doi: 10.1111/j.1467-842X.XXX

Semi-parametric small-area estimation by combining time-series and cross-sectional data methods

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

Concordia University and University of Manitoba

Summary

In survey sampling, policymaking regarding the allocation of resources to subgroups (called small areas) or the determination of subgroups with specific properties in a population should be based on reliable estimates. Information, however, is often collected at a different scale than that of these subgroups, hence the estimation can only be obtained on finer scale data. Parametric mixed models are commonly used in small-area estimation. The relationship between predictors and response, however, may not be linear in some real situations. Recently, small-area estimation under the generalized linear mixed model (GLMM) with a penalised spline (P-spline) regression model, for the fixed part of the model, has been proposed to analyse cross-sectional responses, both normal and non-normal. However, there are many situations in which we have time-related responses in small areas such as an annual dataset on the number of asthma physician visits in different areas of Manitoba, Canada. In cases where covariates that can possibly predict the asthma physician visits (such as age and genetic and environmental factors) may not have a linear relationship with the response, new models for analysing such datasets are required. In the current work, using both time-series and cross-sectional data methods, we propose P-spline regression models for small-area estimation under the GLMMs. Our proposed model covers both normal and non-normal responses. In particular, the empirical best predictors of small-area parameters and their corresponding prediction intervals are studied where the maximum likelihood estimation approach is used to estimate the model parameters. The performance of the proposed approach is evaluated using some simulations and also by analysing two real datasets (precipitation and asthma).

Key words: data cloning; exponential family; maximum likelihood estimation; penalised spline; random effects

*Author to whom correspondence should be addressed.
[1] Department of Mathematics and Statistics, Concordia University, Montreal, QC, H3G 1M8, Canada
[2] Departments of Community Health Sciences and Statistics, University of Manitoba, MB, R3E 0W3, Canada, Email: Mahmoud.Torabi@umanitoba.ca
Acknowledgment. This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) to M. Torabi.
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.

1. Introduction

Sample surveys are commonly conducted to provide reliable estimates of finite population parameters such as totals, means, counts, quantiles, etc. In recent years, there has been an increasing demand for such estimates for sub-populations (small areas), such as counties or gender-age groups, to use in formulating policies and programs, allocating government funds, regional planning, and making decisions at a local level, amongst other uses. The sample sizes within areas, however, are often too small to warrant the use of the traditional area-specific direct estimates.

To produce reliable estimates of characteristics of interest for small areas and obtain measures of error associated with each estimate, a number of methods have been proposed in the literature. These include, among others, the use of synthetic, composite and/or model-based estimators (Jiang & Lahiri 2006; Pfeffermann 2013; Rao & Molina 2015). Model-based estimators borrow strength from related areas both by defining a set of assumptions for modelling the stochastic behaviour of the variables in the underlying population and by introducing random effects into the model. In the context of mixed models, such small-area models may be classified into two broad types: (i) Area-level models (Fay & Herriot 1979) that relate small-area direct estimates to area-specific covariates; such models are used if unit-level data are not available. (ii) Unit-level models (Battese, Harter & Fuller 1988) that relate the unit values to associated unit-level covariates with known area means and area-specific covariates.

Parametric models have been extensively used in small-area estimation. On the other hand, research which investigates non- or semi-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 hold using penalised splines (P-splines) regression. Torabi & Shokoohi (2015) proposed generalised linear mixed models (GLMMs) using P-spline regression to unify the analysis of normal and non-normal responses. From a very different perspective, Chambers & Tzavidis (2006) studied an approach for small-area estimation that is based on M-quantile regression which allows for models that are robust to the distributional assumptions on the errors and area effects. However, when the functional form of the relationship between q-th M-quantile and the covariates is not linear, this approach 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 using a non-parametric specification of the conditional M-quantile of the response variable given the covariates has been also studied (Pratesi, Ranalli & Salvati 2008, 2009; Salvati, Ranalli & Pratesi 2011). Jiang, Nguyen & Rao (2010) developed an adaptive fence procedure (Jiang et al. 2008) for selecting semi-parametric models using P-splines. Sperlich & José Lombardía (2010) used local polynomial inference in the context of small-area estimation.

There has been a limited amount of research based on time series in the context of small-area estimation. Scott & Smith (1974) and Jones (1980), among others, used time series models to develop efficient estimates of aggregated parameters in the repeated survey setting. Tiller (1992) used the Kalman filter to combine a current-period state-wide estimate from the US Current Population Survey with the past estimates for the same state. However, these authors did not investigate the idea of effecting small-area estimation by combining cross-sectional and time-series data. Pfeffermann & Burck (1990) and Singh, Mantel & Thomas (1991), among others, studied cross-sectional and time-series models for small-area estimation using the Kalman filter by assuming specific models for the sampling errors over time. Rao & Yu (1994) proposed a combined cross-sectional and time-series model involving autocorrelated random effects and sampling errors with an arbitrary covariance matrix over time. Datta, Lahiri & Maiti (2002) applied a similar model to the Rao-Yu model having replaced autoregressive (AR) random effects part with a random walk model. Datta et al. (1999) considered a similar model but added extra terms to reflect seasonal variation. Torabi (2012) extended the Datta et al. (1999) model to account for spatial variation over areas/regions. Torabi & Shokoohi (2012) considered cross-sectional and time-series models for both normal and non-normal responses in a specific parametric model. Recently, Boubeta, Lombardia & Morales (2017) also used a time-related response to study empirical best predictors under area-level Poisson mixed models.

The contribution of the current paper is two-fold. The first aim of this paper is to develop semi-parametric models to unify the analysis of both discrete and continuous responses in the class of GLMMs for time-series and cross-sectional data. It is well known that frequentist analysis of these models is computationally difficult. There are some approximate methods based on the frequentist paradigm for analysing mixed models, such as Penalised quasi-likelihood (PQL), Laplace approximation and Gauss-Hermite quadrature, among other approaches. Recently, Lele, Dennis & Lutscher (2007) introduced an approach, called data cloning (DC), to compute maximum likelihood estimates (MLEs) and their corresponding standard errors for general hierarchical models. Data cloning is a computational algorithm based on Markov chain Monte Carlo (MCMC) which yields the MLE. Lele, Nadeem & Schmuland (2010) used the DC method to compute point predictions and prediction intervals for random effects in the class of GLMMs. As the second aim of this paper, we propose to use the DC method to make inference for our proposed semi-parametric mixed models for normal and non-normal responses by combining time-series and cross-sectional data methods in the context of small-area estimation.

The rest of this manuscript is organised as follows. Semi-parametric mixed models for combined time-series and cross-sectional data are introduced in Section 2. In Section 3, we describe how the DC method can be used for obtaining parameter estimates and predictions with their corresponding standard errors. We report the results of simulation studies for evaluating the performance of the proposed approach in Section 4. In Section 5, we consider analyses of two real datasets, a Canadian precipitation dataset and an asthma physician visits dataset from the Canadian province of Manitoba. Finally, some concluding remarks are given in Section 6.

2. Semi-parametric mixed models

A semi-parametric model for time-series and cross-sectional data utilising P-splines on the covariates is described as follows. Let Y_{it} denote the variable of interest in area $i(= 1, ..., m)$ at time $t(= 1, ..., T)$. The Y_{it} values are assumed to be conditionally independent, given the random effects, with exponential family density

$$f_Y(y_{it}|\zeta_{it}, \varsigma_{it}) = \exp\left(\frac{y_{it}\zeta_{it} - a(\zeta_{it})}{\varsigma_{it}} + b(y_{it}, \varsigma_{it})\right),\ i = 1, ..., m;\ t = 1, ..., T \quad (1)$$

The density (1) is parameterised with respect to the canonical parameters ζ_{it}, known scale parameters ς_{it} and known functions $a(\cdot)$ and $b(\cdot)$. The exponential family (1) covers well-known distributions including normal, binomial and Poisson. The natural parameters ζ_{it} for semi-parametric regression model are then modelled as

$$\zeta_{it} = h(\theta_{it}) = m_0(x_{it}) + \nu_i + u_{it},\ i = 1, ..., m;\ t = 1, ..., T, \quad (2)$$

where h is a strictly increasing function to guarantee a one-to-one relationship between θ_{it} and natural parameters $\zeta_{it}, \theta_{it} = E(y_{it}|\zeta_{it}, \varsigma_{it})$, and $\nu_i \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2_\nu), i = 1, \dots, m$, are area specific random effects. We assume that u_{it}'s follow a common $AR(1)$ process for each area i; that is,

$$u_{it} = \rho u_{i(t-1)} + \epsilon_{it}, \quad |\rho| < 1, \quad (3)$$

with $\epsilon_{it} \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2_\epsilon)$. The function $m_0(x_{it})$ is unknown but it is assumed that it can be approximated sufficiently well by following P-spline:

$$m_0(x_{it}) \approx \beta_0 + \beta_1 x_{it} + \cdots + \beta_p x^p_{it} + \sum_{l=1}^{L} \gamma_l (x_{it} - \kappa_l)^p_+. \quad (4)$$

In the formula above, p is the degree of the spline, $(x)_+^p$ denotes the function $x^p \mathbb{1}(x > 0)$, where $\mathbb{1}(.)$ denotes the indicator function, x_{it} is a known value (covariate), $\{\kappa_1, \ldots, \kappa_L\}$ is a set of fixed knots, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_L)^\top$ are the regression coefficients and P-spline part of the model, respectively and L is the number of spline knots. We assume that $\gamma_l \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma_\gamma^2)$, $l = 1, \ldots, L$. The random variables ν_i, γ_l, and ϵ_{it} are also assumed to be independent of each other. Inference is then carried out based on the P-spline model (4).

It is well known that if the location of the knots is sufficiently spread out over the range of x_{it} and if L is sufficiently large, then the class of models which are adequately approximated using P-splines is very large and includes most smooth functions (Eilers & Marx 1996; De Boor 2001). It is recommended to use the minimum of 40 and $n_c/4$, where n_c is the number of unique values of x_{it}, as the number of spline knots L (Ruppert 2002). We follow this recommendation in this paper. We refer the readers to Ruppert, Wand & Carroll (2003) for more details on P-spline regression models.

A special case of model (2) is $h(\theta_{it}) = \theta_{it}$. The area-level mixed model can be written

$$y_{it} = \theta_{it} + e_{it} = m_0(x_{it}) + \nu_i + u_{it} + e_{it}, \;\; u_{it} = \rho u_{i(t-1)} + \epsilon_{it},$$

and if $m_0(x_{it})$ is approximated sufficiently well, then the area-level semi-parametric mixed model is given by

$$y_{it} \approx \beta_0 + \beta_1 x_{it} + \cdots + \beta_p x_{it}^p + \sum_{l=1}^{L} \gamma_l (x_{it} - \kappa_l)_+^p + \nu_i + u_{it} + e_{it}, \;\; u_{it} = \rho u_{i(t-1)} + \epsilon_{it},$$

for $i = 1, ..., m,\ t = 1, ..., T$, when, given the θ_{it}, $\boldsymbol{e} = (e_{11}, ..., e_{mT})^\top$ is a vector of normally distributed sampling errors, given θ_{it}'s, with zero means and a known (to avoid identifiability issues) block diagonal covariance matrix Ψ with blocks Ψ_i.

3. Likelihood-based estimation

Let $\boldsymbol{\alpha} = (\boldsymbol{\beta}^\top, \rho, \sigma_\nu^2, \sigma_\gamma^2, \sigma_\epsilon^2)^\top$ denote the unknown parameters in the model described by (1)-(4). The marginal likelihood of the data denoted by $L(\boldsymbol{\alpha}; \boldsymbol{y})$ is obtained by integrating conditional probabilities of responses over the distribution of random effects as follows:

$$L(\boldsymbol{\alpha}; \boldsymbol{y}) = \int\int\int \prod_{i=1}^{m} \prod_{t=1}^{T} f(y_{it} | \zeta_{it}, \varsigma_{it}) g(\zeta_{it} | \rho, \sigma_\nu^2, \sigma_\gamma^2, \sigma_\epsilon^2) d\nu_i du_{it} d\gamma_l, \tag{5}$$

where $f(\cdot)$ is the semi-parametric mixed model defined as (1)-(4), and $g(\cdot)$ is a multivariate normal distribution with appropriate mean and covariance matrix.

We use the DC method to obtain the MLE of the parameters which appear in (5). The DC method uses the Bayesian computational approach for frequentist purposes. To understand the logic behind the DC method, imagine a hypothetical situation where the observations $\boldsymbol{y} = (y_{11}, ..., y_{mT})^\top$ are repeated independently by K different individuals, and all these individuals happen to result in exactly the same set of observations $\boldsymbol{y}$. We denote these repeated datasets by $\boldsymbol{y}^{(K)} = (\boldsymbol{y}^\top, \boldsymbol{y}^\top, ..., \boldsymbol{y}^\top)^\top$. The likelihood function for the combination of the data from these K independent experiments is then given by $\{L(\boldsymbol{\alpha}; \boldsymbol{y})\}^K = L^K(\boldsymbol{\alpha}; \boldsymbol{y})$. Note that this likelihood function has two important features:

1. The location of the maximum of this function is exactly equal to the location of the maximum of $L(\boldsymbol{\alpha}; \boldsymbol{y})$.
2. The Fisher information matrix based on this likelihood is K times the Fisher information matrix based on $L(\boldsymbol{\alpha}; \boldsymbol{y})$.

Let $\hat{\boldsymbol{\alpha}}$ be the MLE and $\mathfrak{I}(\hat{\boldsymbol{\alpha}})$ be the corresponding Fisher information matrix based on $L(\boldsymbol{\alpha}; \boldsymbol{y})$. We assume that the model is identifiable and there is a unique mode (but possibly multiple smaller peaks) for the likelihood function. The posterior distribution of $\boldsymbol{\alpha}$ conditional on the data $\boldsymbol{y}^{(K)}$ is then given by

$$\pi_K(\boldsymbol{\alpha}|\boldsymbol{y}^{(K)}) = \frac{L^K(\boldsymbol{\alpha}; \boldsymbol{y})\pi(\boldsymbol{\alpha})}{C(\boldsymbol{y}^{(K)})}, \tag{6}$$

where $\pi(\boldsymbol{\alpha})$ is the prior distribution and $C(\boldsymbol{y}^{(K)}) = \int L^K(\boldsymbol{\alpha}; \boldsymbol{y})\pi(\boldsymbol{\alpha})d\boldsymbol{\alpha}$ is the normalising constant. The following theorem guarantees that inference based on $L^K(\boldsymbol{\alpha}; \boldsymbol{y})$, the likelihood of K copies of the original data, is closely related to inference based on $L(\boldsymbol{\alpha}; \boldsymbol{y})$:

Theorem 1. *Consider the general model described by (1)-(4). Under some mild regularity conditions, as K becomes large, the posterior distribution of $\sqrt{K}\Sigma^{-1/2}(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})|\boldsymbol{y}^{(K)}$ converges to a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{I}$ which is the identity matrix with the dimension of $\boldsymbol{\alpha}$, $\hat{\boldsymbol{\alpha}}$ is the MLE, and Σ is the inverse of the Fisher information matrix for the MLE.*

Proof. It suffices to show that the distributions considered in our model satisfy the assumptions A.1-A.3 considered in Lele, Nadeem & Schmuland (2010). First, it is obvious that each sampling distribution $f_Y(.)$ (i.e. normal, binomial and Poisson), as a function of θ, has a local maximum which we shall denote by $\boldsymbol{\theta}_\infty$, and that $f_Y(\boldsymbol{\theta}_\infty) > 0$ and $\pi(\boldsymbol{\theta}_\infty) > 0$. The maximum likelihood estimator is then $\boldsymbol{\theta}_\infty$. Second, for each pair of functions $\pi(.)$ and $f_Y(.)$, the function $\pi(.)$ is continuous at any interior point of parameter space and is thus continuous at $\boldsymbol{\theta}_\infty$. Likewise the function $f_Y(.)$ has continuous second derivatives in a neighbourhood of any interior point as well as at $\boldsymbol{\theta}_\infty$, and $D^2 f_Y(\boldsymbol{\theta}_\infty)$ is strictly negative definite since it belongs to exponential family. Third, since the sampling functions $f_Y(.)$ belong to the exponential family and have local maxima, for any $\delta > 0$, we have $\gamma(\delta) = \sup\{f_Y(\boldsymbol{\theta}) : \|\boldsymbol{\theta} - \boldsymbol{\theta}_\infty\| > \delta\} < f(\boldsymbol{\theta}_\infty)$. Therefore, the rest of proof follows along the lines of Lele, Nadeem & Schmuland (2010).

Theorem 1 assures that the sample mean vector of the generated random numbers from the posterior distribution (6) provides the MLE of the model parameters $\boldsymbol{\alpha}$, and furthermore K times their sample covariance matrix is an estimate of the asymptotic covariance matrix of the MLE $\hat{\boldsymbol{\alpha}}$.

Lele, Nadeem & Schmuland (2010) also provided various checks to determine the value of K which constitutes an adequate number of clones. For instance, one may plot the ratio of the 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. By investigating the graph one can determine if the posterior distribution has become nearly degenerate. As another criterion, it is approximately true that as we increase the number of clones, $(\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}})^\top V^{-1}(\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}}) \sim \chi^2_q$, where $\bar{\boldsymbol{\alpha}}$ and 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) $\zeta = \sum_{b=1}^{B}(O_b - E_b)^2/B$, where O_b and E_b are observed and estimated quantiles for χ^2_q random variable, and (b) $\tilde{r}^2 = 1 - \tau^2$, where τ is the correlation between O and E. If these statistics are close to zero, it indicates that the foregoing χ^2 approximation is reasonable. Note that the foregoing three criteria have been implemented in the `dclone` package (Sólymos 2010), which is freely available in R (R Development Core Team 2016). We use these criteria to obtain the appropriate number of clones in our simulations and in the data analyses.

3.1. Prediction of small-area parameters

The main goal in small area estimation is to predict small-area parameters θ_{it} and to determine the precision of these predictions. Following Hamilton (1986) and Lele, Nadeem & Schmuland (2010), based on the MLE of $\boldsymbol{\alpha}$, the prediction of (and the prediction interval for) θ_{it}, conditional on the observed data, is obtained using MCMC algorithm under the following posterior density

$$\frac{\int f(\boldsymbol{y}|\zeta_{it}, \boldsymbol{\beta}) g(\zeta_{it}|\rho, \sigma^2_\nu, \sigma^2_\gamma, \sigma^2_\epsilon)\phi(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}, \mathfrak{I}^{-1}(\hat{\boldsymbol{\alpha}}))d\boldsymbol{\alpha}}{C(\boldsymbol{y})}. \tag{7}$$

In (7) $f(\cdot)$ and $g(\cdot)$ are as in (5), and $\phi(., \mu, \Sigma)$ denotes a multivariate normal density with mean μ and covariance Σ, which are set equal here to the MLE of μ and the inverse of the Fisher information matrix. Also in (7) $C(\boldsymbol{y}) = \int L(\boldsymbol{\alpha}; \boldsymbol{y})\pi(\boldsymbol{\alpha})d\boldsymbol{\alpha}$ is the normalising constant.

The prior distributions $\pi(\boldsymbol{\alpha})$ in the DC method are chosen as $\beta_j \sim \mathrm{N}(0, 10^6), j = 0, \ldots, p$, $\sigma_\nu \sim \text{Uniform}(0, 1000)$, $\sigma_\gamma \sim \text{Uniform}(0, 1000)$, $\sigma_\epsilon \sim \text{Uniform}(0, 1000)$ and $\rho \sim \text{Uniform}(-1, 1)$. Note that the results in the DC method are invariant to the choice of priors. However, if one uses appropriate/informative priors, a smaller number of clones (K) will be needed in order to achieve convergence. To monitor the convergence of the algorithm, we use several diagnostic methods implemented in the Bayesian output analysis (BOA) program (Smith 2007) in R. We also use diagnostic methods implemented in the `dclone` package (Sólymos 2010) to monitor the convergence of the algorithm in terms of the number of clones K as described in Section 3. We have also provided the R code for the simulation studies and data analyses as supplementary materials; please contact the first author for questions related to the R code.

4. Simulation study

4.1. Normal mixed model

We conducted a simulation study to evaluate performance of the proposed approach in the semi-parametric normal mixed model set-up. We used the following semi-parametric area-level model as the true model under which the samples for the simulation study were generated. We used the following set-up for our simulation study:

$$\begin{aligned} y_{it} &= m_0(x_{it}) + \nu_i + u_{it} + e_{it}, \quad i = 1, ..., m;\ t = 1, ..., T, \\ u_{it} &= \rho u_{i,t-1} + \epsilon_{it}, \quad |\rho| < 1. \end{aligned}$$

We set $m = 50$, $T = 5$, $\rho = 0.4$, $e_{it} \overset{i.i.d.}{\sim} \mathrm{N}(0, 1)$, $\nu_i \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2_\nu)$ and $\epsilon_{it} \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2_\epsilon)$ where $\sigma^2_\nu = \sigma^2_\epsilon = 1$. Following Breidt, Claeskens & Opsomer (2005) and Rao, Sinha & Dumitrescu (2014), we considered these three different choices of $m_0(x_{it})$:

1) Linear: $m_0(x_{it}) = 1 + x_{it}$,
2) Quadratic: $m_0(x_{it}) = 1 + x_{it} + 0.5\, x^2_{it}$,
3) Exponential: $m_0(x_{it}) = 1 - x_{it} + 0.5\, \exp(x_{it})$.

We generated x_{it} from a normal distribution with mean 0 and variance 1 once and treated them as fixed in the simulation study. Throughout the simulation study, we used the linear P-spline approximation ($p = 1$) for $m_0(x_{it})$. Following Ruppert (2002); Ruppert, Wand & Carroll (2003), we set the number of knots to be $L = 40$. We generated $R = 1000$

independent samples

$$\{(y_{it}^{(r)}, x_{it}), i = 1, ..., m; t = 1, ..., T; r = 1, ..., R\},$$

assuming

$$y_{it}^{(r)} = m_0(x_{it}) + \nu_i^{(r)} + u_{it}^{(r)} + e_{it}^{(r)},$$

where $\nu_i^{(r)}, \epsilon_{it}^{(r)}$ and $e_{it}^{(r)}$ were generated from the corresponding normal distributions with $\sigma_\nu^2 = \sigma_\epsilon^2 = \sigma_e^2 = 1$. For each simulated run, we applied the DC method to get the MLE of the model parameters and also to provide the prediction and prediction intervals of the empirical best linear unbiased predictor (EBLUP) of small-area means. That is, we calculated

$$\theta_{it}^{(r)} = m_0(x_{it}) + \nu_i^{(r)} + u_{it}^{(r)},$$

using

$$\hat{\theta}_{it}^{(r)} = \hat{\beta}_0^{(r)} + \hat{\beta}_1^{(r)} x_{it} + \sum_{l=1}^{40} E[\gamma_l^{(r)}(x_{it} - \kappa_l)_+ | \boldsymbol{y}_i]_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}} + E[\nu_i^{(r)} + u_{it}^{(r)} | \boldsymbol{y}_i]_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}},$$

with $\gamma_l \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma_\gamma^2)$ and $\boldsymbol{y}_i = (y_{i1}, ..., y_{iT})^\top$. We also compared our proposed P-spline regression model with the corresponding parametric model which is simply $m_0(x_{it}) = x_{it}^\top \boldsymbol{\beta}$. For each iteration, we then have

$$\tilde{\theta}_{it,p}^{(r)} = \hat{\beta}_0^{(r)} + \hat{\beta}_1^{(r)} x_{it} + E[\nu_i^{(r)} + u_{it}^{(r)} | \boldsymbol{y}_i]_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}}.$$

For this simulation set-up, the *average* number of clones needed to obtain the MLE was $K = 20$, and the *average* number of iterations needed to achieve convergence was about 10,000. We calculated the empirical mean squared prediction error (EMSPE) of small-area means as

$$\text{EMSPE}(\hat{\theta}_{it}) = \frac{1}{R} \sum_{r=1}^{R} \{\hat{\theta}_{it}^{(r)} - \theta_{it}^{(r)}\}^2.$$

Also, the relative bias (RB) of an estimator of the MSPE, say mspe, was calculated as

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

where $\hat{\theta}_{it}^{(r)}, \theta_{it}^{(r)}$, and $\text{mspe}^{(r)}(\hat{\theta}_{it})$ are the values of $\hat{\theta}_{it}, \theta_{it}$, and $\text{mspe}(\hat{\theta}_{it})$ for the *r*-th simulation run, respectively. Note that $\text{mspe}(\hat{\theta}_{it})$ is the variance of $\hat{\theta}_{it}$ whence this quantity can be computed under the posterior distribution (7).

Table 1. Average EMSPE of small-area means $\hat{\theta}_{iT}$ over areas in the case of the P-spline normal mixed model.

True model	Approach	
	P-spline	Parametric
Linear	0.585	0.584
Quadratic	0.593	0.708
Exponential	0.610	0.706

Table 2. Percent AARB of estimators of MSPE of small-area means $\hat{\theta}_{iT}$ over areas in the case of the P-spline normal mixed model.

True model	Approach	
	P-spline	Parametric
Linear	4.94	4.15
Quadratic	5.13	11.18
Exponential	5.81	9.25

The average EMSPE of small-area means $\hat{\theta}_{iT}$ (for the current time T) over areas for all three pre-specified models $m_0(x_{iT})$ (linear, quadratic, exponential) for both P-spline and parametric models are reported in Table 1. The results show that the values of EMSPE are stable for the P-spline method for all three pre-specified models $m_0(x_{iT})$ while these values increase for the quadratic and exponential parametric models. Table 2 reports the average absolute relative bias in percent (AARB) of mspe over areas for the three different models $m_0(x_{iT})$ for both P-spline and parametric models. The proposed P-spline model performs reasonably well in terms of AARB (AARB $< 6\%$) for the all three models $m_0(x_{iT})$. The parametric model, however, gives much higher values than the semi-parametric model for both the quadratic and exponential models.

We are also interested in obtaining prediction intervals for the small-area means. To this end, for each simulation run r, we calculate $\theta_{iT}^{(r)}$ and compute appropriate quantiles α and $(1-\alpha)$ of $\hat{\theta}_{iT}^{(r)}$. In particular, the coverage probability of $\hat{\theta}_{iT}$ is calculated as the proportion of the times (over $R = 1000$) that $\theta_{iT}^{(r)}$ falls within $(\hat{\theta}_{iT}^{(r)}(\alpha), \hat{\theta}_{iT}^{(r)}(1-\alpha))$. Table 3 shows the coverage probabilities and the average lengths of the prediction intervals for $\hat{\theta}_{iT}$ for the P-spline and parametric models for all three pre-specified models $m_0(x_{iT})$. The proposed P-spline model performs well in terms of the average coverage probabilities of the prediction intervals $\hat{\theta}_{iT}$ for all three pre-specified models $m_0(x_{iT})$. The corresponding parametric model performs well in terms of the coverage probabilities but the P-spline method produces slightly shorter confidence intervals.

Table 3. Average coverage probabilities (and average lengths) of prediction intervals for small-area means $\hat{\theta}_{iT}$ over areas in the case of the P-spline normal mixed model.

True model	Approach	Confidence coefficient (average lengths)			
		0.90	0.95	0.98	0.99
Linear	P-spline	0.892 (2.458)	0.944 (2.927)	0.976 (3.471)	0.987 (3.838)
	Parametric	0.893 (2.468)	0.944 (2.940)	0.977 (3.485)	0.988 (3.849)
Quadratic	P-spline	0.892 (2.471)	0.942 (2.943)	0.976 (3.490)	0.988 (3.860)
	Parametric	0.888 (2.666)	0.940 (3.175)	0.973 (3.762)	0.985 (4.155)
Exponential	P-spline	0.890 (2.484)	0.941 (2.958)	0.975 (3.508)	0.986 (3.879)
	Parametric	0.890 (2.642)	0.940 (3.146)	0.971 (3.728)	0.982 (4.117)

4.2. Logistic mixed model

We also conducted a simulation study to evaluate performance of the proposed approach in the semi-parametric logistic mixed model context. To that end, we first generated $R = 1000$ independent samples from the following model:

$$y_{it}^{(r)} \sim \text{Binomial}(n_{it}, \theta_{it}^{(r)}), \tag{8}$$
$$\log(\frac{\theta_{it}^{(r)}}{1-\theta_{it}^{(r)}}) = m_0(x_{it}) + \nu_i^{(r)} + u_{it}^{(r)}, \quad i = 1, ..., m; t = 1, ..., T; r = 1, ..., R,$$

where $\nu_i^{(r)} \overset{i.i.d.}{\sim} \text{N}(0, \sigma_\nu^2)$, $u_{it}^{(r)}$ were generated from an AR(1) model with $(\rho, \sigma_\epsilon^2)$, $\epsilon_{it}^{(r)} \overset{i.i.d.}{\sim} \text{N}(0, \sigma_\epsilon^2)$. Three different choices of $m_0(x_{it})$, linear $(0.1 + 0.01x_{it})$, quadratic $(0.1 + 0.01x_{it} + 0.5\ x_{it}^2)$, and exponential $(0.1 - 0.01x_{it} + 0.5\ \exp(x_{it}))$ were used. We set $m = 50, T = 5, n_{it} = 5, \sigma_\nu^2 = \sigma_\epsilon^2 = 1$, and $\rho = 0.4$. The values of x_{it}'s were generated once from the Uniform$(-10, 0)$ distribution and they were then treated them as fixed in the simulation study.

Using the simulated datasets $\{(y_{it}^{(r)}, x_{it}), i = 1, ..., m; t = 1, ..., T; r = 1, ..., R\}$, we applied the DC method to estimate the model parameters and also to predict the small-area proportion θ_{it} for each simulation run r using

$$\log(\frac{\hat{\theta}_{it}^{(r)}}{1-\hat{\theta}_{it}^{(r)}}) = \hat{\beta}_0^{(r)} + \hat{\beta}_1^{(r)} x_{it} + \sum_{l=1}^{40} E[\gamma_l^{(r)}(x_{it} - \kappa_l)_+ | \boldsymbol{y}_i]_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}} + E[\nu_i^{(r)} + u_{it}^{(r)} | \boldsymbol{y}_i]_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}},$$

where $\gamma_l \overset{i.i.d.}{\sim} \text{N}(0, \sigma_\gamma^2)$. The *average* number of clones needed to obtain the MLE was $K = 20$ and the *average* number of iterations required for convergence was about 10,000.

Similarly to the normal mixed model setting, we studied the EMSPE of $\hat{\theta}_{it}$, the RB of mspe$(\hat{\theta}_{it})$, and the average coverage probabilities of $\hat{\theta}_{it}$. Note that in the case of logistic and Poisson (Section 4.3) mixed models, we have only reported performance of our proposed

Table 4. Average EMSPE of small-area proportions $\hat{\theta}_{iT}$ over areas for the P-spline logistic mixed model.

True model	Average EMSPE
Linear	0.020
Quadratic	0.020
Exponential	0.020

Table 5. AARB of estimators of MSPE of small-area proportions $\hat{\theta}_{iT}$ over areas for the P-spline logistic mixed model.

True model	AARB (in %)
Linear	5.22
Quadratic	4.95
Exponential	5.67

Table 6. Average coverage probabilities (and average lengths) of small-area proportions $\hat{\theta}_{iT}$ over areas with different confidence coefficients for the P-spline logistic mixed model.

True model	Confidence coefficient (average lengths)			
	0.90	0.95	0.98	0.99
Linear	0.888 (0.444)	0.942 (0.517)	0.974 (0.595)	0.986 (0.642)
Quadratic	0.891 (0.441)	0.942 (0.513)	0.974 (0.590)	0.986 (0.638)
Exponential	0.889 (0.441)	0.940 (0.514)	0.974 (0.591)	0.985 (0.639)

P-spline model and have not provided the results of the corresponding parametric models as they had behaviour similar to that observed in the normal mixed model (Section 4.1). Table 4 shows the average EMSPE of the small-area proportions $\hat{\theta}_{iT}$ (for the current time T) over areas for all three pre-specified models $m_0(x_{iT})$. As shown in Table 4, the values of average EMSPE are small and stable for all of the models. The AARB of mspe$(\hat{\theta}_{iT})$ over areas is reported in Table 5. Similarly to the normal mixed model setting, these results show that the proposed P-spline model works reasonably well in terms of the AARB (AARB $\leq 6\%$). The average coverage probabilities and the average lengths of prediction intervals of small-area proportions $\hat{\theta}_{iT}$ over areas for different coefficients are given in Table 6. The proposed P-spline model also performs well in terms of the average coverage probabilities and the average lengths of prediction intervals of the small-area proportions $\hat{\theta}_{iT}$ over areas for all of the models considered.

4.3. Poisson mixed model

We also conducted a simulation study to evaluate performance of the proposed approach in the semi-parametric Poisson mixed model set-up. To that end, we first generated $R = 2000$ independent samples from the following model:

$$\begin{aligned} y_{it}^{(r)} &\sim \text{Poisson}(N_{it}\theta_{it}^{(r)}), \\ \log(\theta_{it}^{(r)}) &= m_0(x_{it}) + \nu_i^{(r)} + u_{it}^{(r)}, \quad i = 1, ..., m; t = 1, ..., T; r = 1, ..., R, \end{aligned} \tag{9}$$

where $\nu_i^{(r)} \overset{i.i.d.}{\sim} \text{N}(0, \sigma_\nu^2)$, $u_{it}^{(r)}$ were generated from AR(1) with $(\rho, \sigma_\epsilon^2)$, $\epsilon_{it}^{(r)} \overset{i.i.d.}{\sim} \text{N}(0, \sigma_\epsilon^2)$. Three different choices of $m_0(x_{it})$ as linear $(0.1 + 0.01x_{it})$, quadratic $(0.1 + 0.01x_{it} + 0.1\, x_{it}^2)$, and exponential $(0.1 + 0.01x_{it} + 0.1\exp(x_{it}))$ were used. We chose $m = 50, T = 5, N_{it} = 3, \rho = 0.4$, and $\sigma_\nu^2 = \sigma_\epsilon^2 = 1$. We generated the x_{it}'s from normal distribution with mean 0 and variance 1, and then treated them as fixed in the simulation study.

Using the simulated datasets $\{(y_{it}^{(r)}, x_{it}), i = 1, ..., m; t = 1, ..., T; r = 1, ..., R\}$, we applied the DC method to estimate the model parameters and also to predict the small-area rate θ_{it} for each simulation run r using

$$\log(\hat{\theta}_{it}^{(r)}) = \hat{\beta}_0^{(r)} + \hat{\beta}_1^{(r)} x_{it} + \sum_{l=1}^{40} E[\gamma_l^{(r)}(x_{it} - \kappa_l)_+ | \boldsymbol{y}_i]_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}} + E[\nu_i^{(r)} + u_{it}^{(r)} | \boldsymbol{y}_i]_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}},$$

where $\gamma_l \overset{i.i.d.}{\sim} \text{N}(0, \sigma_\gamma^2)$. The *average* number of clones needed to obtain the MLE was $K = 20$ and the *average* number of iterations required for convergence was about 10,000.

Similarly to the other simulation studies in this work, we studied the EMSPE of $\hat{\theta}_{it}$, the RB of mspe$(\hat{\theta}_{it})$, and the average coverage probabilities of $\hat{\theta}_{it}$. Table 7 shows the average EMSPE of the small-area rates $\hat{\theta}_{iT}$ (for the current time T) over areas for all three pre-specified models $m_0(x_{iT})$. As shown in Table 7, the values of average EMSPE increase from the linear to the quadratic and to the exponential model. The AARB of mspe$(\hat{\theta}_{iT})$ over areas is reported in Table 8. Similarly to the other mixed models considered in our simulation studies, the proposed P-spline model performs reasonably well in terms of the AARB (AARB $\leq 9.4\%$). We note that when the number of simulations was increased from 2000 to 5000 in the case of the exponential model, we even got better AARB $\leq 4.30\%$ (These results are not shown here). The average coverage probabilities and the average lengths of the prediction intervals of small-area rates $\hat{\theta}_{iT}$ over areas for different coefficients are given in Table 9. The proposed P-spline model performs well in terms of the average coverage probabilities and the average lengths of the prediction intervals of the small-area rates $\hat{\theta}_{iT}$ over areas for different confidence coefficients and for the all three pre-specified models $m_0(x_{iT})$.

Table 7. Average EMSPE of small-area rates $\hat{\theta}_{iT}$ over areas for the P-spline Poisson mixed model.

True model	Average EMSPE
Linear	9.24
Quadratic	10.97
Exponential	11.89

Table 8. AARB of estimators of MSPE of small-area rates $\hat{\theta}_{iT}$ over areas for the P-spline Poisson mixed model.

True model	AARB (in %)
Linear	6.38
Quadratic	9.39
Exponential	7.41

Table 9. Average coverage probabilities (and average lengths) of small-area rates $\hat{\theta}_{iT}$ over areas with different confidence coefficients for the P-spline Poisson mixed model.

True model	Confidence coefficient (average lengths)			
	0.90	0.95	0.98	0.99
Linear	0.897 (7.303)	0.947 (8.777)	0.978 (10.536)	0.989 (11.758)
Quadratic	0.897 (7.851)	0.947 (9.426)	0.978 (11.300)	0.988 (12.601)
Exponential	0.898 (8.188)	0.948 (9.827)	0.978 (11.774)	0.989 (13.124)

5. Applications

5.1. Homogenized and adjusted Canadian climate data (HACCD)

The website of Environment and Climate Change Canada provides homogenized and adjusted climate datasets for many climatological stations in Canada. The homogenized surface air temperature for Canada (HSATC2) data provides monthly, seasonal and annual means of the daily maximum, minimum and mean temperatures (Vincent et al. 2012). The adjusted precipitation for Canada (APC2) dataset provides adjusted daily rainfall, snowfall and total precipitation for many locations in Canada (Mekis & Vincent 2011). These datasets have been discussed and analyzed in a number of papers, for example, Mekis & Hogg (1999), Zhang et al. (2000), Alexander et al. (2006),Vincent & Mekis (2006), among others.

We used the annual mean temperature in HSATC2 and the annual total precipitation in APC2 for those stations that appear in both datasets. As a result 29 locations were selected. Only records from the years 1967 – 1976 (30 years in all) were used due to incompleteness of the data for other years. We refer to the resulting combined data set as the “HACCD” data.

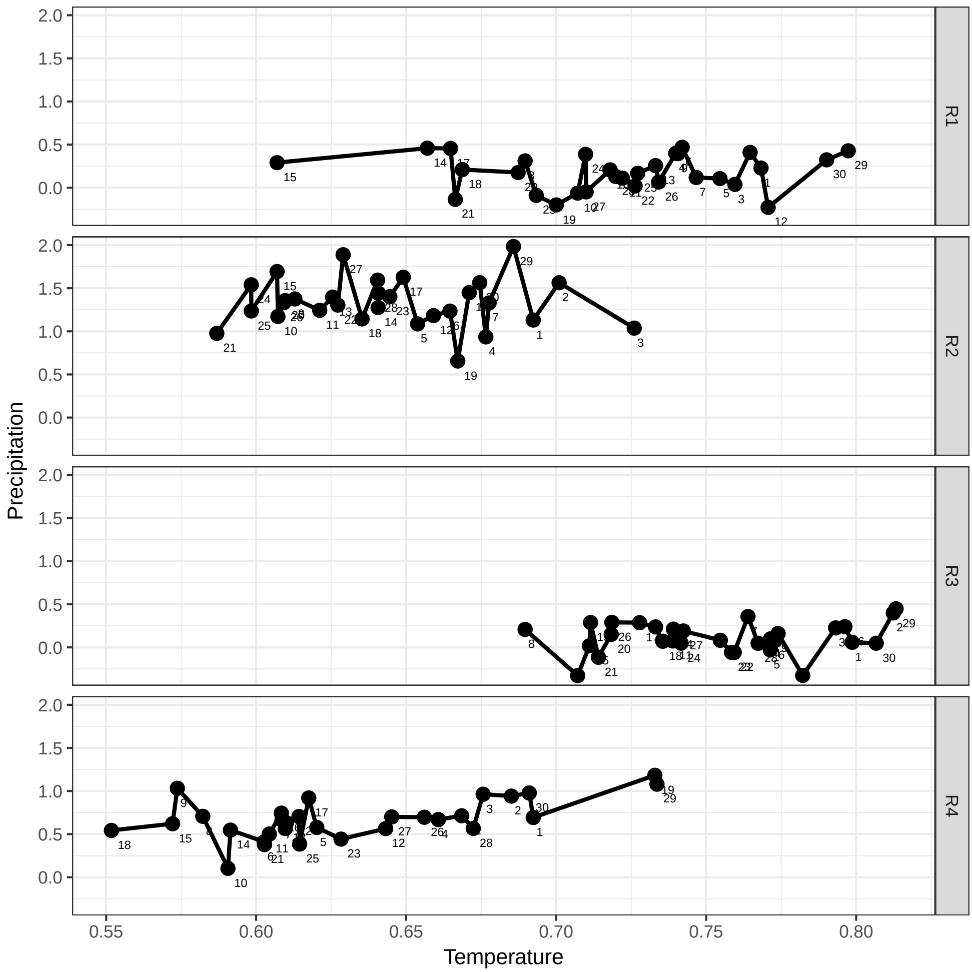


Figure 1. Annual precipitation versus annual temperature for four selected locations after normalising the data. Note that the numbers 1 to 30, with which the graphs are annotated, refer to years 1967 to 1996, and R refers to location.

We were interested in the relationship between annual total precipitation and annual mean temperature for each location in Canada. Note that other datasets, e.g. homogenized surface pressure data and homogenized surface wind speed data, were available, however, in this analysis we focused only on HSATC2 and APC2. Figure 1 depicts the relationship between annual precipitation and annual temperature for selected locations. From this graph, one may conclude that a parametric linear mixed model cannot describe the relationship between annual precipitation and annual temperature.

After normalising the response and covariate, we fitted the following model to the data:

Table 10. Model parameter estimates and corresponding standard errors (SE) and 95% confidence intervals (LCI, UCI) for precipitation at 29 locations in Canada during the years 1967–1996. The estimates were calculated using the P-spline normal mixed model.

Parameter	β_0	β_1	ρ	σ^2_ϵ	σ^2_ν	σ^2_γ
Estimate	0.043199	-0.066147	-0.216700	0.000369	0.904384	0.001008
SE	0.032795	0.050246	0.287821	0.000126	0.055443	0.000548
LCI	-0.009677	-0.188412	-0.674034	0.000193	0.802029	0.000324
UCI	0.097085	0.026536	0.460943	0.000673	1.018250	0.002288

$$\begin{aligned} y_{it} &= \beta_0 + \beta_1 x_{it} + \sum_{l=1}^{40} \gamma_l (x_{it} - \kappa_l)_+ + \nu_i + u_{it} + e_{it}, \quad (i = 1, ..., 29; t = 1, ..., 30), \\ u_{it} &= \rho u_{i,t-1} + \epsilon_{it}, \quad |\rho| < 1, \end{aligned}$$

The model parameter estimates, standard errors, and their corresponding 95% confidence intervals are reported in Table 10. For the HACCD dataset, the number of clones needed to obtain the MLE was $K = 30$ and the number of iterations required for convergence was 50,000. For the variance of the sampling errors we used all available data to obtain a smooth estimate which turned out to be approximately 1. This value was used in the analysis. Figure 2 shows predictions and corresponding 95% prediction intervals for the precipitation at the 29 locations in question, for the year 1996. The predictions are expressed in terms of the normalised data.

5.2. Asthma physician visits

We used our proposed approach to analyse the dataset of annual physician visits relating to Total Respiratory Morbidity (TRM) condition. Such visits consist of visits by patients diagnosed with any of the following respiratory diseases: asthma, chronic or acute bronchitis, emphysema, or chronic airway obstruction, and chronic obstructive pulmonary disease. These data were collected in the Canadian province of Manitoba during the 2000-2010 fiscal years. The population of Manitoba was reasonably stable during the study period, varying only from 1.15 million individuals in 2000 to 1.20 million individuals in 2010. The province is subdivided into five Regional Health Authorities that are responsible for the delivery of health care services. These five regions are further sub-divided into 222 Regional Health Authorities Districts (RHADs). Through the expedient of removing missing values, 217 of these RHADs became available for use in the analysis. For simplicity, we used these RHADs as areas; we denoted them by R1, R2,...,R217. In this analysis, our interest was to study the effect of age

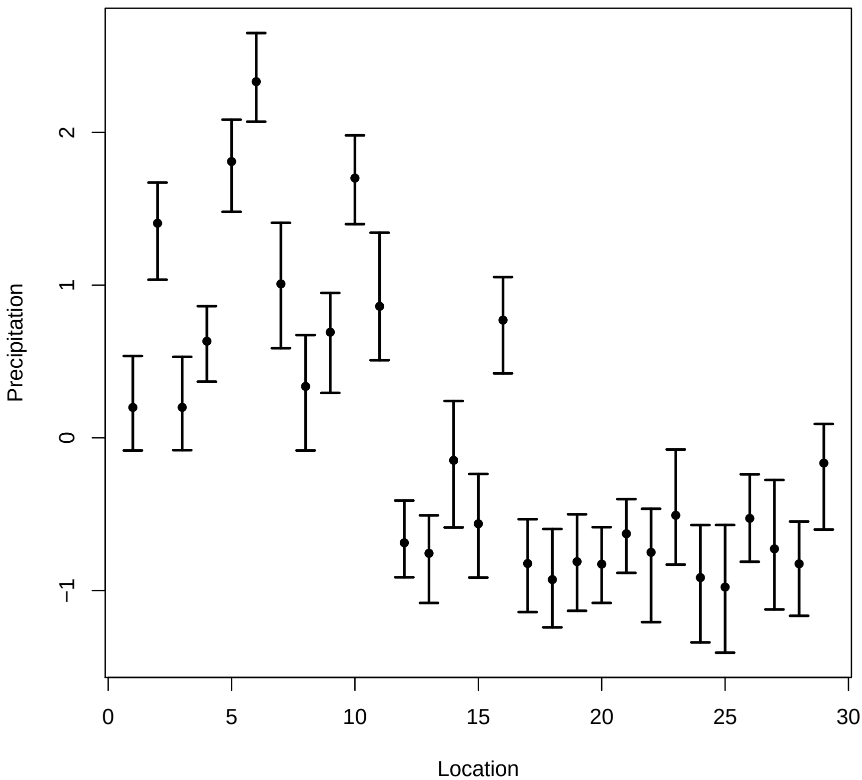


Figure 2. Ninety-five % prediction intervals for average precipitation in the year 1996 at different locations in Canada; the data were normalised. The bullets represent point predictions; the error bars constitute the corresponding prediction intervals.

as a risk factor on the TRM condition. Figure 3 depicts the complex relationship between logit of physician rate and the average age for some selected RHADs. From this figure we can argue that a parametric model is not suitable to fit the data and we therefore turned to a semi-parametric model. Our interest was in using the P-spline logistic mixed model to make inferences about the rate of physician TRM visits in all of the 217 RHADs in different years. The sample sizes for some areas were not large enough to produce reliable estimates. Hence we applied the following model:

$$\begin{aligned} & y_{it} \sim \text{Binomial}(n_{it}, \theta_{it}), \\ & \log\left(\frac{\theta_{it}}{1-\theta_{it}}\right) = \beta_0 + \beta_1 z_{it} + \beta_2 x_{it} + \sum_{l=1}^{L} \gamma_l (x_{it} - \kappa_l)_+ + \nu_i + u_{it}, \\ & i = 1, ..., 217; \quad t = 1, ..., 10, \end{aligned} \tag{10}$$

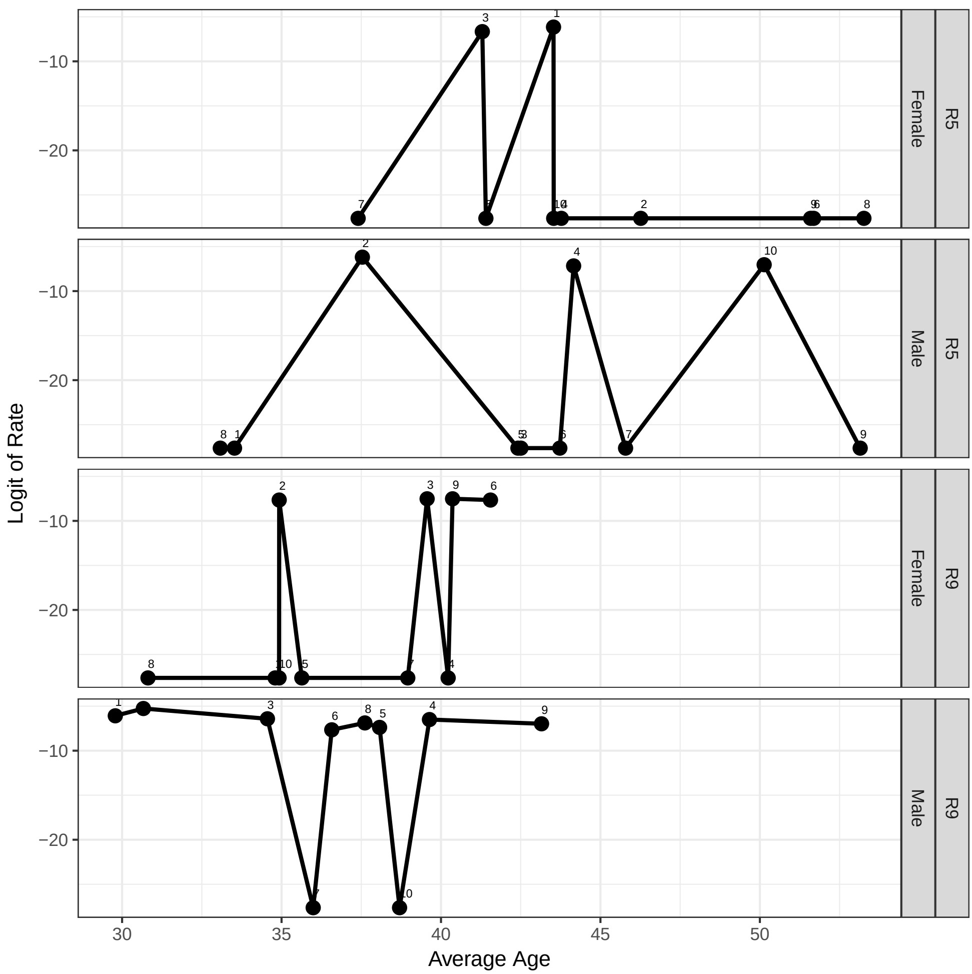


Figure 3. Logit of physician rate versus the average age for females and males for some selected RHADs; zero rates were replaced by 10^{-20} to avoid ∞. Note that 1 to 10 refer to years 2000 to 2010.

where y_{it} and n_{it} are the total number of physician TRM visits and the corresponding population at risk in area Ri at time t, respectively. The quantity θ_{it} is the rate of physician TRM visits in area Ri at time t; β_0 is the overall mean of the log-odds over areas and times; z_{it} and x_{it} are the percentage of females and average age in area Ri at time t, respectively, with the corresponding coefficients β_1 and β_2; $L = 40$ is the number of knots. We assumed that $\gamma_l \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma_\gamma^2), \nu_i \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma_\nu^2)$, $u_{it} = \rho u_{i,t-1} + \epsilon_{it}$ with $|\rho| < 1$ and $\epsilon_{it} \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma_\epsilon^2)$. The model parameters estimate, standard errors, and their corresponding 95% confidence intervals are reported in Table 11. For this particular dataset, the number of clones needed to obtain the MLE was $K = 20$ and the number of iterations required to achieve convergence was 20,000. One of the main features of the DC method is the ability to provide predictions

Table 11. Model parameter estimates and corresponding standard errors (SE) and 95% confidence intervals (LCI, UCI) for physician TRM visits in Manitoba during 2000–2010. The estimates were calculated using the P-spline logistic mixed model.

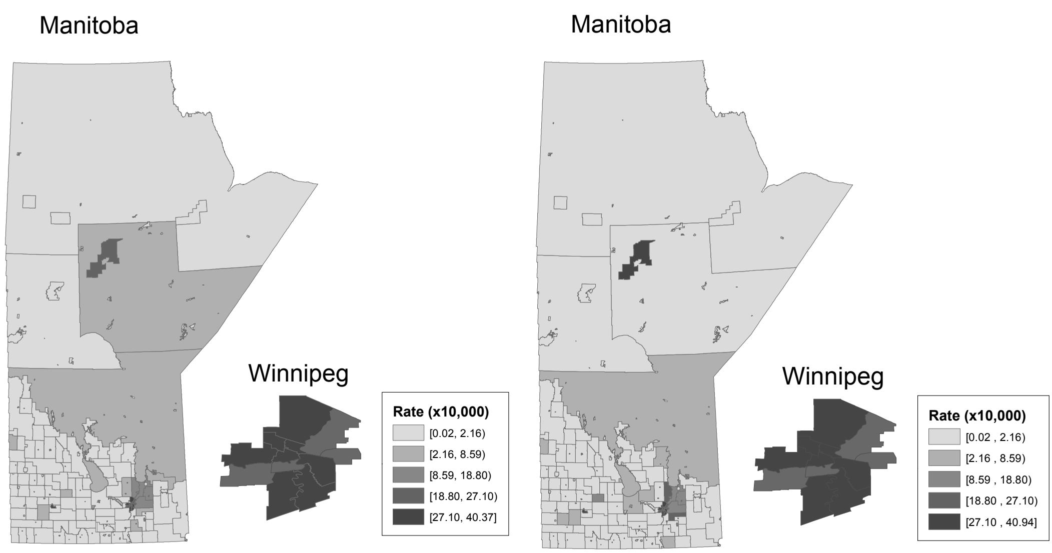
Parameter	β_0	β_1	β_2	ρ	σ^2_ϵ	σ^2_ν	σ^2_γ
Estimate	-11.951	0.300	0.703	0.758	1.850	13.470	0.011
SE	0.092	0.040	0.087	0.006	0.025	0.477	0.004
LCI	-12.131	0.218	0.218	0.747	1.800	12.570	0.004
UCI	-11.766	0.377	0.866	0.770	1.900	14.438	0.020

and prediction intervals for random effects. We provide predictions (Figure 4) and 95% prediction intervals (Figure 5) for the physician TRM visit rates for different RHADs in 2010 for both females and males. Overall our analysis suggests that Winnipeg and some areas in southern Manitoba have larger rates of asthma visits compared to other parts of the province. These findings may represent real increases or different distributions of important covariates that are unmeasured and unadjusted for in our modelling. Further investigation is needed to explore these findings.

6. Concluding comments

Mixed models using penalised spline (P-spline) regression models have previously been studied in the context of small-area estimation for the cross-sectional data. There are, however, many real situations in small-area estimation in which the response variables are serially dependent over time. Models accommodating such serial dependence have not previously been developed. In this paper we propose semi-parametric mixed models which combine time-series and cross-sectional data methodology, using P-spline regression models for both normal and non-normal responses.

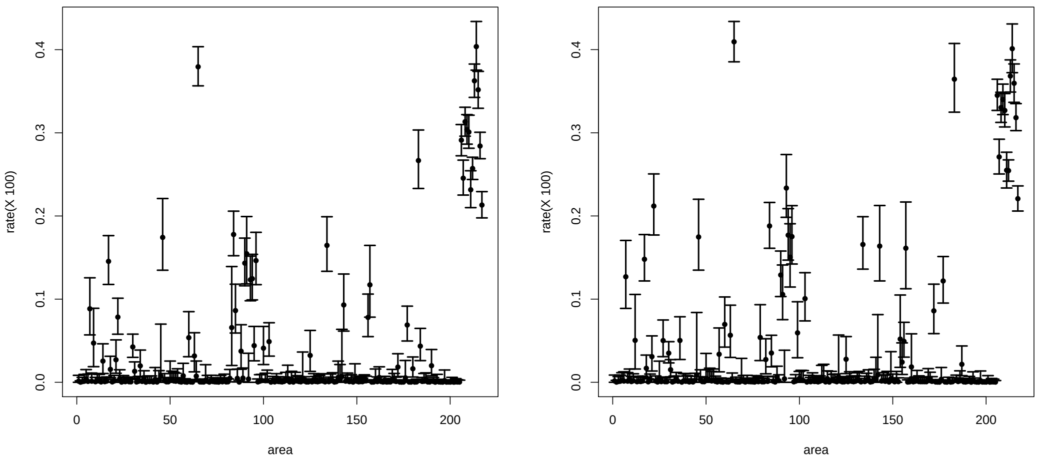
We make use of a data cloning approach to inference in order to obtain maximum likelihood estimates of the parameters of the proposed P-spline mixed models. Under the semi-parametric normal mixed model set-up, we study finite sample properties of our proposed approach. Our approach appears to work reasonably well in terms of the coverage probabilities of the small-area means. We also studied finite sample properties of our proposed approach in the context of semi-parametric logistic and Poisson mixed models. Our approach also appears to work well in this context, in terms of the coverage probabilities of small-area proportions and rates, respectively. We used our proposed approach to analyse two real datasets, consisting of observations of precipitation and of physician visits, using semi-parametric normal and logistic mixed models, respectively.



(a) Females (b) Males

Figure 4. Prediction of TRM visit rates in 217 RHADs for females and males in Manitoba in 2010. The predictions were made using the P-spline logistic mixed model

To accommodate serial dependence we used an AR(1) model in our procedure. However other time series models such as random walks, higher orders of AR, Ornstein-Ulhenbeck models, Polya tree processes, and other smoothing approaches could be used. We have considered only a single covariate in our model; however our model could easily be extended to multiple covariates, which would be more applicable in real life situations. We also chose the number of knots in our model based on the approach proposed by Ruppert (2002) which is not the only possibility. One could also use the fence method introduced by Jiang et al. (2008) and Jiang, Nguyen & Rao (2010) to determine the number of spline knots L and the degree of spline p. Alternatively, one could use a Bayesian framework through the Reversible Jump MCMC scheme (Green 1995). Our semi-parametric area-level time-series model could also be extended to a semi-parametric unit-level time-series model which might be suitable for some applications. Our univariate model could also be extended to a multivariate version to investigate the multiple responses including mixed responses/outcomes (continuous and discrete). We plan to study these approaches in the future.



(a) Females (b) Males

Figure 5. Ninety-fifty% prediction intervals of rates of physician TRM visit for 217 RHADs (females and males) in 2010 in Manitoba. The predictions were made using the P-spline logistic mixed model. The bullets represent point predictions of rate; the error bars constitute the corresponding prediction intervals.

References

ALEXANDER, L.V., ZHANG, X., PETERSON, T.C., CAESAR, J., GLEASON, B., KLEIN TANK, A.M.G. et al. (2006). Global observed changes in daily climate extremes of temperature and precipitation. *Journal of Geophysical Research: Atmospheres* **111**.

BATTESE, G., HARTER, R. & FULLER, W. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association* **83**, 28 – 36.

BOUBETA, M., LOMBARDIA, M.J. & MORALES, D. (2017). Poisson mixed models for studying the poverty in small areas. *Computational Statistics & Data Analysis* **107**, 32 – 47.

BREIDT, F., CLAESKENS, G. & OPSOMER, J. (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika* **92**, 831 – 846.

CHAMBERS, R. & TZAVIDIS, N. (2006). M-quantile models for small area estimation. *Biometrika* **93**, 255 – 268.

DATTA, G., LAHIRI, P., MAITI, T. & LU, K. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association* **94**, 1074 – 1082.

DATTA, G.S., LAHIRI, P. & MAITI, T. (2002). Empirical Bayes estimation of median income of four-person families by state using time series and cross-sectional data. *Journal of Statistical Planning and Inference* **102**, 83 – 97.

DE BOOR, C. (2001). *A practical Guide to Splines*. Berlin: Springer-Verlag, 2nd edn.

EILERS, P.H. & MARX, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical science* **11**, 89 – 121.

FAY, I.R.E. & HERRIOT, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association* **74**, 269 – 277.

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

Hamilton, J.D. (1986). A standard error for the estimated state vector of a state-space model. *Journal of Econometrics* **33**, 387 – 397.

Jiang, J. & Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test* **15**, 1 – 96.

Jiang, J., Nguyen, T. & Rao, J.S. (2010). Fence method for nonparametric small area estimation. *Survey Methodology* **36**, 3 – 11.

Jiang, J., Rao, J.S., Gu, Z. & Nguyen, T. (2008). Fence methods for mixed model selection. *The Annals of Statistics* **36**, 1669 – 1692.

Jones, R.H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics* **22**, 389 – 395.

Lele, S.R., Dennis, B. & Lutscher, F. (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecology Letters* **10**, 551 – 563.

Lele, S.R., Nadeem, K. & Schmuland, B. (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association* **105**, 1617–1625.

Mekis, E. & Hogg, W.D. (1999). Rehabilitation and analysis of Canadian daily precipitation time series. *Atmosphere-Ocean* **37**, 53 – 85.

Mekis, É. & Vincent, L.A. (2011). An overview of the second generation adjusted daily precipitation dataset for trend analysis in Canada. *Atmosphere-Ocean* **49**, 163 – 177.

Opsomer, J., Claeskens, G., Ranalli, M. & Kauermann, G. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society: Series B* **70**, 265 – 286.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science* **28**, 40 – 68.

Pfeffermann, D. & Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology* **16**, 217 – 237.

Pratesi, M., Ranalli, M.G. & Salvati, N. (2008). Semiparametric M-quantile regression for estimating the proportion of acidic lakes in 8-digit HUCs of the northeastern US. *Environmetrics* **19**, 687 – 701.

Pratesi, M., Ranalli, M.G. & Salvati, N. (2009). Nonparametric M-quantile regression using penalised splines. *Journal of Nonparametric Statistics* **21**, 287 – 304.

R Development Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rao, J. & Molina, I. (2015). *Small Area Estimation*. New Jersey: John Wiley & Sons, Inc, 2nd edn.

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

Rao, J.N.K. & Yu, M. (1994). Small-area estimation by combining time-series and cross-sectional data. *The Canadian Journal of Statistics* **22**, 511 – 528.

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

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

Salvati, N., Ranalli, M.G. & Pratesi, M. (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**, 945 – 964.

Scott, A.J. & Smith, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association* **69**, 674 – 678.

SINGH, A., MANTEL, H. & THOMAS, B. (1991). Time series generalizations of Fay-Herriot estimator for small areas. In *Proceedings of Survey Research Methods Section*. American Statistical Association, pp. 455 – 460.

SMITH, B.J. (2007). BOA: an R package for MCMC output convergence assessment and posterior inference. *Journal of Statistical Software* **21**, 1 – 37.

SÓLYMOS, P. (2010). dclone: data cloning in R. *The R Journal* **2**, 29 – 37. URL http://journal.r-project.org/.

SPERLICH, S. & JOSÉ LOMBARDÍA, M. (2010). Local polynomial inference for small area statistics: estimation, validation and prediction. *Journal of Nonparametric Statistics* **22**, 633 – 648.

TILLER, R.B. (1992). Time series modelling of sample survey data from the U.S. current population survey. *Journal of Official Statistics* **8**, 149 – 166.

TORABI, M. (2012). Hierarchical Bayes estimation of spatial statistics for rates. *Journal of Statistical Planning and Inference* **142**, 358 – 365.

TORABI, M. & SHOKOOHI, F. (2012). Likelihood inference in small area estimation by combining time-series and cross-sectional data. *Journal of Multivariate Analysis* **111**, 213 – 221.

TORABI, M. & SHOKOOHI, F. (2015). Non-parametric generalized linear mixed models in small area estimation. *The Canadian Journal of Statistics* **43**, 82 – 96.

VINCENT, L.A. & MEKIS, É. (2006). Changes in daily and extreme temperature and precipitation indices for canada over the twentieth century. *Atmosphere-Ocean* **44**, 177 – 193.

VINCENT, L.A., WANG, X.L., MILEWSKA, E.J., WAN, H., YANG, F. & SWAIL, V. (2012). A second generation of homogenized canadian monthly surface air temperature for climate trend analysis. *Journal of Geophysical Research: Atmospheres* **117**. doi:10.1029/2012JD017859.

ZHANG, X., VINCENT, L.A., HOGG, W.D. & NIITSOO, A. (2000). Temperature and precipitation trends in canada during the 20th century. *Atmosphere-Ocean* **38**, 395 – 429.