1 INTEGRATED DISTRIBUTION MODELS

Three sources of data have been obtained:

- Pan traps will collect occupancy data at species level.
- transects will include some species level information but ID will be at most taxonomic levels e.g. hoverfly.
- flower-visit observations will be entirely at broad taxonomical level.

1.1 Appraoch

We assert that the total hoverfly count does tell us something about the diversity of hoverflies on the site. It would be possible to write an integrated model which the group-level counts (from transects and sunflower observations) and species-specific occupancy patterns (at species level) are separate realisations of some site-level state variable representing alpha diversity. The model could be specified in a way that species-level information could be used where available.

The advantage of this approach is that:

- We use the available data. It should therefore be sensitive to change.
- We are able to also estimate the detection probability.

The disadvantage of this approach is that:

- The model has not yet been attempted and can pose technical and inferential challenges.
- In fact we have face a lot of them. We have had to try a lot of options and tweaked the models in various ways. Still needs improvements.

1.2 Methodology

1.2.1 Model Assumptions

- There are i number of locations under the study, where $i = 1,2,\ldots,n.sites$.
- There are j number of species at each site, $j = 1,2,\ldots,n.sites$.
- There are k number of visit to each site, $k = 1,2,\ldots,n.visit$.
- There are p number of visit to each site, $p = 1,2,\ldots,n.cov$.
- There is a constant detection probability p for all the observations.
- Each species at each site has its own occupancy probability (i.e. ψ_{ij}, for $i = 1,2,\ldots,n.sites; i = 1,2,\ldots,n.sites$).
- Data collected at broader taxonomic levels are counts. It is worth stating that the model will also work for occurrence probabilities.
- Each site has replicated trials (n.replicates) for the collection of data.
- Assume that all the species that make up a particular group constitute the genus information. No provision is therefore made for the species that are not captured in the genus data. This can possibly be included by adding an error term to the observations Y.

Random variable	Description	Dimension
$\mathbf{X}$	Data collected at species level	$n.sites \times n.species \times n.visit$
$\mathbf{C}$	Covariates matrix	$n.sites \times n.cov$
ψ	Occupancy probability	$n.sites \times n.species$
λ	Mean Count of species	$n.sites \times n.species$
$\mathbf{Y}$	Taxon group data	$n.sites \times n.visits$

Table 1. Dimension and description of the random variables used in the model.

1.2.2 Dimensions of Variables

1.2.3 Model

Let

X be the occupancy data collected at the species level.

Y be the count data collected at the genus level.

Occupancy Model

$$\begin{aligned} X_{ijk} &\sim Binomial(n.replicates, z_{ij}p) \\ z_{ij} &\sim Bernoulli(\psi_{ij}) \end{aligned} \tag{1}$$

Abundance Model

$$Y_{ik} \sim Poisson(\sum_{j}^{n.species} \lambda_{ij}) \tag{2}$$

Link between abundance and occurrence Model

$$log(\lambda_{ij}) = cloglog(\psi_{ij}); \tag{3}$$

where $i = 1,2,\ldots,n.sites$, $j = 1,2,\ldots,n.species$ and $k = 1,2,\ldots,n.visit$.

1.2.4 Display of Model

The data is displayed in figure 1.

1.2.5 Treating of missing data

The missing data are treated in the model by assigning "NA" to them. The model treats them as observations that were not actually obtained, rather than 0 which indicates that there was no actual observation. Thus, the true state may be present but we could not obtain information during the data collection.

1.2.6 Diversity Measure

The number of species per taxonomic group is greater than 10. We considered evaluating the model in terms of some site-level metric of alpha diversity, rather than λ, ie. the Shannon index (properly referred to as Shannon's entropy) was used. For each of the study site, the Shannon index was calculated as:

$$H^1 = -\sum p_i log(p_i); \quad where \quad p_i = \frac{\lambda_{ij}}{\sum_j \lambda_{ij}} \tag{4}$$

where $i = 1,2,\ldots,n.sites$ and $j = 1,2,\ldots,n.species$. Note that $\sum_j \lambda_{ij}$ is the total number of species in the particular taxon groups in site i.

1.2.7 Coding

The code has been written in R. Provided is an except of the code used to run the MCMC in nimble. The code for the nimble has been sent to the GitHub repository.

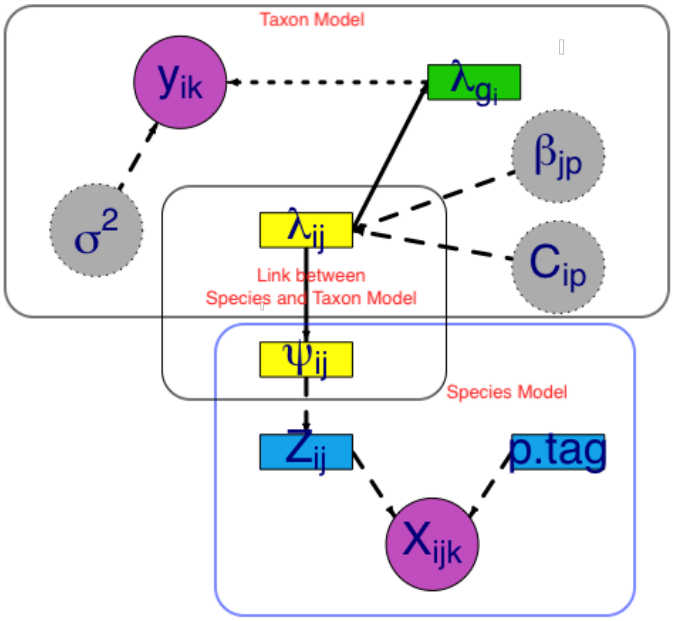

Figure 1. Display of the model with both covariate and no covariate effects. The dotted edges shows the stochastic relationship between the nodes and the solid edges shows the stochastic nodes. The dotted circles shows the nodes that have values chosen for the simulation. The nodes shaded orchid are the data that the model receives, and the nodes shaded yellow represents the link between the species and taxon level.

1.2.8 Application to Real data.

The model was applied to the UK pollinator monitoring schemes. The data contains species-level data from 75 locations (1 km grid cells) with pan traps. The data records number of individuals, but due to their low numbers, we treated them as presence-absence since they reflect local flower density rather than the true abundance count. The data also contains group-level counts from across the UK (with the 1km grid cells). The groups of the data are bumblebees, solitary bees, hoverflies and other insects.

Both dataset have replication within grid cells. There are five pan traps stations per grid cell, and each grid cell was surveyed on 1-4 occasions each year. At each survey visit, the surveyor is expected to do two FIT counts. In the case of the data analysis, I have taken out honeybees which has only one specie and the other flies since they do not have any species counterpart.

Taxon name	No. of species in 2017	No. of species in 2018
Bumblebees	13	13
Hoverflies	64	65
Solitary bees	47	62

Table 2. Number of species recorded for the taxon groups in 2017 and 2018.

From the figure below, we realised that quite a number of the counts were zero. Moreover, we observed that the variance of the data was high. This seems to fit into our model since it accounts for overdispersion in our model by the parameter σ^2.

1.2.9 Challenges faced and methods tried

A lot of challenges has been encountered working with this project. I have enumerated quite a few of them and some of the ways I attempted to solve them.

Data Formatting

It was a real issue trying to sort out the data in the format that I had specified for the model. It took quite sometime to understand the data and them format them into the required format needed for the analysis. Moreover, I failed to do exploratory analysis of the data before fitting the model. This may have been because I tried writing the model before I received the data. Upon receiving the data and later doing the exploratory analysis, some changes and alterations were made to the model.

Overdispersion

It was noted after the exploration analysis that there was excess zeroes, or possibly an overdispersion (if there were low counts at the loations). In an attempt to deal with this, the lognormal poisson distribution was used for the genus count data.

That is:

$$\begin{aligned} y_i &\sim Poisson(\lambda) \\ log(\lambda) &= \mathbf{C}\beta + \varepsilon \\ \varepsilon &\sim N(0,\sigma^2) \quad \text{where } \sigma^2 > 0. \end{aligned} \tag{5}$$

In the future, I hope to make much enquiries about the data.

Convergence of the model

The biggest issue has been the convergence of the model. Varied approaches and parameter values have been chosen, but there seems to be some challenges with convergence. Hopefully, it will not be a bug in the code. The initial values are monitored after the mcmc was run and the trace plot, posterior distribution and running means using the *mcmcplot* function in *R*.

Some of the different alterations done to the model includes the following:

- Assume that the covariate effect is at the species level, i.e.

$$\begin{aligned} cloglog(\psi) = \mathbf{C}\beta; \quad \beta \sim N(0,1) \\ \lambda = -log(1-\psi) \end{aligned} \tag{6}$$

  where β is a $n.species \times n.cov$ matrix.

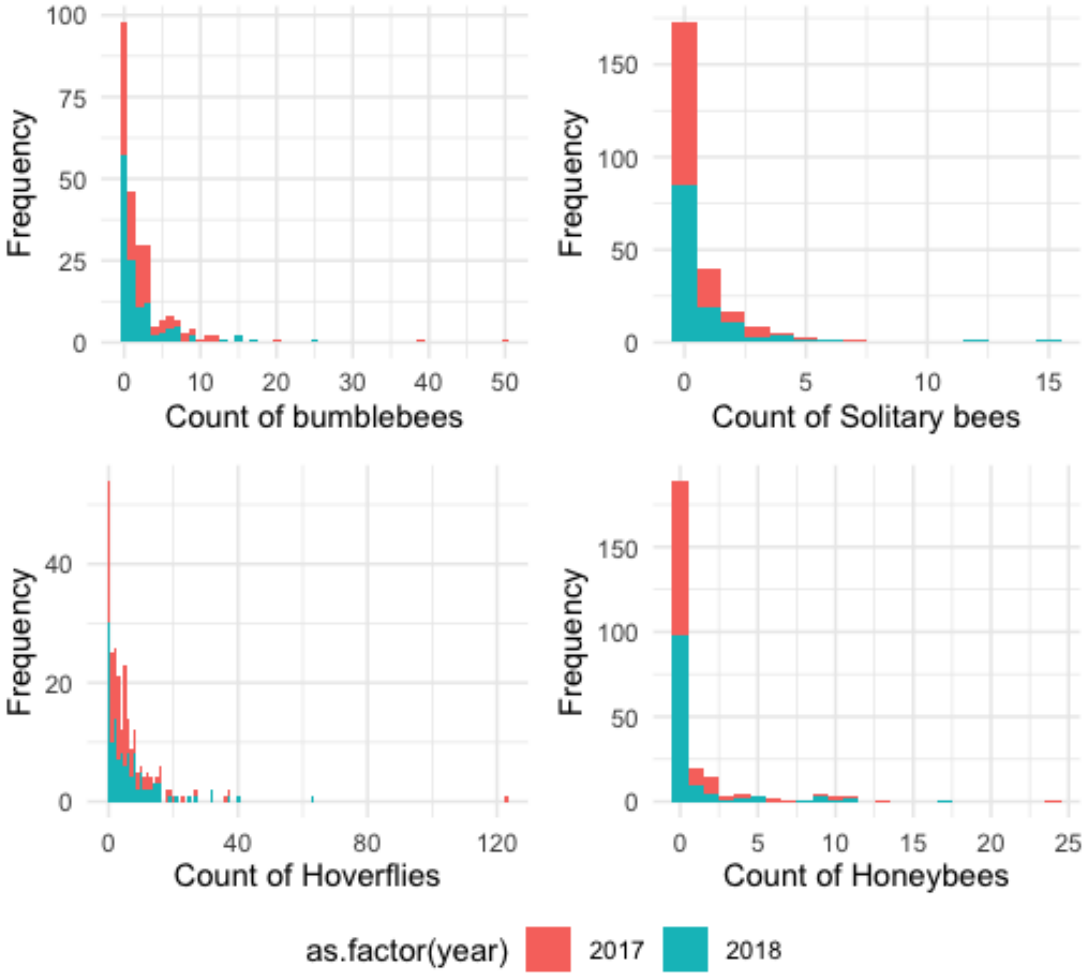

Figure 2. Distribution of the taxon groups.

If we assume an intercept model, then:

$$\begin{aligned} cloglog(\psi) &= \beta; \quad \beta \sim N(0,1) \\ \lambda &= -log(1-\psi) \end{aligned} \tag{7}$$

where β is a $n.sites \times n.species$ matrix.

- Another alternative used for the intercept model was:

$$\begin{aligned} \psi &\sim U(0,1) \\ \lambda &= -log(1-\psi) \end{aligned} \tag{8}$$

where β is a $n.sites \times n.species$ matrix. The convergence in this case was not very bad, but there seemed to be many parameters the model was estimating. In this model, we are estimating $n.sites * n.species + p.tag + \sigma^2$ number of parameters, which gets very large and time consuming when the number of species and sites increases.

- The model as specified in Equation (5) also had challenges with its convergence.
- Another alternative we have explored is to reduce the number of parameters that are estimated by the intercept only model. To do this, we re-write the problem as follows:

$$\begin{aligned} log(\lambda_{ij}) &= \alpha + \beta_i + \gamma_j \\ cloglog(\psi_{ij}) &= log(\lambda_{ij}) \\ \beta_i &\sim N(0, \sigma^2_{site}) \\ \gamma_j &\sim N(0, \sigma^2_{species}) \\ \alpha &\sim N(0, \sigma^2_{\alpha}) \end{aligned} \tag{9}$$

THE VALUES CHOSEN FOR SIMULATION ARE: $\sigma^2_{site} = 1, \sigma^2_{species} = 2; \sigma^2_{\alpha} = 0.4$.

THE PRIOR DISTRIBUTION FOR THIS MODEL WAS:

$$\begin{aligned} \sigma^2_{site} &\sim Gamma(1,1) \\ \sigma^2_{species} &\sim Gamma(1,1) \\ \sigma^2_{\alpha} &\sim Gamma(1,1) \end{aligned} \tag{10}$$

There seems to be problems with the convergence.

THIS IS THE MODEL IN THE CODE.

1.2.10 Future work

In the future, the following will be done:

- We reparametrise the equation (9) as follows:

$$\begin{aligned} \hat{\alpha} &= \alpha + \frac{\sum_i \beta_i}{n} + \frac{\sum_j \gamma_j}{n} \\ \hat{\beta}_i &= \beta_j - \frac{\sum_i \beta_i}{n} \\ \hat{\gamma}_j &= \gamma_i - \frac{\sum_j \gamma_j}{n} \end{aligned} \tag{11}$$

The estimates $\hat{\alpha}$, $\hat{gamma}$, $\hat{\beta}$ are estimated after the MCMC has been run. Hopefully, this should converge.

- I would like to test whether removing the sites without any information has any effects on the diversity measure or not.
- Run the converged model on the data.
- Hopefully we can get a paper out of this.

REFERENCES