Supplementary Information 1

Kwaku Peprah Adjei

Introduction

An overlooked step in developing integrated distribution models is fitting models for each dataset to explore which models best fit each dataset. IDM takes advantage of the shared parameters to get better inference, and is not an escape from doing exploratory analysis on each dataset.

This supplementary material performs exploratory analysis on the UK pollinator monitoring scheme (Breeze et al. 2021). As mentioned in the main paper, we have data from three insect groups: bumblebees, solitarybees and hoverflies; and the models are fit independently for each insect group. The exploration is therefore also done for each insect group.

A Poisson and negative binomial generalised linear (mixed) models (GLMM hereafter) were fit for each insect group in the group counts dataset and binomial GLMM to the species occupancy data. The Poisson and logistic regression are the simplest model to fit to the count and occupancy data respectively (Fahrmeir et al. 2022). When there is overdispersion in the count data, the negative binomial regression will fit the count data better (Fahrmeir et al. 2022). The negative binomial regression adds an extra parameter ϕ to model the extra variation in the data. Values of ψ closer to 0 indicate overdispersion in the count data. A likelihood ratio test of significance for the overdispersion parameter was performed. For each insect group, the log-likelihoods from the Poisson and negative Binomial regression models ($l(\hat{\beta}_{NB})$ and $l(\hat{\beta}_{Poi})$ respectively) using the negative binomial regression model were retrieved and test statistic is:

$$LRT = -2 \times (l(\hat{\beta}_{NB}) - l(\hat{\beta}_{Poi})) \sim \chi^2_1.$$

The negative binomial distribution is chosen for the count data if the p-value is less than 0.05 and Poisson is chosen when the p-value if greater than 0.05.

We performed stepwise model selection technique to obtain the best model for the count and occupancy data. Four model defined in Equation 1 are fit for the exploration analysis. The first model M_0 is an intercept-only model; the second model M_1 has an intercept and latitudinal gradient as a covariate; the third model M_2 adds visits as a random effect to model M_1; and

Table S1-1: Coefficients from the best models (model equations defined above for BestModelIndex) for the group count and species occupancy data. The BestModelIndex shows which of the model fitted were selected as the best model (where best model is defined as the fitted model with the lowest BIC value); the Intercept and Latitude are fixed effect estimates; VisitVar and SpeciesVar are visit and species random effect variance estimates respectively; Psi is the overdispersion parameter of the negative Binomial and the P.value is the p-value estimate from likelihood ration test of overdispersion (the difference between the Poisson and Negative Binomial models for each insect group).

Model	Data	BestModel	Intercept	Latitude	VisitVar	SpeciesVar	Psi	P.value
Bumblebees								
Binomial	Occupancy	M3	-4.54	0.08	0.07	1.95	-	-
Negative Binomial	Counts	M1	0.14	-0.38	-	-	1.42	0
Hoverflies								
Binomial	Occupancy	M3	-5.16	0.23	0.02	2.70	-	-
Negative Binomial	Counts	M2	0.96	-0.13	0.03	-	1.36	0
Solitarybees								
Binomial	Occupancy	M3	-5.99	-0.82	0.09	1.49	-	-
Negative Binomial	Counts	M1	-0.99	-0.53	-	-	1.37	0

the last model M_3 adds species as a random effect to model M_2. Model M_3 was for the species occupancy data only, but the others were for both datasets. We fit the models M_0 and M_1 with the lme4 package (Bates et al. 2009) and M_2 and M_3 with the MASS package (Venables and Ripley 2002) in R. The models were compared using their BIC values. We chose BIC over AIC values because they have high penalty for model complexity. The best model was the model with the lowest BIC value.

$$\begin{aligned}
M_0 &: \text{intercept} \\
M_1 &: \text{intercept} + \text{latitude} \\
M_2 &: \text{intercept} + \text{latitude} + (1|visit) \\
M_3 &: \text{intercept} + \text{latitude} + (1|visit) + (1|species)
\end{aligned} \tag{1}$$

Results and Interpretation

Figures Figure 1, Figure 3 and Figure 5 shows the distribution the bumblebees, hoverflies and solitarybees group count data for each visit to the 74 PoMS sites; Figures Figure 2, Figure 4 and Figure 6 shows the distribution of the proportion of species occupying a given PoMS sites at each survey visit. The figures Figure 1 to Figure 6 indicate a significant missing observation across the strudy regions and visits; with possible overdispersed group counts. The best models for the group count data showed significant overdispersion ($\phi_{BB} = 0.53$, $\phi_{HV} = 0.66$ and $\phi_{SB} = 0.37$ with p-values closer to 0; Table 1).

Furthermore, there were significant intercept and latitudinal gradient effect to both the group counts and species occupancy data (Table 1). Specifically, average abundance decreased as the latitudinal gradient increased for all insect groups ($\beta_{latitude} = -0.47, -0.83, -0.56$ for bumblebees, hoverflies and solitary bees respectively; Table 1); occupancy probability increased as latitudinal gradient increased for bumblebees and hoverflies ($\beta_{latitude} = 0.08, 0.23$ for bumblebees and hoverflies respectively; Table 1), but the opposite for the solitarybees ($\beta_{latitude} = -0.82$; Table 1).

In addition, there was significant species effect and visit effect on species occupancy probability (since model M_3 was chosen for all the insect groups as the best model; Table 1), but insignificant visit effects on bumblebees and solitary bees abundance (since model M_1 was the best model; Table 1).

Combining both datasets

From the results and discussions above, we model the group count with a negative binomial GLMM with latitudinal gradient as a covariate and visit random effect. The species occupancy probabilities were modeled with a binomial GLMM with latitudinal gradient as a covariate and species and visit random effect. We allow both models to share latitudinal gradient effect as well as species and visit random effects. The details are decribed under section … in the main paper.

References

Bates, Douglas, Martin Maechler, Ben Bolker, Steven Walker, Rune Haubo Bojesen Christensen, Henrik Singmann, Bin Dai, Fabian Scheipl, and Gabor Grothendieck. 2009. "Package 'Lme4'." *URL Http://Lme4. R-Forge. R-Project. Org.*

Breeze, Tom D, Alison P Bailey, Kelvin G Balcombe, Tom Brereton, Richard Comont, Mike Edwards, Michael P Garratt, et al. 2021. "Pollinator Monitoring More Than Pays for Itself." *Journal of Applied Ecology* 58 (1): 44–57.

Fahrmeir, Ludwig, Thomas Kneib, Stefan Lang, and Brian D Marx. 2022. "Regression Models." In *Regression: Models, Methods and Applications*, 23–84. Springer.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s.* Fourth. New York: Springer. https://www.stats.ox.ac.uk/pub/MASS4/.

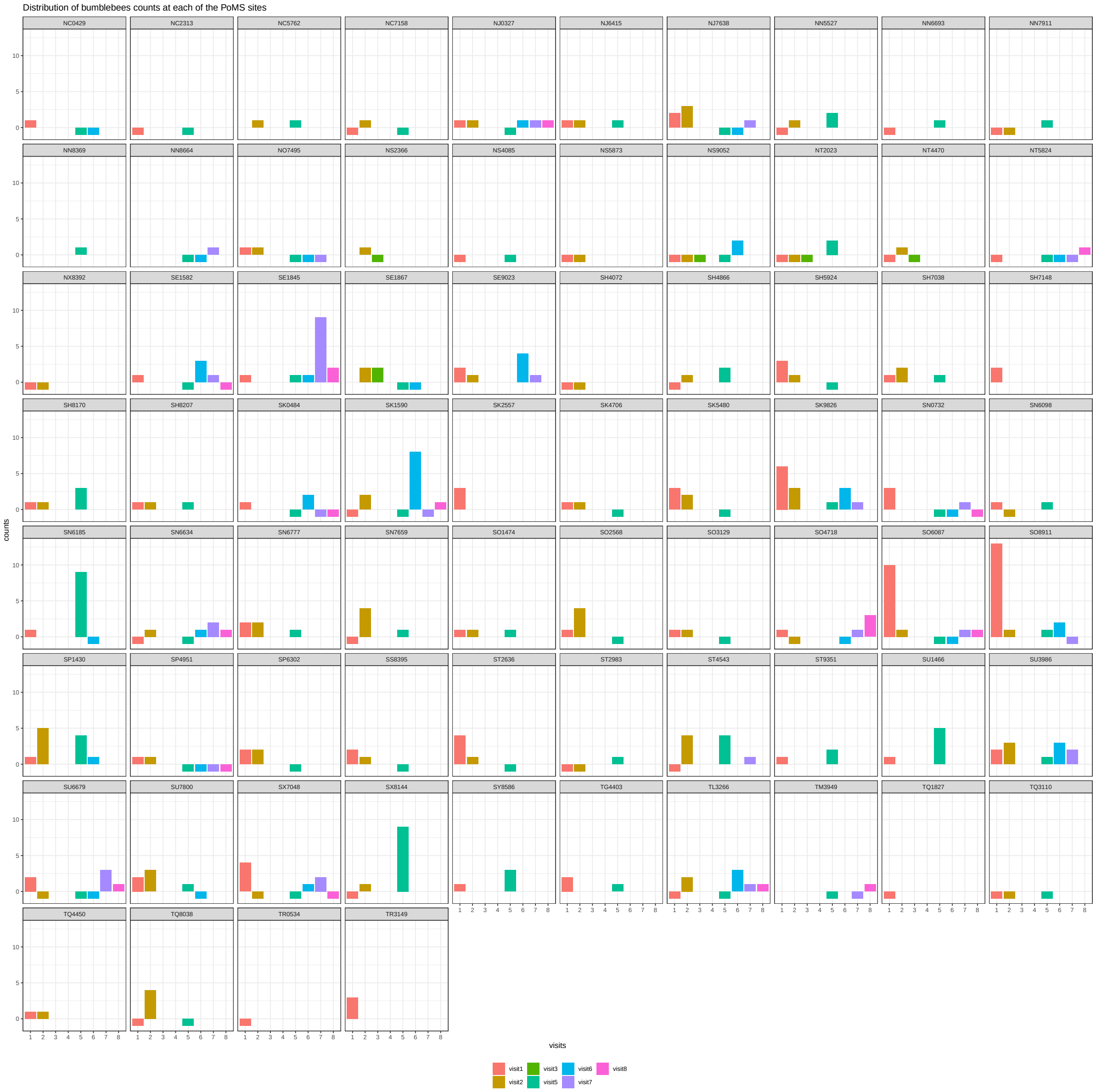

Figure S1-1: Distribution of bumblebess counts for the 74 PoMS sites. The counts are facetted by the PoMS site name and colored by the visit number the observations was made. The columns of visits without no bars represent the visit without no group count obervation ('NA').

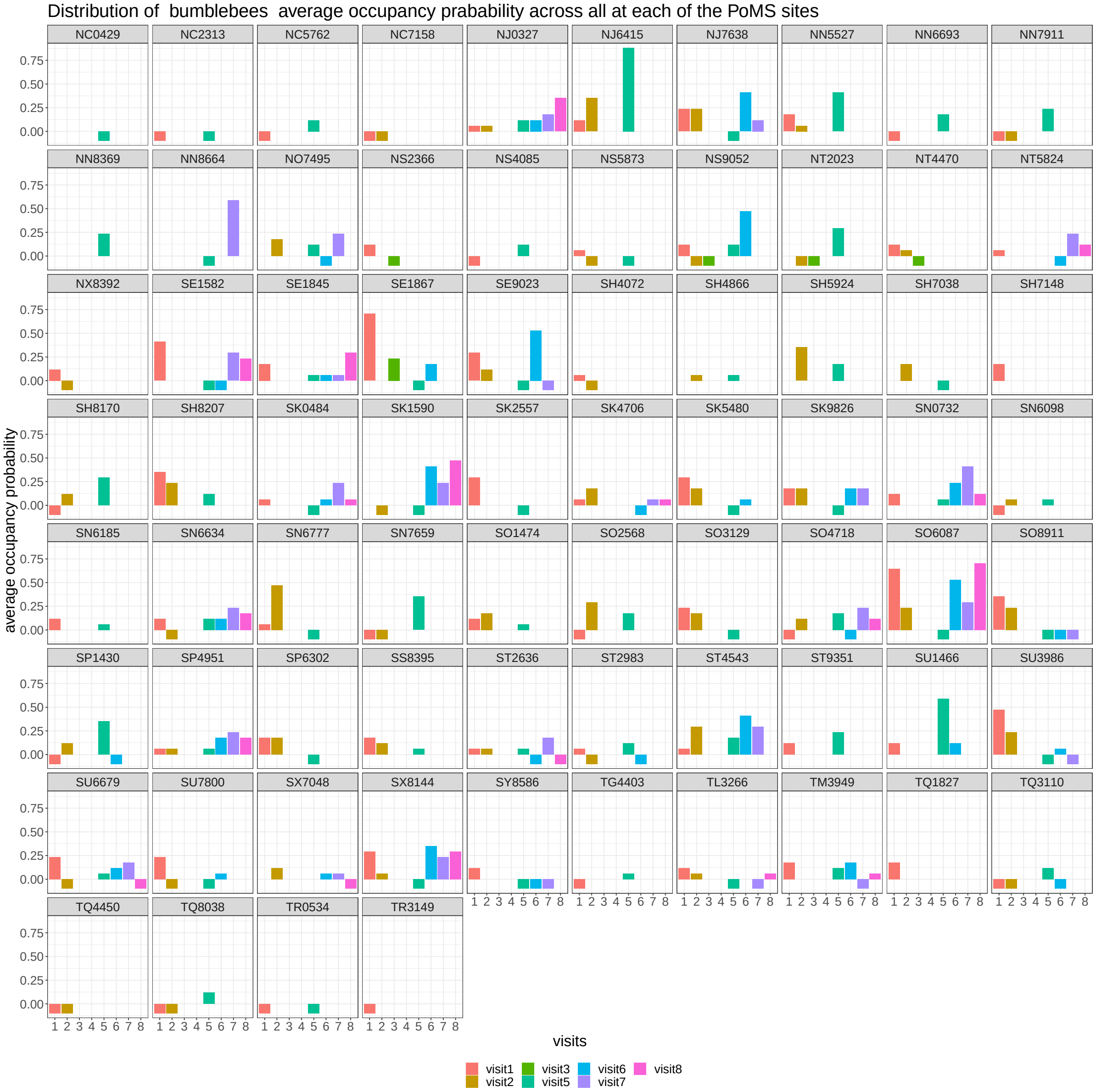

Figure S1-2: Distribution of bumblebees occupancy probability for the 74 PoMS sites. The probabilities for each site are estimated as the proportion of sites with at most one pantrap being occupied by the species. The occupancy probabilities are facetted by the PoMS site name and colored by the visit number the observations was made. The columns of visits without no bars represent the visit without no group count obervation (‘NA’).

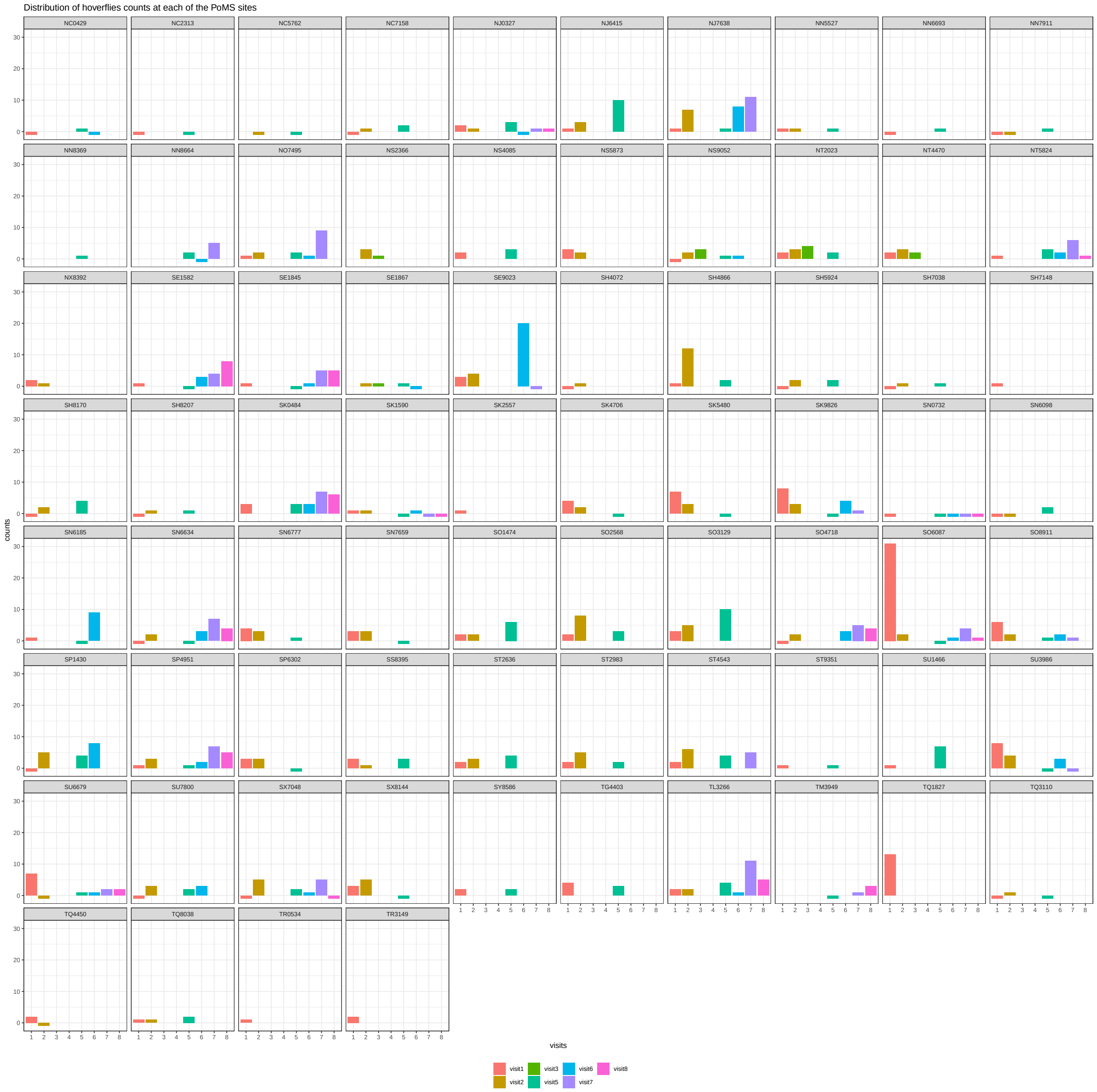

Figure S1-3: Distribution of hoverflies counts for the 74 PoMS sites. The counts are facetted by the PoMS site name and colored by the visit number the observations was made. The columns of visits without no bars represent the visit without no group count obervation ('NA').

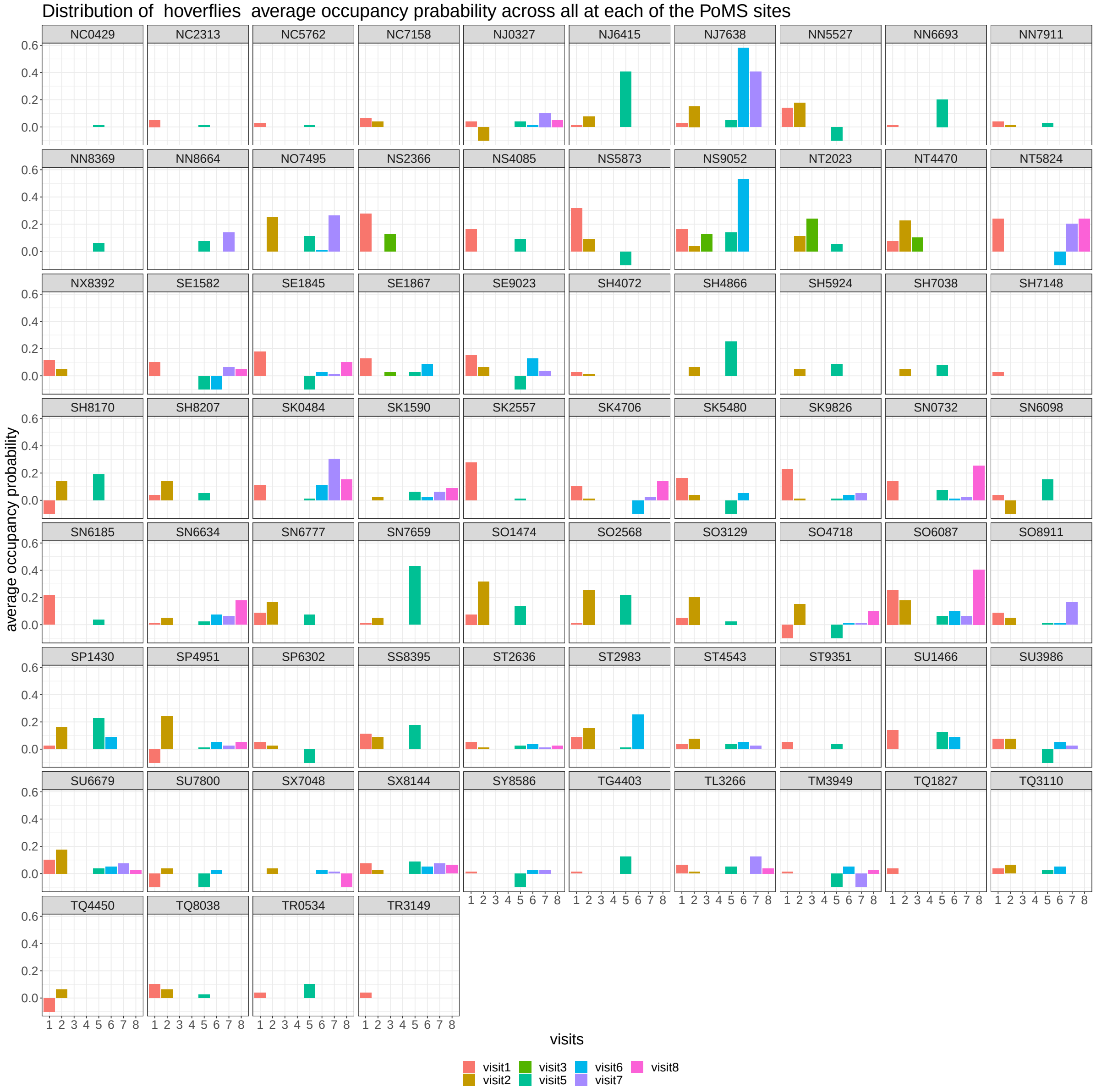

Figure S1-4: Distribution of hoverflies occupancy probability for the 74 PoMS sites. The probabilities for each site are estimated as the proportion of sites with at most one pantrap being occupied by the species. The occupancy probabilities are facetted by the PoMS site name and colored by the visit number the observations was made. The columns of visits without no bars represent the visit without no group count obervation ('NA').

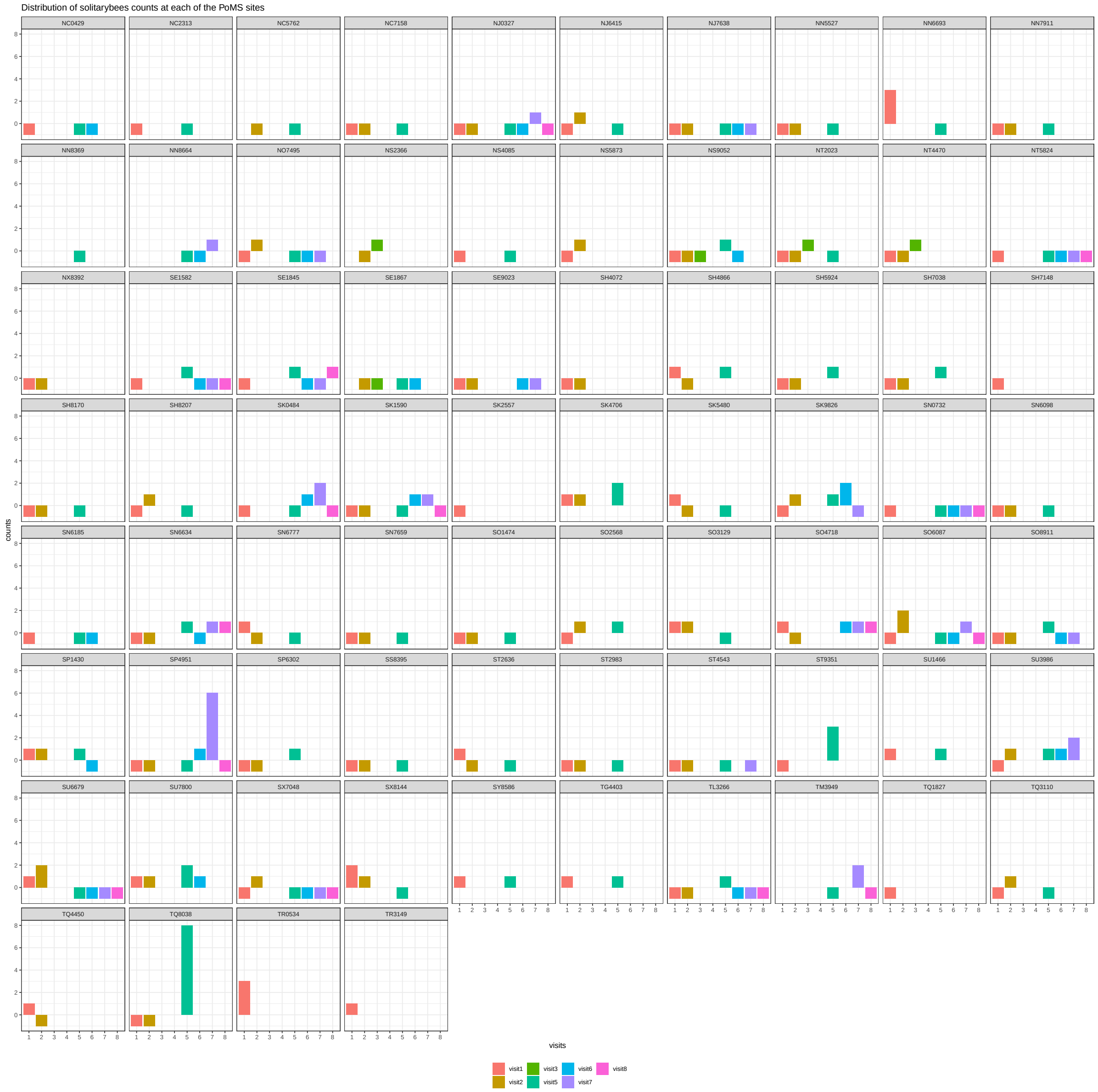

Figure S1-5: Distribution of solitarybees counts for the 74 PoMS sites. The counts are facetted by the PoMS site name and colored by the visit number the observations was made. The columns of visits without no bars represent the visit without no group count obervation ('NA').

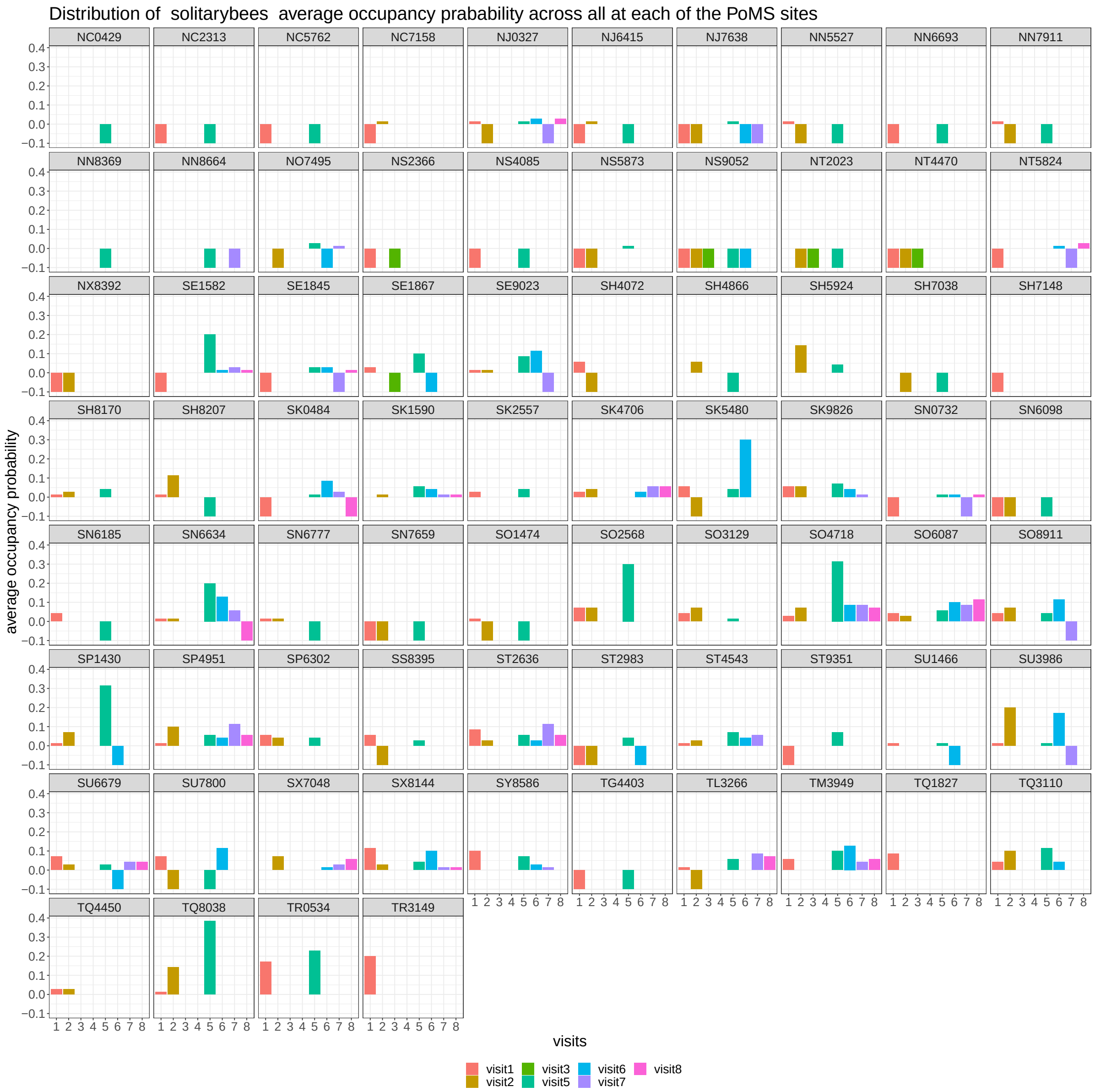

Figure S1-6: Distribution of solitarybees occupancy probability for the 74 PoMS sites. The probabilities for each site are estimated as the proportion of sites with at most one pantrap being occupied by the species. The occupancy probabilities are facetted by the PoMS site name and colored by the visit number the observations was made. The columns of visits without no bars represent the visit without no group count obervation ('NA').