Bayesian Factor Analysis for Spatially Correlated Data, With Application to Summarizing Area-Level Material Deprivation From Census Data

Joseph W. HOGAN and Rusty TCHERNIS

This article describes a Bayesian hierarchical model for factor analysis of spatially correlated multivariate data. The first level specifies, for each area on a map, the distribution of a vector of manifest variables conditional on an underlying latent factor; at the second level, the area-specific latent factors have a joint distribution that incorporates spatial correlation. The framework allows for both marginal and conditional (e.g., conditional autoregressive) specifications of spatial correlation. The model is used to quantify material deprivation at the census tract level using data from the 1990 U.S. Census in Rhode Island. An existing and widely used measure of material deprivation is the Townsend index, an unweighted sum of four standardized census variables (i.e., Z scores) corresponding to area-level proportions of unemployment, car ownership, crowding, and home ownership. The Townsend and many related indices are computed as linear combinations of measured census variables, which motivates the factor-analytic structure adopted here. The model-based index is the posterior expectation of the latent factor, given the census variables and model parameters. Index construction based on a model allows several improvements over Townsend's and similarly constructed indices: (1) The index can be represented as a weighted sum of (standardized) census variables, with data-driven weights; (2) by using posterior summaries, the indices can be reported with corresponding measures of uncertainty; and (3) incorporating information from neighboring areas improves precision of the posterior parameter distributions. Using data from Rhode Island census tracts, we apply our model to summarize variations in material deprivation across the state. Our analysis entertains various spatial covariance structures. We summarize the relative contributions of each census variable to the latent index, suggest ways to report material deprivation at the area level, and compare our model-based summaries with those found by applying the standard Townsend index.

KEY WORDS: Conditional autoregressive model; Health inequalities; Latent variable; Posterior rank; Small-area estimation; Socioeconomic status; Townsend index.

1. FACTOR ANALYSIS AND SPATIAL DATA

Factor-analytic models are useful for summarizing variance and covariance patterns in multivariate data. A common formulation of factor analysis assumes that measurable variables, such as scores on a test, are manifestations of an underlying latent construct, such as ability or intelligence. The latent variable formulation can be useful for data reduction, that is, summarizing multivariate observations using a lower-dimensional variable. A thorough review has been given by Bartholomew and Knott (1999, chaps. 1–3). Recent work from a Bayesian perspective has been done by Geweke and Zhou (1996), Press and Shigamesu (1997), Aguilar and West (2000), and Rowe (2002).

Multivariate spatial data can arise in a number of applied contexts. Wang and Wall (2001, 2003) studied multivariate indicators of cancer risk across counties in Minnesota. Samet, Dominici, Curriero, Coursac, and Zeger (2000) considered the effect on mortality of multiple measurements of air pollution exposure in 20 U.S. cities. Lee, Murie, and Gordon (1995) summarized a variety of methods for combining multiple area-specific census variables into scalar measures of socioeconomic standing. The need to analyze or summarize multivariate data that are spatially aligned suggests the utility of factor-analytic models that can incorporate spatial covariation. Wang and Wall (2001, 2003) recently introduced a factor-analytic model for spatially correlated multivariate cause-specific mortality, which provides area-specific scalar summaries of mortality via factor scores. Their first article studied large-sample properties of the model from a frequentist perspective, and the second developed a Bayesian approach that allows discrete manifest variables. Christensen and Amemyia (2002, 2003) developed semiparametric latent variable models for rectangular grids and provided key references to related work in geography and geology.

In this article we use a Bayesian factor-analytic model of spatially correlated data to summarize area-specific material deprivation from multiple census variables. Spatial correlation is modeled on the latent variable scale and can be specified either marginally or conditionally (e.g., conditional autoregressive structures). The Bayesian approach is natural, because the factor-analytic model is hierarchical in nature; furthermore, it confers the distinct advantage that uncertainty about factor scores, which in many cases are of direct interest, can be accurately summarized as a natural byproduct of the posterior parameter distribution. Moreover, constraints on the variance structure that must be imposed for certain CAR specifications pose no special problems in the estimation routines.

The rest of the article is organized as follows. Section 2 is devoted to providing background on the uses of material deprivation indices, and motivates their construction via factor analysis. Section 3 describes a general approach to incorporating spatial correlation in a factor-analytic model, gives several examples of parameterizations, and highlights key differences between marginal and conditional specifications. The development in Section 3 is given in context of measuring material de-

Joseph W. Hogan is Associate Professor, Center for Statistical Sciences, Department of Community Health, Brown University, Providence, RI 02912 (E-mail: *jwh@brown.edu*). Rusty Tchernis is Assistant Professor, Department of Economics, Wylie Hall, 100 S. Woodlawn, Bloomington, IN 47405 (E-mail: *rtcherni@indiana.edu*). The authors are grateful to several colleagues for helpful discussions and/or comments on earlier drafts, including Michael Daniels, Peter Diggle, Constantine Gatsonis, Xiao-Li Meng, Vincent Mor, Melanie Wall, Lance Waller, Michael White, Scott Zeger, and Sally Zierler. The authors also thank the editor, associate editor, and three anonymous referees for their helpful critiques and suggestions. Census data were formatted in SAS by Jarvis Chen, Kerry Lemieux, and Pamela Waterman at Harvard School of Public Health. Lynn Carlson and Annie Gjelsvik at Brown University provided assistance with mapping. Work on this project was partially funded through National Institute of Health grant R01-HD-36865 via subcontract to Brown University.

Journal of the American Statistical Association
June 2004, Vol. 99, No. 466, Applications and Case Studies
DOI 10.1198/016214504000000296

privation from multivariate census variables. Section 4 applies several models to census data from Rhode Island, and Section 5 concludes.

2. CONSTRUCTION AND USE OF MATERIAL DEPRIVATION INDICES

2.1 Conceptualization and Measurement of Material Deprivation

Deprivation itself is typically viewed as an underlying construct rather than a measurable characteristic, although it manifests in measurable ways. Deprivation is related to but distinct from *poverty*, which is a measurable quantity defined in terms of per capita income. Material deprivation is meant to encompass a broader spectrum of access to the materials needed for daily living, and thus is a potentially richer construct than poverty per se.

Indices of material deprivation usually are derived from multiple components and computed in uncomplicated ways; a recent comprehensive review of indices used in the United Kingdom has been given by Lee et al. (1995). One widely used measure, the Townsend index (Townsend, Simpson, and Tibbs 1985) can be considered a metric on which to measure material resources and is representative. For a given area (e.g., census tract or zip code), the Townsend index uses information on four area-level measures: percent unemployed, percent of households without a car, percent of households not owner-occupied, and percent of households with more than one person per room (i.e., percent crowding) (Townsend et al. 1985; Lee et al. 1995). Natural log transformations are applied to the unemployment and crowding variables, and each variable is standardized in terms of a Z score. The area-specific index is computed as an unweighted sum of the Z scores, with higher values representing greater deprivation; that is, for area i, the index is $T_i = Z_{1i} + Z_{2i} + Z_{3i} + Z_{4i}$. Krieger et al. (2001) described a Townsend index derived from U.S. Census variables, which we use for our analyses in Section 4. Other indices of material deprivation are based on similar constructions (Hutchinson, Foy, and Sandhu, 1989; Jarman 1983; Carstairs 1995; Lee et al. 1995).

2.2 Uses of Deprivation Indices

Area-level indices of material deprivation play an important role in public health and demographic research and in governmental policy. Public health researchers use these indices to study relationships between, for example, health outcomes and social status (see, e.g., Eachus et al. 1996; Wilson, Chen, Taylor, McCracken, and Copeland 1999; Sundquist, Malmström, and Johansson 1999). Governmental agencies, particularly in the United Kingdom, incorporate information about area-level deprivation together with mortality rates for making decisions about resource allocation in relation to health care, because these are considered reliable proxies for morbidity and health services use (Hutchinson et al. 1989; Carstairs 1995). The indices are derived from publicly available data and can be calibrated using data on morbidity and/or specific health outcomes, such as cancer, that are directly related to use of services but typically are available only from sources, such as registries, that are not publicly available. Thus one objective in this area of research is to develop indices with high construct validity as it relates to use of services (see, e.g., Carstairs 1995; Malmström, Sundquist, Bajekal, and Johansson 1998; Barnett, Roderick, Martin, Diamond, and Wrigley 2002). Barnett et al. (2002) provided a recent review of the key issues related to index use in resource allocation, and carried out a detailed analysis showing the relationships between mortality, morbidity, and several deprivation indices. A related and recently opened line of research in the United States concerns the use of area-based measures—in particular, those based on census variables—to quantify and monitor socioeconomic inequalities in health (Krieger et al. 2001, 2002).

Although some researchers object to deprivation indices as proxies of morbidity and health (e.g., Sheldon, Davey-Smith, and Bevan 1993), they remain in widespread use as explanatory variables for epidemiological studies and as components of resource allocation algorithms; furthermore, they show considerable promise for monitoring health inequalities. The primary motivation for our proposed methodology, therefore, is to formalize index development in the context of a model so that the resulting index makes better use of its component variables compared with the empirical indices, for example, by using spatial correlations to incorporate data from neighboring areas into the index, and by calculating measures of uncertainty for each area's index. A potentially important and useful output from the model is a database containing area-specific indices derived from the model, which can subsequently be used by researchers for validation studies. The discussion in Section 5 contains some specific recommendations in this regard.

2.3 Limitations of Existing Measures and Motivation for Model-Based Indices

Despite their intuitive appeal, empirical indices such as Townsend's are based on a number of important structural assumptions. Because the index is an unweighted sum of the Z scores, each variable contributes equally to measured deprivation; furthermore, even though areas may have different population sizes, no adjustment is made for differential precision of the component variables across areas. Moreover, it is assumed that for a specific area, information about deprivation depends exclusively on variables from that area, and not (for example) on variables from neighboring areas. Finally, once computed, the index lacks a measure of uncertainty. For policy-making decisions in particular, this last feature may be problematic, for example, if decisions about resource allocation are based on cutoff values or percentiles of the index.

An effective way to address these concerns is to cast the relationship between the component variables (Z's) and the deprivation index in a modeling framework. We adopt a factor-analytic structure, implemented in a fully Bayesian framework, in which the component variables are the "manifest variables" and deprivation is the underlying latent factor. The deprivation index is defined as the posterior expectation of the latent factor given the manifest variables and model parameters. Under certain distributional assumptions, this approach confers several advantages: first, the model-based deprivation index retains its simple structure as a weighted average of census variables, but the weights are functions of model parameters and therefore are informed by data; second, different sample sizes across areas are incorporated naturally and reflected in posterior variability of the indices; and third, deprivation indices are summarized not as single numbers, but rather in terms of posterior distributions that reflect important uncertainties about deprivation status.

A final limitation of empirical indices is that only data from the specific area are used. We address this shortcoming by introducing spatial dependencies between the latent deprivation indices, thereby incorporating information from census data of neighboring areas. If the underlying latent factor has multivariate normal distribution across the spatially aligned areas, then a wide variety of parameterizations incorporating spatial correlation are available, each with particular implications for model identification and posterior sampling. We consider both marginally and conditionally specified correlation structures; conditional specifications in particular imply marginal variance structures that are not compatible with the usual set of constraints in factor analysis, but the difficulty can be resolved by rescaling.

To illustrate this approach, we use data from the 1990 census to characterize deprivation in 228 of Rhode Island's 232 census tracts. (Two tracts are excluded because they are outlying islands and lack natural contiguous neighbors; two others are excluded because the observed value of one or more manifest variables is an extreme outlier.) We compare the Townsend index and associated ranked values with posterior distributions of model-based indices—and the posterior distribution of their ranks—under several different spatial correlation models. We illustrate two ways to summarize area-level deprivation so as to communicate both ordering and uncertainty.

3. FACTOR–ANALYTIC MODELS OF MATERIAL DEPRIVATION

Factor analysis provides a natural structure within which to generalize the Townsend and related indices because the latent index (or, more precisely, its posterior expectation conditional on manifest variables and model parameters) is a linear combination of the manifest variables (Bartholomew and Knott 1999). After defining relevant notation, we describe a factor-analytic characterization of deprivation index under spatial independence and assuming that area-level samples leading to the manifest variables are of equal size.

The second part of this section is devoted to describing various parameterizations of spatial correlation and their implications for both model interpretation and posterior sampling. We describe both marginally and conditionally specified correlation structures, focusing on Gaussian conditional autoregressive (CAR) parameterizations for the latter.

3.1 Notation and General Model Structure

We describe the model in terms of census variables, but the ideas here apply more generally. Census variables used here represent proportions (e.g., percent unemployed) estimated from a fractional subsample of the full population. For spatial location i (e.g., block group, census tract), where $i = 1, \ldots, N$, let S_{ij} denote the numerator, and m_i denote the denominator, for constructing the realized census variable $Z_{ij} = S_{ij}/m_i$ ($j = 1, \ldots, J$). The denominator is the sampling fraction times the census tract population. For most areas, the sampling fraction is $1/6$ (U.S. Census Bureau 1999); for our example we assume a sampling fraction of $1/6$ for all areas. Further define $Y_{ij} = g(Z_{ij})$ as an appropriate transformation of the census variable, and let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iJ})^{\mathrm{T}}$.

Following Cressie and Chan (1989), we use a square root transformation, $Y_{ij} = Z_{ij}^{1/2} = (S_{ij}/m_i)^{1/2}$, to stabilize variances. Exploratory analyses of the Rhode Island census data indicate that the variance–covariance matrices computed from data grouped by quartiles of m_i are approximately equal, suggesting that $\mathrm{var}(Y_{ij}) \approx \tau_j^2/m_i$ and, more generally, $\mathrm{var}(\mathbf{Y}_i) \approx \mathbf{T}/m_i$, where $\mathbf{T}$ is a 4×4 variance–covariance matrix. Defining $\mathbf{M} = \mathrm{diag}(m_1, \ldots, m_N)$ as the $N \times N$ matrix with m_i along the diagonal and 0's elsewhere, we have $\mathrm{var}(\mathbf{Y}) \approx \mathbf{M}^{-1} \otimes \mathbf{T}$, where $\otimes$ denotes the Kronecker product.

The general factor-analytic model assumes that each area has an L-dimensional ($L < J$) latent variable, $\boldsymbol{\delta}_i = (\delta_{i1}, \ldots, \delta_{iL})^{\mathrm{T}}$, that fully characterizes socioeconomic characteristics, which in turn are manifest through $\mathbf{Y}_i$. For our application to the Rhode Island data, we focus on the case of a univariate index ($L = 1$) and represent the model in hierarchical form. At level I, $\mathbf{Y}_i|\delta_i \sim \mathcal{N}(\boldsymbol{\mu}_i + \boldsymbol{\lambda}\delta_i, \boldsymbol{\Sigma}/m_i)$, where $\boldsymbol{\mu}_i$ is a $J \times 1$ mean vector, $\boldsymbol{\lambda}$ is a $J \times 1$ vector of *factor loadings*, and $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_J^2)$ is a diagonal matrix measuring residual variation in $\mathbf{Y}_i$. That $\boldsymbol{\Sigma}$ is diagonal implies independence between elements of $\mathbf{Y}_i$ conditionally on δ_i. At level II, which in our case characterizes between-area variation, we make the customary assumption that $\delta_i \sim \mathcal{N}(0, 1/m_i)$. The assumption of known variance ensures identifiability of $\boldsymbol{\lambda}$ and $\boldsymbol{\Sigma}$ (Bartholomew and Knott 1999). Level III specifies prior distributions on the unknown parameters $\{\boldsymbol{\mu}_i\}$, $\boldsymbol{\lambda}$, and $\boldsymbol{\Sigma}$.

It is convenient to write the model in compact form. Let $\mathbf{Y} = (\mathbf{Y}_1^{\mathrm{T}}, \ldots, \mathbf{Y}_N^{\mathrm{T}})^{\mathrm{T}}$ represent the $NJ \times 1$ stacked vector of manifest variables (with a corresponding $NJ \times 1$ vector $\boldsymbol{\mu}$ similarly defined), and let $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_N)^{\mathrm{T}}$ denote the $N \times 1$ vector of area-specific latent variables. Then levels I and II can be rewritten as

$$\begin{aligned} &\text{I.} \quad (\mathbf{Y}|\boldsymbol{\delta}) \sim \mathcal{N}(\boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\delta}, \mathbf{M}^{-1} \otimes \boldsymbol{\Sigma}), \\ &\text{II.} \quad \boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}_N, \mathbf{M}^{-1}), \end{aligned} \tag{1}$$

where $\boldsymbol{\Lambda} = \mathbf{I}_N \otimes \boldsymbol{\lambda}$ is the $NJ \times N$ matrix of factor loadings and $\mathbf{I}_N$ denotes an identity matrix with dimension N. This is the spatial independence model, model 1.

3.2 Incorporating Spatial Dependencies and Computing a Model-Based Index

For geographic data, there are at least two motivations for incorporating possible spatial dependencies in the model. First, it is highly likely that adjacent areas have similar socioeconomic characteristics, so that ignoring spatial correlation may result in incorrect posteriors, particularly with respect to uncertainty measures. Second, if spatial dependence is present, then estimation of the posterior index distribution can incorporate information from neighboring areas; specifically, under multivariate normality in levels I and II, the expectation of δ_i given the full vector of manifest variables $\mathbf{Y}$ will be a linear combination of both $\mathbf{Y}_i$ and $\mathbf{Y}_j$ for areas $j \neq i$. A potential byproduct of incorporating spatial correlation is increased precision for parameter information gained by "borrowing of information" from neighboring tracts (see Sec. 4).

Similar to the approaches of Christensen and Amemiya (2002, 2003) and Wang and Wall (2001, 2003), we induce spatial correlation at level II. Generically, we assume that

$$\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}_N, \mathbf{M}^{-1/2}\boldsymbol{\Psi}\mathbf{M}^{-1/2}), \tag{2}$$

where $\boldsymbol{\Psi}$ is an $N \times N$ correlation matrix with 1's on the diagonal and $\psi_{ij} = \text{corr}(\delta_i, \delta_j)$ on the off-diagonal. When $\boldsymbol{\Psi} = \mathbf{I}_N$, (2) reduces to model (1). The motivation for constraining $\boldsymbol{\Psi}$ to have 1's along the diagonal is to maintain the same scaling for δ_i in models with and without spatial correlation, that is, $\text{var}(m_i^{1/2}\delta_i) = 1$. The model-based deprivation index for area i is summarized by the conditional distribution of the "factor score" δ_i given $\mathbf{Y}$ and $(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\Sigma})$. Following standard arguments for multivariate normal distribution, $(\boldsymbol{\delta}|\mathbf{Y}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\Sigma}) \sim \mathcal{N}(\mathbf{d}, \mathbf{D})$, where

$$\begin{aligned} \mathbf{D} &= \left\{\mathbf{M}^{-1/2}\boldsymbol{\Psi}\mathbf{M}^{-1/2} + \boldsymbol{\Lambda}^{\mathrm{T}}(\mathbf{M}^{-1} \otimes \boldsymbol{\Sigma})^{-1}\boldsymbol{\Lambda}\right\}^{-1}, \\ \mathbf{d} &= \mathbf{D}\boldsymbol{\Lambda}^{\mathrm{T}}(\mathbf{M}^{-1} \otimes \boldsymbol{\Sigma})^{-1}(\mathbf{Y} - \boldsymbol{\mu}). \end{aligned} \tag{3}$$

This implies that area-specific indices ($\mathbf{d}$) are linear combinations of the (mean-centered) manifest variables. When $\boldsymbol{\Psi} = \mathbf{I}_N$ (spatial independence), $\mathbf{D}$ is block diagonal, and the area-specific posterior expectation of δ_i given $\mathbf{Y}_i$ and $(\boldsymbol{\Lambda}, \boldsymbol{\Sigma}, \boldsymbol{\mu})$ depends linearly on manifest variables from area i exclusively. More generally, when spatial dependence is nonzero, $\mathbf{D}$ is not block diagonal, and the index for area i may depend on manifest variables outside area i.

The matrix $\boldsymbol{\Psi}$ of spatial correlations can be parameterized either marginally (Cressie 1993; Conlon and Waller 1998) or conditionally (Besag 1974; Sun, Tsutakawa, and Speckman 1999). We use both approaches in our application. For the marginal specification, we assume that $\psi_{ij} = \text{corr}(\delta_i, \delta_j) = \phi^2 \exp(-\zeta d_{ij})$, where $\zeta \geq 0$, to ensure that $\psi_{ij} < 1$, and d_{ij} is the Euclidean distance between centroids of areas i and j, with $d_{ii} = 0$ by definition (Conlon and Waller 1998). For model identifiability, we further assume that ϕ^2 is known ($\phi = 1$) and refer to this specification as model 2.

Conditional autoregressive specifications of spatial dependency are useful for normally distributed data, because when the conditional distributions are assumed normal, the corresponding joint marginal distributions are multivariate normal and can be derived directly (Besag 1974). Sun et al. (1999) described a rather general structure for Gaussian CAR models. Let $\mathcal{R}_i$ denote the set of indices for areas that are neighbors of area i; typically, but not necessarily, neighbors are defined by adjacency. If

$$\delta_i|\{\delta_j : j \in \mathcal{R}_i\} \sim \mathcal{N}\left(\sum_{j\in\mathcal{R}_i} \beta_{ij}\delta_j, \nu/\alpha_i\right), \tag{4}$$

then the joint marginal distribution of $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_N)^{\mathrm{T}}$ follows $\mathcal{N}(\mathbf{0}, \nu\mathbf{B}^{-1})$, where $\mathbf{B}$ is an $N \times N$ matrix with $\{\alpha_1, \ldots, \alpha_N\}$ along the diagonal and $-\alpha_i\beta_{ij}$ on the off-diagonal, provided that $\mathbf{B}$ is symmetric and positive definite (Besag 1974; Sun et al. 1999). A simple and common parameterization of the CAR model gives some insight about parameter interpretation. Let $R_{ij} = I(j \in \mathcal{R}_i)$ be the indicator that area j is a neighbor of area i, set $\beta_{ij} = \omega R_{ij}$, and hold α_i constant (e.g., $\alpha_i = 1$). Then ω measures degree of spatial correlation and ν measures residual variation. Elaborations allow overall residual variation ν/α_i to depend, through α_i, on area-level characteristics such as number of neighbors (Bernardinelli, Clayton, and Montomoli 1995).

The factor-analytic model requires that the marginal variance of δ_i be known. In our CAR models, we fix conditional variance, setting $\nu = 1$. We then rescale the implied marginal model such that $\text{var}(\delta_i) = 1$, so posterior inferences about factor loadings from CAR models can be directly compared with models where spatial correlation is specified marginally. Moreover, one or more parameters must be constrained to ensure that $\mathbf{B}$ is positive definite (Sun et al. 1999); the constraints are model specific. We use several CAR formulations to analyze the Rhode Island census data:

Model 3A defines $\mathcal{R}_i$ as the set of adjacent census tracts and sets $\beta_{ij} = \omega R_{ij}$, $\alpha_i = 1$, and $\nu = 1$, leading to $\mathbf{B} = \mathbf{I}_N - \omega\mathbf{R}$, where $\mathbf{R}$ is an adjacency (weight) matrix with $R_{ii} = 0$ and indicators $R_{ij} = R_{ji}$ on the off-diagonal, that is, $R_{ij} = I(j \in \mathcal{R}_i)$. Let $\xi_1, \ldots, \xi_N$ denote the ordered eigenvalues of $\mathbf{R}$; then $\mathbf{B}$ will be positive definite only if $\xi_1^{-1} < \omega < \xi_N^{-1}$ (Sun et al. 1999). The rationale for setting $\nu = 1$ is that if $\omega = 0$, then $\mathbf{B} = \mathbf{I}_N$, consistent with the convention of assuming known variance for $\boldsymbol{\delta}$ to ensure identifiability. However, it is important to recognize that the variance constraint is being imposed on the conditional distribution of the δ_i given its neighbors, which in general is not consistent with assuming that $\boldsymbol{\Psi}$ has 1's on the diagonal. Setting $\omega = 0$ implies model 1.

Model 3B defines $\mathcal{R}_i$ the same as in 3A, but sets $\beta_{ij} = \omega R_{ij}(n_j/n_i)^{1/2}$, $\alpha_i = n_i$, and $\nu = 1$ (where n_i is number of neighbors for area i). This yields $\mathbf{B} = \text{diag}(n_i) - \omega(n_i n_j)^{1/2}\mathbf{R}$. To ensure that $\mathbf{B}$ is positive definite, ω has the same constraints as for model 3A. This model was suggested by a referee on the grounds that conditional spatial correlation $\text{corr}^2(\delta_i, \delta_j|\{\delta_k : k \in \mathcal{R}_i, k \neq i, j\}) = \omega^2$ is not dependent on neighborhood size (Cressie 1993, sec. 7.6; Stern and Cressie 1999).

We also consider three CAR specifications that incorporate distance between tracts. The first, model 3C, is based purely on distance between neighbors (Best, Arnold, Thomas, Waller, and Conlon 1999), wherein $\mathcal{R}_i$ includes all areas except area i, $\beta_{ij} = \exp(-\gamma d_{ij})$, $\alpha_i = 1$, and γ is unknown. The parameterization of β_{ij} as exponential decay is informed by empirical correlograms for the manifest variables (Fig. 1). The remaining models, 3D and 3E, combine information on adjacency and distance. Centroids of adjacent Rhode Island census tracts are geographically very close in urban areas but not in rural areas, and spatial dependencies appear more pronounced in areas with closer neighbors (Fig. 2). Neighborhoods are defined by adjacency, and spatial correlation may depend on distance within neighborhood (Cressie and Chan 1989). If spatial correlation is positive, then neighbors whose geographic centers are close (e.g., northeastern urban centers) will exhibit greater spatial dependency than those whose centers are further apart (e.g., rural western part of the state). In model 3D, $\beta_{ij} = \omega \exp(-\gamma d_{ij}) R_{ij}$ and $\alpha_i = 1$. It is necessary to fix one of the two unknown parameters; following Cressie and Chan (1989), we fit the model using point mass priors for γ to moderate the role of distance within neighborhood structure. In model 3E we use normalized weights $\beta_{ij} = \omega \exp(-\gamma d_{ij}) R_{ij} / \sum_j \exp(-\gamma d_{ij}) R_{ij}$ and $\alpha_i = \sum_j \exp(-\gamma d_{ij}) R_{ij}$. As with 3E, we use point mass priors for γ.

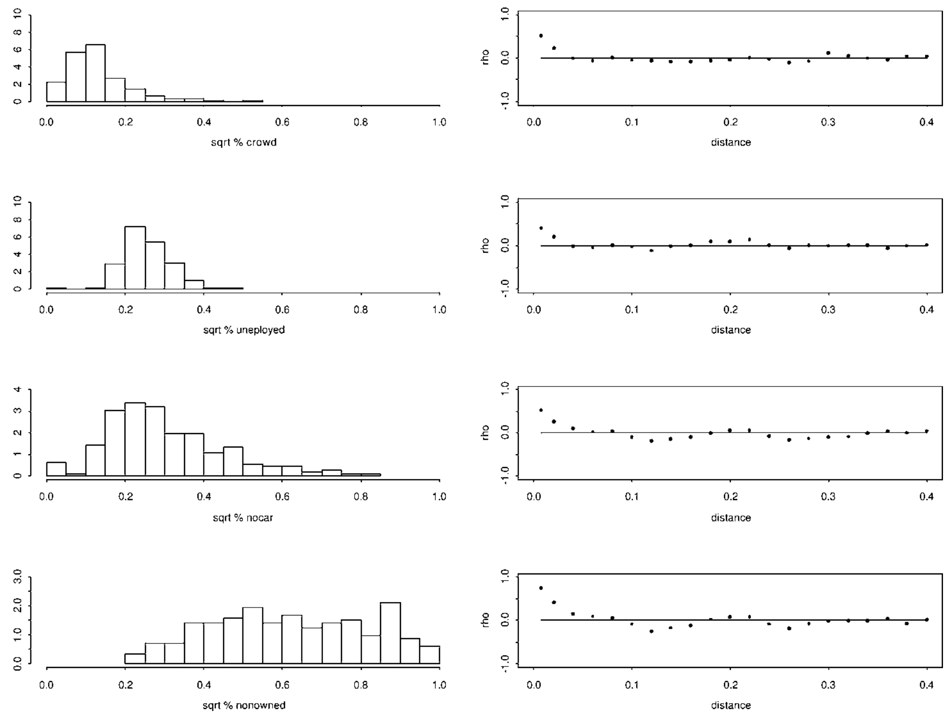

Figure 1. Histograms and Correlograms for Each Transformed Census Variable; From Top to Bottom: Percent of Housing Units With Crowding, Percent Unemployed, Percent Without a Car, Percent Housing Not Owner-Occupied.

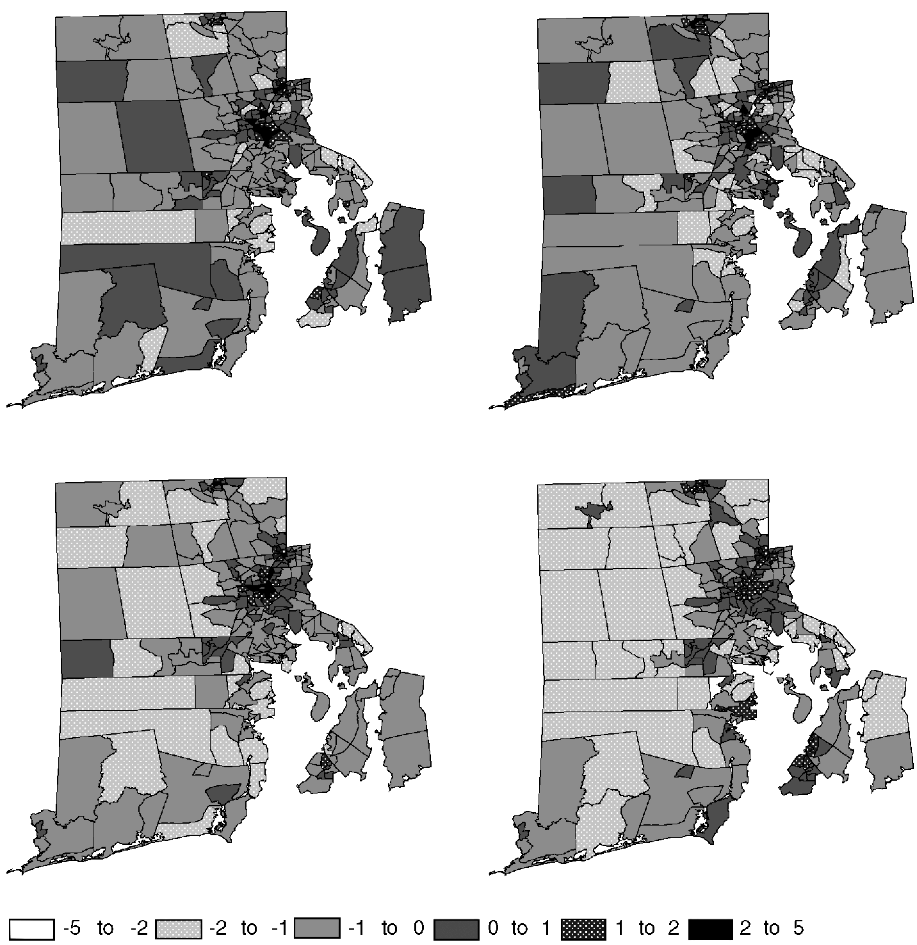

Figure 2. Spatial Distribution of Each Transformed Census Variable, in Terms of Standard Deviations From the Mean in Square Root Scale. Percent housing units with crowding (Z_1, top left), percent unemployed (Z_2, top right), percent without a car (Z_3, bottom left), percent housing not owner-occupied (Z_4, bottom right). Shading indicates number of standard deviations from the mean.

3.3 Compatibility Between Marginal and CAR Models

Posterior factor loadings from CAR models cannot be directly compared with those from marginally specified models because they are measured on different scales, due to the latent factors $\boldsymbol{\delta}$ having different variances in the two types of models. At the second level of marginally specified models (e.g., models 1 and 2), we assume that $\text{var}(\boldsymbol{\delta}) = \boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ is a correlation matrix having 1's along its diagonal. On the other hand, the second level of a CAR model implies $\text{var}(\boldsymbol{\delta}) = \mathbf{B}^{-1}$, which in turn implies that $\text{var}(\mathbf{Y}) = \boldsymbol{\Lambda}\mathbf{B}^{-1}\boldsymbol{\Lambda}^{\mathrm{T}} + \mathbf{I}_N \otimes \boldsymbol{\Sigma}$. Note, however, that because parameters in CAR models 3A–3E are constrained such that $\mathbf{B}^{-1}$ is symmetric and positive definite, we can write $\mathbf{B}^{-1} = \mathbf{V}^{1/2}\boldsymbol{\Psi}\mathbf{V}^{1/2}$, where $\mathbf{V}$ is an $N \times N$ diagonal matrix of variances and $\boldsymbol{\Psi}$ is the corresponding correlation matrix. Hence, $\text{var}(\mathbf{Y}) = \boldsymbol{\Lambda}\mathbf{V}^{1/2}\boldsymbol{\Psi}\mathbf{V}^{1/2}\boldsymbol{\Lambda}^{\mathrm{T}} + \mathbf{I}_N \otimes \boldsymbol{\Sigma} = \widetilde{\boldsymbol{\Lambda}}\boldsymbol{\Psi}\widetilde{\boldsymbol{\Lambda}}^{\mathrm{T}} + \mathbf{I}_N \otimes \boldsymbol{\Sigma}$, where $\widetilde{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}\mathbf{V}^{1/2}$, and factor loadings $\widetilde{\boldsymbol{\Lambda}}$ from CAR models will have the same scale as those derived from marginally specified models.

4. ANALYSIS OF RHODE ISLAND CENSUS DATA

4.1 Objectives

The main objective of our analysis is to use of Bayesian factor analysis for computing and summarizing material deprivation at the census tract level. This involves fitting several models to the manifest variables, selecting a model that best represents the observed data, and then computing the posterior conditional distribution $(\boldsymbol{\delta}|\mathbf{Y}, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{\Psi})$ using the chosen model. Our summary of material deprivation is in terms of posterior ranks of individual δ_i drawn from this distribution, and in terms of posterior probability of deprivation. Our suggestion is to define "deprivation" in terms of percentiles of the latent index distribution, for example, the upper 20% of census tracts. Model-based summaries are compared in detail to Townsend indices.

4.2 Manifest Variables From Rhode Island Census Data

Each census-level variable is the square root of a percentage, and these variables are labeled as follows. For census tract (area) i, Y_{i1} corresponds to the (square root) of percentage of "crowded" households (more than one person per room), Y_{i2} denotes percent unemployed, Y_{i3} is percent of individuals without a car, and Y_{i4} is percentage of households not owned by the occupant(s). Figure 1 shows, for the transformed variables, histograms in the left panels and correlograms in the right panels. For census variable j, the correlograms estimate the function $\rho_j(h) = C_j(h)/C_j(0)$, where $C_j(h) = \text{cov}\{Y_{ij}, Y_{i'j}|d_{ii'} = h\}$, that is, the average between-area correlation as a function of Euclidean distance between area centroids. The correlogram indicates substantial correlation for areas close together. On the map that we used, maximum Euclidean distance between centroids of two areas is .801, and maximum distance between centroids of two adjacent areas is .180. The spatial distribution of each standardized variable is shown in Figure 2. From Figures 1 and 2, spatial clustering appears evident for all variables.

4.3 Model Selection and Posterior Parameter Distributions

Each of the models described in Section 3 was fit to the manifest variables; details on prior selection and posterior sampling are given in the Appendix. Model selection was made using a posterior predictive criterion that balances goodness of fit and predictive variance under a squared error loss function (Gelfand and Ghosh 1998; see also Spiegelhalter, Best, Carlin, and van der Linde 2002 for a criterion based on deviance). Define, for model q, $\theta_{ij}(q) = E_q(Y_{ij}^{\text{rep}}|\mathbf{Y})$ and $\nu_{ij}(q) = \text{var}_q(Y_{ij}^{\text{rep}}|\mathbf{Y})$, where Y_{ij}^{rep} is an observation from the posterior predictive distribution (i.e., a future observation that has the same distribution as the observed Y_{ij}). The model choice criterion for model q is $C(q,k) = G(q) + \{k/(k+1)\}P(q)$, where

$$G(q) = \sum_{i=1}^{N}\sum_{j=1}^{4}\{\theta_{ij}(q) - Y_{ij}\}^2$$

and

$$P(q) = \sum_{i=1}^{N}\sum_{j=1}^{4}\nu_{ij}(q). \tag{5}$$

This criterion penalizes both for lack of fit and for high posterior predictive variance due to underparameterization or overparameterization. The constant k can be used to calibrate the role of P; that is, when $k = 0$, model choice is based purely on fit, and when $k = \infty$, fit and variance are weighted equally. Table 1 summarizes $C(q, \infty)$ for the models listed in Section 3. For many models, incorporating spatial variation actually leads to poorer model performance relative to the spatial independence model, but in nearly all cases it reduces predictive variance by borrowing information from neighboring areas. The best model with respect to fit and variability is 3B, a CAR formulation with spatial correlation based in adjacency (this holds for several other choices of k).

The posterior distribution of ω suggests strong spatial association on the latent variable scale (median .150, 95% posterior

Table 1. Posterior Predictive Selection Criteria for Nine Models Fitted to the Census Data

q	Model	γ	G(q)	P(q)	C(q)
1	1 (spatial independence)		16.8	18.0	34.8
2	2 (marginal correlation)		19.2	14.2	33.4
3	3A (CAR based on adjacency, $\alpha_i = 1$)		16.8	13.4	30.1
4	3B (CAR based on adjacency, $\alpha_i = n_i$)		**16.6**	12.8	**29.4**
5	3C (CAR based on distance only)		17.9	13.8	31.6
6	3D (CAR based on distance and adjacency)	1	17.0	15.3	32.4
7		40	18.0	14.7	32.7
8	3E (model 3D with normalized weights)	1	17.0	13.6	30.6
9		40	19.3	**12.6**	31.9

NOTE: See (5). Lowest (best) values for each column are given listed in bold.

Table 2. Summary of Posterior Distribution of Factor Loadings λ_j ($j = 1, 2, 3, 4$) for Models 1 (spatial independence) and 3B (CAR), and Spatial Dependence Parameter ω for Model 3B

		Independence (model 1)				CAR (model 3B)			
j	*Factor*	*2.5%*	*50%*	*97.5%*	*IQR*	*2.5%*	*50%*	*97.5%*	*IQR*
1	Crowding	1.33	1.57	1.85	.19	1.00	1.22	1.49	.15
2	Unemployment	.89	1.06	1.24	.13	.69	.84	1.03	.11
3	No car	2.96	3.35	3.76	.28	2.38	2.70	3.16	.22
4	Housing not owned	3.78	4.32	4.92	.37	3.00	3.45	4.08	.33
Spatial dependence (ω)						.145	.150	.154	.003

NOTE: The IQR reflects variation of posterior distribution.

interval .145 to .154; this is very close to its upper bound). The clearest manifestation of incorporating spatial correlation is reduced variability in posterior parameter distributions. Table 2 summarizes the posterior distribution of factor loadings. Variability in the posterior distribution of each λ_j is measured using the interquartile range (IQR); in each case the IQR is smaller for the model incorporating spatial variations. Reductions in posterior IQR range from 14% for λ_4 (housing not owned) to 23% for λ_2 (unemployment). The attenuation of factor loadings in model 3B occurs because each factor loading is a variance component—$\lambda_j = \text{cov}(Y_{ij}, \delta_i)$—and some of the total variation in each Y_{ij} is absorbed by the spatial parameter.

Because residual variances σ_j^2 differ across j, the factor loadings themselves measure covariance on different scales and cannot be directly compared to assess strength of association between census variables and material deprivation. Instead we can examine squared correlation coefficients $\lambda_j^2/(\lambda_j^2 + \sigma_j^2)$, for $j = 1, 2, 3, 4$, which represent proportion of variation in the jth manifest variable Y_{ij} that is explained by the latent index δ_i. Table 3 gives a comparison of these between models 1 and 3B. Under both spatial independence and CAR dependence, the correlation is greatest between lack of car ownership and material deprivation, followed by nonownership of housing, unemployment, and crowding. The contribution of the last two variables is essentially equal.

4.4 Summarizing Distribution of Material Deprivation

Material deprivation across areas is summarized in two ways. First, for each area, we computed the posterior distribution of the rank of $E(\delta_i|\mathbf{Y}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{\Sigma})$, including its mean and 99% probability interval. Using output from the Markov chain Monte Carlo algorithm, we first drew one replicate $(\boldsymbol{\lambda}^*, \boldsymbol{\Sigma}^*, \boldsymbol{\mu}^*, \boldsymbol{\Psi}^*)$ from its joint posterior distribution, then drew the vector $\{\mathbf{d}^*|\boldsymbol{\lambda}^*, \boldsymbol{\Sigma}^*, \boldsymbol{\mu}^*, \boldsymbol{\Psi}^*, \mathbf{Y}\}$ according to (3), and then assigned ranks to the N elements of $\mathbf{d}^*$. The simulated posterior rank distribution was determined from 1,000 replicates of $\mathbf{d}^*$ (see also Laird and Louis 1989). Figure 3 summarizes area-specific posterior ranks, indicating wide variability in precision of measuring deprivation, and makes clear that fine distinctions in deprivation level are unlikely to be possible.

Second, to simplify the reporting of deprivation, it is possible to derive ordered categories that represent degree of deprivation on a reduced scale, for instance, based on quintiles of the distribution of $(\delta_i|\mathbf{Y}, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}, \boldsymbol{\mu})$. The vertical lines on Figure 3 indicate quintile-based categories. The variation exhibited by the rank distribution makes clear the potential difficulties of classifying all areas because many 99% intervals straddle two categories (and in two cases, three categories). An alternative to strict classification is a summary of posterior probability of class membership, that is, the posterior probability that the rank for area i falls between a specific pair of quintiles. In cases where resource allocation (for example) is based on identifying only the 10% or 20% most deprived areas, posterior probability summaries can be used to avoid arbitrary cutoffs and artificial distinctions between areas that may be separated by a small number of ranks. If those above the highest quintile are to be identified as "materially deprived," then any area with positive posterior probability of being ranked above that quintile would be included. Figure 4 illustrates this method of reporting using maps of the posterior probability of inclusion in quintile-based groupings 1, 3, and 5 (5 = greatest deprivation). Darker shading indicates higher probability.

4.5 Model-Based Versus Townsend Indices

Limitations of the Townsend index, spelled out in Section 1, include (a) that each census variable contributes with equal weight to the summary index; (b) census tracts are considered independent, despite similarities between neighboring tracts, possibly resulting in less efficient use of information; and (c) measures of uncertainty, particularly due to varying census tract population, are not part of the index. The model-based

Table 3. Summary of Posterior Distribution of Squared Correlations $\lambda_j^2/(\lambda_j^2 + \sigma_j^2)$ ($j = 1, 2, 3, 4$) for Models 1 (spatial independence) and 3B (CAR)

		Independence (model 1)			CAR (model 3B)		
j	*Factor*	*2.5%*	*50%*	*97.5%*	*2.5%*	*50%*	*97.5%*
1	Crowding	.39	.50	.61	.30	.41	.54
2	Unemployment	.40	.51	.60	.29	.39	.53
3	No car	.75	.82	.88	.71	.79	.87
4	Housing not owned	.66	.75	.83	.54	.65	.73

NOTE: Squared correlations reflect percent variation in an area-specific manifest variable that is explained by variation in material deprivation for that area. Under model 3B, variation in an area-specific manifest variable is explained by material deprivation from other areas as well.

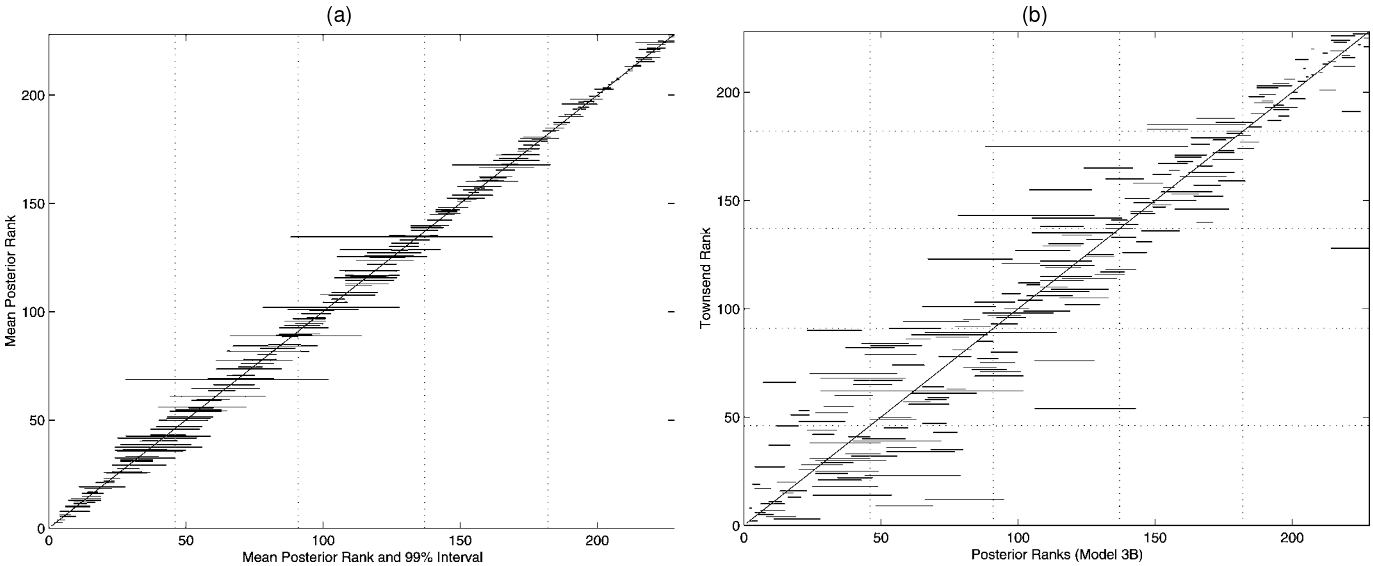

Figure 3. (a) Posterior Ranks and Associated 99% Intervals for Material Deprivation Indices Derived Under Model 3B and (b) Comparison of Distribution of Ranks for Townsend Index (vertical axis) and Model-Based Index From Model 3B (horizontal axis). Horizontal grid lines represent quintiles of the Townsend index, and vertical grid lines represent quintiles of the model-based index.

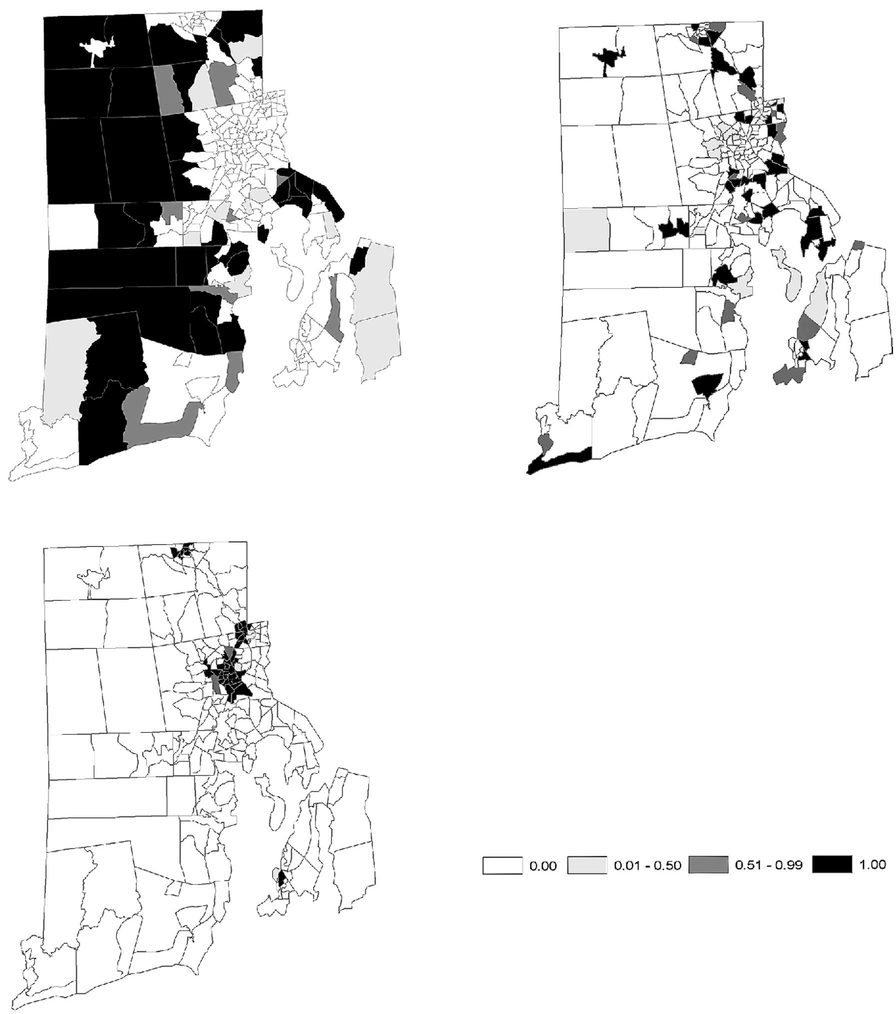

Figure 4. Posterior Probability of Being in Quintile-Based Material Deprivation Category 1 (top left), 3 (top right), and 5 (bottom left), Based on Posterior Distribution of Index Ranks From Model 3B. Darker shading indicates greater posterior probability of membership; higher categories represent greater material deprivation.

Table 4. Summary of Five Areas With Greatest Discrepancy in Rank Between Townsend and Model-Based Indices

			Census variables				Rank		
ID	m_i	n_i	Z_1	Z_2	Z_3	Z_4	Townsend	Model 3B	
28	3,665	7	.5	.3	−2.0	−1.0	66	12.9	(7.0, 19.0)
55	4,157	6	.2	.7	−1.1	−1.2	90	29.2	(23.0, 43.0)
79	7,645	7	−1.5	−1.6	−.1	−.9	12	81.6	(66.0, 95.0)
61	3,504	5	−1.5	−1.2	.5	−.4	54	128.6	(106.0, 143.0)
72	610	5	−1.5	−4.3	3.6	1.9	128	224.1	(214.0, 228.0)

NOTE: Standardized census variables constructed from square root percentages for crowding (Z_1), unemployment (Z_2), no car (Z_3), and housing not owned (Z_4). Higher ranks indicate greater material deprivation. Model-based summary of posterior rank is mean and 99% interval.

method of summarizing information on deprivation addresses each of these shortcomings while retaining the structural format of the Townsend index as a (weighted) linear combination of census variables. When spatial correlation is incorporated, the index for area i may depend on census variables recorded in other areas.

A direct comparison of area-specific indices derived using the Townsend formula and using model 3B is shown in Figure 3(b), which plots ranks of the Townsend index against posterior rank intervals derived from the model. As expected there is generally positive correlation between the ranks, but some notable discordancies are evident. Table 4 presents data from the five most discordant areas. Apparently, discordancies are due in large measure to differential weighting of census variables. In areas 61 and 72, for example, discordancy can be attributed to a higher-than-average percent with no car ($Z_3 > 0$), leading to model-based deprivation indices that are much greater than the Townsend index. (Recall that percent with no car contributes the greatest amount of information to the deprivation index.) The distribution of ranks is much less variable in the upper right tail, where in general there is more agreement between the Townsend and model-based indices.

5. DISCUSSION

We have used a fully Bayesian factor analysis model for spatially correlated data. Spatial correlation is introduced on the latent variable (factor) distribution, and many common parameterizations of spatial covariation can be used, including direct and conditional (CAR) specifications. The model is applied to the problem of calculating an index of material deprivation from census variables, using tract-level data in Rhode Island from the 1990 U.S. Census.

Measuring area-level material deprivation is important in both policy and research. Standard indices, usually constructed as unweighted sums of census variables, are used for health services resource allocation and as proxies of socioeconomic position in epidemiologic research, and more recently have been proposed as public health monitoring tools for small areas (Krieger et al. 2002). The potential use of material deprivation indices as area-level health status surrogates holds great interest because the indices can be constructed from publicly available census data; however, proper validation requires the use of health outcomes data, such as from a nonpublic cancer registry.

Although any measure of a latent construct like deprivation is bound to be imperfect, our model-based approach provides a method for maximizing available information about a fixed area. The model-based index given here is a weighted average of census variables from the target area and from surrounding areas. Spatial correlation allows manifest variables from surrounding areas to inform the index. An advantage of allowing different weights for each variable is that the similar information from different manifest variables does not contribute more than once to the index. The Bayesian formulation leads naturally to indices that are reported with measures of uncertainty, such as posterior probability intervals associated with deprivation rank. In principle, there is no reason why deprivation cannot be summarized on the scale of the δ_i themselves. Although the use of ranks does not convey scaling, it must be noted that a priori scaling of δ_i in terms of the standard normal distribution is itself somewhat arbitrary.

An important question arises as to how researchers and policy makers should use this technology. A major consideration in both policy and research is whether model-based indices lead to increases in validity, whose definition depends on the intended use of the index. Toward this end, the model can be used to generate a public-use database consisting of census variables and appropriate summaries of the posterior index distribution. In the simplest case, the database could contain a posterior mean (i.e., scalar value) for each area, $\varphi_i = E(\delta_i|\mathbf{Y}) = \int E(\delta_i|\mathbf{Y})\, dF(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\Psi}, \boldsymbol{\Sigma})$. Alternately (or additionally), the posterior mean rank could be listed. Using standard analytic tools, researchers could then compare the validity of standard indices, such as the Townsend index, to the model-based indices. Another way to make the data available is in terms of replicate datasets, in the style of multiply imputed versions of census data (cf. Little and Rubin 1987), to convey uncertainty in the deprivation measure. A database of this type would list for each area (say) 10 draws of $E(\delta_i|\mathbf{Y}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\Psi}, \boldsymbol{\Sigma})$ from its posterior distribution. An immediate application of these data would be to regression analyses where the index is used as a proxy for socioeconomic position; inferences would be drawn by combining parameter estimates derived from each imputed dataset using formulas given by Little and Rubin (1987).

Our work can be extended in several directions. First, factor analysis can be used as an exploratory tool to discover multivariate structure in deprivation, or more generally in socioeconomic position (Kreiger et al. 2001), by including more census variables. Thus one natural extension would permit multivariate factors with a spatial correlation structure. Second, as suggested earlier, the advantage gained by using a model versus an empirical index could be quantified through validation studies, such as the one carried out by Barnett et al. (2002). The analyses by Barnett et al. (2002) also suggest that the components of material deprivation may differ between urban and rural areas, suggesting the need to incorporate covariates, which could enter either at the first or the second level (Sammel and Ryan 1996). To validate measurement of deprivation per se, one can use external indicators measured at the neighborhood level (as in Raudenbush and Sampson 1999); to validate the index as a surrogate of morbidity or health services use, data on the specific health outcome is needed (e.g., Krieger et al. 2002). In principle, our model also could be elaborated to include temporal trends when data from different time points are available (e.g., Waller, Carlin, Xia, and Gelfand 1997), for example, to study the relationship between changes in deprivation and corresponding changes in morbidity or health services use.

APPENDIX: POSTERIOR SAMPLING ALGORITHMS

The factor-analytic model can be stated in hierarchical fashion as follows: level I (manifest variables): $(\mathbf{Y}|\boldsymbol{\delta}) \sim \mathcal{N}(\boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\delta}, \mathbf{M}^{-1} \otimes \boldsymbol{\Sigma})$; level II (latent indices): $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}_N, \mathbf{M}^{-1/2}\boldsymbol{\Psi}\mathbf{M}^{-1/2})$; level III (priors): $\lambda_j \sim \mathcal{N}(g, G)I(\lambda_1 > 0)$, $\sigma_j^2 \sim \text{IG}(\alpha/2, \beta/2)$, and $\mu_j \sim \mathcal{N}(0, V_{\boldsymbol{\mu}})$, where $\text{IG}(\cdot,\cdot)$ denotes inverse gamma distribution. Priors are specified using $g = 0$, $G = 10{,}000$, $\alpha = 1/1{,}000$, $\beta = 1/1{,}000$, and $V_{\boldsymbol{\mu}} = 1{,}000$. We simulate from the posterior parameter distribution using Metropolis–Hastings steps within a Gibbs sampling algorithm, which comprises five steps. The first four steps are identical for all the models of spatial correlation, whereas the fifth step depends on the spatial correlation specification. From starting values for each parameter, the algorithm proceeds as follows:

1. Sample elements of $\boldsymbol{\lambda}$ from the conditional distribution $\mathcal{N}(\widehat{\lambda}_j, \widehat{V}_j)I(\lambda_1 > 0)$, where $\widehat{V}_j = (1/G + \boldsymbol{\delta}^{\mathrm{T}}\mathbf{M}\boldsymbol{\delta}/\sigma_j^2)^{-1}$ and $\widehat{\lambda}_j = \widehat{V}_j\{g/G + \boldsymbol{\delta}^{\mathrm{T}}\mathbf{M}(\mathbf{Y}_j - \mathbf{1}_N\mu_j)/\sigma_j^2\}$. Here $\mathbf{Y}_j$ is the $N \times 1$ vector of manifest variable j.
2. Sample elements of $\boldsymbol{\mu}$ from $\mathcal{N}(\widehat{\mu}_j, \widehat{V}_{\mu_j})$, where $\widehat{V}_{\mu_j} = \sigma_j^2 V_{\boldsymbol{\mu}}/\{\sigma_j^2 + V_{\boldsymbol{\mu}}\sum m_i\}$ and $\widehat{\mu}_j = \widehat{V}_{\mu_j}\mathbf{1}_N\mathbf{M}(\mathbf{Y}_j - \boldsymbol{\delta}\lambda_j)/\sigma_j^2$.
3. Sample $\boldsymbol{\delta}$ from $\mathcal{N}(\mathbf{d}, \mathbf{D})$, where $\mathbf{D} = \{\mathbf{M}^{-1/2}\boldsymbol{\Psi}\mathbf{M}^{-1/2} + \boldsymbol{\Lambda}^{\mathrm{T}}(\mathbf{M} \otimes \boldsymbol{\Sigma})^{-1}\boldsymbol{\Lambda}\}^{-1}$ and $\mathbf{d} = \mathbf{D}\boldsymbol{\Lambda}^{\mathrm{T}}(\mathbf{M}^{-1} \otimes \boldsymbol{\Sigma})^{-1}(\mathbf{Y} - \boldsymbol{\mu})$.
4. Sample elements of $\boldsymbol{\Sigma}$ from $\text{IG}(\widehat{\alpha}, \widehat{\beta}_j)$, where $\widehat{\alpha} = (\alpha + N)/2$ and $\widehat{\beta}_j = \{(\mathbf{Y}_j - \mathbf{1}_N\mu_j - \boldsymbol{\delta}\lambda_j)^{\mathrm{T}}\mathbf{M}(\mathbf{Y}_j - \mathbf{1}_N\mu_j - \boldsymbol{\delta}\lambda_j) + \beta\}/2$.
5. Sample the spatial correlation parameter (here a) depending on the specification of $\boldsymbol{\Psi}$. The prior distribution of a is $\pi(a)$. Draw a' from the autoregressive proposal density $q(a'|a) \sim \mathcal{N}(a, \nu^2)$, where ν^2 is a tuning parameter, and accept the new draw, a' with probability $\min\{1, f(\boldsymbol{\delta}|\boldsymbol{\Psi}')\pi(a')q(a|a')/f(\boldsymbol{\delta}|\boldsymbol{\Psi})\pi(a)q(a'|a)\}$, where $f(\boldsymbol{\delta}|\boldsymbol{\Psi}^{(\cdot)})$ is the kernel of the distribution of $\boldsymbol{\delta}$ conditional on $\boldsymbol{\Psi}^{(\cdot)}$ from level II.

   For model 2, $\boldsymbol{\Psi} = \exp(-a\mathbf{W}^d)$, $\mathbf{W}_{ij}^d = \{d_{ij}\}$, and $\pi(a) = \mathcal{N}(z, V_\zeta)I(\zeta > 0)$. For model 3A, $\boldsymbol{\Psi} = \mathbf{A}^{-1/2}(\mathbf{I} - a\mathbf{R})^{-1}\mathbf{A}^{-1/2}$, $\mathbf{A} = \text{diag}\{(\mathbf{I} - a\mathbf{R})^{-1}\}$, and $\pi(a) = N(0, V_a)I(a \in r)$, where $r = (\xi_1^{-1}, \xi_N^{-1})$. In model 3B, $\boldsymbol{\Psi} = \mathbf{A}^{-1/2}(\mathbf{O} - a\mathbf{W})^{-1}\mathbf{A}^{-1/2}$, where $\mathbf{O} = \text{diag}(n_i)$, $W_{ij} = (n_i n_j)^{1/2}I(j \in \mathcal{R}_i)$, $\mathbf{A} = \text{diag}\{(\mathbf{O} - a\mathbf{W})^{-1}\}$, and $\pi(a) = N(0, V_a)I(a \in r)$, where $r = (\xi_1^{-1}, \xi_N^{-1})$. For model 3C, $\boldsymbol{\Psi} = (\mathbf{O} - \mathbf{W}^d)^{-1}$, $W_{ij}^d = \exp(-ad_{ij})$, $\mathbf{O} = \text{diag}\{\sum\exp(-ad_{ij})\}$, and $\pi(a) = N(0, V_a) \times I(a > 0)$. For model 3D, $\boldsymbol{\Psi} = \mathbf{A}^{-1/2}(\mathbf{I} - a\mathbf{W})^{-1}\mathbf{A}^{-1/2}$, $W_{ij} = \exp(-d_{ij}\gamma)I(j \in \mathcal{R}_i)$, $\mathbf{A} = \text{diag}\{(\mathbf{I} - a\mathbf{W})^{-1}\}$, γ is fixed, $\pi(a) = \mathcal{N}(0, V_a)I(a \in r)$, and $r = (\xi_1^{-1}, \xi_N^{-1})$. For model 3E, $\boldsymbol{\Psi} = \mathbf{A}^{-1/2}(\mathbf{O} - a\mathbf{W})^{-1}\mathbf{A}^{-1/2}$, $\mathbf{O} = \text{diag}\{\sum\exp(-d_{ij}\gamma) \times I(j \in \mathcal{R}_i)\}$, $\mathbf{W}_{ij} = \exp(-d_{ij}\gamma)I(j \in \mathcal{R}_i)$, $\mathbf{A} = \text{diag}\{(\mathbf{O} - a\mathbf{W})^{-1}\}$, γ is fixed, $\pi(a) = \mathcal{N}(0, V_a)I(a \in r)$, and $r = (-1, 1)$.

[Received December 2002. Revised December 2003.]

REFERENCES

Aguilar, O., and West, M. (2000), "Bayesian Dynamic Factor Models and Portfolio Allocation," *Journal of Business & Economic Statistics*, 18, 338–357.

Barnett, S., Roderick, P., Martin, D., Diamond, I., and Wrigley, H. (2002), "Interrelations Between Three Proxies of Health Care Need at the Small Area Level: An Urban/Rural Comparison," *Journal of Epidemiology and Community Health*, 56, 754–761.

Bartholomew, D. J., and Knott, M. (1999), *Latent Variable Models and Factor Analysis* (2nd ed.), New York: Oxford University Press.

Bernardinelli, L., Clayton, D., and Montomoli, C. (1995), "Bayesian Estimates of Disease Maps: How Important Are Priors?" *Statistics in Medicine*, 14, 2411–2431.

Besag, J. (1974), "Spatial Interaction and the Statistical Analysis of Lattice Systems" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 35, 192–236.

Best, N. G., Arnold, R. A., Thomas, A., Waller, L. A., and Conlon, E. M. (1999), "Bayesian Models for Spatially Correlated Disease and Exposure Data," in *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Clarendon Press, pp. 131–156.

Carstairs, V. (1995), "Deprivation Indices: Their Interpretation and Use in Relation to Health," *Journal of Epidemiology and Community Health*, 49 (Suppl. 2), S3–S8.

Christensen, W. F., and Amemiya, Y. (2002), "Latent Variable Analysis of Multivariate Spatial Data," *Journal of the American Statistical Association*, 97, 302–317.

——— (2003), "Modeling and Prediction for Multivariate Spatial Factor Analysis," *Journal of Statistical Planning and Inference*, 115, 543–564.

Conlon, E. M., and Waller, L. A. (1998), "Flexible Neighborhood Structures in Hierarchical Models for Disease Mapping," Report 1998-018, University of Minnesota, Division of Biostatistics.

Cressie, N. (1993), *Statistics for Spatial Data* (rev. ed.), Somerset, NJ: Wiley-Interscience.

Cressie, N., and Chan, N. H. (1989), "Spatial Modeling of Regional Variables," *Journal of the American Statistical Association*, 84, 393–401.

Eachus, J., Williams, M., Chan, P., Smith, G. D., Grainge, M., Donovan, J., and Frankel, S. (1996), "Deprivation and Cause-Specific Morbidity: Evidence From the Somerset and Avon Survey of Health," *British Medical Journal*, 312, 287–292.

Gelfand, A. E., and Ghosh, S. K. (1998), "Model Choice: A Minimum Posterior Predictive Loss Approach," *Biometrika*, 85, 1–11.

Geweke, J., and Zhou, G. (1996), "Measuring the Pricing Error of the Arbitrage Pricing Theory," *The Review of Financial Studies*, 9, 557–587.

Hutchinson, A., Foy, C., and Sandhu, B. (1989), "Comparison of Two Scores for Allocating Resources to Doctors in Deprived Areas," *British Medical Journal*, 299, 1142–1144.

Jarman, B. (1983), "Identification of Underprivileged Areas," *British Medical Journal*, 286, 1705–1709.

Krieger, N., Chen, J. T., Waterman, P. D., Soobader, M. J., Subramanian, S. V., and Carson, R. (2002), "Geocoding and Monitoring of U.S. Socioeconomic Inequalities in Mortality and Cancer Incidence: Does the Choice of Area-Based Measure and Geographic Level Matter?" *American Journal of Epidemiology*, 156, 471–482.

Krieger, N., Zierler, S., Hogan, J. W., Waterman, P., Chen, J., Lemieux, K., and Gjelsvik, A. (2001), "Geocoding and Measurement of Neighborhood Socioeconomic Position: A U.S. Perspective," in *Neighborhoods and Health*, ed. I. Kawachi, Oxford, U.K.: Oxford University Press, pp. 147–178.

Laird, N. M., and Louis, T. A. (1989), "Empirical Bayes Ranking Methods," *Journal of Educational Statistics*, 14, 29–46.

Lee, P., Murie, A., and Gordon, D. (1995), "Area Measures of Deprivation: A Study of Current Methods and Best Practices in the Identification of Poor Areas in Great Britain," Centre for Urban and Regional Studies, University of Birmingham, U.K.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis With Missing Data*, New York: Wiley.

Malmström, M., Sundquist, J., Bajekal, M., and Johansson, S. E. (1998), "Indices of Need and Social Deprivation for Primary Health Care," *Scandinavian Journal of Social Medicine*, 26, 124–130.

Press, S. J., and Shigemasu, K. (1997), "Bayesian Inference in Factor Analysis (Revised)," Technical Report 243, University of California-Riverside, Dept. of Statistics.

Raudenbush, S. W., and Sampson, R. (1999), "Assessing Direct and Indirect Effects in Multilevel Designs With Latent Variables," *Sociological Methods & Research*, 28, 123–153.

Rowe, D. B. (2002), *Multivariate Bayesian Statistics: Models for Source Separation and Signal Unmixing*, Boca Raton, FL: CRC Press.

Samet, J. M., Dominici, F., Curriero, F., Coursac, I., and Zeger, S. L. (2000), "Particulate Air Pollution and Mortality: Findings From 20 U.S. Cities," *New England Journal of Medicine*, 343, 1742–1757.

Sammel, M. D., and Ryan, L. M. (1996), "Latent Variable Models With Fixed Effects," *Biometrics*, 52, 650–663.

Sheldon, T. A., Davey-Smith, G., and Bevan, G. (1993), "Weighting in the Dark: Resource Allocation in the New NHS," *British Medical Journal*, 306, 835–839.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 64, 583–639.

Stern, H., and Cressie, N. (1999), "Inference for Extremes in Disease Mapping," in *Disease Mapping and Risk Assessment for Public Health*, ed. A. B. Lawson, New York: Wiley, pp. 63–84.

Sun, D., Tsutakawa, R. K., and Speckman, P. L. (1999), "Posterior Distribution of Hierarchical Models Using CAR(1) Distributions," *Biometrika*, 86, 341–350.

Sundquist, J., Malmström, M., and Johansson, S. E. (1999), "Cardiovascular Risk Factors and the Neighbourhood Environment: A Multilevel Analysis," *International Journal of Epidemiology*, 28, 841–845.

Townsend, P., Simpson, D., and Tibbs, N. (1985), "Inequalities in Health in the City of Bristol: A Preliminary Review of Statistical Evidence," *International Journal of Health Services*, 15, 637–663.

U.S. Census Bureau (1999), "1990 Census of Population and Housing–CPH-4," available at *http://www.census.gov/prod/1/90dec/cph4/90cph4.htm*, accessed September 19, 2003, App. C.

Waller, L. A., Carlin, B. P., Xia, H., and Gelfand, A. E. (1997), "Hierarchical Spatio-Temporal Mapping of Disease Rates," *Journal of the American Statistical Association*, 92, 607–617.

Wang, F., and Wall, M. M. (2001), "Modelling Multivariate Data With a Common Spatial Factor," Report 2001-008, University of Minnesota, Division of Biostatistics.

——— (2003), "Generalized Common Spatial Factor Model," *Biostatistics*, 4, 569–582.

Wilson, K.C., Chen, R., Taylor, S., McCracken, C. F., and Copeland, J. R. (1999), "Socio-Economic Deprivation and the Prevalence and Prediction of Depression in Older Community Residents: The MRC-ALPHA Study," *British Journal of Psychiatry*, 175, 549–553.