The Annals of Applied Statistics
2019, Vol. 13, No. 4, 2341–2362
https://doi.org/10.1214/19-AOAS1262

A NONPARAMETRIC SPATIAL TEST TO IDENTIFY FACTORS THAT SHAPE A MICROBIOME

BY SUSHEELA P. SINGH*, ANA-MARIA STAICU*, ROBERT R. DUNN*, NOAH FIERER† AND BRIAN J. REICH*

*North Carolina State University** *and University of Colorado*†

The advent of high-throughput sequencing technologies has made data from DNA material readily available, leading to a surge of microbiome-related research establishing links between markers of microbiome health and specific outcomes. However, to harness the power of microbial communities we must understand not only how they affect us, but also how they can be influenced to improve outcomes. This area has been dominated by methods that reduce community composition to summary metrics, which can fail to fully exploit the complexity of community data. Recently, methods have been developed to model the abundance of taxa in a community, but they can be computationally intensive and do not account for spatial effects underlying microbial settlement. These spatial effects are particularly relevant in the microbiome setting because we expect communities that are close together to be more similar than those that are far apart. In this paper, we propose a flexible Bayesian spike-and-slab variable selection model for presence-absence indicators that accounts for spatial dependence and cross-dependence between taxa while reducing dimensionality in both directions. We show by simulation that in the presence of spatial dependence, popular distance-based hypothesis testing methods fail to preserve their advertised size, and the proposed method improves variable selection. Finally, we present an application of our method to an indoor fungal community found within homes across the contiguous United States.

1. Introduction. The development and increased accessibility of high-throughput sequencing technologies have steadily decreased the cost of studying DNA (Reuter, Spacek and Snyder (2015), Heather and Chain (2016)). This has made analysis of microbial communities found in environmental samples easier. Armed with previously cost-prohibitive data, investigators have published a flurry of work leveraging microbiome information with applications in varied fields including forensics, ecology, archeology and public health. To date, much of this work has focused on studying abiotic and biotic factors that structure microbial communities and on identifying links between microbiome characteristics (e.g., composition or diversity) with specific outcomes. For example, studies have

Received February 2018; revised April 2019.
Key words and phrases. Bayesian nonparametrics, Dirichlet process, high dimensional data, spatial modeling, spike-and-slab prior, variable selection.

shown that microbiome composition can identify the source of a sample (Grantham et al. (2015)), linked changes in the gut microbiome to immune system dysfunction (Round and Mazmanian (2009)), tied reduced microbial diversity to obesity (Turnbaugh et al. (2009)), and connected imbalances in composition to Type 2 diabetes (Qin et al. (2012)). Though there has been an increased focus on defining the characteristics and markers of "healthy" microbiome communities for various systems within the body (Human Microbiome Project Consortium (2012), Ravel et al. (2011)), the tools to understand which factors may exert influence on microbiome composition are limited.

In this paper, we consider data from Barberán et al. (2015), which contains presence-absence indicators for over 57,000 fungal taxa based on dust samples from 1331 homes in the contiguous United States. In addition, we have geographic, climatic and household covariate information at each sampling location covering a wide range of explanatory variables. Our objective is to develop a testing procedure to identify covariates that influence microbiome composition that is applicable to high-dimensional, spatial, binary data and leverages the multivariate dependence between microorganisms.

Previous studies have demonstrated that a home's location, design, its occupants, and their activities, can all influence the microbiome composition present in dust within the home (Barberán et al. (2015), Kettleson et al. (2015), Dannemiller et al. (2016)). These studies generally reduce the data to summary measures (e.g., richness, Shannon Diversity index) or a measurement of dissimilarity in composition between samples such as Bray–Curtis dissimilarity (Bray and Curtis (1957)). Often, investigators then test for association between environmental covariates and these summaries using nonparametric permutation-based tests, the most popular of which are "ANalysis Of SIMilarities" (ANOSIM; Clarke (1993)) and "PERmutational Multivariate ANalysis Of VAriance" (PERMANOVA; Anderson (2001), McArdle and Anderson (2001)). A tenuous assumption of these tests is exchangeability across sampling locations; we show that violation of this assumption inflates Type I error rates. This is of particular importance in our motivating example because Barberán et al. (2015) note that nearby sampling locations exhibit more similar fungal communities than those that are far apart, and thus the assumption of exchangeability is known to be violated. In addition, Warton, Wright and Wang (2012) notes that ecological count data often does not conform to the mean-variance relationship implicitly assumed in distance metrics, and that this misspecification can make distance-based approaches unreliable for these applications.

Distance-based methods are also limited in interpretability. Because they partition the pairwise distances between samples, we cannot determine precisely how a covariate affects the composition or which taxa are directly affected. In a setting where an investigator may endeavor to target an intervention at a specific taxon or group of taxa, these tests are insufficient. Techniques such as redundancy analysis and canonical correspondence analysis are commonly used tools

that can allow these relationships to be specified, but they too rely on permutation-based tests with an underlying assumption of independence across sampling locations. Recently, methods addressing similar concerns have been developed for use on the compositional taxa counts (Chen and Li (2013), Grantham et al. (2017), Wadsworth et al. (2017), Zhao et al. (2015), Wang and Zhao (2017)). However, these methods are not appropriate for binary data and do not address spatial dependence in the data. Additionally, the proposed methods in Chen and Li (2013) and Wang and Zhao (2017) rely on optimization routines that may not be suitable for problems with thousands of sample locations and tens of thousands of taxa. Grantham et al. (2017) introduces a mixed effects model that accounts for correlation between taxa, but not between sampling locations.

Thorson et al. (2015) provides a spatial factor analysis approach to modeling the distribution of compositional taxa counts that provides computational advantages to the methods above. However, this method uses constructed spatial factors rather than environmental covariates directly, making it ill-suited to identify particular covariates that influence microbiome composition. Warton (2011) proposes a permutation-based test that analyzes the community response and is applicable to presence-absence data, but it too relies on an assumption of spatial independence and is computationally expensive, and thus it is infeasible for large problems. Ovaskainen, Hottola and Siitonen (2010) apply multivariate logistic regression to the binary community response, but their focus is on estimating the correlation between taxa and not on variable selection or covariate testing. Additionally, the direct estimation of the cross-species correlation matrix makes this approach infeasible for a problem of this size. Clark et al. (2017) provides a framework to unify disparate data types, including presence-absence indicators, but it does not account for spatial dependence, does not incorporate dimension reduction, and does not perform variable selection or covariate testing. Ovaskainen et al. (2016) and Ovaskainen et al. (2017) propose a spatial model for covariate effects on multiple species, but the model is parametric, full rank, and does not perform global hypothesis testing. Shirota, Gelfand and Banerjee (2017) proposes a nonparametric model for presence-absence data, but their aim is prediction rather than variable selection and testing for covariate effects.

As an alternative, we propose a flexible Bayesian variable selection method that uses a spike-and-slab prior and accounts for spatial dependence between nearby samples and cross-dependence between taxa. In particular, we model the spatial dependence using a flexible, principal components based approach, which is nonstationary in general. A unique feature of microbiome data is the large number of taxa, and we exploit this feature to estimate a nonstationary spatial covariance function using data-driven basis functions (Lorenz (1956)) and to relax the normality assumption common in spatial analysis (Gelfand, Kottas and MacEachern (2005), Nelsen (1999), Petrone, Guindani and Gelfand (2009), Reich and Fuentes (2007), Rodríguez, Dunson and Gelfand (2010)). We provide a global test of whether or not environmental covariates affect microbiome composition that

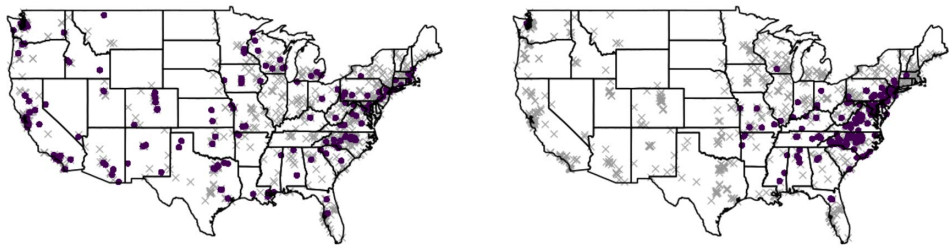
is interpretable, reliable and has fully characterized uncertainty. In addition, our method produces clusters of taxa and tests for covariate effects on individual taxa.

The remainder of the paper is structured as follows: in Section 2, we further describe the data; in Section 3, we detail the modeling procedure; in Section 4, we propose a procedure to estimate data-driven basis functions; in Section 5, we present a simulation study comparing our proposed method to several competitors; and in Section 6, we apply the proposed method to an indoor fungal community and compare our results to a previous study. Finally, we conclude with a brief summary in Section 7.

2. Motivating data. Wild Life of Our Homes (WLOH; yourwildlife.org) is a citizen-science project focused on studying microbial diversity in and around our homes. As part of the project, participants received sampling kits and instructions specifying nine standardized locations around their homes at which samples should be taken (Dunn et al. (2013)). The returned swabs were prepared using the direct PCR approach (Flores, Henley and Fierer (2012)), which amplifies the DNA present in the samples and allows them to be sequenced and classified into Operational Taxonomic Units (OTUs). The total amount of genetic information in a sample is an artifact of the sequencing process, and as a result, the raw number of sequenced reads identified for a given OTU is not comparable across samples. Thus, rather than analyzing the read counts directly, we consider the presence-absence indicators for each taxon. This transformation to presence-absence does not entirely remove the effects of the sequencing process from the data. For example, a sample with a low total number of reads may still incorrectly consider too many taxa as absent. However, the transformation tempers the effect in most other cases.

In addition to supplying sample swabs, participants were asked to complete a questionnaire providing details about the home's location, design features and its occupants. Geographic and climatic information were collected based on latitude and longitude from the Climate Research Unit Time Series v3.21 Dataset (Harris et al. (2014)) and the National Land Cover Database (Fry et al. (2011)) for a total of over 170 covariates.

From samples collected between 2012 and 2015, data was successfully sequenced for 1331 homes spanning the 48 contiguous United States and the District of Columbia indicating the presence of 57,304 distinct fungal taxa. Of these, we focus on $m = 763$ taxa identified in Barberán et al. (2015) as being more prevalent indoors than outdoors and on a set of $p = 20$ potentially influential covariates similar to those in their analysis. The presence or absence at each sampling location for two of these taxa are mapped in Figure 1. In the left panel, *Trichosporon asahii*, which is commonly found living on human skin, is seen to be widespread while in the right panel, *Perenniporia narymica* is seen to occur mainly in the mid-Atlantic region. Thus, there is evidence both that there is spatial dependence underlying the presence of fungal taxa and that the strength of that dependence varies across taxa.



(a) *Trichosporon asahii* (b) *Perenniporia narymica*

FIG. 1. *Map of presence* (*purple circle*) *or absence* (*gray* ×) *for two primarily indoor fungal taxa at each sampling location.*

3. Nonparametric spatial model. Let $Y_j(\mathbf{s})$ be the binary indicator that OTU $j = 1, \ldots, m$ is present in the sample at spatial location $\mathbf{s}$. Suppose that we have a set of p covariates, $\mathbf{X}(\mathbf{s}) = [X_1(\mathbf{s}), \ldots, X_p(\mathbf{s})]$, such as those described in Section 2. We assume there exists a latent continuous process $Z_j(\mathbf{s})$ such that $Y_j(\mathbf{s}) = \mathbb{1}\{Z_j(\mathbf{s}) > 0\}$. The latent process is modeled as

$$Z_j(\mathbf{s}) = \beta_{j0} + \mathbf{X}(\mathbf{s})\boldsymbol{\beta}_j + e_j(\mathbf{s}), \tag{3.1}$$

where β_{j0} is an intercept and $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp})'$ are regression coefficients that together model the probability that OTU j is present in a particular location. The final term, $e_j(\mathbf{s})$, is a multivariate Gaussian spatial process with $\mathrm{E}[e_j(\mathbf{s})] = 0$ and $\mathrm{Var}[e_j(\mathbf{s})] = 1$ that models dependence not captured in the covariates between spatial locations and between OTUs. This defines a probit link for the binary responses, $\mathrm{P}[Y_j(\mathbf{s}) = 1 \mid \mathbf{X}(\mathbf{s})] = \Phi[\beta_{j0} + \mathbf{X}(\mathbf{s})\boldsymbol{\beta}_j]$, where Φ is the standard normal cumulative density function. The assumption that $\mathrm{Var}[e_j(\mathbf{s})] = 1$ is necessary because the covariate magnitudes are identifiable only up to the ratio of effect size to variance.

Our primary goal is to develop a test to identify factors that influence microbiome composition. A covariate influences the composition if it affects the probability that *any* of the taxa will be present in a location, and thus we test the global hypotheses

$$\mathrm{H}_{0r} : \beta_{jr} = 0 \quad \text{for all } j \quad \text{versus} \quad \mathrm{H}_{1r} : \beta_{jr} \neq 0 \quad \text{for some } j. \tag{3.2}$$

The structure of this global test provides a means to identify an influential factor even if it affects only a small subset of the OTUs.

It remains to describe the modeling procedure for the individual components identified in (3.1). In Section 3.1 we specify a Bayesian variable selection model for the regression coefficients, $\boldsymbol{\beta}_j$, and in Section 3.2 we specify a nonparametric Bayesian model for the multivariate spatial process, $e_j(\mathbf{s})$.

3.1. *Identifying influential covariates.* We use a spike-and-slab prior for the coefficients, β_{jr}, to perform variable selection (George and McCulloch (1993), Mitchell and Beauchamp (1988), Kuo and Mallick (1998)). We assume that each coefficient can be written as $\beta_{jr} = \gamma_{jr}\delta_{jr}$ for an inclusion indicator, $\gamma_{jr} \in \{0, 1\}$, and magnitude, $\delta_{jr} \in \mathbb{R}$. This formulation allows us to simplify the hypotheses in (3.2) in terms of the number of OTUs for which the rth covariate is included, $M_r = \sum_{j=1}^{m} \gamma_{jr}$:

$$\mathrm{H}_{0r} : M_r = 0 \quad \text{versus} \quad \mathrm{H}_{1r} : M_r > 0. \tag{3.3}$$

To evaluate this, we calculate the posterior probability of the null hypothesis, $\mathrm{P}(M_r = 0 \mid \mathbf{Y})$, and compare to a threshold $t \in [0, 1]$. If the posterior probability of the null hypothesis is below the threshold, then the covariate is deemed influential.

Because we do not want to include the intercept in the variable selection process, we give it a separate prior $\beta_{j0} \overset{\text{iid}}{\sim} \mathrm{N}(0, \tau_0^{-1})$ with $\tau_0 \sim \mathrm{Gamma}(a_0, b_0)$. Similarly, the magnitudes have the standard conjugate formulation, $\delta_{jr} \overset{\text{indep}}{\sim} \mathrm{N}(0, \tau_r^{-1})$ with $\tau_r \overset{\text{indep}}{\sim} \mathrm{Gamma}(a_r, b_r)$. The inclusion indicators are distributed $\gamma_{jr} \overset{\text{indep}}{\sim} \mathrm{Bernoulli}(\pi_r)$, where π_r is the prior inclusion probability for the associated covariate.

The prior on π_r is chosen to induce sparsity in the coefficients such that the prior probability of the global null hypothesis in (3.3) is 0.5, reflecting no prior knowledge of whether or not a covariate is influential. In particular, the inclusion probabilities have prior density

$$\mathrm{P}(\pi_r) = \omega\left[\frac{1}{\mathrm{B}(1, \theta)}(1 - \pi_r)^{\theta-1}\right] + (1 - \omega), \tag{3.4}$$

a mixture of $\mathrm{Beta}(1, \theta)$ and $\mathrm{U}(0, 1)$ distributions weighted by $\omega \in [0, 1]$ and with $\theta \geq 1$. This prior has large mass on the sparse model with π_r near 0, as is common in high-dimensional Bayesian variable selection (Castillo and van der Vaart (2012), Zhou et al. (2015), Ročková and George (2018)), but remains flexible enough to allow substantial probability for large values of π_r. As ω approaches 1, the prior inclusion probabilities are driven toward 0, leading to sparser coefficient vectors as in the often used $\mathrm{Beta}(1, \theta)$ special case, and as ω decreases to 0 the uniform component dominates and covariates will be added more readily. We can also influence the level of sparsity in the coefficients through the parameter characterizing the Beta distribution, θ. If $\theta = 1$ then the prior is simply $\mathrm{U}(0, 1)$, and the coefficient vectors will not be sparse. As θ increases, the density associated with large values of π_r decays sharply, while density associated with small values changes less drastically, leading to a steeper density curve. As a reasonable default, fix $\omega = 0.5$ and set $\theta = m^2$, where m is the number of taxa under consideration, which gives $\mathrm{P}(M_r = 0) = 0.5$ *a priori* for each covariate, as desired.

3.2. *Capturing residual dependence.* As we show in Section 5, properly accounting for residual dependence is necessary for valid statistical inference. To model the residual dependence in (3.1), we assume that $e_j(\mathbf{s})$ can be decomposed into a structural component, $\xi_j(\mathbf{s})$, and an independent component (or nugget), $\epsilon_j(\mathbf{s})$, such that $e_j(\mathbf{s}) = \xi_j(\mathbf{s}) + \epsilon_j(\mathbf{s})$. The structural component contributes variance $\rho \in [0, 1]$, leaving the nugget distributed $\epsilon_j(\mathbf{s}) \overset{\text{iid}}{\sim} \mathrm{N}(0, 1-\rho)$ to satisfy the identifiability constraint that $\mathrm{Var}[e_j(\mathbf{s})] = 1$. We use a basis expansion model for $\xi_j(\mathbf{s})$ and write $\xi_j(\mathbf{s}) = \boldsymbol{\Psi}(\mathbf{s})\boldsymbol{\alpha}_j$, where $\boldsymbol{\Psi}(\mathbf{s}) = [\psi_1(\mathbf{s}), \ldots, \psi_L(\mathbf{s})]$ are orthogonal spatial basis functions common to all taxa and $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jL})'$ are their associated loadings, for L finite or infinity. The model for the process now becomes $e_j(\mathbf{s}) = \boldsymbol{\Psi}(\mathbf{s})\boldsymbol{\alpha}_j + \epsilon_j(\mathbf{s})$.

We use a Dirichlet process prior (Ferguson (1973)) for the distribution of the loadings, which can be written as $\boldsymbol{\alpha}_j \overset{\text{iid}}{\sim} f(\boldsymbol{\alpha})$, where f is the infinite mixture

$$f(\boldsymbol{\alpha}) = \sum_{k=1}^{\infty} p_k \mathbb{1}\{\boldsymbol{\alpha} = \boldsymbol{\mu}_k\}. \tag{3.5}$$

The mixture means have priors $\boldsymbol{\mu}_k \overset{\text{iid}}{\sim} \mathrm{N}(\boldsymbol{\mu}_0, \rho\mathbf{I}_L)$, where $\boldsymbol{\mu}_0 \sim \mathrm{N}(\mathbf{0}, \tau_{\mu_0}^{-1}\mathbf{I}_L)$, $\rho \sim \mathrm{U}(0, 1)$ and $\tau_{\mu_0} \sim \mathrm{Gamma}(a_{\mu_0}, b_{\mu_0})$. The mixture probabilities, p_k, are modeled using the stick-breaking representation (Sethuraman (1994)) wherein $p_1 = V_1$, $p_k = V_k \prod_{u<k}(1 - V_u)$ for $k > 1$, and $V_u \overset{\text{iid}}{\sim} \mathrm{Beta}(1, D)$. This ensures that $p_k > 0$ for all k and $\sum_{k=1}^{\infty} p_k = 1$ almost surely. Rather than fix the Dirichlet process precision parameter, we assign it an uninformative positive prior, $D \sim \mathrm{Gamma}(a_d, b_d)$. With this infinite mixture model, our prior for the distribution of the spatial random effects, $\xi_j(\mathbf{s})$, has large support in the class of spatial processes (Gelfand, Kottas and MacEachern (2005)). In practice, the infinite mixture model in (3.5) is truncated at K terms for computational purposes. That is, we assume $g_k \in \{1, \ldots, K\}$ for $K \leq m$ by setting $V_K = 1$, giving $f(\boldsymbol{\alpha}) = \sum_{k=1}^{K} p_k \mathbb{1}\{\boldsymbol{\alpha} = \boldsymbol{\mu}_k\}$.

The Dirichlet process prior can be viewed as a clustering model for the spatial loadings over the OTUs. If we let $g_j \in \{1, 2, \ldots\}$ denote the cluster label for OTU j, then the mixture probability, p_k, can be interpreted as $\mathrm{P}(g_j = k)$, the probability that OTU j will be assigned to cluster k. Then, given that OTU j has been assigned to cluster k, its associated spatial loading vector is the group mean for that cluster, that is, $\boldsymbol{\alpha}_j \mid g_j = k$ is $\boldsymbol{\mu}_k$. In the microbiome setting, it is reasonable to believe that taxa exhibit different spatial patterns, as in Figure 1, and that groups of taxa will behave similarly. For example, one may expect that organisms with similar functions or that require the same nutrients might be found in close proximity to one another. This leads to a natural expectation of clustering in the spatial effects over the OTUs.

In combination with the assumptions from the previous section, the model for the latent process becomes

$$\begin{aligned} Z_j(\mathbf{s}) &= \beta_{j0} + \mathbf{X}(\mathbf{s})\boldsymbol{\beta}_j + \boldsymbol{\Psi}(\mathbf{s})\boldsymbol{\alpha}_j + \epsilon_j(\mathbf{s}) \\ &= \beta_{j0} + \sum_{r=1}^{p} X_r(\mathbf{s})\gamma_{jr}\delta_{jr} + \sum_{l=1}^{L} \psi_l(\mathbf{s})\alpha_{jl} + \epsilon_j(\mathbf{s}), \end{aligned}$$

where $\boldsymbol{\beta}_j$ captures the covariates' effect on the probability that OTU j will be present at location $\mathbf{s}$, $\boldsymbol{\Psi}(\mathbf{s})\boldsymbol{\alpha}_j$ captures residual spatial trends and $\epsilon_j(\mathbf{s})$ are independent errors. The adoption of this flexible, nonstationary covariance model provides key advantages over a parametric alternative. First, the model can accommodate a wide range of covariance models, which promotes a model that is robust to covariance misspecification. Second, the principal components based approach reduces the dimensionality of the problem, allowing for fast computation even when a parametric model, be it stationary or nonstationary, may be computationally infeasible.

The details of the full proposed model and its implementation, as well as a discussion of its properties, are contained in the Supplementary Material (Singh et al. (2019)). We also show in the supplement that the covariance structure induced by our model is nonstationary in general, and that the strength of the Dirichlet process clustering controls the dependence between OTUs.

4. Estimating the spatial basis functions. The model detailed in Section 3.2 requires the construction of a set of spatial basis functions, $\boldsymbol{\Psi}(\mathbf{s})$, that are orthogonal and capable of reflecting nonstationarity. While there are several approaches available to estimate spatial basis functions from binary data (e.g., Lee, Huang and Hu (2010)), we follow ideas from functional principal component analysis for binary-valued functional data and estimate the basis functions as the eigenfunctions of an estimated covariance function of the spatial latent process (Hall, Müller and Yao (2008), Serban, Staicu and Carroll (2013)).

Let $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ be the set of spatial locations at which the binary $Y_j(\mathbf{s})$ are observed. Our goal is to construct an estimator of the covariance of the latent process, $Z_j(\mathbf{s})$. To do so, we follow the Taylor approximation technique of Hall, Müller and Yao (2008). Let $\sigma(\mathbf{s}, \mathbf{s}')$ be the covariance between $\mathbf{Z}(\mathbf{s})$ and $\mathbf{Z}(\mathbf{s}')$, which for $\mathbf{s} \neq \mathbf{s}'$ is estimated as

$$\hat{\sigma}(\mathbf{s}, \mathbf{s}') = \frac{\hat{\vartheta}(\mathbf{s}, \mathbf{s}')}{\phi\{\hat{\nu}(\mathbf{s})\}\phi\{\hat{\nu}(\mathbf{s}')\}}, \tag{4.1}$$

where $\phi(\cdot)$ is the standard normal density function. This is akin to equation (10) in Hall, Müller and Yao (2008), where the numerator, $\vartheta(\mathbf{s}, \mathbf{s}')$, represents $\mathrm{Cov}[\mathbf{Y}(\mathbf{s}), \mathbf{Y}(\mathbf{s}')]$, and the denominator acts as a scaling factor, with $\nu(\cdot)$ denoting the mean of the latent process. As we detail below, we use the residuals from probit regressions to estimate $\mathrm{Cov}[\mathbf{Y}(\mathbf{s}), \mathbf{Y}(\mathbf{s}')]$, which ensures that the eigenvectors of

this estimated covariance will explain mostly variation that cannot be explained by the covariates. This allows us to alleviate complications from spatial confounding.

However, the component estimators differ from Hall, Müller and Yao (2008) because we cannot assume that the latent processes share a smooth mean process. In our setting, the mean process may differ across taxa or may be nonsmooth due to its dependence on nonsmooth covariates. We first obtain $\hat{\eta}_j(\mathbf{s})$, the predicted probability that $Y_j(\mathbf{s}) = 1$ from separate probit regressions of $\mathbf{Y}_j$ onto $\mathbf{X}$ for each taxon. Then we smooth $m^{-1}\sum_{j=1}^{m}\hat{\eta}_j(\cdot)$ over 2-D space using a bivariate kernel smoother to obtain an "average" mean process $\bar{\eta}(\cdot)$, and let $\hat{\nu}(\cdot) = \Phi^{-1}\{\bar{\eta}(\cdot)\}$, where $\Phi^{-1}(\cdot)$ is the standard normal quantile function. In order to obtain the estimated covariance of $\mathbf{Y}(\mathbf{s})$ and $\mathbf{Y}(\mathbf{s}')$, we calculate $m^{-1}\sum_{j=1}^{m}[Y_j(\mathbf{s})Y_j(\mathbf{s}') - \hat{\eta}_j(\mathbf{s})\hat{\eta}_j(\mathbf{s}')]$ and smooth these estimates using a four-dimensional kernel smoother. The resulting smoothed estimates are collected as $\hat{\vartheta}(\mathbf{s}, \mathbf{s}')$. As is typical in nonparametric statistics, the optimal bandwidths are chosen using generalized cross-validation (Craven and Wahba (1978), Hastie, Tibshirani and Friedman (2009)).

Applying this procedure to the variances will result in biased estimates (Hall, Müller and Yao (2008)). To remove this bias, we consider a modified estimator, $\hat{\sigma}(\mathbf{s}, \mathbf{s})$, and use the intercept of the weighted linear model

$$\hat{\sigma}(\mathbf{s}, \mathbf{s}') = \beta_0 + w(\mathbf{s}, \mathbf{s}')\,\mathrm{d}(\mathbf{s}, \mathbf{s}')\beta + \epsilon,$$

for $\mathbf{s} \neq \mathbf{s}'$ and with weights $w(\mathbf{s}, \mathbf{s}') = \exp[-\frac{d(\mathbf{s},\mathbf{s}')}{d_{10}}]\mathbb{I}(d(\mathbf{s}, \mathbf{s}') \leq d_{10})$, where d_{10} is the distance between $\mathbf{s}$ and its 10th closest neighbor for some distance measure d. In our application, we use the great-circle distance in miles.

Let $\hat{\boldsymbol{\Sigma}}$ be the initial estimate of the spatial covariance matrix with elements $\hat{\sigma}(\mathbf{s}, \mathbf{s}')$. By construction, $\hat{\boldsymbol{\Sigma}}$ is symmetric. However, to ensure that it is positive semidefinite, we consider its low rank approximation. Let $\tilde{\phi}_1(\mathbf{s}), \ldots, \tilde{\phi}_L(\mathbf{s})$ be the leading L eigenvectors of $\hat{\boldsymbol{\Sigma}}$, scaled by the square root of their associated eigenvalues, such that they account for a specified percentage of explained variance. In our application, we use 90%. To preserve the variance structure described in Section 3.2 (i.e., $\mathrm{Var}[\xi_j(\mathbf{s})] = \rho$), we need to ensure that $\sum_{l=1}^{L}\tilde{\phi}_l^2(\mathbf{s}) = 1$. If $L < n$, this will require scaling the eigenvectors to obtain

$$\psi_l(\mathbf{s}) = \left[\frac{1}{\sum_{l=1}^{L}\tilde{\phi}_l^2(\mathbf{s})}\right]^{\frac{1}{2}}\tilde{\phi}_l(\mathbf{s}).$$

Let $\boldsymbol{\Psi} = [\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_L]$, where $\boldsymbol{\psi}_l = \{\psi_l(\mathbf{s}_1), \ldots, \psi_l(\mathbf{s}_n)\}'$ for $l = 1, \ldots, L$. After this scaling process, $\boldsymbol{\Psi}$ is no longer orthogonal on $\mathbb{R}^L$, and thus we rotate by its right singular vectors to obtain the proposed basis functions.

Now, $\boldsymbol{\Psi}$ is scaled appropriately to preserve the variance structure we require, rotated to preserve orthogonality between basis functions, and reflects the nonstationarity we expect in the data. The estimated basis functions are available only at the locations in $\mathcal{S}$, and extrapolation would be required to make spatial predictions

beyond the n sample locations. However, our objective is not spatial prediction, but rather to account for the complex dependence structure at the sampling locations to give a valid global test of covariate effects.

Because of the reliance on generalized cross-validation to select the bandwidth parameter, the four-dimensional smoothing step to obtain the $\hat{\vartheta}(\mathbf{s}, \mathbf{s}')$ estimates can be prohibitively expensive. Two approaches to alleviating this burden are either to use a different method to select the bandwidth or to make the cross-validation less computationally intensive. As an example, a reasonable approach that avoids cross-validation might be to construct a variogram, identify the distance at which the correlation decays, and use that distance to set a bandwidth. Alternatively, if the data contains sampling locations that are close to one another, one could downsample the locations while approximately preserving the spatial coverage of the data. Then, generalized cross-validation can be done quickly on this smaller, representative set of locations to obtain an estimated optimal bandwidth. This latter approach is utilized in our data application in Section 6.

5. Simulation study. In this study, we consider generating data while varying the type of spatial dependence in the latent process, the existence of cross-dependence between OTUs in the latent process, the magnitude of covariate effect size, and the degree of prevalence in covariate effects, and evaluate how these factors influence the true and false positive rates of the global test in (3.3).

5.1. *Methods*. We generate data on a 15×15 grid on the unit square for a total of $n = 225$ spatial locations. For each of $m = 50$ OTUs, we draw the latent process as $\mathbf{Z}_j \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}_j, 0.95\boldsymbol{\Sigma}_z + 0.05\mathbf{I}_n)$. The structure of $\boldsymbol{\Sigma}_z$ varies based on the type of spatial dependence:

(Ind) Independence: $\boldsymbol{\Sigma}_z = \mathbf{I}_n$,

(Exp) Stationary dependence: $\boldsymbol{\Sigma}_z$ is populated by the exponential covariance function with spatial range set such that the correlation between the two closest sites is 0.75 and

(Nonstat) Nonstationary dependence: where $\boldsymbol{\Sigma}_z(\mathbf{s}, \mathbf{s}') = \cos(2\pi s_1) \times \cos(2\pi s_1') + \sin(2\pi s_2)\sin(2\pi s_2')$ for $\mathbf{s} = (s_1, s_2)$.

This specification of a nonstationary covariance function is chosen to mimic the pockets of positive and negative correlation that may reflect the dependence on habitats expected in the microbiome setting. Plots of this covariance function evaluated at four points throughout the spatial domain are provided in the Supplementary Material (Singh et al. (2019)). When the setting calls for multivariate dependence in the latent process, we assume a separable covariance function and define $\mathrm{Cov}[Z_j(\mathbf{s}), Z_{j'}(\mathbf{s}')] = c(j, j')\boldsymbol{\Sigma}_z(\mathbf{s}, \mathbf{s}')$, where $c(j, j') = 0.8^{|j-j'|}$ is the cross-dependence function. In reality, we do not expect a meaningful ordering of the OTUs, but this covariance is used to generate data with a reasonable range of cross-correlations. The $p = 20$ covariates are drawn from a mean-zero Gaussian

process with separable covariance function $\mathrm{Cov}[X_r(\mathbf{s}), X_{r'}(\mathbf{s}')] = c(r, r')\boldsymbol{\Sigma}_x(\mathbf{s}, \mathbf{s}')$ where $c(r, r')$ is as above, and $\boldsymbol{\Sigma}_x$ is the exponential covariance with spatial range set such that the correlation between the two closest sites is 0.5.

Of the covariates, $p_0 = 6$ are influential (i.e., β_{jr} is nonzero for some j) and the remainder are unimportant for all OTUs (i.e., $\beta_{jr} = 0$ for all j). In order to examine the ability of the algorithm to detect covariate effects across prevalences and magnitudes, the six influential covariates are randomized such that two of them affect all OTUs, two affect a randomly selected 50% of OTUs, and the final two affect a randomly selected 10% of OTUs. Then, within each prevalence-based pair of nonnull covariates, the first is assigned a large magnitude of $\beta_{jr} = 0.5$ for its nonzero effects, and the second is assigned a small magnitude of $\beta_{jr} = -0.25$ for its nonzero effects.

Under each of the simulation settings we generate $N = 100$ replicate datasets and fit the proposed spatial nonparametric model and several competing models:

(PERM) PERMANOVA (Anderson (2001), McArdle and Anderson (2001)), a permutation-based hypothesis test as implemented in the `R` package `vegan 2.4-3` using Bray–Curtis dissimilarity.

(NS) Nonspatial variable selection model, that is, $\rho = 0$.

(NSM) Nonspatial variable selection model with multivariate random effects, where the model from Section 3.1 is adopted for variable selection but we adjust the model for residual dependence, $e_j(\mathbf{s}_i) = e_{ij}$, to model dependence across taxa. In particular, we still assume that we can decompose e_{ij} and write $e_{ij} = \xi_{ij} + \epsilon_{ij}$. However, in this adaptation, the structural component, ξ_{ij}, captures dependence across taxa but *not* spatial locations. We write $\xi_{ij} = \boldsymbol{\alpha}_i'\boldsymbol{\Psi}_j$, where $\boldsymbol{\Psi}_j$ are the logistic principal components of the $m \times m$ sample covariance matrix. The loadings are given the conjugate prior distribution, $\boldsymbol{\alpha}_i \overset{\text{iid}}{\sim} \mathrm{N}_L(0, \rho\mathbf{I}_L)$, and we choose the number of basis functions such that 90% of variance is explained.

(Mat) Parametric spatial model where $\boldsymbol{e}_j = [e_j(\mathbf{s}_1), \ldots, e_j(\mathbf{s}_n)]'$ from (3.1) is modeled using a Matérn covariance function. The smoothness has prior $\kappa \sim \mathrm{U}(0, 2)$ (Stein (1999), Banerjee (2005)), and the range has prior $\log(\zeta) \sim \mathrm{N}(0, \sigma_\zeta^2)$ where σ_ζ^2 is set such that the 99th percentile of the prior distribution for the range is the maximum observed distance. The computing details for this model are included in the Supplementary Material (Singh et al. (2019)).

(SNP) Proposed nonparametric spatial model using the nonstationary basis detailed in Section 4, with the maximum number of groups set to $K = m$.

For each of the Bayesian models (NS, NSM, Mat and SNP), we fit the model using a special case of (3.4) where $\omega = 1$ and $\theta = m$, which simplifies the prior to $\pi_r \overset{\text{iid}}{\sim} \mathrm{Beta}(1, m)$. This commonly used prior on the inclusion probabilities will make it more likely for π_r to be close to 0 than in the mixture setting. Our focus is on identifying covariates that are borderline cases, for example, factors that influence only a few taxa. The sharper cut of this simplified prior near the origin makes

the sampler less likely to include these covariate spuriously. To determine sensitivity to this prior specification, we also ran the simulation using the mixture prior in (3.4) with the recommended default values. The results are qualitatively the same, with improved performance for Mat in identifying small magnitude covariates but a reduced ability to identify low prevalence covariates. The model performance for SNP is broadly unchanged. The remainder of the prior specifications are detailed in the Supplementary Material (Singh et al. (2019)). The models are run for a total of 40,000 iterations with a burn-in period of 10,000, and the posterior samples are thinned by 2. We deem the rth covariate to be influential if the associated posterior probability of the null is below 0.05, that is, $\mathrm{P}(M_r = 0 \mid \mathbf{Y}) < 0.05$, for the Bayesian models, or if its p-value from PERMANOVA is below 0.05.

For each dataset, we evaluate the models using true positive rate (TPR) and false positive rate (FPR), presented in Table 1. Let M_r^* be the indicator that the rth covariate is truly influential. The true positive rate is the percent of truly influential covariates correctly classified as influential by the model for a given threshold t,

$$\mathrm{TPR}(t) = \frac{\sum_{r=1}^{p} M_r^* \mathbb{1}\{\mathrm{P}(M_r = 0 \mid \mathbf{Y}) < t\}}{p_0}.$$

TABLE 1

Summary of true positive rate (TPR), false positive rate (FPR) and average model fitting time in minutes for PERMANOVA (PERM), the nonspatial (NS), nonspatial multivariate (NSM), parametric Matérn (Mat) and proposed nonparametric (SNP) models

		Dependence Between Taxa					
		Independence			Autoregressive		
Spatial Dependence	Model	TPR	FPR	Time	TPR	FPR	Time
Independence	PERM	0.61	0.05	0.61	0.48	0.06	0.60
	NS	0.33	0.00	9.90	0.33	0.00	9.84
	NSM	0.26	0.00	10.94	0.21	0.01	10.88
	Mat	0.32	0.00	61.21	0.33	0.00	61.22
	SNP	0.28	0.00	22.19	0.33	0.00	22.22
Exponential	PERM	0.97	0.79	0.61	0.87	0.62	0.61
	NS	0.78	0.37	9.98	0.73	0.34	9.97
	NSM	0.54	0.18	10.88	0.41	0.12	10.81
	Mat	0.45	0.02	67.16	0.45	0.03	67.02
	SNP	0.56	0.06	23.02	0.54	0.06	22.50
Nonstationary	PERM	0.85	0.53	0.61	0.81	0.49	0.60
	NS	0.90	0.42	9.99	0.87	0.37	9.99
	NSM	0.61	0.16	10.82	0.52	0.09	10.79
	Mat	0.79	0.01	67.65	0.80	0.00	67.69
	SNP	0.90	0.02	22.84	0.90	0.04	22.47

TABLE 2

Inclusion rate for influential covariates for the parametric Matérn (Mat) and proposed nonparametric (SNP) models in the case of nonstationary spatial dependence and independence betweeen taxa, broken out by covariate magnitude (S = Small, L = Large) and prevalence (100%, 50%, 10%)

	Covariate Prevalence and Magnitude					
Model	100%L	100%S	50%L	50%S	10%L	10%S
Mat	1.00	1.00	1.00	0.79	0.82	0.12
SNP	1.00	1.00	1.00	0.99	0.97	0.43

The false positive rate is the percent of truly unimportant covariates that are incorrectly classified by the model as influential,

$$\mathrm{FPR}(t) = \frac{\sum_{r=1}^{p}(1 - M_r^*)\mathbb{1}\{\mathrm{P}(M_r = 0 \mid \mathbf{Y}) < t\}}{p - p_0}.$$

Finally, in Table 2, we consider the inclusion rate for the influential covariates for each model, broken out by magnitude of the covariate effect, small (S) or large (L), and the prevalence of the covariate effect, 100%, 50%, or 10%. The inclusion rate (IR) is defined as the proportion of the N simulation runs for which the method correctly classified the covariate as influential,

$$\mathrm{IR}_{r'}(t) = \frac{1}{N}\sum_{s=1}^{N}\mathbb{1}\{\mathrm{P}(M_{s,r'} = 0 \mid \mathbf{Y}) < t\},$$

for each of the $r' = 1, \ldots, p_0$ influential covariates. As in the global results presented in Table 1, we use a fixed threshold of $t = 0.05$.

5.2. *Results*. As is evident in Table 1, in the case of no spatial dependence in the data, PERM outperforms the Bayesian models. The Bayesian tests are overly conservative and struggle to overcome the decreased signal to noise ratio in this setting. The false positive rate for PERM is well controlled even in the face of multivariate dependence, which is reasonable given that the permutation is done at the sampling location level and thus the structure of any cross-dependence between taxa is preserved.

However, in the presence of spatial dependence, PERMANOVA fails to preserve the size of the hypothesis test and has false positive rates an order of magnitude higher than expected. This is perhaps not unexpected as the pseudo-F test is built on the assumption of exchangeability across sampling locations. Blind application of these permutation-based methods in settings where spatial independence across sampling locations is not a reasonable assumption will result in misleading conclusions.

When the data are spatially dependent, NS and PERM have high true positive rates accompanied by high false positive rates, indicating that the models favor including all covariates rather than discriminating between important and unimportant factors. The addition of correlation across taxa in NSM helps to ameliorate some of this effect, but the false positive rates are still much higher in NSM than in Mat or SNP. Thus, while modeling the dependence between taxa improves the nonspatial model, it is clear that it is not sufficient. Under the exponential correlation structure, Mat and SNP perform similarly, though Mat is too conservative and thus sacrifices some power. Under the nonstationary correlation structure, SNP provides a marked improvement in power while Mat is again overly conservative.

In Table 2, we present the inclusion rates for Mat and SNP broken out by prevalence and magnitude. We focus on the setting with nonstationary spatial dependence and independence between taxa, simply to keep the false positive rates relatively evenly matched so as to make a fair comparison. We exclude PERM, NS and NSM from this discussion because of their outsized false positive rates as discussed in the previous paragraph. Breaking out the model performance in this way allows us to see the contrast between the Bayesian spatial models. In particular, we can see that SNP outperforms the parametric model in identifying covariates with low prevalence and/or small magnitudes, which is our primary focus. SNP picks up the low prevalence, small magnitude covariate 43% of the time, whereas the parametric model selects it in only 12% of the replications.

In addition, the spatial parametric model takes $3\times$ longer to fit than the other models on average, and this is a relatively small problem with only 225 locations and 50 taxa. Mat requires several inversions of an $n \times n$ matrix during each MCMC iteration, and it is clear that this becomes computationally infeasible for problems much larger than this simulated setting. The proposed nonparametric model reduces the dimensionality of the problem for both large numbers of observations and a large number of observed taxa without sacrificing its aptitude to discern influential covariates from unimportant ones.

6. Data analysis. In light of PERMANOVA's demonstrated failure to preserve the size of the hypothesis test in the face of spatial and multivariate dependence, we revisit the analysis of Barberán et al. (2015) in which the authors determined which, if any, of a set of environmental and household covariates affect the indoor fungal community composition of homes. The covariates of interest included mean annual precipitation (MAP), mean annual temperature (MAT), net primary productivity (NPP), elevation, age of the home, number of bedrooms, number of inhabitants, female-to-male ratio of the home's inhabitants, smoking status, number of dogs/cats/birds, whether or not the home has a basement and number of days with the windows open. Using PERMANOVA, they find that the effects of outdoor variables and geographic location are more pronounced than the household covariates, but note that the presence of a basement in the home, the age

of the home, and the presence of a dog also affect the composition of the indoor fungal microbiome.

We follow the intuition of Barberán et al. (2015) and compile a similar list of covariates. In addition to those listed above, we include an indicator that the land is designated as forested, an indicator that the home is a rental unit, and the type of home (single family detached, multi-family dwelling, mobile). We replace the number of days with the windows open with the type of ventilation (central air-conditioning, central heat, window air-conditioning). NPP was missing for 81 of the sampling locations, and when considering only indoor fungal taxa, an additional 24 sampling locations had no present taxa. These locations have been removed, leaving $n = 1226$ locations and $p = 20$ covariates in the analysis. Maps of all of the included covariates are provided in the Supplementary Material (Singh et al. (2019)).

Using both PERMANOVA and the proposed nonparametric method, we investigated each covariate's ability to affect the composition of the taxa identified as the indoor fungal microbiome. SNP was run for 90,000 total iterations, keeping the final 62,000 posterior samples. Unlike in the simulation study, the maximum number of groups is set to $K = 500 < m$. We utilized the downsampling strategy discussed in Section 4 to build the spatial basis functions. We used an 80% threshold for explained variance, resulting in a total of $L = 137$ basis functions. The first few estimated basis functions are mapped in Figure 2. The first several

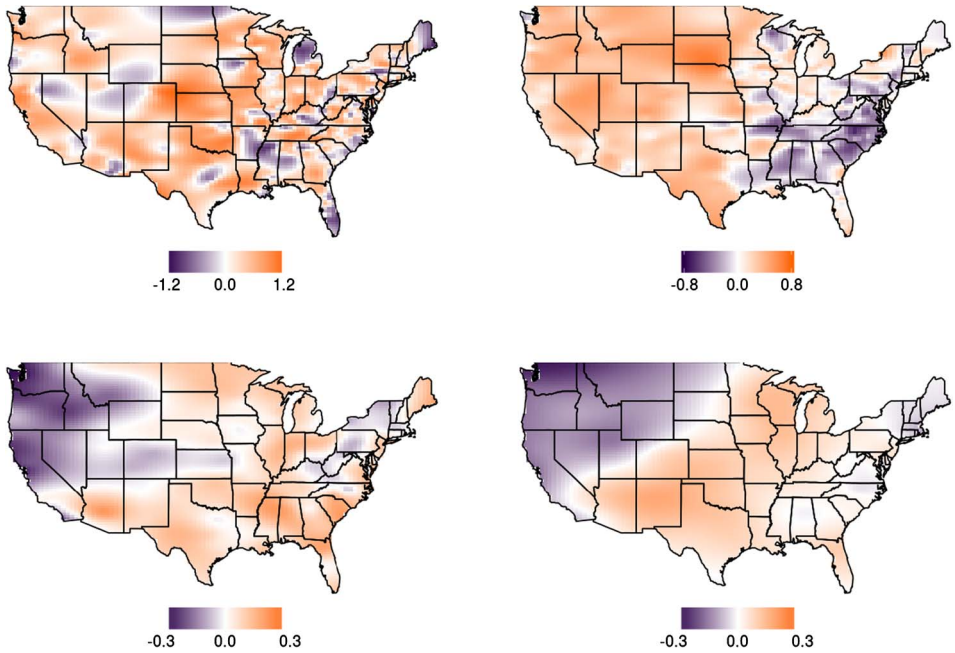


FIG. 2. *Maps of the first four spatial basis functions estimated from the WLOH data.*

TABLE 3

Summary of variable selection results from PERMANOVA (PERM) and the proposed spatial nonparametric method (SNP). p-values are reported from PERM, and the posterior probability of the null hypothesis, the expected number of taxa for which the covariate is included, and the number of taxa for which the coefficient value is positive or negative are reported for SNP

	PERM	SNP			
Covariate	p-value	$P(M_r = 0 \mid \mathbf{Y})$	$E[M_r \mid \mathbf{Y}]$	#Positive	#Negative
MAT	<0.001	0.00	439	41	163
NPP	<0.001	0.00	391	15	105
MAP	<0.001	0.00	339	27	57
Multifamily dwelling	0.038	0.00	91	12	0
Central A/C	<0.001	0.00	78	2	0
Forested	<0.001	0.00	29	0	0
Elevation	<0.001	0.00	19	0	0
Window A/C	<0.001	0.00	8	0	0
Older home	0.078	0.03	11	0	0
Central heat	0.015	0.03	8	0	0
Mobile home	0.289	0.28	2	0	0
Smoking status	0.756	0.33	1	0	0
Number of bedrooms	0.386	0.54	1	0	0
Basement	<0.001	0.60	1	0	0
Number of dogs	0.152	0.61	1	0	0
Rental home	0.075	0.71	1	0	0
Percentage of females	0.735	0.78	0	0	0
Number of occupants	0.016	0.79	0	0	0
Number of birds	0.627	0.86	0	0	0
Number of cats	0.558	0.93	0	0	0

functions reflect the nonstationarity in the data, while later basis functions reflect smooth spatial variation. Reported in Table 3 for each covariate are the p-value from PERMANOVA, the posterior probability of the null hypothesis, the posterior expected number of taxa for which the covariate is selected, and a count of the number of taxa for which the associated coefficient value is positive or negative, assessed as $\sum_{j=1}^{763} \mathbb{1}\{P(\beta_{jr} > 0 \mid \mathbf{Y}) > 0.975\}$ and $\sum_{j=1}^{763} \mathbb{1}\{P(\beta_{jr} < 0 \mid \mathbf{Y}) > 0.975\}$, respectively, for the proposed model. Note that the final two columns will not necessarily sum to the third, because the posterior distribution underlying them differ. The middle column is determined by the posterior distribution of the M_r statistics, while the final two columns are determined by the posterior distributions of the β_{jr}.

Comparing the p-values from PERMANOVA and the posterior probability of the null hypothesis from SNP, we see that the two models largely agree, but we can identify a few covariates that PERMANOVA includes at either the 0.05 or 0.10 significance level that would not be included in the SNP model. Given the inflated Type I error rates of the PERMANOVA test under spatial dependence in

the simulation study, it seems likely that these are false positives. The proposed method is able to identify both covariates that are important to many taxa (e.g., MAT) and those that are important only to a few (e.g., whether or not a home is older). In addition, we are able to precisely describe *how* covariates influence particular taxa. For example, as one would expect, we note that most fungal taxa prefer cooler climes, but that there are some taxa that seem to thrive in the warmer temperatures. Generally, we corroborate the findings of Barberán et al. (2015) and conclude that geographic and climatic factors are most influential to the indoor fungal microbiome composition. The household covariates that appear as influential are whether or not the home is older, whether or not the home is a multifamily dwelling, and whether or not the home has air-conditioning or central heating, all of which play a role in increasing the interaction between the indoor environment and the outdoors.

The 763 species are grouped into an estimated (posterior mean) 50 clusters. The largest clusters, based off of a k-means clustering algorithm with 50 clusters and using $1 - \mathrm{P}(g_j = g_{j'})$ as the dissimilarity matrix, contain taxa that exhibit little spatial clustering and tend to be present across the country. The smaller clusters tend to group together taxa that exhibit more localized presence. For example, in Figure 3, the left panel displays the presence for the 113 taxa assigned to the largest cluster and the right panel displays the presence for the eight taxa assigned to a smaller cluster.

In as much as our results add to those of previous analyses using data from the WLOH project, it is worth commenting about the additional biological insights our approach offers. Barberán et al. (2015) found that, compared to bacteria, the composition of fungi in homes tended to be much more strongly driven by outdoor environmental conditions. In our analysis, this conclusion is even more strongly supported. The primary factors associated with differences in the composition of indoor fungi among households were those associated with climate and its effects, and nearly all (96.7%) significant associations of individual taxa with particular covariates were associations with these environmental factors.

Net Primary Productivity (NPP) was a particularly important correlate of the composition of indoor fungi. In the United States, NPP is highly correlated with

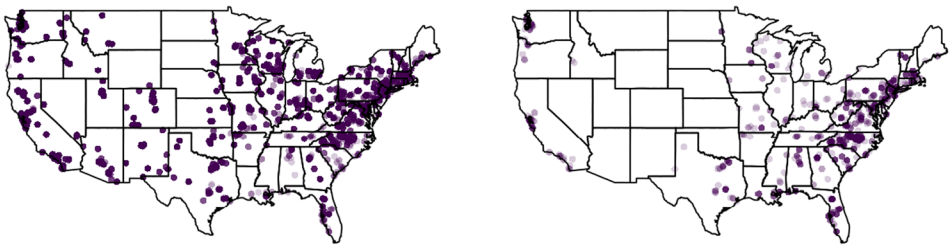


FIG. 3. *Map of presence for taxa assigned to a large cluster of* 113 *taxa and a small cluster of* 8 *taxa. A darker point indicates that a higher number of taxa are present in a location.*

forest cover, such that areas with higher NPP are almost always forests. In this light, it is perhaps not surprising that species more common in regions with high NPP were species associated with forests and dead and down wood, including multiple taxa of the species *Xylobolus annosus*. Conversely, species that became less common under high NPP tended to be from the genera *Alternaria*, *Cladosporium*, *Aspergillus* and *Phoma*, many of which are associated with decaying plant material. Fungi from decaying plant material, much of which is in leaf litter, might be more likely to become airborne in open habitats such as grasslands. Many species were also influenced by the direct effects of the mean annual temperature or precipitation in the region in which a house was located.

One of the few nonenvironmental covariates identified as influential was whether or not the home is a multifamily dwelling. Multifamily dwellings tended to favor fungi associated with human bodies or foods. These included three *Candida* taxa, *Cryptococcus oeirensis*, *Penicillium concetricum*, and the brewer's yeast (*Saccharomyces cerevisae*). Also more common in these homes were *Rhodotorula mucilaginosa*, which does well under stressful conditions, such as those associated with bathrooms that are frequently cleaned. The way in which a house was heated or cooled also influenced which species were present. In particular, as has been noted in smaller scale studies (Hamada and Fujita (2002)), we confirm here that houses with air conditioning tend to be more likely to have *Cladosporium* fungi, which are known to grow in air conditioning units and then spread through houses.

Considering that the homes we studied differed greatly in their size, number of occupants, age, design, and much more, the fact that these variables influence so very little of fungal composition is striking. Houses, in general, favor some fungi relative to others and yet just which species appears to depend nearly exclusively on where the house is built.

7. Discussion. In this paper, we introduced a nonparametric Bayesian model for identifying factors that influence microbiome composition, as well as a covariance estimator amenable to high-dimensional, binary data akin to that of Hall, Müller and Yao (2008). The proposed model uses spike-and-slab variable selection to identify covariates that influence the occupancy probability of even a small subset of the taxa. It also utilizes a set of orthogonal, data-driven spatial basis functions and a Dirichlet process prior over their associated loadings to cluster the OTUs into groups of taxa that exhibit similar spatial responses, allowing dimension reduction in both the number of spatial locations and the number of taxa under consideration, greatly alleviating the computational burden compared to a parametric spatial model.

We demonstrated via simulation that the proposed model outperforms naïve nonspatial models, with and without considering dependence between taxa, and PERMANOVA in identifying influential covariates, and showed that violating the assumption of exchangeability of sampling locations underlying PERMANOVA leads to Type I error rates that are not well controlled. We also showed that the

proposed model is able to better identify low prevalence and/or small magnitude covariate effects as compared to a parametric spatial competitor.

We applied our proposed model to the indoor fungal microbiome from the Wild Life of Our Homes project as identified in Barberán et al. (2015). We were able to broadly substantiate their conclusion that geography and climate are the most influential factors affecting indoor fungal communities, and we provided additional detail in describing how factors affect particular taxa rather than simply classifying factors as influential or unimportant.

This work primarily focused on the global hypothesis of whether or not a covariate influences microbiome composition as a whole. However, the model also allows for local hypothesis tests of individual covariate values, which have not been fully explored here, though they are reported in the last two columns of Table 3. We discussed the application and potential of these local tests, but did not rigorously test the true and false positive rates for covariate effects on individual taxa. An additional area of focus for future work is to expedite and improve the covariance estimation process to scale with large problems.

SUPPLEMENTARY MATERIAL

Supplement to "A nonparametric spatial test to identify factors that shape a microbiome" (DOI: 10.1214/19-AOAS1262SUPP; .pdf). We summarize the proposed model, as well as some of its properties, and provide computational details. In the remainder of the supplement, we present maps of the spatial covariates used in the analysis of Section 6 of the main article.

REFERENCES

ANDERSON, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26** 32–46.

BANERJEE, S. (2005). On geodetic distance computations in spatial modeling. *Biometrics* **61** 617–625. MR2140936

BARBERÁN, A., DUNN, R. R., REICH, B. J., PACIFICI, K., LABER, E. B., MENNINGER, H. L., MORTON, J. M., HENLEY, J. B., LEFF, J. W. et al. (2015). The ecology of microscopic life in household dust. *Proc. R. Soc. Lond., B Biol. Sci.* **282** 212–220.

BRAY, J. R. and CURTIS, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.* **27** 325–349.

CASTILLO, I. and VAN DER VAART, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.* **40** 2069–2101. MR3059077

CHEN, J. and LI, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Ann. Appl. Stat.* **7** 418–442. MR3086425

CLARK, J. S., NEMERGUT, D., SEYEDNASROLLAH, B., TURNER, P. J. and ZHANG, S. (2017). Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data. *Ecol. Monogr.* **87** 34–56.

CLARKE, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Aust. J. Ecol.* **18** 117–143.

CRAVEN, P. and WAHBA, G. (1978). Smoothing noisy data with spline functions. *Numer. Math.* **31** 377–403.

DANNEMILLER, K. C., GENT, J. F., LEADERER, B. P. and PECCIA, J. (2016). Influence of housing characteristics on bacterial and fungal communities in homes of asthmatic children. *Indoor Air* **26** 179–192.

DUNN, R. R., FIERER, N., HENLEY, J. B., LEFF, J. W. and MENNINGER, H. L. (2013). Home life: Factors structuring the bacterial diversity found within and between homes. *PLoS ONE* **8** e64133.

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

FLORES, G. E., HENLEY, J. B. and FIERER, N. (2012). A direct PCR approach to accelerate analyses of human-associated microbial communities. *PLoS ONE* **7** e44563.

FRY, J. A., XIAN, G., JIN, S., DEWITZ, J. A., HOMER, C. G., LIMIN, Y., BARNES, C. A., HEROLD, N. D. and WICKHAM, J. D. (2011). Completion of the 2006 national land cover database for the conterminous United States. *Photogramm. Eng. Remote Sens.* **77** 858–864.

GELFAND, A. E., KOTTAS, A. and MACEACHERN, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *J. Amer. Statist. Assoc.* **100** 1021–1035. MR2201028

GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

GRANTHAM, N. S., REICH, B. J., PACIFICI, K., LABER, E. B., MENNINGER, H. L., HENLEY, J. B., BARBERÁN, A., LEFF, J. W., FIERER, N. et al. (2015). Fungi identify the geographic origin of dust samples. *PLoS ONE* **10** e0122605.

GRANTHAM, N. S., REICH, B. J., BORER, E. T. and GROSS, K. (2017). MIMIX: A Bayesian mixed-effects model for microbiome data from designed experiments. Manuscript in review.

HALL, P., MÜLLER, H.-G. and YAO, F. (2008). Modelling sparse generalized longitudinal observations with latent Gaussian processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 703–723. MR2523900

HAMADA, N. and FUJITA, T. (2002). Effect of air-conditioner on fungal contamination. *Atmos. Environ.* **36** 5443–5448.

HARRIS, I., JONES, P. D., OSBORN, T. J. and LISTER, D. H. (2014). Updated high-resolution grids of monthly climatic observations—the CRU TS3.10 dataset. *Int. J. Climatol.* **34** 623–642.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2722294

HEATHER, J. M. and CHAIN, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics* **107** 1–8.

HUMAN MICROBIOME PROJECT CONSORTIUM (2012). Structure, function and diversity of the healthy human microbiome. *Nature* **486** 207–214.

KETTLESON, E. M., ADHIKARI, A., VESPER, S., COOMBS, K., INDUGULA, R. and REPONEN, T. (2015). Key determinants of the fungal and bacterial microbiomes in homes. *Environ. Res.* **138** 130–135.

KUO, L. and MALLICK, B. (1998). Variable selection for regression models. *Sankhya B* **60** 65–81. MR1717076

LEE, S., HUANG, J. Z. and HU, J. (2010). Sparse logistic principal components analysis for binary data. *Ann. Appl. Stat.* **4** 1579–1601. MR2758342

LORENZ, E. N. (1956). Empirical orthogonal functions and statistical weather prediction.

MCARDLE, B. H. and ANDERSON, M. J. (2001). Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology* **82** 290–297.

MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83** 1023–1036. With comments by James Berger and C. L. Mallows and with a reply by the authors. MR0997578

NELSEN, R. B. (1999). *An Introduction to Copulas*. *Lecture Notes in Statistics* **139**. Springer, New York. MR1653203

OVASKAINEN, O., HOTTOLA, J. and SIITONEN, J. (2010). Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology* **91** 2514–2521.

OVASKAINEN, O., ROY, D. B., FOX, R. and ANDERSON, B. J. (2016). Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. *Methods Ecol. Evol.* **7** 428–436.

OVASKAINEN, O., TIKHONOV, G., NORBERG, A., GUILLAUME BLANCHET, F., DUAN, L., DUNSON, D., ROSLIN, T. and ABREGO, N. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecol. Lett.* **20** 561–576.

PETRONE, S., GUINDANI, M. and GELFAND, A. E. (2009). Hybrid Dirichlet mixture models for functional data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 755–782. MR2750094

QIN, J., LI, Y., CAI, Z., LI, S., ZHU, J., ZHANG, F., LIANG, S., ZHANG, W., GUAN, Y. et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490** 55–60.

RAVEL, J., GAJER, P., ABDO, Z., SCHNEIDER, G. M., KOENIG, S. S. K., MCCULLE, S. L., KARLEBACH, S., GORLE, R., RUSSELL, J. et al. (2011). Vaginal microbiome of reproductive-age women. *Proc. Natl. Acad. Sci. USA* **108** 4680–4687.

REICH, B. J. and FUENTES, M. (2007). A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. *Ann. Appl. Stat.* **1** 249–264. MR2393850

REUTER, J. A., SPACEK, D. V. and SNYDER, M. P. (2015). High-throughput sequencing technologies. *Molecular Cell* **58** 586–597.

ROČKOVÁ, V. and GEORGE, E. I. (2018). The spike-and-slab LASSO. *J. Amer. Statist. Assoc.* **113** 431–444. MR3803476

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

ROUND, J. L. and MAZMANIAN, S. K. (2009). The gut microbiota shapes intestinal immune responses during health and disease. *Nat. Rev., Immunol.* **9** 313–323.

SERBAN, N., STAICU, A.-M. and CARROLL, R. J. (2013). Multilevel cross-dependent binary longitudinal data. *Biometrics* **69** 903–913. MR3146786

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. MR1309433

SHIROTA, S., GELFAND, A. E. and BANERJEE, S. (2017). Spatial joint species distribution modeling using Dirichlet processes.

SINGH, S. P., STAICU, A., DUNN, R. R., FIERER, N. and REICH, B. J. (2019). Supplement to "A nonparametric spatial test to identify factors that shape a microbiome." DOI:10.1214/19-AOAS1262SUPP.

STEIN, M. L. (1999). *Interpolation of Spatial Data*: *Some Theory for Kriging*. *Springer Series in Statistics*. Springer, New York. MR1697409

THORSON, J. T., SCHEUERELL, M. D., SHELTON, A. O., SEE, K. E., SKAUG, H. J. and KRISTENSEN, K. (2015). Spatial factor analysis: A new tool for estimating joint species distributions and correlations in species range. *Methods Ecol. Evol.* **6** 627–637.

TURNBAUGH, P. J., HAMADY, M., YATSUNENKO, T., CANTAREL, B. L., DUNCAN, A., LEY, R. E., SOGIN, M. L., JONES, W. J., ROE, B. A. et al. (2009). A core gut microbiome in obese and lean twins. *Nature* **457** 480–484.

WADSWORTH, W. D., ARGIENTO, R., GUINDANI, M., GALLOWAY-PENA, J., SHELBURNE, S. A. and VANNUCCI, M. (2017). An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinform.* **18** 94.

WANG, T. and ZHAO, H. (2017). A Dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms. *Biometrics* **73** 792–801. MR3713113

WARTON, D. I. (2011). Regularized sandwich estimators for analysis of high-dimensional data using generalized estimating equations. *Biometrics* **67** 116–123. MR2898823

WARTON, D. I., WRIGHT, S. T. and WANG, Y. (2012). Distance-based multivariate analyses confound location and dispersion effects. *Methods Ecol. Evol.* **3** 89–101.

ZHAO, N., CHEN, J., CARROLL, I. M., RINGEL-KULKA, T., EPSTEIN, M. P., ZHOU, H., ZHOU, J. J., RINGEL, Y., LI, H. et al. (2015). Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *Am. J. Hum. Genet.* **96** 797–807.

ZHOU, J., BHATTACHARYA, A., HERRING, A. H. and DUNSON, D. B. (2015). Bayesian factorizations of big sparse tensors. *J. Amer. Statist. Assoc.* **110** 1562–1576. MR3449055

DEPARTMENT OF STATISTICS
NORTH CAROLINA STATE UNIVERSITY
RALEIGH, NORTH CAROLINA 27695
USA
E-MAIL: susheelapsingh@gmail.com
astaicu@ncsu.edu
rob_dunn@ncsu.edu
noah.fierer@colorado.edu
bjreich@ncsu.edu