866 by University Of Oslo, Wiley Online Library on [02/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

had 5 × S top parameters, which for our dataset, S = 21, would have translated to 105 top parameters. Even so, there was no way of dealing with the genera that have unidentified individuals in that model. For details of our choice of prior distributions and the robustness of our model to our choice of prior, see the Supporting information.

Simulation 1: new model performance

To explore the performance of the new model, specifically to examine the accuracy of the inference of not just relative abundance but site-occupancy, regional occupancy and abundance-given-occupancy using individual counts, we set up simulations. We also incorporated all the extensions, namely unidentified individuals, regional occupancy and extra sources pertaining to regional occupancy, in order to test whether the model was able to handle these challenges. See the Supporting information for details.

Simulation 2: are individual counts better than subsample counts?

We use a different set of simulations to test if individual counts perform measurably better than subsample count data (Extension 1). Here, our simulated datasets had a specified site-occupancy probability and abundance-given-occupancy, which gives the relative abundance. We sampled simulated data on the subsample level and then aggregated these to site-level in the form of both individual counts and subsample presence counts. We also wanted to see how well relative abundance estimated from simple ratios worked (i.e. 'raw estimates' as opposed to model estimates). We used the occupancy model from Reitan et al. (2022) for the subsample presence counts data and the new model described here for the individual counts. In addition, we used this set of simulations to examine the effect of different levels of observational error (i.e. missing individuals, double counting of individuals and misclassification of species). We judged how well these methods worked using the root-mean-squared-error (RMSE) of the relative abundances. See the Supporting information for details. All data and code are supplied in the folder 'RAMU-MSOM' at https://github.com/trondreitan/RAMU-MSOM.

Results

Simulation 1

The relative abundance estimates correspond well with the true relative abundance and respond well to regional absence (Fig. 2). The modelled relative abundance estimates had an RMSE ≈ 0.016. When the existence of unidentified individuals was ignored, RMSE ≈ 0.021. Thus, the effort to compensate for the unidentified individuals did improve accuracy, as hypothesized. Raw estimates had RMSE ≈ 0.024, both when attempting to compensate for unidentified individuals (by dividing by the ratio of unidentified individuals in each genus) and when not attempting this, suggesting that it is not so easy to do this type of compensation using raw estimates. One cannot expect the latter to converge to true values with increasing data size, though from theory alone we would expect raw estimates corrected for unidentified individuals to converge. However, with our current data volume, identifiability correction in raw estimates do not work better than those without such corrections. Even if the corrected raw estimates do converge, one would need 2.3 times as many data points (sites) to obtain errors as small as the model estimated ones, regardless of absolute data volume (assuming that the squared error is inversely proportional to the dataset size).

Site-occupancy dynamics are quite well-estimated for the most abundant species (first in each simulated genus) while the least abundance species (last in each simulated genus) which likewise had a very dynamic true site-occupancy trend, were not (e.g. compare G01_S01 and G01_S04 in Fig. 3). Although the site-occupancy dynamics of species with intermediate abundance (e.g. G01_S02 and G01_S03) are also not too well-captured by the estimates, some of it is absorbed into estimated abundance-given-occupancy (Supporting information). Regional occupancy probability was also sometimes estimated to be low for some species + formation combinations in particular datasets where there were no detections, even though the region was actually occupied. However, when looking at the average score over all datasets, the regional occupancy probabilities are reasonable (Supporting information).

Simulation 2

The RMSE of the relative abundance estimates were smallest for model estimates of individual count data (RMSE ≈ 0.023). Compared to the model estimates for individual count data, the RMSE's for raw estimates for individual count data, for model estimates for subsample count data and the raw estimates for subsample presence count data were 26, 59 and 285% higher, respectively. Thus, as hypothesized, the accuracy improved using the occupancy model instead of raw data, and when using individual counts instead of subsample count data. We would need 59, 153 and 1382% more data points for raw estimates on individual count data, model estimates on subsample count data and raw estimates on subsample count data, respectively, to lower the errors to the level of model estimates on individual count data. Here, we assume the standard error to be inversely proportional to the square root of the number of measurements. However, raw estimates on subsample count data cannot be expected to converge towards unbiased results when the number of data increases, as the ratio of subsamples having presence of a given species does not scale linearly with abundance-given-occupancy (Reitan et al. 2022).

The observational error simulations suggested that the relationship between the various RMSEs does not substantially change when the probability of observational errors increased. (see the Supporting information for details).

16000587, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ecog.06866 by University Of Oslo, Wiley Online Library on [02/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

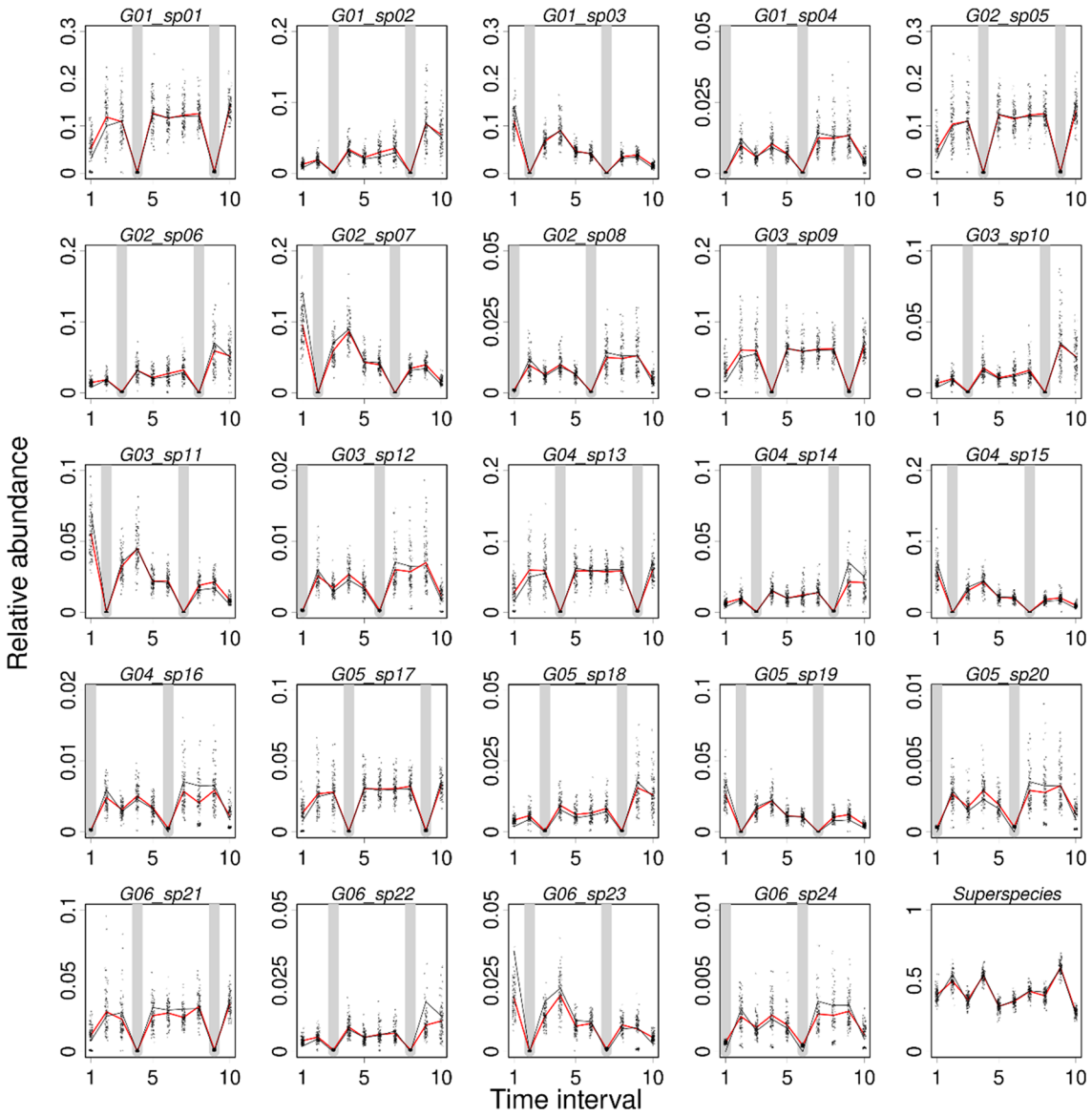

Figure 2. Relative abundance estimates for simulated data. Relative abundance estimates for simulated data (Simulation 1) for individual species are presented in each panel. Solid black lines = true values, red lines = average estimates from 100 simulations, dots = estimates for each simulated dataset, grey vertical bars = true regional absence. Note the different y-axes. The designated species names are shown on top of each panel.

Empirical results

While this work focuses on the details of the new model and simulations for understanding the performance of the model, it was of interest to ensure that the model has empirical relevance. Very briefly, there are clear species-specific temporal dynamics (i.e. non-overlapping credibility bands) in both estimated relative abundance (Fig. 4) and occupancy (Fig. 5) in our empirical dataset. The dynamics of relative abundance and occupancy are appreciably different for species within the same genera (e.g. compare *Microporella speculum*, *M. agonistes*, *M. discors*; compare *Escharoides excavata* and *E. angela*). Our model-estimated relative abundances are also robust to different prior widths (Supporting information).

Discussion

Hierarchical site-occupancy modelling is currently still rarely applied to paleoecological datasets, yet prevailing issues of incomplete detection in paleoecology is rampant, just like in ecological studies where occupancy modelling is more commonly applied. Replicate sampling and subsampling within formations is currently not standard practice in paleoecology. We have shown that there are measurable differences in face-value (raw ratios) and model estimates that will impact not just quantitative but also qualitative inferences. Unlike in the macroecological literature (Brown et al. 1984) where abundance and occupancy are separately measured and then tested for their relationships, we start from first principles (where

16000587, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ecog.06866 by University Of Oslo, Wiley Online Library on [02/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

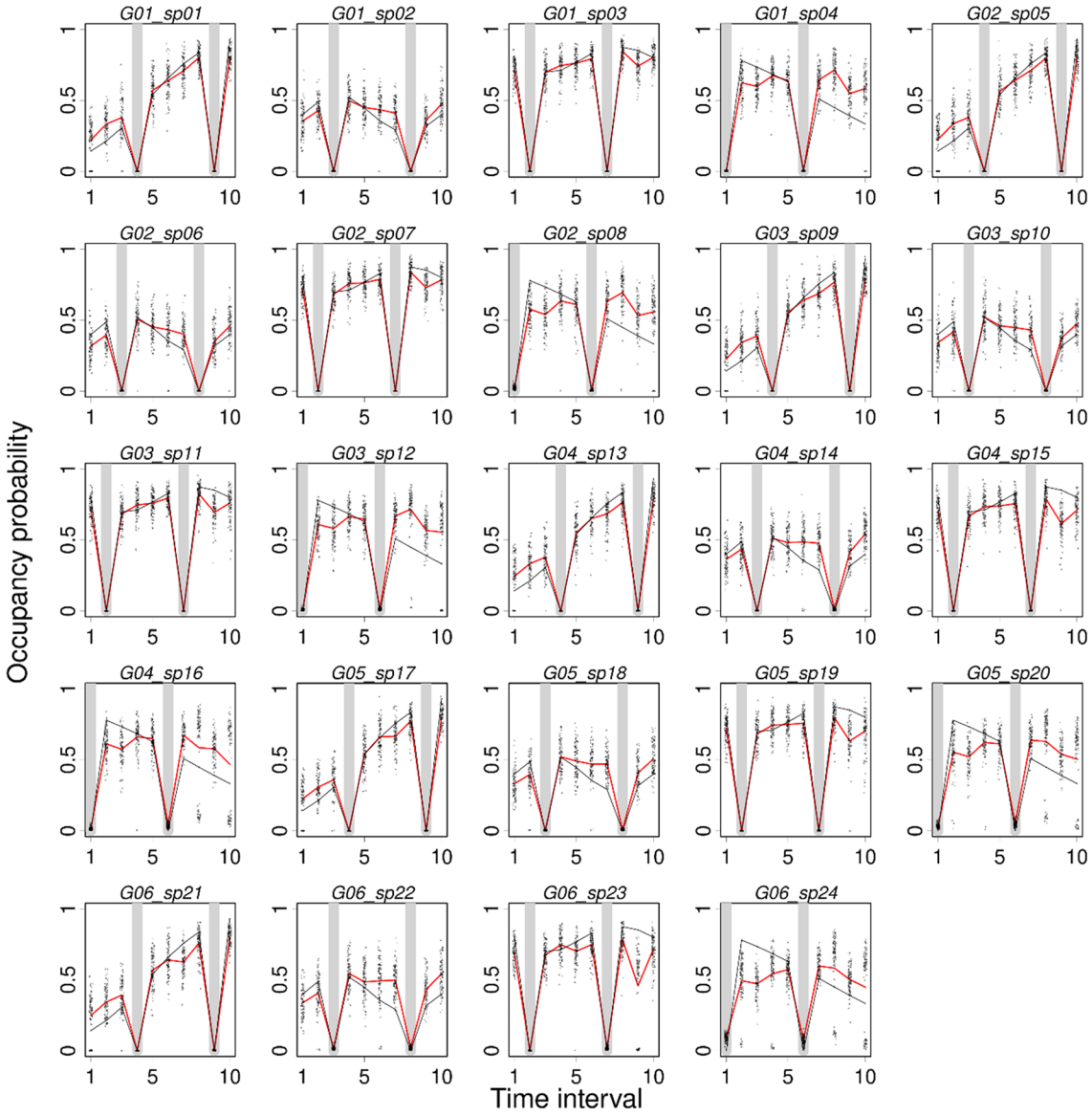

Figure 3. Occupancy probability estimates for simulated data. Occupancy estimates for simulated data (Simulation 1) for individual species are presented in each panel. Solid black lines = true values, red lines = average estimates from 100 simulations, dots = estimates for each simulated dataset, grey vertical bars = true regional absence. The designated species names are shown on top of each panel.

a site has to be occupied before considering abundance) to estimate these two related but different parameters within the same framework.

There is a practical need to strike a balance between the precision and accuracy of parameter estimation and the effort required for data collection in relative abundance estimation. For instance, it is quicker to count subsamples containing focal species, rather than painstakingly counting individuals of those species. This is a good strategy when habitat-use inference via occupancy-modelling is the end-goal (MacKenzie et al. 2017). However, as we hypothesized, much is gained in counting individuals rather than just occupied subsamples when estimating relative abundance. In addition, individual counts are crucial when there are individuals unassignable to genera, a situation common in paleoecology. A question to be considered is how many subsamples and sites are needed, although this will vary depending on the system. In general, increasing the number of subsamples (thus also increasing the total number of individuals found) and sites will help boosting the accuracy of the estimates. The latter has the advantage of increasing the accuracy of the occupancy probabilities as well as averaging out overdispersion. On the other hand, increasing the number of focal species will improve the accuracy of the distribution of random factors, while the accuracy of the abundance estimates themselves will improve only to a very limited extent. Hence, one might consider only adding the focal species of interest for the study/analysis.

Relative abundance and occupancy in our empirical data have clear temporal changes, i.e. dynamics are recovered

16000587, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ecog.06866 by University Of Oslo, Wiley Online Library on [02/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

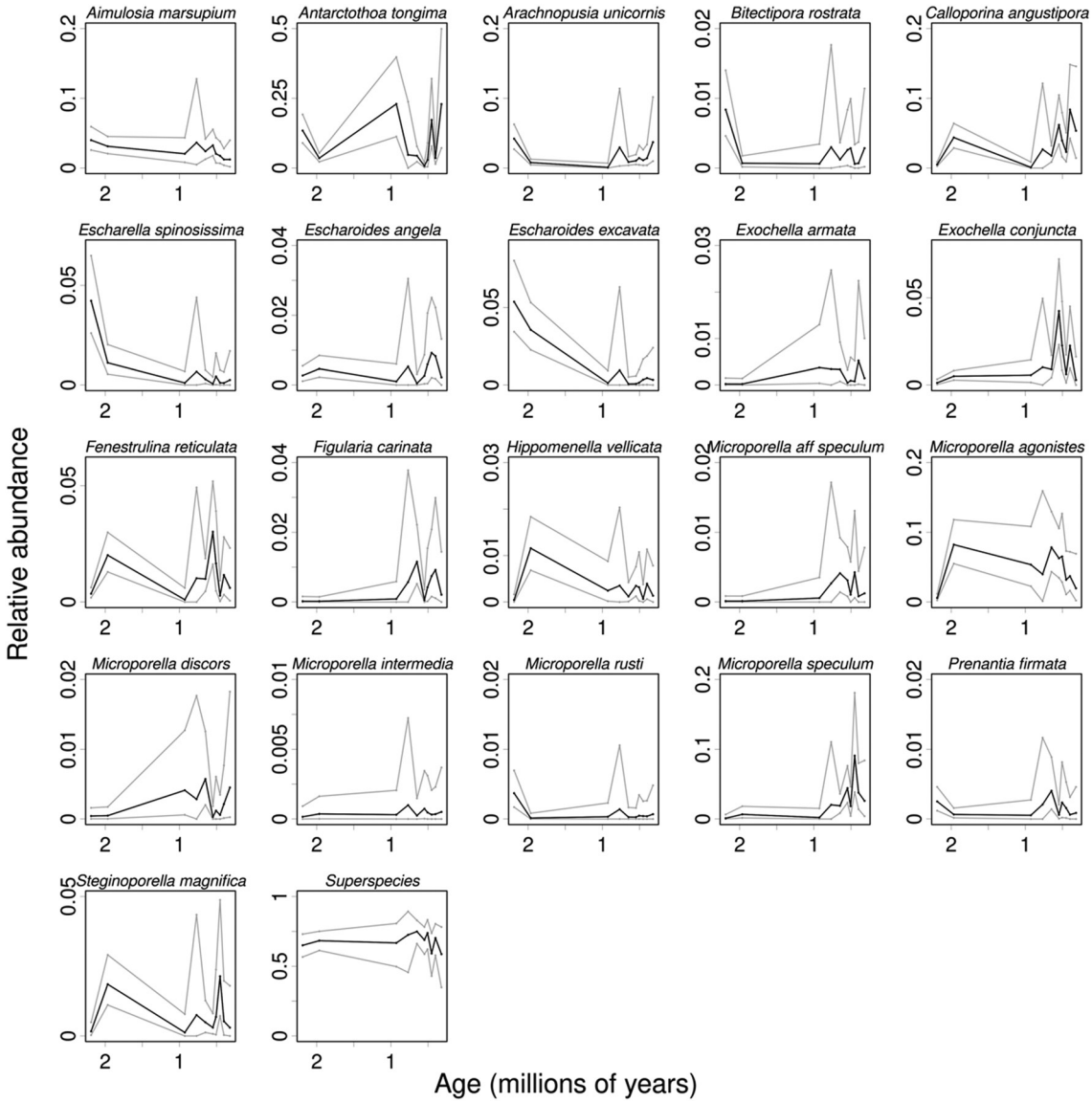

Figure 4. Relative abundance estimates for the empirical dataset. Each panel shows the point estimates (black lines) and 95% credibility bands (grey lines) for the 21 focal species plus the superspecies. Note the different y-axes.

(Fig. 4–5, e.g. species of *Microporella*). Note that the relative abundance estimates for a more abundant species will be more accurate compared to a less abundant species (e.g. compare *Microporella intermedia* with *Aimulosia marsupium* or the superspecies in Fig. 4).

As far as we are aware, ours is the first attempt at explicitly incorporating information on individual unassignable to species while estimating relative abundance and occupancy using paleontological data. Encouragingly, our simulations show that we can recover relative abundance dynamics by explicitly incorporating information on individuals identified to genus- but not species-level. Even if there are many unidentified individuals belonging to a certain genus whose species could potentially be identified as A, B and C, but no specimens of C are recognized, these data are still useful as long as the existence of species C is acknowledged in the modelling. If the existence of species C is unrecognized and C is also very common, the estimates will be biased. However, if C is unrecognized and rare, only a small bias will be introduced, a likely case for a well-studied system like our empirical data.

While we argue that one should always account for detection probabilities, some practical advice is perhaps necessary. As alluded to in methods and implicit in the assumptions of the model, the approach cannot be soundly applied to a community of organisms with highly heterogeneous preservation modes or probabilities (e.g. jellyfish and bivalves at the same time). Each fossil observation should approximate an 'individual' (e.g. an echinoid test) rather than fragments of an individual (e.g. spines of an echinoid). Caveats are evident from both simulations and the empirical data analyses. Most notably, as already alluded to, dynamics are most recoverable for species that are most commonly observed (i.e. the

16000587, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ecog.06866 by University Of Oslo, Wiley Online Library on [02/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

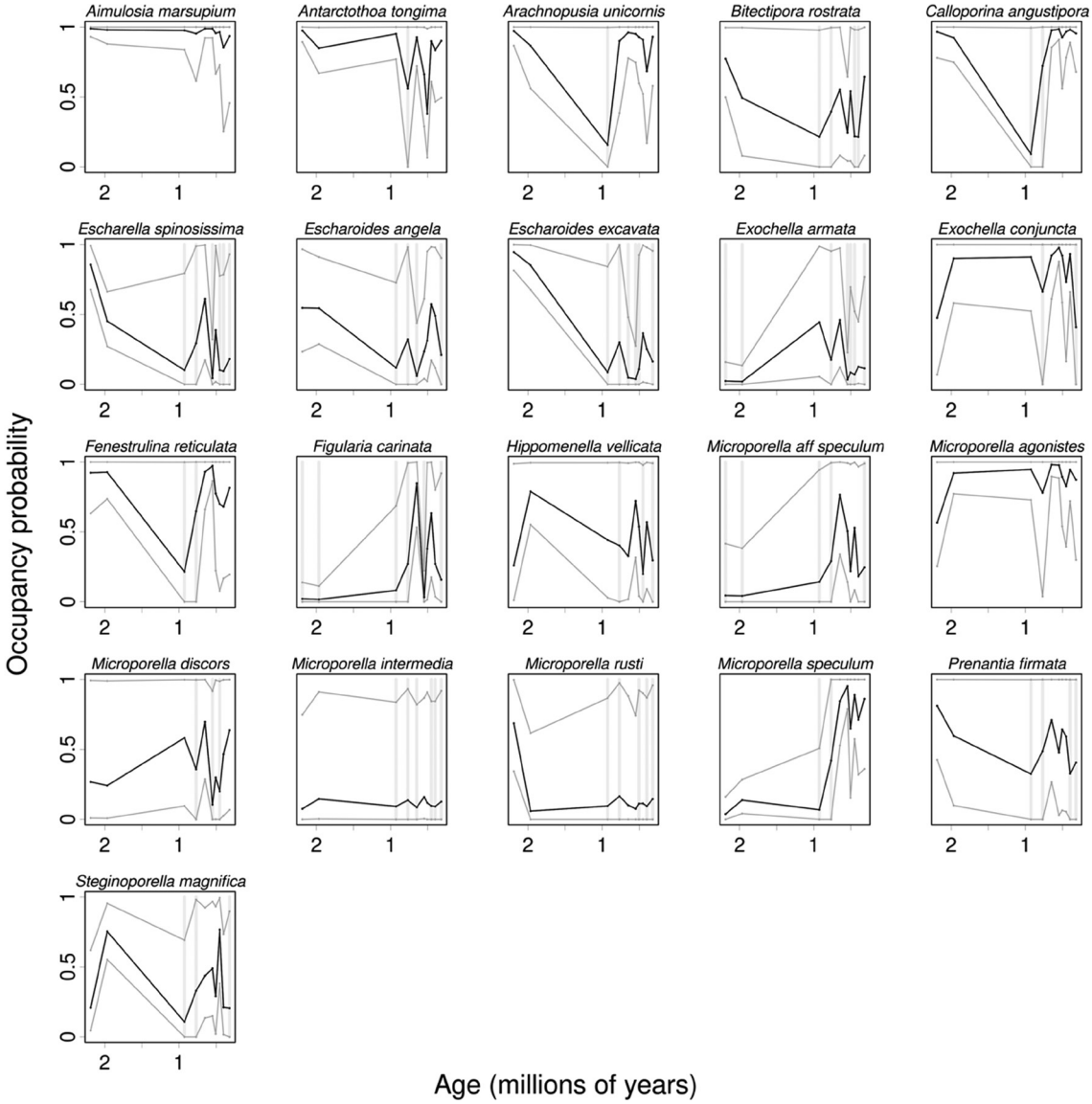

Figure 5. Occupancy estimates for the empirical dataset. Each panel shows the point estimates (black lines) and 95% credibility bands (grey lines) for the 21 focal species. Light grey bars indicate no detections in that formation and no indication of presence from external sources, i.e. situations where regional absence is a real possibility. Note the different *y*-axes.

most prevalent species) in the simulations, hence we have to assume that this is the case also for the empirical dataset. That said, less prevalent species still contribute to information important for estimation of more prevalent species through parameters common to all species. How important regional occupancy modelling is depends both on the occupancy data and the 'external information' available, which will vary from dataset to dataset. In any case, evidence for regional absence in our empirical system is weak (Fig. 5), as can be seen from our top parameter posterior distributions and robustness analyses (Supporting information). Absence is in general more difficult to infer than presence, since some observed absences are due to detection probability rather than true absence. But absence is not impossible to estimate, as we have shown.

Lest one erroneously concludes that a simpler model can be used for estimating relative abundance in a given area, let us be clear that site-occupancy modelling that teases apart occupancy and detection is a necessary component in estimating abundance. Additionally, one in general does not know whether regional absence is possible before analysing the empirical occupancy data. It is important to replicate sampling in ways that will capture variation in detection since absence of information cannot be proof of absence. In our case, we found clear indications of site absence, but not regional absence. Our model can be applied more widely in paleoecology than is perhaps apparent with our example empirical dataset. For instance, deep-sea cores can be subsampled within time-intervals, as estimated by a combination of depth information and age-models based on sedimentation

16000587, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ecog.06866 by University Of Oslo, Wiley Online Library on [02/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

rates, as can be lake sediment cores. More generally, any regional system where multiple outcrops in which sampling can be replicated will be amendable to this occupancy modelling. We recommend subsampling/replicate-sampling sites within formations/time-intervals for occupancy and abundance estimation for paleoecological systems, even when multiple sites cannot easily be sampled within formations. We also urge detailed documentation of individuals. These data, while requiring a bit more work to collect, can yield vastly better estimates of key ecological parameters.

Acknowledgements – We would like to thank M. H. Ramsfjell for contributing to species identifications and curation and for suggesting the availability of external data for regional occupancy. We would also like to thank the field parties of WABO I, II and III for fossil collection, with special thanks to Alan Beu (GNS, New Zealand) for his insights on the Wanganui Basin, GNS for export permits and P. Nordenhaug for collecting data on shell weights. We thank D. Polly for constructive criticisms.
Funding – This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 724324 toLHL).

Author contributions

Trond Reitan: Conceptualization (equal); Formal analysis (lead); Methodology (lead); Software (lead); Writing – original draft (equal); Writing – review and editing (equal). **Emanuela Di Martino**: Data curation (lead); Investigation (supporting); Writing – original draft (supporting); Writing – review and editing (equal). **Lee Hsiang Liow**: Conceptualization (equal); Data curation (supporting); Formal analysis (supporting); Funding acquisition (lead); Methodology (supporting);, Project administration (lead); Writing – original draft (equal); Writing – review and editing (equal).

Transparent peer review

The peer review history for this article is available at https://publons.com/publon/10.1111/ecog.06866.

Data availability statement

Data are available from the Github Repositoryy: https://github.com/trondreitan/RAMU-MSOM (Reitan et al. 2023).

Supporting information

The Supporting information associated with this article is available with the online version.

References

Brown, J. H. 1984. On the relationship between abundance and distribution of species. – Am. Nat. 124: 255–279.

Carter, R. M. and Naish, T. R. 1998. A review of Wanganui Basin, New Zealand: global reference section for shallow marine, Plio-Pleistocene (2.5–0 Ma) cyclostratigraphy. – Sediment. Geol. 122: 37–52.

Dillon, E. M. et al. 2022. What is conservation paleobiology? Tracking 20 years of research and development. – Front. Ecol. Evol. 10: 1031483.

Dussex, N. et al. 2021. Integrating multi-taxon palaeogenomes and sedimentary ancient DNA to study past ecosystem dynamics. – Proc. R. Soc. B 288: 20211252.

Foote, M. 2016 On the measurement of occupancy in ecology and paleontology. – Am. Nat. 42: 707–729.

Hoban, S., Dawson, A., Robinson, J. D., Smith, A. B. and Strand, A. E. 2019. Inference of biogeographic history by formally integrating distinct lines of evidence: genetic, environmental niche and fossil. – Ecography 42: 1991–2011.

Jackson, J. B. C. and Cheetham, A. H. 1990. Evolutionary significance of morphospecies – a test with cheilostome Bryozoa. – Science 248: 579–583.

Kidwell, S. M. 2001 Preservation of species abundance in marine death assemblages. – Science 294: 1091–1094.

Lawing, A. M., Blois, J. L., Maguire, K. C., Goring, S. J., Wang, Y. and McGuire, J. L. 2021. Occupancy models reveal regional differences in detectability and improve relative abundance estimations in fossil pollen assemblages. – Q. Sci. Rev. 253: 106747.

Liow, L. H. 2013. Simultaneous estimation of occupancy and detection probabilities: an illustration using Cincinnatian brachiopods. – Paleobiology 39: 193–213.

Liow, L. H., Di Martino, E., Voje, K. L., Rust, S. and Taylor, P. D. 2016. Interspecific interactions through 2 million years: are competitive outcomes predictable? – Proc. R. Soc. B 283: 20160981.

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Royle, J. A. and Langtimm, C. A. 2002. Estimating site occupancy rates when detection probabilities are less than one. – Ecology 83: 2248–2255.

MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L. and Hines, J. E. 2017. Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence. – Academic Press.

Nichols, J. D. and Pollock, K. H. 1983. Estimating taxonomic diversity, extinction rates, and speciation rates from fossil data using capture–recapture models. – Paleobiology 9:150–163.

Pillans, B. 2017. Quaternary stratigraphy of Whanganui Basin – a globally significant archive. – In: Shulmeister, J. (ed.), Landscape and quaternary environmental change in New Zealand. Atlantis Press, pp. 141–170.

Proust, J. N., Lamarche, G., Nodder, S. and Kamp, P. J. J. 2005. Sedimentary architecture of a Plio-Pleistocene proto-back-arc basin: Wanganui Basin, New Zealand. – Sediment. Geol. 181: 107–145.

Reitan, T., Ergon, T. H. and Liow, L. H. 2022. Relative species abundance and population densities of the past: developing multispecies occupancy models for fossil data. – Paleobiology 49: 23–38.

Reitan, T., Di Martino, E. and Liow, L. H. 2023. Data from: Estimating relative species abundance using fossil data identified to different taxonomic levels. – Github Repository, https://github.com/trondreitan/RAMU-MSOM.

16000587, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ecog.06866 by University Of Oslo, Wiley Online Library on [02/11/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License