A joint spatial factor analysis model to accommodate data from misaligned areal units with application to Louisiana social vulnerability

RACHEL C. NETHERY

*Department of Biostatistics, Harvard University, T.H. Chan School of Public Health,
677 Huntington Ave, Boston, MA 02115, USA*

DALE P. SANDLER

*Epidemiology Branch, National Institute of Environmental Health Sciences,
111 TW Alexander Dr, Durham, NC 27709, USA*

SHANSHAN ZHAO

*Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences,
111 TW Alexander Dr, Durham, NC 27709, USA*

LAWRENCE S. ENGEL

*Epidemiology Branch, National Institute of Environmental Health Sciences, 111 TW Alexander Dr,
Durham, NC 27709, USA and Department of Epidemiology, University of North Carolina at Chapel
Hill, Gillings School of Global Public Health, 135 Dauer Dr, Chapel Hill, NC 27599, USA*

RICHARD K. KWOK*

*Epidemiology Branch, National Institute of Environmental Health Sciences,
111 TW Alexander Dr, Durham, NC 27709, USA
richard.kwok@nih.gov*

SUMMARY

With the threat of climate change looming, the public health community has an interest in identifying communities at the highest risk of devastation based not only on geographic features but also on social characteristics. Indices of community social vulnerability can be created by applying a spatial factor analysis to a set of relevant social variables measured for each community; however, current spatial factor analysis methodology is ill-equipped to handle spatially misaligned data. We introduce a joint spatial factor analysis model that can accommodate spatial data from two distinct partitions of a geographic space and identify a common set of latent factors underlying them. By defining the latent factors over the intersection of the two partitions, the model minimizes loss of information. Using simulated data constructed to mimic the spatial structure of our real data, we confirm the reliability of the model and demonstrate its superiority over competing *ad hoc* methods for dealing with misaligned data in spatial factor analysis. Finally, we

*To whom correspondence should be addressed.

construct an index of community social vulnerability for each census tract in Louisiana, a state prone to environmental disasters, which could be exacerbated by climate change, by applying the joint spatial factor analysis model to a set of misaligned social indicator data from the state. To demonstrate the utility of this index, we integrate it with Louisiana flood insurance claims data to identify communities that may be at particularly high risk during natural disasters, based on both social and geographic features.

Keywords: Climate change; Louisiana; Social indicators; Social vulnerability; Spatial factor analysis; Spatial misalignment.

1. INTRODUCTION

In recent years, advances in global positioning system and geographic information system technologies have simplified the process of collecting and analyzing spatially referenced data. The consequent explosion of spatially based research led to an increased capacity to holistically characterize places and communities. The assessment of social indicators across space is a topic that has a long and rich academic history (Taylor and Hudson, 1970; Smith, 1973; Duncan, 1974; Smith, 1981), but, recently, the mining of enormous quantities of data on social indicators has propelled this topic beyond the realm of the purely academic (Cutter and others, 2003). One obstacle to the analysis of trends in social indicators across space is the need to accommodate spatially referenced variables collected at differing spatial levels—a problem known as spatial misalignment (Banerjee and others, 2014).

Much of the academic community's recent interest in social indicators has focused on quantifying the “social vulnerability” of places/communities to climate change and natural disasters (Cutter and others, 2003). Because the extent to which a community is able to prepare for and recover from disasters is largely determined by social factors, socially vulnerable areas may be more severely impacted; thus, identification of these areas is critical to disaster preparation (Cutter and others, 2003). Several groups have developed indices of social resilience/vulnerability to natural disasters and environmental changes and have estimated them across the US (Cutter and others, 2003, 2008; Cutter and Finch, 2008).

Community social vulnerability is difficult to measure directly, but an abundance of social indicator variables are readily available, many of which are highly correlated thanks to their common association with this broader concept of social vulnerability. Thus, an index of social vulnerability can be constructed as a latent factor (or set of latent factors) underlying a relevant set of observed social indicator variables, through the use of factor analysis or principal component analysis (Cutter and others, 2003; Cutter and Finch, 2008). The standard factor analysis model, however, is potentially inappropriate because it fails to properly account for the spatial correlation that is likely present in spatially referenced data. We propose instead using spatial factor analysis to create social vulnerability indices, and we extend spatial factor analysis to address the issue of spatially misaligned data.

When the units of study in an analysis are spatially or geographically defined, units that are closer together are likely to be more similar than units further apart, leading to spatial correlation. In this case, spatial factor analysis (Wang and Wall, 2003) should be used instead of the standard factor analysis to appropriately account for the spatial correlation and improve estimation and model fit (Wang and Wall, 2003; Nethery and others, 2015). The spatial factor analysis model takes the form

$$Y(S_i) = \Lambda \eta(S_i) + \epsilon(S_i), \quad (1.1)$$

where $Y(S_i)$ is the $p \times 1$ vector of continuous, observed variables for the i^{th} geographic location/region, denoted S_i , for $i = 1, \dots, N$ (Banerjee and others, 2014). Letting m be a prespecified number of latent factors ($m \ll p$), $\eta(S_i)$ is the $m \times 1$ vector of latent factor scores for S_i , Λ represents the $p \times m$ matrix of factor loadings, and $\epsilon(S_i) \stackrel{\text{i.i.d.}}{\sim} \text{MVN}(\mathbf{0}, \Sigma)$ represents the vector of errors for S_i . Σ is a diagonal matrix

with $(j, j)^{th}$ entry equal to σ_j^2 , and $\mathbf{\Lambda}$ is constrained to be lower triangular with diagonal entries $\lambda_{jj} > 0$ for identifiability purposes (Bollen, 1989). (Note that, with this constraint, the factors are still only identified up to a permutation and sign change.) After applying spatial factor analysis to the observed data, the predicted latent factors quantify the latent constructs underlying the observed data, and, in our setting, are used to construct community social vulnerability scores.

Herein, we focus on a Bayesian approach. Spatial factor analysis methodology in the Bayesian setting has been extended to accommodate various data types and analysis goals and has been applied to a wide range of problems, including the reduction of social indicator data (Rowe, 1998; Wang and Wall, 2003; Hogan and Tchernis, 2004; Liu and others, 2005; Lopes and others, 2008; Mezzetti, 2012; Stakhovych and others, 2012; Nethery and others, 2015). The standard Bayesian factor analysis prior distribution specifications, which lead to semi-conjugacy, are discussed by Rowe (1998) and Ghosh and Dunson (2009). In the spatial factor analysis model, spatial correlation is introduced in the factor scores through the prior distribution placed on $\boldsymbol{\eta} = [\boldsymbol{\eta}(S_1)' \cdots \boldsymbol{\eta}(S_N)']'$ (Wang and Wall, 2003). The prior distribution has the form $\boldsymbol{\eta} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_S \otimes \mathbf{I}_m)$, where $\boldsymbol{\Sigma}_S$ is the $N \times N$ spatial covariance matrix, containing a spatial parameter, ϕ (Wang and Wall, 2003; Banerjee and others, 2014). Throughout this article, we assume a conditionally auto-regressive covariance structure, but other spatial covariance structures can easily be accommodated. For more details, see Banerjee and others (2014).

Although the Bayesian spatial factor analysis model allows for a good deal of flexibility and a wide range of spatial correlation structures, it has not yet been extended to accommodate spatial misalignment, a common obstacle to the analysis of social indicator data. Henceforth, we will consider a single type of spatial data, known as areal data. Areal data are counts or averages of a measure collected over geographically defined regions formed by the partitioning of the area of interest. Areal social indicator data are often spatially misaligned (Cressie, 1996), as indicators may be collected by different organizations over distinct geographic partitions that correspond to each organization's specific goals. Social vulnerability research, due to its reliance on areal social indicator data, may be dramatically impacted by spatial misalignment.

Our objective is to assess the social vulnerability of census tracts. Census tracts are areal units defined by the US Census Bureau that contain approximately 2500–8000 residents and are subsets of counties (US Bureau of the Census, 1994). Although many social vulnerability indices are developed at the county level (Cutter and others, 2003; Cutter and Finch, 2008), effects are commonly obscured at this level, particularly for counties that contain large cities, where many communities with different vulnerability levels may be grouped together. Thus, performing inference over smaller regions with more homogeneity in population size, such as census tracts, may provide more insight.

In pursuit of this goal, we wish to create a general social vulnerability index (i.e., not specific to any one type of event/disaster) by analyzing a set of social indicators whose associations with social vulnerability are widely agreed upon in the literature, e.g., socioeconomic and demographic variables (Diener and Suh, 1997; Cutter and others, 2003; Cutter and Finch, 2008; Agency for Toxic Substances and Disease Registry, 2014). However, because crime is known to be an important component of community vulnerability to disasters (Perdikaris, 2014), we also hope to improve on existing indices by including crime data in the construction of our index. The US Census Bureau collects socioeconomic and demographic data for each census tract in the US and makes these data publicly available (US Bureau of the Census, 1994). The US Federal Bureau of Investigation (FBI) makes crime data publicly available for each county in the US through its Uniform Crime Report (UCR) (US Federal Bureau of Investigation, 2010), but crime data are not consistently collected at any finer spatial level. To use both the census tract level socioeconomic and demographic data and the county level crime data together in a spatial factor analysis to create social vulnerability scores for each census tract, it is imperative to address the issue of spatial misalignment.

While there are limited instances of spatial misalignment addressed specifically in the context of spatial factor analysis, the topic of spatial misalignment in general has received a great deal of attention (Gotway and Young, 2002; Trevisani and Gelfand, 2013). Some of the general techniques for handling

misalignment can be trivially extended to the spatial factor analysis setting. For instance, one approach to resolving misalignment, which we call pre-analysis data alignment, is to align the data to a common set of areal units prior to application of a statistical model (Banerjee and others, 2014). Most pre-analysis alignment methods rely on the assumption that the distribution of each variable is constant across some partition of the units over which it is observed. In our application, we would need to assume the constancy of the crime variables across census tracts within a given county, which is likely to be an unreasonable assumption.

In this article, we propose a joint spatial factor analysis model in the Bayesian setting that can accommodate a set of spatially referenced variables recorded at misaligned areal units. The model identifies and predicts a common set of latent factors underlying all the data from two levels of areal units by sharing information between spatial factor analysis models constructed for each spatial level. In Section 2, we introduce the joint model using what is known as the “top-down approach” for multi-scale spatial analyses (Bradley and others, 2015), and we present the necessary assumptions and the recommended approach to model fitting. Although the general multi-scale form of the model is introduced for completeness, we focus the remainder of the paper on the nested setting, i.e., the case when one set of areal units is fully contained within the other, as motivated by our real data application. We test our method on simulated data in Section 3 and compare those results to results from the pre-analysis alignment method. In Section 4, we apply the joint model to a set of misaligned social indicator data from the state of Louisiana to create a general use, data driven social vulnerability index, and we combine it with Louisiana flood vulnerability data to identify the highest risk communities in a region prone to natural disasters. Finally, we discuss results and possible extensions in Section 5.

2. METHODS

2.1. The general multi-scale joint spatial factor analysis model

In the “top-down approach” to spatial multi-scale analyses, models are defined using a partition of the region of support of the data. The alternative “bottom-up approach” defines models at the point level and integrates up to the partition desired for inference (Bradley and others, 2015). Although these approaches have been shown to perform comparably, we have elected to present our model using the top-down approach, because many of the variables used in our social vulnerability application are not interpretable at the point level (e.g., median household income).

Consider two partitions of the region of support, $A = \{A_1, \dots, A_J\}$ and $B = \{B_1, \dots, B_K\}$, corresponding to the two sets of areal units at which variables are observed. In particular, assume a set of p_1 variables, denoted by $\mathbf{Y}$, is recorded over A and a set of p_2 variables, denoted by $\mathbf{Z}$, is recorded over B . Now define a partition at the intersection of A and B , $S = \{S_1, \dots, S_N\}$. The areal units in this intersection partition are sometimes referred to as the “atom” (Trevisani and Gelfand, 2013). Our goal is to construct spatial factor analysis models that relate the observed variables across A and B to a common set of spatially correlated latent factors across S . In other words, we want to write both $\mathbf{Y}(A_j)$ and $\mathbf{Z}(B_k)$ in terms of latent factors $\boldsymbol{\eta}(S_i)$, $j = 1, \dots, J$, $k = 1, \dots, K$, $i = 1, \dots, N$. Then these models can be fit jointly.

To construct models relating $\mathbf{Y}(A_j)$ and $\mathbf{Z}(B_k)$ to $\boldsymbol{\eta}(S_i)$, we rely on the assumption that the values of the observed variables are weighted sums of their values at all the atom units, i.e.,

$$\begin{aligned} \mathbf{Y}(A_j) &= \sum_{S_i \in A_j} \mathbf{W}_{ji}^Y \boldsymbol{\eta}(S_i) \\ \mathbf{Z}(B_k) &= \sum_{S_i \in B_k} \mathbf{W}_{ki}^Z \boldsymbol{\eta}(S_i) \end{aligned} \tag{2.1}$$

for $j = 1, \dots, J, k = 1, \dots, K, i = 1, \dots, N$, where $\mathbf{Y}(S_i)$ is a $p_1 \times 1$ vector of the possibly unobserved values of the variables in $\mathbf{Y}$ at S_i and $\mathbf{Z}(S_i)$ is a $p_2 \times 1$ vector of the possibly unobserved values of the variables in $\mathbf{Z}$ at S_i . $\mathbf{W}_{ji}^Y$ and $\mathbf{W}_{ki}^Z$ are diagonal matrices of size $p_1 \times p_1$ and $p_2 \times p_2$, respectively, with known weights on the diagonals. For instance, the $(q, q)^{th}$ component of $\mathbf{W}_{ji}^Y$ might be assumed to be 1 if the q^{th} variable in $\mathbf{Y}$ is a sum/count or to be the proportion of the land area or population in A_j that is contained in S_i if the q^{th} variable is an average/rate (and analogously for $\mathbf{W}_{ki}^Z$). This assumption can also be written in matrix form as

$$\begin{aligned} \mathbf{Y}(A_j) &= \mathbf{W}_j^Y \begin{bmatrix} \mathbf{Y}(S_1) \\ \vdots \\ \mathbf{Y}(S_N) \end{bmatrix} \\ \mathbf{Z}(B_k) &= \mathbf{W}_k^Z \begin{bmatrix} \mathbf{Z}(S_1) \\ \vdots \\ \mathbf{Z}(S_N) \end{bmatrix} \end{aligned} \quad (2.2)$$

where $\mathbf{W}_j^Y$ is a block matrix such that $\mathbf{W}_j^Y = [\mathbf{W}_{j1}^Y \dots \mathbf{W}_{jN}^Y]$ with $\mathbf{W}_{ji}^Y$ is the $p_1 \times p_1$ diagonal weight matrix defined above if $S_i \in A_j$ and $\mathbf{W}_{ji}^Y = \mathbf{0}_{p_1 \times p_1}$ otherwise. $\mathbf{W}_k^Z$ is defined analogously.

Now, if the $\mathbf{Y}(S_i)$ and $\mathbf{Z}(S_i)$ were observed, we could fit a spatial factor analysis model at the S_i as in equation 1.1:

$$\begin{bmatrix} \mathbf{Y}(S_i) \\ \mathbf{Z}(S_i) \end{bmatrix} = \begin{bmatrix} \mathbf{\Lambda}_1 \\ \mathbf{\Lambda}_2 \end{bmatrix} \boldsymbol{\eta}(S_i) + \begin{bmatrix} \boldsymbol{\epsilon}_1(S_i) \\ \boldsymbol{\epsilon}_2(S_i) \end{bmatrix} \quad (2.3)$$

where $\mathbf{\Lambda}_1$ is a $p_1 \times m$ matrix of factor loadings corresponding to the variables in $\mathbf{Y}$, $\mathbf{\Lambda}_2$ is a $p_2 \times m$ matrix of factor loadings corresponding to the variables in $\mathbf{Z}$, $\boldsymbol{\eta}(S_i)$ is an $m \times 1$ vector of the common factors at location S_i (these factors are correlated across locations), and $\boldsymbol{\epsilon}_1(S_i)$ and $\boldsymbol{\epsilon}_2(S_i)$ are location-specific error vectors for the variables in $\mathbf{Y}$ and $\mathbf{Z}$, respectively. It is assumed that $\boldsymbol{\epsilon}_1(S_i) \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_1)$ and $\boldsymbol{\epsilon}_2(S_i) \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_2)$.

We now return to the construction of our two models. We must use the model specification in equation 2.3 combined with our weighted sum assumption in equation 2.1 to write

$$\begin{aligned} \mathbf{Y}(A_j) &= \sum_{S_i \in A_j} \mathbf{W}_{ji}^Y \mathbf{Y}(S_i) = \sum_{S_i \in A_j} \mathbf{W}_{ji}^Y \mathbf{\Lambda}_1 \boldsymbol{\eta}(S_i) + \sum_{S_i \in A_j} \mathbf{W}_{ji}^Y \boldsymbol{\epsilon}_1(S_i) \\ \mathbf{Z}(B_k) &= \sum_{S_i \in B_k} \mathbf{W}_{ki}^Z \mathbf{Z}(S_i) = \sum_{S_i \in B_k} \mathbf{W}_{ki}^Z \mathbf{\Lambda}_2 \boldsymbol{\eta}(S_i) + \sum_{S_i \in B_k} \mathbf{W}_{ki}^Z \boldsymbol{\epsilon}_2(S_i). \end{aligned} \quad (2.4)$$

In this way, $\mathbf{Y}(A_j)$ and $\mathbf{Z}(B_k)$ can be modeled in terms of a common set of latent factors at the S_i , allowing us to specify our joint model.

Borrowing notation from equation 2.3, the two models can be written separately as
Model 1:

$$\mathbf{Y}(A_j) = \sum_{S_i \in A_j} \mathbf{W}_{ji}^Y \mathbf{\Lambda}_1 \boldsymbol{\eta}(S_i) + \boldsymbol{\epsilon}_1^*(A_j) \quad (2.5)$$

Model 2:

$$\mathbf{Z}(B_k) = \sum_{S_i \in B_k} \mathbf{W}_{ki}^Z \mathbf{\Lambda}_2 \boldsymbol{\eta}(S_i) + \boldsymbol{\epsilon}_2^*(B_k) \quad (2.6)$$

where $\boldsymbol{\epsilon}_1^*(A_j) = \sum_{S_i \in A_j} \mathbf{W}_{ji}^Y \boldsymbol{\epsilon}_1(S_i)$, $\boldsymbol{\epsilon}_2^*(B_k) = \sum_{S_i \in B_k} \mathbf{W}_{ki}^Z \boldsymbol{\epsilon}_2(S_i)$, $\boldsymbol{\epsilon}_1^*(A_j) \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_1^*)$, and $\boldsymbol{\epsilon}_2^*(B_k) \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_2^*)$. Using the matrix notation for the weighting assumption introduced in equation 2.2, these models can be written jointly in the following way:

$$\begin{bmatrix} \mathbf{Y}(A_1) \\ \vdots \\ \mathbf{Y}(A_J) \\ \mathbf{Z}(B_1) \\ \vdots \\ \mathbf{Z}(B_K) \end{bmatrix} = \begin{bmatrix} \mathbf{W}_1^Y \\ \vdots \\ \mathbf{W}_J^Y \\ \mathbf{W}_1^Z \\ \vdots \\ \mathbf{W}_K^Z \end{bmatrix} \begin{bmatrix} \mathbf{I}_N \otimes \mathbf{\Lambda}_1 \\ \mathbf{I}_N \otimes \mathbf{\Lambda}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\eta}(S_1) \\ \vdots \\ \boldsymbol{\eta}(S_n) \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_1^*(A_1) \\ \vdots \\ \boldsymbol{\epsilon}_1^*(A_J) \\ \boldsymbol{\epsilon}_2^*(B_1) \\ \vdots \\ \boldsymbol{\epsilon}_2^*(B_K) \end{bmatrix}. \quad (2.7)$$

To finalize the model structure, the number of latent factors, m , must be specified. Although model selection techniques could be integrated into the estimation procedure, we forego the discussion of such an extension here and assume that the user has specified m in advance. Methods to guide the choice of m are discussed in Section E of the [supplementary material](#) available at *Biostatistics* online. Having specified m , a joint likelihood may be specified and prior distributions assigned. As in the classic spatial factor analysis model, this model allows for spatial correlation in the factor scores through a spatial covariance matrix in the prior distribution for the factors, with spatial parameter ϕ embedded in this matrix. Standard Markov Chain Monte Carlo (MCMC) techniques can be employed for the joint model estimation. By estimating the model jointly, the data from both sets of areal units inform the construction of the latent factors. Posterior summaries for each parameter and for the factor scores can be computed from the MCMC output.

While the joint spatial factor analysis model can easily be formulated for general areal misalignment settings, the most judicious application of this model is in the special case of nested areal units. When the misaligned units are non-nested, the latent factors are being predicted over a partition on which none of the data are directly observed; therefore, obtaining model convergence may be difficult. Moreover, a great deal of tuning or the use of informative prior distributions may be needed in order to obtain reliable inference for the S_i . Our motivating application involves nested data, for which some variables are observed on the intersection partition, and we do not wish to get sidetracked with performance issues specific to the non-nested case. Thus, we turn our attention to the special case of nested areal units for the remainder of this paper and leave further investigation into the non-nested misalignment setting for future work.

2.2. The nested areal units case

In nested areal misalignment, each areal unit of one partition is fully contained within a single unit of the other partition. Using the notation above, the nested case corresponds to $A \equiv S$ or, for all j , $A_j \subseteq B_k$ for some k . In our application, some data are recorded at census tracts and some data are recorded for counties, and census tracts (the S and A) are fully nested within counties (the B). Henceforth, for clarity, we will refer to the S as the small areal units (SAUs) and the B as the large areal units (LAUs).

In this case, our two models are
Model 1:

$$\mathbf{Y}(S_i) = \mathbf{\Lambda}_1 \boldsymbol{\eta}(S_i) + \boldsymbol{\epsilon}_1(S_i) \quad (2.8)$$

Model 2:

$$\mathbf{Z}(B_k) = \sum_{S_i \in B_k} \mathbf{W}_{ki}^Z \mathbf{\Lambda}_2 \boldsymbol{\eta}(S_i) + \boldsymbol{\epsilon}_2^*(B_k). \quad (2.9)$$

The likelihood can be specified jointly, as in Section 2.1. We provide this joint likelihood and the recommended prior distributions for Bayesian estimation in [Appendices A.1 and A.2](#) in the [supplementary material](#) available at *Biostatistics* online, respectively. MCMC samples may be drawn from the full conditional distribution of each parameter using a Gibbs sampler with a Metropolis step ([Metropolis and others, 1953](#); [Hastings, 1970](#); [Geman and Geman, 1984](#); [Gelfand and Smith, 1990](#)). The full conditional distributions and sampling algorithm corresponding to the recommended prior distributions are provided in [Appendix A.3](#) in the [supplementary material](#) available at *Biostatistics* online. Note that, in the nested case, this model automatically provides factor scores at the SAU level, which corresponds to our desired census tract level social vulnerability scores.

3. SIMULATION STUDY

3.1. Assessment of joint model performance

Simulations are carried out using R ([R Core Team, 2014](#)) and MATLAB ([The MathWorks Inc., 2015](#)) statistical software. Misaligned, nested areal data are simulated from a lattice with 125 LAUs and 625 nested SAUs, with each LAU containing exactly 5 SAUs, as shown in [Figure S1](#) of the [supplementary material](#) available at *Biostatistics* online ($K = 125$, $N = 625$). We construct this lattice so that the simulated data reflect anticipated real data, with many fewer LAUs than SAUs. A single latent factor is simulated over the SAUs from a normal distribution, with spatial correlation introduced by assigning a covariance of 0.15 to the factor scores of any pairs of SAUs sharing a boundary ($m = 1$, $\phi = -0.15$).

Define λ_{1k} as the k^{th} element of $\mathbf{\Lambda}_1$ (note that with $m = 1$ $\mathbf{\Lambda}_1$ has dimensions $p_1 \times 1$), λ_{2k} as the k^{th} element of $\mathbf{\Lambda}_2$, σ_{1k}^2 as the element in the (k, k) position of $\mathbf{\Sigma}_1$, and σ_{2k}^2 as the element in the (k, k) position of $\mathbf{\Sigma}_2^*$. The simulated latent factor and fixed loadings, along with randomly generated error terms from normal distributions with fixed variances ([Table 1](#)), are used to generate six observed variables at the SAUs ($p_1 = 6$), by plugging into equation 2.8. Six distinct observed variables are generated at the LAUs ($p_2 = 6$) by plugging the latent factor, fixed loadings, randomly generated errors from normal distributions, and a common set of weights, $\mathbf{W}_k^Z = \text{diag}([.2, .15, .3, .1, .25])$ into equation 2.9 ([Table 1](#)). This process is repeated 1000 times to generate 1000 simulated misaligned datasets. We apply the Gibbs sampler described above, with hyperparameters chosen to create non-informative prior distributions, to each dataset for 100 000 iterations. We discard the first 50 000 samples as burn-in and retain 50 000.

Our primary interest is in accurate prediction of the latent factor scores, as these will form the social vulnerability index. Across the 1000 simulations, the average correlation between the true and predicted factor scores is 0.99, and the average coverage rate of the 95% credible interval for each factor score across all simulations is 0.93, indicating that this model performs extremely well for our purposes. Average parameter estimates and 95% credible interval coverage rates are provided in [Table 1](#). The results of additional simulations with different levels of spatial correlation are presented in [Section C](#) of the [supplementary material](#) available at *Biostatistics* online. All parameters are estimated consistently with credible interval coverage very close to the desired 95%, with the exception of the parameters impacted by the identifiability constraint, i.e., λ_{11} , λ_{21} , σ_{11}^2 , σ_{21}^2 , whose estimation suffer somewhat due to this constraint (note that such performance issues are common to any factor analysis procedure that utilizes identifiability constraints, which is standard practice in the Bayesian setting). However, extensive investigation has revealed that the sub-optimal performance of these parameter estimates has little, if any, impact on the performance of the other parameter estimates and the factor score predictions in a single factor model.

Table 1. True values for all parameters in the misaligned data simulation and average posterior means (PM) and 95% credible interval coverage rates from the joint model across 1000 simulations

	Truth	Average PM	Coverage
λ_{11}	0.50	0.40	0.78
λ_{12}	0.14	0.14	0.95
λ_{13}	-0.63	-0.63	1.00
λ_{14}	1.20	1.21	0.99
λ_{15}	0.25	0.25	0.95
λ_{16}	-0.62	-0.62	0.98
λ_{21}	0.37	0.30	0.74
λ_{22}	0.55	0.56	0.94
λ_{23}	0.43	0.44	0.96
λ_{24}	-0.23	-0.23	0.95
λ_{25}	1.13	1.16	0.95
λ_{26}	0.29	0.30	0.95
σ_{11}^2	0.27	0.32	0.75
σ_{12}^2	0.39	0.39	0.94
σ_{13}^2	0.01	0.01	0.96
σ_{14}^2	0.38	0.39	0.95
σ_{15}^2	0.87	0.87	0.95
σ_{16}^2	0.34	0.34	0.95
σ_{21}^2	0.48	0.50	0.94
σ_{22}^2	0.60	0.61	0.95
σ_{23}^2	0.49	0.50	0.95
σ_{24}^2	0.19	0.19	0.94
σ_{25}^2	0.83	0.84	0.94
σ_{26}^2	0.67	0.68	0.94
ϕ	-0.15	-0.11	0.96

While the effects appear to be minor, it may be wise to order the variables such that the first variable at each spatial level is one not believed to carry great importance or to run the model with different variable orderings to ensure results are robust. Greater caution should be exercised in the multi-factor setting, where the constraints are more stringent and are more likely to impact factor identification. Further discussion of this issue and the results of a small simulation study with two latent factors ($m = 2$) are provided in [Section D](#) of the [supplementary material](#) available at *Biostatistics* online.

This joint model works by sharing information from both the SAU variables and the LAU variables to predict the values of the common latent factor(s). The unique parameters for each model are estimated using the data corresponding to that model as well as these predicted factor values; thus, the parameter estimation also incorporates information from both the SAU and LAU data. Given that the SAU data will provide the most direct information about the characteristics of the SAUs and, thereby, the appropriate latent factor score for each SAU, a pertinent question for users of this method might be related to model performance when the number of SAU variables is small. To test this performance, we conduct 1000 simulations with misaligned data constructed using the same procedure as described above, except with $p_1 = 3$ rather than $p_1 = 6$. While the model identifies the correct factor and performs comparably to the previous simulation in parameter estimation, precision in the latent factor prediction is reduced, as expected when less information is available at the SAU level. For instance, average correlation between

the true latent factor and the predicted latent factor declines from 0.99 in the previous simulation to 0.73. Thus, we believe this joint model will be most advantageous when used on data involving many SAU variables. This investigation also suggests that performance in the non-nested misalignment setting may decline, given that none of the data are directly observed over the partition on which the latent factors are defined.

A highly contentious issue in factor analysis is how to choose m , the number of latent factors to model. This question may become even more complicated in the spatial misalignment setting, when variables do not have identical dimensions. Guidelines for choosing m in the nested misalignment setting are provided in [Section E](#) of the [supplementary material](#) available at *Biostatistics* online.

3.2. Comparison of the joint model with pre-analysis alignment

We also want to compare our joint model to the simple method of pre-analysis data alignment described in [Section 1](#). To the best of our knowledge, alignment methods like this one would typically be the only viable competitors to our model that would allow for factor scores to be predicted and inference performed at the level of the SAUs. We implement pre-analysis alignment on the first 1000 simulated datasets described above ($p_1 = 6$). To do so, for each variable collected at LAUs, we assign the value of each LAU to all of its nested SAUs, and we use these constructed variables at the SAUs together with the variables originally collected at the SAUs in a standard spatial factor model. A spatial factor model with the number of factors correctly specified and prior distributions analogous to those used in the joint model is applied to each aligned dataset using a Gibbs sampler with a Metropolis step. The sampler is run for 100 000 iterations with the first 50 000 samples discarded as burn-in.

When $m = 1$, the pre-analysis alignment model is able to identify the factor underlying the simulated data by relying on the information contained in the SAU data (average correlation of 0.99 between the true and predicted factor); however, the estimated factor loadings demonstrate that the information contained in the LAU data has little to no impact on the factor score predictions. The average posterior means of the LAU factor loadings are $\Lambda_2 = [0.09 \ 0.13 \ 0.11 \ -0.05 \ 0.27 \ 0.07]'$. The very small estimates of these loadings compared with their true values ([Table 1](#)) show that the LAU variables, likely distorted through the alignment procedure, no longer contribute appropriately to the factor score predictions, rendering this approach ineffective for integrating information from misaligned data for factor score prediction.

The superiority of the joint model over pre-analysis alignment is even more evident when a multi-factor simulation is assessed, which we demonstrate using the 10 $m = 2$ datasets generated for the simulations presented in [Section D](#) of the [supplementary material](#) available at *Biostatistics* online. We again fit both the joint model and the pre-analysis alignment model to each dataset, with $m = 2$ correctly specified. Because the factors can be estimated in a different order in each simulation, we investigate factor score predictive performance using the maximum of the pairwise correlations between each of the predicted factors and each of the true factors. For the joint model, the average maximum correlation across the 10 simulations was 0.86 and 0.83 for factors one and two, respectively. For the pre-analysis model, the average maximum correlation was 0.79 and 0.25 for factors one and two, respectively. This demonstrates that, while the pre-analysis alignment model can again use the SAU variables to identify one factor reasonably well, the distortion of the LAU variables misguides the model in the identification of the second factor.

4. APPLICATION TO LOUISIANA SOCIAL VULNERABILITY

Since the US Gulf Coast Region is particularly susceptible to environmental disasters, which could be compounded by climate change, there is a great deal of interest in developing indices of social vulnerability for this region in order to identify high risk communities and implement measures to reduce the impact of future disasters ([Oxfam America Inc, 2009](#)). [Oxfam America Inc \(2009\)](#) has analyzed the Social

Table 2. Variables included in the joint spatial factor analysis model

Variable names	Definitions
Census tract level	
Age	Proportion of the population under 65 years old
Race	Proportion of the population caucasian/white
Education	Proportion of the population 25 or older with a bachelor's degree or higher
Marital status	Proportion of the population 15 or older married
Employed	Proportion of the civilian population 16 or older in the labor force employed
Median household income	Median household income in 2012 inflation adjusted dollars
Rent as proportion of income	Median gross rent in 2012 inflation adjusted dollars as a proportion of median household income in 2012 inflation adjusted dollars
Household size	Average household size
Health insurance	Proportion of the population with health insurance
Gini index of inequality	Measure of income inequality developed by Gini (1912)
Household makeup	Proportion of the population living in married family households
County level	
Violent crime rate	Rate of violent crimes reported per 100 000 population (violent crimes include murders, rapes, robberies, and aggravated assaults)
Property crime rate	Rate of property crimes reported per 100 000 population (property crimes include burglaries, larcenies, and motor vehicle thefts)

Vulnerability Index created by [Cutter and others \(2003\)](#) in the Gulf Coast Region and has integrated it with an indicator of susceptibility to climate hazards. The US National Oceanic and Atmospheric Administration is developing a Coastal Resilience Index to measure the social vulnerability of disaster-prone coastal communities in the region ([Simpier and others, 2010](#)). The US Centers for Disease Control and Prevention (CDC) have also constructed a Social Vulnerability Index, which ranks the social vulnerability of the census tracts in each state to disaster and disease and can be used to identify the most vulnerable communities in the Gulf Coast Region ([Agency for Toxic Substances and Disease Registry, 2014](#)).

However, none of these existing indices take into account crime, which is an important component of social vulnerability to environmental hazards ([Perdikaris, 2014](#)). Furthermore, the construction of each of these indices is a highly subjective process, making them difficult to generalize when new information becomes available. In this section, we present an example of how our joint spatial factor analysis model can be invoked to address these limitations of existing indices while providing high resolution results. We apply the joint model to misaligned data from Louisiana to create a social vulnerability index at the census tract level that incorporates county level crime information. Besides the inevitable subjectivity in the variable selection, the construction of our index is fully data driven, as we model a single latent factor and use the resulting factor scores as the social vulnerability scores. This approach minimizes the subjectivity in our index, rendering it useful in the context of a broad range of disaster types in its current form and easily extended when more tailored information is available for specific disaster types.

Louisiana socioeconomic and demographic data, collected by the US Census Bureau, and crime data, collected by the US FBI, were accessed using Social Explorer ([US Bureau of the Census, 2016](#); [US Federal Bureau of Investigation, 2016](#)). Socioeconomic and demographic measures come from the 2008 to 2012 American Community Survey ([US Bureau of the Census, 2013](#)) and are obtained at the census tract level. The crime data being used are from the FBI's UCR in the year 2010 and are obtained at the county level (counties are referred to as parishes in Louisiana), as they are unavailable at the census tract level, as described in Section 1. The names and descriptions of all variables being included in the factor analysis appear in Table 2.

Census tracts for which some or all variables are missing are excluded from the analysis. Crime data are not missing for any parish. The final dataset contains a total of 1110 census tracts and 64 parishes ($N = 1110$, $K = 64$). Adhering to common practice in factor analysis, each variable is centered and scaled before applying the model (Ghosh and Dunson, 2009; Nethery and others, 2015). Hyperparameters are chosen to create vague prior distributions, and we let $m = 1$, as discussed above. Finally, the MCMC sampler for the joint model is run for 100 000 iterations, discarding the first 90 000 iterations as burn-in. Larger burn-ins are often required in the analysis of real data when compared with simulated data, as real data typically exhibit more messiness, causing samplers to converge more slowly. Model convergence was assessed through traceplots and found to be acceptable. Posterior means are used as estimates for all parameters.

Figure 1a provides a map of the posterior means and standard deviations of the latent factor scores for each census tract in Louisiana, with zooming for only Orleans Parish, Louisiana, which contains much of the city of New Orleans, in Figure 1b. Census tracts mapped in white were eliminated from the analysis due to missing data. The pattern in the factor loadings, shown in Figure 1c, suggests that census tracts with higher predicted latent factor scores may, indeed, have higher levels of social vulnerability.

As further evidence that the latent factor from the joint model, which we call the joint model index, measures the social vulnerability of the Louisiana census tracts, we have plotted it against the CDC's social vulnerability rankings for the census tracts in Louisiana from 2014 in Figure 2. This plot demonstrates that the joint model index is correlated with a respected social vulnerability index (correlation of 0.63), but it also illustrates the effect of including crime in the scores. The ordinary least squares (OLS) regression line from the regression of the joint model index on the CDC's social vulnerability ranking is included on the plot. Census tracts from counties with high crime rates, defined as having both violent and property crime rates above the 75th percentile for the state, largely fall above the OLS regression line, indicating that these tracts are typically assigned relatively higher vulnerability scores from our index compared with the CDC's index. This suggests that our index appropriately integrates information about crime and gives higher scores to communities with higher crime rates when compared with existing social vulnerability indices. This application of the joint model demonstrates how it can be used to incorporate data at different spatial scales to improve on existing social vulnerability indices.

Given that these results suggest that the joint model index constitutes an index of social vulnerability for Louisiana, we now provide an example of how it can be integrated with historic natural disaster data for Louisiana to identify the communities that are at high risk geographically for natural disasters and are highly socially vulnerable, which exacerbates the impacts of such disasters. During future natural disasters, this type of information can be consulted to help allocate resources in a way that alleviates the burden on the highest risk communities. We choose to focus on geographic vulnerability to flooding, because Louisiana has been historically hard hit by tropical storms and floods. To measure flood vulnerability, we employ data from the Federal Emergency Management Agency (FEMA). FEMA offers low-cost flood insurance to property owners nationwide through its National Flood Insurance Program (NFIP), and it makes historic policy and claims data available for each county in the US, summarized January 1, 1978 to January 30, 2017 (Federal Emergency Management Agency, 2017a).

As a proxy for flood vulnerability, we investigate the rate of losses closed with payment (per 100 000 residents) from NFIP for each county in Louisiana. This variable is chosen because a loss closed with payment indicates that flood damages to property were confirmed by multiple sources—both the property owner and the insurance assessor—making it a more reliable measure of flood vulnerability than other NFIP statistics such as rate of policies or rate of claims made (Federal Emergency Management Agency, 2017b). Finally, using thresholds corresponding to the 75th percentile of both the flood vulnerability and social vulnerability measures for Louisiana, we create binary classifiers of flood risk and social vulnerability, i.e., any county above the 75th percentile of losses closed with payment is considered high flood risk and any census tract above the 75th percentile of social vulnerability is considered high social vulnerability. These

Fig. 1. Predicted factor scores (left panel) and standard deviations (right panel) for each census tract from the joint model mapped across all of Louisiana (a) and Orleans Parish only (b) and a heat map of the estimated factor loadings (c).

thresholds were chosen due to their practical results—they identified a moderate number of communities that might reasonably be able to be prioritized for receipt of limited resources during a disaster.

The interaction between these classifiers is mapped across Louisiana, with zooming for Orleans Parish (Figure 3). This indicator shows that the city neighborhoods in New Orleans are both highly socially vulnerable and highly geographically vulnerable to flooding. This is not a surprising finding given the extent of the flooding damages in New Orleans following Hurricane Katrina in 2005, which exacerbated the already high social vulnerability in many of these city neighborhoods.

Fig. 2. The index created by the joint model plotted against the US Centers for Disease Control and Prevention's Social Vulnerability Index for Louisiana census tracts. Points plotted as 'x' represent census tracts from counties with both violent and property crime rates above the 75th percentile for the state. All other census tracts are plotted with an 'o'. The ordinary least squares regression line from the regression of the joint model index on the CDC's social vulnerability index is included.

5. DISCUSSION

Given the vast amount of spatially referenced data now available, the issue of using these data to provide meaningful and concise characterizations of places and communities is of great interest to researchers and policy makers alike. However, these data may be spatially misaligned. For this reason, we have developed a joint spatial factor analysis model to handle data from misaligned areal units, which can be used to produce a common set of latent factors underlying all the data.

In Section 3, we demonstrated the effectiveness of the model and its superiority over naive methods for dealing with spatial misalignment in factor analysis. Finally, the model was applied to misaligned social indicator data from Louisiana to develop social vulnerability scores for each census tract in the state. We provided evidence that the joint model produces a social vulnerability index for Louisiana that improves on existing indices by incorporating information from different spatial levels while yielding high spatial resolution results. While our index is intended to be general enough to provide information on vulnerability to a wide range of disaster types, thanks to its data driven construction it can easily be extended for use in the context of specific disasters for which additional variables are known to be relevant to social vulnerability. Note that, although our social vulnerability index appears promising in initial investigations, formal validation of its ability to measure social vulnerability is still needed.

When integrated with information about past or future geographic vulnerability to environmental disasters, our social vulnerability index can be used to help policy makers and disaster responders identify communities likely to need the most assistance in future disasters, as we demonstrated through a joint assessment of flood vulnerability and social vulnerability in Louisiana. Future work could combine our index with climate change disaster predictions to prepare for impacts on the population in the high risk

Fig. 3. The interaction between flood vulnerability and social vulnerability classifiers mapped across all of Louisiana (a) and Orleans Parish only (b).

US Gulf Coast Region. Data from the Gulf Long Term Follow-Up Study (*Kwok and others, 2017*), a study tracking the long term health of workers on the 2010 Deepwater Horizon oil spill, when combined with our social vulnerability index, provide an opportunity to investigate whether people living in socially vulnerable communities were differentially impacted by the oil spill.

The key methodological direction for future work presented by this article is investigation of the performance of this model on both simulated and real non-nested misaligned areal data. Moreover, methodological extensions could be implemented to allow for greater model flexibility. For instance, three or more models could be fit jointly to allow for more spatial levels. When $m > 1$, separate spatial parameters can be specified for each latent factor in order to allow the latent factors to contain different degrees of spatial correlation. This method could also be integrated with a model-based approach to choosing m , such as the reversible jump MCMC method of *Lopes and West (2004)*. The applications of this method extend well beyond the social vulnerability application emphasized here. For example, environmental pollutant data are often measured at different scales. This model could be applied to misaligned pollutant data to develop environmental exposure scores across a region of interest.

6. SOFTWARE

MATLAB code is available on Github at https://github.com/rachelnethery/fa_misalign.

SUPPLEMENTARY MATERIAL

Supplementary material is available at <http://biostatistics.oxfordjournals.org>.

ACKNOWLEDGMENTS

We thank the reviewers for their thoughtful suggestions, which improved the manuscript tremendously.
Conflict of Interest: None declared.

FUNDING

This study was funded in part by the Intramural Research Program of the NIH, National Institute of Environmental Sciences (ZO1 ES 102945) and the NIH Summer Internship Program in Biomedical Research.

REFERENCES

- AGENCY FOR TOXIC SUBSTANCES AND DISEASE REGISTRY (2014). The Social Vulnerability Index. The Centers for Disease Control and Prevention. <https://svi.cdc.gov/Index.html> (accessed 22 December 2016).
- BANERJEE, S., CARLIN, B. P. AND GELFAND, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data, 2nd edition*. Boca Raton, FL: CRC Press.
- BOLLEN, K. A. (1989). *Structural Equations with Latent Variables*. Hoboken, NJ: John Wiley and Sons, Inc.
- BRADLEY, J. R., WIKLE, C. K. AND HOLAN, S. H. (2015). Multiscale analysis of survey data: recent developments and exciting prospects. *Statistics Views*. Hoboken, NJ: Wiley.
- CRESSIE, N. A. (1996). Change of support and the modifiable areal unit problem. *Geographical Systems* **3**, 159–180.
- CUTTER, S. L., BARNES, L., BERRY, M., BURTON, C., EVANS, E., TATE, E. AND WEBB, J. (2008). A place-based model for understanding community resilience to natural disasters. *Global Environmental Change* **18**, 598–606.
- CUTTER, S. L., BORUFF, B. J. AND SHIRLEY, W. L. (2003). Social vulnerability to environmental hazards. *Social Science Quarterly* **84**, 242–261.
- CUTTER, S. L. AND FINCH, C. (2008). Temporal and spatial changes in social vulnerability to natural hazards. *Proceedings of the National Academy of Sciences of the USA* **105**, 2301–2306.
- DIENER, E. AND SUH, E. (1997). Measuring quality of life: economic, social, and subjective indicators. *Social Indicators Research* **40**, 189–216.
- DUNCAN, O. D. (1974). Developing social indicators. *Proceedings of the National Academy of Sciences of the USA* **71**, 5096–5102.
- FEDERAL EMERGENCY MANAGEMENT AGENCY (2017a). The National Flood Insurance Program. <https://www.fema.gov/national-flood-insurance-program> (accessed 21 March 2017).
- FEDERAL EMERGENCY MANAGEMENT AGENCY (2017b). The National Flood Insurance Program Data Definitions. http://bsa.nfipstat.fema.gov/reports/data_definitions.html (accessed 10 April 2017).
- GELFAND, A. E. AND SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.
- GEMAN, S. AND GEMAN, D. (1984). Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.
- GHOSH, J. AND DUNSON, D. B. (2009). Default prior distributions and efficient posterior computation in Bayesian factor analysis. *Journal of Computational and Graphical Statistics* **18**, 306–320.

- GINI, C. (1912). Variabilità e mutabilità. *Reprinted in Memorie di metodologica statistica* (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi.
- GOTWAY, C. A. AND YOUNG, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association* **97**, 632–648.
- HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.
- HOGAN, J. W. AND TCHERNIS, R. (2004). Bayesian factor analysis for spatially correlated data, with application to summarizing area-level material deprivation from census data. *Journal of the American Statistical Association* **99**, 314–324.
- KWOK, R. K., ENGEL, L. S., MILLER, A. K., BLAIR, A., CURRY, M. D., JACKSON II, W. B., STEWART, P. A., STENZEL, M. R., BIRNBAUM, L. S. and others. (2017). The GuLF STUDY: A prospective study of persons involved in the Deepwater Horizon oil spill response and clean-up. *Environmental Health Perspectives (Online)* **125**, 570.
- LIU, X., WALL, M. M. AND HODGES, J. S. (2005). Generalized spatial structural equation models. *Biostatistics* **6**, 539–557.
- LOPES, H. F., SALAZAR, E. AND GAMERMAN, D. (2008). Spatial dynamic factor analysis. *Bayesian Analysis* **3**, 759–792.
- LOPES, H. F. AND WEST, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica* **14**, 41–68.
- METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. AND TELLER, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* **21**, 1087–1092.
- MEZZETTI, M. (2012). Bayesian factor analysis for spatially correlated data: application to cancer incidence data in Scotland. *Statistical Methods and Applications* **21**, 49–74.
- NETHERY, R. C., WARREN, J. L., HERRING, A. H., MOORE, K. A. B., EVENSON, K. R. AND DIEZ-ROUX, A. V. (2015). A common spatial factor analysis model for measured neighborhood-level characteristics: the multi-ethnic study of atherosclerosis. *Health & Place* **36**, 35–46.
- OXFAM AMERICA INC. (2009). Exposed: Social Vulnerability and Climate Change in the US Southeast. <https://policy-practice.oxfamamerica.org/static/oa3/files/Exposed-Social-Vulnerability-and-Climate-Change-in-the-US-Southeast.pdf> (accessed 09 November 2016).
- PERDIKARIS, J. (2014). *Physical Security and Environmental Protection*. Boca Raton, FL: CRC Press.
- R CORE TEAM (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- ROWE, D. B. (1998). Correlated bayesian factor analysis, [PhD Thesis]. Citeseer.
- SIMPIER, T. T., SWANN, D.L., EMMER, R., SIMPIER, S.H. AND SCHNEIDER, M. (2010). Coastal Resilience Index: A Community Self-Assessment. MASGP-08-014. http://masgc.org/assets/uploads/publications/662/coastal_community_resilience_index.pdf (accessed 09 November 2016).
- SMITH, D. M. (1973). *The Geography of Social Well-Being in the United States: An Introduction to Territorial Social Indicators*. New York, New York: McGraw-Hill.
- SMITH, T. W. (1981). Social indicators. *Journal of Social History* **14**, 739–747.
- STAKHOVYCH, S., BIJMOLT, T. H. A. AND WEDEL, M. (2012). Spatial dependence and heterogeneity in Bayesian factor analysis: a cross-national investigation of Schwartz values. *Multivariate Behavioral Research* **47**, 803–839.
- TAYLOR, C. L. AND HUDSON, M. C. (1970). World handbook of political and social indicators II. Section 1. Cross-national aggregate data. *Technical Report*. Michigan University Ann Arbor Department of Political Science.
- THE MATHWORKS INC. (2015). *MATLAB Release 2015b*. Natick, MA: The MathWorks Inc.

- TREVISANI, M. AND GELFAND, A. (2013). *Spatial Misalignment Models for Small Area Estimation: A Simulation Study*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 269–279.
- US BUREAU OF THE CENSUS (1994). Geographic Areas Reference Manual. <https://www.census.gov/geo/reference/garm.html> (accessed 09 November 2016).
- US BUREAU OF THE CENSUS (2013). American Community Survey Information Guide. https://www.census.gov/content/dam/Census/programs-surveys/acs/about/ACS_Information_Guide.pdf (accessed 18 November 2016).
- US BUREAU OF THE CENSUS (2016). American Community Survey Estimates, 2008–2012. Prepared by Social Explorer. <http://www.socialexplorer.com/> (accessed 29 July 2016).
- US FEDERAL BUREAU OF INVESTIGATION (2010). Uniform Crime Reporting Data Online. www.ucrdatatool.gov (accessed 09 November 2016).
- US FEDERAL BUREAU OF INVESTIGATION (2016). Uniform Crime Report Data. Prepared by Social Explorer. <http://www.socialexplorer.com/> (accessed 11 October 2016).
- WANG, F. AND WALL, M. M. (2003). Generalized common spatial factor model. *Biostatistics* **4**, 569–582.

[Received April 13, 2017; revised February 13, 2018; accepted for publication February 25, 2018]