Unifying occupancy-detection and local frequency scaling (Frescalo) models

Oliver L. Pescott*

[a] UK Centre for Ecology and Hydrology, Benson Lane, Crowmarsh Gifford, OX10 8BB, Oxfordshire, UK

Abstract

Frescalo's "local frequency scaling" and classical occupancy-detection models both seek to recover true species-occurrence signals from imperfect data. In this paper, we show that the two approaches rest on the same underlying detection mathematics. Occupancy models treat each site's repeat visits as independent detection trials and separately estimate occupancy probability and per-visit detectability. Frescalo, by contrast, pools data across ecologically defined neighbourhoods and infers a single combined detection rate and a temporal "time-factor" to capture trends. We demonstrate that the Bernoulli-trial formulation of occupancy-detection converges to Frescalo's Poisson-process framework, with occupancy and detectability collapsing into a single rate parameter. This equivalence clarifies how Frescalo's neighbourhood and time corrections function as a coarser-scale analogue of repeat-visit models. By casting Frescalo in occupancy modelling terms, we hope to promote further investigation into the adoption of occupancy-model diagnostics, extensions and covariate tests within Frescalo analyses, improving transparency and rigour when working with opportunistic biodiversity data.

Keywords: occupancy models, sampling effort, effort correction, citizen science, unstructured data, Frescalo

1. Introduction

Occupancy-detection models [9] and the Frescalo "local frequency scaling" method [6] both aim to correct raw biological records (i.e. species occurrence) data for imperfect sampling. Classical occupancy models do this at the scale of repeated visits to individual sites, explicitly estimating true presence probabilities (ψ) and detectability (p) via a hierarchical likelihood. Frescalo was designed to work at larger spatio-temporal scales, exploiting emergent patterns of relative frequency in "neighbourhoods" to derive Poisson-process-based scaling factors (α) and species' relative "time factors" indexing true fluctuations in site occupancy. Given that many datasets lack repeat-visit structure, and/or may exhibit variation in the detection process that is unmodellable due to a

*Corresponding author
Email address: olipes@ceh.ac.uk (Oliver L. Pescott)

lack of knowledge of its determinants [11], understanding how Frescalo recovers effort-adjusted trends from aggregated data can broaden the toolkit of ecologists.

Although the two models can appear quite different, Pescott et al. [11] informally suggested that Frescalo could be seen as a type of occupancy-detection model "where an adjustment for overlooked species is made in relation to spatial rather than temporal replication, whilst simultaneously adjusting for variable regional effort". We here show that this suggestion can be formalised due to the two model types' reliance on the same core mathematics of Bernoulli versus Poisson detections. Below we (1) recall each framework, (2) write down their key equations, and (3) algebraically map one onto the other, demonstrating that Frescalo time trends are based on an implicit occupancy-detection model whose "visits" and "occupancy" are folded into a single site/species discoverability rate parameter λ and standardised neighbourhood effort index.

2. Occupancy-detection models

2.1. Basic single-season model

Following MacKenzie et al. [9], at each site i for species j assume a latent occupancy indicator

$$z_{ij} \sim Bernoulli(\psi_{ij}).$$

Conditional on presence, v total survey visits indexed by k produce

$$y_{ij1}, ..., y_{ijv} \mid z_{ij} = 1 \sim Bernoulli(p_{ij})$$

where p is detectability. If $z_{ij} = 0$ (i.e. species absent), then all $y_{ijk} = 0$. Marginalising out z_{ij}, it is well-known that the probability of at least one detection across v visits is

$$P(\max_k y_{ijk} = 1) = \psi_{ij}[1 - (1 - p_{ij})^v].$$

Thus the model simultaneously estimates

$$\psi_{ij} = \Pr(\text{occupied}), \quad p_{ij} = \Pr(\text{detect}|\text{occupied}),$$

and inference proceeds via the full likelihood over all sites and detection histories.

3. Frequency scaling using local occupancy (Frescalo)

3.1. Neighbourhood frequencies

Frescalo [6] pools presence-only data across a neighbourhood around target site i. We denote the observed proportion of neighbourhood sites in which species j was recorded by f_{ij} (in practice this frequency may relate to a weighted neighbourhood as per Hill [6], but this detail is not crucial for what follows). Under a Poisson-process model of species discovery with rate λ_{ij} and unknown total neighbourhood-level sampling effort $s_{i(N)}$, one has

$$f_{ij} = 1 - \exp(-\lambda_{ij} s_{i(N)}).$$

Subsequently, a frequency-weighted neighbourhood index

$$\phi_i = \frac{\sum_j f_{ij}^2}{\sum_j f_{ij}}$$

is then "standardised" to a target value Φ by solving for a site-specific effort multiplier α_i such that

$$\phi_i(\alpha_i) = \frac{\sum_j [1 - (1 - f_{ij})^{\alpha_i}]^2}{\sum_j [1 - (1 - f_{ij})^{\alpha_i}]} = \Phi.$$

Mathematically, Φ is chosen so that every neighbourhood's weighted-mean frequency $\phi_i = \sum_j f_{ij}^2 / \sum_j f_{ij}$ equals Φ. Hill [6] showed that ϕ_i is the ratio of the mean species richness to the 'effective number of common species' (often called N_2, the reciprocal of Simpson's index; Hill [5]), which means that ϕ_i isolates sampling intensity from true differences in richness and evenness. By fixing $\phi_i = \Phi$, we therefore align all neighbourhoods to the same effort scale without erasing real ecological differences.

This process yields the standardised neighbourhood frequencies

$$\tilde{f}_{ij} = 1 - (1 - f_{ij})^{\alpha_i}$$

which are independent of time (i.e. they are calculated with respect to the entire time period under consideration, rather than any subdivisions of this used for trend calculations), and serve as a proxy for the "true" discoverability- or effort-standardised neighbourhood species rank-frequency curve.

3.2. Temporal correction

Within each time period t, one chooses a set of "benchmark" species [8] and computes the proportion recorded per site and time period (Hill's s_{it}) as an index of site-level recording effort. (Note that there are potentially many ways to choose ones' neighbourhood benchmarks, but Hill [6] proposed a fixed proportion R^* of the standardised species rank-frequency curve after an additional normalisation step involving the division of species' ranks by the expected species count $\sum_j \tilde{f}_{ij}$; however, the precise method of choosing benchmarks does not affect what follows). For each species j in period t, Hill then defines a Poisson-link intensity

$$Q_{ijt} = -\ln[1 - s_{it}\tilde{f}_{ij}],$$

The modelled "discovery" probability is then

$$P_{ijt}(x_{jt}) = 1 - \exp(-Q_{ijt}x_{jt}).$$

Hill [6] estimates the time-factor x_{jt} by matching the *total* modelled to *total* observed presences y_{ijt}:

$$\sum_i y_{ijt} = \sum_i P_{ijt}(x_{jt}).$$

In practice one iterates x_{jt} in the exact Poisson form above until those sums coincide (e.g. see the *R* code of Pescott [10]). The difference between the (summed) observed presences y_{ijt} and the model's baseline expectation after standardising time-independent neighbourhood effort α_i and adjusting for site/time specific effort s_{it} is therefore captured by the time factor x_{jt}. Frescalo can thus deliver detection-corrected trends from unstructured data when its core assumptions are met.

4. Bridging the gap

4.1. Static occupancy and detection

We can compare the static (i.e. single season) single-species occupancy-detection model probability of at least one detection in v visits

$$\psi[1-(1-p)^v]$$

with the Poisson-process discovery rate (conditional on the all-time frequency curve) used in Frescalo

$$1-e^{-\lambda s_{i(N)}}.$$

For small pv, $(1-p)^v \approx e^{-pv}$, hence $\psi[1-\exp(-pv)] \approx 1-\exp(-\psi pv)$; now identifying $\lambda = \psi p$ and $v = s_{i(N)}$ recovers the approximate Frescalo detection probability $1-\exp(-\lambda s_{i(N)})$. For any value of p, exact equivalence can be found by solving

$$1-e^{-\lambda v} = \psi[1-(1-p)^v]$$

for

$$\lambda = -\frac{1}{v}\ln[1-\psi(1-(1-p)^v)],$$

but this only reduces to ψp in the limit $pv \to 0$. Frescalo's λ therefore combines occupancy and detectability into one Poisson rate, approximating the occupancy-detection model probability of at least one detection across v visits. Occupancy ψ and per-visit detectability p collapse into $\lambda_{ij} = \psi_{ij}p_{ij}$ and the unobserved number of site visits v in the Frescalo context becomes the continuous neighbourhood effort index $s_{i(N)}$, standardised across neighbourhoods via α_i. Frescalo can therefore be interpreted as an occupancy-detection analogue at the neighbourhood scale: it replaces the two-parameters (ψ, p) and discrete v with a Poisson rate λ and a continuous effort-multiplier α equalising variable survey effort s_i (inferred by the neighbourhood level $s_{i(N)}$) across sites.

A key step in recognising the equivalent elements of these models is to appreciate that Frescalo applies its discoverability standardisation at a large scale: not only is the adjustment done with respect to the multi-site neighbourhood and across all species, but it is also calculated across all time periods in the analysis. The standardised neighbourhood frequencies $\tilde{f}_{ij}$ and the species rank-frequency curve they form is estimated once, independently of time, before temporal change is examined.

4.2. Time trend interpretation

A time trend in occupancy derived from a classical occupancy-detection model is modelled simply by letting ψ_{ij} vary linearly or non-linearly over time, conditional on both model-specific [16] and other standard survey sampling assumptions [3] being reasonable. Frescalo, by contrast, posits a single time-independent set of discoverability-adjusted baseline frequencies $\tilde{f}_{ij}$, and then uses benchmarks and the site/period effort index s_{it} to compute expected frequencies under an assumption of stasis, subsequently letting the time factors x_{jt} absorb any residual differences as true ecological change.

This underscores a key difference in how effort-adjustment processes function in each model type. Occupancy-detection models assume that true site occupancies, and so trends in these, are directly recoverable from visit-level information. Frescalo assumes that fine-scale visit data is generally unavailable and/or uninformative for all or part of the time series of interest, and so models species' discoverability at a much larger scale. The main aim of this adjustment is to ensure a common scale across which neighbourhoods, and therefore sites, can be compared: without the harmonisation of effort across neighbourhoods, the time factors estimated for each site for a species would not be comparable, making average time factors and trends in these meaningless.

Another fundamental difference is the meaning of the site occupancy values produced. As noted, ψ_i has the simple meaning of predicted site occupancy under the classical model (notwithstanding debates around usage versus occupancy when these types of models are applied at different scales; [14]). The Frescalo time factor x_{jt} is, however, defined by the benchmark average, and values >1 or <1 indicate that a species is at a higher or lower average frequency relative to the common species where it occurs, rather than in absolute occupancy probability. This may be an important limitation to inferring effort via observable recording outcomes, rather than having knowledge of those factors that directly map onto effort, such as the actual number of visits and covariates that are known to explain an important portion of observed variance in species' visit-level detectability [6, 15, 7].

One way around this issue is the observation of Bijlsma [1] that site occupancy probabilities can actually be derived from Frescalo via the combination of the standardised species' frequencies $\tilde{f}_{ij}$ and the time-factors x_{jt}, and this has been exploited in at least one published analysis [4]. However, this requires a note of caution: whilst sensitivity analyses published in Hill [6] suggested that the trends in time-factors estimated by Frescalo can be relatively insensitive to the choice of R^*, the benchmark threshold (variation in this parameter changing the intercept of estimated trends but not their slope), the same is not true of back-calculated site occupancy probabilities. Because the relationship between time-factors and species frequencies is non-linear, the shifts in time trend intercept seen using different values of R^* will not translate into the same proportional changes in predicted site occupancies over time (Pescott, pers. obs.) This may be particularly important when these trends are used to classify species' into risk categories, as for example happens in Red Listing exercises [e.g. 12].

5. Conclusions

Unstructured species occurrence data are too valuable to ignore, especially for historical periods where no information about the visit-level data collection process survives [11]. Hill's "frequency scaling using local occupancy" or Frescalo method allows the careful analyst to infer a large-scale detectability or effort metric that can subsequently be used to place neighbourhoods on a common footing for the estimation of distribution trends. The large-scale formulation of this approach not only allows for the potential inclusion of more data sources (e.g. records extracted from Atlases or museums), but may also act to reduce the actual error in species' trends intrinsically [2, 13].

By demonstrating how Frescalo collapses the classical occupancy-detection model's ψ and p into λ, and how it infers visit-related effort via an emergent community-level mean rate, the approach performs an occupancy-detection-type correction even when explicit or informative temporal repeat-visit data are lacking. By highlighting this link, we hope to promote the development of additional diagnostics, extensions and more rigorous uncertainty quantification for the frequency scaling using local occupancy method.

References

[1] Bijlsma, R.J., 2013. The Estimation of Species Richness of Dutch Bryophytes between 1900 and 2011. Technical Report 15. Dutch Bryological and Lichenological Society. Gouda, the Netherlands.

[2] Boyd, R.J., Bowler, D.E., Isaac, N.J.B., Pescott, O.L., 2024. On the trade-off between accuracy and spatial resolution when estimating species occupancy from geographically biased samples. Ecological Modelling 493, 110739. doi:10.1016/j.ecolmodel.2024.110739.

[3] Boyd, R.J., Powney, G.D., Pescott, O.L., 2023. We need to talk about nonprobability samples. Trends in Ecology & Evolution doi:10.1016/j.tree.2023.01.001.

[4] Eichenberg, D., Bowler, D.E., Bonn, A., Bruelheide, H., Grescho, V., Harter, D., Jandt, U., May, R., Winter, M., Jansen, F., 2021. Widespread decline in Central European plant diversity across six decades. Global Change Biology 27, 1097–1110. doi:10.1111/gcb.15447.

[5] Hill, M.O., 1973. Diversity and evenness: A unifying notation and its consequences. Ecology 54, 427–432.

[6] Hill, M.O., 2012. Local frequency as a key to interpreting species occurrence data when recording effort is not known. Methods in Ecology and Evolution 3, 195–205.

[7] Kéry, M., Royle, J.A., Schmid, H., Schaub, M., Volet, B., Häfliger, G., Zbinden, N., 2010. Site-Occupancy Distribution Modeling to Correct Population-Trend Estimates Derived from Opportunistic Observations. Conservation Biology 24, 1388–1397. doi:10.1111/j.1523-1739.2010.01479.x.

[8] Latour, J., van Swaay, C., 1992. Dagvlinders als indicatoren voor de regionale milieukwaliteit. De Levende Natuur 93, 19–22.

[9] MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, J.A., Langtimm, C.A., 2002. Estimating Site Occupancy Rates When Detection Probabilities Are Less Than One. Ecology 83, 2248–2255. doi:10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2.

[10] Pescott, O., 2025. An R translation of Hill's fortran code for Frescalo. doi:10.5281/zenodo.15305437.

[11] Pescott, O.L., Humphrey, T.A., Stroh, P.A., Walker, K.J., 2019. Temporal changes in distributions and the species atlas: How can British and Irish plant data shoulder the inferential burden? British & Irish Botany 1, 250–282. doi:10.33928/bib.2019.01.250.

[12] Stroh, P., Leach, S., August, T., Walker, K., Pearman, D., Rumsey, F., Harrower, C., Fay, M., Martin, J., Pankhurst, T., Preston, C., Taylor, I., 2014. A Vascular Plant Red List for England. Botanical Society of Britain and Ireland, Bristol.

[13] Stroh, P., Walker, K., Humphrey, T., Pescott, O., Burkmar, R. (Eds.), 2023. Plant Atlas 2020. Mapping Changes in the Distribution of the British and Irish Flora. Botanical Society of Britain and Ireland & Princeton University Press, Princeton.

[14] Valente, J.J., Jirinec, V., Leu, M., 2024. Thinking beyond the closure assumption: Designing surveys for estimating biological truth with occupancy models. Methods in Ecology and Evolution 15, 2289–2300. doi:10.1111/2041-210X.14439.

[15] van Strien, A.J., van Swaay, C.A., Termaat, T., 2013. Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. Journal of Applied Ecology 50, 1450–1458. doi:10.1111/1365-2664.12158.

[16] Warton, D.I., Stoklosa, J., Guillera-Arroita, G., MacKenzie, D.I., Welsh, A.H., 2017. Graphical diagnostics for occupancy models with imperfect detection. Methods in Ecology and Evolution 8, 408–419. doi:10.1111/2041-210X.12761.