Clustering metabarcoding data: a model-based approach

Luisa Ferrari[1], Maria Franco-Villoria[1], Garritt Page[2], Massimo Ventrucci[3], Alex Laini [4]

[1] University of Modena & Reggio Emilia, Italy
[2] Brigham Young University, USA
[3] University of Bologna, Italy
[4] University of Turin, Italy

E-mail for correspondence: `luisaferrari@unimore.it`

Abstract: Metabarcoding is a highly efficient molecular technique that provides large species occurrence datasets. However, it presents a major limitation as only presence/absence of a species, not abundance, is detectable. Therefore, metabarcoding data requires the use of statistical tools designed for multivariate binary data. We aim to develop a model-based clustering strategy for metabarcoding data. Following a comparison of the methods from the literature, we propose to investigate an extension towards the inclusion of environmental covariates that often accompany occurrence data. In summary, this project seeks to maximize the utility of metabarcoding data with a context-appropriate clustering technique.

Keywords: Clustering; Mixture model; Metabarcoding.

1 Metabarcoding data

Metabarcoding is a molecular technique that allows for the simultaneous genetic identification of multiple organisms within biological samples (Taberlet et al. 2018). This novel methodology has revolutionized the field of community ecology, as it enables the rapid and simultaneous identification of a large number of species. The final product of the metabarcoding process consists of a series of DNA variations called *Operational Taxonomic Units* (OTUs), which are assigned to the most likely taxa on the basis of genetic similarity using open-access databases. However, metabarcoding can only offer information about occurrence on a presence-absence scale (i.e. binary), due to the specific nature of the molecular identification procedure.
Hence, metabarcoding results from multiple sites form a typical $N \times P$ *occurrence matrix* where each cell n, p reports the occurrence $Y_{np} \in \{0, 1\}$ for species n

This paper was published as a part of the proceedings of the 39th International Workshop on Statistical Modelling (IWSM), Limerick, Ireland, 13–18 July 2025. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

observed in the sample from site p. In most cases, the spatio-temporal coordinates of each sampling site are available $(\mathbf{S}_p, T_p)$, along with a set of environmental covariates $(\mathbf{X}_p)$ believed to affect occurrence.
In most applications, ecologists are primarily interested in building a distribution model for the observed data, able to consider the impact of environmental factors, as well as potential interactions between species. The binary nature of metabarcoding data does not prevent the applicability of many Joint Species Distribution Model (JSDM) frameworks presented in the literature. We shall focus in this project on a frequent secondary objective in community ecology analysis, namely, the identification of meaningful clusters among either sampling sites or species (Legendre & Legendre, 2012).

2 Clustering presence-absence data

Clustering is often used on ecological data for the identification of occurrence patterns and structures in complex datasets. First, obtaining sites' clusters is desirable for the delimitation of biogeographical regions, i.e. regions with similar species composition (e.g., Kreft & Jetz, 2010). This type of analysis is useful to explore how compositional differences reflect ecological and historical processes, such as climatic changes, which is fundamental for ecosystem-based management and conservation planning.
On the other hand, clustering species provides groups with similar geographical/temporal distributions, which offer insights into biotic interactions and the species-environment relationships (e.g. Pang et al. 2023). Although less common in the community ecology literature, clustering species could be particularly useful in the context of metabarcoding datasets, which are characterized by a large number of species. In this case, clustering could offer a way to simplify and summarize the community structure into a manageable number of species subsets with similar geographical distributions.
While algorithm-based approaches to clustering have traditionally been used in ecology, we believe it would be more convenient to employ a more sophisticated model-based approach commonly used in many other fields, which could incorporate probabilistic assumptions about the data and account for hierarchical structures and potential exogenous covariates.
Among the numerous alternatives proposed in the literature for the case of multivariate binary data, it is worth mentioning the popular Bernoulli Mixture Model (BMM), which assumes a mixture of Multivariate Bernoulli distributions on the data (Juan & Vidal, 2002). The BMM model has been thoroughly studied, and it is implemented in multiple R packages (e.g. `BayesBinMix` by Papastamoulis & Rattray, 2017).
A recently proposed alternative to the BMM is the Hamming mixture model (HMM) by Argiento et al. (2024), which introduced a novel multivariate distribution based on the Hamming distance for generic categorical data: although not directly developed for binary data, we argue it would be interesting to compare the impact between the two different probability distribution assumptions. Finally, a different approach has been proposed by Cagnone & Viroli (2012), who made use of a latent factor model where the factors are distributed as a finite mixture of multivariate Gaussians. This option moves the mixture assumption to a lower level of the model hierarchy. Additionally, it also allows the retrieval

of latent traits that may explain the clustering. While the examples mentioned are among the most notable, note they do not encompass the entirety of existing methods.

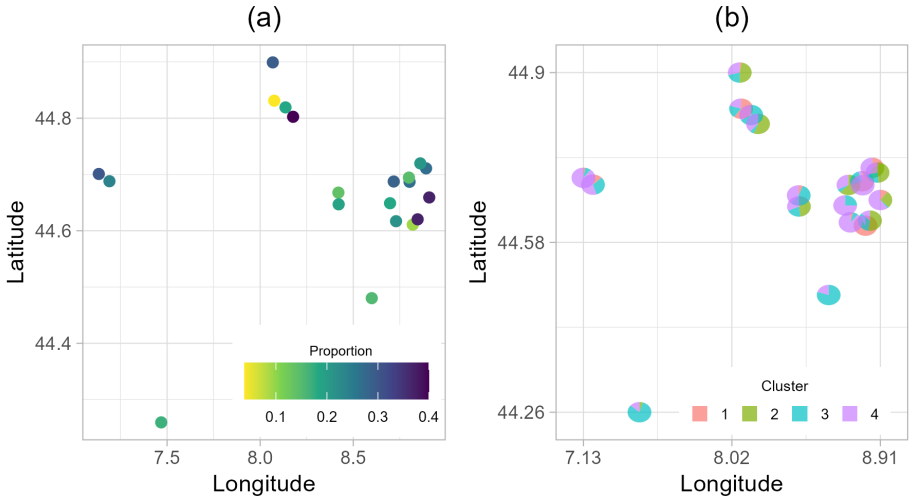

FIGURE 1. Example of a metabarcoding dataset. (a) Proportion of species present in each site; (b) pie charts about cluster membership proportions.

3 Preliminary results

We present some preliminary results on a metabarcoding dataset containing presence/absence of aquatic macroinvertebrates (OTUs) in 20 river sites in North-West Italy. Panel (b) of Figure 1 reports BMM cluster results using the `BayesBinMix` package: the plot shows the site composition in terms of species clusters. The results show a potential spatial correlation in site composition. However, the current BMM specification is unable to consider either available environmental covariates or geographical proximity. We propose to compare via simulation the top model-based binary data clustering approaches. We then aim to extend the best candidate by considering the inclusion of environmental covariates. This extension would provide ecologists with an *explanation* for clustering in terms of species-environment interactions, without requiring the estimation of the full joint species correlation matrix. Furthermore, we propose to explore a Bayesian implementation with a more refined prior choice on the number of clusters, such as the asymmetric Dirichlet prior from Page et al. (2023). Finally, we will use efficient MCMC algorithms to analyze large metabarcoding datasets.

Acknowledgments: Work partly funded by the European Union under the NextGeneration EU Programme within the Plan PNRR - Missione 4 “Istruzione e Ricerca” - Componente C2 Investimento 1.1 “Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN)” by the Italian Ministry of University and Research (MUR), Project title: “METAbarcoding for METAcommunities: towards a genetic approach to community ecology

(META2)", Project code: 2022PA 3BS2 (CUP E53D23007580006), MUR D.D. financing decree n. 1015 of 07/07/2023

References

Argiento, R., Filippi-Mazzola, E., & Paci, L. (2024). Model-Based Clustering of Categorical Data Based on the Hamming Distance. *Journal of the American Statistical Association*, 1–23.

Cagnone, S., & Viroli, C. (2012). A factor mixture analysis model for multivariate binary data. *Statistical Modelling*, 12(3), 257-277.

Juan, A., & Vidal, E. (2002). On the use of Bernoulli mixture models for text classification. *Pattern Recognition*, 35(12), 2705-2710.

Legendre, P., & Legendre, L. (2012). *Numerical ecology* (Vol. 24). Elsevier.

Kreft, H., & Jetz, W. (2010). A framework for delineating biogeographical regions based on species distributions. *Journal of Biogeography*, 37(11), 2029-2053.

Page, G. L., Ventrucci, M., & Franco-Villoria, M. (2023). Informed Bayesian Finite Mixture Models via Asymmetric Dirichlet Priors. *arXiv preprint arXiv:2308.00768*.

Pang, S. E., Slik, J. F., Zurell, D., & Webb, E. L. (2023). The clustering of spatially associated species unravels patterns in tropical tree species distributions. *Ecosphere*, 14(6), e4589.

Papastamoulis, P., & Rattray, M. (2017). BayesBinMix: an R package for model based clustering of multivariate binary data. *The R Journal*.

Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018). *Environmental DNA: For biodiversity research and monitoring*. Oxford University Press.