Submitted to the Annals of Applied Statistics

A JOINT BAYESIAN MODEL FOR CLUSTERED AND NON-CLUSTERED TAXA IN MICROBIOME SAMPLES

BY SUMAN MAJUMDER[1], BRENT A. COULL[1], JESSICA L. MARK WELCH[2], PATRICK J. LA RIVIERE[3], FLOYD E. DEWHIRST[4], JACQUELINE R. STARR[5] AND KYU HA LEE[1]

[1]*Harvard University, smajumder@hsph.harvard.edu; bcoull@hsph.harvard.edu; klee@hsph.harvard.edu*

[2]*Marine Biological Laboratory, jmarkwelch@mbl.edu*

[3]*University of Chicago, pjlarivi@uchicago.edu*

[4]*Forsyth Institute, fdewhirst@forsyth.org*

[5]*Brigham and Women's Hospital, spjst@channing.harvard.edu*

Nature abounds in instances where one type of object spatially clusters around another. This is certainly the case in microbial datasets. We have scenarios where multiple taxa cluster around a central taxon, we have scenarios where there is a layered, almost hierarchical, structure of clustering among different taxa. An example is presented here where we have *Streptococcus* and *Porphyomonas* clustering around *Corynebacterium* while *Porphyomonas* clusters around *Streptococcus*. None of the existing methods is a good fit for analyzing such datasets. A simultaneous modeling of such spatially clustered and non-clustered taxa in a human dental plaque sample is presented here. The proposed fully Bayesian method uses Neyman-Scott process with parent location information, whenever available, resulting in huge gain in performance in the simulation study. The method is reasonably fast and works well on the real datasets as well.

1. Introduction.

Examples of one type of object spatially orienting around another are plentiful in nature. Celestial objects cluster around one another such as planets revolving around a star. Animals of different types congregate around source of nutrition such as water, food or warmth. In the human body itself, immune cells gather around foreign organisms to fight them. A ubiquitous structure as spatial clustering therefore needs to be well quantified for us to understand these structures and make inference about and based on them. These structures are present in cosmic to micro scale objects around us and understanding them would benefit us in understanding cosmological design, solving ecological problems, performing microbiome analysis, to name a few.

In this paper, we focus on the analysis of microbial data where spatial clustering has occurred. These datasets differ from the others as they bring their own unique features. One key aspect is that the cluster centers are often observed here as they are usually another taxon residing in the area. We call them 'parent' taxon and the clustering taxon is called 'offspring' taxon. Another challenge here is that these dataset consists of multiple taxa not all of which are spatially clustered. It is important to model all of these taxa simultaneously to get a sense of

Keywords and phrases: Dental plaque sample, Imaging, Parent-offspring models, Point processes, Spatial statistics, Thomas process.

correlation between them. This therefore becomes a multivariate inference problem. Apart from these, each dataset comes with its unique challenges. In particular, the human dental plaque dataset that we analyze in this paper, has multiple taxa clustering around the same parent taxon. There is also a layered spatial clustering structure where taxon A has clustered around taxon B which in turn is clustered around taxon C. This makes the problem much more complex than the usual parent-offspring clustering modeling.

Spatial modeling of microbial data is usually done using Poisson processes and most popularly by using log-Gaussian Cox process (LGCP) models (Møller, Syversveen and Waagepetersen, 1998). Multi-taxa models are also available to be modeled by multivariate LGCP models (Møller, Syversveen and Waagepetersen, 1998). However, typical LGCP models are not well-suited here as they do not incorporate any spatial clustering information. Furthermore, the approximation methods employed for analysis using LGCP models (such as Diggle et al., 2013) would distort the spatial clustering information resulting in possibly invalid results.

A more suited approach is to use Neyman-Scott processes (Neyman and Scott, 1958; Illian et al., 2008; Chiu et al., 2013) or shot noise Cox process (SNCP) (Møller, 2003) models. However, these methods are also not exactly suited to handle the unique challenges the dataset present. Inference methods using Neyman-Scott models and its extensions (Moller and Waagepetersen, 2003; Møller and Waagepetersen, 2007; Waagepetersen, 2007; Diggle, 2013; Guan, 2006; Tanaka, Ogata and Stoyan, 2008; Guttorp and Thorarinsdottir, 2012; Mrkvička, Muška and Kubečka, 2014; Kopeckỳ and Mrkvička, 2016) do not use parent information in the model as their applications did not call for it. Moreover, Neyman-Scott process model is univariate in nature. Multiple offspring versions of Neyman-Scott process model have been proposed (Tanaka and Ogata, 2014). However this assumes separate and independent parent taxon for each offspring taxon. A multivariate SNCP model was proposed by Jalilian et al. (2015). But this also does not incorporate parent information and neither is there any guarantee that the multiple offspring taxa considered would come from the same parent taxon.

In this paper, we propose a novel methodology to model multiple taxa, spatially clustered or otherwise, simultaneously using Poisson processes such that it uses available parent information to more accurately capture the spatial clustering structure. The proposed model is flexible and can be molded to be used in various scenarios for both univariate and multivariate analyses. In this paper, we discuss in detail about the form of the model that suits the human dental plaque data that we analyze. The other potential models that can arise from the proposed framework have been discussed in Section 6. Simulation studies show improved estimation using our model in comparison to minimum contrast method (Diggle et al., 2013). The analysis of the human dental plaque sample also shows interesting results.

2. Dental Plaque Sample Data.

Dental plaque sample was collected from a donor who was asked not smoke, floss or brush for 12 hours and to not eat for 2 hours to allow the microbial community to grow and consolidate. The sample was then sequenced and probed. The identified taxa in the image are *Actinomyces*, *Capnocytophaga*, *Corynebacterium*, *Fusobacterium*, *Leptotrichia*, *Neisseriaceae*, *Pasteurellaceae*, *Porphyomonas*, *Streptococcus* and *Eubacterium* (used for probing). Figure 1 shows the composition of these taxa in the dental plaque sample. The black space in the image indicate regions where no taxon was observed. This image was then used to determine the centroids of the taxa to be used as locations for these observations. The image was scaled such that 1 unit is equivalent to 1μm. This is the scale of the microbial community in human dental plaque and so doing the analysis in this scale is both meaningful and convenient. The resulting image is presented in Figure 2.

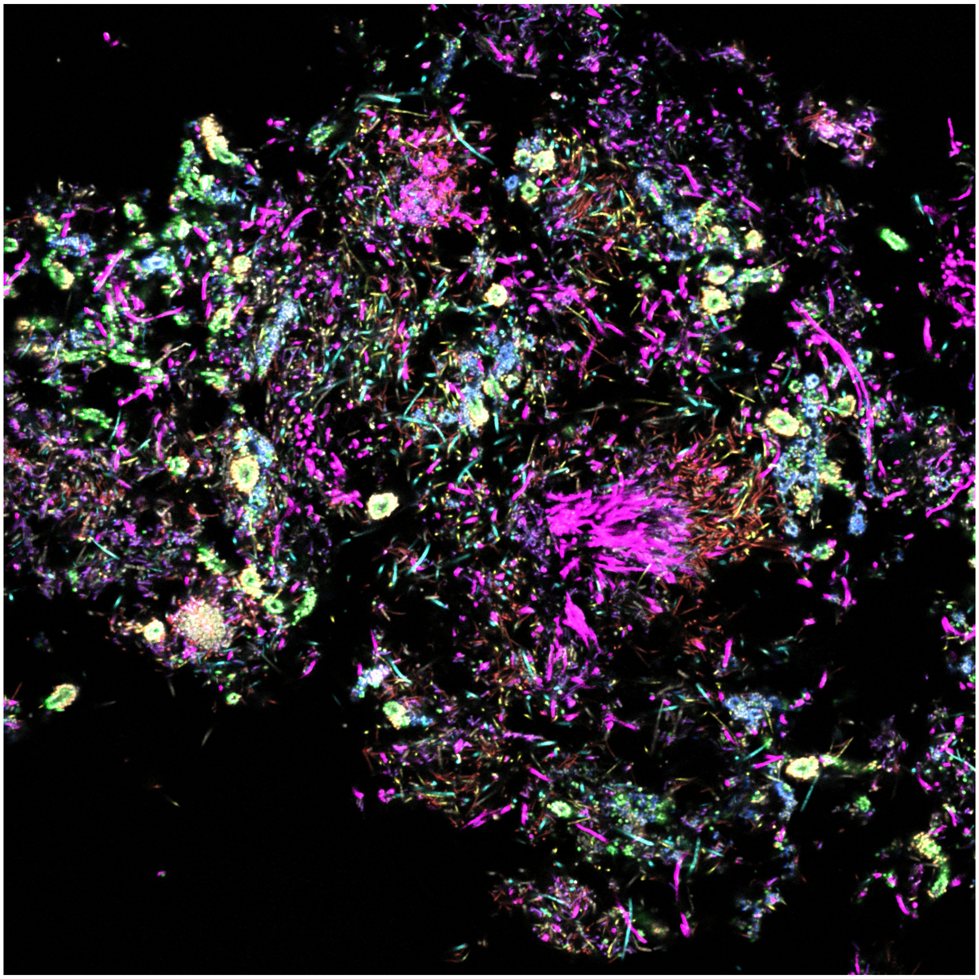

Fig 1: RGB image of a dental plaque sample from from a human donor. The black space indicates regions where no observations are made. Important taxa in the image are *Corynebacterium* (Pink), *Streptococcus* (Green), *Porphyomonas* (Blue) and *Pasteurellaceae* (Orange).

We see the phenomenon of *Streptococci* clustering around the tips of *Corynebacterium*, as mentioned in Mark Welch et al. (2016), in Figure 1. We also see *Porphyomonas* forming clusters around *Corynebacterium* in the same way. Figure 3 illustrates this further. This means that we have a scenario where there are multiple offspring taxa clustering around the same parent taxon. We also see *Pasteurellaceae* clustering around *Streptococcus* as well. This is evident in Figure 4 and was also mentioned in Mark Welch et al. (2016) and Perera et al. (2020). This is a unique situation where observations from one offspring taxon clusters around another taxon that is an offspring process itself. The other taxa, excluding *Eubacterium*, are scattered around homogeneously. Individual images for these are presented in the Appendix. *Eubacterium* is excluded from the analysis as they were included for the probing and are of not much importance.

3. Methodology. Let $Y_1, Y_2, \ldots, Y_m$ be m processes that are present in the data with $\lambda_i(\mathbf{s})$ is the intensity function for process Y_i at location $\mathbf{s} \in \mathcal{W} \subset \mathbb{R}^d$. The observation window is denoted by $\mathcal{W}$ and d can be 2 or 3 depending on the image in question. Define $Y = \cup_{i=1}^{m} Y_i$ to be the superimposed process of all these processes. We will model the superimposed Y

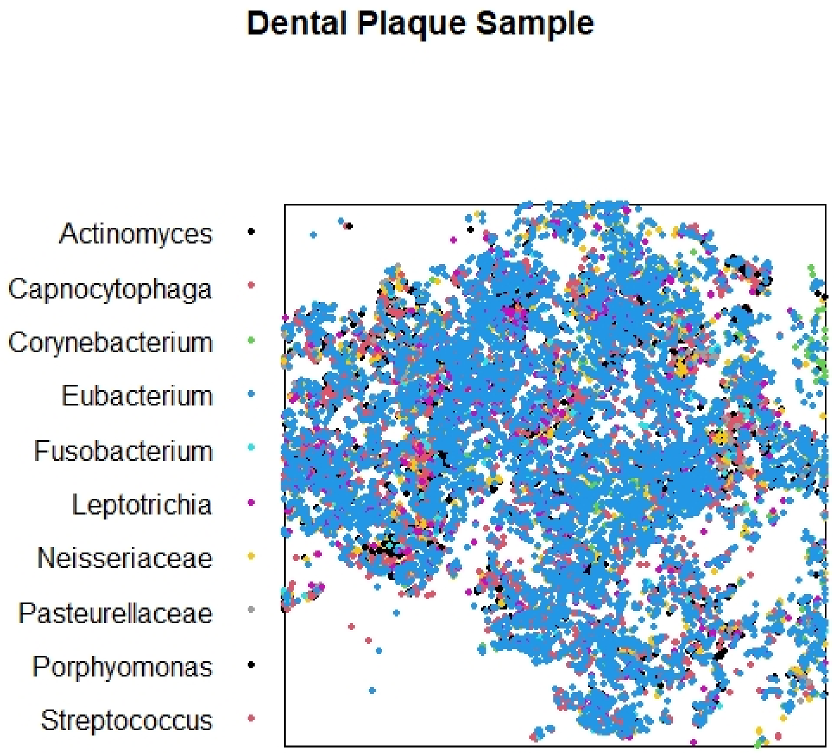

Fig 2: Object scale (1μm) image of the centroids of the different taxa in the dental plaque sample.

process as a Poisson process with intensity function $\rho(\mathbf{s})$ at location $\mathbf{s} \in \mathcal{W} \subset \mathbb{R}^d$ defined as

$$\rho(\mathbf{s}) = \sum_{i=1}^{m} \lambda_i(\mathbf{s}) \mathbb{I}(\mathbf{s} \in Y_i), \tag{3.1}$$

where $\mathbb{I}(\mathbf{s} \in Y_i)$ is the indicator function of the i-th process being observed at location $\mathbf{s} \in \mathcal{W}$. The indicator terms serve two important purpose in the model. Firstly, they ensure that in one location there are never more than one taxon observed. The second purpose the indicator function serves is to ensure that in each location of the observation window, there is always one taxon present. This means the model avoids black spaces in the image and is not suitable to be applied to them. This forces us to reduce unnecessary black spaces in the image, in line with the recommendations of the practitioners, to get a proper analysis.

In Eq. 3.1, we do not put any restrictions on the form of $\lambda_i(\mathbf{s})$ for $i = 1, \ldots, m$ and $\mathbf{s} \in \mathcal{W}$. This allows us the flexibility to choose our models as the data necessitates. In our case, we

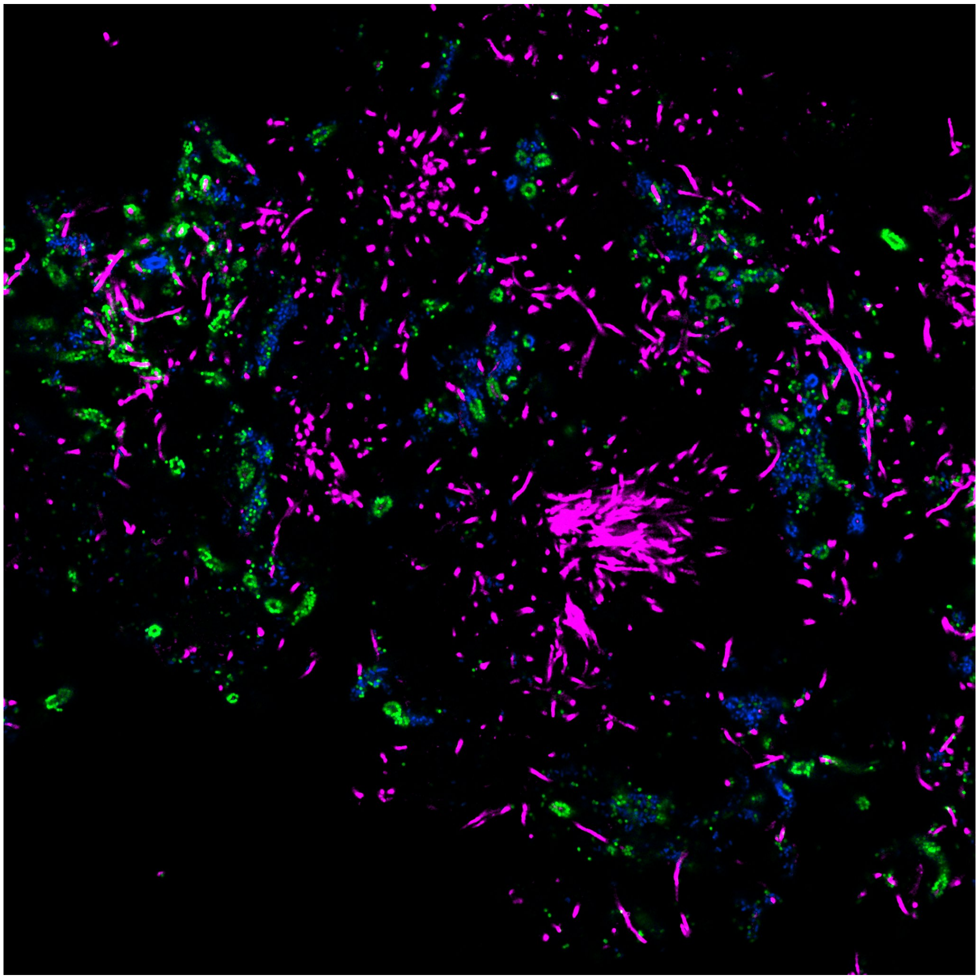

Fig 3: RGB image of only *Corynebacterium* (Pink), *Streptococcus* (green) and *Porphyomonas* (Blue) in the dental plaque sample. The corncob structures are present in areas of the image corresponding to the tip of the *Corynebacterium*. The central portion consisting of stems of the *Corynebacterium* are relatively empty.

would be using a combination of homogeneous Poisson processes (HPP) and Neyman-Scott processes (NSP). We discuss these in relative detail next. other possibilities of models that can arise from this framework is discussed in Section 6.

Homogeneous Poisson Process: If we assume the intensity of the subprocesses $Y_1, \ldots, Y_m$ are constant over space, i.e., $\lambda_i(s) = \lambda_i$, then each of the subprocesses are considered to be HPPs. The entire process is then a inhomogeneous Poisson process (IHPP) resulting from superimposition of HPPs.

Neyman-Scott Process: We can have observations from an offspring taxon X_c grow around a parent c which comes from the parent process C. The resulting offspring process $Z = \cup_{c \in C} X_c$ is called a Neyman-Scott process. Classical modeling takes C to be a homogeneous Poisson process and the offspring locations are assumed to be distributed around the parent $c \in C$ according to some rule $\alpha k(\cdot - c, h)$ with $k(\cdot, \cdot)$ being a kernel, h being the bandwidth parameter and α denoting the average number of offspring per parent. Typical choices of $k(\cdot, \cdot)$ are Gaussian (referred to as Thomas process), uniform (Matérn process), Cauchy and

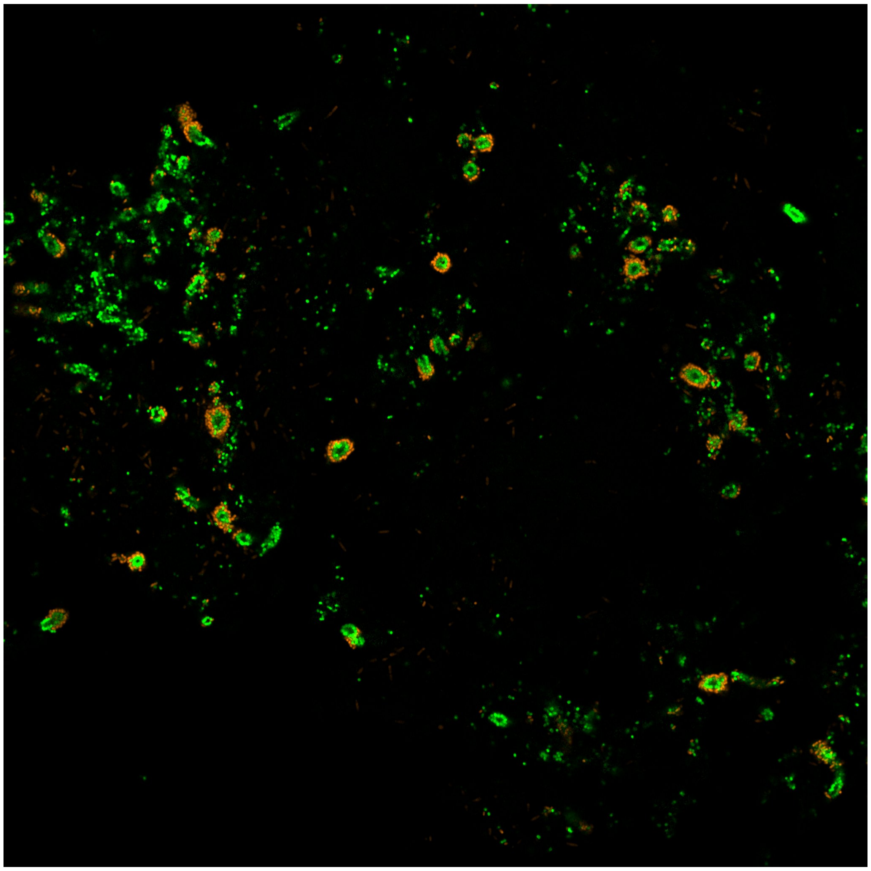

Fig 4: RGB image of *Pasteurellaceae* (Orange) clustering around *Streptococcus* (Green) in the human dental plaque sample.

variance-gamma kernels (See Chiu et al., 2013; Illian et al., 2008, for details). To achieve this, we can have $m = 2$ with Y_1 being a parent process, modeled by a homogeneous Poisson process (HPP) with intensity λ^C and Y_2 being an offspring to $C = Y_1$ with intensity $\alpha k(\cdot, h)$. In this scenario, Eq. 3.1 can be written as

$$\rho(\mathbf{s}) = \lambda^C \mathbb{I}(\mathbf{s} \in C) + \alpha \sum_{\mathbf{c} \in C} k(\mathbf{s} - \mathbf{c}, h) \mathbb{I}(\mathbf{s} \in Y_2). \tag{3.2}$$

Multi-layered Parent-Offspring Model: We can have a scenario where there are $p(< m)$ many processes $Y_1, \dots, Y_p$ which are all offspring processes from parent process $C_1, \dots, C_p$. However, some of $C_1, \dots, C_p$ may be the same process and some of these may even be processes included in $Y_1, \dots, Y_p$. Let us denote $Y_{p+1}, \dots, Y_{p+q}$ to be the $q(< m, p+q \leq m)$ unique parent processes that are not offspring processes themselves which can be modeled as HPPs themselves with intensities λ_v^C for $v = p+1, \dots, p+q$. The remainder of the $m-p-q$ taxa are also modeled as HPPs with intensities λ_j for $j = p+q+1, \dots, m$. In such a scenario, Eq. 3.1 would be written as

$$\rho(\mathbf{s}) = \sum_{l=1}^{p} \alpha_l \sum_{\mathbf{c}_l \in C_l} k_l(\mathbf{s} - \mathbf{c}_l, h_l) \mathbb{I}(\mathbf{s} \in Y_l) + \sum_{v=p+1}^{p+q} \lambda_v^C \mathbb{I}(\mathbf{s} \in Y_v) + \sum_{j=p+q+1}^{m} \lambda_j \mathbb{I}(\mathbf{s} \in Y_j). \tag{3.3}$$

This is the scenario we see in the human dental plaque sample we collected. We have three offspring taxa, namely *Streptococcus*, *Porphyomonas* and *Pasteurellaceae*. The parent taxon for the first two is *Corynebacterium* and for the third is *Streptococcus* which itself is an offspring taxon. So, according to notation in Eq. 3.3, Y_1, Y_2, Y_3 are offspring taxa processes corresponding to *Streptococcus*, *Porphyomonas* and *Pasteurellaceae*. $C_1 = C_2 = Y_4$ is the parent process associated with *Corynebacterium* and $C_3 = Y_1$. Moving forward, we will call the model detailed in Eq. 3.3 as our proposed method and expand upon its components and the necessity computing required for this model.

Since the superposed process Y is modeled as a Poisson process with intensity function $\rho(\mathbf{s})$, the corresponding log-likelihood function is

(3.4)

$$\begin{aligned} l(Y|\boldsymbol{\theta}) &\simeq -\int\limits_{\mathcal{W}} \rho(\mathbf{u})d\mathbf{u} + \sum_{\mathbf{y}\in Y} \log \rho(\mathbf{y}), \\ &\simeq \sum_{l=1}^{p} \alpha_l \sum_{\mathbf{c}_l \in C_l} \int\limits_{\mathcal{W}} k_l(\mathbf{u}-\mathbf{c}_l, h_l)d\mathbf{u} - \sum_{v=p+1}^{p+q} |\mathcal{W}|\lambda_v^C - \sum_{j=p+q+1}^{m} |\mathcal{W}|\lambda_j \\ &+ \sum_{l=1}^{p} \sum_{\mathbf{y}\in Y_l} \log\left(\alpha_l \sum_{\mathbf{c}_l\in C_l} k_l(\mathbf{y}-\mathbf{c}_l, h) \right) + \sum_{v=p+1}^{p+q} n_v \log \lambda_v^C + \sum_{j=p+q+1}^{m} n_j \log \lambda_j, \end{aligned}$$

where $\simeq$ implies equal up to a constant and $\boldsymbol{\theta}$ is the collection of all parameters in the model, namely $\{\alpha_l\}_{l=1}^{p}$, $\{h_l\}_{l=1}^{p}$, $\{\lambda_v^C\}_{v=p+1}^{p+q}$ and $\{\lambda_j\}_{j=p+q+1}^{m}$. Also, n_i denotes the numer of observation from the i-th process in the observation window $\mathcal{W}$ which has area $|\mathcal{W}|$.

Given a prior specification, the log-posterior density can now be computed as

$$\log p(\boldsymbol{\theta}|Y) \simeq l(Y|\boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta}), \tag{3.5}$$

where $l(Y|\boldsymbol{\theta})$ is as given in Eq. 3.4 and $\pi(\boldsymbol{\theta})$ is the product of all the individual prior densities. We then proceed to draw samples from this posterior distribution using a simple MCMC method. Since the parent locations are known, we do not require a reversible jump MCMC mechanism here. The most complicated term to compute are the integral terms which are approximated using Monte Carlo methods since $\int_{\mathcal{W}} k_l(\mathbf{u}-\mathbf{c}, h_l)d\mathbf{u} = \mathbb{E}\left(\mathbb{I}_{\{\mathbf{X}_l \in \mathcal{W}\}}\right)$, where $\mathbf{X}_l$ is a $\mathbb{R}^d$-valued random variable with density $k_l(\cdot - \mathbf{c}, h_l)$. For our applications we use $k_l(\cdot - \mathbf{c}, h_l)$ to be bivariate Gaussian with mean $\mathbf{c}$ and variance $h_l^2\mathbf{I}$.

3.1. *Prior Settings and Practical Considerations.* We now specify the prior distributions for the model parameters to complete the Bayesian model specification. We use

$$\begin{aligned} \alpha_l &\overset{iid}{\sim} \text{Gamma}(a_Y, b_Y)\,,\ l = 1, \ldots, p; \\ h_l &\overset{iid}{\sim} \text{Half-Normal}(\sigma)\,,\ l = 1, \ldots, p; \\ \lambda_v^C &\overset{iid}{\sim} \text{Gamma}(a_C, b_C)\,,\ v = p+1, \ldots, p+q; \\ \lambda_j &\overset{iid}{\sim} \text{Gamma}(a, b)\,,\ j = p+q+1, \ldots, m; \end{aligned} \tag{3.6}$$

as our priors, where $\overset{iid}{\sim}$ denotes independent and identically distributed. Choices for a, b, a_Y, b_Y, a_C, b_C are to be made such the corresponding Gamma priors are diffuse enough. Typical values used for them in our analysis is 0.01. Using half-normal prior for the bandwidth parameters ensures that the bulk of the mass remain on the smaller values of h_l, apriori.

TABLE 1

Description of the settings for the 12 different scenarios considered for the simulation study. The offspring density is controled by setting $(\alpha_1, \alpha_2) = (1.5, 1)$ *for 'Sparse',* $(4, 3)$ *for 'Dense' and* $(4, 1)$ *for 'Mixed' setting. Bandwidth 'Low' setting sets* $(h_1, h_2) = (0.01, 0.02)$ *and the 'High' setting sets it to* $(0.1, 0.01)$*. The setting 'Extra Taxon' refers to whether there is an extra taxon present in the data or not.*

Scenario	Extra Taxon	Offspring Density	Bandwidth
1	No	Sparse	Low
2	No	Sparse	High
3	No	Dense	Low
4	No	Dense	High
5	No	Mixed	Low
6	No	Mixed	High
7	Yes	Sparse	Low
8	Yes	Sparse	High
9	Yes	Dense	Low
10	Yes	Dense	High
11	Yes	Mixed	Low
12	Yes	Mixed	High

This is in-line with the idea that the bandwidth parameter value must be low to obtain clustering and the general idea apriori is that there is clustering. We choose the value of σ appropriately to reflect this idea. The choice is made such that 99-th percentile of the half-normal prior would be 10μm, since beyond that, clustering relationship would not exist.

3.2. *Computational Scheme.* We use Markov chain Monte Carlo (MCMC) methods for drawing samples from the posterior distribution. Bayesian implementation of typical NSP models require reversible jump MCMC (Green, 1995) or birth-death-move algorithm (Moller and Waagepetersen, 2003) since the parent locations are unknown. That is not the case here as we know the locations of the parent taxa for all the offspring processes. Modeling HPPs are simple and require no additional tricks either.

We use a Metropolis within Gibbs sampling algorithm here to draw samples from the full conditional distribution of each parameter. With the prior specification in Eq. 3.6, we have closed and conjugate Gamma full conditional densities for each of $\alpha_l\ ,\ l = 1, \ldots, p$; $\lambda_v^C\ ,\ v = p+1, \ldots, p+q$ and $\lambda_j\ ,\ j = p+q+1, \ldots, m$. The full conditional densities for $h_l\ ,\ l = 1, \ldots, p$ are not easy to draw from and requires a Metropolis-Hastings step. We use a Gaussian proposal density with a fixed proposal variance for these with the added restriction that the parameter must be positive. The details of the sampling algorithm, along with the full conditional densities are presented in the Appendix.

4. Simulation Study. We perform simulations studies to benchmark the performance of proposed method. For simplicity, we take the unit square as our window of observations. We generate observations from the model described in Eq. 3.3 with 2 offspring taxa from the same parent taxa. In half of the scenarios we have an additional taxa present in the data while for the other half, there is no extra taxon. Under these models we consider different scenarios by varying offspring density (by controlling the parameters α_1, α_2) and bandwidth (the parameters h_1, h_2). The various scenarios considered in the simulation studies are detailed in Table 1.

The methods we use to analyze these datasets are the proposed method as in Eq. 3.3, the univariate Neyman-Scott model as in Eq. 3.2 for each offspring separately and minimum contrast Diggle (2013) method of estimation which does not use parent location information. For the last method, we use the function `thomas.estK` available in the R package

TABLE 2

Mean absolute percentage bias for estimating parameter values of $\alpha_1, \alpha_2, h_1, h_2$ and λ^C under different settings for the proposed method in multivariate (MM) and univariate (UM) set-up and minimum-contrast (MC) method when there is no extra taxon present in the data. Setting 'Sparse' refers to when both α values are low, while in 'Dense', both of them are high and in 'Mixed' one of them is high while the other one is low. Setting 'Low Bandwidth' means when both h values are small while in 'High Bandwidth' h_1 is high but h_2 is low.

		Sparse			Dense			Mixed		
		MM	UM	MC	MM	UM	MC	MM	UM	MC
Low Bandwidth	α_1	6.29	6.61	159.63	3.34	3.40	21.22	3.23	3.45	223.56
	α_2	7.20	7.41	52.36	7.04	7.37	862.75	3.28	3.52	26.46
	h_1	3.23	3.22	101.69	1.59	1.61	14.11	1.51	1.53	154.08
	h_2	4.73	4.71	2886.59	3.95	3.90	214.01	2.68	2.67	18.40
	λ^C	17.96	17.20	17.96	6.14	6.20	14.80	5.97	6.43	19.46
High Bandwidth	α_1	21.94	23.24	10476.94	34.54	37.63	7769.80	34.57	38.03	8761.49
	α_2	6.62	6.90	314.00	5.35	5.56	47.57	3.37	3.74	145.76
	h_1	69.41	72.92	4732.73	111.50	121.11	499.37	112.38	122.30	508.81
	h_2	3.95	3.97	102.09	3.39	3.38	24.11	1.75	1.75	91.35
	λ^C	18.70	18.29	297.95	5.67	5.66	87.59	6.12	6.17	83.68

TABLE 3

Mean absolute bias for estimating parameter values of $\alpha_1, \alpha_2, h_1, h_2$ and λ^C under different settings for the proposed method in multivariate (MM) and univariate (UM) set-up when there is no extra taxon present in the data. Setting 'Sparse' refers to when both α values are low, while in 'Dense', both of them are high and in 'Mixed' one of them is high while the other one is low. Setting 'Low Bandwidth' means when both h values are small while in 'High Bandwidth' h_1 is high but h_2 is low. Figures in brackets indicate standard error.

		Sparse		Dense		Mixed	
		MM	UM	MM	UM	MM	UM
Low Bandwidth	α_1	0.09(0.11)	0.10(0.11)	0.13(0.16)	0.14(0.16)	0.13(0.16)	0.14(0.15)
	α_2	0.07(0.09)	0.07(0.09)	0.07(0.09)	0.07(0.09)	0.10(0.11)	0.11(0.11)
	h_1	0.03(0.04)	0.03(0.04)	0.02(0.02)	0.02(0.02)	0.02(0.02)	0.02(0.02)
	h_2	0.09(0.11)	0.09(0.11)	0.08(0.10)	0.08(0.10)	0.05(0.07)	0.05(0.07)
	λ^C	35.91(13.30)	34.40(13.52)	12.29(15.41)	12.39(15.35)	11.95(14.29)	12.86(14.20)
High Bandwidth	α_1	0.33(0.35)	0.35(0.37)	1.38(1.53)	1.51(1.54)	1.38(1.58)	1.52(1.66)
	α_2	0.07(0.08)	0.07(0.08)	0.05(0.06)	0.06(0.07)	0.10(0.12)	0.11(0.12)
	h_1	6.94(7.97)	7.29(8.21)	11.15(12.75)	12.11(12.88)	11.24(13.11)	12.23(13.72)
	h_2	0.04(0.05)	0.04(0.05)	0.03(0.04)	0.03(0.04)	0.02(0.02)	0.02(0.02)
	λ^C	37.39(13.52)	36.57(13.56)	11.34(13.88)	11.32(13.92)	12.24(15.32)	12.35(15.33)

`spatstat` (Baddeley and Turner, 2021). The R codes for a Bayesian implementation such as in (Kopeckỳ and Mrkvička, 2016) was not available at the time of this analysis.

We see from the simulation results that the proposed method (denoted as MM) and its univariate counterpart (denoted as UM) work much better than the minimum contrast method (denoted as MC) in terms of relative percent bias in estimating the parameters. In fact, the latter often fail to converge and produce nonsensical result making it extremely unreliable. Not using the parent information for estimation does hamper its performance as was mentioned in Section 1. These results are presented in Table 2 for the case where there is no extra taxon present in the data. The standard errors for the estimates obtained using the proposed method (MM) and its univariate counterpart (UM) are presented in Table 3. The corresponding results for the cases where there was an extra taxon present in the data are pushed to the Appendix as they tell a similar story and the estimates are also very similar.

The relatively poor performance for the method in the high bandwidth setting for estimating α_1 and h_1 is expected as in that scenario the clusters are very dispersed and therefore the essential assumption of clustering for these process gets violated. This phenomenon is quite

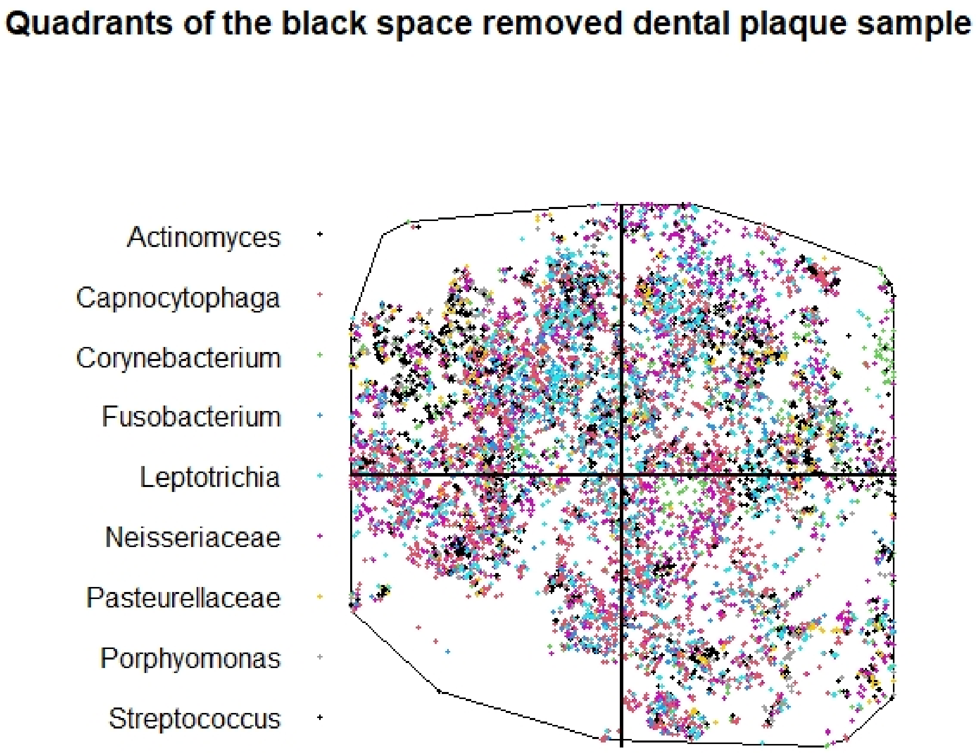

Fig 5: Division of the locations of the dental plaque sample image in first (bottom left), second (top left), third (bottom right) and fourth (top right) segments.

common in cluster data analysis and the minimum contrast estimation also suffers from this. However, the proposed model performs much better in even that scenario compared to the minimum contrast method.

5. Analysis of the Human Dental Plaque Sample. The clustering of *Streptococcus* and *Porphyomonas* around *Corynebacterium* only happen on the tips (Mark Welch et al., 2016). This brings an inhomogeneity in the dataset as we see both tip and stem of the *Corynebacterium* in Figure 3. This requires careful manual subsetting of the dataset to successfully apply the method. Another important aspect is the black space in the image which needs to be removed. Instead of using the image as it is, we use a convex hull of the observed locations as our analysis window to reduce black space. Because we lack a more sophisticated subsetting method, we simply subset the image in four segments as presented in Figure 5 and use the method on the segments separately.

TABLE 4

Estimates for parameters associated with Corynebacterium (λ_C), Streptococcus (α_1, h_1), Porphyomonas (α_2, h_2) and Pasteurellaceae (α_3, h_3) obtained by applying the proposed method on each of the four segments of the dental plaque sample image. The figures in parenthesis denote the corresponding standard errors. All results are rounded to three decimal places.

Segment	α_1	α_2	α_3	h_1	h_2	h_3	λ_C
I	2.215 (0.224)	5.05 (0.341)	0.571 (0.079)	8.315 (0.563)	7.234 (0.5)	3.823 (0.4)	0.006 (0.001)
II	1.976 (0.102)	2.907 (0.132)	0.355 (0.031)	11.232 (0.578)	15.211 (0.791)	4.230 (0.464)	0.021 (0.001)
III	1.039 (0.083)	1.797 (0.113)	0.521 (0.06)	7.619 (0.706)	9.830 (0.722)	4.495 (0.43)	0.017 (0.001)
IV	1.215 (0.079)	2.105 (0.106)	0.457 (0.045)	9.413 (0.812)	11.087 (0.721)	4.624 (0.422)	0.024 (0.002)

TABLE 5

Estimates for parameters associated with Neisseriaceae (λ_1), Capnocytophaga (λ_2), Actinomyces (λ_3), Fusobacterium (λ_4) and Leptotrichia (λ_5) obtained by applying the proposed method on each of the four segments of the dental plaque sample image. The figures in parenthesis denote the corresponding standard errors. All results are rounded to three decimal places.

	λ_1	λ_2	λ_3	λ_4	λ_5
Segment 1	0.037 (0.002)	0.056 (0.002)	0.013 (0.001)	0.010 (0.001)	0.021 (0.002)
Segment 2	0.045 (0.002)	0.072 (0.003)	0.027 (0.002)	0.024 (0.001)	0.039 (0.002)
Segment 3	0.037 (0.002)	0.054 (0.002)	0.014 (0.001)	0.013 (0.001)	0.022 (0.001)
Segment 4	0.048 (0.002)	0.056 (0.002)	0.022 (0.001)	0.017 (0.001)	0.033 (0.002)

The estimates and their standard errors (rounded to three decimal places) for the four different segments are presented in the Tables 4 and 5. The trace plots for the Markov chains show good mixing and hence well convergence. They are presented in the supplementary materials.

We see from Tables 4 and 5 that the relationship varies over different quadrants of the image. But so does the abundance of *Corynebacterium* and proportion of *Corynebacterium* stem regions. Segment 1 has the least proportion of stem region of *Corynebacterium* but the overall abundance is also low. The estimates of 8.32μm bandwidth for *Streptococcus* and 7.23μm bandwidth for *Porphyomonas* seem to be most reliable. However, they are still a bit larger than one would expect them to be as these estimates tend to imply clustering on the weaker side. A proper manual subsetting of the tip region of the *Corynebacterium* taxon only would help the analysis and make the estimates more reliable and perhaps smaller. The estimated bandwidths of $3.82 - 4.62\mu$m for *Pasteurellaceae* seem to be close to what is typically observed and expected for the taxon being clustered around *Streptococcus*.

The estimates for α_1 and α_2 also seem to be affected by the presence of the stem region of *Corynebacterium*. Higher stem regions with no offspring around them forces a lower average number of offspring per parent to be estimated in Segments 2, 3 and 4. The estimate of λ_C being so small across board may seem problematic, but λ_C represents the average number of parents per square unit area ($1\mu\text{m}^2$ in our case). Given the size of each quadrant being $106.275\mu\text{m} \times 106.275\mu\text{m}$ and the number of observed *Corynebacterium* taxon, these estimates make sense. The same is true for estimates of λ_1 through λ_5 in Table 5.

6. Discussion and Conclusion. We propose a novel method for quantifying various relationships between different taxa present in a microbiome sample. The proposed method makes use of the parent information, typically present in microbiome data, for the Neyman-Scott models which makes the inference better and computationally easier. It clearly outperformed the much used minimum contrast method for Thomas process models in every scenario in the simulation study. The anticipated behaviors of parent-offspring relationship is captured in the real data analysis. However, better subsetting methods would make the estimates more reliable.

TABLE 6

Mean absolute percentage bias for estimating parameter values of $\alpha_1, \alpha_2, h_1, h_2$ and λ^C under different settings for the proposed method in multivariate (MM) and univariate (UM) set-up and minimum-contrast (MC) method when there is one extra taxon present in the data. Setting 'Sparse' refers to when both α values are low, while in 'Dense', both of them are high and in 'Mixed' one of them is high while the other one is low. Setting 'Low Bandwidth' means when both h values are small while in 'High Bandwidth' h_1 is high but h_2 is low.

		Sparse			Dense			Mixed		
		MM	UM	MC	MM	UM	MC	MM	UM	MC
Low Bandwidth	α_1	5.72	6.35	544.26	2.98	3.13	637.41	3.04	3.22	17.04
	α_2	7.55	8.19	116.50	6.07	6.27	224.11	3.90	4.19	101.32
	h_1	2.91	2.93	113.59	1.51	1.50	115.94	1.41	1.44	15.86
	h_2	4.99	5.04	47.03	4.65	4.60	60.50	2.57	2.57	44.97
	λ^C	19.47	17.13	14.39	5.49	5.75	17.72	5.67	6.19	15.74
High Bandwidth	α_1	20.31	24.59	10691.11	35.59	39.12	6896.32	35.45	38.40	10630.23
	α_2	7.09	7.63	47.07	6.15	6.98	65.93	3.54	3.82	172.42
	h_1	69.63	80.54	3828.96	114.89	124.86	402.17	116.67	123.72	487.02
	h_2	3.44	3.45	24.24	3.37	3.37	40.41	1.84	1.84	84.97
	λ^C	19.87	18.37	258.27	6.18	6.19	77.22	5.74	5.72	73.35

As mentioned in Section 3, the proposed model is flexible and can accommodate different types of models and their combinations. We discussed how we can incorporate homogeneous Poisson processes and Neyman-Scott processes in the model. Choosing the form of $\lambda_i(\mathbf{s}), i = 1, \ldots, m$, we can set this model to be an LGCP or a multivariate LGCP model. We can in fact replace the HPP components of the proposed model in Eq. 3.3 by multivariate LGCP component to make more sophisticated model that may better capture the correlation between various taxa. Such improvements over the model in Eq. 3.3 and usage of the general model in Eq. 3.1 in other scenarios is a future avenue for such works.

The proposed method can benefit from better subsetting methods being available. It boasts a possibility of being used in a variety of microbiome data set-up. An R package for the proposed method would be very helpful to the practitioners. One important aspect of microbiome data is that they may come from multiple donors in multiple sittings. An appropriate meta-analysis method is required to apply this method to each of the datasets and consolidate the individual estimates to a single estimate for more general purpose use.

APPENDIX A: ADDITIONAL IMAGES OF THE DENTAL PLAQUE DATA

Additional images for distribution of *Neisseriaceae*, *Capnocytophaga*, *Actinomyces*, *Fusobacterium* and *Leptotrichia* in the human dental plaque sample are presented here in Figure 6. There are no particular pattern in any of them (except Eubacterium) and therefore they were modeled as homogeneous Poisson process in the data analysis. Eubacterium was used for probing and therefore takes the shape of every available structure. It was excluded from the analysis.

APPENDIX B: ADDITIONAL TABLES FROM SIMULATION STUDY

here we present results from the simulation study of cases where there was an additional taxon present in the data. The results in Tables 6 and 7 tell a similar story as in Section 4. The proposed method (MM) and its univariate counterpart (UM) perform much better compared to minimum contrast method (MC) in terms of relative percent bias in estimating the model parameters. All the models perform poorly when the bandwidth is high. However, the proposed model works much better compared to the minimum contrast method.

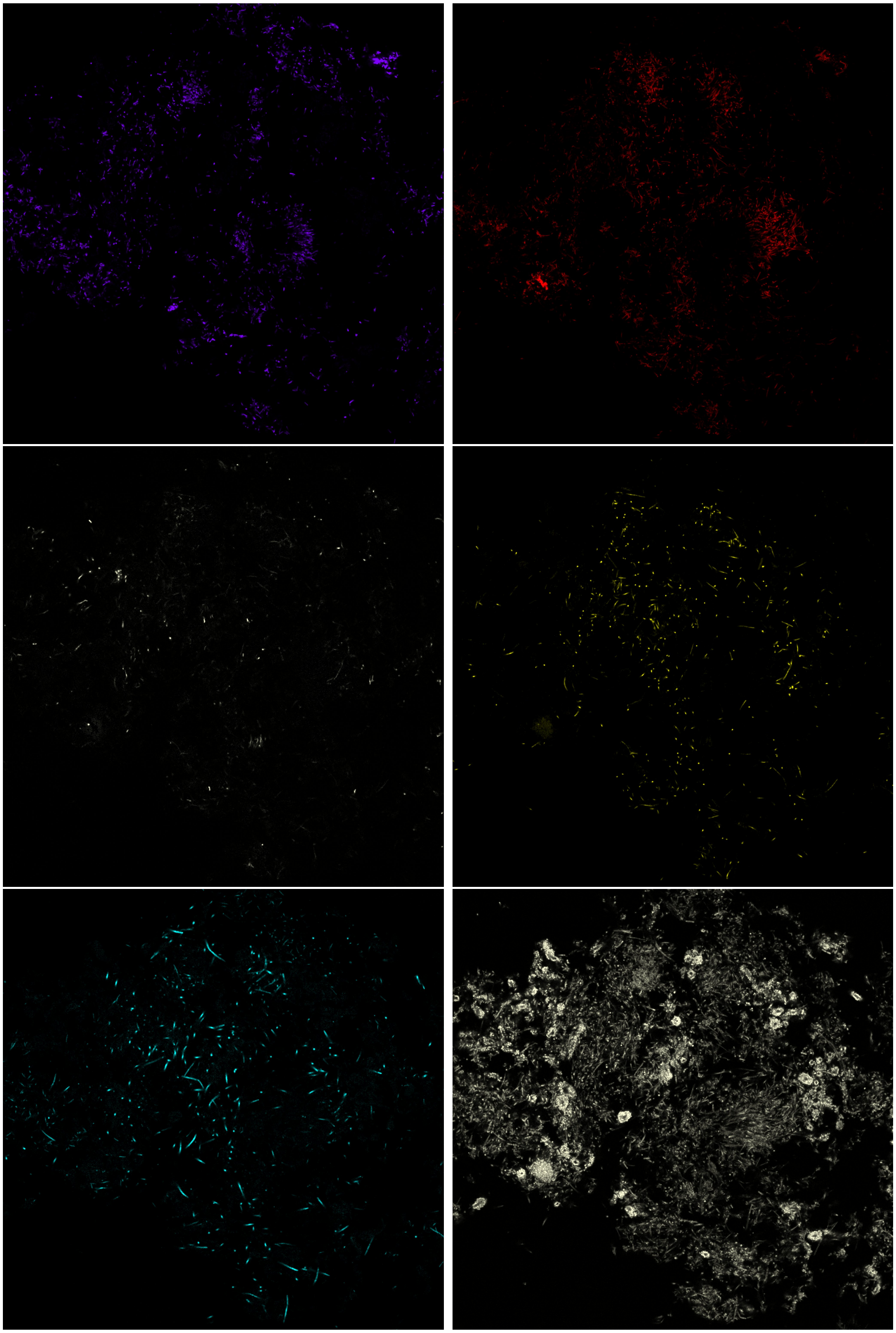

Fig 6: RGB images of *Neisseriaceae* (top left), *Capnocytophaga* (top right), *Actinomyces* (middle left), *Fusobacterium* (middle right), *Leptotrichia* (bottom left) and *Eubacterium* (bottom right) distribution in the dental plaque sample. *Eubacterium* was used for probing and hence is left out of the analysis. The other taxa presented here are scattered randomly over the region.

TABLE 7

Mean absolute bias for estimating parameter values of $\alpha_1, \alpha_2, h_1, h_2$ and λ^C under different settings for the proposed method in multivariate (MM) and univariate (UM) set-up when there is one extra taxon present in the data. Setting 'Sparse' refers to when both α values are low, while in 'Dense', both of them are high and in 'Mixed' one of them is high while the other one is low. Setting 'Low Bandwidth' means when both h values are small while in 'High Bandwidth' h_1 is high but h_2 is low. Figures in brackets indicate standard error.

		Sparse		Dense		Mixed	
		MM	UM	MM	UM	MM	UM
Low Bandwidth	α_1	0.09(0.10)	0.10(0.10)	0.12(0.14)	0.13(0.14)	0.12(0.15)	0.13(0.15)
	α_2	0.08(0.09)	0.08(0.09)	0.06(0.08)	0.06(0.08)	0.12(0.13)	0.13(0.14)
	h_1	0.03(0.04)	0.03(0.04)	0.02(0.02)	0.01(0.02)	0.01(0.02)	0.01(0.02)
	h_2	0.10(0.12)	0.10(0.12)	0.09(0.12)	0.09(0.12)	0.05(0.06)	0.05(0.06)
	λ^C	38.94(12.21)	34.26(12.43)	10.98(14.10)	11.50(14.10)	11.34(14.27)	12.38(14.45)
High Bandwidth	α_1	0.30(0.31)	0.37(0.35)	1.42(1.56)	1.56(1.60)	1.42(1.59)	1.54(1.65)
	α_2	0.07(0.09)	0.08(0.09)	0.06(0.07)	0.07(0.07)	0.11(0.14)	0.11(0.14)
	h_1	6.96(7.33)	8.05(8.14)	11.48(13.04)	12.49(13.41)	11.67(13.33)	12.37(13.76)
	h_2	0.03(0.04)	0.03(0.04)	0.03(0.04)	0.03(0.04)	0.02(0.02)	0.02(0.02)
	λ^C	39.74(12.58)	36.74(12.66)	12.37(15.95)	12.38(15.90)	11.49(14.47)	11.44(14.39)

APPENDIX C: COMPUTATIONAL DETAILS OF THE SAMPLING ALGORITHM

Acknowledgments. The authors would like to thank ...

Funding. The first author was supported by ...

The second author was supported in part by ...

REFERENCES

BADDELEY, A. and TURNER, R. (2021). Package 'spatstat'.

CHIU, S. N., STOYAN, D., KENDALL, W. S. and MECKE, J. (2013). *Stochastic geometry and its applications*. John Wiley & Sons.

DIGGLE, P. J. (2013). *Statistical analysis of spatial and spatio-temporal point patterns*. CRC press.

DIGGLE, P. J., MORAGA, P., ROWLINGSON, B. and TAYLOR, B. M. (2013). Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. *Statistical Science* **28** 542–563.

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732.

GUAN, Y. (2006). A composite likelihood approach in fitting spatial point process models. *Journal of the American Statistical Association* **101** 1502–1512.

GUTTORP, P. and THORARINSDOTTIR, T. L. (2012). Bayesian inference for non-Markovian point processes. In *Advances and Challenges in Space-time Modelling of Natural Events* 79–102. Springer.

ILLIAN, J., PENTTINEN, A., STOYAN, H. and STOYAN, D. (2008). *Statistical analysis and modelling of spatial point patterns* **70**. John Wiley & Sons.

JALILIAN, A., GUAN, Y., MATEU, J. and WAAGEPETERSEN, R. (2015). Multivariate product-shot-noise Cox point process models. *Biometrics* **71** 1022–1033.

KOPECKỲ, J. and MRKVIČKA, T. (2016). On the Bayesian estimation for the stationary Neyman-Scott point processes. *Applications of Mathematics* **61** 503–514.

MARK WELCH, J. L., ROSSETTI, B. J., RIEKEN, C. W., DEWHIRST, F. E. and BORISY, G. G. (2016). Biogeography of a human oral microbiome at the micron scale. *Proceedings of the National Academy of Sciences* **113** E791–E800.

MØLLER, J. (2003). Shot noise Cox processes. *Advances in Applied Probability* **35** 614–640.

MØLLER, J., SYVERSVEEN, A. R. and WAAGEPETERSEN, R. P. (1998). Log gaussian cox processes. *Scandinavian journal of statistics* **25** 451–482.

MOLLER, J. and WAAGEPETERSEN, R. P. (2003). *Statistical inference and simulation for spatial point processes*. CRC Press.

MØLLER, J. and WAAGEPETERSEN, R. P. (2007). Modern statistics for spatial point processes. *Scandinavian Journal of Statistics* **34** 643–684.

MRKVIČKA, T., MUŠKA, M. and KUBEČKA, J. (2014). Two step estimation for Neyman-Scott point process with inhomogeneous cluster centers. *Statistics and Computing* **24** 91–100.

NEYMAN, J. and SCOTT, E. L. (1958). Statistical approach to problems of cosmology. *Journal of the Royal Statistical Society: Series B (Methodological)* **20** 1–29.

PERERA, D., MCLEAN, A., LOPEZ, V. M., CLOUTIER-LEBLANC, K., ALMEIDA, E., CABANA, K., MARK WELCH, J. L. and RAMSEY, M. M. (2020). Mechanisms underlying proximity between oral commensal bacteria. *bioRxiv*.

TANAKA, U., OGATA, Y. and STOYAN, D. (2008). Parameter estimation and model selection for Neyman-Scott point processes. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **50** 43–57.

TANAKA, U. and OGATA, Y. (2014). Identification and estimation of superposed Neyman–Scott spatial cluster processes. *Annals of the Institute of Statistical Mathematics* **66** 687–702.

WAAGEPETERSEN, R. P. (2007). An estimating function approach to inference for inhomogeneous Neyman–Scott processes. *Biometrics* **63** 252–258.