The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1496–1516
DOI: 10.1214/16-AOAS944

BAYESIAN NONPARAMETRIC DEPENDENT MODEL FOR PARTIALLY REPLICATED DATA: THE INFLUENCE OF FUEL SPILLS ON SPECIES DIVERSITY[1]

BY JULYAN ARBEL[2,*], KERRIE MENGERSEN AND JUDITH ROUSSEAU

Collegio Carlo Alberto and Bocconi University, Queensland University of Technology and Université Paris-Dauphine

We introduce a dependent Bayesian nonparametric model for the probabilistic modeling of membership of subgroups in a community based on partially replicated data. The focus here is on species-by-site data, that is, community data where observations at different sites are classified in distinct species. Our aim is to study the impact of additional covariates, for instance, environmental variables, on the data structure, and in particular on the community diversity. To this end, we introduce dependence a priori across the covariates and show that it improves posterior inference. We use a dependent version of the Griffiths–Engen–McCloskey distribution defined via the stick-breaking construction. This distribution is obtained by transforming a Gaussian process whose covariance function controls the desired dependence. The resulting posterior distribution is sampled by Markov chain Monte Carlo. We illustrate the application of our model to a soil microbial data set acquired across a hydrocarbon contamination gradient at the site of a fuel spill in Antarctica. This method allows for inference on a number of quantities of interest in ecotoxicology, such as diversity or effective concentrations, and is broadly applicable to the general problem of community response to environmental variables.

1. Introduction. This paper was motivated by the ecotoxicological problem of studying communities, or groups of species, observed as counts of species at a set of sites, where the composition and distribution of species may differ among sites, and for which the sites are indexed by a contaminant. More specifically, the soil microbial data set we are focusing on in this paper was acquired at different sites of a fuel spill region in Antarctica. Although there is now much greater awareness of human impacts on the Antarctic, substantial challenges remain. One

Received July 2015; revised February 2016.

[1]Cross-references with the prefix "S" contained in this article refer to the online supplementary material [Arbel, Mengersen and Rousseau (2016)].

[2]Supported by the European Research Council (ERC) through StG "N-BNP" 306406, by ANR BANDHITS and by the Australian Research Council. Department of Decision Sciences, BIDSA and IGIER, Bocconi University.

*Also at CREST, Paris, at the beginning of this project.

Key words and phrases. Bayesian nonparametrics, covariate-dependent model, Gaussian processes, Griffiths–Engen–McCloskey distribution, partially replicated data, stick-breaking representation.

of these is the containment of historic buried station waste, chemical dumps and fuel spills. These wastes do not break down in such extreme environments and their spread is exacerbated by melting ice in summer. Developing effective containment strategies requires an understanding of the impact of these incursions on the natural environment. The data set considered here consists of soil microbial counts of operational taxonomic units, OTUs, as well as a site contaminant level measured by the total petroleum hydrocarbon, TPH. Thus, the aim is to model the probabilities of occurrence associated with the species at the different sites and to be able to interpret the impact of the contaminant on the community as a whole or on a particular species.

This specific case study gives rise to a more general problem that can be described as modeling the probability of membership of subgroups of a community based on partially replicated data obtained by observing different subsets of the subgroups at different levels of a covariate. The problem can also be considered as the analysis of compositional data in which the data points represent so-called compositions, or proportions, that sum to one. A typical example is the chemical composition of rock specimens in the form of percentages of a prespecified number of elements [see, e.g., Aitchison (1982), Barrientos, Jara and Quintana (2015)]. This problem is endemic in many fields, for example, meat composition in biology [Alston, Mengersen and Gardner (2011)], consumer demand in economics [Pawlowsky-Glahn and Buccianti (2011)], food composition in nutrition [van den Boogaart and Tolosana-Delgado (2013)], genotype frequency in genetics [George, Mengersen and Davis (2000)], bacterial composition in microbiomics [Li (2015)] and so on. Despite this, the solution to that problem remains a challenge [Aitchison (1986, 1994), Lovell et al. (2015)]. Common approaches are typically based on parametric assumptions and require prespecification of the number of subgroups (e.g., species) in the community. In this paper, we suggest an alternative that overcomes this drawback. The method is described in terms of species for reasons of intuitiveness in description, but the approach is generally applicable far beyond the species sampling framework.

We propose a Bayesian nonparametric approach to both the specific and general problems described above using a covariate dependent random probability measure as a prior distribution. Dependent extensions of random probability measures, with respect to a covariate such as time or position, have been extensively studied recently under three broad constructions. First, a class of solutions is based on the Chinese Restaurant process; see, for instance, Caron, Davy and Doucet (2007) and Johnson et al. (2013). These are oriented toward in-line data collection and fast implementation. Second, some approaches use completely random measures; see, for example, Lijoi, Nipoti and Prünster (2014a, 2014b). An appealing feature of this approach is analytical tractability which allows for more elaborate studying of the distributional properties of the measures. Third, many strategies make use of the stick-breaking representation based on the line of research pioneered by MacEachern (1999, 2000), which defines dependent Dirichlet processes. There are

many variants, including those suggested by Chung and Dunson (2011), Dunson and Park (2008), Dunson, Pillai and Park (2007), Griffin and Steel (2006, 2011), among others. The success of the stick-breaking constructions stems from their attractiveness from a computational point of view as well as their great flexibility in terms of full support, which we prove for our model in Section S.3.2 of the Supplementary Material. This is the approach that we follow here.

We define a dependent version of the Griffiths–Engen–McCloskey distribution (hereafter denoted GEM), which is the distribution of the weights in a Dirichlet process, for modeling presence probabilities. Dependence is introduced via the covariance function of a Gaussian process, which allows dependent Beta random variables to be defined by inverse cumulative distribution functions transforms. The main appeal of introducing dependence in the model is to allow predictions to be made at any value of the covariate.

The resulting model is not confined to the estimation of diversity indices, but could also utilize the predictive structure yielded by specific discrete nonparametric priors to address issues such as the estimation of the number of new species (subgroups) to be recorded from further sampling, the probability of observing a new species at the $(n + m + 1)$th draw conditional on the first n observations, or of observing rare species, defined as occurring with frequency less than a certain threshold [see, e.g., Favaro, Lijoi and Prünster (2012), Lijoi, Mena and Prünster (2007)].

The paper is organized as follows. In Section 2 we describe our case study, review the ecotoxicological literature and background, and discuss diversity and effective concentration estimation. Section 3 describes the Bayesian nonparametric model, posterior sampling and most useful properties of the model. Estimation results and ecotoxicological guidelines are given in Section 4. A discussion on model considerations is given in Section 5, and Section 6 concludes this paper with a general discussion. Extended results, details of posterior computation and the proofs of our results are available in the Supplementary Material available as Arbel, Mengersen and Rousseau (2016).

2. Case study and ecotoxicological context.

2.1. *Case study and data.* As already sketched in the Introduction, our case study consists of a soil microbial data set acquired across a hydrocarbon contamination gradient at the location of a fuel spill at Australia's Casey Station in East Antarctica (110°32′E, 66°17′S) along a transect at 22 locations. Microbes are classified as Operational Taxonomic Units (OTU), that we also generically refer to as species throughout the paper. OTU sequencing was processed on genomic DNA using the mothur software package; see Schloss et al. (2009). We refer to Snape et al. (2015) for a complete account of the data set acquisition. The total number of species recorded at least once at one site is 1, 800+. All species were included in the analysis and estimation. However, we have noticed that it is possible to work

with a subset of the data, consisting of those species with abundance over all measurements exceeding a given low threshold (in our case up to ten), without altering significantly the results. A crucial point for the subsequent analyses is that we order the species by *decreasing overall abundance* (i.e., species $j = 1$ is the most numerous species in the whole data set). The variations of sampling across the sites explain why the species are not strictly ordered when considered site by site; see Figure 1.

OTU measurements are paired with a contaminant called Total Petroleum Hydrocarbon [TPH, see Siciliano et al. (2014)], suspected to impact OTU diversity. The contamination TPH level recorded at each site ranges from 0 to 22,000 mg TPH/kg soil. Ten sites were actually recorded as uncontaminated with TPH equal to zero. We call the microbial communities associated to these sites *baseline communities*, and use them to define effective concentrations EC_x; see Section 2.4. Although a continuous variable, TPH is recorded with ties that we interpret as due to measurement rounding. Our methodology cannot handle ties unless they are genuine ties which could be collapsed together into one single site. Such a collapsing would not be biologically meaningful, hence, we jitter TPH concentrations with a random Gaussian noise (absolute value for the case TPH $= 0$) to account for measurement errors and to discriminate the ties. This noise can be incorporated in the probabilistic model. A sensitivity assessment of the impact of the variance of the noise, using a range of moderate values compared with the variability of TPH, showed little to no impact on the estimates of interest. An alternative to allow for multiple observations at the same level of covariate can be achieved by adding a small amount to the diagonal of the Gaussian process correlation matrix, known as the nugget term [see, e.g., Andrianakis and Challenor (2012), Cressie (1993)]. This is left for future investigation.

2.2. *Ecotoxicological context*. This paper focuses on an ecotoxicological case study where the goal is to predict the impact of a contaminant on an ecosystem. The common treatment of this question relies on toxicity tests, either on single species (called populations) or on multiple species (called communities). The need for appropriate modeling techniques is apparent due to data limitations, for instance, in our case where data acquisition in Antarctica is extremely expensive. Although single species modeling methods are now well understood, many community modeling methods are less strongly theoretically endorsed. There are two alternative community modeling approaches. On the one hand, one can model single species independently and then aggregate the individual predictions into community predictions [e.g., Ellis, Smith and Pitcher (2011)]. A drawback attached to the aggregation is the lack of appropriate uncertainty of the method. Moreover, one necessarily loses crucial information by dismissing interplays across species. On the other hand, the response of the community as a whole is modeled, which generally entails the use of some univariate summaries of community responses, such as compositional dissimilarity [e.g., Ferrier and Guisan (2006), Ferrier et al. (2007)]

or rank abundance distributions [Foster and Dunstan (2010)]. Alternatively, the responses of multiple species can be modeled simultaneously [e.g., Dunstan, Foster and Darnell (2011), Foster and Dunstan (2010), Wang et al. (2012)].

Single species are commonly modeled through the probability of presence p_j of each species j as a function of the environmental parameters. The natural distribution for multiple species is the multinomial distribution, which provides an intuitive framework when the sampling process consists of independent observations of a fixed number of species. Recent literature demonstrates the popularity of the multinomial distribution in ecology [e.g., De'ath (2012), Fordyce et al. (2011), Holmes, Harris and Quince (2012)] and genomics [Bohlin, Skjerve and Ussery (2009), Dunson and Xing (2009)]. Our use of the GEM distribution actually extends the multinomial distribution to cases where the number of species does not need to be fixed or known, that is, where the prior is on infinite vectors of presence probabilities.

2.3. *Diversity.* Modeling presence probabilities provides a clear link to indices that describe various community properties of interest to ecologists, such as species diversity, richness, evenness, etc. The literature on diversity is extensive, not only in ecology [Colwell et al. (2012), De'ath (2012), Foster and Dunstan (2010), Hill (1973), Patil and Taillie (1982)] but also in other areas of science, such as biology, engineering, physics, chemistry, economics, health and medicine [see Borges and Roditi (1998), Havrda and Charvát (1967), Kaniadakis, Lissia and Scarfone (2005), and in more mathematical fields such as probability theory [Donnelly and Grimmett (1993)]. There are numerous ways to study the diversity of a population divided into groups. Examples of predominant indices in ecology include the Shannon index $-\sum_j p_j \log p_j$, the Simpson index (or Gini index) $1-\sum_j p_j^2$, on which we focus in this paper, and the Good index which generalizes both $-\sum_j p_j^\alpha \log^\beta p_j$, $\alpha, \beta \geq 0$ [Good (1953)].

Diversity estimation, and, more generally, estimation of community indices based on species data, has been a statistical problem of interest for a long time. One of the main stumbling blocks is the high variability in species data. For instance, the most obvious estimators, hereafter referred to as *empirical estimators*, which involve plugging in empirical presence probabilities (i.e., observed proportions $\hat{p}_{ij}$ of species j at site i), suffer from that curse. Many treatments have been proposed in the literature to account for this issue. A first approach is the field of occupancy modeling and imperfect detection; see, for instance, the monograph Royle and Dorazio (2008). We provide a concise description of imperfect detection modeling in Section 5.1 and do not pursue this direction here.

Another approach, that we are following in this paper, involves smoothing or regularizing empirical estimates. A Bayesian approach is a natural way to do this. Specifically, Gill and Joanes (1979) show that using a Dirichlet prior distribution over $(p_1, \ldots, p_J)$ in the multinomial model with J species greatly improves

estimation over empirical counterparts. The reason for this is that using a prior prevents pathological behavior due to outliers by smoothing the estimates. The smoothing is controlled by the Dirichlet parameter which can be chosen based on expert information, related literature or other relevant information sources. Our case study exhibits an additional variability in the across-sites dimension compared to the framework of Gill and Joanes (1979). This variability is instantiated by the plot of empirical estimates of Simpson diversity in Figure 4. However, we leverage this additional difficulty by borrowing strength across the sites, following the intuition that neighboring sites should respond similarly to the contaminant. The borrowing of strength is achieved by incorporating dependence across the sites in the prior distribution. Noting that the total number of species is not known a priori, we adopt a Bayesian nonparametric approach. More specifically, we extend the work by Gill and Joanes (1979) from a Dirichlet prior to a covariate-dependent Dirichlet process prior. This is also extending the model of Holmes, Harris and Quince (2012) to a covariate-dependent setting with a priori unknown number of species.

Interestingly, this idea of using a Bayesian nonparametric approach as a smoothing technique for species data was recently adopted in the context of discovery probability, the probability of observing new species or species already observed with a given frequency. Good (1953) proposed smoothed estimators popularized as Good–Turing estimators for discovery probabilities. Good–Turing estimators were shown to have a Bayesian nonparametric interpretation [see Arbel et al. (2016), Favaro, Nipoti and Teh (2016), Lijoi, Mena and Prünster (2007)], thus demonstrating the ability of Bayesian nonparametric methods to regularize species data.

2.4. *Effective concentration.* Highly relevant in terms of protecting an ecosystem, the *effective concentration* at level x, denoted by EC_x, is the concentration of contaminant that causes $x\%$ effect on the population relative to the baseline community [e.g., Newman (2012)]. For example, the EC_{50} is the median effective concentration and represents the concentration of a contaminant which induces a response halfway between the control baseline and the maximum after a specified exposure time. For single species studies, this is commonly assessed by an $x\%$ increase in mortality. In applications with a multi-species response as in this paper, it is the response of the community as a whole that is of interest. The EC_x values are used to derive appropriate protective guidelines on contaminant concentrations, for instance, in terms of waste, chemical dumps and fuel spills containment strategies. Currently, it is not clear how to best calculate EC_x values using whole-community data. The EC_x values can be defined in many ways depending on the specific aspects of interest to the ecological application. We illustrate the use of the Jaccard dissimilarity index, denoted by $\mathrm{Jac}(X)$ (where X here denotes TPH and the Jaccard dissimilarity is seen as a function of TPH), one of the many dissimilarity variants available, as a measure of change in community composition. We defined the baseline community as the set of uncontaminated sites (ten sites) where TPH

equals zero; see Section 2.1. The dissimilarity at TPH zero, denoted by Jac_0, is an estimate of the variability in community composition between uncontaminated sites. The EC_x value is obtained from $\mathrm{Jac}(X)$ values as the smallest TPH value X such that

$$\mathrm{Jac}(X) = 1 - (1 - \mathrm{Jac}_0)(1 - x/100). \tag{1}$$

In this way, $\mathrm{EC}_0 = 0$, the TPH value for which there is no change relative to baseline, is obtained at $\mathrm{Jac}(X) = \mathrm{Jac}_0$, while EC_{100} is obtained at $\mathrm{Jac}(X) = 1$ for a TPH value such that the community composition becomes disjoint with the baseline. We see by equation (1) that intermediate values are obtained by linear interpolation. The smallest TPH value is used so as to provide a conservative EC_x estimate, since the dissimilarity curve is not guaranteed to be monotonic. A particular feature of the model which allows us to follow this methodology is its ability to estimate the community composition between observed TPH values, since it is unlikely that the dissimilarity threshold $\mathrm{Jac}(X)$ sought in equation (1) will coincide exactly with one of the measured TPH levels in the data. 95% credible bands for EC_x values were obtained in a similar fashion, as the smallest and the largest values of, respectively, the 2.5% and 97.5% quantiles of the EC_x value, again so as to provide conservative estimates. Note that both quantities $\mathrm{Jac}(X)$ and EC_x are estimated from the posterior samples obtained in Section 3.3; see also Figure 5(a) for an illustration of the method.

3. Model.

3.1. *Data model.* We describe here the notation and the sampling process of covariate-dependent species-by-site count data. To each site $i = 1, \dots, I$ corresponds a covariate value $X_i \in \mathcal{X}$, where the space $\mathcal{X}$ is a subset of $\mathbb{R}^d$. We focus here on a single covariate ($d = 1$). The general case ($d \geq 1$) is discussed in Section 6. Individual observations $Y_{n,i}$ at site i are indexed by $n = 1, \dots, N_i$, where N_i denotes the total abundance or number of observations. Observations $Y_{n,i}$ take on positive natural number values $j \in \{1, \dots, J_i\}$, where J_i denotes the number of distinct species observed at site i. No hypothesis is made about the unknown total number of species $J = \max_i J_i$ in the community of interest, which might be infinite. We denote by $(\mathbf{X}, \mathbf{Y})$ the observations over all sites, where $\mathbf{X} = (X_i)_{i=1,\dots,I}$, $\mathbf{Y} = (\mathbf{Y}_i^{N_i})_{i=1,\dots,I}$ and $\mathbf{Y}_i^{N_i} = (Y_{n,i})_{n=1,\dots,N_i}$. The abundance of species j at site i, that is, the number of times that $Y_{n,i} = j$ with respect to index n, is denoted by N_{ij} and satisfies $\sum_{j=1}^{J_i} N_{ij} = N_i$.

We model the probabilities of presence $\mathbf{p} = (\mathbf{p}(X_i))_{i=1,\dots,I}$, where $\mathbf{p}(X_i)$ represents the vector of probabilities $p_j(X_i)$ of species j under covariate X_i, by the following:

$$Y_{n,i} | \mathbf{p}(X_i), X_i \overset{\text{ind}}{\sim} \sum_{j=1}^{\infty} p_j(X_i) \delta_j, \tag{2}$$

for $i = 1, \dots, I$, $n = 1, \dots, N_i$, where δ_j denotes a Dirac point mass at j.

3.2. *Dependent prior distribution.* We follow a Bayesian approach which implies that we need to define a prior distribution for the probabilities $\mathbf{p}$. The Dirichlet process [Ferguson (1973)] is a popular distribution in Bayesian nonparametrics which has been used for modeling species data by Lijoi, Mena and Prünster (2007). We extend the methodology developed by Lijoi et al. in building a covariate-dependent prior distribution in a way which is reminiscent of the extension of the classical Dirichlet process to the dependent Dirichlet process by MacEachern (1999). More specifically, the marginal prior distribution on $\mathbf{p}(X)$ for covariate X is defined by the following stick-breaking construction, which introduces Beta random variables $V_j(X) \overset{\text{iid}}{\sim} \text{Beta}(1, M)$ such that $p_1(X) = V_1(X)$ and, for $j > 1$,

$$p_j(X) = V_j(X) \prod_{l<j} (1 - V_l(X)). \tag{3}$$

This prior distribution is called the Griffiths–Engen–McCloskey distribution and denoted by $\mathbf{p}(X) \sim \text{GEM}(M)$, where $M > 0$ is called the precision parameter. The motivation for using the GEM distribution is explained by Figure 1 which shows, for species $j = 1, \ldots, 32$, the observed proportions $(\hat{p}_{ij})$ at site $i = 9$ and draws of (p_j) from the GEM(M) prior with precision parameter $M = 6$. Since the GEM(M) prior on $\mathbf{p}(X_i)$ is *stochastically ordered* [see Pitman (2006)], it puts more mass on the more numerous species of the community. It makes sense to sort the data by decreasing overall abundance, as explained in Section 2.1, and to use a prior with a stochastic order on $\mathbf{p}$ since the data under study are naturally present in a large and small number of species. In Figure 1 we observe the same nonincreasing pattern between the observed frequencies and draws from the GEM prior, which is an argument in favor of the use of the GEM(M) prior for marginal modeling of the probabilities $\mathbf{p}(X)$. For a discussion on the ordering assumption, see Section 5.2.

For an exhaustive description of the prior distribution on $\mathbf{p}$, the marginal description (3) needs to be complemented by specifying a distribution for stochastic processes $(V_j(X), X \in \mathcal{X})$ for any positive integer j. Since (3) requires Beta

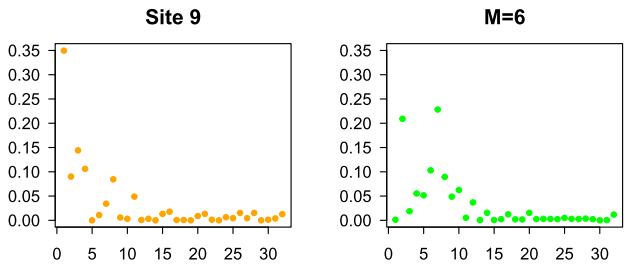


FIG. 1. *Comparison of probabilities of presence in raw data at site $i = 9$ (left) and probabilities sampled from the Griffiths–Engen–McCloskey prior with $M = 6$ (right). The x-axis represents species $j = 1, \ldots, 32$.*

marginals, natural candidates are Beta processes. A simple yet effective construct to obtain a Beta process is to transform a Gaussian process by the inverse cumulative distribution function (CDF) transform as follows. Denote by $Z \sim \mathrm{N}(0, \sigma_Z^2)$ a Gaussian random variable, by Φ_{σ_Z} its CDF and by F_M a Beta$(1, M)$ CDF. Then $V = F_M^{-1} \circ \Phi_{\sigma_Z}(Z)$ is Beta$(1, M)$ distributed, with $F_M^{-1}(U) = 1 - (1 - U)^{1/M}$. Denote by $g_{\sigma_Z, M} = F_M^{-1} \circ \Phi_{\sigma_Z}$. Note that the idea of including a transformed Gaussian process within a stick-breaking process is used in previous articles, including Barrientos, Jara and Quintana (2012), Pati, Dunson and Tokdar (2013), Rodríguez and Dunson (2011), Rodríguez, Dunson and Gelfand (2010).

In our case, we use Gaussian processes $\mathcal{Z}_j$ on the space $\mathcal{X}$, $j = 1, 2, \ldots$, which define Beta processes $\mathbf{V}_j$, which in turn define the probabilities $\mathbf{p}_j$. Although the main parameters of interest are the $\mathbf{p}_j$, we will work hereafter with $\mathcal{Z}_j$ for computational convenience.

The Gaussian process is used as a prior probability distribution over functions. It is fully specified by a mean function m, which we take equal to 0, and a covariance function K defined by

$$K(X_i, X_l) = \mathrm{Cov}\big(\mathcal{Z}_j(X_i), \mathcal{Z}_j(X_l)\big). \tag{4}$$

We control the overall variance of $\mathcal{Z}_j$ by a positive pre-factor $\sigma_{\mathbf{Z}}^2$ and write $K = \sigma_{\mathbf{Z}}^2 \tilde{K}$, where $\tilde{K}$ is normalized in the sense that $\tilde{K}(X_i, X_i) = 1$ for all i. We work with the squared exponential (SE), Ornstein–Uhlenbeck (OU) and rational quadratic (RQ) covariance functions; see Section S.2 in the Supplementary Material for more details. All three involve a parameter λ called the length-scale of the process $\mathcal{Z}_j$. This parameter tunes how far apart two points X_1 and X_2 have to be for the process to change significantly. The shorter λ is, the rougher are the paths of the process $\mathcal{Z}_j$. We adopt the same technique as van der Vaart and van Zanten (2009) who deal with λ by making it random with an inverse-Gamma (denoted IG) prior distribution. They obtain adaptive minimax-optimal posterior contraction rates which indicate that the length-scale parameter λ correctly adapts to the path smoothness. Gibbs (1997) derived a covariance function where the length-scale $\lambda(X)$ is a (positive) function of X. This case is not studied here, although it could result in interesting behavior, as noted in Rasmussen and Williams (2006). Each species j is associated to a Gaussian process $\mathcal{Z}_j$. We have a set of I points $\mathbf{X} = (X_1, \ldots, X_I)$ in the covariate space $\mathcal{X}$ which reduces the evaluation of the whole process $\mathcal{Z}_j$ to its values at $\mathbf{X}$ denoted by $\mathbf{Z}_j = (Z_{1,j}, \ldots, Z_{I,j}) = (\mathcal{Z}_j(X_1), \ldots, \mathcal{Z}_j(X_I))$. We denote also by $\mathbf{Z}$ the matrix of all vectors $\mathbf{Z}_j$, $\mathbf{Z} = (Z_{ij})_{1 \le i \le I, 1 \le j \le J}$. The vector $\mathbf{Z}_j$ is multivariate Gaussian. Its covariance matrix $K(\mathbf{X}, \lambda, \sigma_{\mathbf{Z}}) = (\sigma_{\mathbf{Z}}^2 \tilde{K}_\lambda(X_i, X_l))_{i,l=1,\ldots,I}$ is a Gram matrix with entries given by equation (4). The prior distribution of $\mathbf{Z}_j$ is

$$\log \pi(\mathbf{Z}_j | \mathbf{X}, \lambda, \sigma_{\mathbf{Z}}) = \frac{1}{2} \mathbf{Z}_j^\top K^{-1}(\mathbf{X}, \lambda, \sigma_{\mathbf{Z}}) \mathbf{Z}_j - \frac{1}{2} \log \big| K(\mathbf{X}, \lambda, \sigma_{\mathbf{Z}}) \big| - \frac{I}{2} \log 2\pi$$

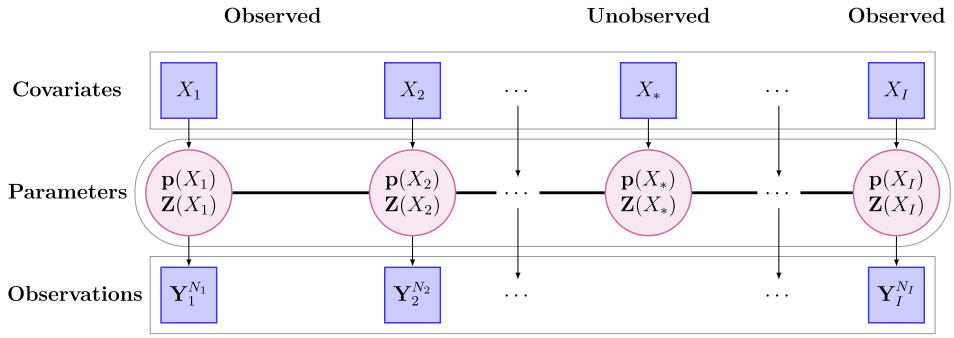


FIG. 2. *Diagrammatic representation for the Dep-GEM model. Squares represent observed data: covariates* $\mathbf{X} = (X_i)_{i=1,\dots,I}$ *and observations* $\mathbf{Y}_i^{N_i} = (Y_{1,i}, \dots, Y_{N_i,i})$. *Circles represent parameters for the Dep-GEM model.*

or, written in terms of $\sigma_{\mathbf{Z}}^2$ and $\tilde{K}_\lambda = (\tilde{K}_\lambda(X_i, X_l))_{i,l=1,\dots,I}$,

$$\pi(\mathbf{Z}_j|\mathbf{X}, \lambda, \sigma_{\mathbf{Z}}) \propto \sigma_{\mathbf{Z}}^{-I} |\tilde{K}_\lambda|^{-1/2} \exp\left(-\frac{\mathbf{Z}_j^\top \tilde{K}_\lambda^{-1} \mathbf{Z}_j}{2\sigma_{\mathbf{Z}}^2}\right).$$

The prior distribution is complemented by specifying the distributions over hyperparameters $\sigma_{\mathbf{Z}}$ the standard deviation, λ the length scale and M the precision parameter of the GEM distribution. We use the following standard hyperpriors:

$$\sigma_{\mathbf{Z}}^2 \sim \mathrm{IG}(a_{\mathbf{Z}}, b_{\mathbf{Z}}), \qquad \lambda \sim \mathrm{IG}(a_\lambda, b_\lambda) \quad \text{and} \quad M \sim \mathrm{Ga}(a_M, b_M). \tag{5}$$

Note that these are also common choices in the absence of dependence since they are conjugate priors, and recall that the inverse-Gamma for λ also proves to lead to good convergence results.

It is convenient to estimate the model in terms of $\mathbf{Z}_j$, and then to use the transform $\mathbf{V}_j = g_{\sigma_{\mathbf{Z}},M}(\mathbf{Z}_j)$. The likelihood is

$$\mathcal{L}(\mathbf{Y}|\mathbf{Z}, \mathbf{X}, \sigma_{\mathbf{Z}}, M) = \prod_{j=1}^{J} \prod_{i=1}^{I} g_{\sigma_{\mathbf{Z}},M}\big(Z_j(X_i)\big)^{N_{ij}} \big(1 - g_{\sigma_{\mathbf{Z}},M}\big(Z_j(X_i)\big)\big)^{\bar{N}_{i,j+1}}, \tag{6}$$

where $\bar{N}_{i,j+1} = \sum_{l>j} N_{il}$. The posterior distribution is then

$$\pi(\mathbf{Z}, \lambda, \sigma_{\mathbf{Z}}, M|\mathbf{Y}, \mathbf{X}) \propto \mathcal{L}(\mathbf{Y}|\mathbf{Z}, \mathbf{X}, \sigma_{\mathbf{Z}}, M)\pi(\mathbf{Z}|\mathbf{X}, \lambda, \sigma_{\mathbf{Z}})\pi(\sigma_{\mathbf{Z}})\pi(\lambda)\pi(M). \tag{7}$$

The model is illustrated by a graphical representation in Figure 2.

3.3. *Posterior computation and inference.* Here we highlight the main points of interest of the algorithm, which is fairly standard; the fully detailed posterior sampling procedure can be found in the Supplementary Material, Section S.1. Inference in the Dep-GEM model is performed via two distinct samplers: (i) first a Markov chain Monte Carlo (hereafter MCMC) algorithm comprising Gibbs and

Metropolis–Hastings steps for sampling the posterior distribution of $(\mathbf{Z}, \sigma_{\mathbf{Z}}, \lambda, M)$. It proceeds by sequentially updating each parameter $\mathbf{Z}, \sigma_{\mathbf{Z}}, \lambda$ and M via its conditional distribution; (ii) second a sampler from the posterior predictive distribution of $\mathbf{Z}_*$, which consists of posterior conditional sampling of the Gaussian process $\mathcal{Z}$ at covariates $\mathbf{X}_* = (X_1^*, \ldots, X_{I_*}^*)$ which are not observed (i.e., such that $\{X_1, \ldots, X_I\}$ and $\{X_1^*, \ldots, X_{I_*}^*\}$ are pairwise distinct). This is achieved by integrating out $\mathbf{Z}$ in the conditional distribution of $\mathbf{Z}_*$ given $\mathbf{Z}$ according to the posterior distribution sampled in (i).

3.4. *Distributional properties*. We provide in Proposition 1 the first prior moments, expectation, variance and covariance, of the diversity. This is of crucial importance for eliciting the values of hyperparameters, or their prior distribution, based on prior information (expert, etc.) Additionally, since the Dep-GEM introduces some dependence across the $p_j(X_i)$ in varying X_i, the question of the dependence induced in a diversity index arises. Denote the Simpson index by $H_{\mathrm{Simp}}(X_i)$; see Section 2.3. An answer is formulated in the next proposition in terms of the covariance between $H_{\mathrm{Simp}}(X_1)$ and $H_{\mathrm{Simp}}(X_2)$. Further notable properties are presented in the Supplementary Material Section S.3, including marginal moments of the Dep-GEM prior and continuity of sample paths in Proposition S.3.1, full support in Proposition S.3.3, a study of the joint distribution of samples from the Dep-GEM prior in Proposition S.3.4, and a discussion on the joint exchangeable partition probability function based on size-biased permutations in Section S.3.4.

PROPOSITION 1. *The expectation and variance of the Simpson diversity, and its covariance at two sites X_1 and X_2, induced by the Dep-GEM distribution, are as follows*:

$$E(H_{\mathrm{Simp}}) = \frac{M}{1+M}, \qquad \mathrm{Var}\big(H_{\mathrm{Simp}}(X)\big) \frac{2M}{(M+1)(M+1)_3}, \tag{8}$$

$$\mathrm{Cov}\big(H_{\mathrm{Simp}}(X_1), H_{\mathrm{Simp}}(X_2)\big) = \frac{\nu_{2,2}(1-\omega_{2,0}) + 2\nu_{2,0}\gamma_{2,2}}{(1-\omega_{2,0})(1-\omega_{2,2})} - \nu_{1,0}^2, \tag{9}$$

where $\nu_{i,j} = E[V^i(X_1)V^j(X_2)]$, $\omega_{i,j} = E[(1-V(X_1))^i(1-V(X_2))^j]$ *and* $\gamma_{i,j} = E[V^i(X_1)(1-V(X_2))^j]$.

The values of $\nu_{i,j}, \omega_{i,j}, \gamma_{i,j}$ cannot be computed in a closed-form expression when $i \times j \neq 0$, but they can be approximated numerically. The same formal computations for the Shannon index lead to somehow more complex expressions which are not displayed here [see also Cerquetti (2014)]. The expressions of Proposition 1 are illustrated in Figure 3.

The precision parameter M has the following impact on the prior distribution and on the diversity: when $M \to 0$, the prior degenerates to a single species with

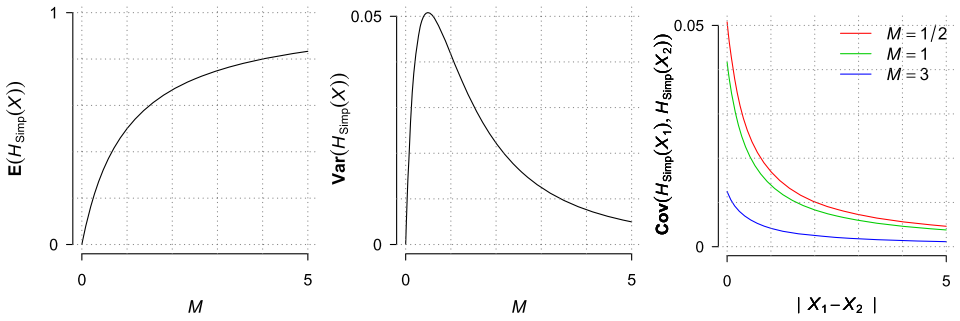


FIG. 3. *Illustration of Proposition 1.* Left: $E(H_{\text{Simp}}(X))$ *w.r.t.* M. Middle: $\text{Var}(H_{\text{Simp}}(X))$ *w.r.t.* M. Right: *three paths of* $\text{Cov}(H_{\text{Simp}}(X_1), H_{\text{Simp}}(X_2))$ *w.r.t.* $|X_1 - X_2|$ *for* $M \in \{1/2, 1, 3\}$.

probability 1, hence, $H_{\text{Simp}} \to 0$, whereas when $M \to \infty$, the prior tends to favor infinitely many species, and $H_{\text{Simp}} \to 1$. In both cases, the variance and the covariance vanish. In between, the variance is maximum for $M \approx 0.49$. The covariance at X_1 and X_2 equals the variance when $X_1 = X_2$ (by continuity of the sample paths), while the covariance vanishes when $|X_1 - X_2| \to \infty$ (which corresponds to independence for infinitely distant covariates).

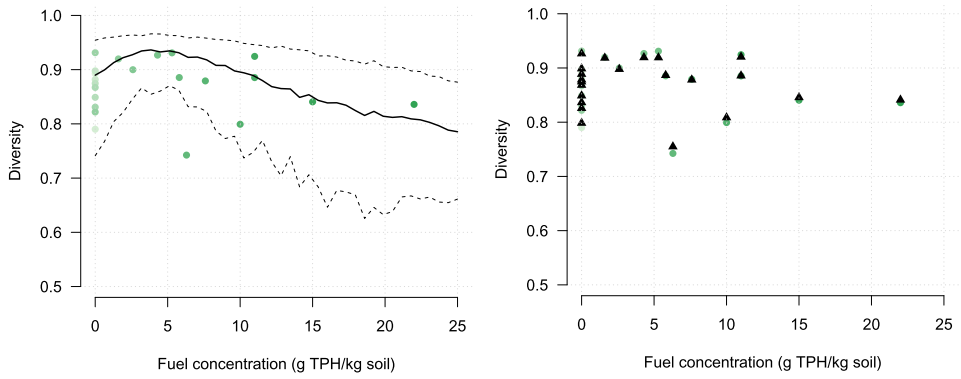
Despite the fact that the first moments of the diversity indices under a GEM prior can be derived, a full description of the distribution seems hard to achieve. For instance, the distribution of the Simpson index involves the small-ball-like probabilities $P(\sum_j p_j^2 < a)$ for which, to the best of our knowledge, no result is known under the GEM distribution.

4. Case study results. We now apply the model to the estimation of diversity and of effective concentrations EC_x as described in Section 2, and assess the goodness of fit of the model and its sensitivity to sampling variation.

4.1. *Results.* The MCMC algorithm is run with squared exponential Gaussian processes for 50,000 iterations thinned by a factor of 5 with a burn-in of 10,000 iterations. The parameters of the hyperpriors (5) are $a_{\mathbf{Z}} = b_{\mathbf{Z}} = 1$, $\eta_\lambda = 1$, $a_\lambda = b_\lambda = 1$ and $a_M = b_M = 1$. The efficiency and convergence of the MCMC sampler were assessed by trace plots and autocorrelations of the parameters.

The results for the Simpson diversity estimation are illustrated in Figure 4 for the Dep-GEM model [left, 4(a)] and for the independent GEM model [right, 4(b)]. The horizontal axis represents the pollution level TPH and the vertical axis represents the Simpson diversity. The posterior mean of the diversity is represented by the solid line, and a 95% credible interval is indicated by dashed lines, for the dependent model only. The dots indicate the empirical estimator of the diversity.

The Dep-GEM model [Figure 4(a)] suggested that diversity first increases with TPH with a maximum at 4000 mg TPH/kg soil, and then decreases with TPH. Such a variation may depict a feature of practical interest for biologists known as



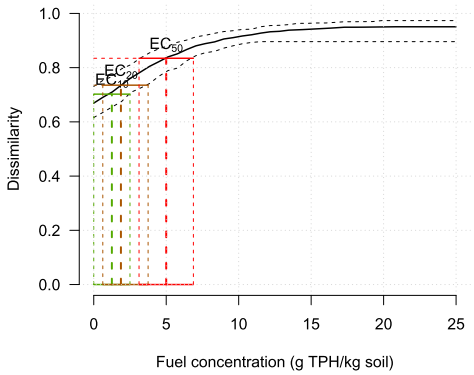
(a) Dep-GEM (b) Independent GEM

FIG. 4. *Diversity estimation results*. (a) *Dep-GEM model estimates* (50,000 *MCMC samples*). *Solid line*: *Simpson diversity estimate*. *Dashed lines*: 95% *credible interval for the Simpson diversity*. *Dots*: *Empirical estimates of Simpson diversity*. (b) *Independent GEM model estimates* (50,000 *MCMC samples*). *Triangles*: *posterior mean estimate of the Simpson diversity*.

a hormetic effect [see Calabrese (2005)]. Hormesis refers to a dose-response phenomenon characterized by favorable responses to low exposures to a pollutant: the species can feed on TPH at limited concentration, hence resulting in an increase of diversity, while above some threshold TPH starts to annihilate most of the species, hence reducing diversity.

The GEM model estimates are shown for comparison in Figure 4(b). These estimates showed more variability with respect to TPH in that they are closer to the empirical estimates of the diversity. Note that the GEM estimates were only available at levels of the covariate that were present in the data because of the independent nature of the model specification. The Dep-GEM, in contrast, provided predictions across the full range of TPH values. The credible bands are narrowest for TPH between 3000–5000 mg TPH/kg soil, due to borrowing of information between concentrated points, and they widen both at TPH $= 0$, due to a lot of data points with high variability, and at large TPH, due to few data points.

The Jaccard dissimilarity curve with respect to TPH is shown in Figure 5(a). The EC_x values are estimated as explained in Section 2.4 and provided in Table 5(b). Dissimilarity increased with TPH, illustrating that the contaminant alters community structure. Typically, EC_{10}, EC_{20} and EC_{50} values of Table 5(b) are reported in toxicity studies to be used in the derivation of protective concentrations in environmental guidelines, see Section 2.4. EC_{10}, EC_{20} and EC_{50} values estimated from this model are 1250, 1875 and 5000 mg TPH/kg soil respectively. For small x (less than 10%), the lower bound of the credible interval on the EC_x value is zero because both TPH and dissimilarity values are bounded below by zero. Conversely, for large x (more than 75%), the upper bound on the credible interval is 25,000, which is the limit of the TPH range in our analysis.



(a) Illustration of EC_x and Jaccard dissimilarity

x	EC_x	min	max
10	1250	0	2500
20	1875	625	3750
50	5000	3125	6875

(b) EC_x estimates and 95% credible intervals (min, max)

FIG. 5. *Jaccard dissimilarity and* EC_x *estimation results*. (a) *Posterior distribution* (*Dep-GEM model*) *of Jaccard dissimilarity between the control community where TPH equals zero and communities where* TPH > 0. *Solid line*: *mean estimate*. *Dashed lines*: 95% *credible intervals of the dissimilarity estimate*. *Horizontal segments and vertical dashed lines*: *illustration of estimation of* EC_x *values and their credible intervals*. (b) *Estimates of* EC_x *values and their credible intervals*.

4.2. *Posterior predictive checks*. Since we aim to compare the performance of the model in terms of diversity estimates, we also need to specify measures of goodness of fit. We resort to the conditional predictive ordinates (CPOs) statistics, which are now widely used in several contexts for model assessment; see, for example, Gelfand (1996). For each species j, the CPO statistic is defined as follows:

$$\mathrm{CPO}_j = f(\mathbf{Y}_j|\mathbf{Y}_{-j}) = \int \mathcal{L}(\mathbf{Y}_j|\theta)\pi(\mathrm{d}\theta|\mathbf{Y}_{-j}),$$

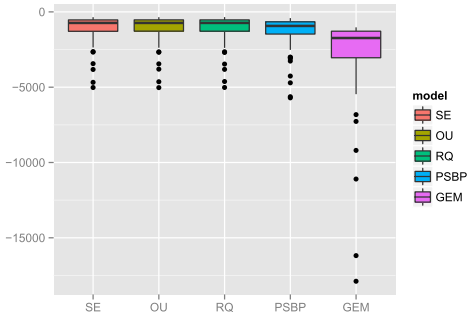
where $\mathcal{L}$ represents the likelihood (6), $\mathbf{Y}_j$ denotes data for species j over all sites, $\mathbf{Y}_{-j}$ denotes the observed sample $\mathbf{Y}$ with the jth species excluded, $\pi(\mathrm{d}\theta|\mathbf{Y}_{-j})$ is the posterior distribution of the model parameters $\theta = (\mathbf{Z}, \sigma_{\mathbf{Z}}, \lambda, M)$ based on data $\mathbf{Y}_{-j}$ and f denotes the (cross-validated) posterior predictive distribution. By rewriting the statistic CPO_j as

$$\mathrm{CPO}_j = \left(\int \frac{1}{\mathcal{L}(\mathbf{Y}_j|\theta)}\pi(\mathrm{d}\theta|\mathbf{Y})\right)^{-1},$$

it can be easily approximated by Monte Carlo as

$$\widehat{\mathrm{CPO}}_j = \left(\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\mathcal{L}(\mathbf{Y}_j|\theta^{(t)})}\right)^{-1},$$

where $\{\theta^{(t)}, t = 1, 2, \ldots, T\}$ is an MCMC sample from $\pi(\mathrm{d}\theta|\mathbf{Y})$. We illustrate the logarithm of the CPO_j, $j = 1, \ldots, J$, by boxplots in Figure 6(a) and pro-



(a)

Model	Mean.CPO	Median.CPO
Dep-GEM SE	−1131.3	−732.9
Dep-GEM OU	−1131.5	−732.7
Dep-GEM RQ	−1131.8	−731.4
PSBP	−1373.5	−932.3
Independent GEM	−2910.1	−1734.1
Single GEM	−34,606.2	−28,773.6

(b)

FIG. 6. *Log-conditional predictive ordinates* (*log-CPO*) *for different models and prior specifications* (*see text*). (a) *Boxplots of log-CPO*. (b) *Summaries of log-CPO*, *mean and median*.

vide the corresponding average and median values in Table 6(b). For the purpose of comparison, we have estimated six models. The first three are the Dep-GEM model with squared-exponential (SE), Ornstein–Uhlenbeck (OU) and rational quadratic (RQ) covariance functions; see Section S.2 in the Supplementary Material. The fourth is the probit stick-breaking process (PSBP) by Rodríguez and Dunson (2011). For the purpose of comparison, we set the hyperparameters of the PSBP to match the expected number of clusters of the Dep-GEM prior. Last, we used two variants of the GEM prior: first independent GEM priors at each site, as in Figure 4, and second a single GEM prior where the presence probabilities are all drawn from the same GEM distribution.

The single GEM is used as a very crude baseline (not shown in the boxplots) which does poorly compared to the other five models. As expected, the dependence induced by the Dep-GEM and the PSBP greatly improves the predictive quality of the model compared to the independent GEM. The Dep-GEM model has a slightly better predictive fit than the PSBP, which seems to indicate that the total ordering of the species that we use helps as far as prediction is concerned.

4.3. *Sensitivity to sampling variation*. A thorough sensitivity analysis to sampling variation was conducted in Arbel et al. (2015). It consists of estimating the model on modified data by (i) deleting the least abundant species; (ii) including additional species; (iii) excluding sites randomly. The model provided consistent results under these modifications, thus supporting some robustness to sampling variation.

5. Model considerations and extensions. In addition to a sensitivity analysis to sampling variation as in Section 4.3, here we consider sensitivity with respect to the model itself, which could be extended in a number of ways.

5.1. *Imperfect detection*. As pointed out in Section 2.3, we do not connect our model to the fields of occupancy modeling and imperfect detection, as developed, for instance, by Royle and Dorazio (2008). A possible extension to the current model is by accounting for imperfect detection. Following Dorazio et al. (2008), Royle and Dorazio (2006), a simple yet effective way to handle this extension is to define a probability of detection θ_i fixed for each site i, and to model the variability of θ_i across i by an exchangeable prior. It is anticipated that introducing such an underlying amount of nondetection in the model would increase the credible bands, but essentially not affect point estimates. Our model is clearly robust to the case where the probability of detection is homogeneous across sites, corresponding to a baseline $\theta_i = \theta$, and where a complete census of species is acquired. However, this is rarely the case in practice. The more realistic case of varying θ_i is harder to tackle in our case study, notably due to the lack of a rationale for detection. There is a growing body of literature aimed at diversity estimation when detection is properly accounted for in situations where replicate data are available; see, for instance, Broms, Hooten and Fitzpatrick (2015) for a thorough review. Such a methodology is based on Bayesian hierarchical models and could be applied in quite a natural way to our model (which is also a Bayesian hierarchical model) if replicate data were to be available. When replicates are not available, Broms, Hooten and Fitzpatrick (2015) recommend splitting sites in order to artificially create replicates. In our study, there was a concern about doing this due to the limited number of sites, but adding this layer is not prevented by our model and can be considered as an avenue for future investigation.

5.2. *Assumption on data, stochastic decrease of the $\hat{p}_j$'s*. We have assumed that after ordering with respect to overall abundance, the $\hat{p}_j$'s display a stochastically decreasing pattern as in Figure 1. In our experience, this assumption turns out to be satisfied with most of the species data sets, where species can be microbes, animals, words in text, DNA sequences, etc. However, this assumption proves to be overly restrictive in the following cases, (i) data might be subject to detection error: this is covered in the previous section by changing the prior adequately; (ii) there are outlier species which contradict the assumption: this could be addressed by adding a mixture layer in the prior specification; (iii) the underlying assumption itself is not true: this is, for instance, the case when all species are overall evenly distributed. A treatment would be context-specific and depend on the field.

5.3. *Comparison to other models*. In Section 4 we have compared the Dep-GEM model to other models: two GEM priors and the probit stick-breaking prior (PSBP) of Rodríguez and Dunson (2011). The benefits of the Dep-GEM over the first two is apparent in terms of smoothing of the estimates due to the a priori dependence; see Figure 4. It also carries over better predictive fit [see Figure 6(a) and Table 6(b)] and, most importantly, allows us to assess the response of species to any value of the contaminant, including unsampled values. With respect to the

PSBP, the CPOs indicate a slightly better predictive fit of the Dep-GEM prior, at least for the case study at hand.

6. Discussion. We have presented a Bayesian nonparametric dependent model for species data based on the distribution of the weights of a Dependent Dirichlet process, named Dep-GEM distribution, which is constructed by appeal to Gaussian processes. A fundamental advantage of our approach based on the stick-breaking is that it brings considerable flexibility in defining the dependence structure, which is defined by the kernel of a Gaussian process.

In terms of model fit, we have shown that the Dep-GEM model improves estimation compared to an independent GEM model. This was evaluated by computing conditional predictive ordinates (CPOs). In addition, our dependent model allows predictions at arbitrary covariate levels (not just those that were in the data). It allows, for example, estimation of the diversity and the dissimilarity across the full range of covariates. This is an essential feature in applications where the experimental data are sparse and is instrumental in estimating the EC_x values.

There are computational limitations to the use of this model. The estimation can deal with a large number of observations since the complexity grows linearly with the number of different observed species J. However, the number of unique covariate values I represents the limiting factor of the algorithm, and may lead to dimensionality problems. One could consider the use of INLA approximations [see Rue, Martino and Chopin (2009)] in the case of prohibitively large I.

Possible extensions of the present paper include the following. First, extra flexibility would be guaranteed by using the two-parameter Poisson–Dirichlet distribution instead of the GEM distribution, since it controls more effectively the posterior distribution of the number of clusters [Lijoi, Mena and Prünster (2007)]. This can be done at almost no extra cost since it only requires one additional step in the Gibbs sampler. Second, the Dep-GEM model is tested on univariate variables only, but could be extended to multivariate variables $X \in \mathbb{R}^d$, $d > 1$. Instead of a Gaussian process $\mathcal{Z}$, one would use a Gaussian random field $\mathcal{Z}^d$. To that purpose, all the methodology presented in Section 3 remains valid. The algorithm can become computationally challenging in the case of large dimensional covariates, but it does not carry additional difficulty for limited dimensions. Applications of such an extension are promising, such as testing joint effects in dynamic models (time $\times$ contaminant), spatial models (position $\times$ contaminant), etc.

Acknowledgments. The problem of estimating change in soil microbial diversity associated with TPH was motivated by discussions with the Terrestrial and Nearshore Ecosystems research team at the Australian Antarctic Division (AAD). The case study data set used in this paper was provided by the AAD, with particular thanks to Tristrom Winsley. We acknowledge the generous technical assistance of researchers at the AAD, in particular, Ben Raymond, Catherine King,

Tristrom Winsley and Ian Snape. We also wish to thank Antonio Canale for providing us with the implementation code for the probit stick-breaking process, Nicolas Chopin and Annalisa Cerquetti for helpful discussions, as well as the Editor, Karen Kafadar, an Associate Editor and three referees for their constructive feedback. Part of the material presented here is contained in the Ph.D. thesis Arbel (2013) defended at the University of Paris-Dauphine in September 2013.

SUPPLEMENTARY MATERIAL

Supplement to "Bayesian nonparametric dependent model for partially replicated data: The influence of fuel spills on species diversity" (DOI: 10.1214/16-AOAS944SUPP; .pdf). The supplementary material contains details about posterior computation and inference in the Dep-GEM model, additional results and omitted proofs that complement the analysis of the main text. It is available as Arbel, Mengersen and Rousseau (2016).

REFERENCES

AITCHISON, J. (1982). The statistical analysis of compositional data. *J. Roy. Statist. Soc. Ser. B* **44** 139–177. MR0676206

AITCHISON, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall, London. MR0865647

AITCHISON, J. (1994). Principles of compositional data analysis. In *Multivariate Analysis and Its Applications* (*Hong Kong*, 1992). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **24** 73–81. IMS, Hayward, CA. MR1479457

ALSTON, C. L., MENGERSEN, K. L. and GARDNER, G. E. (2011). Bayesian mixture models: A blood-free dissection of a sheep. In *Mixtures*: *Estimation and Applications* (K. Mengersen, C. P. Robert and M. Titterington, eds.) 293–308. Wiley, Chichester. MR2883358

ANDRIANAKIS, I. and CHALLENOR, P. G. (2012). The effect of the nugget on Gaussian process emulators of computer models. *Comput. Statist. Data Anal.* **56** 4215–4228. MR2957866

ARBEL, J. (2013). Contributions to Bayesian nonparametric statistics. Ph.D. thesis, Univ. Paris-Dauphine.

ARBEL, J., MENGERSEN, K. and ROUSSEAU, J. (2016). Supplement to "Bayesian nonparametric dependent model for partially replicated data: The influence of fuel spills on species diversity." DOI:10.1214/16-AOAS944SUPP.

ARBEL, J., KING, C. K., RAYMOND, B., WINSLEY, T. and MENGERSEN, K. L. (2015). Application of a Bayesian nonparametric model to derive toxicity estimates based on the response of Antarctic microbial communities to fuel-contaminated soil. *Ecol. Evol.* **5** 2633–2645.

ARBEL, J., FAVARO, S., NIPOTI, B. and TEH, Y. W. (2016). Bayesian nonparametric inference for discovery probabilities: Credible intervals and large sample asymptotics. *Statist. Sinica*. To appear. Available at arXiv:1506.04915.

BARRIENTOS, A. F., JARA, A. and QUINTANA, F. A. (2012). On the support of MacEachern's dependent Dirichlet processes and extensions. *Bayesian Anal.* **7** 277–309. MR2934952

BARRIENTOS, A. F., JARA, A. and QUINTANA, F. A. (2015). Bayesian density estimation for compositional data using random Bernstein polynomials. *J. Statist. Plann. Inference* **166** 116–125. MR3390138

BOHLIN, J., SKJERVE, E. and USSERY, D. (2009). Analysis of genomic signatures in prokaryotes using multinomial regression and hierarchical clustering. *BMC Genomics* **10** 487.

BORGES, E. P. and RODITI, I. (1998). A family of nonextensive entropies. *Phys. Lett. A* **246** 399–402. MR1649464

BROMS, K. M., HOOTEN, M. B. and FITZPATRICK, R. M. (2015). Accounting for imperfect detection in Hill numbers for biodiversity studies. *Methods in Ecology and Evolution* **6** 99–108.

CALABRESE, E. J. (2005). Paradigm lost, paradigm found: The re-emergence of hormesis as a fundamental dose response model in the toxicological sciences. *Environ. Pollut.* **138** 378–411.

CARON, F., DAVY, M. and DOUCET, A. (2007). Generalized Pólya urn for time-varying Dirichlet process mixtures. In 23*rd Conference on Uncertainty in Artificial Intelligence* (*UAI'*2007). Vancouver, Canada.

CERQUETTI, A. (2014). Bayesian nonparametric estimation of Patil–Taillie–Tsallis diversity under Gnedin–Pitman priors. Preprint. Available at arXiv:1404.3441.

CHUNG, Y. and DUNSON, D. B. (2011). The local Dirichlet process. *Ann. Inst. Statist. Math.* **63** 59–80. MR2748934

COLWELL, R. K., CHAO, A., GOTELLI, N. J., LIN, S.-Y., MAO, C. X., CHAZDON, R. L. and LONGINO, J. T. (2012). Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology* **5** 3–21.

CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*. Wiley, New York. MR1239641

DE'ATH, G. (2012). The multinomial diversity model: Linking Shannon diversity to multiple predictors. *Ecology* **93** 2286–2296.

DONNELLY, P. and GRIMMETT, G. (1993). On the asymptotic distribution of large prime factors. *J. Lond. Math. Soc.* (2) **47** 395–404. MR1214904

DORAZIO, R. M., MUKHERJEE, B., ZHANG, L., GHOSH, M., JELKS, H. L. and JORDAN, F. (2008). Modeling unobserved sources of heterogeneity in animal abundance using a Dirichlet process prior. *Biometrics* **64** 635–644, 670–671. MR2432438

DUNSON, D. B. and PARK, J.-H. (2008). Kernel stick-breaking processes. *Biometrika* **95** 307–323. MR2521586

DUNSON, D. B., PILLAI, N. and PARK, J.-H. (2007). Bayesian density regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 163–183. MR2325270

DUNSON, D. B. and XING, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *J. Amer. Statist. Assoc.* **104** 1042–1051. MR2562004

DUNSTAN, P. K., FOSTER, S. D. and DARNELL, R. (2011). Model based grouping of species across environmental gradients. *Ecol. Model.* **222** 955–963.

ELLIS, N., SMITH, S. J. and PITCHER, C. R. (2011). Gradient forests: Calculating importance gradients on physical predictors. *Ecology* **93** 156–168.

FAVARO, S., LIJOI, A. and PRÜNSTER, I. (2012). A new estimator of the discovery probability. *Biometrics* **68** 1188–1196. MR3040025

FAVARO, S., NIPOTI, B. and TEH, Y. W. (2016). Rediscovery of Good–Turing estimators via Bayesian nonparametrics. *Biometrics* **72** 136–145.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

FERRIER, S. and GUISAN, A. (2006). Spatial modelling of biodiversity at the community level. *J. Appl. Ecol.* **43** 393–404.

FERRIER, S., MANION, G., ELITH, J. and RICHARDSON, K. (2007). Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Divers. Distrib.* **13** 252–264.

FORDYCE, J. A., GOMPERT, Z., FORISTER, M. L. and NICE, C. C. (2011). A hierarchical Bayesian approach to ecological count data: A flexible tool for ecologists. *PLoS ONE* **6** e26785.

FOSTER, S. D. and DUNSTAN, P. K. (2010). The analysis of biodiversity using rank abundance distributions. *Biometrics* **66** 186–195. MR2756705

GELFAND, A. E. (1996). Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice* 145–161. Chapman & Hall, London. MR1397969

GEORGE, A. W., MENGERSEN, K. and DAVIS, G. P. (2000). Localization of a quantitative trait locus via a Bayesian approach. *Biometrics* **56** 40–51.

GIBBS, M. N. (1997). Bayesian Gaussian processes for regression and classification. Ph.D. thesis, Citeseer.

GILL, C. A. and JOANES, D. N. (1979). Bayesian estimation of Shannon's index of diversity. *Biometrika* **66** 81–85. MR0529150

GOOD, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40** 237–264. MR0061330

GRIFFIN, J. E. and STEEL, M. F. J. (2006). Order-based dependent Dirichlet processes. *J. Amer. Statist. Assoc.* **101** 179–194. MR2268037

GRIFFIN, J. E. and STEEL, M. F. J. (2011). Stick-breaking autoregressive processes. *J. Econometrics* **162** 383–396. MR2795625

HAVRDA, J. and CHARVÁT, F. (1967). Quantification method of classification processes. Concept of structural a-entropy. *Kybernetika* (*Prague*) **3** 30–35. MR0209067

HILL, M. O. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology* **54** 427–432.

HOLMES, I., HARRIS, K. and QUINCE, C. (2012). Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PLoS ONE* **7** e30126.

JOHNSON, D. S., REAM, R. R., TOWELL, R. G., WILLIAMS, M. T. and LEON GUERRERO, J. D. (2013). Bayesian clustering of animal abundance trends for inference and dimension reduction. *J. Agric. Biol. Environ. Stat.* **18** 299–313. MR3110895

KANIADAKIS, G., LISSIA, M. and SCARFONE, A. M. (2005). Two-parameter deformations of logarithm, exponential, and entropy: A consistent framework for generalized statistical mechanics. *Phys. Rev. E* (3) **71** 046128, 12. MR2139991

LI, H. (2015). Microbiome, metagenomics and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application* **2** 73–94.

LIJOI, A., MENA, R. H. and PRÜNSTER, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **94** 769–786. MR2416792

LIJOI, A., NIPOTI, B. and PRÜNSTER, I. (2014a). Bayesian inference with dependent normalized completely random measures. *Bernoulli* **20** 1260–1291. MR3217444

LIJOI, A., NIPOTI, B. and PRÜNSTER, I. (2014b). Dependent mixture models: Clustering and borrowing information. *Comput. Statist. Data Anal.* **71** 417–433. MR3131980

LOVELL, D., PAWLOWSKY-GLAHN, V., EGOZCUE, J. J., MARGUERAT, S. and BÄHLER, J. (2015). Proportionality: A valid alternative to correlation for relative data. *PLoS Comput. Biol.* **11** e1004075.

MACEACHERN, S. N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science* 50–55. Amer. Statist. Assoc., Alexandria, VA.

MACEACHERN, S. N. (2000). Dependent Dirichlet processes. Technical report, Dept. Statistics, The Ohio State Univ.

NEWMAN, M. C. (2012). *Quantitative Ecotoxicology*. CRC Press, Boca Raton, FL.

PATI, D., DUNSON, D. B. and TOKDAR, S. T. (2013). Posterior consistency in conditional distribution estimation. *J. Multivariate Anal.* **116** 456–472. MR3049916

PATIL, G. P. and TAILLIE, C. (1982). Diversity as a concept and its measurement. *J. Amer. Statist. Assoc.* **77** 548–567. MR0675883

PAWLOWSKY-GLAHN, V. and BUCCIANTI, A., eds. (2011). *Compositional Data Analysis*: *Theory and Applications*. Wiley, Chichester. MR2920574

PITMAN, J. (2006). *Combinatorial Stochastic Processes*. *Lecture Notes in Math.* **1875**. Springer, Berlin. MR2245368

RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA. MR2514435

RODRÍGUEZ, A. and DUNSON, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Anal.* **6** 145–177. MR2781811

RODRÍGUEZ, A., DUNSON, D. B. and GELFAND, A. E. (2010). Latent stick-breaking processes. *J. Amer. Statist. Assoc.* **105** 647–659. MR2724849

ROYLE, J. A. and DORAZIO, R. M. (2006). Hierarchical models of animal abundance and occurrence. *J. Agric. Biol. Environ. Stat.* **11** 249–263.

ROYLE, J. A. and DORAZIO, R. M. (2008). *Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities*. Academic Press, San Diego, CA.

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 319–392. MR2649602

SCHLOSS, P. D., WESTCOTT, S. L., RYABIN, T., HALL, J. R., HARTMANN, M., HOLLISTER, E. B., LESNIEWSKI, R. A., OAKLEY, B. B., PARKS, D. H., ROBINSON, C. J., SAHL, J. W., STRES, B., THALLINGER, G. G., HORN, D. J. V. and WEBER, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75** 7537–7541.

SICILIANO, S. D., PALMER, A. S., WINSLEY, T., LAMB, E., BISSETT, A., BROWN, M. V., VAN DORST, J., JI, M., FERRARI, B. C., GROGAN, P., CHU, H. and SNAPE, I. (2014). Soil fertility is associated with fungal and bacterial richness, whereas pH is associated with community composition in polar soil microbial communities. *Soil Biol. Biochem.* **78** 10–20.

SNAPE, I., SICILIANO, S. D., WINSLEY, T., VAN DORST, J., MUKAN, J., PALMER, A. S. and LAGEREWSKIJ, G. (2015). *Operational Taxonomic Unit* (*OTU*) *Microbial Ecotoxicology Data from Macquarie Island and Casey Station*: *TPH*, *Chemistry and OTU Abundance Data*. Australian Antarctic Data Centre.

VAN DEN BOOGAART, K. G. and TOLOSANA-DELGADO, R. (2013). Fundamental concepts of compositional data analysis. In *Analyzing Compositional Data with R*, *Use R!*. Springer, Heidelberg. MR3099409

VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.* **37** 2655–2675. MR2541442

WANG, Y., NAUMANN, U., WRIGHT, S. T. and WARTON, D. I. (2012). mvabund—An R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution* **3** 471–474.

J. ARBEL
COLLEGIO CARLO ALBERTO
VIA REAL COLLEGIO, 30
10024, MONCALIERI
ITALY
E-MAIL: julyan.arbel@carloalberto.org

K. MENGERSEN
QUEENSLAND UNIVERSITY OF TECHNOLOGY
ARC CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS
GPO BOX 2434, BRISBANE
AUSTRALIA
E-MAIL: k.mengersen@qut.edu.au

J. ROUSSEAU
CEREMADE—UNIVERSITÉ PARIS-DAUPHINE
PLACE DU MARÉCHAL DE LATTRE DE TASSIGNY
75775 PARIS CEDEX 16
FRANCE
E-MAIL: rousseau@ceremade.dauphine.fr