Multilayer Adjusted Cluster Point Process Model: Application to Microbial Biofilm Image Data Analysis

Suman Majumder
Harvard T. H. Chan School of Public Health, Boston, Massachusetts, U.S.A.

Brent A. Coull
Harvard T. H. Chan School of Public Health, Boston, Massachusetts, U.S.A.

Jessica L. Mark Welch
Marine Biological Laboratory, Woods Hole, Massachusetts, U.S.A.

Patrick J. La Riviere
University of Chicago, Chicago, Illinois, U.S.A.

Floyd E. Dewhirst
The Forsyth Institute, Cambridge, Massachusetts, U.S.A.

Jacqueline R. Starr *
Brigham and Women’s Hospital, Boston, Massachusetts, U.S.A.

Kyu Ha Lee *
Harvard T. H. Chan School of Public Health, Boston, Massachusetts, U.S.A.

February 10, 2022

*Co-Senior Author
Abstract

A common problem in spatial statistics tackles spatial distributions of clusters of objects. Such clusters of similar or dissimilar objects are encountered in many fields, including field ecology, astronomy, and biomedical imaging. Still challenging is to quantify spatial clustering when one or more entities clusters around a different entity in multiple layers. Such multi-entity and multi-layered structures are observed, for example, in human dental plaque biofilm images, which exhibit multi-species structures in corncob-like arrangements. We propose a novel, fully Bayesian, multivariate spatial point process model to quantify corncob-like arrangements with “parent-offspring” statistical approaches. The proposed multilayer adjusted cluster point process (MACPP) model departs from commonly used approaches in that it exploits the locations of the central “parent” object in clusters and accounts for multilayered multivariate parent-offspring clustering. In simulated datasets, the MACPP outperforms the classical Neyman-Scott process model, a univariate model for modeling spatially clustered processes, by producing decisively more accurate and precise parameter estimates. We analyzed data from a human dental plaque biofilm image in which Streptococcus and Porphyromonas simultaneously cluster around Corynebacterium and Pasteurellaceae clusters around Streptococcus. The proposed MACPP model successfully captured the parent-offspring structure for all the taxa involved.

Keywords: dental plaque sample, imaging, microbiome, parent-offspring models, point processes, spatial statistics, Thomas process

1 Introduction

Nature abounds in instances where one type of object is dispersed around another type of object. At interplanetary and field ecologic scales, respectively, radio galaxies (Yates et al., 1989; Hill and Lilly, 1991) (Bornancini et al., 2004, 2006) and forest songbirds (Tarof and Ratcliffe, 2004; Bourque and Desrochers, 2006; Melles et al., 2009) exemplify this type of spatial clustering. At microscopic scales, among single-celled organisms and human cells, such arrangements are common. For example, in human dental plaque biofilm, spherical Streptococcus cells often cluster around the ends of filamentous Corynebacterium cells (Jones, 1972; Mark Welch et al., 2016). These corncob-like arrangements have been observed in dental plaque for decades and are likely to hold clues about interspecies microbial interactions. To describe and quantify such arrangements, we have developed a new multilayer adjusted cluster point process model (MACPP).

Most biofilm image analyses have focused primarily on macro-level structural characteristics or derived features, such as biofilm volume, thickness, or surface roughness (Vorregaard, 2008). Less commonly, spatial point process models have been used to analyze the spatial patterning of microbial cells: how cells of one taxonomic class (or taxon) are distributed in relation to other cells of the same or different taxa. However, standard point process models, such as the log-Gaussian Cox process model (Møller et al., 1998), do not account for the complex spatial clustering arrangements often present in biofilm images.
The Neyman-Scott process model (NSP) (Neyman and Scott, 1958) is a classic statistical model used to quantify spatial clustering relationships, most often applied to true parent-offspring clusters (such as trees in a forest). In this approach, locations of the central object in each cluster are treated as latent, in part because in a cluster comprising similar objects (such as a grove of trees or a school of fish), it may be impossible to identify any true “parent.” In contrast, in corncob arrangements in dental plaque biofilm, the locations of the “parent” Corynebacterium cells are often known. While reasonable for within-taxon clustering, naïve application of the NSP to investigate between-taxon relationships (i.e. one taxon clustering around the other) is inappropriate because it ignores the taxon in the center of the cluster.

Corncob arrangements in dental plaque biofilm also have other characteristics that preclude direct application of the NSP or the related shot noise Cox process model (Møller, 2003). First, the “parent” and “offspring” cells are of different bacterial taxa, thus requiring a multivariate extension of univariate approaches. Second, the corncob arrangements sometimes include multiple “offspring” taxa in that both Streptococcus and Porphyromonas are observed around Corynebacterium “parents.” In this regard, existing multivariate cluster point process models (Tanaka and Ogata, 2014; Jalilian et al., 2015) are not applicable to microbial corncob-like arrangements because they do not model the taxa in the cluster center and have no scope to enforce multiple offspring taxa having the same parent taxon. Moreover, neither of the existing approaches address a third challenge, that the corncobs can be multilayered in nature (e.g., Figure 1). And, fourth, the corncob arrangements themselves are part of a more complexly organized microbial community that includes other taxa unrelated to the arrangements.

The primary innovation of our proposed MACPP model is that it exploits the known locations of central objects in clusters, and it also addresses the above-described challenges. Specifically, it can quantify multivariate and multilayered clustering and inter-process relationships. The MACPP is flexible in that it can simultaneously model clustered and non-clustered processes and can be applied in multivariate or univariate contexts. In this paper, we describe the model in detail, evaluate its performance through simulation studies, and demonstrate the feasibility of applying the model to both real and synthetic datasets.
Figure 2: A biofilm image of a dental plaque sample from a human donor: RGB image - taxa relevant to the multilayered concob arrangements are Corynebacterium (Pink), Streptococcus (Green), Porphyromonas (Blue) and Pasteurellaceae (Orange); Spatial locations for the centroids of the taxa identified from the segmented image.

2 Microbiome Biofilm Image Data from Dental Plaque Samples

Dental plaque samples were collected from a donor who was asked not to smoke, floss, or brush for 12 hours and not to eat for 2 hours prior to collection. The sample was suspended in ethanol, an aliquot mounted on a glass slide, and dried. Multi-spectral fluorescence in situ hybridization (FISH) imaging was applied to the microbiota of the dental plaque sample by including multiple genus-level probes to identify several distinct microbial taxa simultaneously: Actinomyces, Capnocytophaga, Corynebacterium, Fusobacterium, Leptotrichia, Neisseriaceae, Pasteurellaceae, Porphyromonas, Streptococcus, and Eubacterium (Figure 2(a)). We performed segmentation of the biofilm image in FIJI (Schindelin et al., 2012) by applying the $3 \times 3$ median filter and “Auto Local” thresholding function with the Bersen method. Spatial coordinate information for centroids of the taxa were identified by applying the “Analyze Particles” function using size filter with a 0.5 $\mu$m diameter threshold in FIJI (Figure 2(b)). More details on collection and processing of dental plaque, fixation on a microscopy slide, and multi-spectral FISH imaging process are described in Mark Welch et al. (2016).

The sampled image is representative of similar samples from the same and other donors without active tooth decay (not shown). Filamentous Corynebacterium cells (Figure 3, pink) are among the most common. Members of this genus are now believed to help establish the “healthy” dental microbiota, providing a scaffold around which other community members assemble. In some parts of the image, cells of Streptococcus (Figure 3, green) and
Figure 3: RGB images of only Corynebacterium (Pink), Streptococcus (green) and Porphyromonas (Blue) (a) in the entire dental plaque sample; (b) in a sub-region of the sample focusing on the corncob structure. The corncob structures are present in areas of the image corresponding to the tip of the Corynebacterium. The central portion of the entire image in (a) consisting of stems of the Corynebacterium are relatively empty without any taxa surrounding them.

Porphyromonas (Figure 3, blue) seem to surround the tips of Corynebacterium filaments (Mark Welch et al., 2016), sometimes both in the same clusters. Adding another layer and further complexity, Pasteurellaceae cells (Figure 4, orange) sometimes surround Streptococcus cells (Mark Welch et al., 2016; Perera et al., 2020). As outlined in Section 1, these arrangements give Streptococcus the potential to be either parent, offspring, or both in the same cluster. Other taxa are scattered around seemingly homogeneously and may or may not have additional spatial relationships with Corynebacterium, Streptococcus, Porphyromonas, or Pasteurellaceae (individual images in Section A of the Supplementary Materials).

3 Multilayer Adjusted Cluster Point Process Model

We consider a multivariate process $Y = \bigcup_{i=1}^{m} Y_i$, $i = 1, \ldots, m$, at location $s \in \mathcal{W} \subset \mathbb{R}^2$, where each component $Y_i$ is a Poisson process characterized by an intensity function $\lambda_i(s)$ and $\mathcal{W}$, the observation window. In application to the biofilm sample, each $Y_i$ will capture the spatial distribution of a taxon, $i$. The intensity function $\rho(s)$ of the superimposed process is defined as

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

(1)
Figure 4: RGB images of *Pasteurellaceae* (Orange) clustering around *Streptococcus* (Green) (a) in the entire human dental plaque sample; (b) in a sub-region of the sample focusing on the clustering.

where \( \mathbb{I}(s \in Y_i) \) is the indicator function of the \( i \)-th process being observed at location \( s \in W \). The indicator terms ensure that only one process can be observed at each location.

### 3.1 Model Formulations

Suppose that \( Y_1, \ldots, Y_p \) are homogeneous Poisson processes with intensities \( \lambda_v^C \) for \( v = 1, \ldots, p \) \((< m)\) and serve only as parent processes. Then we consider the \( q \) \((\leq m - p)\) processes, \( Y_{p+1}, \ldots, Y_{p+q} \), that behave as offspring processes. Furthermore, let \( C_l \) for \( l = p + 1, \ldots, p + q \) denote the corresponding parent processes with the following properties:

i) \( C_l = C_{l'} \) for some \( l \neq l' \), implying multiple offspring processes can share a parent process;

ii) \( C_l = Y_j \) for some \( l \) and \( j = 1, \ldots, p + q \), implying some offspring processes can function as parent processes.

We assume that the offspring points are distributed around parent point \( c \in C_l, l \in \{p + 1, \ldots, p + q\} \) according to the rule \( \alpha_l k_l(\cdot - c, h_l) \), where \( \alpha_l \) is the average number of offspring per parent, \( k_l(\cdot, \cdot) \) is a kernel (e.g. Gaussian), and \( h_l \) is a bandwidth parameter that controls the distance between the parent and its offspring locations for the \( l \)-th offspring process (Chiu et al., 2013; Illian et al., 2008). The parameter \( h_l \) should be interpreted as the distance at which offspring are most likely to be situated relative to their corresponding parents. The remaining \( m - p - q \) types (e.g. taxa) that are unrelated to multilayered arrangements are modeled as homogeneous Poisson processes with intensities \( \lambda_j \) for \( j = p + q + 1, \ldots, m \).
Therefore, under the proposed specification, the intensity function in (1) can be written as
\[
\rho(s) = \sum_{v=1}^{p} \lambda^C_v I(s \in Y_v) + \sum_{l=p+1}^{p+q} \alpha_l \sum_{c_l \in C_l} k_l(s - c_l, h_l) I(s \in Y_l) + \sum_{j=p+q+1}^{m} \lambda_j I(s \in Y_j). \tag{2}
\]

Since the superimposed process \(Y\) is modeled as a Poisson process with intensity function \(\rho(s)\) in (2), the likelihood as a function of the unknown parameters \(\theta = \{\alpha_{p+1}, \ldots, \alpha_{p+q}, h_{p+1}, \ldots, h_{p+q}, \lambda^C_1, \ldots, \lambda^C_p, \lambda_{p+q+1}, \ldots, \lambda_m\}\), is given by
\[
L(Y|\theta) = \exp \left\{ \int_{\mathcal{W}} 1 - \rho(u) du \} \prod_{y \in Y} \rho(y) \right. \\
\times \exp \left\{ -|\mathcal{W}| - \sum_{v=1}^{p} |\mathcal{W}| \lambda^C_v - \sum_{l=p+1}^{p+q} \alpha_l \sum_{c_l \in C_l} \int_{\mathcal{W}} k_l(u - c_l, h_l) du - \sum_{j=p+q+1}^{m} |\mathcal{W}| \lambda_j \right\} \\
\times \exp \left\{ \sum_{v=1}^{p} n_v \log \lambda^C_v + \sum_{l=p+1}^{p+q} \sum_{y \in Y_l} \log \left( \alpha_l \sum_{c_l \in C_l} k_l(y - c_l, h_l) \right) + \sum_{j=p+q+1}^{m} n_j \log \lambda_j \right\}, \tag{3}
\]

where \(|\mathcal{W}|\) is the area of \(\mathcal{W}\) and \(n_i\) denotes the number of observations from the \(i\)-th process in \(\mathcal{W}\).

### 3.2 Prior Distributions and Practical Considerations

We outline priors for the unknown model parameters to complete the Bayesian specification of the MACPP. Specifically, We consider the following priors:
\[
\begin{align*}
\alpha_l & \sim^{iid} \text{Gamma}(a_Y, b_Y), \ l = p + 1, \ldots, p + q, \\
h_l & \sim^{iid} \text{Half-Normal}(\sigma), \ l = p + 1, \ldots, p + q, \\
\lambda^C_v & \sim^{iid} \text{Gamma}(a_C, b_C), \ v = 1, \ldots, p, \\
\lambda_j & \sim^{iid} \text{Gamma}(a, b), \ j = p + q + 1, \ldots, m,
\end{align*} \tag{4}
\]

where \(\sim^{iid}\) denotes independent and identically distributed, and \((a, b, a_Y, b_Y, a_C, b_C, \sigma)\) are hyperparameters to be specified.

It is well appreciated in the literature on point process models that using a noninformative prior for spatial bandwidth or scale parameters is impossible due to numerical reasons (Moller and Waagepetersen, 2003; Diggle, 2013; Kopecký and Mrkvička, 2016). We use a half-normal prior for the bandwidth parameters \(h_l\); ensuring the bulk of the mass remains on relatively small positive values enables comparatively easier prior elicitation. Specifically, the value of hyperparameter \(\sigma\) can be set such that the 99-th percentile of the half-normal prior corresponds to the maximum extent of the distance \(d\) between a parent point and offspring.
points. This choice helps sensitize the proposed Bayesian framework to stronger clustering within a small radius, the type of clustering we seek to detect and quantify when, for example, cells of two bacterial taxa directly interact. In general, choosing the value of \( d \) should depend on the context of application.

### 3.3 Computational Scheme

Combining (3) and (4), the joint posterior density for the proposed MACPP can be written as

\[
\pi(\theta | Y) \propto L(Y | \theta) \pi(\theta),
\]

where \( \pi(\theta) \) is the product of all the individual prior densities. We then proceed to draw samples from the posterior distribution by using a Markov chain Monte Carlo (MCMC) algorithm. More conventional NSP-type approaches treat the number of parent points and their locations as random and thus require an additional reversible jump MCMC step to estimate the parameters associated with the latent parent process (Green, 1995; Moller and Waagepetersen, 2003). In contrast, the proposed framework exploits the observed parent locations and proceeds without a complex birth-death-move algorithm. The components of \( \theta \) can then be updated by Gibbs sampling (exploiting conjugacies in the full conditionals) or via Metropolis-Hastings steps (details in Section B of the Supplementary Material).

One practical challenge is that the integral term in (3) does not have a closed-form expression. Therefore, we use Monte Carlo methods to approximate the integral for computational efficiency. In practice, the expression \( \int_W k_l(u - c_l, h_l)du \) can be thought of as the probability of occurrence of \( X_l \) within the observation window \( W \), where \( X_l \) is a bivariate real-valued random variable with density \( k_l(\cdot - c_l, h_l) \). Furthermore, in the proposed framework with Thomas processes, \( k_l(\cdot - c_l, h_l) \) corresponds to a bivariate normal density function with mean \( c_l \) and covariance matrix \( h_l^2 I \). Thus, we draw samples from the bivariate normal distribution and compute the average proportion of points that fall within \( W \), which serves as an approximation to the integral term in (3).

We further optimized the code by using C language based coding. An user-friendly version of the R software package is available in the online repository ([https://github.com/SumanM47/PO_Modeling.git](https://github.com/SumanM47/PO_Modeling.git)). The implementation of the model in the R package can generate 10,000 scans in 1.5 minutes for a dataset with \( \sim 150 \) points from a single parent process with a total of \( \sim 750 \) points from two offspring processes on a Dell Latitude 7210 laptop with i5 cores and 16 gigabytes of memory.

### 3.4 Model validation

We assess goodness-of-fit of the models by comparing the observed and estimated counts of different entities. Specifically, for the parent species, the estimated intensity parameter \( \lambda^C_v, v = 1, \ldots, p \), per unit area can be scaled by the area of the observation window; \( \lambda^C_v | W \) should be close to the number of the \( v \)-th parent cells counted in the observation window. For approximating the \( l \)-th offspring count, one can compute \( \alpha_l \sum_{c \in C_l} \int_W k_l(u - c, h_l)du \) for \( l = p + 1, \ldots, p + q \). The integral term can be approximated by Monte Carlo method
as described in Section 3.3. This expression involves both the offspring density parameter and the bandwidth parameter and therefore helps validate the joint estimate of the $(\alpha_l, h_l)$ pair. In our example, $\lambda^C|W|$ should provide a good estimate of the number of *Corynebacterium* cells; and $\alpha_2 \sum_{c \in C_2} \int_W k_2(u - c, h_2)du$, $\alpha_3 \sum_{c \in C_3} \int_W k_3(u - c, h_3)du$ and $\alpha_4 \sum_{c \in C_4} \int_W k_4(u - c, h_4)du$ should provide good estimates of the observed number of *Streptococcus* clustered around *Corynebacterium*, the observed number of *Porphyromonas* clustered around *Corynebacterium*, and the observed number of *Pasteurellaceae* clustered around *Streptococcus*, respectively.

4 Simulation Studies

4.1 Simulation Set-up

We performed simulation studies to benchmark the performance of the MACPP. For simplicity, the unit square was taken as the observation window. We generated data under the model outlined in Section 3, with $q = 2$ offspring taxa $B$ and $C$ around the same parent taxon $A$, which was the only parent taxon ($p = 1$). We considered twelve data scenarios, in Table 1, that varied in terms of offspring density $(\alpha_2, \alpha_3)$, bandwidth $(h_2, h_3)$, and presence of a taxon spatially unrelated to the multilayered arrangement (hereinafter referred to as “unrelated” taxon). Throughout the scenarios, the intensity of the parent process $(\lambda^C_1)$ and that of the process for the unrelated taxon $(\lambda^4)$ were set to 150 and 95, respectively. For each of the twelve scenarios, we generated 100 images, each analyzed as an independent dataset.

4.2 Analysis Plan and Hyperparameters

We applied the following three approaches to each simulated dataset:

(i) *The proposed MACPP model.* Multilayered taxa $(A, B, C)$ were jointly analyzed in one framework.

(ii) *A set of two univariate MACPP models.* Two separate group of taxa $(A, B)$ and $(A, C)$ were analyzed respectively.

(iii) *A set of two univariate NSP models.* Ignoring the parent taxon $A$, only offspring taxa $B$ and $C$ were analyzed, separately. We applied the method of minimum contrast (Diggle, 2013), using the R package spatstat (Baddeley et al., 2015).

The parameters estimated by the three approaches have different interpretations. For example, the bandwidth parameter $h_2$ in NSP models from the analysis (iii) is interpreted as the distance scale for the unobserved cluster formed by the offspring taxon $B$, ignoring the observed parent taxon $A$. Despite this distinction, we included analysis (iii) using the NSP because it is among the most relevant cluster point process models applied to this class of problems. In simulation studies, we primarily focused on the numerical performance of the methods in estimating the parameters rather than on their interpretations.

For analyses (i) and (ii), we set the hyperparameters $(a_Y, b_Y, a_C, b_C, a, b)$ to 0.01. The hyperparameter $\sigma$ was set to 0.02 so that the 99-th percentile of the prior distribution of $h_l$’s
was approximately 0.05 (i.e., 5% of the length of the observation window). The posterior means were used as estimates of parameters.

We estimated the offspring intensity ($\alpha_2, \alpha_3$) and bandwidth parameters ($h_2, h_3$) for the offspring processes (taxa $B$ and $C$) and the intensity parameter ($\lambda_{T'}$) for the parent process (taxon $A$). For the MACPP, the estimates are the average of the posterior mean for the parameters for each of the datasets in a given scenario. We also computed the posterior standard deviation (SD) averaged over 100 datasets for each scenario as well as the empirical standard error (SE) of the estimates for the different datasets. With the NSP, no uncertainty measure is available for the individual estimates for each dataset, so we computed only the SE for these estimates between different datasets. However, the NSP method failed to converge in multiple scenarios. We report the percentage of instances where estimates of parameters are not reliable due to the numerical issue for each scenario. Therefore, for the NSP, the average estimates and SE were computed based only on the datasets in which the model converged.

### 4.3 Primary Results

Results were very similar between multivariate and univariate MACPP models. Hereafter, we refer to tabular results from the multivariate MACPP analyses. The MACPP method performed well in estimating the true parameter values with both a small SD and a small SE (Tables 2 and C.1 of Supplementary Materials). The NSP method often failed to converge, in up to 70% of datasets, depending on the scenario. When the NSP did converge, it produced results that are nonsensical, with SEs too high to give any credibility to these estimates.

For both methods, standard errors increased in the high-bandwidth scenarios compared with the low-bandwidth scenarios. The parent process intensities were also captured better by the MACPP than by the classical NSP, especially in high-bandwidth scenarios. Lastly, performance of the MACPP was not affected by the presence of a taxon spatially unrelated to the multilayered arrangement (Tables 2 and C.1 of Supplementary Materials).

### 4.4 Sensitivity Analyses

As outlined in Section 3.2, we chose a half-normal prior for the bandwidth parameter. We conducted comprehensive sensitivity analyses (detailed in Section D of the Supplementary Materials) to examine the extent to which conclusions are robust with respect to the choice of prior distribution for the bandwidth parameter. In these analyses, we compared parameter estimation performance under three different prior distributions (half-normal, uniform, log-normal). In summary, the proposed framework was not sensitive to the choice of the prior distribution for the bandwidth parameter under scenarios with low-bandwidth. Under high-bandwidth data scenarios, however, elicitation of an informative prior seemed to help the proposed framework become more numerically stable, increasing power.
Table 1: A summary of twelve simulation scenarios considered in Section 4. The offspring density is controlled by setting \((\alpha_2, \alpha_3) = (1.5, 1)\) for ‘Sparse’, \((4, 3)\) for ‘Dense’ and \((4, 1)\) for ‘Mixed’ setting. Bandwidth ‘Low’ setting sets \((h_2, h_3) = (0.01, 0.02)\) and the ‘High’ setting sets it to \((0.1, 0.01)\). The setting “Unrelated taxon” refers to whether there exists a taxon in the data spatially unrelated to the multilayered arrangement.

| Scenario | Unrelated taxon | Offspring density | Bandwidth |
|----------|-----------------|-------------------|-----------|
| 1        | Absent          | Sparse            | Low       |
| 2        | Absent          | Sparse            | High      |
| 3        | Absent          | Dense             | Low       |
| 4        | Absent          | Dense             | High      |
| 5        | Absent          | Mixed             | Low       |
| 6        | Absent          | Mixed             | High      |
| 7        | Present         | Sparse            | Low       |
| 8        | Present         | Sparse            | High      |
| 9        | Present         | Dense             | Low       |
| 10       | Present         | Dense             | High      |
| 11       | Present         | Mixed             | Low       |
| 12       | Present         | Mixed             | High      |

5 Analysis of Human Microbiome Biofilm Image Data

5.1 Hyperparameters and Analysis Settings

From the \(m = 9\) different taxa probed in the human dental plaque sample, we analyzed the data for locations of three offspring taxa, namely Streptococcus, Porphyromonas and Pasteurellaceae. The parent taxon for the first two taxa is Corynebacterium and for the third taxon, it is Streptococcus which itself is an offspring taxon (Figure 1). Therefore, according to the notations introduced in Section 3.1, the process \(Y_1\) corresponding to Corynebacterium, which functions only as a parent taxon \((p = 1)\). \(Y_2, Y_3,\) and \(Y_4\) represent the three offspring processes \((q = 3)\) for Streptococcus, Porphyromonas, and Pasteurellaceae, respectively. The corresponding parent processes are \(C_2 = Y_1, C_3 = Y_1\) (Corynebacterium as a parent), and \(C_4 = Y_2\) (Streptococcus as a parent). The remainder of the taxa are modeled as homogeneous Poisson processes. In the MACPP analyses, we set the hyperparameters at \(a_Y = b_Y = a_C = b_C = a = b = 0.01\) and \(\sigma\) at 2.06, such that the 99th percentile for the bandwidth parameters was approximately 5\(\mu\)m.

Streptococcus and Porphyromonas clustered around Corynebacterium only at its tip — not along its length and not at its base (Figure 3 and Mark Welch et al., 2016). Because all parts of Corynebacterium cells were observed in the image, there is inhomogeneity in the clustering relationships. Thus, we manually subset the dataset to create more homogeneous sub-datasets in which we could test performance of the MACPP. We simply subset the image in four equally sized quadrants (Figure 5) and analyzed data from each quadrant independently. Taxa’s abundance varied across the quadrants, allowing for some comparison of performance (Table 3). The inclusion of black space, where no taxa are observed, can
Table 2: The true value, estimates and uncertainty measures for the parameters $(\alpha_2, \alpha_3, h_2, h_3, \lambda^C)$ using MACPP and NSP models for the first six scenarios. For MACPP, the estimates (EST) are the posterior means averaged over different datasets, the SD is computed by averaging the posterior standard deviation over different datasets and SE is computed as the standard deviation of the estimates over the datasets. For NSP, the estimates (EST) are the outputs of the minimum contrast method and SE is calculated similarly using these estimates. SD for NSP is not computed as the method does not provide an uncertainty measure. The last column (%F) refers to the percentage of times the NSP model failed to converge in that particular scenario.

| Scenario | True value | MACPP | NSP |
|----------|------------|-------|-----|
|          | EST        | SD    | SE  | EST | SE | %F |
| $\alpha_2$ | 1.50       | 1.54  | 0.10 | 0.11 | 2.34 | 7.71 |
| $\alpha_3$ | 1.00       | 1.03  | 0.08 | 0.09 | 1.97 | 6.84 |
| $h_2$     | 0.01       | 0.01  | < 0.01 | < 0.01 | 0.01 | 0.04 | 6 |
| $h_3$     | 0.02       | 0.02  | < 0.01 | < 0.01 | 0.70 | 6.43 |
| $\lambda^C$ | 150.00     | 164.07 | 13.16 | 13.28 | 170.78 | 52.51 |
| $\alpha_2$ | 1.50       | 1.50  | 0.11 | 0.11 | 291.16 | 383.69 |
| $\alpha_3$ | 1.00       | 1.02  | 0.08 | 0.08 | 1.06 | 0.40 |
| $h_2$     | 0.10       | 0.08  | 0.01 | 0.01 | 8.33 | 23.82 | 26 |
| $h_3$     | 0.01       | 0.01  | < 0.01 | < 0.01 | 0.01 | < 0.01 |
| $\lambda^C$ | 150.00     | 162.74 | 13.05 | 13.52 | 679.72 | 2378.68 |
| $\alpha_2$ | 4.00       | 4.05  | 0.14 | 0.16 | 13.30 | 66.75 |
| $\alpha_3$ | 3.00       | 3.06  | 0.13 | 0.11 | 10.60 | 75.03 |
| $h_2$     | 0.01       | 0.01  | < 0.01 | < 0.01 | 0.02 | 0.10 | 1 |
| $h_3$     | 0.02       | 0.02  | < 0.01 | < 0.01 | 0.03 | 0.08 |
| $\lambda^C$ | 150.00     | 204.84 | 14.50 | 14.29 | 209.22 | 49.75 |
| $\alpha_2$ | 4.00       | 4.01  | 0.17 | 0.17 | 710.62 | 648.03 |
| $\alpha_3$ | 3.00       | 3.04  | 0.13 | 0.12 | 3.00 | 0.62 |
| $h_2$     | 0.10       | 0.09  | 0.01 | 0.01 | 1.21 | 0.77 | 56 |
| $h_3$     | 0.01       | 0.01  | < 0.01 | < 0.01 | 0.01 | < 0.01 |
| $\lambda^C$ | 150.00     | 202.08 | 14.34 | 15.32 | 20.33 | 45.79 |
| $\alpha_2$ | 4.00       | 4.05  | 0.15 | 0.16 | 4.07 | 0.68 |
| $\alpha_3$ | 1.00       | 1.03  | 0.17 | 0.09 | 3.06 | 14.91 |
| $h_2$     | 0.01       | 0.01  | < 0.01 | < 0.01 | 0.01 | < 0.01 | 1 |
| $h_3$     | 0.02       | 0.02  | < 0.01 | < 0.01 | 1.10 | 10.15 |
| $\lambda^C$ | 150.00     | 203.82 | 14.51 | 15.40 | 203.47 | 35.16 |
| $\alpha_2$ | 4.00       | 4.02  | 0.17 | 0.14 | 685.54 | 694.99 |
| $\alpha_3$ | 1.00       | 1.01  | 0.07 | 0.06 | 2.49 | 10.25 |
| $h_2$     | 0.10       | 0.09  | 0.01 | 0.01 | 1.48 | 1.21 | 53 |
| $h_3$     | 0.01       | 0.01  | < 0.01 | < 0.01 | 0.02 | 0.07 |
| $\lambda^C$ | 150.00     | 198.19 | 14.22 | 13.86 | 35.99 | 99.33 |
Figure 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.

deflate density estimates and induce spurious spatial correlations. We minimized unnecessary black space by using a convex hull of the observed locations as the analysis window (Figure 5). We also applied the classical NSP model individually on the three offspring processes, *Streptococcus*, *Porphyromonas* and *Pasteurellaceae*, for each of the four quadrants.

5.2 Results

The MACPP approach successfully identified the multilayered arrangement in which *Streptococcus* and *Porphyromonas* clustered around *Corynebacterium* and *Pasteurellaceae* around *Streptococcus*. The estimated bandwidth parameters \( h_2, h_3 \) and \( h_4 \) for the four quadrants ranged between 7.62–11.23\( \mu m \), 7.23–15.21\( \mu m \) and 3.82–4.62\( \mu m \), respectively. The corresponding estimated offspring density parameters \( \alpha_2, \alpha_3 \) and \( \alpha_4 \) were between 1.04–2.22, 1.80–5.05 and 0.36–0.57, respectively. The estimated intensity parameter for the process related to *Corynebacterium* was 0.01–0.02 per unit area (Table 4). The estimates of intensity parameters for the processes corresponding to the other taxa also varied among the
Table 3: The abundance of different taxa of interest in the human dental plaque sample dataset and its subdivisions

| Taxon             | Quadrant | Total |
|-------------------|----------|-------|
|                   | I        | II     | III    | IV     |       |
| Actinomyces       | 119      | 280    | 154    | 223    | 776   |
| Capnocytophaga    | 512      | 755    | 574    | 573    | 2414  |
| Corynebacterium   | 58       | 219    | 186    | 245    | 708   |
| Fusobacterium     | 92       | 250    | 141    | 173    | 656   |
| Leptotrichia      | 191      | 411    | 234    | 339    | 1175  |
| Neisseriaceae     | 339      | 479    | 402    | 491    | 1711  |
| Pasteurellaceae   | 53       | 130    | 76     | 106    | 365   |
| Porphyromonas     | 227      | 525    | 269    | 420    | 1441  |
| Streptococcus     | 98       | 379    | 163    | 249    | 889   |

quadrants (Table E.3).

The estimates for the parent intensity and the offspring density parameters varied among quadrants and among the parent-offspring pairs. The estimates for $\lambda^C$ in each quadrant were consistent with the observed counts. For example, in the second quadrant, there are 219 observed *Corynebacterium* cells, and the estimated intensity parameter of 0.02 per unit area corresponds to a count of $\lambda^C|W| \approx 219$. For each of the three offspring taxa, the estimated counts matched the observed counts rounded to the nearest integer. For example, the estimated taxon counts for *Streptococcus*, *Porphyromonas* and *Pasteurellaceae* in the third quadrant comes out to be 163, 269 and 76 respectively, which matches the corresponding counts in Table 3.

The estimated bandwidth parameters $h_2, h_3$ and $h_4$ also varied by quadrant, exhibiting consistent patterns for the different offspring-parent pairs. Across all quadrants, for example, the estimated bandwidth parameter for the process related to *Pasteurellaceae* clustering around *Streptococcus* was smaller than those estimated for the processes concerning *Streptococcus* and *Porphyromonas* clustering around *Corynebacterium*. The estimated bandwidth parameters for the two processes that have *Corynebacterium* as their parent process were especially high and more variable. The NSP analysis of each of the offspring processes separately for each of the quadrants resulted in very low estimated bandwidth parameter values (Table 4). The NSP-based estimated number of cluster centers (parents) was extremely low, in fact, underestimated compared with the observed number of parent cells. On the other hand, the corresponding estimates of offspring density parameters $\alpha_2, \alpha_3$ and $\alpha_4$ from the NSP analyses were substantially larger than those from the MACPP analysis.

6 Discussion and Conclusion

We have developed a multivariate, multilayer adjusted model to estimate parent-offspring clustering relationships. The proposed MACPP framework produces model parameters that directly quantify multilayered structural arrangements, in which locations of one type of
Table 4: The estimates and uncertainty measures for the parameters \((\alpha_2, \alpha_3, \alpha_4, h_2, h_3, h_4, \lambda^C)\) using MACPP and NSP model for the four segments of the human dental plaque biofilm data. For MACPP, the estimates (EST) are the posterior means and SD is computed as the posterior standard deviation for each of the parameters. For NSP, the estimates are the output of the minimum contrast method. SD is not computed for NSP as the method does not provide one.

| Segment | MACPP | NSP |
|---------|-------|-----|
|         | EST   | SD  | EST  |
|         |       |     |      |
| I       |       |     |      |
| \(\alpha_2\) | 2.22  | 0.22 | 9.94 |
| \(\alpha_3\) | 5.05  | 0.34 | 24.19|
| \(\alpha_4\) | 0.57  | 0.08 | 15.39|
| \(h_2\) | 8.32  | 0.56 | 2.88 |
| \(h_3\) | 7.23  | 0.50 | 2.55 |
| \(h_4\) | 3.82  | 0.40 | 9.39 |
| \(\lambda^C\) | 0.01  | <0.01| <0.01|
| II      |       |     |      |
| \(\alpha_2\) | 1.98  | 0.10 | 41.89|
| \(\alpha_3\) | 2.91  | 0.13 | 35.43|
| \(\alpha_4\) | 0.36  | 0.03 | 10.12|
| \(h_2\) | 11.23 | 0.58 | 7.07 |
| \(h_3\) | 15.21 | 0.79 | 4.57 |
| \(h_4\) | 4.23  | 0.46 | 6.72 |
| \(\lambda^C\) | 0.02  | <0.01| <0.01|
| III     |       |     |      |
| \(\alpha_2\) | 1.04  | 0.08 | 9.34 |
| \(\alpha_3\) | 1.80  | 0.11 | 18.66|
| \(\alpha_4\) | 0.52  | 0.06 | 3.78 |
| \(h_2\) | 7.62  | 0.71 | 3.36 |
| \(h_3\) | 9.83  | 0.72 | 1.92 |
| \(h_4\) | 4.50  | 0.43 | 2.67 |
| \(\lambda^C\) | 0.02  | <0.01| <0.01|
| IV      |       |     |      |
| \(\alpha_2\) | 1.22  | 0.08 | 32.29|
| \(\alpha_3\) | 2.11  | 0.11 | 51.56|
| \(\alpha_4\) | 0.46  | 0.05 | 9.93 |
| \(h_2\) | 9.41  | 0.81 | 5.84 |
| \(h_3\) | 11.09 | 0.72 | 4.14 |
| \(h_4\) | 4.62  | 0.42 | 4.42 |
| \(\lambda^C\) | 0.02  | <0.01| <0.01|
object depend on locations of another, central object. This task cannot be achieved with a traditional NSP approach because it ignores the locations of parent objects. In simulation studies, the MACPP produced less bias and much lower empirical standard deviations for parameter estimates compared with the NSP model which produced rather nonsensical results. Although development of the proposed MACPP model was motivated by oral microbiome biofilm image data, the approach is not specific to biofilm and can be used for other biomedical, nonclinical, or nonbiological image applications.

We demonstrated feasibility and utility of the proposed method in application to real biofilm image data that exhibit complex arrangements with nine taxa. Broadly speaking, the MACPP successfully captured the multilayered corncob-like structure among a group of four taxa, despite the presence of five other spatially unclustered taxa. Yet, the estimated bandwidth parameters were much greater than the approximately sub 5-micron distances expected from cell-to-cell (or nearly cell-to-cell) contact apparent in the visible corncob arrangements and likely when cells physico-chemically interact. Because of this seeming discrepancy, it could be tempting to conclude that the low parent-offspring bandwidths estimated from the NSP framework make the NSP a more valid and preferred approach than the MACPP. However, because the classical NSP ignores the location of parent cells, it highly underestimates the number of cluster centers (parents). To compensate, it greatly overestimates the offspring density. Additionally, in ignoring parent locations, the NSP models self-clustering instead of parent-offspring clustering. Therefore, though appealing, the estimates produced by applying the NSP are not appropriate to quantify the multilayered intertaxon relationships we seek to investigate.

It would also be inappropriate to interpret the MACPP’s high average estimated parent-offspring bandwidth strictly as bias or as reflecting a universal drawback to the proposed approach. Rather, the seeming discrepancy suggests the need for careful interpretation of model parameters from such a complex, heterogeneous image. It also highlights a need for modifications to the approach if a simple, direct interpretation about cell-to-cell (parent-to-offspring) contact is the goal.

Two characteristics of the dental plaque biofilm data present particular challenges. First, cells of *Corynebacterium* are filamentous. When the locations of imputed centroids are used to estimate average distances to neighboring cells of a different taxon, centroid-based methods may give misleading estimates of true cell-to-cell distances. This form of bias may be exacerbated by the specific biological organization, in that the spherical offspring taxa (*Streptococcus* and *Porphyromonas*) cluster around only one end of the parent *Corynebacterium* filaments. Estimates of average *Pasteurellaceae-Streptococcus* distances were much smaller and closer to the expected range for cell-to-cell contact, likely because these clusters involved two types of similarly sized, spherical organisms.

One way to improve the MACPP performance, therefore, might be to model the shapes of cells by bi-axial spheroids. Another might be to use outline-based approaches, such as the Hausdorff distance (*Huttenlocher et al., 1993*), rather than relying on the geographic coordinates of imputed centroids to locate each filamentous cell. Nevertheless, such complicated adaptations of the MACPP are likely to be unnecessary if qualitative inference about the clustering arrangement suffices. Another option we are exploring is to condition models on
features empirically identified in the image. Limiting the MACPP analysis only to parent or offspring cells within some short distance of each other (e.g., radius = 10 microns) might improve its usefulness in similar applications, compared with analyses yielding only marginal bandwidth estimates over a heterogeneous spatial structure.

A second challenge is higher-order spatial structure, such as is evident in the image of a dental plaque biofilm community. Among other reasons for macro-level community structure, the environment at the outer edge of dental biofilm differs from that near the tooth surface. Image sampling and processing further contribute to heterogeneity, especially in that the image is of a two-dimensional slice from a three-dimensional structure. Some areas of the image have a higher concentration of cross-sections in which Corynebacterium cells themselves appear spherical, whereas other areas display large numbers of lengthwise Corynebacterium filaments. For the purposes of pilot-testing model performance, we divided the image into four quadrants, each with more homogeneous spatial patterns than the whole. Between-quadrant variability in parameter estimates from both the MACPP and NSP approaches suggest this approach was helpful. However, for many applications this ad hoc approach is unlikely to be sufficient, and better subsetting methods could be considered.

Even more powerful would be incorporation of the MACPP into a broader modeling framework that could capture higher-order spatial structure, for example, through regression parameters. The proposed model in (1) is flexible in that it can accommodate different standard modeling frameworks. For example, one can choose the form of $\lambda_i(s), i = 1, \ldots, m$, and replace the homogeneous Poisson process components in (2) by multivariate log-Gaussian Cox process components. Such an extension would enable the characterization and quantification of more complex spatial correlation structures among multiple taxa or other types of objects.

We are actively pursuing this line of model development.

A further avenue that will enhance usefulness of the approach is development of a meta-analytic approach to combine data from multiple images. In 100 sampled images of tongue biofilm from five donors, both within-sample and across-sample variability of inter-taxon spatial relationships is apparent. This variability can be quantified via a meta-analytic, multivariate, log-Gaussian Cox process model that we have developed (manuscript in preparation). To date, it has not been standard to apply formal unified models to combine data across multiple biomedical images. Instead, most practitioners have relied on post hoc comparisons, such as through ANOVA or non-parametric two-group comparisons. An efficient meta-analytic approach would increase the scope and applicability of the proposed method and would help in development of spatial analysis of multiple microbiome samples in general and may therefore prove to be a more fruitful line of model development.

Model validation for MACPP is not straightforward. The complex spatial structure does not permit out-of-sample prediction or split-sample cross-validation. In the context of traditional Bayesian point process models, the empirical spatial distributions can be compared with those based on posterior predictive samples (Leininger and Gelfand, 2017). However, this observed-versus-expected approach is challenged by the complexity of the data. Residuals for each of the sub-processes in (2) are easily obtained. Yet, a good match of observed and expected counts for one process can be misleading about overall model fit if another process is poorly estimated. This is exactly what happened with the NSP analysis of real data,
where offspring predicted counts were accurate and predicted parent counts were grossly underestimated. In contrast, the MACPP produced nearly perfect prediction of counts of different taxa in the observation window. To our knowledge, there is as yet, no valid method to combine the multiple residuals to produce a summary statistic reflecting overall goodness-of-fit.

We have proposed a novel MACPP method for quantifying multilayer, multivariate spatial relationships that we applied to analyze spatial arrangements of microbial cells in dental plaque biofilm image data. The proposed method exploits information about locations of objects at the center of a cluster of unlike objects, providing distinct advantages over the classic NSP model when parent information is available. The MACPP clearly outperformed the existing cluster point process model in every scenario in our numerical studies.

References

Adrian Baddeley, Ege Rubak, and Rolf Turner. *Spatial point patterns: methodology and applications with R*. CRC press, 2015.

Carlos G Bornancini, Héctor J Martínez, Diego G Lambas, Wim De Vries, Wil Van Breugel, Carlos De Breuck, and Dante Minniti. Clustering and light profiles of galaxies in the environment of 20 ultra-steep-spectrum radio sources. *The Astronomical Journal*, 127(2):679, 2004.

Carlos G Bornancini, Nelson D Padilla, Diego G Lambas, and Carlos De Breuck. Spatial clustering of ultra steep spectrum sources and galaxies. *Monthly Notices of the Royal Astronomical Society*, 368(2):619–622, 2006.

Julie Bourque and André Desrochers. Spatial aggregation of forest songbird territories and possible implications for area sensitivity. *Avian conservation and ecology*, 1(2), 2006.

Sung Nok Chiu, Dietrich Stoyan, Wilfrid S Kendall, and Joseph Mecke. *Stochastic geometry and its applications*. John Wiley & Sons, 2013.

Peter J Diggle. *Statistical analysis of spatial and spatio-temporal point patterns*. CRC press, 2013.

Peter J Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

Gary J Hill and Simon J Lilly. A change in the cluster environments of radio galaxies with cosmic epoch. *The Astrophysical Journal*, 367:1–18, 1991.

Daniel P Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9):850–863, 1993.

Janine Illian, Antti Penttinen, Helga Stoyan, and Dietrich Stoyan. *Statistical analysis and modelling of spatial point patterns*, volume 70. John Wiley & Sons, 2008.
Abdollah Jalilian, Yongtao Guan, Jorge Mateu, and Rasmus Waagepetersen. Multivariate product-shot-noise cox point process models. *Biometrics*, 71(4):1022–1033, 2015.

Sheila J Jones. A special relationship between spherical and filamentous microorganisms in mature human dental plaque. *Archives of Oral Biology*, 17(3):613–IN27, 1972.

Jiří Kopecký and Tomáš Mrkvicka. On the bayesian estimation for the stationary neyman-scott point processes. *Applications of Mathematics*, 61(4):503–514, 2016.

Thomas J Leininger and Alan E Gelfand. Bayesian inference and model assessment for spatial point patterns using posterior predictive samples. *Bayesian Analysis*, 12(1):1–30, 2017.

Jessica L Mark Welch, Blair J Rossetti, Christopher W Rieken, Floyd E Dewhirst, and Gary G Borisy. Biogeography of a human oral microbiome at the micron scale. *Proceedings of the National Academy of Sciences*, 113(6):E791–E800, 2016.

SJ Melles, D Badzinski, M-J Fortin, F Csillag, and K Lindsay. Disentangling habitat and social drivers of nesting patterns in songbirds. *Landscape ecology*, 24(4):519–531, 2009.

Jesper Møller. Shot noise cox processes. *Advances in Applied Probability*, 35(3):614–640, 2003.

Jesper Møller and Rasmus Plenge Waagepetersen. *Statistical inference and simulation for spatial point processes*. CRC Press, 2003.

Jesper Møller, Anne Randi Syversveen, and Rasmus Plenge Waagepetersen. Log gaussian cox processes. *Scandinavian journal of statistics*, 25(3):451–482, 1998.

Jerzy Neyman and Elizabeth L Scott. Statistical approach to problems of cosmology. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(1):1–29, 1958.

Dasith Perera, Anthony McLean, Viviana Morillo Lopez, Kaileigh Cloutier-Leblanc, Eric Almeida, Kiana Cabana, Jessica L Mark Welch, and Matthew M Ramsey. Mechanisms underlying proximity between oral commensal bacteria. *bioRxiv*, 2020.

Johannes Schindelin, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, Curtis Rueden, Stephan Saalfeld, Benjamin Schmid, et al. Fiji: an open-source platform for biological-image analysis. *Nature methods*, 9(7):676–682, 2012.

Ushio Tanaka and Yosihiko Ogata. Identification and estimation of superposed neyman-scott spatial cluster processes. *Annals of the Institute of Statistical Mathematics*, 66(4):687–702, 2014.

Scott A Tarof and Laurene M Ratcliffe. Habitat characteristics and nest predation do not explain clustered breeding in least flycatchers (empidonax minimus). *The Auk*, 121(3):877–893, 2004.
Martin Vorregaard. Comstat2-a modern 3d image analysis environment for biofilms. Master’s thesis, Citeseer, 2008.

MG Yates, L Miller, and JA Peacock. The cluster environments of powerful, high-redshift radio galaxies. *Monthly Notices of the Royal Astronomical Society*, 240(1):129–166, 1989.
Supplementary materials for ‘Multilayer Adjusted Cluster Point Process Model: Application to Microbial Biofilm Image Data’

Suman Majumder
Harvard T. H. Chan School of Public Health, Boston, Massachusetts, U.S.A.

Brent A. Coull
Harvard T. H. Chan School of Public Health, Boston, Massachusetts, U.S.A.

Jessica L. Mark Welch
Marine Biological Laboratory, Woods Hole, Massachusetts, U.S.A.

Patrick J. La Riviere
University of Chicago, Chicago, Illinois, U.S.A.

Floyd E. Dewhirst
The Forsyth Institute, Cambridge, Massachusetts, U.S.A.

Jacqueline R. Starr
Brigham and Women’s Hospital, Boston, Massachusetts, U.S.A.

Kyu Ha Lee
Harvard T. H. Chan School of Public Health, Boston, Massachusetts, U.S.A.
Contents

A Images of the Remaining Taxa from the Human Dental Plaque Biofilm Data 3

B Computational Details of the Sampling Algorithm 4
   B.1 Updating parameters associated with offspring densities 4
   B.2 Updating intensity parameters in homogeneous Poisson processes 4
   B.3 Updating bandwidth parameters 4

C Table of Estimates for the Remaining Scenarios from the Simulation Study 5

D Sensitivity Analyses 7

E Table of MACPP based Parameter Estimates for the Remaining Taxa in Dental Plaque Biofilm Data 8
A  Images of the Remaining Taxa from the Human Dental Plaque Biofilm Data

Additional images for distribution of *Neisseriaceae, Capnocytophaga, Actinomyces, Fusobacterium* and *Leptotrichia* in the human dental plaque sample are presented here in Figure A.1. 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.

Figure A.1: RGB images of *Neisseriaceae* (top left), *Capnocytophaga* (top middle), *Actinomyces* (top right), *Fusobacterium* (bottom left), *Leptotrichia* (bottom middle) 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.
B Computational Details of the Sampling Algorithm

We use a Markov chain Monte Carlo (MCMC) method to draw samples from the joint posterior distribution of \( \theta \). In the MCMC scheme, parameters are updated by either exploiting conjugacies inherent to the proposed model or using a Metropolis-Hastings algorithm.

B.1 Updating parameters associated with offspring densities

Let \( \theta^{-(\alpha)} \) denote a set of parameters \( \theta \) with \( \alpha \) removed. The full conditional distribution for \( \alpha_l, \ l = p+1, \ldots, p+q \) is

\[
\alpha_l | \theta^{-(\alpha)} \sim \text{Gamma}(a_Y + n_l, b_Y + \sum_{c_l \in C_l} \int_{W} k_l(u - c_l, h_l) \, du),
\]

where \( n_l \) is the number of observations in the window of taxon \( l \).

B.2 Updating intensity parameters in homogeneous Poisson processes

Posterior conjugacy is also achieved in the full conditional distributions of intensity parameters, \( \lambda_{v}^{C}, \ v = 1, \ldots, p \) and \( \lambda_{j}, \ j = p+q+1, \ldots, m \), which are given by

\[
\lambda_{v}^{C} | \theta^{-(\lambda_{v})} \sim \text{Gamma}(a_C + n_v, b_C + |W|), \ v = 1, \ldots, p;
\]

\[
\lambda_{j} | \theta^{-(\lambda_{j})} \sim \text{Gamma}(a + n_j, b + |W|), \ j = p+q+1, \ldots, m,
\]

where \( n_v \) and \( n_j \) are the numbers of observations for taxon \( v \) and taxon \( j \) within the window, respectively.

B.3 Updating bandwidth parameters

Since the full conditionals of the bandwidth parameters do not have standard forms, we use a random walk Metropolis-Hastings step to update each of \( h_l, \ l = 1, \ldots, p \). Given \( h_j(t) \), the sample for \( h_j, \ j = p+1, \ldots, p+q \) from iteration \( t \), for iteration \( (t+1) \), we propose a candidate sample \( h_j^* \) as a random draw from \( N(h_j(t), \sigma_{prop}^2) \), where \( \sigma_{prop}^2 \) is the prespecified variance of the proposal density. The corresponding acceptance ratio computes to be

\[
R = \frac{\exp\left(-\alpha_l \sum_{c_l \in C_l} \int_{W} k(u - c_l, h_l^*) \, du\right) \prod_{y \in \mathcal{Y} \setminus Y_l} \left(\sum_{c_l \in C_l} \int_{W} k(u - c_l, h_l^*) \right) \exp\left(-h_j^* \frac{2}{\sigma^2} \right) \mathbb{I}(h_j^* > 0)}{\exp\left(-\alpha_l \sum_{c_l \in C_l} \int_{W} k(u - c_l, h_l(t)) \, du\right) \prod_{y \in \mathcal{Y} \setminus Y_l} \left(\sum_{c_l \in C_l} \int_{W} k(u - c_l, h_l(t)) \right) \exp\left(-h_j(t) \frac{2}{\sigma^2} \right)}.
\]

Then we accept the proposed candidate \( h_j^* \) as \( h_j^{(t+1)} \) with probability \( \min\{R, 1\} \) or keep \( h_j^{(t+1)} = h_j(t) \).
C Table of Estimates for the Remaining Scenarios from the Simulation Study

Here we present results from the simulation study of cases 7 through 12 as numbered in Table 1 of the main manuscript. The results in Table C.1 tell a similar story as in Section 4. The MACPP performs much better compared to the NSP in terms of both estimating the model parameters as well as producing credible uncertainty quantification. NSP often fails to converge and produces nonsensical results.
Table C.1: The true value, estimates and uncertainty measures for the parameters \((\alpha_2, \alpha_3, h_2, h_3, \lambda^C)\) using MACPP and NSP models for the last 6 scenarios. For MACPP, the estimates (EST) are the posterior means averaged over different datasets, the SD is computed by averaging the posterior standard deviation over different datasets and SE is computed as the standard deviation of the estimates over the datasets. For NSP, the estimates (EST) are the outputs of the minimum contrast method and SE is calculated similarly using these estimates. SD for NSP is not computed as the method does not provide an uncertainty measure. The last column (%F) refers to the percentage of times the NSP model failed to converge in that particular scenario.

| Scenario | True Value | MACPP EST | MACPP SD | MACPP SE | NSP EST | NSP SE | %F |
|----------|------------|-----------|----------|----------|---------|--------|-----|
|          | \(\alpha_2\) | 1.50      | 1.53     | 0.10     | 0.10    | 1.46   | 0.33 |
|          | \(\alpha_3\) | 1.00      | 1.02     | 0.08     | 0.09    | 3.31   | 20.25|
| 7        | \(h_2\)    | 0.01      | 0.01     | <0.01    | <0.01   | 0.01   | <0.01|
|          | \(h_3\)    | 0.02      | 0.02     | <0.01    | <0.01   | 0.04   | 0.09 |
|          | \(\lambda^C\)| 150.00   | 161.06   | 12.91    | 12.20   | 171.35 | 34.72|
|          | \(\alpha_2\) | 4.00      | 4.02     | 0.14     | 0.15    | 8.77   | 48.78|
|          | \(\alpha_3\) | 3.00      | 3.05     | 0.13     | 0.13    | 2.91   | 0.77 |
| 8        | \(h_2\)    | 0.10      | 0.08     | <0.01    | <0.01   | 10.30  | 28.72|
|          | \(h_3\)    | 0.01      | 0.01     | <0.01    | <0.01   | 0.01   | <0.01|
|          | \(\lambda^C\)| 150.00   | 160.25   | 12.86    | 12.57   | 171.35 | 34.72|
|          | \(\alpha_2\) | 4.00      | 4.00     | 0.17     | 0.17    | 613.34 | 569.20|
|          | \(\alpha_3\) | 3.00      | 3.02     | 0.13     | 0.14    | 2.93   | 0.53 |
| 9        | \(h_2\)    | 0.10      | 0.09     | 0.01     | 0.01    | 1.15   | 0.64 |
|          | \(h_3\)    | 0.01      | 0.01     | <0.01    | <0.01   | 0.01   | <0.01|
|          | \(\lambda^C\)| 150.00   | 202.78   | 14.38    | 14.29   | 208.52 | 39.37|
|          | \(\alpha_2\) | 4.00      | 4.05     | 0.15     | 0.14    | 18.45  | 87.48|
|          | \(\alpha_3\) | 1.00      | 1.00     | 0.07     | 0.08    | 2.05   | 10.04|
| 10       | \(h_2\)    | 0.01      | 0.01     | <0.01    | <0.01   | 0.03   | 0.11 |
|          | \(h_3\)    | 0.02      | 0.02     | <0.01    | <0.01   | 0.47   | 4.41 |
|          | \(\lambda^C\)| 150.00   | 201.50   | 14.37    | 14.09   | 203.36 | 53.30|
|          | \(\alpha_2\) | 4.00      | 4.02     | 0.17     | 0.17    | 547.61 | 553.61|
|          | \(\alpha_3\) | 1.00      | 1.02     | 0.07     | 0.07    | 0.97   | 0.24 |
| 11       | \(h_2\)    | 0.10      | 0.09     | 0.01     | 0.01    | 1.04   | 0.82 |
|          | \(h_3\)    | 0.01      | 0.01     | <0.01    | <0.01   | 0.01   | <0.01|
|          | \(\lambda^C\)| 150.00   | 199.05   | 14.23    | 15.95   | 26.62  | 41.46|
D Sensitivity Analyses

We perform sensitivity analyses based on the simulated datasets used in Section 4. For the sensitivity analysis, we only looked at two scenarios: numbered 5 and 6 in Table 1 in the main paper. These scenarios allow us to judge the performance of the model both high and low offspring densities as well as in scenarios where the parent-offspring clustering is tight or loose.

Since we are interested in testing the prior sensitivity of the bandwidth parameters, we used 4 different choices for the prior distribution of the bandwidth parameters, namely 1) Half-normal prior, 2) Uniform prior, 3) Log-normal prior with a flat tail and high variance and 4) Log-normal prior with a slim tail and higher peak. For the Uniform prior, the lower and upper bounds were taken to be 0 and 0.2. Both the log-normal priors had \( \mu = \log 0.05 \) but the flat-tailed one had \( \sigma = 1 \) while the high-peaked one had \( \sigma = 0.1 \) as their hyperparameter setting. The hyperparameter setting for the half-normal prior was the same as in Section 4.

We report the mean absolute percentage bias for estimating the corresponding parameters in the two scenarios for the four different prior settings. The half-normal prior based MACPP works best and the performance is similar to what was observed in Section 4. When the true bandwidth is low, all the models, irrespective of prior choice, work great and their performances are similar to each other. But when the true bandwidth is high, clearly having a tighter prior helps (Table D.2). However, having an informative log-normal prior backfired for even low bandwidth scenario when the offspring density is low, as is the case for the second offspring process. Further sensitivity analysis on choice of \( \sigma \) parameter for the log-normal prior may be needed to understand its effects.

Table D.2: Mean absolute percentage bias for estimating parameter values of \( \alpha_1, \alpha_2, h_1, h_2 \) and \( \lambda^C \) based on posterior means of each of the 100 datasets for different choices of priors for the bandwidth parameters for MACPP when there is no extra taxon present in the data and when the first offspring process has true offspring density set at 4 while the for the second one it is 1. Setting ‘Low Bwd’ means when both \( h \) values are small while in ‘High Bwd’ \( h_1 \) is high but \( h_2 \) is low. Log-normal (flat) refers to the prior setting where a flat-tailed log-normal prior was used for the bandwidth parameters while Log-normal (tight) refers to the scenario where the said log-normal prior was tight and more informative.

|             | Half-normal | Uniform | Log-normal (flat) | Log-normal (tight) |
|-------------|-------------|---------|-------------------|-------------------|
| Low \( \alpha_1 \) | 0.03 | 0.03 | 0.03 | 0.03 |
| Low \( \alpha_2 \) | 0.07 | 0.07 | 0.07 | 0.07 |
| Low \( h_1 \) | 0.02 | 0.02 | 0.02 | 0.05 |
| Low \( h_2 \) | 0.04 | 0.04 | 0.04 | 0.24 |
| Low \( \lambda^C \) | 0.06 | 0.06 | 0.06 | 0.06 |
| High \( \alpha_1 \) | 0.03 | 0.31 | 0.52 | 0.03 |
| High \( \alpha_2 \) | 0.05 | 0.05 | 0.05 | 0.05 |
| High \( h_1 \) | 0.09 | 0.99 | 1.37 | 0.10 |
| High \( h_2 \) | 0.03 | 0.03 | 0.03 | 0.20 |
| High \( \lambda^C \) | 0.06 | 0.06 | 0.06 | 0.06 |
Table E.3: Estimates for parameters associated with *Neisseriaceae* ($\lambda_5$), *Capnocytophaga* ($\lambda_6$), *Actinomyces* ($\lambda_7$), *Fusobacterium* ($\lambda_8$) and *Leptotrichia* ($\lambda_9$) obtained by applying the proposed method on each of the four segments of the dental plaque sample image. All results are rounded to two decimal places. The standard errors were all smaller than 0.01 and hence are not reported separately.

| Segment | $\lambda_5$ | $\lambda_6$ | $\lambda_7$ | $\lambda_8$ | $\lambda_9$ |
|---------|-------------|-------------|-------------|-------------|-------------|
| 1       | 0.04        | 0.06        | 0.01        | 0.01        | 0.02        |
| 2       | 0.05        | 0.07        | 0.03        | 0.02        | 0.04        |
| 3       | 0.04        | 0.05        | 0.01        | 0.01        | 0.02        |
| 4       | 0.05        | 0.06        | 0.02        | 0.02        | 0.03        |