# Feature allocation models with imperfect detection for ecological applications

Federica Stolf

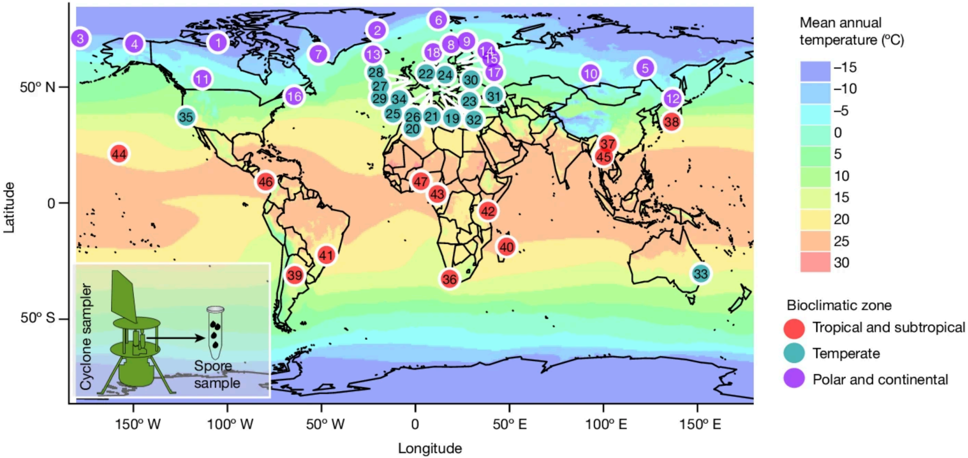
federica.stolf@duke.edu

Joint work with Tommaso Rigon and David Dunson

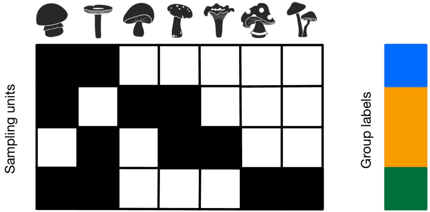
SISBayes – September 4, 2025

# Motivating data: Global Spore Sampling Project data

GSSP data from Abrego et al. (2024)



# Motivating application: species co-occurrence data (incidence data)




Species co-occurrence data:

- $Y^{(q)}$ is a $n_q \times p$ binary matrix with $q = 1, \ldots, Q$;
- $y_{ij}^{(q)} = 1$ if species $j$ was found in sample $i$ of group $q$ and 0 otherwise.

Challenges

- The number of species is huge ($p \approx 1,000 - 200,000$).
- Many rare species.
- **Detectability** issue: detectability of species occurrences is rarely perfect. Does the absence of a species mean that the species is not present or that we are unable to observe it?

# Feature allocation models

- **Indian Buffet Processes** (IBPs) (Teh and Gorur, 2009; Griffiths and Ghahramani, 2011; Broderick et al., 2013) are popular Bayesian nonparametric models designed for binary latent feature matrices with a potentially infinite number of columns.
- IBPs sequentially sample the latent feature matrix, allowing the discovery of new binary features as more data becomes available.
- In biodiversity studies, **features = observed species** $\longrightarrow$ we can include an ever-growing number of species.

**Limitations** of current IBPs for biodiversity data:

1. Exchangeability assumption for samples
2. They do not account for imperfect detection

# Beta bernoulli models

- We focus on a specific instance of the IBP with a finite but unknown number of species, the **beta Bernoulli (BB)** models (Ghilotti et al. 2024).
- $N$ is the unknown total number of species.
- The BB model for multiple groups with parameters $(N, \alpha_q, \theta_q)$ such that $N \in \mathbb{N}$, $\alpha_q < 0$ and $\theta_q > -\alpha_q$ assume

$$y_{ij}^{(q)} \mid \pi_{jq} \overset{\text{ind}}{\sim} \mathrm{Bernoulli}(\pi_{jq}), \qquad \pi_{jq} \overset{\text{ind}}{\sim} \mathrm{Beta}(-\alpha_q, \alpha_q + \theta_q),$$

for $j = 1, \ldots, N$, $i = 1, \ldots, n_q$ and $q = 1, \ldots, Q$.
- The observed number of species for each group $K_{nq}$, is such that $K_{nq} \leq N$.

# BB models with imperfect detection (i)

The BB model with imperfect detection with parameters $(N, \alpha_q, \theta_q, \sigma_q)$ such that $N \in \mathbb{N}$, $\alpha_q < 0$, $\theta_q > -\alpha_q$ and $\sigma_q \in (0,1)$ assume for $j = 1, \dots, N$ and $q = 1, \dots, Q$

$$
\begin{aligned}
y_{ij}^{(q)} \mid \eta_{jq}, \gamma_{jq} &\overset{\text{ind}}{\sim} \mathrm{Bernoulli}(\eta_{jq}\gamma_{jq}), \\
\eta_{jq} &\sim \mathrm{Beta}\{-\alpha_q, \sigma_q(\alpha_q + \theta_q)\}, && \leftarrow \textbf{Occupancy} \\
\gamma_{jq} &\sim \mathrm{Beta}\{-\alpha_q + \sigma_q(\alpha_q + \theta_q), (1 - \sigma_q)(\alpha_q + \theta_q)\}. && \leftarrow \textbf{Detectability}
\end{aligned}
$$

Detectability parameter $\sigma_q$

- $\sigma_q \to 0$ means that no species can be detected.
- $\sigma_q \to 1$ implies perfect detectability, indeed

$$
E(\eta_{jq}) = -\alpha_q/\theta_q, \qquad E(\gamma_{jq}) = 1.
$$

# BB models with imperfect detection (ii)

- In this framework $N$ is the number of total species potentially detected $\longrightarrow$ global species-richness.
- We assume $N \sim \mathrm{Poisson}(\lambda)$.
- Each group has a different number of species observed, $K_{n1}, \ldots, K_{nQ}$ (with $K_{nq} \sim \mathrm{Poisson}$).

**Theorem 1**

*Marginally the BB model with imperfect detection is equivalent to the BB model, i.e.*

$$\pi_{jq} = \eta_{jq}\gamma_{jq}, \qquad \pi_{jq} \sim \mathrm{Beta}(-\alpha_q, \alpha_q + \theta_q).$$

# Prior selection and posterior computations

Prior choice for the BB model with imperfect detection with parameters $(N, \alpha_q, \theta_q, \sigma_q)$

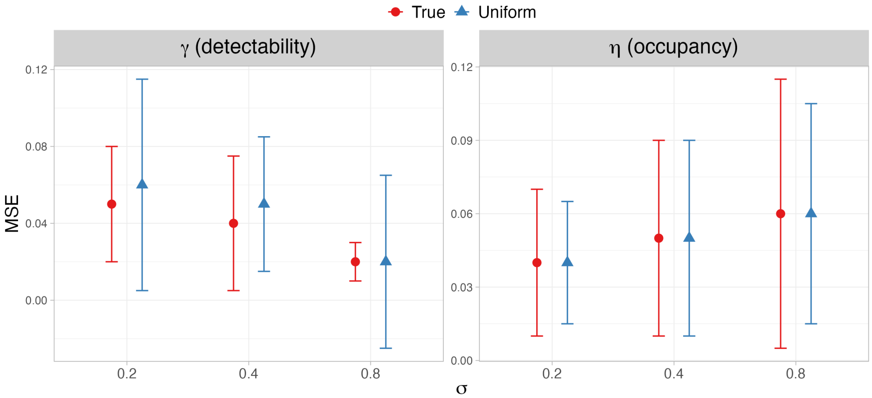
- We take hierarchical priors on $\alpha_q$ and $\theta_q$ to borrow information across sites.
- The detectability parameter $\sigma_q$ is not identifiable. We set $\sigma_q \sim \mathrm{Unif}(a_\sigma, b_\sigma)$ with $a_\sigma \to 0$ and $b_\sigma \to 1$.

Posterior computation

1. Obtain the posteriors of $(N, \alpha_q, \theta_q)$ from the marginal model.
2. Obtain the posteriors of the detectability and occupancy probabilities $(\eta_{jq}, \gamma_{jq})$ via a collapsed Gibbs sampler with data augmentation.

# Simulations for detectability parameters

MSE for $\gamma$ and $\eta$ vectors with $\sigma$ sampled from a uniform distribution or fixed to its true value.



# How to measure biodiversity?

The most common measure for biodiversity is **alpha-diversity**: species diversity of a local community or habitat.

- Typically measured as species richness, i.e. the total number of species in a community.
- In the BB models the species richness is $N$, that is unknown and it can be estimated employing a prior distribution for $N$.

We will focus on **beta-diversity**: heterogeneity of species across different sampling regions.

- There are many ways to quantify it and little agreement about which is best.
- Often estimated using dissimilarity indexes (e.g., Jaccard, Sørensen, Bray–Curtis).

# Beta diversity

- We propose a definition of $\beta$-diversity under a coherent probabilistic framework.
- We define the $\beta$-diversity between groups $p$ and $q$ as

$$\beta_{pq} = 1 - \frac{\sum_{j=1}^{N} \eta_{jq}\eta_{jp} - \sum_{j=1}^{N} \eta_{jq} \sum_{j=1}^{N} \eta_{jp}}{\left\{\left[\sum_{j=1}^{N} \eta_{jq}^2 - (\sum_{j=1}^{N} \eta_{jq})^2\right]\left[\sum_{j=1}^{N} \eta_{jp}^2 - (\sum_{j=1}^{N} \eta_{jp})^2\right]\right\}^{1/2}}.$$

- It is a function of the posteriors of $\eta$, so we can do uncertainty quantification.
- It is bounded between 0 and 1 and it is a correlation among random vectors.

# Theoretical results related to beta diversity

Consider the BB model with imperfect detection and assume $N \sim \mathrm{Poisson}(\lambda)$. The total number of shared species among two groups $p$ and $q$ $(p \neq q)$ is

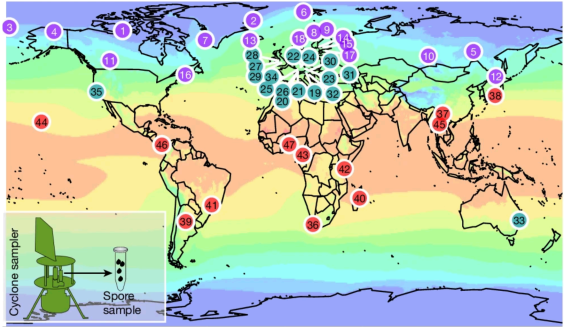
$$C_{pq} \sim Poisson(\lambda \zeta_{pq}),$$

$$\zeta_{pq} = 1 - \frac{(\sigma_q\{\alpha_q + \theta_q\})_{n_q}}{(\sigma_q\{\alpha_q + \theta_q\} - \alpha_q)_{n_q}} - \frac{(\sigma_p\{\alpha_p + \theta_p\})_{n_p}}{(\sigma_p\{\alpha_p + \theta_p\} - \alpha_p)_{n_p}} + \frac{(\sigma_q\{\alpha_q + \theta_q\})_{n_q}(\sigma_p\{\alpha_p + \theta_p\})_{n_p}}{(\sigma_q\{\alpha_q + \theta_q\} - \alpha_q)_{n_q}(\sigma_p\{\alpha_p + \theta_p\} - \alpha_p)_{n_p}}.$$

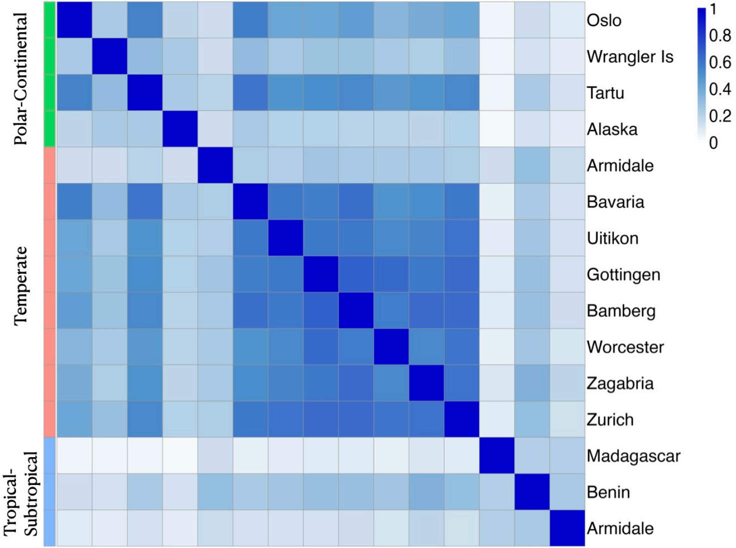
We can obtain the analogous result for the BB marginal model.

# Preliminary results for GSSP data

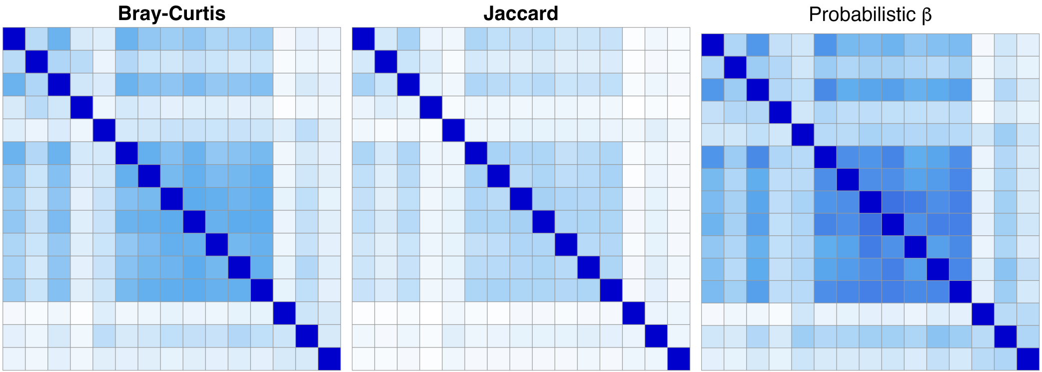
- We analyze the fungi data from Abrego et al. (2024).
- We select the sites that have at least 50 samples, for a total of $Q = 15$ groups. The groups considered cover all three climatic zones (polar-continental, temperate and tropical-subtropical) and all the continents.
- The number of species identified is $p = 17170$.



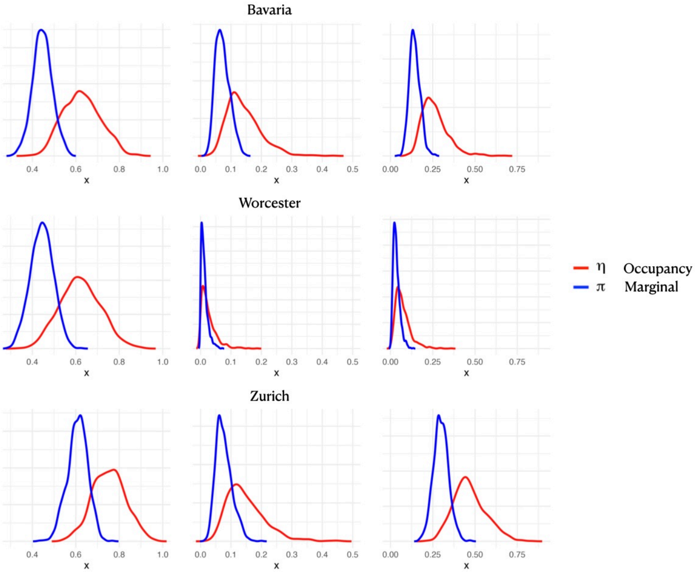
# GSSP data - beta similarity



# Beta similarity - comparison



# Posterior marginal and occupancy probabilities



# Conclusion and future directions

- We introduce a feature allocation model that accounts for imperfect detection in species co-occurrence data and accommodates partially exchangeable data.
- The proposed framework offers an ecologically meaningful interpretation, separating occupancy and detectability components.
- We provide both theoretical insights for assessing biodiversity and applied results to demonstrate the method's utility.
- Future directions include studying the beta diversity index through correlation structures among random measures, related to the framework of Franzolini et al. (2025) for species sampling models.

# References

Abrego, N., B. Furneaux, B. Hardwick, P. Somervuo, I. Palorinne, C. A. Aguilar-Trigueros, N. R. Andrew, U. V. Babiy, T. Bao, G. Bazzano, et al. (2024). Airborne dna reveals predictable spatial and seasonal dynamics of fungi. *Nature* 631 (8022), 835–842.

Broderick, T., J. Pitman, and M. Jordan (2013). Feature allocations, probability functions, and paintboxes. *Bayesian Analysis* 8 (4), 801–836.

Franzolini, B., Lijoi, A., Prünster, I., and Rebaudo, G. (2025). Multivariate species sampling models. arXiv:2503.24004

Ghilotti, L., F. Camerlenghi, and T. Rigon (2024). Bayesian analysis of product feature allocation models. arXiv:2408.15806

Griffiths, T. and Z. Ghahramani (2011). The Indian Buffet Process: An Introduction and Review. *Journal of Machine Learning Research* 12 (32), 1185–1224.

Teh, Y. and D. Gorur (2009). Indian buffet processes with power-law behavior. In *Advances in Neural Information Processing Systems*, Vol 22

Thank you for your attention!