The Canadian Journal of Statistics
Vol. xx, No. yy, 2014, Pages 1–14
La revue canadienne de statistique

Non-parametric generalized linear mixed models in small area estimation

Mahmoud Torabi[1]* and Farhad Shokoohi[2]

[1]*Department of Community Health Sciences, University of Manitoba, Winnipeg, MB, Canada*
[2]*Department of Statistics, The Ohio State University, Columbus, OH, U.S.A.*

Key words and phrases: Bayesian computation; Exponential family; Penalized spline; Prediction interval; Random effect; Small area estimation.

MSC 2010: Primary 62D99; secondary 62G08

Abstract: Mixed models are commonly used for the analysis of small area estimation. In particular, small area estimation has been extensively studied under linear mixed models. Recently, small area estimation under the linear mixed model with penalized spline (P-spline) regression model, for fixed part of the model, has been proposed. However, in practice there are many situations that we have counts or proportions in small areas; for example a dataset on the number of asthma physician visits in small areas in Manitoba. In particular, the covariates age, genetic, environmental factors, among other covariates seem to predict asthma physician visits, however, these relationships may not be linear (see Section 5). In this paper, small area estimation under generalized linear mixed models using P-spline regression models is proposed to cover Normal and non-Normal responses. In particular, the empirical best predictor of small area parameters with corresponding prediction intervals are studied. The performance of the proposed approach is evaluated through simulation studies and also by a real dataset. *The Canadian Journal of Statistics* xx: 1–14; 2014

Résumé: Insérer votre résumé ici. **Abstract in French**. *La revue canadienne de statistique* xx: 1–14; 2014

1. INTRODUCTION

Small area estimation has received considerable attention due to growing demand for reliable small area statistics. Rao (2003), Jiang and Lahiri (2006) and Jiang (2010) have given comprehensive accounts of model-based small area estimation. In particular, area level model (Fay and Herriot, 1979) and nested error linear regression model (Battese, Harter, and Fuller, 1988) are often used in small area estimation to obtain efficient model-based estimators of small area means.

Most of the research in small area estimation has focused on parametric models, and the research based on non-parametric models in the context of small area estimation is limited. Opsomer et al. (2008) extended the linear mixed model approach in the context of small area estimation to the case in which a linear relationship may not be assumed using penalized splines (P-splines) regression. From a very different perspective, Chambers and Tzavidis (2006) studied an approach for small area estimation that is based on M-quantile regression which allows for models robust to outliers and to distributional assumptions on the errors and the area effects. However, when the functional form of the relationship between the q-th M-quantile and the covariates is not linear, it can lead to biased estimates of the small area parameters. An extended version of this approach for the estimation of the small area distribution function us-

* *Author to whom correspondence may be addressed.*
E-mail: torabi@cc.umanitoba.ca

ing a non-parametric specification of the conditional M-quantile of the response variable given the covariates has been also studied (Pratesi, Ranalli, and Salvati, 2008, 2009; Salvati, Ranalli, and Pratesi, 2011). Jiang, Nguyen, and Rao (2010) developed an adaptive fence procedure for the non-parametric model selection using P-spline models. Sperlich and José Lombardía (2010) used local polynomial inference in the context of small area estimation.

However, there are many applications in small area estimation where responses are counts or proportions. In particular, one may be interested to analyze the number of incidences in small areas; for example a dataset on the number of asthma physician visits in small areas in Manitoba (see Section 5) where the covariates age, genetic, environmental factors, among other covariates may not be linear to predict asthma physician visit rates. The aim of this paper is to develop a unified analysis of both discrete and continuous responses using P-spline regression models. These types of models fall in the class of generalized linear mixed models (GLMMs). It is well known that the frequentist analysis of these models is computationally difficult.

There are some approximate methods, based on the frequentist paradigm, for analyzing GLMMs such as Penalized quasi-likelihood (PQL), Laplace approximation, Gauss-Hermite quadrature among other approaches. Recently, Lele, Dennis, and Lutscher (2007) introduced an approach, called data cloning (DC), to compute the maximum likelihood (ML) estimates and their corresponding standard errors for general hierarchical models. Data cloning is a computing algorithm based on Markov Chain Monte Carlo (MCMC) methods. Lele, Nadeem, and Schmuland (2010) described an approach to compute prediction and prediction intervals for the random effects in the class of GLMMs. Torabi and Shokoohi (2012) used DC approach in the context of small area estimation to study cross-sectional and time-series models, and Baghishani and Mohammadzadeh (2011) used the DC method in the context of spatial GLMMs.

We use DC to analyze our proposed non-parametric mixed models for Normal and non-Normal responses in the context of small area estimation. This paper is organized as follows. Non-parametric mixed models are described in Section 2. In Section 3, we describe how DC can be used to estimate model parameters and also to obtain prediction and prediction intervals of small area parameters. The performance of proposed approach is reported through several simulation studies in Section 4 with also comparing DC and its competitor PQL, noting that PQL is the only approximate approach which has the package in R and can handle the complicated structure of our non-parametric GLMMs. In Section 5, a real application of our proposed approach is also explored. Some concluding remarks are given in Section 6.

2. NON-PARAMETRIC MIXED MODELS

A unit level regression model can be described as follows. Let y_{ij} be the variable of interest for the j-th observation in a given area $i (j = 1, ..., n_i; i = 1, ..., m)$. The y_{ij} are assumed to be conditionally independent, given random effects, with exponential family p.d.f.

$$f(y_{ij}|\theta_{ij}, \phi_{ij}) = \exp[\{y_{ij}\theta_{ij} - a(\theta_{ij})\}/\phi_{ij} + b(y_{ij}, \phi_{ij})], \tag{1}$$

$(j = 1, ..., n_i; i = 1, ..., m)$. The density (1) is parameterized with respect to the canonical parameters θ_{ij}, known scale parameters ϕ_{ij} and functions $a(\cdot)$ and $b(\cdot)$. The exponential family (1) covers well-known distributions including Normal, binomial and Poisson distributions. The natural parameters θ_{ij} for non-parametric P-spline regression models are then modeled as

$$h(\mu_{ij}) = \theta_{ij} = m_0(x_{ij}) + \nu_i + u_{ij},\ (j = 1, ..., n_i; i = 1, ..., m), \tag{2}$$

where $\mu_{ij} = E(y_{ij}|\theta_{ij}, \phi_{ij}), h$ is a strictly increasing function, $\nu_i \overset{i.i.d.}{\sim} N(0, \sigma_v^2)$ and $u_{ij} \overset{i.i.d.}{\sim} N(0, \sigma_u^2)$ are area specific and unit-level random effects, respectively, and $m_0(x_{ij})$ is unknown

but, if this function is to be estimated by using P-splines, can be approximated sufficiently well by

$$m_0(x_{ij};\boldsymbol{\beta},\boldsymbol{\gamma}) = \beta_0 + \beta_1 x_{ij} + \cdots + \beta_p x_{ij}^p + \sum_{l=1}^{L} \gamma_l (x_{ij} - \kappa_l)_+^p, \tag{3}$$

where p is the degree of the spline, $(x)_+^p$ denotes the function $x^p I_{\{x>0\}}$, with $I(\cdot)$ as an indicator function, x_{ij} is a known value, $\{\kappa_1, \ldots, \kappa_L\}$ is a set of fixed knots, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_L)'$ are the regression coefficients of parameters and P-spline parts of the model, respectively, L is the number of spline knots, and $\gamma_l \overset{i.i.d.}{\sim} N(0, \sigma_\gamma^2)$, $(l = 1, \ldots, L)$. It is well known that with spreading out the location of knots sufficiently over the range of x_{ij} and with large enough L, the class of P-spline approximation is very large and can approximate most smooth functions (Eilers and Marx, 1996; Boor, 2001). It is recommended to use the number of spline knots (L) as the minimum of 40 and the number of unique x_{ij}'s divided by 4 (Ruppert, 2002); we also use this criterion in our paper. We refer to Ruppert, Wand, and Carroll (2003) for more details of P-spline models.

As a special case, under Normal distribution, $h(\mu_{ij}) = \mu_{ij} = \theta_{ij}$, the unit level non-parametric P-spline linear mixed model is obtained where u_{ij}'s are sampling errors which are normally distributed with zero mean and variance σ_u^2. The random variables $(\gamma_l, \nu_i, u_{ij})$ are also assumed to be independent of each other. Opsomer et al. (2008) used restricted maximum likelihood approach to estimate the model parameters. They calculated empirical best linear unbiased predictor (EBLUP) of small area mean θ_i using

$$\hat{\theta}_i = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_i + \cdots + \hat{\beta}_p \bar{X}_i^p + \sum_{l=1}^{L} \hat{\gamma}_l E(x_{ij} - \kappa_l)_+^p + \hat{\nu}_i,$$

with $\bar{X}_i$ as the mean of population units x_{ij} in area i, and $E(\cdot)$ stands for the expectation. They also provided mean squared prediction error (MSPE) of $\hat{\theta}_i$ and the corresponding MSPE estimation.

3. FREQUENTIST INFERENCE

Let $\boldsymbol{y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_m)'$ be the observed data vector and, conditionally on the random effects, assume that the elements of $\boldsymbol{y}$ are independent and drawn from a distribution in the exponential family with parameters $\boldsymbol{\beta}$ where $\boldsymbol{y}_i = (y_{i1}, ..., y_{in_i})', (i = 1, ..., m)$. It is also assumed that distribution for random effects depends on parameters $(\sigma_\nu^2, \sigma_\gamma^2, \sigma_u^2)$. The goal of the analysis is to estimate the model parameters $\boldsymbol{\alpha} = (\boldsymbol{\beta}, \sigma_\nu^2, \sigma_\gamma^2, \sigma_u^2)'$ and predict the small area parameters θ_{ij} or its variant.

As the DC approach uses Bayesian tools to make an inference, we start with standard Bayesian approach for our hierarchical model. Denote $L(\boldsymbol{\alpha}; \boldsymbol{y})$ as likelihood of $\boldsymbol{\alpha}$ given $\boldsymbol{y}$ and $\pi(\boldsymbol{\alpha})$ as prior distribution on the parameter space. The posterior distribution $\pi(\boldsymbol{\alpha}|\boldsymbol{y})$ is given by

$$\pi(\boldsymbol{\alpha}|\boldsymbol{y}) = \frac{L(\boldsymbol{\alpha}; \boldsymbol{y})\pi(\boldsymbol{\alpha})}{C(\boldsymbol{y})}, \tag{4}$$

where $C(\boldsymbol{y}) = \int L(\boldsymbol{\alpha}; \boldsymbol{y})\pi(\boldsymbol{\alpha}) d\boldsymbol{\alpha}$ is the normalizing constant. There are computational tools, MCMC algorithms, that facilitate generation of random variates from the posterior distribution $\pi(\boldsymbol{\alpha}|\boldsymbol{y})$ without computing the integrals in the numerator or the denominator of (4) (Gilks, Richardson, and Spiegelhalter, 1996; Spiegelhalter, Abrams, and Myles, 2004).

The DC method uses the Bayesian computational approach for frequentist purposes. In DC, the observations $\boldsymbol{y}$ are repeated independently by K different individuals and all these individuals happened to have the same set of observations $\boldsymbol{y}$ called $\boldsymbol{y}^{(K)} = (\boldsymbol{y}, \boldsymbol{y}, ..., \boldsymbol{y})$. The posterior distribution of $\boldsymbol{\alpha}$ conditional on the data $\boldsymbol{y}^{(K)}$ is then given by

$$\pi_K(\boldsymbol{\alpha}|\boldsymbol{y}^{(K)}) = \frac{\{L(\boldsymbol{\alpha};\boldsymbol{y})\}^K \pi(\boldsymbol{\alpha})}{C(\boldsymbol{y}^{(K)})}, \tag{5}$$

where $C(\boldsymbol{y}^{(K)}) = \int \{L(\boldsymbol{\alpha};\boldsymbol{y})\}^K \pi(\boldsymbol{\alpha}) d\boldsymbol{\alpha}$ is the normalizing constant. The expression $\{L(\boldsymbol{\alpha};\boldsymbol{y})\}^K$ is the likelihood for K copies of the original data. The following Theorem states that how one can use the likelihood of K copies of the original data to make an inference based on the MLE.

Theorem 1. *Consider the general models (1)-(3). Under some regularity conditions, as K becomes large, the distribution in (5) converges to a multivariate Normal distribution with mean equal to the MLE of the model parameters and variance-covariance matrix equal to $1/K$ times the inverse of the Fisher information matrix for the MLE.*

Proof. The proof follows along the lines of (Walker, 1969), Lele, Dennis, and Lutscher (2007) and Lele, Nadeem, and Schmuland (2010). ■

Hence, the sample mean vector of the generated random numbers from (5) provides the MLE of the model parameters and K times their sample variance-covariance matrix is an estimate of the asymptotic variance-covariance matrix for the MLE $\hat{\boldsymbol{\alpha}}$. Lele, Nadeem, and Schmuland (2010) also provided various checks to determine the adequate number of clones (K). For instance, one may plot the ratio of largest eigenvalue of the posterior variance of K clones to the eigenvalue of the posterior variance of one clone as a function of the number of clones K to determine if the posterior distribution has become nearly degenerate. As another criterion, it is approximately true that as we increase the number of clones,

$$(\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}})' \boldsymbol{V}^{-1} (\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}}) \sim \chi^2_q, \tag{6}$$

where $\bar{\boldsymbol{\alpha}}$ and $\boldsymbol{V}$ are the mean and the variance of the posterior distribution of $\boldsymbol{\alpha}$, respectively, and q is the dimension of $\boldsymbol{\alpha}$. One may also compute the following two statistics: a) $\zeta = \frac{1}{B}\sum_{b=1}^{B}(O_b - E_b)^2$, where O_b and E_b are observed and estimated quantiles for χ^2_q random variable, and b) $\tilde{r}^2 = 1 - \rho^2$, where ρ is the correlation between (O, E). If these statistics are close to zero, it indicates that the approximation (6) is reasonable. Note that the above three criteria have been implemented in the *dclone* package (Sólymos, 2010), a freely available package created for R (R Development Core Team, 2012), which will be used in our simulation studies as well as in our application.

3.1. Prediction of small area parameters

Prediction of small area parameters (random effects), particularly from the frequentist viewpoint, is problematic. If the parameters $\boldsymbol{\alpha}$ are known, then one can clearly use the conditional distribution of $\boldsymbol{\Theta} = (\theta_{11}, ..., \theta_{mn_m})$, the latent variables, given the observed data. That is, to predict $\boldsymbol{\Theta} = \boldsymbol{\theta}$, one can use $\pi(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{\alpha}^*)$ where $\boldsymbol{\alpha}^*$ is the true value of the parameter. A naive approach, when $\boldsymbol{\alpha}$ is estimated using the data, is to use $\pi(\boldsymbol{\theta}|\boldsymbol{y}, \hat{\boldsymbol{\alpha}})$. However, this approach does not take into account the variability introduced by the model parameters estimate. An approach that has been suggested in the literature (e.g., Hamilton, 1986; Lele, Nadeem, and Schmuland, 2010) to

take into account the variation of the estimators is to use the density:

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\int f(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\beta}) g(\boldsymbol{\theta}|\sigma_\nu^2,\sigma_\gamma^2,\sigma_u^2)\phi(\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}},I^{-1}(\hat{\boldsymbol{\alpha}}))d\boldsymbol{\alpha}}{C(\boldsymbol{y})}, \tag{7}$$

where $g(\cdot)$ is a multivariate Normal distribution in our set-up and $\phi(.,\mu,\Sigma)$ denotes a multivariate Normal density with mean μ and variance-covariance Σ, which are equal to the MLE and the inverse of the Fisher information matrix here. In this paper, we obtain prediction and prediction intervals for random effects $\boldsymbol{\psi}$ using the density in equation (7) along with MCMC sampling, noting that one can use the same approach to predict variants of $\theta_{ij}, (j = 1, ..., n_i; i = 1, ..., m)$.

In this paper, for the DC analysis, the Normal prior distributions were used for fixed parameters with mean 0 and variance 10^6, and uniform distribution between 0 and 1000 for the standard deviations. Since the DC is invariant to the priors, one may use different priors. To monitor the convergence of the model parameters, we used several diagnostic methods implemented in the Bayesian output analysis (*BOA*) program (Smith, 2007) in R. We also used diagnostic methods implemented in the *dclone* package (Sólymos, 2010), which described in Section 3, to monitor the convergence of the model parameters in terms of number of clones K.

4. SIMULATION STUDY

4.1. Non-parametric linear mixed model

We conduct a simulation study to evaluate the performance of proposed approach in the non-parametric linear mixed model set-up. We use the following true non-parametric area-level model to generate samples for the simulation study

$$y_i = m_0(x_i) + \nu_i + u_i, (i = 1, ..., m),$$

where $m = 100$ is number of areas, $\nu_i \overset{i.i.d.}{\sim} N(0, \sigma_\nu^2)$ and $u_i \overset{i.i.d.}{\sim} N(0, \sigma_u^2)$ with known $\sigma_u^2 = 1$, and three different choices of $m_0(x)$: 1. $m_0(x) = 1 + x$ (linear), 2. $m_0(x) = 1 + x + x^2$ (quadratic), and 3. $m_0(x) = 1 - x + 0.5\ \exp(x)$ (exponential 1) as used by Breidt, Claeskens, and Opsomer (2005) as well as Rao, Sinha, and Dumitrescu (2014). We also generate the $x_i (i = 1, ..., m)$ from Normal distribution with mean one and variance one once and treat them fixed in the simulation study. For the simulation study, we use the linear P-spline approximation ($p = 1$) for $m_0(x_i)$. Following Ruppert (2002), the number of knots set to be $L = 25$. We generate $R = 1000$ independent samples $\{(y_i^{(r)}, x_i), i = 1, ..., 100; r = 1, ..., 1000\}$ where $y_i^{(r)} = m_0(x_i) + \nu_i^{(r)} + u_i^{(r)}$, $\nu_i^{(r)}$ and $u_i^{(r)}$ are generated from the corresponding Normal distributions of ν_i and u_i with $\sigma_\nu^2 = \sigma_u^2 = 1$. For each simulated run, we apply the method of DC to get the MLE of the model parameters and also to provide the prediction and prediction intervals of the EPLUP of small area means $\theta_i^{(r)} = m_0(x_i) + \nu_i^{(r)}, (r = 1, ..., R)$, using $\hat{\theta}_i^{(r)} = \hat{\beta}_0^{(r)} + \hat{\beta}_1^{(r)} x_i + \sum_{l=1}^{25} \hat{\gamma}_l^{(r)} (x_i - \kappa_l)_+ + \hat{\nu}_i^{(r)}$ with $\gamma_l \overset{i.i.d.}{\sim} N(0, \sigma_\gamma^2)$. We also compare our proposed P-spline regression model with the corresponding parametric model $\hat{\theta}_{i,p}^{(r)} = \hat{\beta}_0^{(r)} + \hat{\beta}_1^{(r)} x_i + \hat{\nu}_i^{(r)}$. For this simulation set-up, the average number of clones was $K = 10$ to obtain MLE, and the average number of iterations for convergence of the model parameters was about 50,000 for our P-spline regression model. We calculate the empirical MSPE of $\hat{\theta}_i$ as

$$\text{EMSPE}(\hat{\theta}_i) = \frac{1}{R}\sum_{r=1}^{R} \{\hat{\theta}_i^{(r)} - \theta_i^{(r)}\}^2,$$

TABLE 1: Average EMSPE of small area means by data cloning approach, P-spline and parametric approaches for three different models (linear, quadratic, and exponential 1) in the case of linear mixed model.

True model	Method	
	P-spline	Parametric
Linear	0.526	0.525
Quadratic	0.541	0.785
Exponential 1	0.552	0.864

TABLE 2: Percent average absolute relative bias of estimators of MSPE of small area means by data cloning approach, P-spline and parametric approaches for three different models (linear, quadratic, and exponential 1) in the case of linear mixed model.

True model	Method	
	P-spline	Parametric
Linear	7.32	7.26
Quadratic	6.20	13.16
Exponential 1	6.89	12.73

and the relative bias of an estimator of the MSPE, say $mspe$, as

$$\text{RB}[mspe(\hat{\theta}_i)] = \Big\{\frac{1}{R}\sum_{r=1}^{R} mspe^{(r)}(\hat{\theta}_i) - \text{EMSPE}(\hat{\theta}_i)\Big\}/\text{EMSPE}(\hat{\theta}_i),$$

where $\hat{\theta}_i^{(r)}, \theta_i^{(r)}$, and $mspe^{(r)}(\hat{\theta}_i)$ are the values of $\hat{\theta}_i, \theta_i$, and $mspe(\hat{\theta}_i)$ for the *r*-th simulation batch, respectively. Note that $mspe(\hat{\theta}_i)$ is the variance of $\hat{\theta}_i$ calculated by (7).

The results of average EMSPE over areas for the three different models (linear, quadratic, exponential 1) of P-spline and parametric approaches are reported in Table 1. As shown in Table 1, the values of EMSPE are stable for the P-spline approach for the three different models while these values are increased for the parametric approach from the linear to the quadratic and also from the quadratic to the exponential 1 model. The results of average absolute relative bias (AARB) of $mspe$ over areas for the three different models and the two approaches (P-spline and parametric) are also reported in Table 2. The proposed P-spline approach performs very well in terms of AARB($< \%8$) for the all three models while these values for the quadratic and exponential 1 in the case of parametric approach are 13.2% and 12.7%, respectively.

We also study the performance of precision of EBLUP of small area means by providing corresponding prediction intervals using DC approach. To this end, for each simulation run r, we calculate $\theta_i^{(r)}$ and compute appropriate quantiles α and $(1-\alpha)$ of the posterior distribution of $\hat{\theta}_i^{(r)}$ using (7). In particular, the coverage probabilities of the $\hat{\theta}_i$ is the proportion of the times (over $R = 1000$) that $\theta_i^{(r)}$ falls within $(\hat{\theta}_i^{(r)}(\alpha), \hat{\theta}_i^{(r)}(1-\alpha))$. Table 3 shows the average coverage probabilities and average lengths of prediction intervals of the estimates of small area means for the P-spline and parametric approaches for the three different models. The proposed P-spline approach also performs very well in terms of average coverage probabilities of the EBLUP of small area means for the all three different models. The corresponding parametric method also performs well in terms of coverage probabilities of the $\hat{\theta}_i$.

TABLE 3: Average coverage probabilities (and average lengths) of small area means with different confidence coefficients by data cloning approach, P-spline and parametric approaches for three different models (linear, quadratic, and exponential 1) in the case of linear mixed model.

True model	Method	Confidence coefficient (average lengths)			
		0.90	0.95	0.98	0.99
Linear	P-spline	0.882 (2.288)	0.936 (2.726)	0.971 (3.235)	0.983 (3.582)
	Parametric	0.882 (2.286)	0.936 (2.725)	0.971 (3.234)	0.983 (3.581)
Quadratic	P-spline	0.886 (2.335)	0.939 (2.783)	0.972 (3.303)	0.984 (3.657)
	Parametric	0.895 (2.868)	0.945 (3.417)	0.976 (4.055)	0.987 (4.489)
Exponential 1	P-spline	0.884 (2.339)	0.937 (2.787)	0.971 (3.308)	0.984 (3.663)
	Parametric	0.897 (2.962)	0.945 (3.529)	0.973 (4.188)	0.983 (4.636)

We should point out that in the non-parametric linear mixed model proposed by Opsomer et al. (2008), we need to analytically derive tedious algebra to get $mspe(\hat{\theta}_i)$, while in DC approach, not only can we easily get $mspe(\hat{\theta}_i)$, but also we can get the corresponding prediction intervals for $\hat{\theta}_i$.

4.2. Non-parametric logistic mixed model

We also conduct a simulation study to evaluate the performance of proposed approach in the non-parametric logistic mixed model set-up. To our knowledge, the $\mathrm{glmmPQL}$ function in MASS package in R is the only existing package to be able to handle our non-parametric GLMMs. We then compare the performance of both DC and PQL methods in our non-parametric logistic mixed model. To that end, we first generate $R = 2000$ independent samples:

$$y_{i,s}^{(r)} \sim Binomial(n_i, \mu_i^{(r)}) \tag{8}$$

$$\log(\frac{\mu_i^{(r)}}{1-\mu_i^{(r)}}) = m_0(x_i) + \nu_i^{(r)}, \quad (i = 1, ..., m),$$

with $\nu_i^{(r)} \overset{i.i.d.}{\sim} N(0, \sigma_\nu^2)$, and three different choices of $m_0(x)$ as linear $(-0.1 + 0.01x)$, quadratic $(-0.1 + 0.01x + 0.01x^2)$ and exponential 1 $(-0.1 + 0.01x + 0.1\exp(x))$. We set $n_i = 5, m = 100$, and x_i's are generated from uniform distribution between -10 and 0 once and treat them fixed in the simulation study. Using the simulated datasets $\{(y_{i,s}^{(r)}, x_i), i = 1, ..., 100; r = 1, ..., 2000\}$ with $\sigma_\nu^2 = 1$, we apply the approaches of DC and PQL to estimate the model parameters and also to predict the small area proportion $\hat{\mu}_i$ for each simulation run r using

$$\log(\frac{\hat{\mu}_i^{(r)}}{1-\hat{\mu}_i^{(r)}}) = \hat{\beta}_0 + \hat{\beta}_1 x_i + \sum_{l=1}^{25} \hat{\gamma}_l (x_i - \kappa_l)_+ + \hat{\nu}_i, \quad (i = 1, ..., 100).$$

For the DC in this simulation set-up, the average number of clones was $K = 20$ to obtain MLE and the average number of iterations for convergence of the model parameters was about 50,000.

Similar to the linear mixed model, we also study the EMSPE of $\hat{\mu}_i$, RB of $mspe(\hat{\mu}_i)$, and the average coverage probabilities of $\hat{\mu}_i$. We report the average EMSPE of small area proportions over areas for the three different models and two approaches PQL and DC (Table 4). It is clear

TABLE 4: Average EMSPE of small area proportions over areas by penalized quasi-likelihood (PQL) and data cloning (DC) approaches for our P-spline logistic mixed model.

True model	Average EMSPE	
	DC	PQL
Linear	0.022	0.244
Quadratic	0.022	0.207
Exponential 1	0.022	0.245

TABLE 5: Percent average absolute relative bias of estimators of MSPE of small area proportions over areas by data cloning approach for our P-spline logistic mixed model.

True model	RB(%)
Linear	6.63
Quadratic	6.98
Exponential 1	7.43

TABLE 6: Average coverage probabilities (and average lengths) of small area proportions over areas with different confidence coefficients by data cloning approach for our P-spline logistic mixed model.

True model	Confidence coefficient (average lengths)			
	0.90	0.95	0.98	0.99
Linear	0.885 (0.468)	0.939 (0.543)	0.974 (0.621)	0.986 (0.669)
Quadratic	0.884 (0.467)	0.938 (0.542)	0.972 (0.620)	0.985 (0.668)
Exponential 1	0.884 (0.467)	0.938 (0.542)	0.972 (0.620)	0.985 (0.668)

from Table 4 that the EMSPE of PQL is almost 10 times bigger than corresponding values in the DC method. The AARB of MSPE estimator of small area proportions over areas for the three different models for the DC approach is also provided in Table 5, noting that the PQL ($\mathrm{glmmPQL}$ function in R) can't provide the MSPE estimator of small area proportions. Similar to the linear mixed model, the DC also performs very well for our P-spline approach in terms of AARB ($< \%8$) as shown in Table 5 for the all three models. The results of the average coverage probabilities and average lengths of small area proportions and different coefficients are also given in Table 6 for the three different models. The DC approach also performs very well for our P-spline approach in terms of coverage probabilities and average lengths of prediction intervals of the small area proportions for different confidence coefficients and for the all three different models, noting that we are unable to provide prediction intervals of small area proportions for the PQL method using $\mathrm{glmmPQL}$ function in R.

4.3. Non-parametric Poisson mixed model

We also conduct a simulation study to evaluate the performance of proposed approach in the non-parametric Poisson mixed model set-up. To that end, we first generate $R = 1000$ independent samples:

$$y_i^{(r)} \sim Poisson(N_i \mu_i^{(r)}) \tag{9}$$

TABLE 7: Average EMSPE of small area rates over areas by data cloning approach for our P-spline Poisson mixed model.

True model	Average EMSPE
Linear	3.691
Quadratic	4.472
Exponential 1	3.753

TABLE 8: Percent average absolute relative bias of estimators of MSPE of small area rates over areas by data cloning approach for our P-spline Poisson mixed model.

True model	RB(%)
Linear	8.12
Quadratic	6.74
Exponential 1	7.46

$$\log(\mu_i^{(r)}) = m_0(x_i) + \nu_i^{(r)}, \quad (i = 1, ..., m),$$

with $\nu_i^{(r)} \overset{i.i.d.}{\sim} N(0, \sigma_\nu^2)$, offset $N_i = 3$, number of areas $m = 50$, and three different choices of $m_0(x)$ as linear $(-0.1 + 0.01x)$, quadratic $(-0.1 + 0.01x + 0.1x^2)$ and exponential 1 $(-0.1 + 0.01x + 0.1\exp(x))$. The x_i's are generated from uniform distribution between -10 and 0 once and treat them fixed in the simulation study. The true small area rate of $i-$th area for each simulation run r is $\mu_i^{(r)}$. Using the simulated datasets $\{(y_i^{(r)}, x_i), i = 1, ..., 50; r = 1, ..., 1000\}$ with $\sigma_\nu^2 = 1$, we apply the DC approach to estimate the model parameters and also to predict the small area rate μ_i for each simulation run r using

$$\log(\hat{\mu}_i^{(r)}) = \hat{\beta}_0^{(r)} + \hat{\beta}_1^{(r)} x_i + \sum_{l=1}^{13} \hat{\gamma}_l^{(r)} (x_i - \kappa_l)_+ + \hat{\nu}_i^{(r)}, \quad (i = 1, ..., 50).$$

For the DC approach in this simulation set-up, the average number of clones was $K = 20$ to obtain MLE and the average number of iterations for convergence of the model parameters was about 50,000 in our P-spline model.

Similar to the other simulation studies (Sections 4.1 and 4.2), we also study the EMSPE of $\hat{\mu}_i$, RB of $mspe(\hat{\mu}_i)$, and the average coverage probabilities of $\hat{\mu}_i$. We report the average EMSPE of small area rates over areas for the three different models using the DC approach for our P-spline model (Table 7). The AARB of MSPE estimator of small area rates over areas for the three different models for the DC approach is also provided in Table 8. Similar to the previous simulation studies, the DC approach also performs very well for our P-spline model in terms of AARB ($< \%9$) as shown in Table 8 for the all three models. The results of the average coverage probabilities and average lengths of small area rates and different coefficients are also given in Table 9 for our P-spline model in the case of three different models (linear, quadratic, and exponential 1). The DC approach also performs very well for our P-spline model in terms of coverage probabilities and average lengths of prediction intervals of the small area rates for different confidence coefficients and for the all three different models.

TABLE 9: Average coverage probabilities (and average lengths) of small area rates over areas with different confidence coefficients by data cloning approach for our P-spline Poisson mixed model.

True model	Confidence coefficient (average lengths)			
	0.90	0.95	0.98	0.99
Linear	0.886 (5.185)	0.939 (6.261)	0.973 (7.562)	0.986 (8.480)
Quadratic	0.900 (9.750)	0.952 (11.620)	0.981 (13.793)	0.990 (15.264)
Exponential 1	0.886 (5.253)	0.939 (6.341)	0.974 (7.656)	0.985 (8.585)

TABLE 10: Parameters estimate and corresponding standard error (SE) of Total Respiratory Morbidity study in Manitoba using data cloning approach for our P-spline logistic mixed model.

Parameter	β_0	β_1	σ_ν^2	σ_γ^2
Estimate	-3.01	-0.78	0.05	0.01
SE	0.03	0.14	0.01	0.002

5. APPLICATION

The performance of the proposed approach is also evaluated by using a real dataset of logistic mixed model. We study physician visits for Total Respiratory Morbidity (TRM) conditions (a patient diagnosed with any of the following respiratory diseases: asthma, chronic or acute bronchitis, emphysema, or chronic airway obstruction, and chronic obstructive pulmonary disease) in the Canadian province of Manitoba during 2000 – 2010 fiscal years. The population of Manitoba was stable during the study period from 1.15 million in 2000 to 1.20 million in 2010. The province consisted of eleven Regional Health Authorities that were responsible for the delivery of health care services. These eleven regions were further sub-divided into 67 Regional Health Authorities Districts (RHAD) and these RHADs were used as areas in our model. Our interest is to use the non-parametric logistic mixed model to make an inference on the rate of physician TRM visits in the 67 RHADs. Let y_i and n_i be the number of physician TRM visits and corresponding population at risk, respectively, and x_i is the average age in the ith area. Let $y_i \sim Binomial(n_i, \mu_i)$ and consider the following P-spline logistic mixed model:

$$\log(\frac{\mu_i}{1-\mu_i}) = \beta_0 + \beta_1 x_i + \sum_{l=1}^{L} \gamma_l (x_i - \kappa_l)_+ + \nu_i, \quad (i = 1, \ldots, m),$$

where μ_i is the rate of physician TRM visits in area i, β_0 is overall mean log-odds over areas, β_1 is the coefficient of age, $m = 67$ is the number of areas, $L = 17$ is the number of fixed knots, $\gamma_l \overset{i.i.d.}{\sim} N(0, \sigma_\gamma^2)$ and $\nu_i \overset{i.i.d.}{\sim} N(0, \sigma_\nu^2)$.

The estimate of model parameters and associated standard errors are reported in Table 10. For this specific application, the number of clones was $K = 100$ to obtain MLE with number of iterations 20,000 for the convergence of the model parameters. As mentioned in Section 3, if scaled variances decrease at a $1/K$ rate and have reached a lower bound (say < 0.05), the DC approach has converged (Figure 1). One of the main features of the DC approach is the ability to provide the prediction (and prediction interval) of random effects. We provide prediction (Figure 2) and 95% prediction intervals (Figure 3) of the physician TRM visit rates.

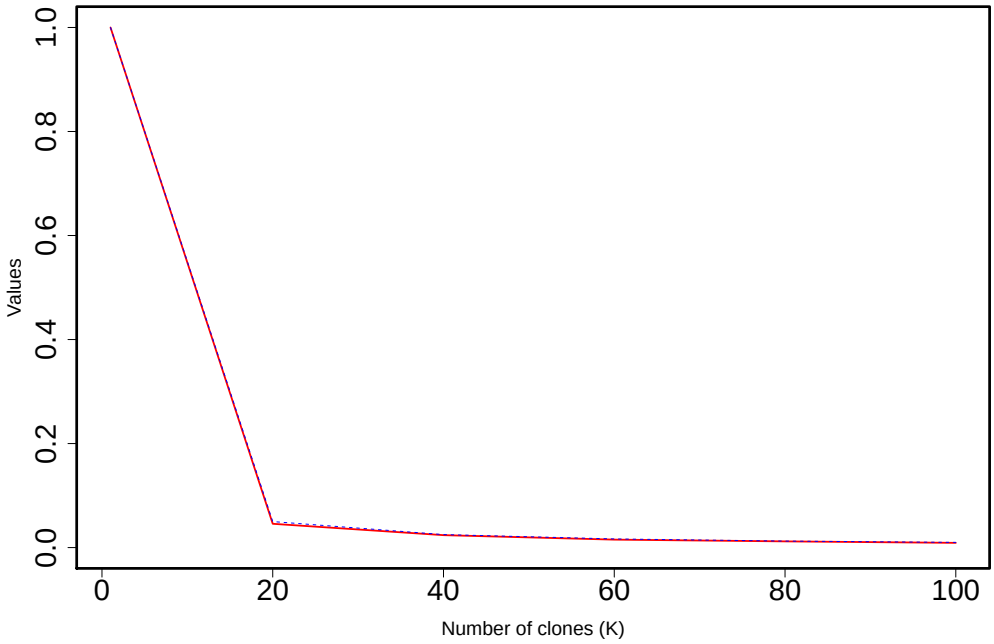

FIGURE 1: Data cloning convergence diagnostics for Total Respiratory Morbidity study dataset. The standardized maximum eigenvalues (solid line) converge to zero at the expected rate $1/K$ (dashed line).

6. CONCLUDING REMARKS

Linear mixed models using penalized spline (P-spline) regression models have been previously studied in the context of small area estimation. There are, however, many applications in small area estimation where response variables are counts or proportions. In this paper, small area estimation under generalized linear mixed models (GLMMs) using P-spline regression models has been proposed to cover Normal and non-Normal responses.

The analysis of these non-parametric mixed models is extremely difficult using frequentist paradigm. Analysis based on data cloning has overcome the computational difficulties of the maximum likelihood method. We have used the data cloning for our non-parametric GLMMs to unify the analysis of Normal and non-Normal responses from frequentist perspective. Under the linear mixed model set-up, we have shown by simulation study that our proposed approach works very well in terms of coverage probabilities of small area means; noting that one needs to do tedious algebra to get an estimator of mean squared prediction errors (MSPE) of small area means (Opsomer et al., 2008), while using data cloning not only can we easily get MSPE estimation of small area means, but also we can get the corresponding prediction intervals of small area means. Under the non-parametric logistic and Poisson mixed models, we have also shown by simulation studies that our proposed approach works very well in terms of coverage probabilities of small area proportions and rates, respectively. Our proposed approach was also evaluated by a real dataset of physician visits due to Total Respiratory Morbidity in 67 health regions in Manitoba under the P-spline logistic mixed model.

In our Total Respiratory Morbidity study, we assumed that the physician visits were independent from each other; however, we had some cases who visited physicians multiple times over the study period. We have planned to study this kind of data more appropriately in a separate

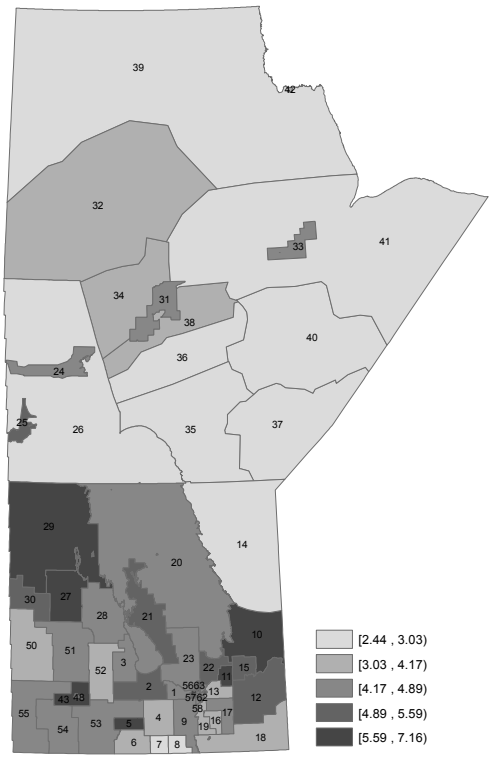

FIGURE 2: Percent prediction of Total Respiratory Morbidity visit rates in 67 health regions in Manitoba using data cloning approach for our P-spline logistic mixed model.

manuscript.

We have considered a single covariate in our proposed model; however, it can be easily extended to multiple covariates which is more applicable in real life situations. In our proposed model, we assumed that random effects have Normal distribution; however, one can relax this assumption and consider mixture of Normal distributions (or other appropriate distributions) and study the misspecification of our proposed model. Another extension of our work would be to consider P-spline regressions in non-linear mixed models. We have planned to study these models in our future study.

Recently, Baghishani, Rue, and Mohammadzadeh (2012) used the DC method via Integrated Nested Laplace Approximation (Rue, Martino, and Chopin, 2009) in the class of GLMMs called hybrid DC. They showed that this approximation approach is faster than the usual MCMC and with same efficiency, noting that it is only applicable for Normal random effects. One can use this approach for our proposed model to speed up the computing time of model parameters estimate.

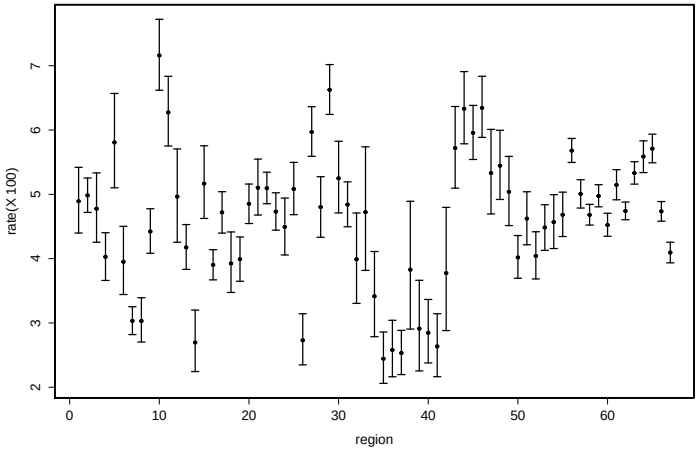

FIGURE 3: The 95% prediction bands of Total Respiratory Morbidity visit rates in 67 health regions in Manitoba using data cloning approach for our P-spline logistic mixed model. The bullet represents prediction rates with corresponding lower and upper prediction bands.

ACKNOWLEDGEMENTS

We would like to thank the Editor, Associate Editor, and two referees for constructive comments and suggestions, which led to an improved version of the manuscript. F. Shokoohi was a post-doctoral fellow and M. Torabi is a full-time faculty member at the University of Manitoba, Canada. This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) and Manitoba Health Research Council (MHRC).

Disclaimer: The interpretations, conclusions and opinions expressed in this paper are those of the authors and do not necessarily reflect the position of the Manitoba Health. This study is based in part on data provided by Manitoba Health through Manitoba Centre for Health Policy. The interpretation and conclusions contained herein are those of the researchers and do not necessarily represent the views of the government of Manitoba.

BIBLIOGRAPHY

Baghishani, H. and M. Mohammadzadeh (2011). A Data Cloning Algorithm for Computing Maximum Likelihood Estimates in Spatial Generalized Linear Mixed Models. *Computational and Data Analysis 55*(4), 1748–1759.

Baghishani, H., H. Rue, and M. Mohammadzadeh (2012). On a hybrid data cloning method and its application in generalized linear mixed models. *Statistics and Computing 22*(1), 597–613.

Battese, G., R. Harter, and W. Fuller (1988). An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. *Journal of the American Statistical Association 83*(401), 28–36.

Boor, C. D. (2001). *A Practical Guide to Splines*. Springer-Werlag.

Breidt, F. J., G. Claeskens, and J. D. Opsomer (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika 92*(4), 831–846.

Chambers, R. and N. Tzavidis (2006). M-quantile Models for Small Area Estimation. *Biometrika 93*(2), 255–268.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with B-splines and penalties. *Statistical Science 11*(2), 89–121.

Fay, I. R. E. and R. A. Herriot (1979). Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association 74*(366), 269–277.

Gilks, W., S. Richardson, and D. Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC.

Hamilton, J. D. (1986). A Standard Error for the Estimated State Vector of a State-Space Model. *Journal of Econometrics 33*(3), 387 – 397.

Jiang, J. (2010). *Large Sample Techniques for Statistics*. Springer, New York.

Jiang, J. and P. Lahiri (2006). Mixed Model Prediction and Small Area Estimation. *Test 15*(1), 1–96.

Jiang, J., T. Nguyen, and J. S. Rao (2010). Fence Method for Nonparametric Small Area Estimation. *Survey Methodology 36*(1), 3–11.

Lele, S. R., B. Dennis, and F. Lutscher (2007). Data Cloning: Easy Maximum Likelihood Estimation for Complex Ecological Models Using Bayesian Markov Chain Monte Carlo Methods. *Ecology Letters 10*(7), 551–563.

Lele, S. R., K. Nadeem, and B. Schmuland (2010). Estimability and Likelihood Inference for Generalized Linear Mixed Models Using Data Cloning. *Journal of the American Statistical Association 105*(492), 1617–1625.

Opsomer, J., G. Claeskens, M. Ranalli, and G. Kauermann (2008). Non-Parametric Small Area Estimation Using Penalized Spline Regression. *Journal of the Royal Statistical Society: Series B 70*(1), 265–286.

Pratesi, M., M. G. Ranalli, and N. Salvati (2008). Semiparametric M-quantile Regression for Estimating the Proportion of Acidic Lakes in 8-digit HUCs of the Northeastern US. *Environmetrics 19*(7), 687–701.

Pratesi, M., M. G. Ranalli, and N. Salvati (2009). Nonparametric M-quantile Regression Using Penalised Splines. *Journal of Nonparametric Statistics 21*(3), 287–304.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rao, J. N. K. (2003). *Small Area Estimation*. Wiley-Interscience.

Rao, J. N. K., S. K. Sinha, and L. Dumitrescu (2014). Robust small area estimation under semi-parametric mixed models. *Canadian Journal of Statistics 42*(1), 126–141.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(2), 319–392.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics 11*(4), 735–757.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.

Salvati, N., M. G. Ranalli, and M. Pratesi (2011). Small Area Estimation of the Mean Using Non-parametric M-quantile Regression: a Comparison when a Linear Mixed Model does not Hold. *Journal of Statistical Computation and Simulation 81*(8), 945–964.

Smith, B. J. (2007). BOA: An R Package for MCMC Output Convergence Assessment and Posterior Inference. *Journal of Statistical Software 21*(11), 1–37.

Sólymos, P. (2010). dclone: Data cloning in R. *The R Journal 2*(2), 29–37.

Sperlich, S. and M. José Lombardía (2010). Local Polynomial Inference for Small Area Statistics: Estimation, Validation and Prediction. *Journal of Nonparametric Statistics 22*(5), 633–648.

Spiegelhalter, D. J., K. R. Abrams, and J. P. Myles (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley.

Torabi, M. and F. Shokoohi (2012). Likelihood Inference in Small Area Estimation by Combining Time-Series and Cross-Sectional Data. *Journal of Multivariate Analysis 111*(0), 213–221.

Walker, A. M. (1969). On the Asymptotic Behaviour of Posterior Distributions. *Journal of the Royal Statistical Society. Series B (Methodological) 31*(1), 80–88.

Received 24 July 2013
Accepted 12 October 2014