Bayesian Analysis (2026) **TBA**, Number TBA, pp. 1–28

Bayesian Discovery of Species in Multiple Areas

Alessandro Colombi*, Raffaele Argiento†, Federico Camerlenghi‡ and Lucia Paci§

Abstract. In ecology, the description of species composition and biodiversity calls for statistical methods that involve estimating features of interest in unobserved samples based on an observed one. In the last decade, the Bayesian nonparametrics literature has thoroughly investigated the case where data arise from a homogeneous population. In this work, we propose a novel framework to address heterogeneous populations, specifically dealing with scenarios where data arise from two areas. This setting significantly increases the mathematical complexity of the problem and, as a consequence, it has received limited attention in the literature. While early approaches leverage computational methods, we provide a distributional theory for the in-sample analysis of any observed sample and enable out-of-sample prediction for the number of unseen distinct and shared species in additional samples of arbitrary sizes. The latter also extends the frequentist estimators, which solely deal with one-step-ahead prediction. Furthermore, our results can be applied to address sample size determination in sampling problems aimed at detecting distinct and shared species. Our results are illustrated in a real-world dataset concerning a population of ants in the city of Trieste.

Keywords: abundance data, Bayesian nonparametrics, shared species, vector of finite Dirichlet processes.

MSC2020 subject classifications: Primary 62F15, 62M20, 62G05.

1 Introduction

In ecology, there is often a strong interest in describing species composition and biodiversity and, when the study involves more than a single area, in comparing them. Given that the concept of biodiversity lacks a single universally accepted mathematical definition (Colwell et al., 2009), many similarity indexes have been proposed in the literature. For a single assemblage of data, the most widely used indices are Shannon's and Simpson's diversity, while Jaccard's and Sørensen's indices are classic choices when comparisons across multiple areas are investigated. A detailed overview of these and many other alternatives can be found in Chao et al. (2006a). Regardless of the number of areas considered, such investigations generally follow a two-step process. The first step is to choose a sampling strategy to detect as many species as possible. Since both time and financial constraints often limit sampling efforts, it is crucial to develop statistical tools that, based on the current sample, can estimate how many additional

*Bocconi Institute for Data Science and Analytics, Bocconi University, Milan, Italy, alessandro.colombi@unibocconi.it

†Department of Economics, Università degli studi di Bergamo, Italy, raffaele.argiento@unibg.it

‡Department of Economics, Management and Statistics, University of Milano-Bicocca, Italy, federico.camerlenghi@unimib.it

§Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan, Italy, lucia.paci@unicatt.it

© 2026 International Society for Bayesian Analysis https://doi.org/10.1214/26-BA1617

observations are needed to discover new species. These predictions enable researchers to assess whether additional sampling efforts are feasible and justified in terms of time and cost. Due to these practical limitations, collected samples rarely capture the full diversity of the population, leaving some species undetected. Consequently, biodiversity indices estimated from sample data are often negatively biased. Therefore, the second step is usually to estimate, based on the observed data, the species richness, i.e., the total number of species, both observed and unobserved.

This work extends beyond ecology and is relevant to many other disciplines. Thus, the term "species" can be interpreted in a broader sense. For instance, it can refer to fields such as topic modelling (Efron and Thisted, 1976), typos in texts (Nayak, 1988), bugs in computer code (Chao and Yang, 1993), or genomics (Mao, 2004).

1.1 Estimators for a Single Area

The case in which the investigation focuses on a single area is a long-established and extensively discussed problem in statistics, dating back to Fisher et al. (1943). Here, the $n \geq 1$ observed individuals represent a random sample $(X_1, \ldots, X_n)$ drawn from an unknown discrete distribution P. In this setting, the celebrated Good-Turing estimator (Good, 1953) provides an estimate of the probability that the $(n+1)$th observation coincides with a species whose frequency in the original sample is exactly equal to f, with $f \geq 0$. In particular, for $f = 0$, this estimates the probability of observing a new, previously undetected species as the relative frequency of the species observed exactly once in the sample, commonly known as singletons. The Good-Toulmin estimator (Good and Toulmin, 1956) represents a m-steps ahead generalisation for the probability of discovering a new species. These estimators are nonparametric since they do not rely on any assumption for the unknown generating distribution P, making them flexible and easy to use both for deriving stopping rules in sampling strategies (Rasmussen and Starr, 1979) and for estimating species richness (Chao, 1984). See Orlitsky et al. (2016) for a review of generalisations and improvements of the Good-type estimators.

An alternative Bayesian nonparametric approach has been introduced by Lijoi et al. (2007). In this framework, the authors only require $(X_1, \ldots, X_n)$ to be an exchangeable sequence. By de Finetti's representation theorem, this hypothesis is equivalent to assuming that the unknown distribution P is a random probability measure governed by a prior distribution $\mathscr{P}$. Given P, the n observations are assumed to be independent and identically distributed according to P. In particular, Lijoi et al. (2007) studied priors $\mathscr{P}$ belonging to the broad class of Gibbs-type priors (De Blasi et al., 2015), which encompasses notable examples such as the Dirichlet process (Ferguson, 1973) and the two parameters Poisson-Dirichlet model (Pitman, 1995). In this setting, Favaro et al. (2012) presents a Bayesian generalisation of the Good-Turing and Good-Toulimin estimators. Based on an observed sample of size n and an additionally unobserved sample of size $m \geq 0$, they provide an estimator for the probability that the $(n+m+1)$th observation coincides with a species whose frequency, within the sample of size $n+m$, is exactly f. Thus, for $m = 0$ and $f = 0$, this is a Bayesian analogue of the Good-Turing and Good-Toulmin estimators, respectively. As for their frequentist counterparts, these estimators are crucial in sampling problems.

Regarding the estimation of species richness, Favaro et al. (2009) approached the problem as an extrapolation of the unseen species problem. Specifically, for each $m \geq 0$, they derive an estimator for the number of new distinct species that would be observed in an additional unobserved sample $(X_{n+1}, \ldots, X_{n+m})$. As m varies, this estimate forms the so-called extrapolation curve (Gotelli and Colwell, 2001), and species richness can be considered the limit for $m \to \infty$. However, whenever $\mathscr{P}$ is chosen in the family of Gibbs-type priors, this limit may either be infinite or finite, dividing the family into two subgroups. The choice of $\mathscr{P}$ depends on which hypothesis is more realistic for the specific application. Notable examples of the first group, characterised by unlimited growth, include the Dirichlet process and the two-parameter Poisson–Dirichlet model. In contrast, the second group, namely the Gibbs-type priors with negative parameters, encompasses Gnedin's model (Gnedin, 2010) and the finite Dirichlet process (Argiento and De Iorio, 2022). Moving beyond the Gibbs-type priors framework, Zito et al. (2023) proposed a tunable Bayesian nonparametric approach between these two regimes. We refer to Balocchi et al. (2024b) for an exhaustive review on this topic.

1.2 Estimators for Multiple Areas

The study of estimators for multiple areas is much less extensive than for a single population. Even with just $d = 2$ different areas, the mathematical complexity increases significantly because we need to estimate the number of shared species, i.e., those occurring in both areas. Indeed, this statistic is crucial for evaluating the similarity or dissimilarity between the two areas. To get an idea of the increased difficulty, we point out that the total number of distinct species can be expressed as the sum of two quantities: the number of observed distinct species (a known quantity) and the number of unobserved species (an unknown quantity). Rather, for shared species, there are three distinct unknowns: species that are shared but undetected in both areas and species that are shared but observed in only one of the two areas. The exponential increase in the number of unknowns (which is $2^d - 1$) explains why the case of two areas is the most studied in the literature. For scenarios with $d > 2$ areas, comparisons are made between all possible pairs. In fact, commonly used multi-area indices are designed for the two-area case, since summarising comparisons across three or more areas into a single index is challenging, see Pan et al. (2009).

In the frequentist framework, the model assumes two discrete and unknown probability distributions, P_1 and P_2, with potentially different supports, whose intersection is not empty and is ordered in such a way that the first S species are the shared ones. Then, two random samples $\boldsymbol{X}_j = \left(X_{j,1}, \ldots, X_{j,n_j}\right)$, $j = 1, 2$, of sizes n_1 and n_2 are taken from P_1 and P_2, respectively. In this setting, Yue and Clayton (2012), Chuang et al. (2015) and Chao et al. (2017) proposed Turing-type estimators for the probability of discovering a new shared species in the next pair of observations. However, the three studies addressed only the case of one-step-ahead prediction, despite Yue and Clayton (2012) highlighting the need for sampling strategies where $m > 1$ pairs of observations are considered. Consequently, a generalisation of the Good-Toulmin estimator for the shared species problem remains open. On the other hand, various methods have been proposed to estimate shared species richness.

The seminal work by Chao et al. (2000) introduced a method based on sample coverage, which has been improved in two ways. The first improvement relies on Laplace's approximations (Chao et al., 2006b), while the second version presents a lower bound of the initial method (Pan et al., 2009), which proves to be more useful in practice. Finally, Chuang et al. (2015) presents a jackknife estimator for the richness of shared species. All these contributions are inferential in nature: they aim to estimate an unknown population-level quantity and rely on asymptotic assumptions, such as sufficiently large sample sizes. In other words, they answer the question, "How many species that we have not yet observed actually exist?", even though, under their assumptions, some of these species may never be encountered.

In the Bayesian nonparametric framework, a natural extension of the exchangeability assumption to the case of two areas is to assume that observations are partially exchangeable, meaning they are exchangeable within each group but not across the two groups. As in the single-group case, by virtue of de Finetti's representation theorem, this condition implies that the unknown distributions (P_1, P_2) form a vector of dependent random probability measures governed by a prior distribution $\mathscr{Q}_2$. Given (P_1, P_2), the observations are independent between groups and independent and identically distributed according to P_1 and P_2, respectively. Although Gibbs-type priors are an established choice for exchangeable data, the range of choices for $\mathscr{Q}_2$ is much broader in the two-group case. A key difficulty in this setting is that it is not enough to specify solely the marginal priors for P_1 and P_2, but modelling the dependence between these two distributions is crucial. Numerous constructions have been proposed in the literature, see e.g., Müller et al. (2004), Teh et al. (2006), Lijoi et al. (2014), Camerlenghi et al. (2019), and Bassetti et al. (2020). For a comprehensive overview, see Quintana et al. (2022), while we refer to the recent work by Franzolini et al. (2025) for a unified framework of multivariate species sampling models, offering new directions for this area of research. These approaches have been employed as Bayesian nonparametric priors in model-based clustering via mixture models. However, their application in species sampling problems has been limited by the lack of closed-form estimators, particularly for predicting shared species across areas. To the best of our knowledge, the main exceptions are Lijoi et al. (2008), Bacallado et al. (2015), and Camerlenghi et al. (2017), where posterior inference relies mainly on computational methods, while Franzolini et al. (2025) provided a closed-form species-discovery application for distinct species only. However, none of these contributions develops a closed-form multi-step distributional treatment for shared-species discovery, as we propose in the present work.

1.3 Summary of the Contribution

In this paper, we take a step forward and develop a Bayesian nonparametric methodology to study the problem of unseen distinct and shared species when observations are collected in two different areas. Our approach is predictive in nature and aims to answer the question: "How many species not yet observed in the two areas will be discovered in a future sample?" The proposed method is based on the Vector of Finite Dirichlet Process (Vec-FDP), introduced by Colombi et al. (2024), and provides closed-form expressions for estimating the quantity of interest. The Vec-FDP prior assumes a common species composition in the two areas, and that the number of species is finite, yet

random, while allowing for variation in their proportions. Working with finite samples, some species may be observed in only one area; however, they cannot be interpreted as structurally area-specific, since they have positive, albeit possibly negligible, probability of occurring in both areas.

Specifically, the main goals of this work are to derive a distributional theory for: (i) in-sample analysis for any observed samples of sizes n_1 and n_2, and (ii) out-of-sample predictions of the number of unseen distinct and shared species in additional unobserved samples of sizes m_1 and m_2. Hence, our results can be used to address sample size determination in sampling problems designed to detect distinct and shared species. Moreover, we are able to evaluate the sample coverage for distinct and shared species, i.e., the proportion of species observed in a sample (Good, 1953), both in-sample and out-of-sample. A remarkable finding is that our result holds for any finite future sample sizes m_1 and m_2, thereby extending the results in the frequentist literature, which only provides solutions for one-step ahead predictions. Unlike previous Bayesian contributions, our methodology provides a joint description of all the relevant quantities and yields an efficient computational strategy for evaluating the out-of-sample quantities of interest without relying on expensive and approximate computational methods. The proposed estimators are illustrated using both synthetic data, highlighting similarities and differences with frequentist approaches, and a real-world dataset. Specifically, we analyse an ant population in Trieste, sampled from two parks.

The rest of the paper is organized as follows. Section 2 describes the Vec-FDP prior model and its properties. The main theoretical results about the in-sample analysis and the out-of-sample prediction are presented in Section 3 and Section 4, respectively. Section 5 provides an interpretation of the model parameters in terms of diversity indices and outlines two strategies for their estimation. A simulation study is presented in Section 6 while the results of the analysis of a real dataset are given in Section 7. We conclude with a discussion in Section 8.

2 Vector of Finite Dirichlet Processes

Consider a sample $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ from $d = 2$ partially exchangeable sequences, that is, $\boldsymbol{X}_j = \left(X_{j,1}, \ldots, X_{j,n_j}\right)$, for $j = 1, 2$, each taking values in a Polish space $\mathbb{X}$. This is equivalent to assuming that the distribution of $\boldsymbol{X}$ is invariant under permutations occurring within elements of vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, but it is not invariant under permutations among them; see Camerlenghi et al. (2019) for a recent contribution. Hereafter, we refer to observations coming from two areas or groups, keeping in mind that the definition of a 'group' is problem-specific and can be interpreted in a broader sense.

By virtue of de Finetti's representation theorem, assuming partial exchangeability is equivalent to saying that there exists a vector (P_1, P_2) of random probability measures such that

$$\begin{aligned} (X_{1,i_1}, X_{2,i_2}) \mid (P_1, P_2) &\overset{\text{ind}}{\sim} P_1 \otimes P_2, \quad (i_1, i_2) \in \{1, \ldots, n_1\} \times \{1, \ldots, n_2\} \\ (P_1, P_2) &\sim \mathscr{Q}_2, \end{aligned} \tag{1}$$

where $\mathscr{Q}_2$ denotes the prior distribution of the vector in a Bayesian setting. The definition of $\mathscr{Q}_2$ is a crucial issue as it governs the properties of the statistical model that generates the data $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$. In this work, we study the model in Equation (1) under the Vector of Finite Dirichlet Process (Vec-FDP) prior, as introduced in Colombi et al. (2024). In this framework, the random probability measures P_1 and P_2 are constructively defined as follows,

$$P_j(A) \overset{a.s.}{=} \sum_{m=1}^{M} w_{j,m}\delta_{\tau_m}(A), \quad j = 1, 2, \tag{2}$$

where A is a measurable set, δ_{τ_m} stands for the delta-Dirac mass at τ_m, and the nonnegative weights $\boldsymbol{w}_j = (w_{j,1}, \ldots, w_{j,M})$ are the species' proportions specific to population j, which sum up to one almost surely. Then, defining a prior for (P_1, P_2) is equivalent to place a prior over $(M, \tau_1, \ldots, \tau_M, \boldsymbol{w}_1, \boldsymbol{w}_2)$. Here, conditionally on M, $(\tau_1, \ldots, \tau_M)$ are common random atoms across the two random probability measures, which are assumed to be independent and identically distributed with a common distribution P_0, that is a diffuse probability measure on $\mathbb{X}$. Moreover, we assume $\boldsymbol{w}_1$ independent of $\boldsymbol{w}_2$ and such that $\boldsymbol{w}_j \mid M \sim \text{Dir}_M(\gamma_j, \ldots, \gamma_j)$, for $j = 1, 2$, where $\text{Dir}_M(\gamma_j, \ldots, \gamma_j)$ denotes the M-dimensional symmetric Dirichlet distribution with group-specific parameter γ_j.

Finally, M is supposed to be a positive integer-valued random variable whose probability mass function is denoted as q_M. For the sake of simplicity, we follow Colombi et al. (2024) and choose a 1-shifted Poisson distribution, namely, $q_M(m) = e^{-\Lambda}\Lambda^{m-1}/(m-1)!$, for all integers $m \geq 1$, and denote it as $M \sim \text{Pois}_1(\Lambda)$. Nevertheless, we point out that all our results hold for any distribution on positive integers. This completes the prior specification, and we write

$$(P_1, P_2) \sim \text{Vec-FDP}(\Lambda, \boldsymbol{\gamma}, P_0), \tag{3}$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$. By placing a Vec-FDP prior on (P_1, P_2), we assume that the two groups share the same finite, yet random, number of species M that appear with different frequencies $w_{j,m}$ in the two groups, i.e., in the two areas. Note that the assumption of full sharing of atoms is a popular strategy in Bayesian nonparametrics, see e.g., Teh et al. (2006), Denti et al. (2023) and many examples in Franzolini et al. (2025). Since a Dirichlet prior is chosen for the weights, the same M species would eventually be observed in the two groups if we were able to obtain an infinite sample from P_1 and P_2. Thus, species observed only in one area in a finite sample should not be interpreted as structurally area-specific species: under the Vec-FDP prior, all species belong to the support of both populations, although their probability in one of the populations may be very small. In practice, we always work with finite samples of sizes $n_1 \geq 1$ and $n_2 \geq 1$.

Let $n = n_1 + n_2$ be the total number of observations. Let the integers label the observations according to their order of arrival by group, that is, observations are indexed first by group $j = 1, 2$, and then by within-group order of arrival. Within each group j, a random number of K_{j,n_j} distinct species will be observed, which we label $\boldsymbol{X}_j^* = \left\{X_{j,1}^*, \ldots, X_{j,K_{j,n_j}}^*\right\}$. Let $K_{j,n_j} = r_j$ be a realisation of this random variable.

Since P_1 and P_2 share the same support, ties between the two groups are expected, i.e., $\mathbb{P}\left(X^*_{1,k} = X^*_{2,k'}\right) > 0$. Thus, the set of labels for the distinct species in the entire sample is obtained as $\boldsymbol{X}^{**} = \boldsymbol{X}^*_1 \cup \boldsymbol{X}^*_2 = \left\{X^{**}_1, \ldots, X^{**}_{\mathcal{K}_{n_1,n_2}}\right\}$, where $\mathcal{K}_{n_1,n_2}$ denotes the overall number of distinct species, which is, in general, smaller than the sum of K_{1,n_1} and K_{2,n_2}. Such a difference is due to the random number of species shared between the two groups, namely

$$\mathcal{S}_{n_1,n_2} = K_{1,n_1} + K_{2,n_2} - \mathcal{K}_{n_1,n_2}. \tag{4}$$

In the following, we let $\mathcal{K}_{n_1,n_2} = r$ and $\mathcal{S}_{n_1,n_2} = t$ denote the realisations of the distinct and shared number of species, respectively. Moreover, we refer to group-specific quantities as *local* quantities, while we call *global* quantities those related to the joint sequence $\boldsymbol{X}$. Therefore, K_{j,n_j} is also named the local number of distinct species in group j while $\mathcal{K}_{n_1,n_2}$ is the global number of distinct species. We refer to Section S1.1 for a more detailed description of these quantities.

2.1 pEPPF and Predictive Distribution

The marginal law of the sample $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ from model (1) with $\mathscr{Q}_2$ chosen as in Equation (3) is uniquely determined by the species labels $\boldsymbol{X}^{**}$ and their abundances. The latter are defined by the vectors of frequency counts $\boldsymbol{n}_j = (n_{j,1}, \ldots, n_{j,r})$, where $n_{j,l}$ represents the number of observations in the j th group that coincide with the l distinct value, indexed according to the order of arrival by groups. These counts must satisfy the following constraints:

$$n_{j,l} \geq 0, \quad n_{1,l} + n_{2,l} > 0, \quad \sum_{l=1}^{r} n_{j,l} = n_j \quad l = 1, \ldots, r; \; j = 1, 2. \tag{5}$$

Consistent with the description in Section 2, some of the counts $n_{j,l}$ may also be zero.

Extending the results of Pitman (1996), one obtains that the marginal likelihood $\mathscr{L}(\boldsymbol{X})$, which is obtained by integrating (P_1, P_2) out of the model (1), admits the following factorisation: $\mathscr{L}(\boldsymbol{X}) = \mathscr{L}(\boldsymbol{n}_1, \boldsymbol{n}_2, \boldsymbol{X}^{**}) = \mathscr{L}(\boldsymbol{n}_1, \boldsymbol{n}_2) \prod_{l=1}^{r} P_0(\mathrm{d}X^{**}_l)$, see Franzolini et al. (2025). In what follows, we drop the value of the labels $\boldsymbol{X}^{**}$ as it is not relevant for our purposes. Consequently, the main object of interest is the law of abundances $\mathscr{L}(\boldsymbol{n}_1, \boldsymbol{n}_2)$. The law of the random partition induced by a partially exchangeable sequence $(\boldsymbol{X}_1, \boldsymbol{X}_2)$ is described through a probabilistic object called the partially Exchangeable Partition Probability Function (pEPPF); see, e.g., Lijoi et al. (2014) and Camerlenghi et al. (2019). Under a Vec-FDP prior, the pEPPF equals

$$\Pi^{(n)}_r(\boldsymbol{n}_1, \boldsymbol{n}_2) \;=\; V^r_{n_1,n_2} \prod_{j=1}^{d} \prod_{l=1}^{r} (\gamma_j)_{n_{j,l}}, \tag{6}$$

where $d = 2$, $(\boldsymbol{n}_1, \boldsymbol{n}_2)$ satisfy the constraints given in Equation (5) and

$$V^r_{n_1,n_2} \;=\; \sum_{m=1}^{\infty} (m)_{r\downarrow}\, q_M(m) \prod_{j=1}^{d} \frac{1}{(\gamma_j m)_{n_j}}. \tag{7}$$

See Section S3.1 and Colombi et al. (2024) for an alternative expression. In this work, we let $(m)_{r\downarrow} = m(m-1)\dots(m-r+1)$ denote the falling factorial of order r and $(x)_n = \Gamma(x+n)/\Gamma(x)$ is the Pochhammer symbol, also known as the rising factorial when n is a natural number. The sum in Equation (7) can start from $m = r$ since $(m)_{r\downarrow} = 0$ for all $m < r$. See Section S4 for further analysis and properties of the V coefficients, such as convergence, asymptotics, and a recurrence relationship.

The pEPPF in Equation (6) represents the sampling model we assume generates the data. However, since its form is rather involved and hard to interpret, it is common to describe the data-generating mechanism by inspecting the predictive distributions, which are a generalisation of the well-known Chinese restaurant franchise process introduced in Teh et al. (2006). Confining our attention to the first group, Colombi et al. (2024) showed that, taken $(\boldsymbol{n}_1, \boldsymbol{n}_2)$ observations as in Equation (5), the (n_1+1)th observation can either be equal to one of the 'old' (already observed) species, with probability proportional to $V^r_{n_1+1,n_2}(n_{1,l}+\gamma_1)$, or to a 'new' (never observed before) species, whose label is drawn from P_0, with probability proportional to $V^{r+1}_{n_1+1,n_2}\gamma_1$. Namely,

$$\mathbb{P}\left(X_{1,n_1+1} \in A \mid \boldsymbol{X}\right) = \frac{V^r_{n_1+1,n_2}}{V^r_{n_1,n_2}} \sum_{l=1}^{r} (n_{1,l}+\gamma_1)\,\delta_{X^{**}_l}(A) + \frac{V^{r+1}_{n_1+1,n_2}}{V^r_{n_1,n_2}}\gamma_1 P_0(A),$$

The case of a new observation in group 2 trivially follows. In Section S3, we generalise this result and report the predictive distribution for a new pair of observations, one in each group.

3 In-Sample Analysis

3.1 Correlation

A natural question that arises when moving from analysing a single group to multiple groups is quantifying the interaction between these two. Here, we aim to provide a quantitative answer by examining the statistical dependence between the data generating models for the two groups, namely between P_1 and P_2. Expanding the result of Colombi et al. (2024), we obtain a closed-form expression for the correlation between P_1 and P_2 when evaluated on the same measurable set A, namely

$$\operatorname{cor}\left(P_1(A), P_2(A)\right) = \frac{E\left(1/M\right)}{\sqrt{(1+\gamma_1)(1+\gamma_2)}\sqrt{E\left(\frac{1}{1+\gamma_1 M}\right)E\left(\frac{1}{1+\gamma_2 M}\right)}}. \tag{8}$$

The expression (8) does not depend on the choice of the set A. Thus, it may be considered an overall measure of dependence between the two random probability measures. Furthermore, if $M \sim \text{Pois}_1(\Lambda)$, then the numerator in Equation (8) is available in closed-form and it equals

$$E\left(1/M\right) = \Lambda^{-1}\left(1-\mathrm{e}^{-\Lambda}\right). \tag{9}$$

Additionally, the following limiting values hold:

$$\lim_{\gamma_1,\gamma_2\to 0} \operatorname{cor}\left(P_1(A), P_2(A)\right) = E\left(1/M\right), \qquad \lim_{\gamma_1,\gamma_2\to +\infty} \operatorname{cor}\left(P_1(A), P_2(A)\right) = 1.$$

The above limits suggest an interpretation of the γ_j's as homogeneity parameters: large values of γ_j's indicate similar groups that share most of the distinct species. Conversely, small values of γ_j's result in the minimum value of Equation (8). Interestingly, this value is not zero but depends on the choice of the prior distribution of M. Intuitively, larger expected values of M drive the correlation between the two populations toward zero.

3.2 In-Sample Statistics

In this section, we investigate the distributions of the most relevant in-sample statistics for a sample $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ of sizes n_1 and n_2 from model (1) under the Vec-FDP prior given in Equation (3). Mathematically, this is equivalent to studying the properties of the Bayesian nonparametric prior. We recall the quantities of interest introduced in Section 2 and further described in Section S1.1: the local number of distinct species, $\mathcal{K}_{j,n_j}$, one for each group, the global number of distinct species $\mathcal{K}_{n_1,n_2}$, and the number of shared species $\mathcal{S}_{n_1,n_2}$. Previous works derived the marginal distribution of both K_{j,n_j} (Lijoi et al., 2007; Argiento and De Iorio, 2022) and $\mathcal{K}_{n_1,n_2}$ (Colombi et al., 2024), as reported in Equations (S28) and (S31), respectively. Conversely, the distribution of $\mathcal{S}_{n_1,n_2}$ has not yet been derived, although it is linearly related to the number of local and global distinct species by Equation (4). This is also because the joint distribution of K_{1,n_1}, K_{2,n_2} and $\mathcal{K}_{n_1,n_2}$ is required to derive the distribution of $\mathcal{S}_{n_1,n_2}$. Theorem 3.1 overcomes this limitation.

Theorem 3.1. *Let $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ be a sample of sizes n_1 and n_2 from model (1) under the* Vec-FDP *prior in Equation (3). Then, the joint distribution of the local number of distinct species K_{1,n_1} and K_{2,n_2} and the global number of distinct species $\mathcal{K}_{n_1,n_2}$ equals*

$$\mathbb{P}\left(\mathcal{K}_{n_1,n_2} = r,\, K_{1,n_1} = r_1,\, K_{2,n_2} = r_2\right) = V^r_{n_1,n_2} \frac{r_1! r_2!}{r_1^*! r_2^*! t!} \prod_{j=1}^{2} |C(n_j, r_j; -\gamma_j)|, \tag{10}$$

for $r \in \{1, \ldots, r_1 + r_2\}$ and $r_j \in \{\{1, \ldots, \min\{r, n_j\}\}$ $(j = 1, 2)$ and where we defined $r_1^* = r - r_2$, $r_2^* = r - r_1$ and $t = r_1 + r_2 - r$. The coefficient $C(\cdot, \cdot; \cdot)$ in Equation (10) denotes the generalised factorial coefficient, as defined in Charalambides (2002).

The proof of Theorem 3.1 is provided in Section S5.1. The generalised factorial coefficients in Equation (10) can be computed via the triangular recurrence relationships described in Charalambides (2002). The quantities r_1^* and r_2^* represent the number of species observed only in groups 1 and 2, respectively, while t is the number of shared species between the two groups. See Section S2 for further details on generalized factorial coefficients. Equation (10) is a joint distribution and enables the evaluation of all other linearly dependent in-sample statistics. For instance, the distribution of the shared species, $\mathbb{P}\left(\mathcal{S}_{n_1,n_2} = t\right)$, is obtained by summing the expression in Equation (10) for all r, r_1, r_2 such that $t = r_1 + r_2 - r$, $t = 0, \ldots, \min\{n_1, n_2\}$. Alternatively, one may draw Monte Carlo samples from Equation (10) to obtain a Monte Carlo estimation of the distribution of interest. Furthermore, since Equation (4) is a linear relationship, the choice of $\mathcal{S}_{n_1,n_2}$ as a dependent variable is arbitrary. If

one is interested in only the global quantities, it is possible to derive the joint prior $\mathbb{P}\left(\mathcal{K}_{n_1,n_2} = r, \mathcal{S}_{n_1,n_2} = t\right)$, integrating out one of the local quantities. Such expression is reported in Section S5.2. Then, the marginal distribution of $\mathcal{S}_{n_1,n_2}$ is computed exactly as $\mathbb{P}\left(\mathcal{S}_{n_1,n_2} = t\right) = \sum_{r=t}^{n} \mathbb{P}\left(\mathcal{K}_{n_1,n_2} = r, \mathcal{S}_{n_1,n_2} = t\right)$.

4 Out-of-Sample Prediction

The present section addresses the task of out-of-sample prediction of new distinct and shared species. Given $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$, an additional sample comprising m_1 and m_2 individuals is considered, resulting in enlarged samples of sizes $n_1 + m_1$ and $n_2 + m_2$, namely $\left(X_{j,1}, \ldots, X_{j,n_j+m_j}\right)$, for $j = 1, 2$. Specifically, our interest lies in assessing the probability of discovering new local and global distinct species, as well as new shared species. Regarding the new local species, we follow the definition of Lijoi et al. (2007), i.e., $K_{j,m_j}^{(n_j)} = K_{j,n_j+m_j} - K_{j,n_j}$, for $j = 1, 2$. This definition can be extended to the global number of new distinct species, here defined as $\mathcal{K}_{m_1,m_2}^{(n_1,n_2)} = \mathcal{K}_{n_1+m_1,n_2+m_2} - \mathcal{K}_{n_1,n_2}$. We point out that the latter counts only the species that are unobserved in both groups. For example, observing a species that has already been spotted in the first group but was missing in the second would increase $K_{2,m_2}^{(n_2)}$ by one and leave $\mathcal{K}_{m_1,m_2}^{(n_1,n_2)}$ unchanged. Further caution is required when interpreting the number of new shared species, defined as $\mathcal{S}_{m_1,m_2}^{(n_1,n_2)} = \mathcal{S}_{n_1+m_1,n_2+m_2} - \mathcal{S}_{n_1,n_2}$. This not only takes into account the species that are new in both groups, but also the species that were first observed in one group only and belong to the additional sample of the other group. The relationship between such posterior quantities mimics that of their prior counterparts namely,

$$\mathcal{S}_{m_1,m_2}^{(n_1,n_2)} = K_{1,m_1}^{(n_1)} + K_{2,m_2}^{(n_2)} - \mathcal{K}_{m_1,m_2}^{(n_1,n_2)}. \tag{11}$$

We refer to Section S1.2 for a more detailed description of these predictive quantities.

4.1 Posterior of the Total Number of Species

The model introduced in Section 2 assumes a finite number of species M, which is usually unknown and assumed to be random. As data are observed, at least $\mathcal{K}_{n_1,n_2} = r$ species must exist, hence it is convenient to reparametrise the total number of species as $M = r + M^\star$, where $M^\star$ is interpreted as the random number of species yet not discovered. Following the posterior representation theorem for the Vec-FDP prior in Colombi et al. (2024), the posterior distribution of $M^\star$ has a probability mass function $q_{M|\boldsymbol{X}}^\star$ defined as

$$q_{M|\boldsymbol{X}}^\star(m^\star) = \frac{1}{V_{n_1,n_2}^r}(m^\star + r)_{r\downarrow} q_M(m^\star + r) \prod_{j=1}^{d} \frac{1}{(\gamma_j(m^\star + r))_{n_j}}, \tag{12}$$

where $m^\star = 0, 1, 2 \ldots$. Specifically, $M^\star$ may equal zero since $q_{M|\boldsymbol{X}}^\star(0) > 0$. See Section S6.1 for the equivalence between Equation (12) and the corresponding formulation presented in Colombi et al. (2024). Furthermore, in Section S6.2, we show that posterior

expected value of $M^\star$ equals $E\,(M^\star \mid \boldsymbol{X}) = V^{r+1}_{n_1,n_2}/V^{r}_{n_1,n_2}$. In particular, it admits the following asymptotic approximation for large sample sizes:

$$E\,(M^\star \mid \boldsymbol{X}) \;=\; (r+1)\,\frac{q_M(r+1)}{q_M(r)}\,(\gamma_1 r)_{\gamma_1}(\gamma_2 r)_{\gamma_2}\,n_1^{-\gamma_1}n_2^{-\gamma_2}\;(1+o(1))\;. \tag{13}$$

Equation (13) shows that $E\,(M^\star \mid \boldsymbol{X})$ goes to zero when $n_1, n_2 \to \infty$. This aligns with our modelling assumption, i.e., with an infinite amount of data, all possible species would have already been observed, leaving no room for further discoveries. Equation (13) also highlights the crucial role of the prior parameters in governing the discovery rate of new species. Under the choice $q_M(m) = \text{Pois}_1(\Lambda)$, the posterior expected value of $M^\star$ is approximately linear in Λ, since $q_M(r+1)/q_M(r) = \Lambda/r$. Regarding γ_1 and γ_2, values larger than one cause the expected number of undiscovered species to approach zero rapidly, implying that the potential for further discoveries becomes negligible after relatively few observations. Conversely, values substantially smaller than one allow new species to remain discoverable even for large observed sample sizes. This behavior is illustrated in Figure S3, which displays the exact posterior mass function $q^\star_{M|\boldsymbol{X}}$ for different combinations of $(\Lambda, \gamma_1, \gamma_2)$, evaluated at $r = 50$ and $(n_1, n_2) = (200, 200)$.

4.2 Joint Predictive Distribution

Following the same approach as in Section 3.2, it is of primary interest to derive both the marginal and joint distributions of the random variables on the right side of Equation (11). These are stated in the following theoretical results.

Theorem 4.1. *Let $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ be a sample of sizes n_1 and n_2 from model (1) under the* Vec-FDP *prior given in Equation (3). Let $\mathcal{K}_{n_1,n_2} = r$ and $\mathcal{S}_{n_1,n_2} = t$ be the observed number of global distinct and shared species, and let $K_{j,n_j} = r_j$, for $j = 1, 2$ be the observed local distinct species. Then,*

$$\mathbb{P}\left(\mathcal{K}^{(n_1,n_2)}_{m_1,m_2} = k,\; K^{(n_1)}_{1,m_1} = k_1,\; K^{(n_2)}_{2,m_2} = k_2 \mid \boldsymbol{X}\right) =$$
$$\frac{V^{r+k}_{n_1+m_1,n_2+m_2}}{V^{r}_{n_1,n_2}} \prod_{j=1}^{2} |C(m_j, k_j; -\gamma_j, -(\gamma_j r_j + n_j))| \sum_{s^*=0}^{k} \sum_{k_1^*=0}^{k-s^*} \frac{k_1! k_2!}{s^*! k_1^*! k_2^*!} \binom{r_1^*}{s_{12}} \binom{r_2^*}{s_{21}}, \tag{14}$$

for non-negative integers k, k_1, k_2 such that $0 \le k \le k_1 + k_2$ and $0 \le k_j \le m_j$ for $j = 1, 2$. Specifically, in Equation (14) we set $k_2^* = k - s^* - k_1^*$, $s_{12} = k_2 + k_1^* - k$ and $s_{21} = k_1 - k_1^* - s^*$. Finally, the coefficient $C(\,\cdot\,, \cdot\,; \cdot\,, \cdot\,)$ in Equation (14) denotes the non-central generalised factorial coefficient, as defined in Charalambides (2002).

The proof of Theorem 4.1 is provided in Section S6.4. The non-central generalised factorial coefficient satisfies a specific recursive relation that helps with its evaluation. See Charalambides (2002) for details. As discussed in Section S1.2, all auxiliary quantities involved in Equation (14) are interpretable. In fact, s^* is the number of new shared species among the k new distinct species, then k_1^* is the number of new distinct species in group 1 but missing in group 2 while s_{12} is the number of species that were first

only observed in group 1 and then also observed in group 2. Finally, k_2^* and s_{21} are defined accordingly. The marginal distributions for the local and global number of distinct species are reported in Equations (15) and (16). These are the posterior counterparts of Equations (S28) and (S31).

Proposition 1. *Under the same hypothesis of Theorem 4.1, the marginal distribution of the global number of new distinct species* $\mathcal{K}_{m_1,m_2}^{(n_1,n_2)}$ *is*

$$\mathbb{P}\left(\mathcal{K}_{m_1,m_2}^{(n_1,n_2)}=k \mid \boldsymbol{X}\right)=\frac{V_{n_1+m_1,n_2+m_2}^{r+k}}{V_{n_1,n_2}^{r}} \times \sum_{k_1^*=0}^{k}\sum_{k_2^*=0}^{k-k_1^*}\frac{(k_1^*+s^*)!(k_2^*+s^*)!}{k_1^*!k_2^*!s^*!}\prod_{j=1}^{2}|C(m_j,k_j^*+s^*;-\gamma_j,-(\gamma_j r+n_j))|, \tag{15}$$

for $k\in\{0,\dots,m_1+m_2\}$ *and where* $s^*=k-k_1^*-k_2^*$.

The proof of Proposition 1 is provided in Section S6.5. The marginal distribution of the local number of new distinct species in group j, $K_{j,m_j}^{(n_j)}$, follows from Equation (15) and it equals

$$\mathbb{P}\left(K_{j,m_j}^{(n_j)}=k_j \mid \boldsymbol{X}\right) = \frac{V_{n_j+m_j}^{r_j+k_j}}{V_{n_j}^{r_j}}|C(m_j,k_j;-\gamma_j,-(\gamma_j r_j+n_j))|, \tag{16}$$

for $k_j\in\{0,\dots,m_j\}$. The latter coincides with the findings in De Blasi et al. (2015) about Gibbs-type priors with negative parameters. Furthermore, we highlight that Theorem 3.1 requires $K_{j,n_j}\leq\mathcal{K}_{n_1,n_2}$. This condition, however, is not necessary in Theorem 4.1, as it is possible for the number of global discoveries to exceed the number of local discoveries, i.e., $K_{j,m_j}^{(n_j)}>\mathcal{K}_{m_1,m_2}^{(n_1,n_2)}$. An example illustrating this scenario is presented in Section S1.2.

4.3 Discovering Shared Species

Similarly to Section 3.2, Equations (16) and (15) allow us to compute the posterior expected values of the number of new distinct species and, by means of Equation (11), to derive the Bayesian estimator of the number of new shared species, that is

$$E\left[\mathcal{S}_{m_1,m_2}^{(n_1,n_2)} \mid \boldsymbol{X}\right]=E\left[K_{1,m_1}^{(n_1)} \mid \boldsymbol{X}\right]+E\left[K_{2,m_2}^{(n_2)} \mid \boldsymbol{X}\right]-E\left[\mathcal{K}_{m_1,m_2}^{(n_1,n_2)} \mid \boldsymbol{X}\right]. \tag{17}$$

The associated uncertainty is quantified through the posterior distribution, namely, $\mathbb{P}\left(\mathcal{S}_{m_1,m_2}^{(n_1,n_2)}=s \mid \boldsymbol{X}\right)$ that is obtained by summing the expression in Equation (14) for all k,k_1,k_2 such that $s=k_1+k_2-k$ and for all $s=\{0,\dots,m_1+m_2\}$. Alternatively, it can be estimated via Monte Carlo sampling, drawing samples from Equation (14).

The shared species sample coverage is defined as the proportion of shared species that are observed in the sample. Being able to estimate it allows us to assess the number of shared species that have been observed, and therefore, to decide whether it is

worth continuing the experiment and sampling more observations. Moreover, the shared species sample coverage on m-steps ahead facilitates determining the size of the additional sample that ensures the coverage exceeds a specified threshold. In our setting, the shared coverage probability for m-steps ahead, i.e., the probability of not discovering new shared species in the additional sample, follows from Equation (14) and equals

$$\mathbb{P}\left(\mathcal{S}^{(n_1,n_2)}_{m_1,m_2}=0 \mid \boldsymbol{X}\right)=\sum_{k_1=0}^{m_1}\sum_{k_2=0}^{m_2}\frac{V^{r+k_1+k_2}_{n_1+m_1,n_2+m_2}}{V^{r}_{n_1,n_2}}\prod_{j=1}^{d}|C(m_j,k_j;-\gamma_j,-(\gamma_j r_j+n_j))|. \quad (18)$$

In particular, the one-step ahead coverage probability is obtained from Equation (18) by setting $m_1=1$ and $m_2=1$. In this case, we can explicitly write the whole distribution $\mathbb{P}\left(\mathcal{S}^{(n_1,n_2)}_{1,1}=s \mid \boldsymbol{X}\right)$ for each $s\in\{0,1,2\}$ and not just for $s=0$ as happens for $m_j>1$. For the sake of brevity, such distribution is reported in Section S7, alongside the probability of discovering new local and global distinct species. It is worth noting that increasing values of the probability in Equation (18) indicate that the observed sample is sufficiently exhaustive, suggesting that further data collection may not be necessary.

Here, we report the probability of discovering at least one new shared species;

$$\begin{aligned}\mathbb{P}\left(\mathcal{S}^{(n_1,n_2)}_{1,1}>0 \mid \boldsymbol{X}\right) \;=\; & 1-\frac{V^{r}_{n_1+1,n_2+1}}{V^{r}_{n_1,n_2}}(\gamma_1 r_1+n_1)(\gamma_2 r_2+n_2)\\ & -\frac{V^{r+1}_{n_1+1,n_2+1}}{V^{r}_{n_1,n_2}}\left\{\gamma_1(\gamma_2 r_2+n_2)+\gamma_2(\gamma_1 r_1+n_1)\right\}-\frac{V^{r+2}_{n_1+1,n_2+1}}{V^{r}_{n_1,n_2}}\gamma_1\gamma_2.\end{aligned} \quad (19)$$

The ratios of V coefficients in Equation (19) represent the three different scenarios where the new pair of observations yields none, one, or two new distinct global species. Equation (19) can be compared with the Good-Turing estimators proposed in the frequentist literature, as presented in Section 6.3 where a simulation study is carried out. Nevertheless, to the best of our knowledge, neither the frequentist nor the Bayesian literature provides an analogue of the m-step-ahead discovery probability in Equation (18); we illustrate its practical use in Section 7.

5 Diversity Indices

In ecology, the concept of diversity is tied not only to the number of distinct species present in an area but also to their heterogeneity. For instance, having ten equally represented species is intuitively very different from having one highly abundant species and the remaining nine extremely rare (Colwell et al., 2009). In the literature, there is no unique quantitative definition of diversity; instead, a variety of indices have been proposed to measure it. Among these, we focus on Simpson's diversity index (Simpson, 1949), which captures both richness and evenness in species distributions. Assuming that the unknown discrete distribution P_j that generates the population is made up of

M distinct species with species proportions $w_{j,m}$, Simpson's diversity index is

$$\rho_j = \sum_{m=1}^{M} w_{j,m}^2. \tag{20}$$

Useful alternatives are suitable transformations of ρ_j, such as the Gini-Simpson index, defined as $1-\rho_j$, or the inverse-Simpson index, $1/\rho_j$ (Colwell et al., 2009). The index ρ_j ranges between $1/M$ (when all species are uniformly distributed) and one (when one species is abundant and the remaining are negligible). These extreme cases correspond to the cases of maximum and minimum heterogeneity, respectively.

When we assume the population is generated from a vector (P_1, P_2) of unknown discrete distribution, each having M different species whose proportions are $w_{j,m}$, for $j = 1, 2$ and $m = 1, \ldots, M$, the Morisita index (Morisita, 1959) can be used to quantify the similarity between the two areas, namely, $2\rho_{12}/(\rho_1 + \rho_2)$, where

$$\rho_{12} = \sum_{m=1}^{M} w_{1,m} w_{2,m}. \tag{21}$$

Increasing values of the Morisita index show evidence of identical communities, i.e., with the same species proportions. Chao et al. (2017) highlighted an important probabilistic interpretation of the Morisita index. The numerator in Equation (21) represents the probability of selecting the same shared species when two observations, one from each group, are randomly drawn from this population. The denominator of the Morisita index is instead the sum of the Simpson's diversity indices in the two groups. In Section S9, we present the Bayesian estimators of ρ_j, $j = 1, 2$, and ρ_{12} given an observed sample $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ of sizes n_1 and n_2, with r distinct species and t shared species.

5.1 Parameter Estimation

The Bayesian model in Equation (1) under the Vec-FDP prior in Equation (3) is governed by the parameters $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$ and Λ. These parameters are unknown and must be estimated from the data. To this end, we propose two different estimation strategies. The first strategy, following Camerlenghi et al. (2024), is a plug-in approach based on the maximum marginal likelihood estimator. Namely, we find $\hat{\boldsymbol{\gamma}}$ and $\hat{\Lambda}$ that maximise the pEPPF in Equation (6). In the following, we refer to *Bayes I* estimators as the estimators proposed in the previous sections for the global and local quantities when the parameters $\boldsymbol{\gamma}$ and Λ are estimated via the maximum marginal likelihood estimator.

In addition, we provide an alternative estimation strategy that relies on a diversity-based interpretation of $(\boldsymbol{\gamma}, \Lambda)$. To this end, note that Equation (1) does not assume the existence of a unique, unknown vector of probability distributions (P_1, P_2) but rather assumes that (P_1, P_2) is random. Therefore, we can integrate out this source of randomness by taking the expected values, obtaining

$$E(\rho_j) = E\left(\sum_{m=1}^{M} w_{j,m}^2\right) = (1+\gamma_j)\, E\left(\frac{1}{1+\gamma_j M}\right), \tag{22}$$

$$E(\rho_{12}) = E\left(\sum_{m=1}^{M} w_{1,m}w_{2,m}\right) = E(1/M)\ . \tag{23}$$

Proof of Equations (22) and (23) are deferred to Section S8. The expected value of ρ_{12} in Equation (23) solely depends on q_M. This is due to the fact that we model the dependence in the vector of random probability measures (P_1, P_2) through the random number of species M and by imposing common atoms. Note that our choice of q_M as $\text{Pois}_1(\Lambda)$ yields explicit expressions for Equation (23) depending on Λ; see Equation (9). Then, Λ regulates the amount of similarity between the two areas. Furthermore, the limits of the expected Simpson index in Equation (22) further illuminate the interpretation of γ_1 and γ_2 as homogeneity parameters, as described in Section 3.1. In fact, the limits of Equation (22) for $\gamma_j \to 0$ and $\gamma_j \to \infty$ are equal to one and $E(1/M)$, respectively, which represent the cases of minimum and maximum heterogeneity. Therefore, γ_1 and γ_2 are homogeneity parameters since heterogeneity increases as γ_j increases.

Our diversity-based strategy for estimating γ_1, γ_2 and Λ consists of two steps. Firstly, we use the observed data to get estimates of ρ_j, $j = 1, 2$, in Equation (20) and ρ_{12} in Equation (21). This step can be achieved with standard routines, such as using the estimator in Equation (S53). Then, we plug such estimates into the left-hand sides of Equations (22) and (23) and solve with respect to γ_1, γ_2, and Λ. In the following, we refer to *Bayes II* estimators as the estimators for the local and global quantities, when using the strategy just described to estimate $\boldsymbol{\gamma}$ and Λ. In Section 6, we show that the computational effort required by the diversity-based strategy is negligible compared to the alternative based on the maximum marginal likelihood.

We conclude by noting that other estimation procedures can also be considered. For instance, in the case of a single population, Lijoi et al. (2007) and Favaro et al. (2009) suggested maximising the EPPF evaluated at the observed partition. Alternatively, Balocchi et al. (2024a) employed a fully Bayesian approach by specifying suitable hyperpriors and estimating the parameters via Markov chain Monte Carlo (MCMC) methods.

6 Simulation Study

6.1 Data Generation and Competitors

We conduct a simulation study to evaluate our methodology and compare it with existing estimators. The data-generating mechanism, adapted from Yue and Clayton (2012), assumes that (P_1, P_2) are two discrete probability distributions, each consisting of M_{true} species. The group-specific species proportions are denoted by $p_{j,m}$, with $j = 1, 2$ and $m = 1, \ldots, M_{\text{true}}$, which are specified according to several alternatives to cover a wide range of benchmark distributions and misspecification scenarios. In particular, we consider the following alternatives, referred to here as settings:

(i) Dirichlet weights – we set $M_{\text{true}} = 60$ and $p_{j,m}$ to be randomly generated from symmetric Dirichlet distributions with parameters $\gamma_j \in \{0.1, 0.5\}$; all three possible combinations are denoted D_1, D_2, D_3 as explained in Table S4;

(ii) Geometric weights – we set $M_{\text{true}} = 60$ and $p_{j,m}$ to be deterministically assigned according to a geometric decay, i.e., $p_{j,m} \propto \alpha_j^m$, with $\alpha_j \in \{0.8, 0.85, 0.9\}$, as in Yue and Clayton (2012); all possible combinations of these three values are considered, yielding six cases denoted as $G_1, \ldots, G_6$ and explained in Table S5;

(iii) Zipf's weights – we set $M_{\text{true}} = 60$ and $p_{j,m}$ to be deterministically assigned as $p_{j,m} \propto m^{-s_j}$, with $s_j \in \{1.3, 2\}$, similar to Franzolini et al. (2025); all three possible combinations are denoted Z_1, Z_2, Z_3 and explained in Table S6.

In the Geometric and Zipf settings, the M_{true} species in each population are randomly permuted before assigning the probability mass function, see Section S10 for details. The empirical values of the Morisita index after shuffling are reported in Figure S4. Finally, Figures S5–S7 display the accumulation curves for the four statistics of interest for some selected cases (D_3, G_6, Z_3).

All three settings described above satisfy the assumption that the same set of M_{true} species is shared. However, the accumulation curves of shared species show that, for a finite number of observations, the number of observed shared species is smaller than M_{true}. Nevertheless, we consider two further additive settings in which this assumption no longer holds. Specifically, inspired by (Müller et al., 2004), we define the components of (P_1, P_2) as $P_j = cQ_0 + (1-c)Q_j$, for any scalar $c \in [0, 1]$. The shared component Q_0 is assumed to follow a symmetric Dirichlet distribution of size M_0 with parameter δ_0, while the two idiosyncratic components Q_j, $j = 1, 2$, follow symmetric Dirichlet distributions of sizes M_{id}^1 and M_{id}^2 with parameters δ_1 and δ_2, respectively. In particular, we set $\delta_1 = \delta_2$, denoting this common value as δ in the following. Overall, the total number of species is $M_{\text{tot}} = M_0 + M_{\text{id}}^1 + M_{\text{id}}^2$. In the first additive experiment, we fix $M_{\text{tot}} = 80$ and set $M_0 = cM_{\text{tot}}$, rounded down to the nearest integer. The remaining species are equally divided between the two idiosyncratic components, i.e., $M_{\text{id}}^j = (M_{\text{tot}} - M_0)/2$ for $j = 1, 2$, with rounding applied to ensure that the total number of species matches the desired value. Regarding the choice of parameters, we consider all configurations with $c \in \{0, 0.5, 1\}$ and $(\delta_0, \delta) \in \{0.1, 0.5\}^2$, resulting in a total of 12 configurations, denoted as $A_{1,1}, \ldots, A_{3,4}$. The corresponding names are reported in Table S7. Figures S8–S9 display the accumulation curves for the four statistics of interest for some selected cases $(A_{1,4}, A_{3,4})$. A second additive experiment considers instead an unbalanced species composition, with $M_0 = 40$, $M_{\text{id}}^1 = 0$, and $M_{\text{id}}^2 = 40$. Accordingly, the first group contains only the common component, so that $P_1 = Q_0$, whereas the second group is defined as $P_2 = cQ_0 + (1-c)Q_2$. Thus, the species in the first group are a subset of those in the second group, whose total number is twice as large. We consider all configurations with $c \in \{0.25, 0.5, 0.75\}$, $\delta_0 = 0.5$, and $\delta \in \{0.001, 0.5, 2\}$. Note that when δ is very small, a few group-specific species have relatively large probabilities, while many others are extremely rare. Conversely, when δ is larger, the group-specific species are approximately equally likely. The resulting 9 configurations, denoted as $U_{1,2}, \ldots, U_{3,3}$ are explained in Table S8. Figures S10–S11 display the accumulation curves for the four statistics of interest for selected cases $(U_{1,1}, U_{1,2})$. Finally, in each setting, we independently sample n_j species with replacement from P_j, for $j = 1, 2$. Here, we consider the balanced case where $n_1 = n_2$, while we defer the unbalanced case with $n_1 \ll n_2$ to Sections S10.1 and S10.2 of the Supplementary material. In both cases, we denote the total number of observations by $n = n_1 + n_2$.

Whenever possible, we compare the performance of our estimators with that of two simpler benchmark models. The first benchmark models the two populations independently, assuming that each area-specific model coincides with the marginals induced by our proposed framework. Specifically, we consider $P_j = \sum_{m=1}^{M_j} w_{j,m}\delta_{\tau_{j,m}}$, for $j = 1, 2$ where the collections $\{M_j, (w_{j,1}, \ldots, w_{j,M_j}), (\tau_{j,1}, \ldots, \tau_{j,M_j})\}$ are distributed independently as a Finite Dirichlet Process (FDP; Argiento and De Iorio 2022), i.e., $P_j \sim \text{FDP}(\Lambda_j, \gamma_j)$. In the following, we refer to this model and the corresponding estimators as *Independent*. Since any dependence between the two areas is completely ignored, this approach is only suitable for the estimation of local quantities such as K_{j,n_j} and $K^{(n_j)}_{j,m_j}$. The second benchmark model, instead, discards any information about group membership and pools all observations as if they originated from a single area. For this reason, we refer to this approach as *Pooled*. In this case, the merged sample $\boldsymbol{X}_1 \cup \boldsymbol{X}_2$ is assumed to be exchangeable from a single homogeneous population, so that $P_1 = P_2 = P$, with $P \sim \text{FDP}(\Lambda_0, \gamma_0)$. Under this assumption, information on local quantities is irretrievably lost. However, we can still compare the global number of distinct species $\mathcal{K}^{(n_1,n_2)}_{m_1,m_2}$ with the corresponding quantity computed under the *Pooled* model, namely $\mathcal{K}^{(m_1+m_2)}_{n_1+n_2} = K_{m_1+m_2} - K_{n_1+n_2}$. The *Independent* and *Pooled* estimators of the quantities of interest are obtained by estimating the corresponding parameters, (Λ_j, γ_j) for $j = 1, 2$ and (Λ_0, γ_0) via maximum marginal likelihood. An explicit expression for the associated EPPF can be found in Argiento and De Iorio (2022). The performance of our proposed model is evaluated through two separate experiments, described in Sections 6.2 and 6.3, respectively. Additional experiments illustrate the m-steps-ahead discovery probabilities (Section S10.3) and the comparison with the shared-richness estimator of Chao et al. (2000) (Section S10.4).

In the case of one-step-ahead discovery probability of shared species, we compare our approach against two nonparametric frequentist competitors that represent two alternative generalisations of the Good-Turing estimator for the problem of shared species. Specifically, let f_{ν_1,ν_2} be the number of species that appeared exactly ν_1 and ν_2 times in the first and second groups, respectively. Moreover, let $f_{\nu_1,+}$ (f_{+,ν_2}) be the number of species that appeared exactly ν_1 (ν_2) times in the first (second) group and at least once in the second (first) group. Then, the estimator proposed by Yue and Clayton (2012) (*Yue*) is $\mathbb{P}_{\text{Yue}}\left(\mathcal{S}^{(n_1,n_2)}_{1,1} > 0\right) = (f_{1+} + f_{+1} + f_{11})/n_1$, and it is only defined for $n_1 = n_2$. On the other hand, Chao et al. (2017) (*Chao*) proposed $\mathbb{P}_{\text{Chao}}\left(\mathcal{S}^{(n_1,n_2)}_{1,1} > 0\right) = f_{1+}/n_1 + f_{+1}/n_2 + f_{11}/n_1 n_2$.

6.2 Experiment 1

The first experiment assesses the ability of the estimators introduced in Section 4 to predict the number of additional local and global distinct species, as well as the number of shared species, in a future unobserved test set. The experiment proceeds as follows. First, we generate a training dataset according to the data-generating mechanism described in Section 6.1, with balanced sample sizes $n_1 = n_2$ taking values in the set $\{100, 200, \ldots, 800\}$. We then generate $m_1 = m_2 = 200$ additional observations, which constitute the unobserved test set. This procedure is repeated over 100 independently

generated pairs of training and test sets. Each training set is used to estimate the model parameters as described in Section 5.1. We recall that *Bayes I* and *Bayes II* estimators refer to maximum marginal likelihood and diversity-based estimation strategies, respectively. These estimates are subsequently employed to predict the expected number of new species in the test set, following the methodology outlined in Section 4. Specifically, we use Equation (14) to compute $E\left[K_{j,m_j}^{(n_j)} \mid \boldsymbol{X}\right]$ for $j = 1, 2$ as well as $E\left[\mathcal{K}_{m_1,m_2}^{(n_1,n_2)} \mid \boldsymbol{X}\right]$. Finally, $E\left[\mathcal{S}_{m_1,m_2}^{(n_1,n_2)} \mid \boldsymbol{X}\right]$ is computed as in Equation (17). We compare the estimates of the new local and global distinct species with the *Independent* and *Pooled* estimators introduced in Section 6.1. However, we are not aware of any m-steps ahead estimator for the shared species.

Out-of-sample performance is evaluated by computing the Root Mean Squared Error (RMSE) of each estimated quantity with respect to its true value. For the sake of space, Figure 1 reports results only for selected scenarios $(D_3, G_6, Z_3, A_{1,4}, A_{3,4}, U_{1,1}, U_{1,2})$ when the training sample size is fixed at $n_1 = n_2 = 400$. The results for all remaining configurations are reported in the Github repository https://github.com/alessandrocolombi/HSSM. Figure 1 highlights that *Bayes I* and *Bayes II* are the only procedures that provide simultaneous out-of-sample predictions for all four target quantities. In contrast, *Pooled* and *Independent* focus on a single quantity and therefore cannot deliver a coherent joint assessment. Notably, our method is the only one that provides predictions for the number of new shared species. Scenario Z_3 is the most challenging, as evidenced by the uniformly larger RMSEs across all quantities. In this setting, *Bayes II* remarkably outperforms the competitors, while *Bayes I* improves upon *Pooled* for global distinct discoveries and remains comparable to *Independent* for local discoveries. These findings indicate that the predictive performance of our approach is robust to model misspecification, despite the marked difference between the Zipf weights in Z_3 and the Dirichlet-type decay assumed by the model. A similar behavior is observed in scenario $A_{1,4}$, where, by construction, there are no shared species. Nevertheless, under both estimation strategies, the predicted number of new shared species is correctly concentrated at zero, while the remaining quantities exhibit small median RMSEs, typically below one, with slightly greater variability under *Bayes II*. Overall, our estimators remain competitive with the benchmark methods while offering the advantage of relying on a unified model at the cost of only one additional parameter. As expected, the main exception is scenario $A_{3,4}$: this configuration is generated assuming a single common population. Here, *Pooled* estimator achieves more accurate predictions for the number of new global distinct species.

In addition, Table S9 reports the mean and standard deviation of the estimated number of yet-undiscovered species $M^\star$, the total number of species, $K_n + M^\star$, and the model parameters $(\Lambda, \gamma_1, \gamma_2)$. Regarding the estimation of $K_n + M^\star$, the true theoretical value is recovered in settings D_3 and G_6. By contrast, it is apparently underestimated in settings $A_{3,4}$, $U_{1,1}$, and $U_{1,2}$. However, these estimates are consistent with the saturation level of the accumulation curves for the observed global number of species (see Figures S9–S11), suggesting that the total number of species may not be identifiable from the data. Moreover, $K_n + M^\star$ is severely overestimated in settings Z_3 and $A_{1,4}$. These scenarios provide relevant examples of why our model cannot be reliably used to

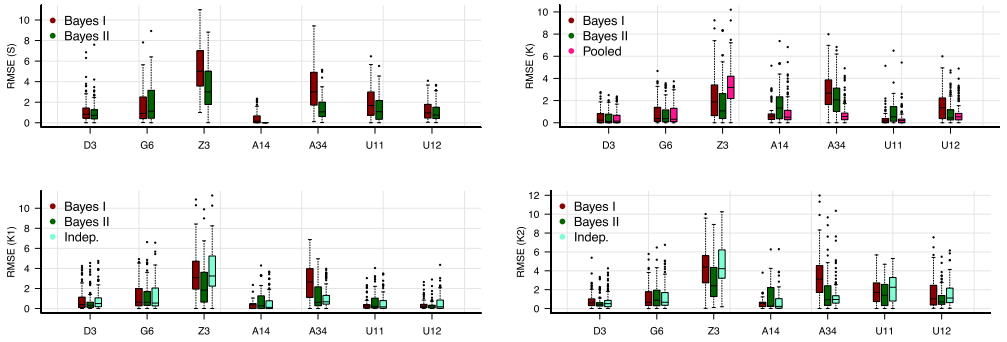


Figure 1: Experiment 1: RMSE of out-of-sample predictions for new shared species (top-left panel), new global distinct species (top-right panel), and new local distinct species (bottom-left panel) and (bottom-right panel) across selected scenarios.

estimate species richness, a task that lies beyond the scope of this work, because of the assumption of full sharing of atoms discussed in Section 2. Nevertheless, such overestimation does not compromise the predictive performance for the quantities of primary interest, as illustrated in Figure 1. Importantly, this robustness concerns prediction over the finite horizons considered here. It should not be interpreted as robustness to unequal supports for arbitrarily long prediction horizons, since, for sufficiently large future samples, predictions necessarily reflect the model assumption that every species has positive probability in both populations and will eventually be observed in both.

Finally, Figure S12 shows that the computational time required by *Bayes II* to fit the model and perform out-of-sample predictions is much shorter than *Bayes I*. Indeed, the latter involves the computation of the marginal likelihood and becomes more expensive as the sample size increases, while *Bayes II* leverages a cost of parameter estimation that is independent of the observed sample size. See Section S10.1 for further details. More broadly, these results highlight the substantial computational benefit of our approach relative to previous solutions available in the literature. Thanks to closed-form expressions, prediction can be performed in less than half a second. Such computational times are typically unattainable for more complex Bayesian nonparametric models that require MCMC methods.

6.3 Experiment 2

The second experiment assesses the ability of our model to estimate the one-step ahead probability of discovering a new local or global distinct species, as well as a new shared species. The estimator for the shared discovery probability is given in Equation (19) while those related to local and global quantities are reported in Equations (S51) and (S52). We compare our estimators with those discussed in Section 6.1.

Given (P_1, P_2) generated as described in Section 6.1, we evaluate all competing methods on a grid of sample sizes $n_1 = n_2 = \{50, 100, \ldots, 400\}$. The experiment is then

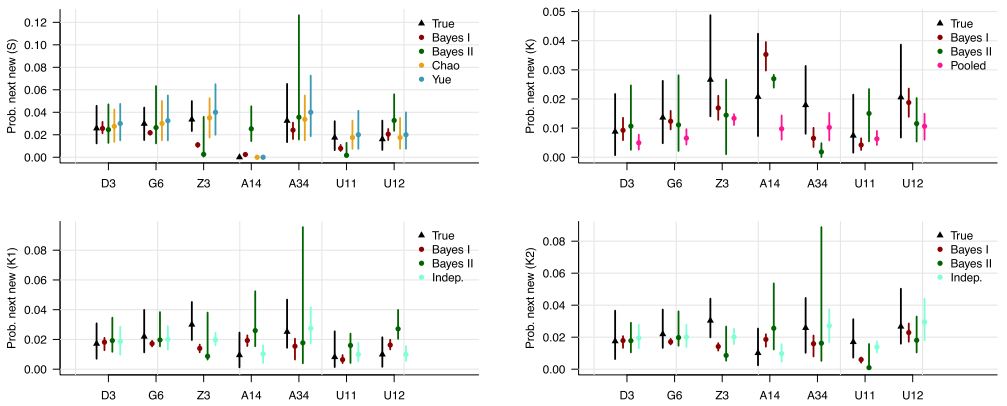


Figure 2: Experiment 2: one-step-ahead prediction probability for new shared species (top-left panel), new global distinct species (top-right panel), and new local distinct species (bottom-left panel) and (bottom-right panel) across selected scenarios.

repeated for 100 independently generated datasets. Results are compared with the true probability of discovering a new species, which is given in Equation (S64). Figure 2 reports the results for selected scenarios $(D_3, G_6, Z_3, A_{1,4}, A_{3,4}, U_{1,1}, U_{1,2})$ and only for sample size $n_1 = n_2 = 400$. Results for all remaining configurations are reported on the Github repository https://github.com/alessandrocolombi/HSSM.

Figure 2 reports the one-step-ahead discovery probabilities, where the black line represents the oracle benchmark; therefore, estimates closer to the black curve indicate better performance. Overall, *Bayes I* exhibits very low variability across replications, whereas *Bayes II* is noticeably more variable, reflecting its different parameter-fitting strategy. In terms of accuracy, both Bayesian estimators consistently outperform the *Pooled* baseline and often achieve results comparable to those of the *Independent* approach, although the latter performs better in some settings. This indicates that the simpler competitor remains a valid alternative when inference is restricted to group-specific quantities. However, it does not allow for a joint assessment of all quantities, which is a key advantage of our model-based approach. Importantly, this experiment also includes competitors for shared-species discovery probabilities: both *Bayes I* and *Bayes II* are broadly in line with the state-of-the-art frequentist estimators. Finally, we note that, unlike the frequentist competitors, our appraoch is not restricted to one-step-ahead prediction but naturally extends to m-step-ahead coverage probabilities, as illustrated in Section S10.3.

7 Analysis of Ants Data

We illustrate the methodology outlined in the previous sections to analyse a real-world dataset coming from a case study conducted by Zara et al. (2021) to evaluate the impact of urbanisation on biodiversity. Although urbanisation is often cited as one of the main

drivers of species extinction due to factors such as pollution, changes in land use, and the introduction and spread of alien species, the authors argue that urban green spaces can serve as important refuges supporting high levels of species diversity. To investigate this topic, the researchers selected several sites in Trieste, a city in north-eastern Italy, representing different levels of urbanisation. For our analysis, we focus on two of these areas because of their ecological distinction: Bosco Bovedo (BB) and Orto Lapidario (OL). The former is a semi-natural urban forest located just outside residential areas, acting as a transition zone between urban and natural environments. The latter is a city park within a local museum complex. The two selected sites are sufficiently far from each other (about 6 kilometres) to rule out migratory contamination between the areas. Hence, we use the methodology introduced in Section 2 to model the two areas using separate distributions, thereby accounting for possible differences in the underlying distributions due to their spatial locations.

Ground-dwelling arthropods were sampled using pitfall traps. Formicidae were identified at the species level on the basis of their morphology, and their abundance was recorded. Overall, a total number of $2,971$ and $3,489$ ants were collected in the first area (BB) and in the second area (OL), respectively. However, 2037 out of 2971 (around 68%) observations in the first group belong to the same species (crematogaster schmidti), which we excluded from the analysis. Similarly, in the second area (OL), 1229 out of 3489 (around 35%) also belong to the same species (pheidole pallidula), which we also excluded. This procedure is common in ecological studies, as highly abundant species are known to contribute little to understanding species diversity. Hence, the dataset analysed consists of $n_1 = 934$ and $n_2 = 2235$ observations. The observed number of local distinct species is $r_1 = 17$ and $r_2 = 23$, while the global number of distinct species and the number of shared species are $r = 30$ and $t = 10$, respectively. The observed species proportions, sorted in decreasing order, are displayed in Figure S19, while the species accumulation curves are reported in Figure S20.

To analyse the ants dataset, we employ the same strategy introduced in the simulation study. First, we evaluate the prediction performance in terms of the number of new shared and distinct species in an additional, unobserved sample. Secondly, we assess the m-steps-ahead prediction performance, starting from $m = 1$. For both cases, the competing approaches are those described in Section 6.1.

Since this is a real dataset and the true future discoveries are not observable, we adopt a training and test split strategy to empirically assess predictive performance. Specifically, we partition the observed data into a training set and a test set: the training set is used to estimate the model parameters, which are then plugged into the proposed estimators to predict the number of new species in the test set. We repeat the analysis for several training set proportions, namely {10%, 30%, 50%, 70%, 90%}. For brevity, we report here only the results corresponding to a 50% training set, with the remaining 50% used as test data. The remaining proportions are deferred to Section S11. The experiments are repeated over 100 independent splits obtained by sampling without replacement from the full dataset. In particular, each split is constructed so as to preserve the original imbalance between the two areas.

Following Section 6.2, the predictive performance on the test set is evaluated using RMSE, reported in the left panel of Figure 3. The benefits of our approach are

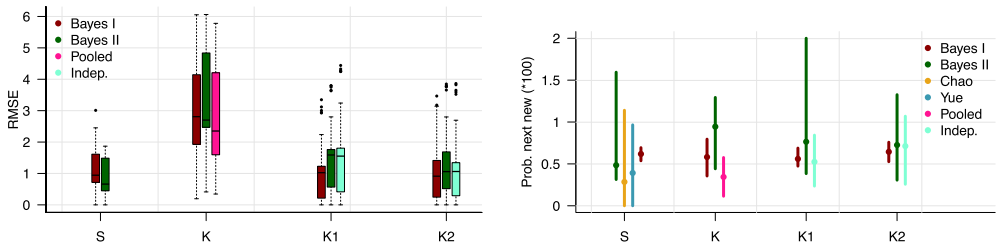


Figure 3: RMSE of out-of-sample predictions (left panel) and one-step-ahead discovery probability (right panel) for shared (S), global distinct (K), and local distinct (K1, K2) species. Probabilities on the rightmost plot have been multiplied by 100.

particularly evident for shared species, for which no natural competitors are available. Improvements are also observed for the prediction of new local distinct species in the first area, which is the one with the smaller sample size. Here, *Bayes I* achieves a lower RMSE than the *Independent* estimator, likely because it borrows strength information from the second area through the dependence structure induced by the model. This interpretation is supported by Figure S21, which reports the additional cases where the training set is smaller (10% and 30%). The proposed estimators, especially *Bayes I*, achieve lower prediction error than the competitor models.

Regarding the one-step-ahead prediction, the right panel of Figure 3 reports the estimated discovery probabilities under *Bayes I*, *Bayes II*, and the competing methods for a total sample size $n = 350$. Additional results for other sample sizes, $n \in \{50, 250, 450, 600\}$, are reported in Figure S22. We did not consider larger sample sizes since the probabilities are too small to be meaningful. Regarding the shared species discovery probability, the lower tails of the intervals for both *Yue* and *Chao* include zero; therefore, they do not provide any guarantee of discovering a new shared species, despite the fact that the observed data suggest that not all shared species have been detected yet. In contrast, our model yields a discovery probability for new shared species that is strictly greater than zero. Furthermore, the probability of discovering a new global distinct species is smaller under the *Pooled* estimator than under our proposed estimators. Overall, the probabilities involved are very close to zero, in agreement with the accumulation curves in Figure S20, which appear close to saturation. This suggests that discovering something new in a single additional draw is unlikely.

For this reason, we strengthen our analysis by considering the m-steps-ahead discovery probability, with $m > 1$, of observing at least one new shared or distinct species. To illustrate this, we propose a two-dimensional visualisation of discovery probabilities across both the current sampling effort and the planned future effort. Specifically, for each quantity of interest, we represent the probability of making at least one new discovery as a function of: (i) the current sample size (horizontal axis), considering equally spaced sizes from $n = 158$ to $n = 2218$, and (ii) the size of the additional sample (vertical axis), from $m = 1$ to $m = 1000$ in steps of 10. The resulting heatmaps provide an immediate summary of the expected gain from further sampling and can be used

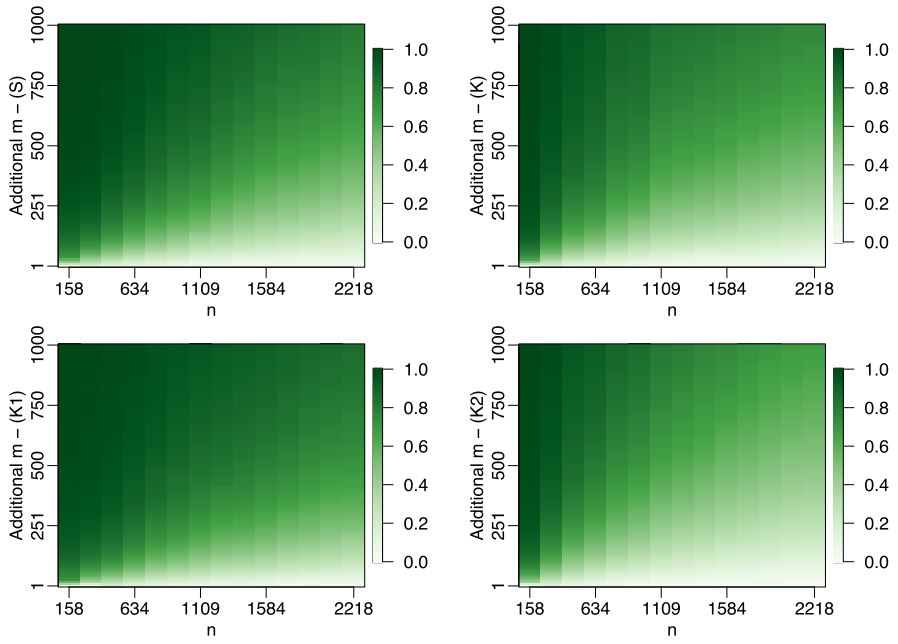


Figure 4: m-steps-ahead prediction probabilities for new shared species (top-left), global distinct species (top-right), and local distinct species (bottom-left and bottom-right). The x-axes display the total number of observations used in the analysis.

in practice to decide whether it is worth investing additional resources or whether the experiment is already sufficiently exhaustive. For the sake of space, we focus only on the *Bayes I* implementation; see Figure 4. The figure confirms that the one-step-ahead analysis is not informative for any practical stopping rule since even at $n = 158$ the discovery probability is already close to zero. As the future sample size increases, the discovery probability grows and eventually approaches one, thereby providing a concrete, interpretable notion of the sampling effort required to achieve a desired chance of observing something new. Moreover, for fixed future effort, discovery probabilities tend to decrease as the sample size n increases, reflecting diminishing returns: once many observations have already been collected, substantially larger additional samples are needed to reach the same probability of discovering new distinct or shared species. Finally, the discovery probabilities for new local distinct species in the second area decrease much more quickly as the observed sample size increases than in the first area, reflecting the marked imbalance between the two sites.

8 Discussion

The increased mathematical complexity has led to limited exploration of the species sampling problem in multiple areas in the literature compared to the case of a single-

area. In this work, we have introduced for the first time a model-based approach within the Bayesian nonparametric setting that relies entirely on closed-form expressions and exact calculations. Specifically, we focused on two primary objectives: predicting the number of distinct and shared species in an additional, unobserved sample of size $m \geq 1$, and estimating the discovery probability of new shared species at m-steps ahead.

The predictive quantities depend on summary statistics of the observed sample and on the three hyperparameters characterizing the Vec-FDP prior, and our experience suggests that the resulting posterior estimates may be sensitive to their values. From a theoretical perspective, we interpreted these parameters by analytically deriving their effects on the correlation, the posterior number of unseen species, and diversity indices. On the practical side, we estimate them from the data using two alternative strategies. Our results show that neither method is uniformly preferable: although *Bayes I* appears to be generally more reliable, *Bayes II* performs better under heavy-tailed data-generating distributions, such as the Zipf setting, and is also substantially faster. As a guideline, we recommend comparing their predictive performance on a held-out validation set and selecting the method that is most appropriate for the data at hand.

We argue that our work provides a foundation for significant future developments. For instance, the exact formulas presented in this work require the computation of the generalised factorial coefficients, whose evaluation scales quadratically with the sample size. We hope that a more in-depth investigation into the asymptotic behaviour of our model may pave the way for suitable approximations that make our solutions more scalable, similar to how Favaro et al. (2009) improved upon Lijoi et al. (2007) in the exchangeable case. In our experience, computing the coefficients V does not constitute an additional computational bottleneck, as they converge rapidly. However, the infinite summation complicates the study of additional properties of our model, such as the expected values of the statistics of interest. It remains an open question whether there exists a distribution q_M such that Equation (7) admits an explicit solution, similarly to Gnedin's model (Gnedin, 2010) in the single-group case. This is not the only open question about these V coefficients. Indeed, in the paper, we began with the model (1), where the P_j's are specified as in Equation (2), to derive the V coefficients, the pEPPF and a system of predictive rules. An alternative approach, similar to Gnedin and Pitman (2006), would focus on the definition of a suitable set of V coefficients that satisfy Equation (S22) to obtain the pEPPF and a system of predictive rules. This strategy would lead to the definition of Gibbs-type priors in a dependent framework, with the model analysed here being a specific example. We wonder if it is possible to identify other tractable examples within this class.

A possible application for our framework is in the multi-armed setting, where the goal is to select the area to sample from in order to maximise the probability of discovering a new global distinct species. A Bayesian nonparametric approach to this problem was first developed in Battiston et al. (2018) under the Hierarchical Pitman-Yor process and subsequently extended to the class of multivariate species sampling models in Franzolini et al. (2025), which includes the Vec-FDP prior. The resulting decision rule can be expressed in terms of a functional of the Vec-FDP posterior, whose explicit form is given in Colombi et al. (2024), including the case of more than two areas ($d > 2$), which is typically the most relevant in these applications.

An open challenge is how to generalise our theoretical results to more than two areas. Notably, the main difficulty is intrinsic to the combinatorial structure of the multiple areas species problem. For $d = 2$, the description of the observed sample can be expressed in terms of the local numbers of distinct species and the global number of distinct species, from which the number of shared species follows deterministically (see, e.g., Equation (11)). However, already for $d = 3$, this relationship no longer holds. Indeed, a joint description of the set of quantities of interest requires additional information beyond the local and global counts, namely the statistics for each pair of areas. As d increases, one must account not only for all pairwise statistics but also for triple-wise statistics, and so on, with the number of required quantities growing exponentially. This rapidly makes the corresponding distributional theory non-scalable in d. A direction for future work is to investigate alternative strategies –beyond the combinatorial marginalization arguments used here– that directly target the global number of distinct and shared species, without requiring integration over the full collection of overlap statistics.

Finally, the full-sharing-of-atoms assumption is crucial for the analytical tractability of the results presented in this work, yet it prevents our model from estimating species richness. Understanding how to balance the mathematical complexity and computational burden arising from relaxing this assumption remains an intriguing direction for future research, which would also enable the estimation of additional ecological indices, including β-diversity (Colwell et al., 1994).

Acknowledgments

This research was largely conducted while Alessandro Colombi was a Ph.D. student at University of Milano-Bicocca, Italy. We would like to thank Giovanni Bacaro (University of Trieste) for supporting with the ants data and Mario Beraha and Tommaso Rigon (University of Milano-Bicocca) for the helpful discussions. We thank the Editorial Board and the two Referees for their valuable comments and suggestions.

Funding

The first and the third authors were supported by the European Union – Next Generation EU funds, component M4C2, investment 1.1., PRIN-PNRR 2022 (P2022H5WZ9). The second author was partially supported by the European Union – Next Generation EU funds, component M4C2, investment 1.1., PRIN-PNRR 2022 (P2022AW4LX).

Supplementary Material

The Supplementary material includes: a detailed description of all in-sample and out-of-sample statistics; a review about generalised factorial coefficients; the proofs of all theoretical results; the Bayesian estimators for diversity indexes; additional details on the simulation study and the real data analysis (DOI: 10.1214/26-BA1617SUPP; .pdf).

References

Argiento, R. and De Iorio, M. (2022). Is infinity that far? A Bayesian nonparametric perspective of finite mixture models. *The Annals of Statistics* 50, 2641–2663. doi:10.1214/22-aos2201. 3, 9, 17

Bacallado, S., Favaro, S., and Trippa, L. (2015). Bayesian nonparametric inference for shared species richness in multiple populations. *Journal of Statistical Planning and Inference* 166, 14–23. doi:10.1016/j.jspi.2014.03.010. 4

Balocchi, C., Camerlenghi, F., and Favaro, S. (2024a). A Bayesian nonparametric approach to species sampling problems with ordering. *Bayesian Analysis* 1, 1–26. doi:10.1214/24-BA1418. 15

Balocchi, C., Favaro, S., and Naulet, Z. (2024b). Bayesian nonparametric inference for species-sampling problems. *Statistical Science*, forthcoming. doi:10.1214/24-STS961. 3

Bassetti, F., Casarin, R., and Rossini, L. (2020). Hierarchical species sampling models. *Bayesian Analysis* 15, 809–838. doi:10.1214/19-BA1168. 4

Battiston, M., Favaro, S., and Teh, Y. W. (2018). Multi-armed bandit for species discovery: A Bayesian nonparametric approach. *Journal of the American Statistical Association* 113, 455–466. doi:10.1080/01621459.2016.1261711. 24

Camerlenghi, F., Favaro, S., Masoero, L., and Broderick, T. (2024). Scaled process priors for Bayesian nonparametric estimation of the unseen genetic variation. *Journal of the American Statistical Association* 119, 320–331. doi:10.1080/01621459.2022.2115918. 14

Camerlenghi, F., Lijoi, A., Orbanz, P., and Prünster, I. (2019). Distribution theory for hierarchical processes. *The Annals of Statistics* 47, 67–92. doi:10.1214/17-AOS1678. 4, 5, 7

Camerlenghi, F., Lijoi, A., and Prünster, I. (2017). Bayesian prediction with multiple-samples information. *Journal of Multivariate Analysis* 156, 18–28. doi:10.1016/j.jmva.2017.01.010. 4

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* 11, 265–270. MR0793175. 2

Chao, A., Chazdon, R. L., Colwell, R. K., and Shen, T.-J. (2006a). Abundance-based similarity indices and their estimation when there are unseen species in samples. *Biometrics* 62, 361–371. doi:10.1111/j.1541-0420.2005.00489.x. 1

Chao, A., Chiu, C.-H., Colwell, R. K., Magnago, L. F. S., Chazdon, R. L., and Gotelli, N. J. (2017). Deciphering the enigma of undetected species, phylogenetic, and functional diversity based on Good-Turing theory. *Ecology* 98, 2914–2929. 3, 14, 17

Chao, A., Hwang, W.-H., Chen, Y.-C., and Kuo, C.-Y. (2000). Estimating the number of shared species in two communities. *Statistica Sinica* 10, 227–246. MR1742110. 4, 17

Chao, A., Shen, T.-J., and Hwang, W.-H. (2006b). Application of Laplace's boundary-mode approximations to estimate species and shared species richness. *Australian & New Zealand Journal of Statistics* 48, 117–128. doi:10.1111/j.1467-842X.2006.00430.x. 4

Chao, A. and Yang, M. C. K. (1993). Stopping rules and estimation for recapture debugging with unequal failure rates. *Biometrika* 80, 193–201. doi:10.1093/biomet/80.1.193. 2

Charalambides, C. A. (2002). *Enumerative combinatorics*. CRC Press. MR1937238. 9, 11

Chuang, C., Shen, T., and Hwang, W. (2015). Estimating the number of shared species by a jackknife procedure. *Environmental and Ecological Statistics* 22, 759–778. doi:10.1007/s10651-015-0318-7. 3, 4

Colombi, A., Argiento, R., Camerlenghi, F., and Paci, L. (2024). Hierarchical mixture of finite mixtures. *Bayesian Analysis* 20, 1–29. doi:10.1214/24-BA1501. 4, 6, 8, 9, 10, 24

Colwell, R. K., Coddington, J. A., and Hawksworth, D. L. (1994). Estimating terrestrial biodiversity through extrapolation. *Philos. Trans. Roy. Soc. London Ser. B* 345, 101–118. 25

Colwell, R. K. et al. (2009). Biodiversity: concepts, patterns, and measurement. *The Princeton guide to ecology* 663, 257–263. 1, 13, 14

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 212–229. 2, 12

Denti, F., Camerlenghi, F., Guindani, M., and Mira, A. (2023). A common atoms model for the Bayesian nonparametric analysis of nested data. *Journal of the American Statistical Association* 118, 405–416. doi:10.1080/01621459.2021.1933499. 6

Efron, B. and Thisted, R. (1976). Estimating the number of unseen species: how many words did Shakespeare know? *Biometrika* 63, 435–447. 2

Favaro, S., Lijoi, A., Mena, R. H., and Prünster, I. (2009). Bayesian non-parametric inference for species variety with a two-parameter Poisson–Dirichlet process prior. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 71, 993–1008. doi:10.1111/j.1467-9868.2009.00717.x. 3, 15, 24

Favaro, S., Lijoi, A., and Prünster, I. (2012). A new estimator of the discovery probability. *Biometrics* 68, 1188–1196. doi:10.1111/j.1541-0420.2012.01793.x. 2

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* 1, 209–230. MR0350949. 2

Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology* 12, 42–58. 2

Franzolini, B., Lijoi, A., Prünster, I., and Rebaudo, G. (2025). Multivariate species sampling models. arXiv:2503.24004. 4, 6, 7, 16, 24

Gnedin, A. (2010). A species sampling model with finitely many types. *Electronic Communications in Probability* 15, 79–88. doi:10.1214/ECP.v15-1532. 3, 24

Gnedin, A. and Pitman, J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences* 138, 5674–5685. doi:10.1007/s10958-006-0335-z. 24

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* 40, 237–264. doi:10.1093/biomet/40.3-4.237. 2, 5

Good, I. J. and Toulmin, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* 43, 45–63. doi:10.1093/biomet/43.1-2.45. 2

Gotelli, N. J. and Colwell, R. K. (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters* 4, 379–391. 3

Lijoi, A., Mena, R. H., and Prünster, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* 94, 769–786. doi:10.1093/biomet/asm061. 2, 9, 10, 15, 24

Lijoi, A., Mena, R. H., and Prünster, I. (2008). A Bayesian nonparametric approach for comparing clustering structures in EST libraries. *Journal of Computational Biology* 15, 1315–1327. doi:10.1089/cmb.2008.0043. 4

Lijoi, A., Nipoti, B., and Prünster, I. (2014). Bayesian inference with dependent normalized completely random measures. *Bernoulli* 20, 1260–1291. doi:10.3150/13-BEJ521. 4, 7

Mao, C. X. (2004). Predicting the conditional probability of discovering a new class. *Journal of the American Statistical Association* 99, 1108–1118. doi:10.1198/016214504000001709. 2

Morisita, M. (1959). Measuring of dispersion of individuals and analysis of the distributional patterns. *Memories of the Faculty of Science, Kyushu University. Series E: Biology*, 215–235. 14

Müller, P., Quintana, F. A., and Rosner, G. (2004). A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66, 735–749. doi:10.1111/j.1467-9868.2004.05564.x. 4, 16

Nayak, T. K. (1988). A note on estimating the number of errors in a system by recapture sampling. *Statistics & Probability Letters* 7, 191–194. doi:10.1016/0167-7152(88)90049-1. 2

Orlitsky, A., Suresh, A. T., and Wu, Y. (2016). Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences of the United States of America* 113, 13283–13288. doi:10.1073/pnas.1607774113. 2

Pan, H.-Y., Chao, A., and Foissner, W. (2009). A nonparametric lower bound for the number of species shared by multiple communities. *Journal of Agricultural, Biological and Environmental Statistics* 14, 452–468. doi:10.1198/jabes.2009.07113. 3, 4

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields* 102, 145–158. doi:10.1007/BF01213386. 2

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. *Statistics, Probability and Game Theory. Papers in honor of David Blackwell* 30, 245–267. doi:10.1214/lnms/1215453576. 7

Quintana, F. A., Müller, P., Jara, A., and MacEachern, S. N. (2022). The dependent Dirichlet process and related models. *Statistical Science* 37, 24–41. doi:10.1214/20-sts819. 4

Rasmussen, S. L. and Starr, N. (1979). Optimal and adaptive stopping in the search for new species. *Journal of the American Statistical Association* 74, 661–667. MR0548262. 2

Simpson, E. (1949). Measurement of diversity. *Nature* 688, 163. 13

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association* 101, 1566–1581. doi:10.1198/016214506000000302. 4, 6, 8

Yue, J. C. and Clayton, M. K. (2012). Sequential sampling in the search for new shared species. *Journal of Statistical Planning and Inference* 142, 1031–1039. doi:10.1016/j.jspi.2011.10.006. 3, 15, 16, 17

Zara, L., Tordoni, E., Castro-Delgado, S., Colla, A., Maccherini, S., Marignani, M., Panepinto, F., Trittoni, M., and Bacaro, G. (2021). Cross-taxon relationships in Mediterranean urban ecosystem: A case study from the city of Trieste. *Ecological Indicators* 125, 107538. 20

Zito, A., Rigon, T., Ovaskainen, O., and Dunson, D. B. (2023). Bayesian modeling of sequential discoveries. *Journal of the American Statistical Association* 118, 2521–2532. doi:10.1080/01621459.2022.2060835. 3