Biostatistics (2015), *pp.* 1–15
doi:10.1093/biostatistics/kxu062

Bayesian two-part spatial models for semicontinuous data with application to emergency department expenditures

BRIAN NEELON*

Department of Public Health Sciences, Medical University of South Carolina, 135 Cannon Street Suite 303, MSC 835, Charleston, SC 29425, USA

neelon@musc.edu

LI ZHU

Department of Biostatistics, University of Pittsburgh, 130 De Soto Street, Pittsburgh, PA 15261, USA

SARA E. BENJAMIN NEELON

Department of Community and Family Medicine, Duke University School of Medicine, 2200 W. Main Street, Durham, NC 27705, USA

Summary

In health services research, it is common to encounter semicontinuous data characterized by a point mass at zero and a continuous distribution of positive values. Examples include medical expenditures, in which the zeros represent patients who do not use health services, while the continuous distribution describes the level of expenditures among users. Semicontinuous data are customarily analyzed using two-part mixture models. In the spatial analysis of semicontinuous data, two-part models are especially appealing because they provide a joint picture of how health services utilization and associated expenditures vary across geographic regions. However, when applying these models, careful attention must be paid to distributional choices, as model misspecification can lead to biased and imprecise inferences. This paper introduces a broad class of Bayesian two-part models for the spatial analysis of semicontinuous data. Specific models considered include two-part lognormal, log skew-elliptical, and Bayesian non-parametric models. Multivariate conditionally autoregressive priors are used to link model components and provide spatial smoothing across neighboring regions, resulting in a joint spatial modeling framework for health utilization and expenditures. We develop a fully conjugate Gibbs sampling scheme, leading to efficient posterior computation. We illustrate the approach using data from a recent study of emergency department expenditures.

Keywords: Bayesian non-parametrics; Dirichlet process mixtures; Semicontinuous data; Skew-elliptical distributions; Spatial data analysis; Two-part model.

*To whom correspondence should be addressed.

1. INTRODUCTION

Emergency department (ED) medical expenditures have been rising steadily in the United States for more than a decade, due in large part to non-urgent ED use. This trend was underscored in a recent study from the National Priorities Partnership (2010), which estimated that inappropriate use of EDs accounted for $38 billion per year in unnecessary expenses. It is also well-established that ED utilization and expenditures vary spatially, even across relatively small geographic regions. A 2013 report by the Institute of Medicine reviewed geographic variation in health expenditures (Newhouse *and others*, 2013), and found disparities even at a local level, suggesting the need for a spatial analysis of health expenditures at a refined geographic scale. Such efforts will enable health officials to effectively allocate resources to stem rising costs.

Using data from the Duke Support Repository (DSR), a rich georeferenced database containing health and financial data for Duke Health System patients residing in Durham County, North Carolina, we describe Bayesian modeling strategies for the spatial analysis of semicontinuous expenditure data characterized by a point mass at zero followed by a continuous distribution with positive support. Semicontinuous data are customarily analyzed using two-part mixture models consisting of a binary component that models the probability of a positive response (in our case, health services use) and a continuous component that models the response distribution among users. In the spatial analysis of health expenditures, two-part models are especially appealing because they allow investigators to characterize spatial regions in terms of both the proportion of ED users and an expenditure distribution conditional on use.

Recently, a great deal of attention has been devoted to selecting appropriate distributions for the continuous part of the data, as model misspecification can lead to incorrect inferences. Recent developments include two-part lognormal (LN) models, two-part log skew-normal (LSN) models, and two-part generalized gamma (GG) models (Liu *and others*, 2012). However, these approaches have typically relied on maximum likelihood for parameter estimation, which can be challenging when fitting the types of spatial models under consideration here.

Our aim, then, is to develop a broad class of Bayesian two-part models for the spatial analysis of areal-referenced, semicontinuous data. To guard against model misspecification, we consider a range of parametric and non-parametric two-part models, including two-part LN, log skew-elliptical, and Dirichlet process mixture models. The proposed models incorporate both patient- and region-level predictors, as well as spatially correlated random effects for each model component. Multivariate conditionally autoregressive priors provide spatial smoothing and link the model components, which has been shown to improve inferences (Su *and others*, 2009). To facilitate Bayesian computation, we develop an efficient Gibbs sampling algorithm based solely on conjugate full-conditional distributions. Together, these features provide a flexible and computationally tractable framework for the analysis of spatially referenced semicontinuous data.

2. TWO-PART SPATIAL MODELS

Let Y_{ij} denote the ED expenditures for the jth patient ($j = 1, \ldots, n_i$) in the ith spatial unit ($i = 1, \ldots, n$). The generic form of the proposed two-part model is given by

$$f(y_{ij}) \stackrel{d}{=} (1 - \pi_{ij})\mathbb{1}_{(y_{ij}=0)} + \pi_{ij} g(y_{ij}|y_{ij} > 0; \mu_{ij}, \sigma, \alpha)\mathbb{1}_{(y_{ij}>0)}, \tag{2.1}$$

where $\pi_{ij} = \Pr(Y_{ij} > 0)$ is the probability of a non-zero expenditure; $\mathbb{1}_{(\cdot)}$ denotes the indicator function; and $g(y_{ij}|y_{ij} > 0; \mu_{ij}, \sigma, \alpha)$ is a positively valued density for the non-zero ED expenditures indexed by location ($\mu_{ij} \in \Re$), scale ($\sigma > 0$), and skewness ($\alpha \in \Re$) parameters. Given the importance of selecting an appropriate functional form for $g(\cdot)$, we consider four distinct two-part models based on different distributional assumptions. In each case, we model the probability parameter, π_{ij}, and the location parameter, μ_{ij}, as functions of patient-level covariates, region-level covariates, and spatially correlated random effects.

2.1 *Two-part LN model*

A common choice for $g(\cdot)$ is the LN density, giving rise to the two-part LN model:

$$\begin{aligned} f(y_{ij}) &\stackrel{d}{=} (1-\pi_{ij})\mathbb{1}_{(y_{ij}=0)} + \pi_{ij}\,\mathrm{LN}(y_{ij}|y_{ij}>0;\,\mu_{ij},\sigma^2)\mathbb{1}_{(y_{ij}>0)},\\ h(\pi_{ij}) &= \boldsymbol{x}_{ij}'\boldsymbol{\gamma} + s_1(v_{ij}) + b_{1i} \quad \text{(binary part)},\\ \mu_{ij} &= \boldsymbol{x}_{ij}'\boldsymbol{\beta} + s_2(v_{ij}) + b_{2i} \quad \text{(continuous part)}, \end{aligned} \tag{2.2}$$

where μ_{ij} and σ^2 are the mean and variance of Y_{ij} (>0) on the log scale; $h(\cdot)$ is a binary link function; $\boldsymbol{x}_{ij}$ is a vector of patient- and region-level predictors (e.g. insurance status and median household income); $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are vectors of fixed effect regression parameters; $s_1(v_{ij})$ and $s_2(v_{ij})$ are smooth, non-linear functions of a patient-level continuous covariate v_{ij} (e.g. age) to be modeled via splines; and b_{1i} and b_{2i} are region-level random intercepts assumed to follow a bivariate, spatially correlated prior distribution. Under the LN model, the skewness parameter, α, is assumed to be zero. Throughout, we assume identical predictors for the two model components, but in general this is not necessary.

The LN distribution is appealing because it leads to straightforward Bayesian computation and easily interpreted parameters. However, by imposing both unimodality and symmetry about μ_{ij} on the log scale, the LN is somewhat restrictive, as these conditions are commonly violated in practice. Below we consider generalizations of the two-part LN model. In each case, the structure of the binary part is identical to that given in (2.2). We therefore focus on various models for the continuous component.

2.2 *Two-part LSN model*

A more flexible choice for $g(\cdot)$ is a logged version of the skew-normal distribution (Azzalini, 1985), which maintains unimodality but permits non-zero skewness on the log scale. The LSN density is given by

$$g(y_{ij}|y_{ij}>0;\,\mu_{ij},\sigma,\alpha) \stackrel{d}{=} \frac{2}{y_{ij}\sigma}\phi\left(\frac{\ln y_{ij}-\mu_{ij}}{\sigma}\right)\Phi\{\alpha\sigma^{-1}(\ln y_{ij}-\mu_{ij})\}, \tag{2.3}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and CDF of a standard normal random variable. Setting $\alpha=0$ yields the LN density. Following Frühwirth-Schnatter and Pyne (2010), we can aid inference by working with the following stochastic representation of the LSN:

$$\ln Y_{ij}|Y_{ij}>0 = \mu_{ij} + \psi Z_{ij} + e_{ij}, \quad e_{ij}\sim N(0,\xi^2), \tag{2.4}$$

where $\psi=\sigma\alpha/\sqrt{1+\alpha^2}$, Z_{ij} is a standard truncated normal $\mathrm{TN}_{[0,\infty)}(0,1)$ random variable that is stochastically independent of e_{ij}, $\xi^2=\sigma^2/(1+\alpha^2)$, and μ_{ij} is modeled as in (2.2). As discussed below in Section 3, this stochastic representation admits conjugate full-conditional distributions for all model parameters, leading to straightforward Bayesian computation.

2.3 *Two-part log skew-t model*

The skew-t distribution (Azzalini and Capitanio, 2003) accommodates both skewness and heavier tails relative the skew normal. The pdf for the logged version of the skew-t is given by

$$g(y_{ij}|y_{ij}>0;\,\mu_{ij},\sigma,\alpha,\nu) \stackrel{d}{=} \frac{2}{y_{ij}\sigma}t_\nu(d_{ij})T_{\nu+1}\left(\alpha d_{ij}\sqrt{\frac{\nu+1}{\nu+d_{ij}^2}}\right), \tag{2.5}$$

where $d_{ij} = (\ln y_{ij} - \mu_{ij})/\sigma$ and t_ν and T_ν denote the pdf and CDF of a standard Student-t random variable with $\nu > 0$ degrees of freedom (df). As $\nu \to \infty$, the log skew-t (LST) distribution converges to the LSN. For Bayesian inference, it is preferable to work with the following stochastic representation of the LST (Frühwirth-Schnatter and Pyne, 2010):

$$\begin{aligned} \ln Y_{ij}|Y_{ij} > 0 &= \mu_{ij} + \psi Z_{ij} + e_{ij}, \quad e_{ij}|W_{ij} \sim N(0, \xi^2/W_{ij}), \\ Z_{ij}|W_{ij} &\sim \text{TN}_{[0,\infty)}(0, 1/W_{ij}), \\ W_{ij} &\sim \text{Ga}\left(\frac{\nu}{2}, \frac{\nu}{2}\right), \end{aligned} \tag{2.6}$$

where $\text{Ga}(\cdot)$ denotes a Gamma distribution, ψ and ξ^2 are defined as in (2.4), and ν is fixed at a user-chosen value. This representation again yields closed-form full conditionals.

2.4 *DP mixture of two-part LN models*

Lastly, we consider a Dirichlet process mixture of two-part LN models (DPLN; Antoniak, 1974). We begin by assuming the following two-part model:

$$\begin{aligned} f(y_{ij}) &\overset{d}{=} (1 - \pi_{ij})\mathbb{1}_{(y_{ij}=0)} + \pi_{ij}\,\text{LN}(y_{ij}|y_{ij} > 0; \mu_{ij}, \sigma^2_{ij})\mathbb{1}_{(y_{ij}>0)}, \\ h(\pi_{ij}) &= \boldsymbol{x}'_{ij}\boldsymbol{\gamma}_{ij} + s_{1ij}(v_{ij}) + b_{1ij}, \\ \mu_{ij} &= \boldsymbol{x}'_{ij}\boldsymbol{\beta}_{ij} + s_{2ij}(v_{ij}) + b_{2ij}. \end{aligned} \tag{2.7}$$

Let $\boldsymbol{\theta}_{ij}$ denote the collection of model parameters for the (ij)th subject. We place a Dirichlet process prior on $\boldsymbol{\theta}_{ij}$:

$$\boldsymbol{\theta}_{ij}|G \sim G, \tag{2.8}$$

$$G|\omega, G_0 \sim \text{DP}(\omega, G_0), \tag{2.9}$$

where G_0 is a joint base distribution and $\omega > 0$ is a concentration parameter. Next, we assume a stick-breaking representation for G (Sethuraman, 1994):

$$\begin{aligned} G &= \sum_{k=1}^{\infty} p_k \delta_{\boldsymbol{\theta}^*_k} \quad \text{where } \delta_{\boldsymbol{\theta}^*_k} \text{ is a point mass at } \boldsymbol{\theta}^*_k, \\ \boldsymbol{\theta}^*_k &\overset{\text{iid}}{\sim} G_0, \\ p_k &= \upsilon_k \prod_{l=1}^{k-1}(1 - \upsilon_l), \quad \text{and} \\ \upsilon_k|\omega &\sim \text{Be}(1, \omega), \end{aligned}$$

where $\text{Be}(\cdot)$ denotes a beta distribution and $\omega > 0$ is fixed at a user-specified value, with smaller values leading to greater sparseness. In practice, we truncate the infinite mixture by a finite upper bound $K \ll N$, where $N = \sum_i n_i$ denotes the total sample size. Let $\{\boldsymbol{\theta}^*_1, \ldots, \boldsymbol{\theta}^*_K\}$ denote the K potential values for $\boldsymbol{\theta}_{ij}$, and let $C_{ij} \in \{1, \ldots, K\}$ be a latent component indexing variable for the (ij)th subject taking the value k if $\boldsymbol{\theta}_{ij} = \boldsymbol{\theta}^*_k$. Then model (2.7) can be recast as a finite DP mixture of two-part LN models with mixing weights, p_k, defined by a stick-breaking process. That is, given $C_{ij} = k$, the distribution of Y_{ij} is given by

(2.2) but now with component-specific parameters:

$$\begin{aligned} f(y_{ij}|C_{ij}=k) &= (1-\pi_{ijk})\mathbb{1}_{(y_{ij}=0)} + \pi_{ijk}\,\mathrm{LN}(y_{ij}|y_{ij}>0;\,\mu_{ijk},\sigma_k^2)\mathbb{1}_{(y_{ij}>0)},\\ h(\pi_{ijk}) &= \boldsymbol{x}_{ij}'\boldsymbol{\gamma}_k + s_{1k}(v_{ij}) + b_{1ik},\\ \mu_{ijk} &= \boldsymbol{x}_{ij}'\boldsymbol{\beta}_k + s_{2k}(v_{ij}) + b_{2ik}, \quad k=1,\ldots,K. \end{aligned} \tag{2.10}$$

As with the previous models, fully conjugate updates are available for the DPLN.

By accommodating both skewness and multimodality on the log scale, the DPLN offers added flexibility over the log skew-elliptical models. In fact, the DPLN relaxes the parametric assumptions for both the continuous and binary parts by establishing a flexible link function that relates predictors to the probability of non-zero expenditures. Specifically, recall that a binary link function can be represented in terms of a latent density, with a normal density corresponding to a probit link, a logistic density representing a logit link, and a scale-mixture of normals defining a class of t-link functions (Albert and Chib, 1993). Because we can represent any latent density as a DP mixture, the DPLN implicitly establishes an arbitrary class of link functions for the binary portion of the two-part model. By accommodating a broader class of link functions, the DPLN improves our capacity to predict probabilities that are close to 0 or 1.

Moreover, by incorporating spatial random effects into each mixture component, the DPLN can generate flexible region-specific density estimates of patient expenditures, which can in turn be used to summarize the unique expenditure patterns for each spatial unit. While this is true of the parametric models as well, those models allow only for location shifts on the log scale, and do not permit spatial variation in the shape of the density of $\ln Y|Y>0$. By introducing spatial heterogeneity into each mixture component, the DPLN offers greater flexibility in which the shape of the density of $\ln Y|Y>0$ is allowed to vary spatially. This added flexibility is carried over to the original (unlogged) scale, and can be used to aid quantile-based inferences. For example, after fitting the DPLN, one can derive region-specific quantiles of interest and use these to examine spatial patterns in the tails of the expenditure distribution. This can prove especially useful in the analysis of expenditure data, where there is often little spatial variation in mean expenditures but pronounced variation in the upper tails—for example, among high-expenditure patients who place the greatest burden on the health system.

3. Bayesian inference

We adopt a Bayesian inferential approach and assign prior distributions to all model parameters. In order to promote smoothing of sparsely populated regions, we assign a bivariate intrinsic conditionally autoregressive (BICAR) prior to the spatial random effects $\boldsymbol{b}_i = (b_{1i}, b_{2i})'$:

$$\boldsymbol{b}_i|\boldsymbol{b}_{-i}, \boldsymbol{\Lambda} \sim N_2\left(\frac{1}{m_i}\sum_{l\in\partial_i}\boldsymbol{b}_l, \frac{1}{m_i}\boldsymbol{\Lambda}\right), \tag{3.1}$$

where m_i is the number of neighbors of ith spatial unit, the ∂_i is the set of neighbors for the ith spatial unit, and $\boldsymbol{\Lambda} = \begin{pmatrix} \Lambda_{11} & \rho\sqrt{\Lambda_{11}\Lambda_{22}} \\ \rho\sqrt{\Lambda_{11}\Lambda_{22}} & \Lambda_{22} \end{pmatrix}$ is the variance–covariance matrix for $\boldsymbol{b}_i$ conditional on the remaining spatial random effects, $\boldsymbol{b}_{-i}$. The BICAR prior is appealing because it links the binary and continuous parts of the model through the correlation parameter ρ, which can improve inferences when the binary and continuous processes are related (Su *and others*, 2009). In our application, for instance, regions with a high proportion of ED users may have high expenditures given ED use; it is therefore desirable to build this association into the model.

Following Brook's Lemma (cf. Banerjee *and others*, 2004), the joint prior of the $2n \times 1$ vector $\boldsymbol{b} = \{(b_{11}, \ldots, b_{1n})', (b_{21}, \ldots, b_{2n})'\}'$ is given by

$$\boldsymbol{b}|\boldsymbol{\Lambda} \propto \exp\{-\tfrac{1}{2}\boldsymbol{b}'(\boldsymbol{\Lambda}^{-1} \otimes \boldsymbol{Q})\boldsymbol{b}\}, \tag{3.2}$$

where $\boldsymbol{Q} = \boldsymbol{M} - \boldsymbol{A}$; $\boldsymbol{M} = \text{diag}(m_1, \ldots, m_n)$; and $\boldsymbol{A}$ is an $n \times n$ adjacency matrix with $a_{ii} = 0$, $a_{il} = 1$ if spatial units i and l are neighbors, and $a_{il} = 0$ otherwise. Because $\boldsymbol{Q}$ is singular, the joint prior distribution in (3.2) is over-parameterized and thus improper, although the posterior of $\boldsymbol{b}$ will remain proper under the proposed models. If a fixed intercept is included in the model, a sum-to-zero constraint must be applied to ensure an identifiable model.

From prior (3.2), we can derive conditional priors for $\boldsymbol{b}_1 = (b_{11}, \ldots, b_{1n})'$ and $\boldsymbol{b}_2 = (b_{21}, \ldots, b_{2n})'$:

$$\boldsymbol{b}_1|\boldsymbol{b}_2 \sim N_n\left(\rho\sqrt{\frac{\Lambda_{11}}{\Lambda_{22}}}\boldsymbol{b}_2, \quad (1-\rho^2)\Lambda_{11}\boldsymbol{Q}^+\right) \tag{3.3}$$

and

$$\boldsymbol{b}_2|\boldsymbol{b}_1 \sim N_n\left(\rho\sqrt{\frac{\Lambda_{22}}{\Lambda_{11}}}\boldsymbol{b}_1, \quad (1-\rho^2)\Lambda_{22}\boldsymbol{Q}^+\right), \tag{3.4}$$

where $\boldsymbol{Q}^+$ is a generalized inverse of the rank-deficient matrix $\boldsymbol{Q}$. This conditional prior specification leads to efficient Gibbs updates for the spatial effects. As detailed in supplementary material available at *Biostatistics* online, the update for $\boldsymbol{b}_1$ depends on likelihood contributions from all N observations, while the update for $\boldsymbol{b}_2$ relies only on contributions from the $N_1 < N$ positive observations. This sample-size imbalance prevents a joint Gibbs update for $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ based on prior (3.2). The conditional priors (3.3) and (3.4) avoid this problem by allowing separate univariate updates for $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$, the former based on all N observations and the latter on the N_1 positive observations.

For the two-part LN model, we complete the prior specification by assuming independent $N(0, 100)$ priors for the fixed effects and spline coefficients, an inverse-Gamma IG(0.01, 0.01) prior for σ^2, and an inverse-Wishart IW(3, $\boldsymbol{I}_2$) prior for the spatial covariance matrix, $\boldsymbol{\Lambda}$. For the LSN and LST models, we additionally assign an $N(0, 100)$ prior to ψ and an IG(0.01, 0.01) prior to ξ^2. For the DPLN, we assign component-specific priors analogous to those for the LN model. Additionally, we assume $\upsilon_k|\omega \sim \text{Be}(1, \omega)$. For the ED expenditure analysis, we used a selection criterion to choose between $\omega = 1$ and $\omega = 5$, the former yielding a sparser mixture.

For posterior computation, we propose Gibbs sampling schemes based exclusively on full-conditional updates. Assuming a probit link, Gibbs sampling for the binary portion of the LN, LSN, and LST two-part models proceeds via data augmentation by first introducing a unit-variance latent normal variable, U_{ij}, such that $Y_{ij} > 0$ if $U_{ij} > 0$ and $Y_{ij} = 0$ if $U_{ij} < 0$ (Albert and Chib, 1993). Next, after initializing parameters, we iterate through following steps:

1. For each i and j, sample u_{ij} from its full-conditional normal distribution truncated below (above) by zero for $y_{ij} > 0$ ($y_{ij} = 0$).

2. Draw $\boldsymbol{\gamma}$ and the spline coefficients associated with $s_1(v_{ij})$ jointly according to a multivariate normal full conditional.
3. Assuming prior specification (3.3), draw $\boldsymbol{b}_1$ from its multivariate normal full conditional.

To update the LN parameters in model (2.2), we iterate through the following Gibbs steps:

1. Draw $\boldsymbol{\beta}$ and the spline coefficients associated with $s_2(v_{ij})$ jointly according to a multivariate normal full conditional.
2. Sample σ^2 from its inverse-Gamma full conditional.
3. Assuming prior specification (3.4), draw $\boldsymbol{b}_2$ from its multivariate normal full conditional.
4. Given $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$, update $\boldsymbol{\Lambda}$ from its inverse-Wishart full conditional.
5. Apply sum-to-zero constraints to $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ to ensure identifiability.

For the two-part LSN model, the updates of the binary part are similar to those given above. For the continuous part, we follow the stochastic representation given in (2.4) and implement the following Gibbs sampler:

1. For all $y_{ij} > 0$, draw the latent variable z_{ij} from its truncated normal full conditional.
2. Draw $\boldsymbol{\beta}$, ψ, and the spline coefficients associated with $s_2(v_{ij})$ jointly according to a multivariate normal full conditional.
3. Draw ξ^2 according to its inverse-Gamma full conditional.
4. Assuming prior specification (3.4), draw $\boldsymbol{b}_2$ from its multivariate normal full conditional.
5. Draw $\boldsymbol{\Lambda}$ from its inverse-Wishart full conditional and apply sum-to-zero constraints.

The LST follows a similar sampling scheme, but with an additional Gamma update for $\{W_{ij}\}$.

For the DPLN, Gibbs sampling proceeds by introducing for each subject a latent component indexing variable, C_{ij}. Next, after truncating the infinite stick-breaking summation by a finite upper bound K, we iterate through the following steps:

1. For all i and j, sample C_{ij} from its closed-form multinomial full conditional.
2. For $k = 1, \ldots, K$, draw the stick-breaking weight, υ_k, from its beta full conditional.
3. Conditional on C_{ij}, update u_{ij} from its truncated normal full conditional.
4. For $k = 1, \ldots, K$, draw the component-specific parameters $\boldsymbol{\gamma}_k$, $\boldsymbol{\beta}_k$, $\boldsymbol{b}_{1k}$, $\boldsymbol{b}_{2k}$, $\boldsymbol{\Lambda}_k$, σ_k^2, and spline coefficients according to their full conditional distributions. These updates are analogous to those for two-part LN model above, but now with component-specific likelihoods.

Full conditionals for the above sampling schemes are provided in supplementary material available at *Biostatistics* online.

Posterior convergence is monitored by inspecting trace plots and computing standard Markov chain Monte Carlo (MCMC) diagnostics such as Geweke's z-score (Geweke, 1992). For model comparison, we adopt the deviance information criterion (DIC; Spiegelhalter *and others*, 2002), which combines a measure of current model fit ($\bar{D}$) with a penalty for model complexity (p_D). For the DPLN, we adopt a modified version, termed DIC_3, proposed by (Celeux *and others*, 2006).

A computational challenge for Bayesian mixture models is "label switching", in which draws of component-specific parameters may be associated with different components during the MCMC run. This does not pose a major concern for the DPLN, since our interest lies in marginal rather than within-component inferences. However, label switching should be addressed prior to assessing posterior convergence; here, we adopt the relabeling algorithm proposed by Stephens (2000).

4. ILLUSTRATIVE EXAMPLE

To evaluate model performance, we conducted a small simulation study. To emulate the DSR data, we generated spatial data according to the Durham County adjacency matrix, which comprises 153 Census block groups. Next, for each block group, we simulated 100 responses according to a two-component finite mixture of two-part GG models. We adopted the parameterization of the GG given by Manning *and others* (2005), which takes the form

$$f(y;\mu,\sigma,\zeta)=\frac{\eta^{\eta}}{\sigma y\sqrt{\eta}\Gamma(\eta)}\exp(z\sqrt{\eta}-u),$$

where $z=\text{sign}(\zeta)\{\ln y-\mu\}/\sigma$, $\sigma>0$ is a scale parameter, μ is a location parameter, $\eta=|\zeta|^{-2}$ is a shape parameter, and $u=\eta\exp(|\zeta|z)$. The linear predictors included two covariates as well as spatial random effects, resulting in the following two-part mixture model:

$$\begin{aligned}f(y_{ij}|C_{ij}=k)&=(1-\pi_{ijk})\mathbb{1}_{(y_{ij}=0)}+\pi_{ijk}\,\text{GG}(y_{ij}|y_{ij}>0;\mu_{ijk},\sigma_k,\zeta_k)\mathbb{1}_{(y_{ij}>0)},\\ \Phi^{-1}(\pi_{ijk})&=\gamma_{0k}+\gamma_{1k}x_{1ij}+\gamma_{2k}x_{2ij}+b_{1ik},\\ \mu_{ijk}&=\beta_{0k}+\beta_{1k}x_{1ij}+\beta_{2k}x_{2ij}+b_{2ik},\quad i=1,\ldots,153,\quad j=1,\ldots,100,\quad k=1,2,\end{aligned}$$

where C_{ij} is the component index, and x_{1ij} and x_{2ij} are from Uniform(−1,1) and Bernoulli(0.5) distributions, respectively. For component 1, we set $\boldsymbol{\gamma}_1'=(\gamma_{01},\gamma_{11},\gamma_{21})=(0.1,0.2,-0.2)$, $\boldsymbol{\beta}_1'=(\beta_{01},\beta_{11},\beta_{21})=(0.3,-0.5,0.4)$, $\sigma_1^2=1$, and $\zeta_1=0.40$; for component 2, we set $\boldsymbol{\gamma}_2'=(0.5,0.1,-0.4)$, $\boldsymbol{\beta}_2'=(4,0.5,1.2)$, $\sigma_2^2=0.5$, and $\zeta_2=0.80$. The spatial random effects were generated according to a BICAR with $\boldsymbol{\Lambda}_1=\left(\begin{smallmatrix}1 & 0.50\\ 0.50 & 1\end{smallmatrix}\right)$ and $\boldsymbol{\Lambda}_2=\left(\begin{smallmatrix}1 & 0.20\\ 0.20 & 1\end{smallmatrix}\right)$. These values resulted in a response distribution that varied flexibly in terms of shape and scale across block groups.

For the two-part LN, LSN, and LST models, we ran the MCMC for 5000 iterations, with 2000 iterations as burn-in. For the LST, we considered models with $\nu=3$ and $\nu=16$ df. For the DPLN, we truncated the infinite mixture to $K=25$ components and compared models with concentration parameters $\omega=1$ and $\omega=5$. We ran 15 000 iterations with a burn-in of 10 000 and retained every fifth iteration to reduce

Table 1. *Model comparison statistics for the illustrative example in Section* 4 *and the ED analysis in Section* 5

	Model	$\bar{D}(\theta)$	p_D	DIC_3	DIC
Illustrative example	LN	98 165	240	98 398	98 404
	LSN	95 904	260	96 120	96 165
	LST ($\nu=3$)	96 652	266	96 900	96 918
	LST ($\nu=16$)	96 035	263	96 258	96 298
	DPLN ($\omega=1$, $K=25$)	90 318	463	90 781	—
	DPLN ($\omega=5$, $K=25$)	90 302	473	90 775	—
ED analysis	LN	231 782	172	231 954	231 953
	LSN	231 597	175	231 772	231 770
	LST ($\nu=3$)	232 167	194	232 345	232 361
	LST ($\nu=16$)	231 384	172	231 556	231 556
	DPLN ($\omega=1$, $K=25$)	228 958	791	229 749	—
	DPLN ($\omega=5$, $K=25$)	229 284	710	229 994	—

LN, two-part lognormal model; LSN, two-part log skew-normal model; LST, two-part log skew-t model; DPLN, Dirichlet process mixture of two-part lognormal models.

auto-correlation. Convergence was monitored by trace plots and Geweke statistics. All models were fit in R version 3.0 (R Core Team, 2013).

As Table 1 indicates, the two DPLN models vastly outperformed the other models in terms of DIC_3 fit. Overall, the DPLN with $\omega = 5$ had the lowest DIC_3 value, followed by the DPLN with $\omega = 1$ and LSN models. Apparently, the added flexibility of the DPLN is needed to fully capture the features of the response distribution. Figure 1 presents estimated densities (on the log scale) for the continuous part of the model for the reference group ($x_1 = x_2 = 0$). Figure 1(a) presents the density for an "average" block group

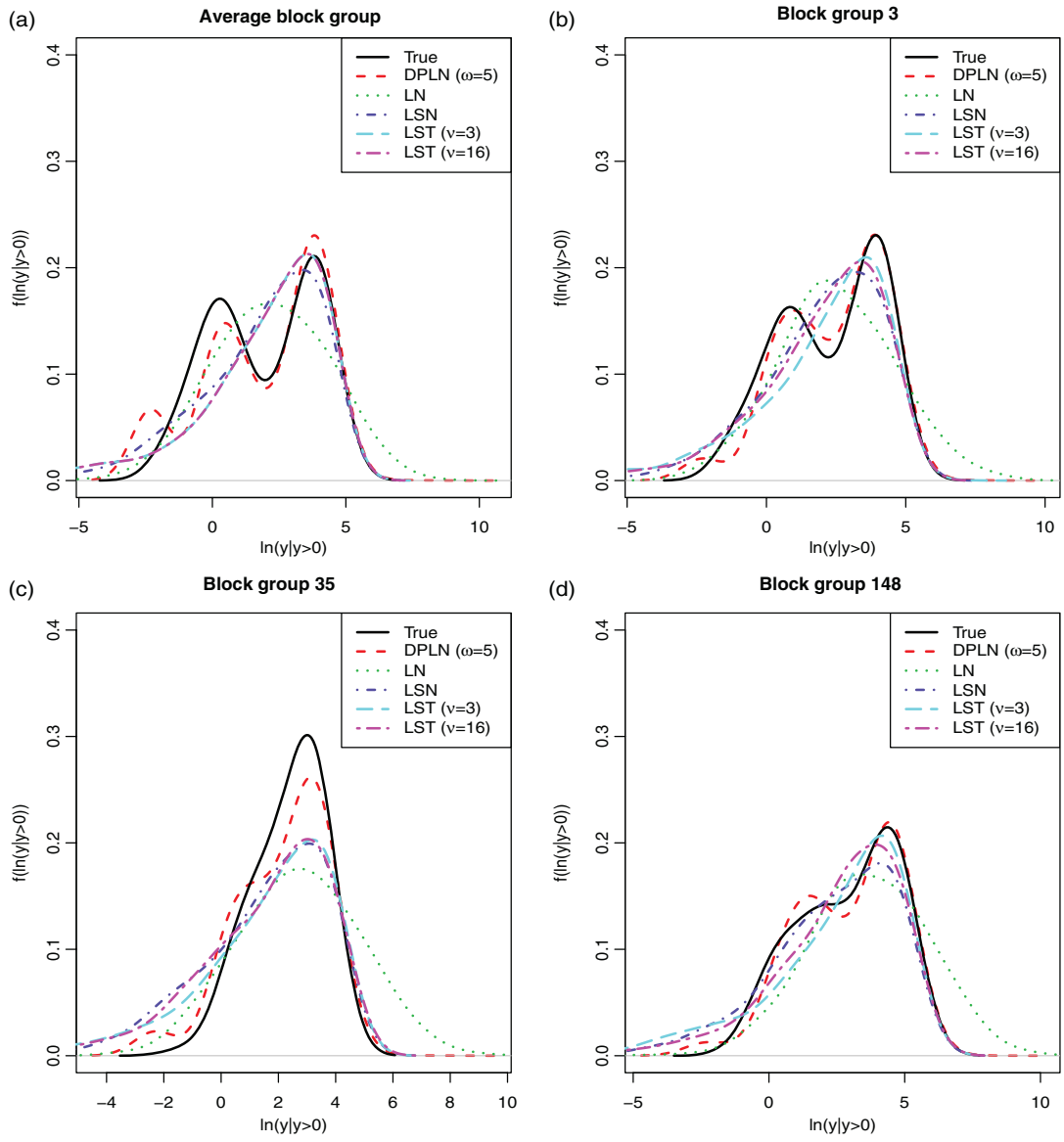

Fig. 1. True and estimated densities of $\ln Y | Y > 0$ for the "reference" group ($x_1 = x_2 = 0$) in the illustrative example. (a) An "average" block group with spatial effects set to zero. (b)–(d) Three selected block groups.

with spatial effects set to zero, and Figure 1(b)–(d) show densities on the log scale for three representative block groups. The DPLN generally provided excellent fit to the true block group densities. More formally, empirical estimates of the Kullback–Leibler divergence indicated that the DPLN was superior to other models in all cases (see Table S1 in supplementary material available at *Biostatistics* online). These results highlight the usefulness of the DPLN for spatial density estimation when the shape of the density varies across spatial units.

Figure 2 maps the true, DPLN-ω_5, and LSN-estimated probabilities of a positive response, expected values among the positive responses, and upper quartiles among the positive responses for the reference group. Also included are the mean squared predictive errors (MSPEs) comparing the two fitted models. The DPLN and LSN were able to replicate the true spatial pattern for $\Pr(Y > 0)$, each with an MSPE of 0.002. However, the DPLN outperformed the LSN for $\mathrm{E}(Y|Y > 0)$ and even more so for the upper quartile, where it led to a nearly 3-fold reduction in MSPE. This suggests that the DPLN can provide

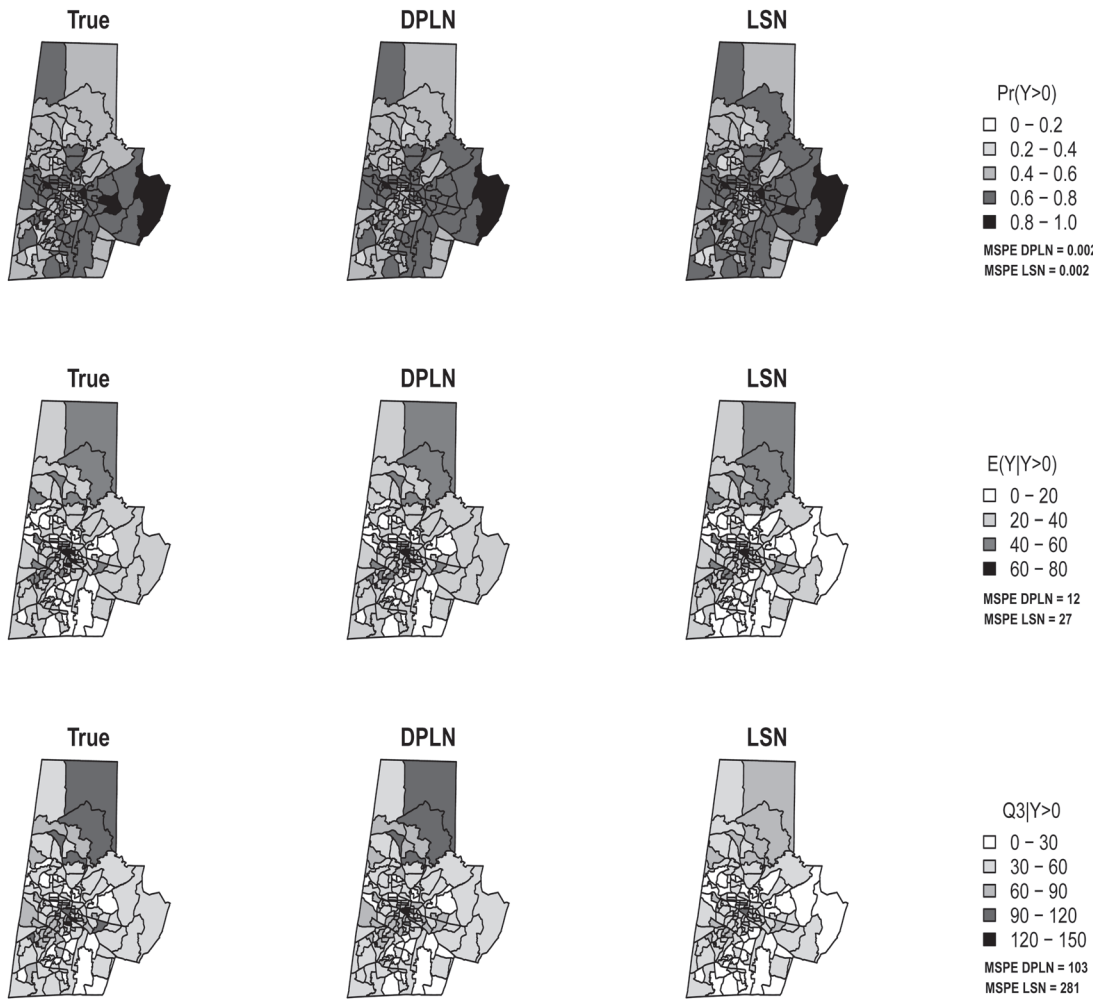

Fig. 2. True, DPLN-ω_5, and LSN-estimated $\Pr(Y > 0)$, $\mathrm{E}(Y|Y > 0)$, and upper quartiles $(\mathrm{Q3}|Y > 0)$ among the positive responses for the reference group $(x_1 = x_2 = 0)$ in the illustrative example. MSPE, mean square predictive error.

substantial improvement in estimating non-central quantiles when standard distributional assumptions such as unimodality fail to hold.

5. Analysis of ED expenditures

We used the proposed two-part models to analyze ED expenditures from the DSR. We extracted records for patients seen at a Duke-affiliated ED or a non-ED clinic during the 2011 calendar year. Patients with at least one ED visit were defined as "ED users" and patients with no ED visits were regarded as "non-ED users". For each ED patient, we identified billing account numbers related exclusively to ED care and summed the corresponding payments to obtain total annual ED expenditures. We further restricted the analysis to non-Hispanic white (NHW), non-Hispanic black (NHB), and Hispanic patients aged 18 or older. The records were georeferenced by residential address and linked at the Census block group level to the 2006–2010 American Community Survey (U.S. Census Bureau, 2010). The final dataset contained 44 722 records and included information on ED use (yes/no); total annual ED expenditures ($\geqslant$0) for 2011; patient demographics, such as age, race, sex, and insurance status; and median household income for each block group. Table S2 in supplementary material available at *Biostatistics* online provides summary statistics for the DSR data. A quarter of the patients were ED users, with a median expenditure of $2347 (IQR = [1256, 4659]). Figure S1 in supplementary material available at *Biostatistics* online presents a map of the Durham city limits and ED locations.

We fit the proposed spatial two-part models to the expenditure data with age, sex, race, insurance, and block group median income as covariates. We modeled age using cubic B-splines with first, second, and third quartiles as interior knots (33, 47, and 61, respectively). We implemented the models using the Gibbs sampling routines described in Section 3. For the two-part LN, LSN, and LST models, we ran 5000 iterations with 2000 as burn-in; for DPLN, we ran 15 000 iterations discarding 10 000 as burn-in, retaining every fifth iteration. Posterior convergence was confirmed by monitoring trace plots and computing Geweke's z-scores.

To select the appropriate model, we first examined the residuals after fitting a spatial LN model to the positive observations. If the LN assumption holds, these residuals should be approximately normal on the log scale for the entire sample as well as for individual block groups, since the LN permits only location shifts on the log scale. As Figure S2 of supplementary material available at *Biostatistics* online demonstrates, there was evidence of skewness and multimodality in the residuals, suggesting the need for a more flexible model. The DIC results further indicated that the DPLN with $\omega = 1$ had optimal fit (Table 1). We therefore present results for this model below.

Table 2 provides estimates of the probability of ED use, $\Pr(Y > 0)$; the conditional mean expenditures among ED users, $\mathrm{E}(Y|Y > 0)$; and the marginal mean expenditures among all patients, $\mathrm{E}(Y)$, for

Table 2. *DPLN-estimated probability of ED use* ($\Pr(Y > 0)$), *mean expenditures among ED users* ($\mathrm{E}(Y|Y > 0)$), *and marginal mean expenditures among ED users and non-users* ($\mathrm{E}(Y)$) *for three patient groups*

Patient group	$\Pr(Y > 0)$	$\mathrm{E}(Y\|Y > 0)$	$\mathrm{E}(Y)$
NHW, female, private insurance	0.05 (0.04, 0.06)	3418 (2975, 3835)	190 (153, 233)
Hispanic, male, uninsured	0.64 (0.60, 0.68)	2889 (2633, 3169)	2176 (1954, 2442)
NHB, female, medicaid	0.31 (0.27, 0.34)	3925 (3069, 4790)	1096 (956, 1253)

All groups are assumed to be of mean age and block group income and to reside in an "average" block group with spatial effects equal to zero. 95% credible intervals are given in parentheses.
NHW, non-Hispanic white race; NHB, non-Hispanic black race.

three patient covariate groups. The first group comprised privately insured NHW females of mean age and household income residing in an "average" block group with spatial effects set to zero. Race, sex, and insurance were allowed to vary for the other two groups. The second group comprised uninsured Hispanic males, and the third group comprised NWB female Medicaid enrollees. As the results indicate, the first group had the lowest probability of ED use; among ED users, however, this group had the second-highest mean expenditures. This suggests that while ED use was relatively rare for this group, ED users in this group tended to receive relatively expensive care. Among ED users and non-users combined, group 2 had the highest marginal mean expenditures, $E(Y)$. This was primarily due to the fact that this group was more likely to use the ED than other groups. Interestingly, group 2 also had the lowest average expenditures among ED users ($E(Y|Y > 0)$). Thus, this group may include patients who rely on the ED for routine but less costly non-urgent care. Finally, group 3 comprised Medicaid enrollees who had the highest mean expenditures among users.

Figure 3 maps the DPLN-estimated $\Pr(Y > 0)$ and $E(Y|Y > 0)$ for patient group 1. The probability of ED use ranged from 0.02 to 0.10 across the block groups, with the southwest tending to have the lowest rates of ED use. The average expenditures among users ranged from $2961 to $4017, again with the southwestern block groups having among the lowest mean expenditures. Interestingly, several block groups in the central part of the county had high rates of ED use but relatively low expenditures, suggesting that this area may comprise a large fraction of patients who use the ED for low-cost routine care. Highlighted on the map are four illustrative block groups representing various use/expenditure patterns; Table S3 of supplementary material available at *Biostatistics* online provides the accompanying expenditure estimates for each block group. Block group 86 in the central east had a comparatively high proportion of ED users and high expenditures among users. In contrast, block group 122 in the southwest had both low ED use and average expenditures among users. The other two block groups represent those with high use, low expenditures (block group 84 in the northeast) and low use, high expenditures (block group 24 in the central west). Block group 84 may comprise routine ED users who rely on the ED for less expensive primary care, and may therefore be amenable to community-based efforts to reduced non-urgent ED use. Figures S3 and S4 in supplementary material available at *Biostatistics* online present analogous maps for patient groups 2 and 3. Note that the spatial patterns vary across the three patient groups, suggesting that there was a spatial interaction among the patient-level covariates race, sex, and insurance.

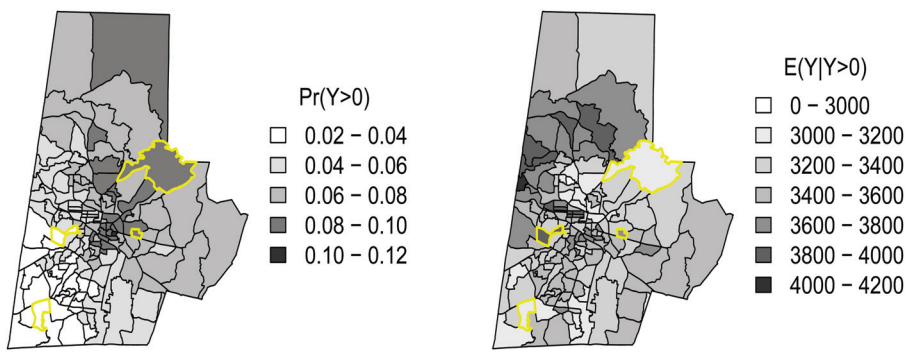

Fig. 3. DPLN-estimated (a) $\Pr(Y > 0)$ and (b) $E(Y|Y > 0)$ for patient group 1 in the ED analysis. Highlighted are four representative block groups with (clockwise from top): high $\Pr(Y > 0)$ and low $E(Y|Y > 0)$ (BG 84); high $\Pr(Y > 0)$, high $E(Y|Y > 0)$ (BG 86); low $\Pr(Y > 0)$, low $E(Y|Y > 0)$ (BG 122); and low $\Pr(Y > 0)$, high $E(Y|Y > 0)$ (BG 24). BG, block group.

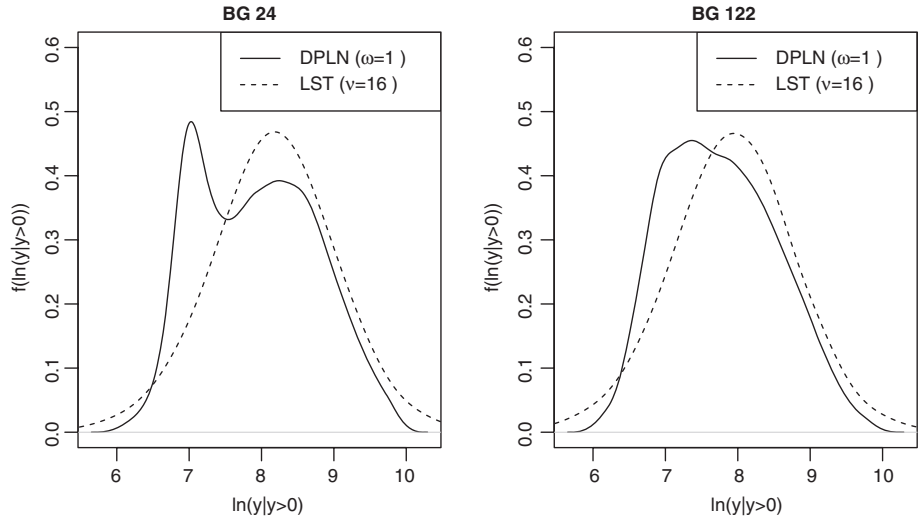

Fig. 4. DPLN- and LST-estimated densities of $\ln Y|Y > 0$ for group 1 patients residing in block groups 24 and 122. DPLN, Dirichlet process mixture of two-part lognormals; LST, log skew-t.

Figure 4 presents the estimated DPLN-ω_1 and 16-df LST densities of the logged expenditures among ED users in patient group 1 who reside in block groups 24 and 122. According to the DPLN, block group 122 consisted of a single, homogenous population of spenders with comparatively low expenditures, whereas block group 24 appeared to comprise two subpopulations: one characterized by low average expenditures and low variability, and a second with high mean expenditures and increased variability. The LST, however, was unable to capture the bimodality identified by the DPLN. We can further summarize the spatial patterns by estimating quantiles of interest. For example, in block group 122, the DPLN-estimated median expenditures among ED users was \$2200 (95% CI = [1760, 2821]) and the 90th quantile was \$6662 (95% CI = [5040, 8642]); for block group 24, the median and 90th quantiles were \$2760 (95% CI = [2160, 3720]) and \$8300 (95% CI = [6320, 10 962]), respectively. Note that the corresponding estimates under the 16-df LST were approximately 25% higher for both block groups (Table S4 of supplementary material available at *Biostatistics* online), indicating that model choice can substantially impacts estimates of expenditure quantiles.

6. Discussion

We have proposed a broad class of Bayesian two-part models for areal-referenced, semicontinuous data. The models have a number of attractive features: (1) they provide a joint spatial assessment of ED use and associated expenditures; (2) they incorporate patient- and Census-level information to better explain spatial trends; (3) their BICAR prior structure links model components and encourages spatial smoothing; and (4) depending on distributional assumptions, they provide a flexible approach to spatial density estimation of semicontinuous data. The models can also be fit within a computationally feasible Bayesian framework.

The choice between models will depend in large part on the aims of the analysis. If the primary aim is to assess fixed covariate effects while treating unobserved spatial heterogeneity (i.e. the spatial random effects) as “nuisance", then a parametric model may be suitable. As a starting point, one can examine the residuals from a spatial LN model fitted to the positive observations. If there is evidence of residual skewness, the two-part LSN or LST may be more desirable. In our experience, one typically needs to

account for log-skewness, which generally rules out the LN model. Model comparison statistics, such as DIC, can be used to further evaluate model fit and select an appropriate model.

If, on the other hand, the focus is on region-specific density estimation, with a particular interest in estimating expenditure quantiles, then the DPLN may be preferable. Recall that the non-mixture models allow only for spatial location shifts on the log scale, whereas the DPLN allows the shape of the density to vary spatially, thereby accommodating additional features such as multimodality. This added flexibility is propagated to the original data scale and can aid region-level inferences. In our application, for instance, the predicted expenditure quantiles for the two block groups in Figure 4 were approximately 25% higher under the LST than under the DPLN, indicating that the choice of model can have a notable impact on inferences.

Future directions might include developing spatiotemporal versions of the proposed models to explore evolving spatial patterns in ED expenditures over time. The models could also be used to examine other types of health expenditures, such as prescription drug spending. More generally, the models developed here should prove useful in many settings where interest lies in examining joint spatial trends in health services use and related medical expenditures. The Bayesian approach outlined in this paper provides a practical framework for fitting such models.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

FUNDING

This work was supported by Grant 1C1CMS331018-01-00 from the Department of Health and Human Services, Centers for Medicare & Medicaid Services, and by the Bristol Myers Squibb Foundation Together on Diabetes program. The contents are solely the responsibility of the authors and have not been approved by the Department of Health and Human Services, Centers for Medicare & Medicaid Services.

REFERENCES

ALBERT, J. H. AND CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669–679.

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

AZZALINI, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* **12**, 171–178.

AZZALINI, A. AND CAPITANIO, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology*) **65**, 367–389.

BANERJEE, S., CARLIN, B. P. AND GELFAND, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton: Chapman & Hall/CRC.

CELEUX, G., FORBES, F., ROBERT, C. P. AND TITTERINGTON, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis* **1**, 651–673.

Frühwirth-Schnatter, S. and Pyne, S. (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-*t* distributions. *Biostatistics* **11**, 317–353.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (editors), *Bayesian Statistics 4*. Oxford: Clarendon Press, pp. 169–193.

Liu, L., Strawderman, R. L., Johnson, B. and O'Quigley, J. M. (2012). Analyzing repeated measures semi-continuous data, with application to an alcohol dependence study. *Statistical Methods in Medical Research*. http://smm.sagepub.com/content/early/2012/04/01/0962280212443324.full.pdf+html

Manning, W. G., Basu, A. and Mullahy, J. (2005). Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics* **24**, 465–488.

National Priorities Partnership. (2010). Reducing emergency department overuse: a $38 billion opportunity. *NQF Report*, National Quality Forum, Washington, D.C.

Newhouse, J. P., Garber, A. M., Graham, R. P., McCoy, M. A., Mancher, M. and Kibria, A. (editors); Committee on Geographic Variation in Health Care Spending & Promotion of High-Value Care; Board on Health Care Services; Institute of Medicine. (2013). *Variation in Health Care Spending: Target Decision Making, Not Geography*. Washington, D.C: The National Academies Press.

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

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

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

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology*) **62**, 795–809.

Su, L., Tom, B. D. M. and Farewell, V. T. (2009). Bias in 2-part mixed models for longitudinal semicontinuous data. *Biostatistics* **10**, 374–389.

U.S. Census Bureau. (2010). *American Community Survey 2006–2010*. Washington, D.C: U.S. Census Bureau.

[*Received April 24, 2014; revised December 11, 2014; accepted for publication December 18, 2014*]