A nonparametric spatial model for periodontal data with non-random missingness

Brian J. Reich[1], **Dipankar Bandyopadhyay** [2 3] and **Howard D Bondell**[1]

February 12, 2013

Abstract

Periodontal disease progression is often quantified by clinical attachment level (CAL) defined as the distance down a tooth's root that is detached from the surrounding bone. Measured at 6 locations per tooth throughout the mouth (excluding the molars), it gives rise to a dependent data set-up. These data are often reduced to a one-number summary, such as the whole mouth average or the number of observations greater than a threshold, to be used as the response in a regression to identify important covariates related to the current state of a subject's periodontal health. Rather than a simple one-number summary, we set forward to analyze all available CAL data for each subject, exploiting the presence of spatial dependence, non-stationarity, and non-normality. Also, many subjects have a considerable proportion of missing teeth which cannot be considered missing at random because periodontal disease is the leading cause of adult tooth loss. Under a Bayesian paradigm, we propose a nonparametric flexible spatial (joint) model of observed CAL and the location of missing tooth via kernel convolution methods, incorporating the aforementioned features of CAL data under a unified framework. Application of this methodology to a data set recording the periodontal health of an African-American population, as well as simulation studies reveal the gain in model fit and inference, and provides a new perspective into unraveling covariate-response relationships in presence of complexities posed by these data.

Keywords: Attachment level; Dirichlet process; Kernel convolution; Non-normality; Non-stationarity

[1]Department of Statistics, North Carolina State University
[2]Division of Biostatistics, School of Public Health, University of Minnesota
[3]Both the first and second author contributed equally.

A nonparametric spatial model for periodontal data with non-random missingness

1 Introduction

Periodontitis, a form of periodontal disease (PD), is a chronic inflammatory disease of the periodontium triggered by bacterial plaque, and is characterized by gingivitis, destruction of the alveolar bone and periodontal ligament, apical migration of the epithelial attachment resulting in formation of periodontal pockets, and ultimately loosening and exfoliation of teeth. According to the Position Paper of the American Academy of Periodontology (Burt et al., 2005), the incidence of severe generalized PD ranges between 5-15% for any population, however a vast majority of the adults remain affected by some moderate form leading to severe impairment in the quality of life. Dental hygienists measure the progression of PD in a subject via the clinical attachment level (CAL), one of the most important (clinical) surrogate endpoints (Nicholls, 2003) of PD.

The motivating example for this work comes from a clinical study conducted at the Medical University of South Carolina (MUSC) on Type-2 diabetic Gullah-speaking African Americans (henceforth GAAD study, more details appear in Section 2). The underlying statistical question is to estimate the connection between CAL and various determinants (covariates) of PD, such as age, gender, body mass index (BMI status), glycemic control, etc. CAL measured throughout the whole mouth (6 locations per tooth excluding the molars) gives rise to a dependent data set-up, and quantifying a subject's latent disease status from these extensive site-level data can be challenging. The whole-mouth average CAL is routinely used as a response to study covariate effects via re-

gression. However, the whole-mouth average may be far from representative of periodontal health when data are missing not at random, as one would expect higher CAL values to be associated with tooth loss. In addition, there is inherent loss of information by pooling.

Rather than aggregating the data, we analyze all available data for each patient. The process of periodontal decay might have a spatial morphology (Reich et al., 2007; Reich and Bandyopadhyay, 2010), i.e., a diseased tooth-site can influence the decay-status of a set of neighboring tooth-sites (and not the whole mouth), and thus the model must account for spatial dependence. In practice, spatial analyses often assume normality of the outcomes; on the contrary, CAL values are often right-skewed, with a majority of the values concentrated around the lower levels but with an important minority having much higher levels (López et al., 2001; Do et al., 2003). In addition to non-normality, we also suspect non-stationarity, e.g., the variance and skewness of periodontal responses for the posteriorly-located molars are quite different than the anterior incisors. Furthermore, CAL values are missing if and only if the tooth is missing, and since extreme PD can lead to missing teeth, the missing-data mechanism is non-ignorable.

Model-based methods to estimate the effects of covariates on periodontal disease should accommodate the aforementioned concerns. Various spatial methods are available to account for non-normality (e.g., Gelfand et al., 2005; Griffin and Steel, 2006; Reich and Fuentes, 2007; Rodriguez and Dunson, 2011; Fonseca and Steel, 2011; Reich, 2012), non-stationarity (e.g., Sampson and Guttorp, 1992; Higdon et al., 1999; Fuentes, 2002; Schmidt and O'Hagan, 2003; Paciorek and Schervish, 2006), and data with informative observation locations (Diggle et al., 2010; Reich and Bandyopadhyay, 2010; Pati et al., 2011). In this paper, we combine important features of several of

these methods, and tailor them to the special features of PD data. An alternative to non-Gaussian modeling is to identify a suitable transformation, and apply a multivariate Gaussian model to the residuals. However, this makes interpretation less clear, and even if the transformed marginal distribution is Gaussian, there is no assurance that the joint distributions are Gaussian (Jara et al., 2008). Therefore, we prefer a flexible model on the original scale. Another approach is to avoid model-based methods altogether using a GEE-type analysis (Diggle et al., 2002). However, these methods may lack power compared to valid likelihood-based approaches, especially in high dimensions (in our case, each subject has 168 measure locations, and thus a 168×168 covariance matrix must be estimated). Our approach centers our prior on the stationary Gaussian model, and allows for flexibility (non-stationary, non-Gaussian) while utilizing subject-matter knowledge (symmetry of the mouth, spatial proximity, etc) in the prior.

Our spatial model builds on the kernel convolution (KC) approach of Higdon et al. (1999), who approximate a Gaussian process as a linear combination of kernel basis functions, and allow for non-stationarity by assigning each kernel a different bandwidth. We also build on Reich and Bandyopadhyay (2010) by jointly modeling the responses and the missing data locations in a multivariate spatial model. To allow for non-Gaussian responses in the KC framework, we model the distribution of the kernel coefficients using non-parametric Bayesian (Hjort et al., 2010) methods. The distribution is allowed to vary spatially to permit different shapes for the response distribution in different regions of the mouth. Rather than allow the kernel bandwidth and coefficient density to vary arbitrarily through space, we exploit the natural symmetry of the mouth (i.e., the mouth can be partitioned into four similar quadrants) to borrow strength across the mouth to efficiently

estimate these complex features. Similarly, we allow the strength of association between the underlying disease status and probability of a missing tooth to vary by tooth type, since the likely causes of missing teeth may be different for molars than incisors. The proposed flexible model which permits non-normality, non-stationarity, and non-random missingness leads to simple conjugate full conditional distributions for all but a few spatial correlation parameters, leading to relatively straight-forward MCMC coding and tuning.

The remainder of paper proceeds as follows. Sections 2 and 3 present the data and the statistical model, respectively. Computing details utilizing a Bayesian framework are given in Section 4. We study the finite sample performance of our methodology using a simulation study in Section 5 which shows that failing to account for the above features (typical for PD data) can dramatically degrade estimation of fixed effects. In Section 6, we analyze the GAAD dataset where we find strong skewness for molars, higher spatial correlation between incisors than other teeth, and strong evidence of non-random missingness. Section 7 provides some concluding remarks.

2 CAL data and exploratory analysis

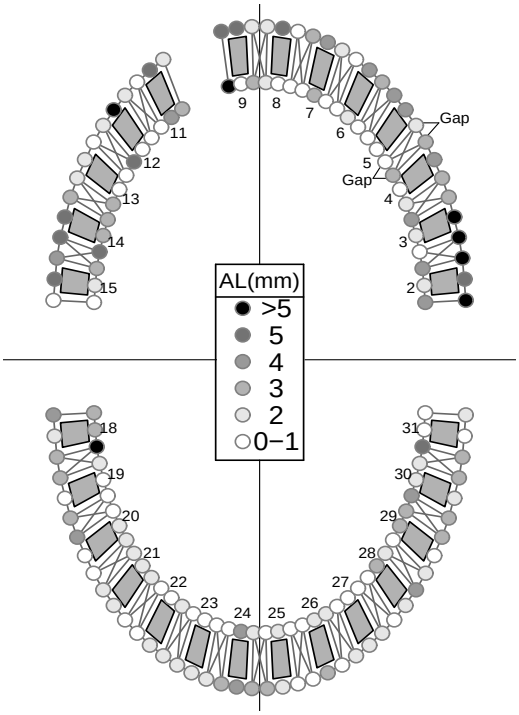
The dataset described in Section 1 was collected as part of a clinical study (Fernandes et al., 2009) conducted at MUSC primarily aimed at exploring the relationship between PD and diabetes (as determined by Hba1c, or glycosylated hemoglobin) in Type-2 diabetic Gullah-speaking African Americans (13 years or older) residing in the coastal islands of South Carolina. Clinical attachment level (CAL), defined as the depth (in mm) measured from the cemento-enamel junction (CEJ) to the bottom of the gingival sulcus for each site corresponding to a tooth was measured for six

pre-specified sites per tooth (excluding the third molars) via a periodontal probe, giving 168 measurement locations. Figure 1 shows the locations of these measurement sites for one subject who has an incisor missing. Of the six sites on each tooth, our model distinguishes between the four in a gap between teeth and the two that are not. Our model also classifies teeth as molars (tooth numbers 2-3, 14-15, 18-19, and 30-31), pre-molars (4-5, 12-13, 20-21, and 28-29), canines (6, 12, 22, 27), and incisors (7-10, 23-26), and into four quadrants: two on the upper jaw (teeth 2-8 and 9-15), and two on the lower jaw (18-24 and 25-31). For any particular tooth, the "tongue side" (lingual) locations refer to the three sites adjacent to (or the direction towards) the tongue that are closer to the center of the oral cavity, while the "cheek side" (buccal) refers to the three sites adjacent to the cheeks/lips that are farther away from the center.

Several subject-level covariates considered as possible determinants of PD such as age (in years), gender (1 = female, 0 = male), body mass index or BMI (in kg/m^2), smoking status (1 = smoker, 0 = never smoker) and HbA1c (1 = high, 0 = controlled) were included as fixed effects. The absolute pairwise correlation between these variable is no more than 0.2 for any pair. We also include spatial covariates such as tooth-type indicators, site in jaw, and site in gap (details appear in Section 3). We selected 199 of the 279 subjects having complete covariate information and at least one non-missing tooth.

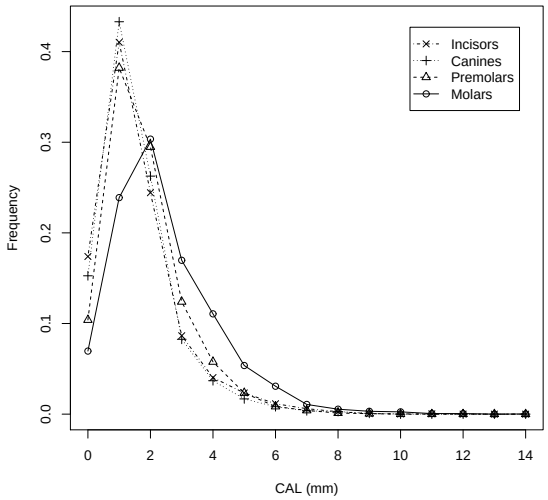
The density plots (collapsed by tooth type) in Figure 2a show that the data's density varies from fairly symmetric (although slightly right-skewed, mostly due to the boundary effect) for incisors to considerably right-skewed for molars. Non-normality persists after log (after adding one to the responses) and square root transformations, therefore we model the untransformed data for

Figure 1: Measurement locations and sample data for one subject. The vertical and horizontal lines separate the mouth into its four quadrants. "Gap" identifies as an example the four sites in the gap between teeth 4 and 5.

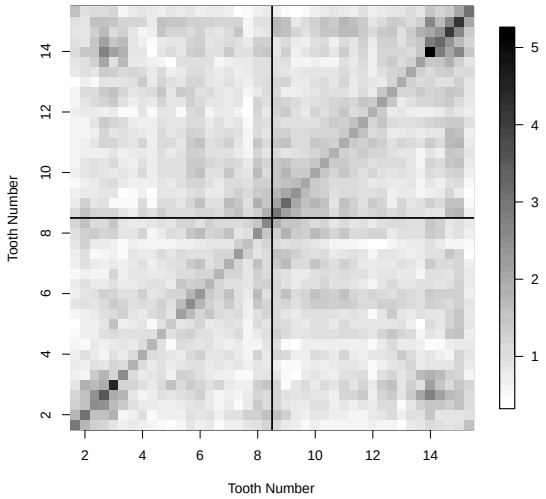


interpretability. In addition to non-normality, there is also evidence of non-stationarity reflecting differential rates of PD for the various types of tooth locations (anteriorly located incisors vs. posteriorly located molars). Figure 2b reveals a complex covariance structure. The variance is larger for molars than the other types of teeth. There is evidence of dependence between adjacent sites, especially for molars and incisors, as well as long-range dependence between molars on both sides of the mouth, for example, sites on teeth 3 and 14. Finally, dependence between adjacent sites depends on whether sites are in the gap between teeth, as evident in the lower covariance that appears in every third column/row for sites that are not in the gap between teeth. In the

Figure 2: Plots of the observed CAL data. Panel (a) gives the sample frequency of CAL for each tooth type (CAL is rounded to the nearest mm), and Panel (b) gives the 42×42 sample covariance of CAL for the 3×14=42 sites the tongue side of upper jaw (i.e., teeth 2-15 in Figure 1, the vertical and horizontal lines separate the two quadrants).



(a) Data by tooth type



(b) Sample covariance for one side of one jaw

next section, we describe a model that allows for the non-normality and the complex covariance structure observed in this exploratory data analysis.

3 Flexible spatial model for CAL

3.1 General framework

Let $y_i(\mathbf{s})$ be the observation for subject $i = 1, ..., n$ at spatial location $\mathbf{s}$. For each subject, there are the same $N = 168$ potential measurement locations $\mathbf{s}_1, ..., \mathbf{s}_N$. Denote $\mathbf{y}_i = [y_i(\mathbf{s}_1), ..., y_i(\mathbf{s}_N)]^T$ as the vector of responses for subject i. To account for non-randomly missing teeth, we jointly model the response and the location of missing teeth. For these data, typical for CAL data, either all the measurements from a tooth are observed or all observations are missing. Therefore, we define the

missing tooth process at the tooth level, rather than site level, with $\delta_i(t) = 1$ if tooth t is missing for subject i and $\delta_i(t) = 0$ otherwise.

We model CAL using a spatial model that exploits the natural symmetry of the mouth. That is, rather than simply modeling correlation via spatial distance, we account for correlation between teeth of the same type but in different quadrants. For example, due to genetic or hygienic factors, it may be that a subject has low CAL in most of the mouth, but high CAL for molars in all four quadrants. This dependence is accounted for by subject-level random effects. Let $\mathbf{Z}$ be the $N \times q$ random effect design matrix. We include $q = 6$ random effects: indicators for the four types of teeth, an indictor of whether the site is located on the upper jaw, and an indictor of whether the site is in a gap between teeth (rather than on the side of a tooth). The fixed effects mentioned in Section 2 are included in the $N \times p$ design matrix $\mathbf{X}_i$. These covariates do not vary spatially, and thus the rows of $\mathbf{X}_i$ are identical. The design matrix $\mathbf{X}_i$ does not include an intercept or spatial covariates such as tooth type, since these effects are captured by the mean of the random effects. The CAL values for subject i is modeled as

$$\mathbf{y}_i = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i \text{ and } \boldsymbol{\mu}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}\mathbf{a}_i + \boldsymbol{\theta}_i \tag{1}$$

where $\boldsymbol{\mu}_i = [\mu_i(\mathbf{s}_1), ..., \mu_i(\mathbf{s}_N)]^T$ is the vector of true CAL values for subject i and $\boldsymbol{\varepsilon}_i \sim \mathrm{N}(0, \sigma^2 I_N)$ is the vector of errors. The true CAL is decomposed as a function of the fixed effects $\mathbf{b}$, subject random effects $\mathbf{a}_i = (a_{1i}, ..., a_{qi})^T$, and spatial random effects $\boldsymbol{\theta}_i = [\theta_i(\mathbf{s}_1), ..., \theta_i(\mathbf{s}_N)]^T$, whose distributions will be discussed in Sections 3.2 and 3.3, respectively. In this model, spatial dependence is split into low-resolution random effects such as those for tooth type and jaw, and high-resolution

spatial effects $\theta_i(\mathbf{s})$ to account for small-scale spatial dependence from one site to the next on the same tooth or neighboring teeth.

For missing data, we introduce a latent continuous spatial process $z_i(t)$, so that $\delta_i(t) = I(z_i(t) > 0)$, for $t = 1, ..., T$. The latent variable is modeled in terms of the average true (latent) CAL values for subject i at the six locations on tooth t, denoted $\mathbf{D}_t^T \boldsymbol{\mu}_i$, where $\mathbf{D}_t$ is the N-vector with $D_t(\mathbf{s})$ equal 1/6 if site $\mathbf{s}$ is on tooth t and zero otherwise. Then

$$z_i(t) \sim \mathrm{N}\left(\mu_i^*(t), 1\right) \quad \text{and} \quad \mu_i^*(t) = \mathbf{D}_t^T[\mathbf{X}_i\mathbf{b}^* + \mathbf{Z}\mathbf{a}_i^*] + c(t)\mathbf{D}_t^T\boldsymbol{\mu}_i \tag{2}$$

independent over t and i, where $\mathbf{b}^*$ and $\mathbf{a}_i^* = (a_{i1}^*, ..., a_{iq}^*)^T$ are the fixed and random effects for missing teeth, respectively. Integrating over the latent $z_i(t)$ gives the usual probit link

$$\mathrm{P}[\delta_i(t) = 1] = \Phi\left[\mu_i^*(t)\right], \tag{3}$$

where Φ is the standard normal distribution function. The relationship between CAL and missing teeth is controlled by $c(t)$, which is allowed to vary by tooth. If $c(t) > 0$, then regions with high CAL are more likely to have missing teeth, and vice versa. We allow $c(t)$ to vary by tooth type, but assume that it is constant for all teeth of the same type to borrow strength across teeth. We note that we have not included a spatial random effect analogous to $\boldsymbol{\theta}_i$ in the missing teeth model. We experimented with this model, but found that these spatial random effects were not well-identified after including subject random effects, likely because with only $T = 28$ teeth for each subject, there is not enough information in the data to identify small-scale spatial dependence within a

subject.

3.2 Low-resolution spatial model

The subject-level random effects a_{ik} and a^*_{ik} control the low resolution spatial trends for subject i. These random effects are modeled as $a_{ik} \overset{iid}{\sim} F_k$ and $a^*_{ik} \overset{iid}{\sim} F^*_k$. Rather than specifying a parametric distribution, we model F_k nonparametrically using a Dirichlet process mixture (DPM) of normals (Ferguson, 1973, 1974; Antoniak, 1974). The DPM model is commonly used to capture uncertainty in the parametric form of a distribution. Below, we specify the DPM prior for an arbitrary distribution function F with associated density $f(y)$. Using the stick-breaking representation (Sethuraman, 1994), the DPM model for F is equivalent to modeling the density as an infinite mixture of normals

$$f(y) = \sum_{j=1}^{\infty} \pi_j \phi(y|\eta_j, \tau_1^2) \quad \text{and} \quad \eta_j \overset{iid}{\sim} \text{N}(m, \tau_2^2), \tag{4}$$

where $\phi(y|m, \tau^2)$ is the Gaussian density function with mean m and variance τ^2, and the mixture weights satisfy $\pi_j > 0$ and $\sum_{j=1}^{\infty} \pi_j = 1$.

The mixture probabilities 'break the stick', i.e., the unit interval, into pieces that sum to one. The proportion of the stick attributed to term j is determined by $v_j \overset{iid}{\sim} \text{Beta}(1, D)$. The first probability is $\pi_1 = v_1$. The remaining terms are $\pi_j = v_j(1 - \sum_{l<j} \pi_l) = v_j \prod_{l<j}(1 - v_l)$, where $\sum_{l<j} \pi_l$ is the proportion of the stick accounted for by the first $j - 1$ terms, and v_j is the proportion of the remaining stick attributed to term j. We denote this model as $F \sim \text{DPM}(m, \tau_1, \tau_2, D)$.

Each CAL random effect has its own distribution, that is $F_k \sim \text{DPM}(m_k, \tau_{1k}, \tau_{2k}, D_k)$. For the missing tooth random effects, the responses are binary and provide less information about the

shape of the random effects distribution. Therefore we model them parametrically by taking F_k to be the normal distribution with mean m_k^* and standard deviation σ_k^*.

3.3 High-resolution spatial model

We model the site-level spatial CAL processes $\theta_i(\mathbf{s})$ using Gaussian processes. To allow for non-stationarity and non-Gaussianity, the spatial terms are modeled using kernel convolution methods. Higdon et al. (1999) show that an arbitrary Gaussian process $\theta(\mathbf{s})$ can be written as a kernel convolution of white noise,

$$\theta(\mathbf{s}) = \int_{\mathcal{R}^2} K(||\mathbf{s} - \mathbf{u}||) Z(\mathrm{d}\mathbf{u}), \tag{5}$$

where K is a kernel function and Z is a white noise process. The kernel function is related to the covariance via

$$\mathrm{Cov}\left[\theta(\mathbf{s}), \theta(\mathbf{s}+\mathbf{h})\right] = \tau_Z^2 \int_{\mathcal{R}^2} K(||\mathbf{u}||) K(||\mathbf{u}+\mathbf{h}||) \mathrm{d}\mathbf{u}, \tag{6}$$

where τ_Z^2 controls the variance. For example, the Gaussian kernel function $K(||\mathbf{u}||) = \exp(-||\mathbf{u}||^2/\rho^2)$ gives the isotropic squared-exponential covariance function $\mathrm{Cov}[\theta(\mathbf{s}), \theta(\mathbf{s}+\mathbf{h})]$ of the form $\exp(-||\mathbf{h}||^2/\psi^2)$ where ψ is a function ρ. The integral (5) permits the approximation

$$\theta(\mathbf{s}) = \sum_{l=1}^{L} K(||\mathbf{s} - \mathbf{u}_l||) \gamma_l \tag{7}$$

where $\{\mathbf{u}_1, ..., \mathbf{u}_L\}$ is a fixed set of knots covering the spatial domain of interest and $\gamma_l \overset{iid}{\sim} \mathrm{N}(0, \tau_\theta^2)$. This yields the covariance function

$$\mathrm{Cov}\left[\theta(\mathbf{s}), \theta(\mathbf{s}+\mathbf{h})\right] = \tau_\theta^2 \sum_{l=1}^{L} K(||\mathbf{s}-\mathbf{u}_l||) K(||\mathbf{s}+\mathbf{h}-\mathbf{u}_l||) \tag{8}$$

which approximates (6) for large L and regular-spaced $\{\mathbf{u}_1, ..., \mathbf{u}_L\}$. Higdon et al. (1999) use this approximation to define non-stationary processes by having a different bandwidth for each term in the sum, and using a second spatial process to spatially smooth the bandwidths. Assuming the subjects share the same kernels, that is,

$$\theta_i(\mathbf{s}) = \sum_{l=1}^{L} K_l(\mathbf{s}) \gamma_{li}. \tag{9}$$

we specify an anisotropic, non-stationary, and non-Gaussian model for the spatial processes. Rather than allowing the kernel functions to vary arbitrarily through the spatial domain, we simply specify a different bandwidth for each tooth type. We use squared-exponential kernels (sometimes called the Gaussian or radial basis function kernel) defined as,

$$K_l(\mathbf{h}) = \begin{cases} \exp\left(-\frac{(s_1-u_{1l})^2+\phi_k(s_2-u_{2l})^2}{\rho_k^2}\right) & \mathbf{s} \text{ and } \mathbf{u}_l \text{ on the same jaw} \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

where knot $\mathbf{u}_l = (u_{1l}, u_{2l})^T$ is on a tooth of type k. The spatial range of the kernel is controlled by the kernel bandwidth ρ_k. The relative strength of correlation in the two directions is determined by ϕ_k. The first spatial coordinate is the distance moving from left to right around the jaw, and the

second coordinate is zero for tongue-side sites and one for cheek-side measurements.

The squared-exponential kernel is restrictive in that its parameters control only the bandwidth of the kernels and not the shape. Richer kernels are available that have both scale and shape parameters, for example those that correspond to a Matérn covariance. However, in our application the data are on a regular grid and we find the kernels are non-negligible for only one or two neighboring sites. Therefore, it does not seem possible to identify the shape of the kernels, and we elect to use the simple squared-exponential kernels.

The distribution of the latent γ_{li} varies by tooth type. To account for non-normality, we model these latent variables using nonparametric methods. For knots on a tooth of type k, we assume that

$$\gamma_{li} \overset{iid}{\sim} G_k \quad \text{where} \quad G_k \sim \text{DPM}(0, \omega_{1k}, \omega_{2k}, H_k). \tag{11}$$

The mean is taken to be zero to identify the means of the tooth-type random effects. By allowing these distributions to vary spatially, we accommodate varying degrees of skewness in different regions of the mouth. Despite having different kernels and random effects distribution for each tooth type, proximal sites of different tooth types remain dependent because the kernels cover teeth of multiple types, and thus random effects are shared by multiple tooth types. The covariance of θ remains (8) if we assume the same density throughout space, i.e., $G_k \equiv G$ for all k, where τ_θ^2 is the variance of γ_{li}. Allowing G_k to vary by region affects the spatial covariance by having a different variance in each region.

Finally, we discuss the marginal distribution (over the latent γ_{li}) of $\theta_i(\mathbf{s})$ induced by convolving over the G_k. For simplicity, assume that $G_k \equiv G$ for all k. Since $\theta_i(\mathbf{s})$ is a linear combination

of terms distributed as G, its distribution will generally be more Gaussian than G. The number of terms with non-negligible weight $K_l(\mathbf{s})$ determines whether $\theta_i(\mathbf{s})$'s distribution more closely resembles a Gaussian distribution or G. To explore this relationship, first assume that G is fixed and consider varying the number of knots. If the knots are sparse so that $K_l(\mathbf{s})$ is positive for only a single term, then $\theta_i(\mathbf{s}) \sim G$; in contrast, in the limiting case with $L \to \infty$, there are many terms with non-negligible $K(\mathbf{s} - \mathbf{u}_l)$ and $\theta_i(\mathbf{s})$ is approximately Gaussian.

A more relevant case from our modeling perspective is holding L and $\{\mathbf{u}_1, ..., \mathbf{u}_L\}$ fixed, and examining the span of marginal distributions that can be fit by varying G, since this is what is done during model fitting. We define the moment generating function corresponding to G as $\mathrm{E}[\exp(t\gamma_{li})] = M(t)$. Then the moment generating function of $\theta_i(\mathbf{s})$ is $M_\theta(t) = \mathrm{E}\{\exp[t\theta_i(\mathbf{s})]\} = \prod_{l=1}^{L} M[K_l(\mathbf{s})t]$. For regular grids of observations and knots, the set of kernel values will be the same for all $\mathbf{s}$, and the moment generating function can be written $M_\theta(t) = \prod_{l=1}^{L} M[w_l t]$ for all $\mathbf{s}$, where $w_l = K_l(\mathbf{s})$. Therefore, identifying the G associated with an arbitrary response distribution requires solving $M_\theta(t) = \prod_{l=1}^{L} M[w_l t]$ for M. In the special case of a uniform kernel, $K_l(\mathbf{s}) = I[||\mathbf{s} - \mathbf{u}_l|| < \rho]$, and regular spacing so that each observation has $K_l(\mathbf{s}) > 0$ for n_K knots, then $M_\theta(t) = M(t)^{n_K}$. Therefore, setting $M_\gamma(t) = M_\theta(t)^{1/n_K}$ gives the desired marginal. This shows that even for a dense grid of knots, the span on densities available for the marginal distribution of $\theta_i(\mathbf{s})$ is the same as the span of the DPM model, which is known to be sufficiently flexible. Although we do not use a uniform kernel in our data example, in kernel smoothing the shape of the kernel is often less important than the bandwidth (Hand et al., 2001), and so this intuition should hold for other kernels as well.

4 Computational details

We use Metropolis-within-Gibbs MCMC (Gilks et al., 1995) to analyze this model. The full conditionals required for MCMC have fairly simple conjugate forms for most of the model's parameters, as described below. A complication is that the DPM priors in (4) are infinite mixtures. In practice it may not be necessary to use an infinite mixture. Note that by construction, the mixture probabilities are stochastically decreasing in j, for example, the prior mean of π_j is $[1/(D+1)][D/(D+1)]^{j-1}$. Therefore, little is lost by truncating the mixture at a fixed number of terms, M, and setting $v_M = 1$ to ensure that the probabilities sum to one. This gives a semiparametric finite mixture model that approximates the full nonparametric DP model. Conveniently, the mass in the final term π_M represents the truncation error, so to determine if the approximation is valid we inspect the posterior of π_M. We find that $M = 10$ mixture components provides a sufficient approximation for our data.

To facilitate MCMC for the non-Gaussian subject random effects, we introduce auxiliary variables (Chen et al., 2000) $g_{ki} \in \{1, ..., M\}$ to indicate the mixture component for the k^{th} random effect for subject i in (4). The auxiliary model for a_{ki} becomes

$$\begin{aligned} a_{ki}|g_{ki} = G &\overset{indep}{\sim} \mathrm{N}(\theta_{kG}, \tau_{1k}^2) \\ \theta_{kj} &\overset{iid}{\sim} \mathrm{N}(m_k, \tau_{2k}^2) \\ P(g_{ki} = j) &= \pi_{kj} = v_{kj} \prod_{l<j} (1 - v_{kl}), \end{aligned} \tag{12}$$

where $v_{kj} \overset{iid}{\sim} \mathrm{Beta}(1, D_k)$. Similarly, for the kernel convolution coefficients, we introduce auxiliary

variables $h_{li} \in \{1, ..., L\}$ such that

$$\begin{aligned} \gamma_{li}|h_{li} = H &\overset{indep}{\sim} \mathrm{N}(\delta_{lH}, \omega_{1k}^2) \\ \delta_{lj} &\overset{iid}{\sim} \mathrm{N}(0, \omega_{2k}^2) \\ P(h_{li} = j) &= u_{lj} \prod_{k<j} (1 - u_{lk}). \end{aligned} \tag{13}$$

where $u_{lj} \overset{iid}{\sim} \text{Beta}(1, H_l)$.

Below we provide the full conditional of the parameters in the auxiliary parameter model that are required for MCMC. The latent variable $z_i(t)$ is updated as

$$z_i(t) \sim \begin{cases} TN_{(-\infty,0)}(\mu_i^*(t), 1), & \delta_i(t) = 0 \\ TN_{(0,-\infty)}(\mu_i^*(t), 1), & \delta_i(t) = 1 \end{cases}, \tag{14}$$

where $TN_A(m, s^2)$ is the truncated normal density with truncation region A, location m, and scale s. The vectors in the mean $\boldsymbol{\mu}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}\mathbf{a}_i + \boldsymbol{\theta}_i$ each have multivariate normal full conditional distributions. Define $\mathbf{z}_i = [z_i(1), ..., z_i(T)]^T$, $\mathbf{D}$ to be the $T \times N$ matrix with t^{th} row $\mathbf{D}_t^T$, and $\mathbf{C}$ as the $T \times T$ diagonal matrix with t^{th} diagonal value $c(t)$. To update $\mathbf{b}$, let $\mathbf{r}_{b1i} = \mathbf{y}_i - \mathbf{Z}\mathbf{a}_i - \boldsymbol{\theta}_i$ and $\mathbf{r}_{b2i} = \mathbf{z}_i - \mathbf{D}[\mathbf{X}_i\mathbf{b}^* + \mathbf{Z}\mathbf{a}_i^*] + \mathbf{CD}(\mathbf{Z}\mathbf{a}_i - \boldsymbol{\theta}_i)]$. Then $\mathbf{b}|\text{rest} \sim \mathrm{N}(\mathbf{V}_b^{-1}\mathbf{M}_b, \mathbf{V}_b^{-1})$, where

$$\mathbf{M}_b = \sigma^{-2} \sum_{i=1}^{n} \mathbf{r}_{b1i}^T \mathbf{X}_i + \sum_{i=1}^{n} \mathbf{r}_{b2i}^T \mathbf{CDX}_i \quad \text{and} \quad \mathbf{V}_b = \sigma^{-2} \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{X}_i + \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{D}^T \mathbf{CCDX}_i + c_b^{-2} \mathbf{I}_p,$$

and $\mathbf{b}$ has prior $\mathbf{b} \sim \mathrm{N}(0, c_b^2 \mathbf{I}_p)$. The random effects have full conditionals $\mathbf{a}_i|\text{rest} \sim \mathrm{N}(\mathbf{V}_a^{-1}\mathbf{M}_a, \mathbf{V}_a^{-1})$,

where

$$\mathbf{M}_a = \sigma^{-2}\mathbf{r}_{a1i}^T\mathbf{Z} + \mathbf{r}_{a2i}^T\mathbf{CDZ} + \Omega_a\Theta_i \quad \text{and} \quad \mathbf{V}_a = \sigma^{-2}\mathbf{Z}^T\mathbf{Z} + \mathbf{Z}^T\mathbf{D}^T\mathbf{CCDZ} + \Omega_a,$$

and $\mathbf{r}_{a1i} = \mathbf{y}_i - \mathbf{X}_i\mathbf{b} - \boldsymbol{\theta}_i$ and $\mathbf{r}_{a2i} = \mathbf{z}_i - \mathbf{D}[\mathbf{X}_i\mathbf{b}^* + \mathbf{Z}\mathbf{a}_i^*] + \mathbf{CD}(\mathbf{X}_i\mathbf{b} - \boldsymbol{\theta}_i)]$, $\Theta_i = (\theta_{1g_{1i}}, ..., \theta_{qg_{qi}})^T$ and $\Omega_a = \text{Diag}(\tau_{11}^{-2}, ..., \tau_{1q}^{-2})$. Also, $\boldsymbol{\gamma}_i = (\gamma_{1i}, ..., \gamma_{Li})^T|\text{rest} \sim \text{N}(\mathbf{V}_\gamma^{-1}\mathbf{M}_\gamma, \mathbf{V}_\gamma^{-1})$, where

$$\mathbf{M}_\gamma = \sigma^{-2}\mathbf{r}_{\gamma1i}^T\mathbf{K} + \mathbf{r}_{\gamma2i}^T\mathbf{CDK} + \Omega_\gamma\Gamma_\gamma \quad \text{and} \quad \mathbf{V}_\gamma = \sigma^{-2}\mathbf{K}^T\mathbf{K} + \mathbf{K}^T\mathbf{D}^T\mathbf{CCDK} + \Omega_\gamma,$$

and $\mathbf{r}_{\gamma1i} = \mathbf{y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}\mathbf{a}_i$ and $\mathbf{r}_{\gamma2i} = \mathbf{z}_i - \mathbf{D}[\mathbf{X}_i\mathbf{b}^* + \mathbf{Z}\mathbf{a}_i^* + \mathbf{C}(\mathbf{X}_i\mathbf{b} - \mathbf{Z}\mathbf{a}_i)]$, $\mathbf{K}$ is the $N \times L$ matrix with elements $K_l(\mathbf{s})$, $\Gamma_i = (\theta_{1h_{1i}}, ..., \theta_{Lh_{Li}})^T$ and $\Omega_\gamma = \text{Diag}(\omega_{1i}^{-2}, ..., \omega_{Ni}^{-2})$, where $\omega_{li} = \omega_j$ if tooth l is of type j. The vectors $\mathbf{b}^*$ and $\mathbf{a}_i^*$ in the missing tooth model have similar full conditionals.

The missing data effects $c(t)$ have Gaussian full conditionals. We assume that $c(t) = C_j$ for all $t \in \mathcal{T}$ where $\mathcal{T}$ is the set of t such that tooth t is of type j. Then C_j has full conditional $C_j|\text{rest} \sim \text{N}(V_C^{-1}M_C, V_C)$ where

$$M_C = \sum_{i=1}^{n}\sum_{t\in\mathcal{D}}(\mathbf{D}_t^T\boldsymbol{\mu}_i)(z_i(t) - \mathbf{D}_t^T[\mathbf{X}_i\mathbf{b}^* + \mathbf{Z}\mathbf{a}_i^*]) \quad \text{and} \quad V_C = \sum_{i=1}^{n}\sum_{t\in\mathcal{D}}(\mathbf{D}_t^T\boldsymbol{\mu}_i)^2 + \sigma_C^{-2},$$

where C_j has prior $C_j \sim \text{N}(0, \sigma_C^2)$. The error variance has full conditionals

$$\sigma^{-2} \sim \text{Gamma}\left[nN/2 + a_\sigma, \sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu}_i)^T(\mathbf{y}_i - \boldsymbol{\mu}_i) + b_\sigma\right],$$

where σ has prior $\sigma^{-2} \sim \text{Gamma}(a_\sigma, b_\sigma)$.

Next, we define the full conditionals for the parameters in the auxiliary model (12); the parameters in (13) are similar. The full conditionals are

$$
\begin{aligned}
\theta_{kj}|\text{rest} &\sim \text{N}\left(\frac{m_k\tau_{2k}^{-2}+\tau_{1k}^{-2}\sum_{i=1}^{n}I(g_{ki}=j)a_{ki}}{\tau_{2k}^{-2}+\tau_{1k}^{-2}\sum_{i=1}^{n}I(g_{ki}=j)}, \frac{1}{\tau_{2k}^{-2}+\tau_{1k}^{-2}\sum_{i=1}^{n}I(g_{ki}=j)}\right) \\
P(g_{ki}=G|\text{rest}) &= \frac{\pi_{kG}\exp\left[-0.5\tau_{1k}^{-2}(a_{ki}-\theta_{kG})\right]}{\sum_{m=1}^{M}\pi_{km}\exp\left[-0.5\tau_{1k}^{-2}(a_{ki}-\theta_{km})\right]} \\
\tau_{1k}^{-2}|\text{rest} &\sim \text{Gamma}\left(n/2+a_\tau, \sum_{i=1}^{n}(a_{ki}-\theta_{kg_{ki}})^2/2+b_\tau\right) \\
\tau_{2k}^{-2}|\text{rest} &\sim \text{Gamma}\left(M/2+a_\tau, \sum_{m=1}^{M}(\theta_{km}-m_k)^2/2+b_\tau\right) \\
m_k|\text{rest} &\sim \text{N}\left(\frac{\tau_{2k}^{-2}\sum_{m=1}^{M}\theta_{km}}{M\tau_{2k}^{-2}+\sigma_m^{-2}}, \frac{1}{M\tau_{2k}^{-2}+\sigma_m^{-2}}\right) \\
v_{kl}|\text{rest} &\sim \text{Beta}\left(1+\sum_{i=1}^{n}I(g_{ki}=l), D+\sum_{i=1}^{n}I(g_{ki}>l)\right)
\end{aligned}
$$

where priors are assumed to be $\tau_{1k}^{-2}, \tau_{2k}^{-2} \sim$ Gamma(a_τ, b_τ) and $m_k \sim$ N$(0, \sigma_m^2)$.

The spatial correlation parameters (ρ_k and ϕ_k) that define the kernel do not have conjugate full conditionals and are updated using Metropolis sampling. We transform to $\log(\rho_k)$ and $\log(\phi_k)$, and use Gaussian candidate distributions. The candidate distributions are adaptively tuned during the burn-in to give acceptance ratio near 0.4.

We sample 5,000 iterations and discard the first 1,000 as burn-in for the simulated data in Section 5. For the real data analysis in Section 2 we sample 25,000 iterations (after thinning by four) and discard the first 10,000 as burn-in. Convergence is monitored with trace plots and ACF plots of several representative parameters. The computing was implemented using `R`, and the code is available upon request from the first author.

5 Simulation study

In most dental studies of CAL, including our analysis of the GAAD data in Section 6, the primary interest is to estimate risk factors associated with periodontal disease, for example the important special case of a clinical trial where one of the covariates is a treatment indicator. Therefore, in this simulation study we explore the effects of invalid modeling assumptions on the estimation of the fixed effects **b**. The data are generated from models (1) and (2). For computational purposes, to allow for a full simulation study we use only spatial locations on the upper jaw, giving $N = 84$ observations per subject. There are $p = 3$ fixed effects, generated as $\mathbf{X}_i \sim \text{N}(0, \mathbf{I}_p)$, with coefficients $\mathbf{b} = (0.0, 0.2, 0.4)^T$ for CAL and $\mathbf{b}^*$ equal zero for missing teeth. The CAL error variance is $\sigma^2 = 0.5^2$. The subject random effects are generated as $a_{ij} = T(A_{ij})$, where T is a transformation described below and $(A_{i1}, ..., A_{iq})^T \overset{iid}{\sim} \text{N}(0, \Sigma)$, where Σ is the compound symmetry correlation matrix with correlation 0.1. The missing tooth random effects are generated as $(a^*_{i1}, ..., a^*_{iq})^T \overset{iid}{\sim} \text{N}(0, \Sigma)$. The high resolution spatial component $\boldsymbol{\theta}_i$ is generated as a Gaussian process with mean zero and covariance $\text{Cov}[\theta_i(\mathbf{s}), \theta_i(\mathbf{s}')] = 0.5^2 \exp(-0.5||\mathbf{s} - \mathbf{s}'||^2)$. We keep the missing data function $c(t)$ constant over tooth types.

Each simulated data set has $n = 100$ subjects. We generate data under four designs, defined by the CAL random effects transformation T and missing data parameters $c(t)$:

1. Gaussian, randomly missing, $T(x) = x$, $c(t) = 0.0$

2. Non-Gaussian, randomly missing, $T(x) = -\log[1 - \Phi(x)]$, $c(t) = 0.0$

3. Gaussian, non-random missingness, $T(x) = x$, $c(t) = 1.0$

Table 1: Mean squared error (times 100, standard errors in parentheses) averaged over the p covariates for the simulation study.

		Statistical model			
		Random missing (RM)		Non-Random missing (NRM)	
Simulation		Gaussian	Non-Gaussian	Gaussian	Non-Gaussian
Design		1	2	3	4
Gaussian, RM	1	0.606 (0.004)	0.698 (0.010)	0.624 (0.007)	0.709 (0.008)
Non-Gaussian, RM	2	0.636 (0.003)	0.454 (0.007)	0.620 (0.005)	0.443 (0.002)
Gaussian, NRM	3	1.364 (0.114)	1.439 (0.123)	0.580 (0.011)	0.624 (0.010)
Non-Gaussian, NRM	4	0.424 (0.015)	0.359 (0.012)	0.380 (0.009)	0.317 (0.007)

4. Non-Gaussian, non-random missingness, $T(x) = -\log[1 - \Phi(x)]$, $c(t) = 0.5$

The transformation $T(x) = -\log[1 - \Phi(x)]$ yields subject random effects with exponential marginal distributions with mean one for the non-Gaussian designs. We use $L = N$ knots for the kernel convolution prior, with knots fixed at the data points $\mathbf{s}_1, ..., \mathbf{s}_N$. For priors, we choose $b_j, b_j^*, m_k, m_k,^*, C_j \sim \text{N}(0, 10^2)$, $\sigma^{-2}, \sigma_k^{*-2}, \tau_{jk}^{-2}, \omega_{jk}^{-2} \sim$ Gamma(0.1,0.1), and $\log(\rho_k), \log(\phi_k) \sim$ N(0,10^2). Finally, we fix the stick-breaking parameters $D_k = H_k = 1$.

Under this set-up, we generate $S = 200$ data sets from each simulation design. For each data set, we fit the Gaussian ($M = 1$) and non-Gaussian ($M = 10$) models, as well as models with ($c(t) \neq 0$) and without ($c(t) = 0$) non-randomly missing teeth. Methods are compared using mean squared error (MSE; computed using posterior means as estimates) for the fixed effects b_j, averaged over the S data sets and p covariates.

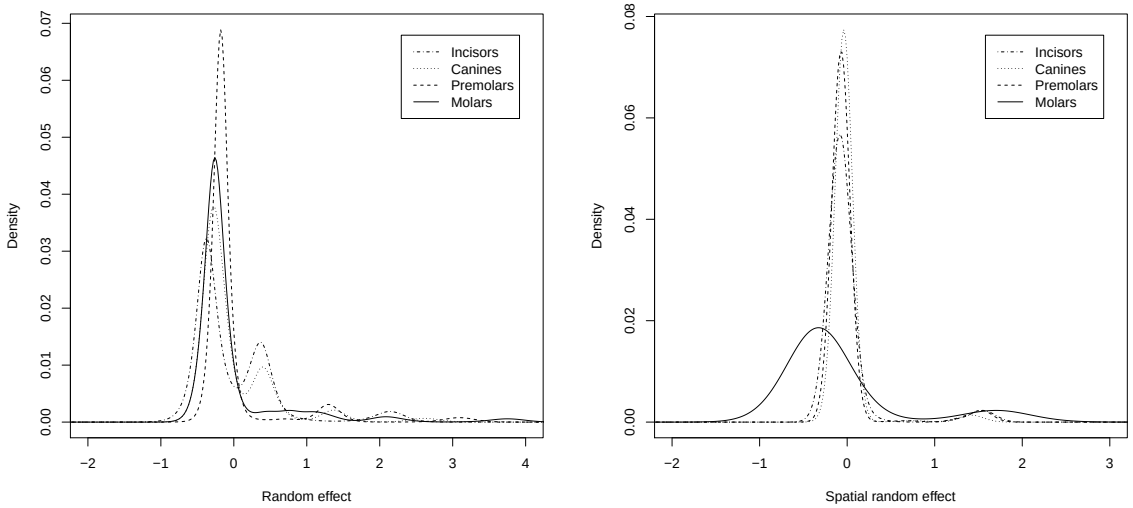
The results in Table 1 confirm that properly accounting for both non-random missingness and non-Gaussian responses provides improved fixed effect estimation. For design 2, Gaussian models have larger MSE than non-Gaussian models (relative MSE = 0.636/0.454 = 1.400 for model 1

compared to model 3). For design 3, models that do not account for non-random missingness have larger MSE than those that do (relative MSE = 1.364/0.580 = 2.352 for model 1 compared to model 2). The final design has both non-Gaussianity and non-random missingness, and in this case the MSE is minimized by the full model. The effect of accounting for non-random missingness is larger with normal data in design 3 than non-normal data in design 4. However, this may confounded with the larger informative missing coefficient ($c(t)$=1.0 for design 3 compared to $c(t) = 0.5$ for design 4).

6 Analysis of GAAD data

We analyze the GAAD data using the model in Section 3 and the same priors as in Section 5's simulation study. The posterior means of densities F_k and G_k are given in Figure 3. For each MCMC iteration, the densities are evaluated on a grid of points after shifting the density to have mean zero. The posterior mean is computed by averaging over MCMC samples. The random effect density for all four tooth types are right-skewed. The skewness is the strongest for pre-molars and molars. The density of the spatial random effects in Figure 3b shows that the variance is much larger for molars than the other tooth types. The density for molars is heavily right-skewed. Combining results in Figures 3a and 3b suggests that there are some subjects with very large CAL values for all molars, as well as some sites with extremely large CAL values after accounting for large-scale effects. We note that the spatial process $\theta_i(\mathbf{s})$ is a linear combination of coefficients for several knots, therefore the distribution of $\theta_i(\mathbf{s})$ is less skewed than the distribution of the γ_{li}, but remains right-skewed.

Figure 3: Posterior mean density estimate for the tooth-type random effects (F_k) and spatial effects (G_k) by tooth type. The densities are shifted to have mean zero.

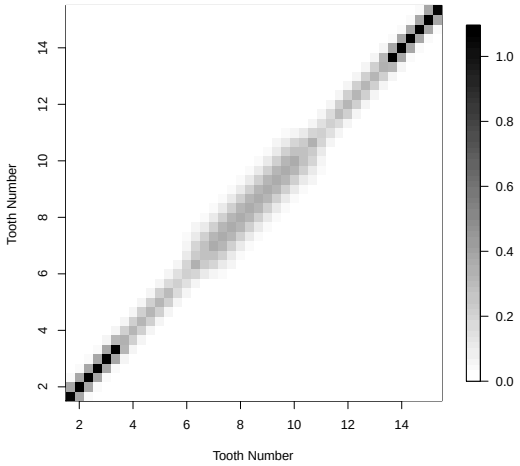


(a) Tooth-type random effect density, F_k (b) Spatial random effect density, G_k

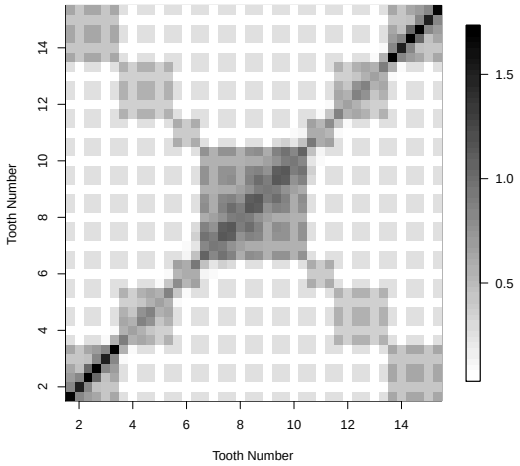
The covariance is summarized in Figure 4 and Table 2. The covariance of the spatial random effects θ in Figure 4a confirms that the variance is higher for molars than other teeth. Also, the covariance has larger spatial range for incisors than other teeth. The difference in spatial range is statistically significant. The bandwidth in Table 2 has 95% credible interval (1.23, 1.37) for incisors compared to (0.76, 0.88) for canines and less for other teeth. Table 2 shows that the covariance is anisotropic. For incisors there is stronger dependence for cheek-side/tongue-side pairs on the same tooth, and for other teeth cheek-side/tongue-side pairs have weak correlation. The sum of covariances of the subject-level random effects and spatial random effects in Figure 4b shows long-range dependence between molars on different sides of the jaw, as well and strong dependence between all incisors, as in Figure 2b's sample covariance.

To examine the effects of the covariance on the fitted values, Figure 5 plots the data and posterior mean of $\mu_i(\mathbf{s})$ for one subject's upper jaw. For comparison, we also fit the stationary Gaussian

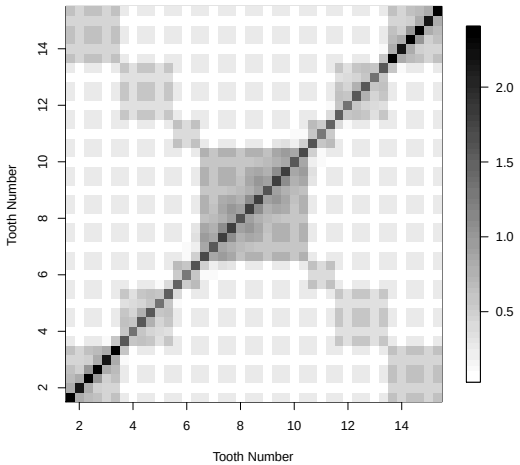
Figure 4: Posterior mean spatial covariance for the tongue side of the upper jaw. The posterior mean covariance is plotted for (a) the spatial effects $\boldsymbol{\theta}_i$, (b) the true CAL $\boldsymbol{\mu}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}\mathbf{a}_i + \boldsymbol{\theta}_i$, and (c) CAL responses $\boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i$. Panel (d) plots the posterior mean of the correlation of true CAL $\boldsymbol{\mu}_i$.



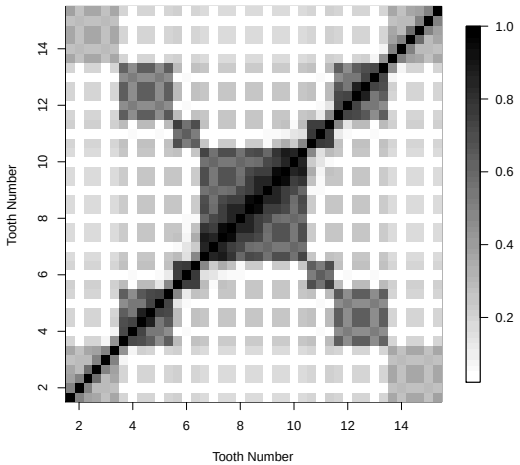
(a) Covariance of the spatial random effects



(b) Covariance of true CAL



(c) Covariance of measured CAL



(d) Correlation of true CAL

Table 2: Median (95% intervals) for the spatial correlation and informative missingness parameters.

	Incisors	Canines	Pre-Molars	Molars
Bandwidth, ρ_k	1.29 (1.23, 1.37)	0.82 (0.76, 0.88)	0.77 (0.72, 0.81)	0.54 (0.52, 0.57)
Anisotropy, ϕ_k	3.03 (2.63, 3.53)	0.00 (0.00, 0.02)	0.03 (0.00, 0.05)	0.59 (0.50, 0.67)
Informative missingness, C_k	0.31 (0.23, 0.39)	0.20 (0.01, 0.38)	0.24 (0.14, 0.35)	0.07 (0.02, 0.12)

model without non-randomly missing data, that is, with $M = 1$, $\rho_k \equiv \rho$, $\phi_k \equiv \phi$, and $c(t) \equiv 0$. For both models, the fitted values are considerably more smooth for incisors than teeth in the back of the mouth, especially for the upper jaw. The largest differences between the full and reduced models are for tooth 2 which is on the edge, and tooth 13 which is missing along with both of its neighbors. This is due to the significant dependence between CAL values and missing teeth, as shown by the positive value of C_k (the value of $c(t)$ for teeth of type k) in Table 2, and driven by relative severity of CAL at sites surrounding missing teeth. We also see the effect of assuming normality in the reduced form model. For example, on the tongue-side of tooth 15 the observed CAL is 3, 4, and 3 for the three sites. The reduced form model smooths the fitted values to be almost identical for these three sites, whereas the non-Gaussian full model fits closer to the observed 4mm measurement.

Table 3 summarizes the posterior of the fixed effects for CAL ($\mathbf{b}$) and missing teeth ($\mathbf{b}^*$). As expected, age is positively associated with both CAL and missing teeth. Similar to age, BMI has positive, but not statistically significant effect for CAL and missing teeth. Females are associated with a lower CAL, but with more missing teeth as compared to males. This may be due to the fact that most of the diseased teeth for females (with high degree of PD) were already absent, leaving

Table 3: Median (95% intervals) for the regression coefficients for the full model and the reduced model that assumes normality, stationarity, and random missingness.

	Full model		Reduced model
	CAL	Missing tooth	CAL
Age	0.06 (0.04, 0.09)	0.16 (0.12, 0.21)	0.16 (0.10, 0.22)
Female	-0.10 (-0.14, -0.07)	0.04 (0.01, 0.09)	-0.21 (-0.27, -0.15)
BMI	0.03 (0.00, 0.06)	0.05 (0.00, 0.09)	0.03 (-0.03, 0.09)
Smoker	0.04 (0.00, 0.07)	-0.04 (-0.08, 0.00)	0.09 (0.03, 0.14)
Hba1c	0.11 (0.07, 0.14)	0.01 (-0.03, 0.06)	0.24 (0.18, 0.30)

behind teeth with lower CAL values (indicative of mild to moderate PD) as compared to males. However, we add a note of caution here in interpreting this result, given that there is a predominance of females in this population (about 75%), and this is a common feature of this population (Johnson-Spruill et al., 2009). In addition, Hba1c (or glycemic control) also remain positively associated with CAL, which supports the initial premise of this study that poorly controlled Type-2 diabetic patients are more likely to develop periodontal disease than well-controlled diabetics. Table 3 also includes the regression coefficients for a the reduced model assuming normality, stationarity, and random missingness. There are several significant differences, for example, the 95% intervals for two models are non-overlapping for age, gender, and Hba1c.

Model comparison is challenging in the presence of informative sampling, since comparisons must account for both fit to observed values as well as the sampling process and relationship between the values and the sampling process. Many common Bayesian model select techniques such as Bayes factors (Carlin and Louis, 2008) and deviance information criteria (Spiegelhalter et al., 2002) require specifying a likelihood for the data, which is hard to define in the presence of informative missingness. Therefore, we inspect the adequacy of the simpler stationary Gaussian

model using the posterior predictive loss (PPL) approach of Daniels et al. (2012), which extends the posterior prediction model assessment approach of Gelfand and Ghosh (1998) to accommodate dependent data with missing responses. The PPL criteria bypasses the likelihood construction and presents a computationally convenient framework for quantifying the fit of the model by comparing features of the posterior predictive distribution to equivalent features of the observed data.

Consider the observed data for subject i as the set $(\boldsymbol{\delta}_i, \mathbf{y}_i)$, where $\boldsymbol{\delta}_i = [\delta_i(\mathbf{s}_1), ..., \delta_i(\mathbf{s}_N)]^T$ and $\mathbf{y}_i = [y_i(\mathbf{s}_1), ..., y_i(\mathbf{s}_N)]^T$. Let $T_i = T(\boldsymbol{\delta}_i, \boldsymbol{\delta}_i \circ \mathbf{y}_i)$ be a one-number summary of the observed outcomes for subject i, where $\boldsymbol{\delta}_i \circ \mathbf{y}_i = [\delta_i(\mathbf{s}_1)y_i(\mathbf{s}_1), ..., \delta_i(\mathbf{s}_N)y_i(\mathbf{s}_N)]^T$. Further let $T_i^* = T(\boldsymbol{\delta}_i^*, \boldsymbol{\delta}_i^* \circ \mathbf{y}_i^*)$ be the corresponding summary from the posterior predictive distribution from a given model (conditional on $\mathbf{X}_i$ and $\mathbf{Z}$, but not the random effects $\mathbf{a}_i$ and $\boldsymbol{\theta}_i$). Then PPL bases its model fit by comparing T_i to $E[T_i^* | \boldsymbol{\delta}, \mathbf{y}]$, where $\boldsymbol{\delta} = [\boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_n]$ and $\mathbf{y} = [\mathbf{y}_1, ..., \mathbf{y}_n]$.

Specifically, for some $k > 0$, define for each model,

$$D_k = \min_{\mathbf{a}} \sum_{i=1}^n E\left[\mathcal{L}\left(T_i^*, a_i, T_i\right) | \boldsymbol{\delta}, \mathbf{y}\right],$$

where $\mathcal{L}(T^*, a, T) = (T^* - a)^2 + k(T - a)^2$, so that $\mathbf{a}$ is the minimizer of this loss with respect to the posterior predictive distribution for the given model. For this choice of loss (see Gelfand and Ghosh, 1998; Daniels et al., 2012), $D_k = V + \frac{k}{k+1} M$, where

$$V = \sum_{i=1}^n \text{Var}\ (T_i^* | \boldsymbol{\delta}, \mathbf{y}),$$

$$M = \sum_{i=1}^n \left\{E\left(T_i^* | \boldsymbol{\delta}, \mathbf{y}\right) - T_i\right\}^2.$$

Models with smaller D_k are thus preferred.

Since our primary interest is in estimating the fixed effects in the mean, we take the summary statistic to be the mean of the observed outcomes, i.e. $T(\boldsymbol{\delta}_i, \boldsymbol{\delta}_i \circ \mathbf{y}_i) = \left\{\sum_{j=1}^{N} \delta_i(\mathbf{s}_j) y_i(\mathbf{s}_j)\right\} / \left\{\sum_{j=1}^{N} \delta_i(\mathbf{s}_j)\right\}$. For the Gaussian and stationary model with informative missingness, we find $M = 117.3$ and $V = 117.0$; for the full non-Gaussian and non-stationary model we find $M = 115.4$ and $V = 73.9$. Therefore, the full model has smaller D_k than the Gaussian and stationary model for all k.

6.1 Spatially-varying regression models

In the analysis above we have assumed that the effects of the covariates are constant throughout the mouth. However, given the dramatic differences in the residual processes in different areas of the mouth, it is natural to question if there are different factors influencing periodontal health in different regions, and thus different covariate effects in different regions. One option to exploring this possibility is to allow the covariate effects $\mathbf{b}$ to vary smoothly over space as in Gelfand et al. (2003). However, given the symmetries of the mouth, we explore spatially-varying effects by simply adding interactions between the subject-level covariate (age, gender, etc.) and spatial covariates indicating molar, pre-molar, and canine tooth types (incisor is the reference group), upper jaw, and a tooth in gap between teeth.

The results for the Gaussian/stationary residual model and our full residual model are given in Figure 6. We find significant interactions between age and smoking status and molar sites, indicating that disease progression for older patients and smokers is more rapid at molars than incisors. There is also some evidence that disease progression for smokers and older patients is slower at sites in the gap between teeth and on the lower jaw. For the predictors not shown

in Figure 6, there are no significant interactions for BMI and gender, and significantly negative (similar for both models) interactions between HbA1c and incisor and pre-molar. Compared to the Gaussian/stationary model, the posterior means for the full model are shrunk towards zero for most parameters, and the posterior standard deviations are consistently smaller for the full model. This could be the effect of dampening the impact of observations in the tail by properly modeling the residual distribution.

Finally, we note while our model treats the CAL data as continuous, it is actually rounded to the nearest millimeter. To test for sensitivity to this assumption, we compare the results with an interval censored model which accounts for this rounding. Denote $Y_i(\mathbf{s}) \in \{0, 1, ...\}$ as the observed value. Then, the unrounded value $y_i(\mathbf{s})$ following (1) is censored on the interval $(Y_i(\mathbf{s}) - 0.5, Y_i(\mathbf{s}) + 0.5)$, which can be modeled by treating $y_i(\mathbf{s})$ as a latent variable similar to z_i in (14). Figures 6e and 6f show the results of using this censored data model to account for rounding compared to modeling CAL directly and ignoring rounding. The results do not appear to be sensitive to the treatment of rounding.

7 Discussion

In this paper, we have proposed a flexible model for spatially-referenced periodontal data. Our framework is computationally convenient, and allows the response distribution to have differing degrees of non-normality in different regions of the mouth, the spatial covariance to be non-stationarity and the effect of non-random missingness to be varying for different tooth types. The simulation study demonstrates failing to account for these complexities leads to inefficient estima-

tion of the fixed effects.

Our current simulation study and data analysis models the spatial dependence and non-normality primarily as a means to obtain precise estimates of covariate effects. However, in other settings, these features may be of interest on their own. For example, one might consider monitoring a patient over time, in which the spatial modeling developed here could be extended to the spatio-temporal setting to detect regions with deteriorating periodontal health. With a single subject, it may be difficult to allow the density of the spatial random effects to vary spatially. In this case, it could be held constant for all sites to permit a stationary non-Gaussian modeling. Also, in settings with replications but without the symmetry exhibited by PD data, it may be possible to let the density of the spatial random effects vary smoothly across space rather than having four distinct models for four distinct regions as considered here. Here the methods for spatially-varying density functions, e.g., Reich and Fuentes (2007) might be useful.

References

Antoniak, C. E. (1974), "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *Annals of Statistics*, 2, 1152–1174.

Burt, B. et al. (2005), "Position paper: Epidemiology of Periodontal Diseases," *Journal of Periodontology*, 76, 1406–1419.

Carlin, B. P. and Louis, T. A. (2008), *Bayesian methods for data analysis*, Chapman & Hall/CRC.

Chen, M., Shao, Q., and Ibrahim, J. (2000), *Monte Carlo methods in Bayesian computation*, Springer Verlag.

Daniels, M. J., Chatterjee, A. S., and Wang, C. (2012), "Bayesian Model Selection For Incomplete Data using the Posterior Predictive Distribution," *Biometrics*, 68, 1055–1063.

Diggle, P., Manezes, R., and Su, T. (2010), "Geostatistical inference under preferential sampling (with discussion)," *Journal of the Royal Statistical Society: Series C*, 59, 191–232.

Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002), *Analysis of longitudinal data*, Oxford University Press.

Do, L., Spencer, J., Roberts-Thomson, K., Ha, D., Tran, T., Trinh, H., et al. (2003), "Periodontal disease among the middle-aged Vietnamese population," *Journal of the International Academy of Periodontology*, 5, 77–84.

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

— (1974), "Prior distribution on spaces of probability measures," *The Annals of Statistics*, 2, 615–629.

Fernandes, J., Wiegand, R., Salinas, C., Grossi, S., Sanders, J., Lopes-Virella, M., and Slate, E. (2009), "Periodontal disease status in Gullah African Americans with type 2 diabetes living in South Carolina," *Journal of Periodontology*, 80, 1062–1068.

Fonseca, T. C. O. and Steel, M. F. J. (2011), "Non-Gaussian spatiotemporal modelling through scale mixing," *Biometrika*, 98, 761–774.

Fuentes, M. (2002), "Spectral methods for nonstationary spatial processes," *Biometrika*, 89, 281–298.

Gelfand, A. E. and Ghosh, S. (1998), "Model choice: A minimum posterior predictive loss approach," *Biometrika*, 85, 1–11.

Gelfand, A. E., Kim, H. K., Sirmans, C. F., and Banerjee, S. (2003), "Spatial modelling with spatially varying coecient processes," *Journal of the American Statistical Association*, 98, 387–396.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005), "Bayesian Nonparametric Spatial Modeling with Dirichlet Process Mixing," *Journal of the American Statistical Association*, 100, 1021–1035.

Gilks, W., Best, N., and Tan, K. (1995), "Adaptive rejection Metropolis sampling within Gibbs sampling," *Applied Statistics*, 455–472.

Griffin, J. E. and Steel, M. F. J. (2006), "Order-based dependent Dirichlet processes," *Journal of the American Statistical Association*, 101, 179–194.

Hand, D., Mannila, H., and Smyth, P. (2001), *Principles of Data Mining*, The MIT Press, Cambridge, MA.

Higdon, D., Swall, J., and Kern, J. (1999), "Non-stationary spatial modeling," in *Bayesian Statistics 6 - Proceedings of the Sixth Valencia Meeting*, J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, (editors). Clarendon Press - Oxford, pp. 761–768.

Hjort, N., Holmes, C., Müller, P., and Walker, S. (2010), *Bayesian Nonparametrics*, Cambridge University Press.

Jara, A., Quintana, F., and Martín, E.-S. (2008), "Linear mixed models with skew-elliptical distributions: a Bayesian approach," *Computational Statistics and Data Analysis*, 52, 5033–2045.

Johnson-Spruill, I., Hammond, P., Davis, B., McGee, Z., and Louden, D. (2009), "Health of Gullah Families in South Carolina With Type 2 Diabetes," *The Diabetes Educator*, 35, 117–123.

López, R., Fernández, O., Jara, G., and Baelum, V. (2001), "Epidemiology of clinical attachment loss in adolescents," *Journal of Periodontology*, 72, 1666–1674.

Nicholls, C. (2003), "Periodontal disease incidence, progression and rate of tooth loss in a general dental practice: The results of a 12-year retrospective analysis of patient's clinical records," *British Dental Journal*, 194, 485–488.

Paciorek, C. J. and Schervish, M. J. (2006), "Spatial Modelling using a new class of nonstationary covariance functions," *Environmetrics*, 17, 483–506.

Pati, D., Reich, B. J., and Dunson, D. B. (2011), "Bayesian geostatistical modeling with informative sampling locations," *Biometrika*, 98, 35–48.

Reich, B. (2012), "Spatiotemporal quantile regression for detecting distributional changes in environmental processes," *Journal of the Royal Statistical Society: Series C*, 64, 535–553.

Reich, B. and Bandyopadhyay, D. (2010), "A latent factor model for spatial data with informative missingness," *Annals of Applied Statistics*, 4, 439–459.

Reich, B. and Fuentes, M. (2007), "A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields," *Annals of Applied Statistics*, 1, 249–264.

Reich, B., Hodges, J., and Carlin, B. (2007), "Spatial analyses of periodontal data using conditionally autoregressive priors having two classes of neighbor relations," *Journal of the American Statistical Association*, 102, 44–55.

Rodriguez, A. and Dunson, D. B. (2011), "Nonparametric Bayesian models through probit stick-breaking processes," *Bayesian Analysis*, 6, 145–178.

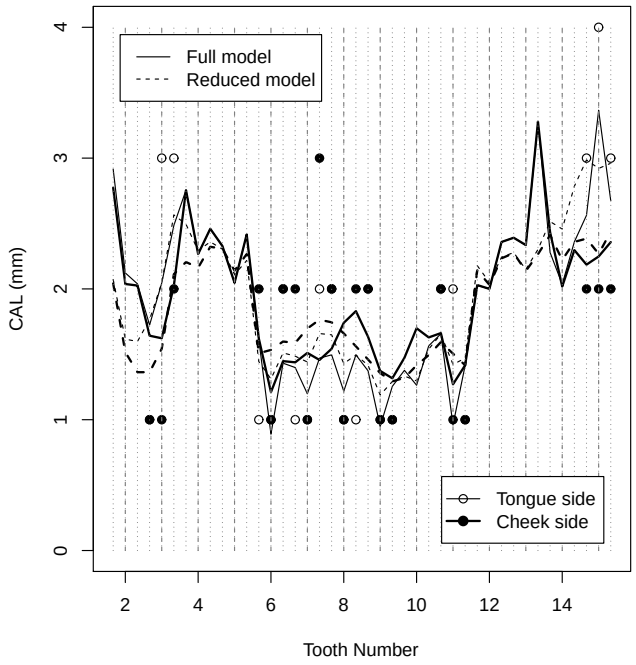
Sampson, P. D. and Guttorp, P. (1992), "Nonparametric estimation of nonstationary covariance structure," *Journal of the American Statistical Association*, 87, 108–119.

Schmidt, A. M. and O'Hagan, A. (2003), "Bayesian Inference for Nonstationary Spatial Covariance Structures via Spatial Deformations," *Journal of the Royal Statistical Society, Series B*, 65, 743–775.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639–650.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002), "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society: Series B*, 64, 583–639.

Figure 5: Data and posterior mean of $\mu(\mathbf{s})$ for the upper jaw (top) and lower jaw (bottom) of one subject. The reduced model is stationary, Gaussian, and does not account for non-random missing teeth, i.e., $\alpha_{ik} \overset{iid}{\sim} \mathrm{N}(m_k, \tau_k^2)$, $\rho_k \equiv \rho$, $\phi_k \equiv \phi$, $\gamma_{li} \overset{iid}{\sim} \mathrm{N}(0, \omega^2)$, and $a(t) \equiv 0$.



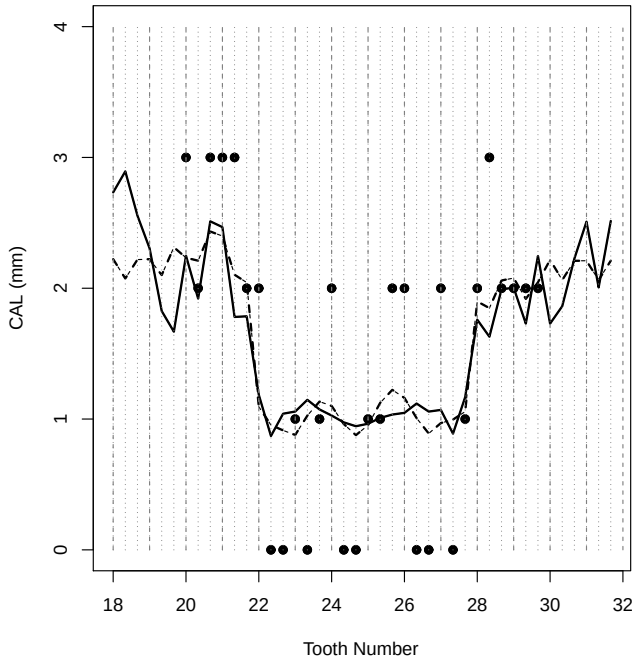
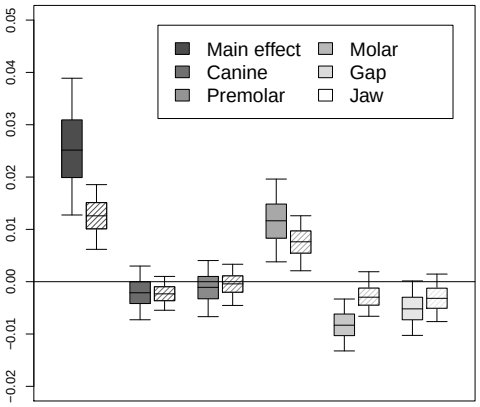
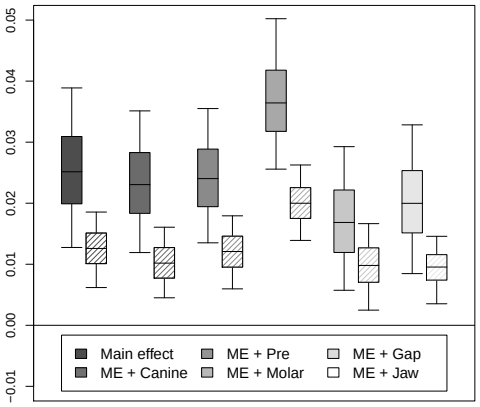


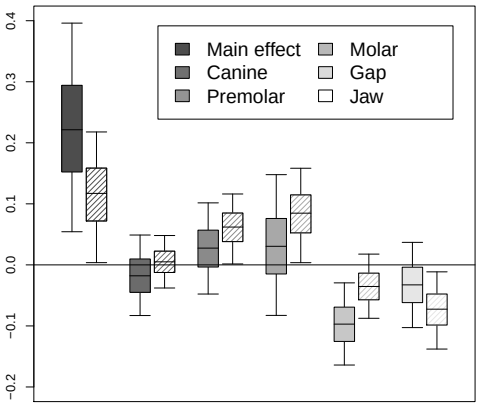
Figure 6: Covariate effects on mean CAL for the simple model with Gaussian and stationary errors and the full model with non-Gaussian and non-stationary errors. Panels (e) and (f) compare results of modeling CAL directly (as in Panels (a)-(d)) and ignoring rounding versus modeling rounding using latent variables.



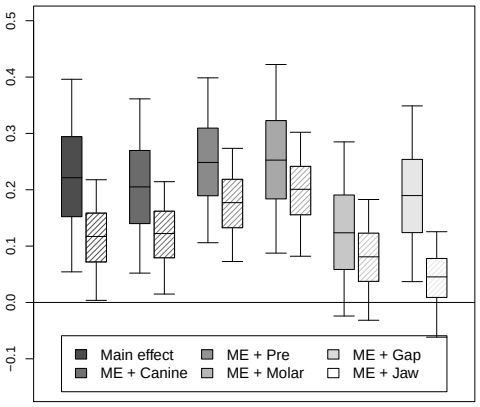
(a) Age effects, simple (solid) vs full (dashed)



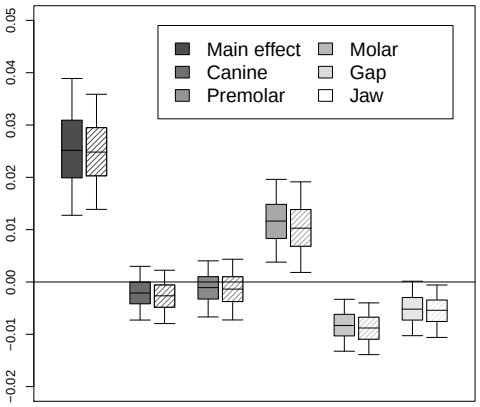
(b) Age effects, simple (solid) vs full (dashed)



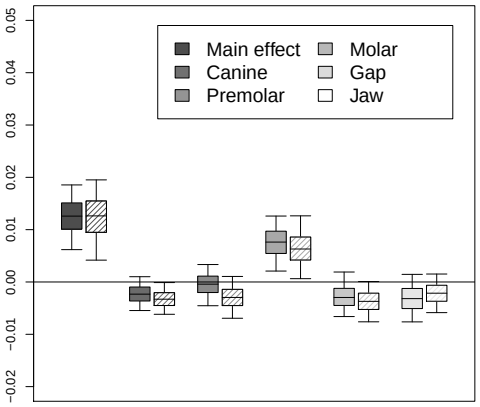
(c) Smoker effects, simple (solid) vs full (dashed)



(d) Smoker effects, simple (solid) vs full (dashed)



(e) Age effects for the simple model ignoring (solid) and accounting for (dashed) rounding



(f) Age effects for the full model ignoring (solid) and accounting for (dashed) rounding