Published in final edited form as:
Stat Methods Med Res. 2011 April ; 20(2): 85–102. doi:10.1177/0962280210372453.

A spatial Beta-Binomial model for clustered count data on dental caries

Dipankar Bandyopadhyay[1,*], **Brian J. Reich**[2], and **Elizabeth H. Slate**[1]

[1]Division of Biostatistics and Epidemiology, Medical University of South Carolina

[2]Department of Statistics, North Carolina State University

Abstract

One of the most important indicators of dental caries prevalence is the total count of decayed, missing or filled (DMF) surfaces in a tooth. These count data are often clustered in nature (several count responses clustered within a subject), over-dispersed, as well as spatially referenced (a diseased tooth might be positively influencing the decay process of a set of neighboring teeth). In this paper, we develop a multivariate spatial Beta-Binomial (BB) model for these data that accommodates both over-dispersion as well as latent spatial associations. Using a Bayesian paradigm, the re-parameterized marginal mean (as well as variance) under the BB framework are modeled using a regression on subject/tooth-specific co-variables and a conditionally autoregressive (CAR) prior that models the latent spatial process. The necessity of exploiting spatial associations to model count data arising in dental caries research is demonstrated using a small simulation study. Real data confirms that our spatial BB model provides a superior estimation and model fit as compared to other sub-models that do not consider modeling spatial associations.

Keywords

Beta-Binomial; conditionally auto-regressive (CAR); count data; dental caries; spatial

1 Introduction

Dental caries, otherwise known as tooth decay, is an infectious localized disease and is one of the most prevalent chronic diseases of people worldwide (affecting around 80% of the world population) and is the primary cause of oral pain and tooth loss, as described in the US Surgeon General's report[1], Fejerskov and Kidd[2], Kidd *et al.*[3]. Its impact on individuals and communities as a result of pain and suffering, impairment of function and reduced quality of life is considerable[4]. In the USA, caries is the most common chronic disease of childhood, and is five times more common than asthma[1]. Caries is primarily influenced by poor oral hygiene (brushing/flossing habits), frequent consumption of sugar based products as well as socio-economic factors such as poverty status, access to care, etc. Despite the widespread decline in caries prevalence and severity in high-income countries over the past few decades, disparities still remain and many children and adults still develop caries[5,6].

Caries is caused by organic acids that originates from the microbial fermentation of carbohydrates[7] which affects the mineral constitution of teeth. Primarily, three essential interactive factors causes caries: the host, represented by teeth and saliva; the oral microbial

* Address of correspondence: Division of Biostatistics and Epidemiology, Department of Medicine, 135 Cannon Street Suite 303, Charleston, SC 29425, U.S.A. bandyopd@musc.edu.

flora; and poor dietary habits[8,9]. The destroyed tooth surfaces can never fully regenerate, though remineralization of very small carious lesions is possible by an optimal level of dental hygiene. Personal hygiene care involves daily proper brushing and flossing, dietary modifications (reduced consumption of sugar), etc. Use of dental sealants and fluoride therapy is recommended to provide protection against dental caries. The cost of treatment of caries is extremely costly, representing the fourth most expensive disease to treat in most of the industralized world[7]. For further details on dental caries, the interested reader is referred to the very interesting seminar article by Selwitz *et al.*[10].

In this paper, we analyze data on 100 subjects from a clinical study[11] conducted at the Medical University of South Carolina (MUSC) among Type-2 diabetic Gullah-speaking African-Americans, whose dental caries status remains highly unexplored. Clinical studies examining dental caries status produces a clustered multivariate set-up[12] where measures are taken for tooth surfaces (whether a particular tooth surface is decayed, or missing or filled), clustered within a tooth, which in turn, remains clustered within the oral cavity of that subject. The total count of decayed, missing and filled surfaces per tooth, also known as the 'DMFS' index, is recorded per tooth. There are 32 teeth in total for the adult dentition) for each subject as a measure of caries status. The DMFS index[13] is the most popular caries marker and is indicative of the cumulative severity of caries status for that particular tooth.

Figure 1 shows the density histogram of tooth level raw DMFS counts (n_s, $s = 1,\ldots,32$) for our study subjects. Because the 'maximum' tooth level DMFS count can be either 4 or 5 (depending on the tooth location, see Section 2 for more details), we have a discrete 'Binomial' count setup. The raw plot reveals a somewhat *U*-shaped structure, indicating some over-dispersion (extra-Binomial variation) in the form of excessive observed DMFS counts of 0 and 5. This raises the question of whether a considerable number of subjects truly have caries-free (DMFS=0) or advanced carious (DMFS=5) teeth, or whether this large number of 0 and 5 is a result of some 'latent heterogeneity', with many subjects simply having a low/high 'true' caries index. Also, the observed (raw) DMFS index of a tooth might be influenced (spatially) by the caries status of a set of neighboring teeth. Here, we develop a spatial Beta-Binomial (BB) model for multivariate count data to address two broad questions related to the dental caries status of this Gullah-speaking study sample: (a) How do the potential covariates like age, gender, smoking habit, brush-floss habits, poverty status, etc influence caries status at the tooth-level after accounting for spatial association? and (b) How does spatial association influence development of caries lesions in a group of neighboring teeth? The BB model has been highly explored in the statistical literature, particularly in toxicology[14,15], disease incidence[16,17] and radiation[18] to model binomial data with excess unobserved heterogeneity. In the area of disease mapping, interest often lies in identification of 'extreme areas', which may arise in proximity to one another either in a smooth spatial surface, or as isolated 'hot spots' or 'low spots' which are quite distinct from the neighboring teeth[19]. It is presumed that a particular tooth within the oral cavity experiencing caries lesions tends to affect a set of neighboring teeth of the same subject. Also, the responses are 'clustered' in nature, i.e. a group of 32 teeth within the oral cavity of a subject seem to share similar subject-level characteristics and can be considered 'stochastically' independent of the set of teeth for another subject. To analyze these correlated (and possibly over-dispersed) count data with latent spatial autocorrelation, we use a conditionally autoregressive (CAR) specification for the spatial random effects, pioneered by Besag[20]. Using a convenient re-parametrization[14] of the BB parameters, we use a CAR logistic regression model to determine the effect of subject/tooth specific covariables on DMFS outcomes. The CAR specification assumes a Markov Random Field (MRF)[21] model such that the conditional distribution of a spatial unit (tooth) depends on only the outcomes of a pre-defined set of neighboring teeth. The (latent) spatial model thus borrows strength across neighboring teeth to provide smoothing and improve estimates of

true caries status for each tooth. MRF's include a wide class of spatial models such as auto-Gaussian models for spatial Gaussian processes, auto-logistic models for binary spatial processes, auto-Gamma models for non-negative continuous processes and auto-Poisson models for spatial count processes. The CAR specification is actually an auto-Gaussian model analogous to an auto-regressive time series model[19]. However, some practical questions associated with incorporating spatial dependencies include proper quantification of spatial autocorrelation through defining a proper neighborhood structure, its impact on covariate assessments and enhancement of statistical efficiency through standard errors reduction. The monograph by Lawson *et al.*[22] provides a comprehensive review of CAR models in the context of disease mapping while Banerjee *et al.*[23] details recent advances in hierarchical Bayesian methodologies for spatial/spatio-temporal data.

Though CAR models have been widely explored in the context of disease mapping, their application for analysis of multivariate dental caries data is relatively sparse and comes with unique challenges. First, there is clustering induced at the subject level. Also, unlike many applications where the spatial structure is clearly defined by a grid or geopolitical boundaries, the spatial structure within a mouth must be carefully chosen. Applications of spatial models for dental disease outcomes are relatively new. Very recently, various CAR specifications within the mouth in the context of periodontal (gum) disease modeling has been explored in[24,25]. To the best of our knowledge, there had been no previous exploration of the CAR specification to model spatially correlated count data on dental caries. The spatial nature of caries progression within a mouth is not established; we adopt a neighborhood structure accomodating adjacent teeth as well as (some) contacting teeth on opposite jaws. We show that accounting for this neighborhood spatial association structure dramatically improves the predictive performance of our model. Our approach is Bayesian, which has the ability to incorporate background (prior) information about the unknown parameters in the model, both for subject/tooth level covariates as well as for spatial dependencies. Bayesian inference with proper prior elicitations, powered by Markov chain Monte Carlo (MCMC) steps provides inference and does not depend on asymptotic calculations. Our models allow us to determine the degree of extra-binomial variation in the data. We also compare our spatial CAR models to alternative models that do not distinguish the effect of spatial associations among teeth using a Bayesian model choice criterion[26] and cross-validatory techniques[27].

The outline of the paper is as follows. Section 2 describes the DMFS count data that motivated this research. In Section 3, we propose our spatial BB model and explore the association structures needed to develop the spatial model. Section 4 proposes Bayesian inference based on the data likelihood, prior and hyperprior specifications as well as various model selection/assessment tools. In Section 5, we apply our spatial model to the dental caries data, use model selection and assessments tools to determine the best model and summarize findings. We explore a small simulation study in Section 6 to study the effect of accounting for spatial associations on estimating covariate effects. Conclusions and future developments are in Section 7.

2 Motivating Data

The motivating data were collected from a clinical study[11] conducted by the Center for Oral Health Research (COHR) of the Medical University of South Carolina (MUSC) as part of the South Carolina Center for Biomedical Research Excellence (COBRE) Program for Oral Health. The study assessed caries status among Type-2 diabetic Gullah-speaking (or simply Gullah) African-Americans (13 years or older) residing in the coastal sea-islands of South Carolina. All subjects answered a detailed questionnaire focussed on their social, medical and dental history and underwent an oral exam. To develop our methods, we selected a

random sample of 100 subjects with complete covariate information and DMFS counts for all the 32 teeth within a subject. Each quadrant of teeth (consisting of a cluster of 8 teeth, 2 in each jaw) in a human mouth[28] is made up of (a) the non-anterior teeth (3 molars and 2 pre-molars) and (b) the anterior teeth (1 incisor and 2 canines). For measuring caries, each non-anterior tooth contributes 5 surfaces (occlusal, mesial, distal, facial and lingual) while an anterior tooth contributes 4 surfaces (the occlusal not being recorded) to the response. Figure 2 illustrates the different surfaces for permanent dentition within a mouth. Handling the *M* part of the DMFS index varies[29] because the extracted/missing tooth calls for an arbitrary allocation of the number of decayed surfaces for that tooth. In this paper, we use the DM_5FS convention, i.e. assign 4 surfaces for an extracted anterior tooth and 5 for an extracted non-anterior tooth when extraction is reported due to caries. If the subject reports a tooth loss due to a cause other than caries, this tooth does not contribute to the data analysis. Additionally, several subject level covariates were collected including Age (in years), Gender (0=Male, 1=Female), Smoking status (0=Never, 1=Smoker), Brush-Floss (1 = Brushed twice and flossed once every day, 0= otherwise) and Poverty (1=Below poverty line, 0=Above poverty line). The mean age of the subjects in the sample is 53 years with a range of 27-73 years. Although study recruitment was gender blind, females participated at a higher rate (75%) than males and this is not unusual among Gullah-speaking African Americans[30]. There are only about 8% 'former smokers' in our sample. Thus, in order to avoid a separate category for 'smokers' and keep our autologistic model relatively simple, we collapsed the groups of 'former smokers' and 'present smokers' into 'smokers', which comprises about 36% of the study sample. The poverty status was determined according to family per capita income level, i.e. total income generated by the family divided by the number of family members. If this value was less than $5000, the subject was classified as below the poverty line. About 34% of the subjects in the sample live below the poverty line and 24% reported to have brushed twice and flossed once every day. All the above covariates are subject-level and do not vary within a mouth. We also included a tooth-level covariate, i.e. Molar Indicator (1=Tooth is molar, 0=Otherwise) to allow assessment of the severity of caries among the molars as compared to non-molars for this sample. In this paper, we explore the degree of tooth-level spatial association influencing DMFS counts while controlling for the effects of the subject/tooth-level covariates.

3 Statistical Model

Let $\{Y_i(s) : (i, s) \in D\}$ be the DMFS count collected for tooth s of subject i, where D denotes the set of indices for the observed data, $i = 1,\ldots,\{m = 100\}$ and $s = 1,\ldots, \{n = 32\}$ and let $p_i(s)$ be the probability of having a diseased (D/M/F) surface in the sth tooth. We assume that there is an equal probability of experiencing a D, M or F surface in a particular tooth. Conditional on $p_i(s)$, we model $Y_i(s)$ as Binomial(n_s, p_i(s)), where $n_s = 4$ where s references on incisors and canines and $n_s = 5$ for pre-molars and molars. To account for over-dispersion, the random variable $p_i(s)$ follows a Beta distribution with parameters $a_i(s)$ and $b_i(s)$, with density function is given by

$$f(p_i(s)|a_i(s), b_i(s)) = \frac{\Gamma(a_i(s)+b_i(s))}{\Gamma(a_i(s))\Gamma(b_i(s))} p_i(s)^{a_i(s)-1}(1 - p_i(s))^{b_i(s)-1} I_{(0,1)}(p_i(s)) \quad (1)$$

where $a_i(s), b_i(s) > 0$, $I_A(x)$ denotes the indicator function of the event $x \in A$, and $\Gamma(.)$ denotes the gamma[31] function. Different choices of $a_i(s)$ and $b_i(s)$ lead to a variety of shapes of the density of $p_i(s)$, viz. U-shaped, J-shaped, reverse-J shaped, as well as constant (i.e. Uniform(0,1), when $a_i(s) = b_i(s) = 1$). The use of a Beta distribution to model variability in the binomial probability parameter was first proposed by Skellam[32], however its extension

to unequal n_s (as in our case) was first considered by Williams[14] in toxicological applications. Then, $Y_i(s)$ is Beta-Binomial (BB) on the support $\{0,\ldots,n_s\}$ with the distribution given by

$$P(Y_i(s)=y_i(s)|a_i(s),b_i(s))=\binom{n_s}{y_i(s)}\frac{B(a_i(s)+y_i(s),n_s+b_i(s)-y_i(s))}{B(a_i(s),b_i(s))} \quad (2)$$

where $B(.,.)$ is the beta[31] function. Conditional on $a_i(s)$ and $b_i(s)$, the mean and variance of $Y_i(s)$ are respectively $n_s\frac{a_i(s)}{a_i(s)+b_i(s)}$ and $\frac{n_s a_i(s)b_i(s)(n_s+a_i(s)+b_i(s))}{(a_i(s)+b_i(s))^2(1+a_i(s)+b_i(s))}$. Our goal is to determine how the subject/tooth-level covariates are predictive of DMFS responses, controlling for spatial associations. Looking into the mean and variance expression above, it is not clear how a regression framework can be developed to model $a_i(s)$ and $b_i(s)$ separately, or forcing dependence between them which will eventually determine $Y_i(s)$. To have meaningful interpretation of the parameters of the Beta density, we re-parameterize[17] $a_i(s)$ and $b_i(s)$ as $\mu_i(s)=\frac{a_i(s)}{\gamma(s)}$, where $\gamma(s)=a_i(s)+b_i(s)$. Thus, the original parameters can now be written as $a_i(s)=\gamma(s)\mu_i(s)$ and $b_i(s)=\gamma(s)(1-\mu_i(s))$. Note that $E(p_i(s)|\gamma(s),\mu_i(s))=\mu_i(s)$ and $\mathrm{Var}(p_i(s)|\gamma(s),\mu_i(s))=\frac{\mu_i(s)(1-\mu_i(s))}{\gamma(s)+1}$. Given the new parametrization, $\mu_i(s)$ is the mean of the probability $p_i(s)$ distributed as Beta, whereas $\gamma(s)$ denotes the shape of the distribution of $p_i(s)$. Thus, the mean and variance of the BB distribution becomes $n_s\mu_i(s)$ and $n_s\mu_i(s)(1-\mu_i(s))\frac{\gamma(s)+n_s}{\gamma(s)+1}$ respectively. Note that the overdispersion parameter for the BB model is given by $\frac{\gamma(s)+n_s}{\gamma(s)+1}\in(1,n_s)$ and the marginal variance of the BB distribution approaches the Binomial variance when $n_s=1$ or $\gamma(s)\to\infty$. However in our case, $n_s\in\{4,5\}$. Using this formulation allows one to add overdispersion directly without perturbing $E(Y_i(s)|.)$[33], thus matching a BB model when there is an excess of 'Binomial' samples collected at some specified level of a covariate X.

Under a generalized linear mixed model (GLMM) framework[34], a regression incorporating heterogeneity at the subject and tooth level is defined as

$$\mu_i(s)=F(\beta_0+\mathbf{X_{1is}}^T\beta+U_{1i}(s)) \quad (3)$$

where the inverse link $F(.)$ can be specified as the symmetric 'logit' link, i.e. $F(x)=\exp(x)/(1+\exp(x))$, however other choices like probit and complimentary log-log are also possible, $\boldsymbol{X_{1is}}$ denotes the vector of subject/tooth-level covariates (including the intercept term) with the corresponding fixed-effects parameter vector β, β_0 is the intercept term and $U_{1i}(s)$ is the (spatial) random effect corresponding to the count response of the $(i,s)^{\text{th}}$ tooth. Choices for $\gamma(s)$ are discussed in Section 4.1. Under a Markov Random Field (MRF) assumption, the full conditional distributions of $U_{1i}(s)$ are specified as,

$$p(U_{1i}(s)|u_{1i}(s'),s\neq s',\rho,\sigma^2_{SP})=N\left(\rho\sum_{s':s'\sim s}\frac{w_{ss'}}{w_s}u_{1i}(s'),\frac{\sigma^2_{SP}}{w_s}\right),\ s,s'=1,\ldots,n \quad (4)$$

independently for each i, where $s \sim s'$ denotes that tooth s is a neighbor of tooth s', $w_{ss'} = 1$ if $s \sim s'$ and = 0 otherwise, $m_s = \Sigma_{s'} w_{ss'}$ is the total number of neighbors of tooth s and $\sigma^2_{SP} > 0$ controls the magnitude of spatial variation. By Brook's Lemma[23], full conditionals in (4) are uniquely determined by the joint density $U_{1i}|\rho, \sigma^2_{sp} \sim \text{MVN}(0, \text{Variance} = \sigma^2_{sp} Q(\rho)^{-1})$, i.e. (unconditionally) the latent vector $\mathbf{U}_{1i}$ follows a zero-mean multivariate normal prior with a 'CAR' covariance structure. Here, $Q(\rho) = \text{Diag}(m_s) - \rho W$, $W = (w_{ss'})$ is the adjacency matrix of the *graph* representing our region and ρ is the smoothing parameter controling the degree of spatial dependence. Because the population is homogenous, the spatial effects $\boldsymbol{U1i}$ for the m subjects are modelled as independently and identically distributed (i.i.d) according to this $\text{CAR}(\rho, \sigma^2_{sp})$ prior distribution. Requiring $\rho \in (0, 1)$ ensures propriety of the CAR distribution. Because one of our motivations is to estimate the degree of spatial association in the data, we did not use the *intrinsic autoregressive* (IAR) formulation (choosing $\rho = 1$) based on pairwise differences[35], as the IAR is improper and does not have a parameter to control the strength of spatial dependence. To avoid the computation of eigenvalues in order to guarantee propriety as pointed out in Cressie[36], we work with the scaled adjacency matrix B defined by $B = D^{-1}W$, where $D = \text{Diag}(m_s)$ as suggested in Carlin and Banerjee[37]. The 'adjacency' map for dental caries outcomes is constructed in Table 1. The tooth numbering starts with the 'third molar' on the mandibular (upper jaw) left quadrant and goes clock-wise, till the 32nd tooth, i.e. the 'third molar' in the maxillary (lower jaw) left quadrant. The neighborhood structure for each tooth consists of at most 2 teeth on each sides within the same jaw. In addition, the non-anterior teeth (molars and pre-molars) with an occlusal surface also have the opposing non-anterior tooth on the other jaw as a neighbor. Thus, $m_s \in \{3, 4, 5\}$, i.e. total number of neighbors for any tooth can be at most 5.

4 Bayesian Inference

4.1 Likelihood, choice of priors and posterior distributions

Considering $\Omega = (\beta_0, \beta, \rho, \sigma^2_{sp}, \gamma(s))$ as the parameter vector in our spatial BB regression model the primary goal is to estimate $\boldsymbol{\Omega}$ and draw inference on these parameters controlling for spatial association. Then from (2) and (3), the joint data-likelihood (conditional on the (spatial) random-effects $\boldsymbol{U}_1 = \{U_i(s) : (i, s) \in D\}$) is given by

$$L(\boldsymbol{\Omega}; \mathrm{U}_1, \mathrm{X}, \mathrm{y}) = \prod_{(i,s)\in D} \left[\binom{n_s}{y_i(s)} \frac{\left[\prod_{k=0}^{y_i(s)-1} (\gamma(s)\mu_i(s)+k)\right] \left[\prod_{k=0}^{n_s-y_i(s)-1} (\gamma(s)(1-\mu_i(s))+k)\right]}{\left[\prod_{k=0}^{n_s-1} (\gamma(s)+k)\right]} \right] \tag{5}$$

where $\prod_{k=0}^{-1} (a+k) \equiv 1$. Clearly, the likelihood in (5) does not belong to the exponential family and is computationally awkward; moreover frequentist maximum likelihood (ML) based estimation using quasi-likelihood[38], or other methods in our spatial set-up is quite complicated, specially for an applied audience, requiring situation specific computer code. This motivates a Bayesian inference for this problem, relying on MCMC techniques which is straightforward to implement using available freeware WinBUGS[39]. The Bayesian method provides the entire posterior distribution of the parameters and any arbitrary parameter functionals considering both the data likelihood and the priors assigned to the parameters. Next, we investigate the choice of prior and hyperprior distributions for the model parameters.

Since we have no prior information from historical data or experiment, we specify weakly-informative prior opinions on the fixed effects regression parameters β to obtain well-defined (proper) posteriors. We assume the elements of the parameter space Ω are independently distributed. Specifically, we assign weakly informative i.i.d. Normal(0, precision = 0.25) priors on the elements of β. This implies that the density of the associated odds-ratio centered at 1 with 95% intervals (e^{-4}, e^{4}), will include a sufficiently wide range of prior guesses[40]. For the intercept term β_0, we use a flat prior (dflat() option in WinBUGS). For prior choices on the spatial CAR parameters, we proceed as follows. We take $\rho \sim$ Beta(5, 1) with mean 0.83 to elicit our prior belief of a reasonably high spatial association in the data and a hyperprior on the spatial precision parameter $\sigma_{sp}^{-2} \sim$ Gamma(0.1, 0.1) (with mean 1, variance 10) to represent a significantly less vague belief[41], often used in spatial modeling practice. We consider three different scenarios for prior on $\gamma(s)(> 0)$. Case (a): $\log(\gamma(s)) = \gamma_0 + U_2(s)$, such that $U_2(s) \sim \mathrm{CAR}(\rho, \sigma_{sp}^2)$ and $\gamma_0 \sim$ N(0, Precision = 0.001); Case (b): $\gamma(s) = \gamma_1 \sim$ diffuse Gamma(0.1, 0.01) and Case (c): $\gamma(s) \sim$ diffuse Gamma(0.1, 0.01), where Gamma(a, b) denote a Gamma density with mean a/b and variance a/b^2. While (a) is primarily motivated by the fact that $\gamma(s)$ can be spatially dependent, (b) and (c) considers a much simple model for $\gamma(s)$ by putting a diffuse Gamma prior.

The posterior conclusions from our Bayesian analysis will be based on the joint posterior distribution of all the model parameters conditional on the data which is obtained by combining the likelihood given in (5) and the joint prior densities using Bayes' theorem:

$$p(\mathbf{\Omega}, \mathrm{U}_1|\mathrm{X}, \mathrm{y}) \propto \mathrm{L}(\mathbf{\Omega};\mathrm{U}_1, \mathrm{X}, \mathbf{y})\times\{\prod_{(i,s)\in D} \pi_0(U_{1i}(s)|\rho, \sigma_{sp}^2)\}\times\pi_1(\rho)\times\pi_2(\sigma_{sp}^2)\times\pi_3(\gamma)\times\pi_4(\beta_0)\times\pi_5(\beta) \quad (6)$$

where $\pi_j(.)$, $j = 0,\ldots, 5$ denote the prior/hyperprior distributions on the model parameters as described above.

The relevant MCMC steps were implemented readily using freeware package WinBUGS. To improve convergence, we used hierarchical centering[42], i.e. all covariates were centered around the mean and divided by their standard deviation. We used 15000 iterations with an initial burn-in of 10000, with 2 different chains with arbitrary starting values to ensure convergence. Based on examination of the trace plots, auto-correlation plots and Gelman-Rubin diagnostic $\hat{\mathrm{R}}$[43], convergence was excellent with proper mixing of the 2 chains. Posterior inference is based on 5000 MC samples after discarding the initial burn-in samples.

4.2 Model Selection and assessment

Our initial model selection was performed using the Deviance Information Criterion (DIC) of Spiegelhalter *et al.*[26]. DIC reflects the goodness of fit as well as the complexity of the hierarchical model within the Bayesian paradigm and is considered to be a Bayesian version of the Akaike Information Criterion (AIC). It is defined as $DIC = \bar{D} + p_D$, where $\bar{D} = E(D(\mathbf{\Theta})|\mathbf{y})$ is the posterior mean of the deviance, $\mathbf{\Theta}$ is the full set of model parameters and p_D is the effective number of parameters in the model. Spiegelhalter et al.[26] showed that p_D can be approximated as $p_D = \bar{D} - D(\hat{\mathbf{\Theta}})$, where $\hat{\mathbf{\Theta}}$ is a suitable 'plug-in' estimate of $\mathbf{\Theta}$, viz. the posterior mean or median. DIC is essentially a single-number summary (lower is better) of the relative fit between the model and the 'true model' generating the data for the purpose of prediction.

Although the DIC provides a measure of relative goodness-of-fit among competing models, it does not provide information on model adequacy. After selecting the best model using DIC, we also perform model assessments through conditional predictive ordinate (CPO) statistics[27] and the associated 'log pseudo-marginal likelihood' (LPML). The CPO is a cross-validation approach and based on the posterior predictive distribution (p.p.d)[43] of the observed data. If $\boldsymbol{\Theta}$ denotes the entire parameter space and $\boldsymbol{y}_{pr}$ denotes the predictive data vector, then the p.p.d is given by:

$$p(\mathrm{y}_{pr}|\mathrm{y}) = \int p(\mathrm{y}_{pr}|\boldsymbol{\Theta}) p(\boldsymbol{\Theta}|\mathrm{y}) d\boldsymbol{\Theta}. \tag{7}$$

One can obtain predictive data easily from a converged posterior sample and samples from the p.p.d are replicates of the observed model generated data. For our observed response $y_i(s)$ from subject i at tooth s with covariate vector $\boldsymbol{X}_{\mathbf{1}\boldsymbol{is}}$, the CPO statistic for observation (i, s) is defined as $\mathrm{CPO}_{is} = f(y_i(s)|\boldsymbol{D}_{(-is)}) = \int f(y_i(s)|\boldsymbol{\Theta}\ \boldsymbol{X}_{\mathbf{1}\boldsymbol{is}}) \pi(\boldsymbol{\Theta}|\boldsymbol{D}_{(-is)}) d\boldsymbol{\Theta}$ where $\pi(\boldsymbol{\Theta}|\boldsymbol{D}_{(-is)})$ is the posterior density of parameter vector $\boldsymbol{\Theta}$ given $\boldsymbol{D}_{(-is)}$, the cross-validated data without the (i, s)th observation. Using a harmonic mean approximation result[27], the CPO_{is} can be easily computed with MCMC samples from the full posterior $\pi(\boldsymbol{\Theta}|\boldsymbol{D})$. Typically, the $\{\mathrm{CPO}_{is}\}$ behave as Bayesian residuals and are plotted against any covariate values x_{1is} (or observed $Y_i(s)$'s) to determine patterns of covariate dependence and to identify possible outliers. Larger values of CPO_{is} indicate better support for the model. A summary measure based on the CPO is the logarithm of the psuedo-marginal likelihood (LPML) defined as LPML = $\Sigma_{(i,s)\in D} \log(\mathrm{CPO}_{is})$, where a higher value of the LPML means better support of the model from the observed data.

5 Data Analysis and findings

We now apply our proposed model to our dental caries data as described in Section 2. Including all subject/tooth-specific covariates (i.e. age, gender, smoking status, brush-floss habits, poverty status and molar indicator), we posit 6 competing models to fit our data and discuss model selection and assessment procedures for these competing models. The models under consideration are:

Model-1 : Simple Binomial regression model with $U_{1i}(s) = U_i \sim$ i.i.d N(0, precision = 0.001)

Model-2 : Simple Binomial regression model with CAR spatial random effects term $U_{1i}(s)$;

Model-3 : BB regression model with $U_{1i}(s) = U_i \sim$ i.i.d N (0, precision = 0.001)

Model-4 : BB regression model with spatial components $U_{1i}(s)$, considering $\gamma(s) = \gamma$ (a constant) ~ Gamma(0.1, 0.01).

Model-5 : BB regression model with spatial components $U_{1i}(s)$, considering $\log(\gamma(s)) = \gamma_0 + U_{2s}$, with the intercept term $\gamma_0 \sim$ N(0, precision = 0.001) and $U_s(s) \sim \mathrm{CAR}(\rho, \sigma^2_{sp})$, as described earlier;

Model-6 : BB regression model with spatial components $U_{1i}(s)$, considering $\gamma(s) \sim$ Gamma(0.1, 0.01).

Table 2 presents the model comparison using the Bayesian model choice criteria described in Section 4.2. With the increase in model complexity using a BB model (instead of a Binomial model) and adding spatial CAR components, we find a substantial improvement in model fit with decreasing DIC and increasing LPML values. The DIC values for Models 3

(BB without CAR, DIC=6560.5) and 4 (BB with CAR, DIC=5198.8) as well as their LPML values differ sufficiently to indicate the superiority of Model 4. Adding the spatial component (Model 4) to the BB structure improves 'leave-one-out' cross validation, as demonstrated by the overall LPML.

Both DIC and LPML favor Model 4 over Models 5 and 6 (which assume different structures for $\gamma(s)$), indicating that a simple 'constant' structure for $\gamma(s)$ fits the data well. To assess model validation in terms of predictive performance, we use the box-plots of log(CPO) statistics to compare between Models 1, 3 and 4 in Figure 3. The median value of log(CPO) for Model 4 is indicated by the horizontal line. The median log(CPO) for Models 3 and 4 are respectively -1.301 and -1.073 and are considerably apart. The cumulative log(CPO), i.e. LPML values are quite different. To summarize, both DIC and LPML favor Model 4 (which assumes $\gamma(s) = \gamma$) among these six models. To determine an 'overall' goodness of fit of Model 4, we also computed the Bayesian p-value[43], which measures the discrepancy between the data and the model by comparing a summary Pearson statistic of the p.p.d to the true distribution of the data. The test quantity for categorical (count) response is a function of data and model parameters and is given as $T(\mathrm{y}, \boldsymbol{\Theta}) = \sum_{(i,s)\in D} \frac{(y_i(s) - \mathrm{E}(y_i(s)))^2}{\mathrm{E}(y_i(s))}$, where the expectation is taken over the p.p.d of the model parameters. Let $\boldsymbol{y}^{sim}$ denote a simulated sample from the p.p.d and $T(\boldsymbol{y}, \boldsymbol{\Theta})$ denote the test quantity used. The Bayesian p-value of the test quantity is defined as p-value = $P(T(\boldsymbol{y}^{sim}, \boldsymbol{\Theta}) \geq T(\boldsymbol{y}, \boldsymbol{\Theta})|\boldsymbol{y})$ and measures whether the variation in the data is consistent with the predicted variation by the model[43]. For Model 4, the mean (for post burn-in MCMC runs) of the Bayesian p-value obtained using 1000 draws of the posterior samples for each MCMC iteration is 0.723, indicating a reasonable overall fit. Henceforth, we discuss our data analysis results related to fixed-effects and spatial parameter estimates based on Model 4. Figure 1 displays the posterior predictive density obtained by fitting both Models 3 and 4, overlayed on the raw data density histogram. Although both the models tend to capture the shape of the density histogram, visual inspection reveals that Model 4 (our spatial BB model) provides a more adequate fit.

Table 3 reports the posterior estimates of the mean, standard deviation and 95% credible intervals (C.I.) of the model parameters for Model 4. Note that the interpretation of spatial association parameters (i.e. ρ and σ^2_{sp}) is highly dependent on the structure of the pre-specified adjacency map as in Table 3. There is a high degree of spatial association (estimate of $\rho = 0.93$) in our data. The posterior mean estimate of the spatial-variance component σ^2_{sp} is 0.113 with 95% C.I. = (0.081, 0.146), which is clearly separated from zero and explains moderate amount of spatial variation. The estimate of γ is 1.37 which estimates the over-dispersion parameter $\frac{\gamma+n_s}{\gamma+1}$ to be 2.25 (for $n_s = 4$) or 2.67 (for $n_s = 5$). Both of the values indicates evidence of some Binomial over-dispersion in the data. The fixed-effects parameters can be interpreted in terms of increase/decrease in odds of the mean probability of having an additional diseased (D/M/F) surface within a tooth, controlling for the spatial random effects. Age (Odds Ratio (OR) = $e^{0.021}$ = 1.02, 95% CI of OR = (1.01, 1.03)) is found to be associated with caries, such that there is a 2% increase in the odds of the (mean) probability of having an additional carious surface with one unit increase in age, and so on. Females and smokers are found to have more diseased teeth, although we admit that our sample has a predominantly high proportion of females, a characteristic common for this population. We did not find any effect of brushing/flossing habits and poverty status on DMFS response for our sample. Molars (OR= 3.26, 95% CI of OR = (2.91, 3.65)) are found to be highly diseased as compared to the non-molar teeth. Specifically in our sample, we have a substantial number of molar teeth missing due to caries, which describes the excess of 5 counts in Figure 1. We conclude that based on the sample of 100 Gullah-speaking

subjects, there is a significant amount of spatial association contributing to dental caries, as determined by the DMFS index. Controlling for spatial association, we also found that age, gender and smoking are associated with caries prevalence.

We also conducted a sensitivity analysis to determine whether there is any significant effect on fixed-effects posteriors with changes in the prior choice for σ_{sp}^{-2}, the spatial precision parameter. We assume $\sigma_{sp}^{-2} \sim \text{Gamma}(k, k)$, where $k \in \{0.01, 0.001, 1\}$. Although some changes are noticed for the posterior of σ_{sp}^{-2} with various choices of k, the posterior estimates of mean and 95% credible interval of the fixed-effects remain similar to those for the weakly-informative $k = 0.1$. Our observation agrees with previous findings in the literature[44], that changes in prior assumptions for the (spatial) scale parameter in Bayesian hierarchical models might lead to substantial changes in posterior estimates of the scale parameter, however the fixed-effects' posteriors remain largely unaffected.

6 Simulation study

In this section we conduct a brief simulation study to illustrate the effect of properly accounting for within- and between-tooth associations on subject-level covariate effect estimation. Each simulated data set consists of $N = 40$ subjects on the same spatial grid defined in Section 3. There are $p = 3$ subject-level covariates, generated as $X_i \sim N(0, I_p)$. The regression coefficients are $\beta_0 = 0$ and $\beta = (\beta_1, \beta_2, \beta_3)' = (0, 1, 2)/3$. For subject i, we simulate data as

$$\begin{aligned} (U_i(1), \ldots, U_i(n))' &\sim \text{CAR}(\rho, \sigma_{sp}^2) \\ \text{logit}(\mu_i(s)) &= X_i'\beta + U_i(s) \\ p_i(s) &\sim \text{Beta}(\tau\mu_i(s), \tau[1 - \mu_i(s)]) \\ Y_i(s) &\sim \text{Binomial}(n_s, p_i(s)) \end{aligned} \tag{8}$$

We generate $S = 100$ data sets from each of three simulation designs:

1. Design 1: 'Basic model' with $\tau = 1.2$, $\rho = 0.99$, and $\sigma_{sp} = 1$
2. Design 2: 'Near-binomial model' with $\tau = 50$, $\rho = 0.99$, and $\sigma_{sp} = 1$
3. Design 3: 'Moderate spatial association model' with $\tau = 1.2$, $\rho = 0.4$, and $\sigma_{sp} = 1$

For Design 1, the data have a fairly strong spatial association as well as overdispersion that is constant across space. The model selection criteria in Section 5's analysis of caries data suggest that these features are present for the caries data. Design 2 maintains the strong spatial association, but has a near-binomial likelihood with $\tau = 50$. The final design is similar to Design 1, except with weak spatial association.

For each simulated data set, we fit four models: the spatial BB model as in Model 4 and termed 'SpaBB', the non-spatial BB model as in Model 3 and termed 'NonSpaBB', the spatial binomial model as in Model 2 and termed 'SpaBin' and non-spatial binomial model as in Model 1 and termed 'NonSpaBin'. For all models we take $\rho \sim \text{Beta}(5,1)$, $\sigma_{sp}^2 \sim \text{InvGamma}(0.1,0.1)$, $\tau \sim \text{Gamma}(0.1,0.01)$, and $\beta_j \sim N(0, \text{Precision} = 0.25)$. Similar to the data analysis, we used 15000 iterations with a burn-in of 10000 and used the remaining 5000 samples to calculate posterior estimates. For each model we compute the posterior mean of β_j, denoted $\widehat{\beta}_j^{(sim)}$ for data set number *sim*, and compute mean square error (MSE):

$MSE_j = \frac{1}{S}\sum_{sim=1}^{S}\left(\beta_j - \widehat{\beta}_j^{(sim)}\right)^2$, where β_j is the true value used to generate the data. In addition, we report the proportion of data sets for which the posterior 95% interval excludes zero (i.e. empirical power), for each method and each regression coefficient.

The results are given in Table 4. For all simulation designs, the non-spatial models have considerably larger *MSE* than the spatial models, illustrating the importance of accounting for spatial structure when estimating subject-level covariates. The non-spatial models also have inflated empirical Type-I error (power for β_1), and smaller power for β_3. The spatial BB model has the smallest MSE for all covariates for the first design with strong spatial correlations and BB likelihood. The spatial BB model also gives similar performance to the binomial model for Design 2's near-binomial data, however the spatial binomial model outperforms the spatial BB model for Design 3 with moderate spatial association. In this case, the spatial random effects in the spatial binomial model may be sufficient to account for the extra-binomial variation for the BB likelihood.

7 Conclusion

In this paper, we develop a Beta-Binomial model for multivariate spatial count data on dental caries that accommodates extra-Binomial variation (heterogeneity) as well as possible spatial clustering. We show, using both simulation studies and real data, that accounting for spatial clustering and over-dispersion in the same model provides a substantial improvement for estimating covariate effects as compared to other standard count data models that do not make this accommodations.

In this study, the total random variation could also be decomposed into both spatially correlated as well as uncorrelated heterogeneities, often referred to as the 'convolution prior'[45], i.e. $\mu_i(s) = F(\beta_0 + \boldsymbol{X_{1is}}^T\boldsymbol{\beta} + U_1(s) + E_i)$. However, using this model for $\mu_i(s)$ leads to convergence issues, specially for the fixed-effects parameters even when a high burn-in period (like 50000) is used. However while fitting separate i.i.d CAR models to each subject (Model 4 in this paper), we find no such problems for the fixed-effects. This issue is currently under investigation. To check whether we need to fit separate i.i.d CAR to each subject, we modeled $\mu_i(s)$ as $\mu_i(s) = F(\beta_0 + \boldsymbol{X_{1is}}^T\boldsymbol{\beta} + U_1(s))$. Because the DIC for this model was 5228.72 (far away from Model 4 but closer to Model 6), we did not proceed with this model.

Our inference is strictly based on (a) the DM_5FS convention[29] for measuring DMFS, (b) a specified spatial model, i.e. conditionally-autoregressive, and (c) a pre-defined adjacency structure. A number of areas can be identified for future work. Whether our inference is robust to various definitions of DMFS, other spatial models (viz. geostatistical, etc) as well as some other adjacency structure needs investigation. Also, because subjects with poor caries status are likely to have fewer teeth than subjects with a 'good' caries status and a tooth missing and located at the back of the mouth might be surrounded with teeth that are also missing, one can also consider 'jointly' modeling the location and presence/absence of a tooth together with its caries status in the spirit of Reich and Bandyopadhyay[25] for discrete count responses using shared (latent) spatial random effects. Additionally, one can also consider modeling the DMFS counts as ordinal data and develop spatial random effects models. Our study population is homogenous (Gullah-speaking African Americans diagnosed with type-2 diabetes) which primarily motivated our prior choice on γ. Generalization of our proposed methods to analyzing dental caries data collected for other groups/races or population would be relatively straightforward, however one needs to re-consider proper parameterizations of the BB likelihood as well as the spatial adjacency structure to accommodate spatial heterogeneity and (possible) over-dispersion.

Acknowledgments

This research was supported by NIH/NCRR Grant P20 RR017696-06. The authors thank the Center for Oral Health Research (COHR) at the Medical University of South Carolina for providing the data and context for this work.

References

1. US Department of Health and Human Services. National Institutes of Health. 2000. Oral Health in America: A report of the Surgeon General, National Institutes of Dental and Craniofacial Research; p. 1-308.
2. Fejerskov, O.; Kidd, EAM., editors. Dental caries: the disease and its clinical management. Copenhagen: Blackwell; 2003.
3. Kidd EA, Giedrys-Leeper E, Simons D. Take two dentists: a tale of root caries. Dental Update. 2000; 27:222–230. [PubMed: 11218479]
4. World Health Organization. The world oral health report 2003. Geneva: WHO; 2003.
5. Curzon ME, Preston AJ. Risk groups: nursing bottle caries/caries in the elderly. Caries Research. 2004; 38(suppl 1):24–33. [PubMed: 14685021]
6. Petersen PE, Yamamoto T. Improving the oral health care of older people: the approach of the WHO global oral health programme. Community Dentistry and Oral Epidemiology. 2005; 33:81–92. [PubMed: 15725170]
7. Vieira AR, Marazita ML, Goldstein-McHenry T. Genome-wide scan finds suggestive caries loci. Journal of Dental Research. 2008; 87(5):435–439. [PubMed: 18434572]
8. Keyes PH. The infectious and transmissible nature of experimental dental caries: Findings and Implications. Archives of Oral Biology. 1960; 13:304–320. [PubMed: 14408737]
9. Keyes PH. Recent advances in dental caries research. International dentistry Journal. 1962; 12:443–463.
10. Selwitz RH, Ismail AI, Pitts NB. Dental caries. Lancet. 2007; 369:51–59. [PubMed: 17208642]
11. Fernandes, J.; Slate, EH.; Wiegand, RE.; London, SD.; Grewal, JS.; Werner, P.; Sanders, JJ.; Lopes-Virella, M.; Salinas, CF. Dental caries in Type 2 Gullah diabetics; Journal of Dental Research. 2007. p. 1054www.dentalresearch.org
12. Burnside G, Pine CM, Williamson PR. The application of multilevel modelling to dental caries data. Statistics in Medicine. 2007; 26:4139–4149. [PubMed: 17340596]
13. Darby, ML.; Walsh, MM. Dental Hygiene: Theory and Practice. 2nd. W. B. Saunders Company; USA: 2003.
14. Williams DA. The Analysis of Binary Responses From Toxicological Experiments Involving Reproduction and Teratogenicity. Biometrics. 1975; 31:949–952. [PubMed: 1203435]
15. Williams DA. Extra-Binomial Variation in Logistic Linear Models. Applied Statistics. 1982; 31:144–148.
16. Lawson, AB. Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology. Boca Raton, FL: CRC Press; 2009.
17. Griffiths DA. Maximum likelihood estimation of the Betabinomial distribution and an application to the household incidence of the total number of cases of a disease. Biometrics. 1973; 29:637–648. [PubMed: 4785230]
18. Prentice RL. Binary regression Using an extended Betabinomial distribution with discussion of correlation induced by covariate measurement errors. Journal of the American Statistical Association. 1986; 81:321–327.
19. Ainsworth, L. Ph D dissertation. Burnaby (BC, Canada): Simon Fraser University; 2007. Models and methods for spatial data: detecting outliers and handling zero-inflated counts.
20. Besag J. Spatial interaction and the statistical analysis of lattice systems (with discussion). Journal of the Royal Statistical Society. 1974; 36:192–236.B
21. Mardia KV. Multidimensional multivariate Gaussian Markov random fields with applications to image processing. Journal of Multivariate Analysis. 1998; 24:45–64.
22. Lawson, AB.; Biggeri, A.; Böhning, D.; Lessafre, E.; Viel, J.; Bertollini, R., editors. Disease Mapping and Risk Assessment for Public Health. New York: Wiley; 1999.

23. Banerjee, S.; Carlin, BP.; Gelfand, AE. Hierarchical Modeling and Analysis for Spatial Data. Boca Raton, FL: Chapman & Hall/CRC; 2004.
24. Reich, BJ. Ph D dissertation. Minneapolis, MN: University of Minnesota; 2005. Neighbor Relation Modelling and Variance Component Identification in Hierarchical Areal Data Models with Application to Periodontology, Cancer Epidemiology, and Sports.
25. Reich BJ, Bandyopadhyay D. A latent factor model for spatial data with informative missingness. Annals of Applied Statistics. 2010 In Press.
26. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit (with discussion). Journal of the Royal Statistical Society. 2002; 64:583–639.B
27. Gelfand AE, Dey DK. Bayesian model choice: asymptotics and exact calculations. Journal of the Royal Statistical Society. 1994; 56:501–514.B
28. Darby, ML.; Walsh, MM. Dental Hygiene: Theory and Practice. 1st. W. B. Saunders Company; USA: 1995.
29. Broadbent JM, Thompson WM. For debate: problems with the DMF index pertinent to dental caries data analysis. Community dentistry and Oral Epidemiology. 2005; 33(6):400–409. [PubMed: 16262607]
30. Johnson-Spruill I, Hammond P, Davis B, McGee Z, Louden D. Health of Gullah families in South Carolina with Type-2 diabetes. The Diabeted Educator. 2009; 35:117–123.
31. Abramowitz, M.; Stegun, IA. Handbook of Mathematical Functions. New York: Dover; 1965.
32. Skellam JG. A probability distribution from the binomial distribution by regarding the probability of success as variable between the sets of trials. Journal of the Royal Statistical Society. 1948; 10:257–265.B
33. MacEachern, S.; Kottas, A.; Gelfand, A. Technical Report 01-10. Institute of Statistics and Decision Sciences, Duke University; 2001. Spatial nonparametric Bayesian models.
34. Jiang, J. Linear and Generalized Linear Mixed Models and Their Applications. New York: Springer-Verlag; 2007.
35. Besag J, York JC, Mollié A. Bayesian image restoration, with two applications in spatial statistics (with discussions). Annals of the Institute of Statistical Mathematics. 1991; 43:1–59.
36. Cressie, NAC. Statistics for spatial data. Revised. New York: Wiley; 1993.
37. Carlin, BP.; Banerjee, S. Hierarchical multivariate CAR models for spatio-temporally correlated survival data (with discussion). In: Bernardo, JM.; Bayarri, MJ.; Berger, JO.; Dawid, AP.; Heckerman, D.; Smith, AFM.; West, M., editors. Bayesian Statistics 7. Oxford: Oxford University Press; 2003. p. 45-63.
38. McCullagh, P.; Nelder, J. Generalized Linear Models. Second. New York: Chapman and Hall/CRC; 1989.
39. Spiegelhalter, D.; Thomas, A.; Best, N.; Lunn, D. MRC Biostatistics Unit, Institute of Public Health and Department of Epidemiology and Public Health, Imperial College School of Medicine; 2005. WinBUGS User Manual, Version 1.4.2. available at http://www.mrc-bsu.cam.ac.uk/bugs
40. Dunson DB, Chen Z, Harry J. A Bayesian approach for joint modeling of cluster size and subunit-specific outcomes. Biometrics. 2003; 59:521–530. [PubMed: 14601753]
41. Best, NG.; Arnold, RA.; Thomas, A.; Waller, LA.; Conlon, EM. Bayesian models for spatially correlated disease and exposure data (with discussion). In: Bernardo, JM.; Berger, JO.; Dawid, AP.; Smith, AFM., editors. Bayesian Statistics 6. Oxford: Oxford University Press; 2003. p. 131-156.
42. Gelfand, AE.; Sahu, SK.; Carlin, BP. Efficient parameterizations for generalised linear models (with discussion). In: Bernardo, JM.; Berger, JO.; Dawid, AP.; Smith, AFM., editors. Bayesian Statistics 5. Oxford: Oxford University Press; 2003. p. 165-180.
43. Gelman, A.; Carlin, JB.; Stern, H.; Rubin, D. Bayesian Data Analysis. 2nd. Florida: Chapman & Hall/CRC; 2004.
44. Daniels MJ, Kass RE. Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. Journal of the American Statistical Association. 1999; 94:1254–1263.

45. Zhou H, Lawson AB, Hebert JR, Slate EH, Hill EG. A Bayesian hierarchical modeling approach for studying the factors affecting the stage at diagnosis of prostate cancer. Statistics in Medicine. 2008; 27:3612–3628. [PubMed: 18416442]

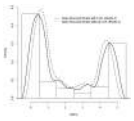

Figure 1.
Density histogram of raw DMFS counts (packed over teeth and subject) overlayed with posterior predictive densities generated using Beta-Binomial models (with and without spatial components)

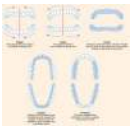

Figure 2.
Classification of tooth surfaces for permanent dentition within a human mouth. This figure was published in '*Dental Hygiene: Theory and Practice*', 3rd edition, Michele L. Darby and Margaret M. Walsh, Chapter 14 Page 237, Copyright Saunders (Elsevier) (2010). Adapted from 'The Art of Dental Scaling' by D. Wotton, 1991, University of Vermont, Burlington

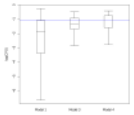

Figure 3.
Box-plots of log(CPO) for Models 1,3 and 4. Larger values of log(CPO) indicate more support for the model. The horizontal line is the median log(CPO) value for Model 4 (the best model).

Table 1

Adjacency map for the dental caries data. Neighbors are denoted by tooth numbers

Tooth number	Neighbors	Tooth number	Neighbors
1	2, 3, 32	17	16, 18, 19
2	1, 3, 4, 31	18	17, 19, 20, 15
3	1, 2, 4, 5, 30	19	17, 18, 20, 21, 14
4	2, 3, 5, 6, 29	20	18, 19, 21, 22, 13
5	3, 4, 6, 7, 28	21	19, 20, 22, 23, 12
6	4, 5, 7, 8	22	20, 21, 23, 24
7	5, 6, 8, 9	23	21, 22, 24, 25
8	6, 7, 9, 10	24	22, 23, 25, 26
9	7, 8, 10, 11	25	23, 24, 26, 27
10	8, 9, 11, 12	26	24, 25, 27, 28
11	9, 10, 12, 13	27	25, 26, 28, 29
12	10, 11, 13, 14, 21	28	26, 27, 29, 30, 5
13	11, 12, 14, 15, 20	29	27, 28, 30, 31, 4
14	12, 13, 15, 16, 19	30	28, 29, 31, 32, 3
15	13, 14, 16, 18	31	29, 30, 32, 2
16	14, 15, 17	32	30, 31, 1

Table 2

Model comparison using DIC and LPML

Model	$\bar{D}$	p_D	DIC	LPML
Model 1	14945.0	7.4	14952.4	-7485.9
Model 2	13764.6	299.3	14063.9	-6270.8
Model 3	5875.6	684.9	6560.5	-4319.2
Model 4	4528.0	670.8	5198.8	-3686.1
Model 5	4593.7	704.6	5298.3	-3758.1
Model 6	4547.9	676.9	5224.8	-3739.3

Table 3

(Conditional) Posterior estimates of fixed-effects and other model parameters with 95% credible intervals (C.I.) for Model 4

Parameter	Mean	Std. Dev.	95% C.I.
Intercept	-2.099	0.155	(-2.425, -1.796)
Age	0.021	0.005	(0.011, 0.031)
Gender	0.313	0.093	(0.11, 0.476)
Smoker	0.434	0.101	(0.201, 0.583)
Brush-Floss	0.109	0.088	(-0.061, 0.275)
Poverty	0.009	0.097	(-0.176, 0.187)
Molar	1.182	0.057	(1.067, 1.294)
ρ	0.935	0.021	(0.887, 0.961)
γ	1.371	0.044	(1.286, 1.461)
σ^2_{sp}	0.113	0.017	(0.081, 0.146)

Table 4

MSE (SE) and empirical power for the simulation study.

		MSE (SE)			Empirical Power		
Design	**Model**	β_1	β_2	β_3	β_1	β_2	β_3
1	SpaBB	0.010 (0.002)	0.021 (0.003)	0.026 (0.004)	0.38	0.89	1.00
	SpaBin	0.019 (0.003)	0.026 (0.003)	0.063 (0.003)	0.63	0.96	1.00
	NonSpaBB	1.974 (0.300)	2.228 (0.277)	2.290 (0.321)	0.87	0.87	0.88
	NonSpaBin	2.983 (0.363)	4.700 (0.691)	3.323 (0.580)	0.95	0.94	0.95
2	SpaBB	0.014 (0.002)	0.019 (0.003)	0.035 (0.004)	0.56	0.92	1.00
	SpaBin	0.013 (0.002)	0.018 (0.002)	0.033 (0.004)	0.67	0.97	1.00
	NonSpaBB	2.318 (0.353)	1.900 (0.357)	2.561 (0.387)	0.83	0.81	0.85
	NonSpaBin	3.435 (0.411)	3.909 (0.693)	4.018 (0.516)	0.98	0.93	0.93
3	SpaBB	0.002 (0.000)	0.005 (0.001)	0.021 (0.002)	0.14	1.00	1.00
	SpaBin	0.002 (0.001)	0.002 (0.000)	0.003 (0.000)	0.24	1.00	1.00
	NonSpaBB	1.717 (0.257)	1.744 (0.286)	1.606 (0.261)	0.94	0.86	0.88
	NonSpaBin	3.165 (0.522)	2.022 (0.346)	2.873 (0.717)	0.94	0.94	0.98