MULTIVARIATE SPATIO-TEMPORAL MODELS FOR HIGH-DIMENSIONAL AREAL DATA WITH APPLICATION TO LONGITUDINAL EMPLOYER-HOUSEHOLD DYNAMICS[1]

BY JONATHAN R. BRADLEY, SCOTT H. HOLAN
AND CHRISTOPHER K. WIKLE

University of Missouri

Many data sources report related variables of interest that are also referenced over geographic regions and time; however, there are relatively few general statistical methods that one can readily use that incorporate these multivariate spatio-temporal dependencies. Additionally, many multivariate spatio-temporal areal data sets are extremely high dimensional, which leads to practical issues when formulating statistical models. For example, we analyze Quarterly Workforce Indicators (QWI) published by the US Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) program. QWIs are available by different variables, regions, and time points, resulting in millions of tabulations. Despite their already expansive coverage, by adopting a fully Bayesian framework, the scope of the QWIs can be extended to provide estimates of missing values along with associated measures of uncertainty. Motivated by the LEHD, and other applications in federal statistics, we introduce the multivariate spatio-temporal mixed effects model (MSTM), which can be used to efficiently model high-dimensional multivariate spatio-temporal areal data sets. The proposed MSTM extends the notion of Moran's I basis functions to the multivariate spatio-temporal setting. This extension leads to several methodological contributions, including extremely effective dimension reduction, a dynamic linear model for multivariate spatio-temporal areal processes, and the reduction of a high-dimensional parameter space using a novel parameter model.

1. Introduction. Ongoing data collection from the private sector along with federal, state, and local governments have produced massive quantities of data measured over geographic regions (areal data) and time. This unprecedented volume of spatio-temporal data contains a wide range of variables and, thus, has created unique challenges and opportunities for those practitioners seeking to capitalize on their full utility. For example, methodological issues arise because these data exhibit complex multivariate spatio-temporal covariances that may involve

nonstationarity and interactions between different variables, regions, and times. Additionally, the fact that these data (with complex dependencies) are often extremely high-dimensional (so called "big data") leads to the important practical issue associated with computation.

As an example, the US Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) program produces estimates of US labor force variables called Quarterly Workforce Indicators (QWIs). The QWIs are derived from a combination of administrative records and data from federal and state agencies [Abowd et al. (2009)]. The sheer amount of QWIs available is unprecedented, and has made it possible to investigate local (in space–time) dynamics of several variables important to the US economy. For example, the average monthly income QWI is estimated quarterly over multiple regions and industries (e.g., education, manufacturing, etc.). In total, there are 7,530,037 quarterly estimates of average monthly income.

The QWIs present interesting methodological challenges. In particular, not every state signs a new Memorandum of Understanding (MOU) each year and, hence, QWIs are not provided for these states [Abowd et al. (2009), Section 5.5.1]. Furthermore, some data are suppressed at certain regions and time points due to disclosure limitations [Abowd et al. (2009), Section 5.6]. Another limitation is that uncertainty measures are not made publicly available. Consequently, it is difficult for QWI data users to assess the quality of the published estimates. Thus, producing a complete set of estimates (i.e., national coverage) that have associated measures of uncertainty is extremely important and provides an unprecedented tool for the LEHD-user community. As such, we take a fully Bayesian approach to estimating quarterly measures of average monthly income and, thus, provide a complete set of estimates that have associated measures of uncertainty.

A fully Bayesian model that can efficiently and jointly model a correlated (over multiple variables, regions, and times) data set of this size (7.5×10^6) is unprecedented. It is instructive to compare the dimensionality of the QWI to data sets used in spatial analyses in other scientific domains. For example, Banerjee et al. (2008) use a fully Bayesian approach to analyze a multivariate spatial agricultural data set consisting of 40,500 observations; Cressie and Johannesson (2008) use an empirical Bayesian approach to analyze a spatial data set of total column ozone with 173,405 observations; Lindgren, Rue and Lindström (2011) use a fully Bayesian approach to analyze climate spatially using approximately 32,000 observations; and Sengupta et al. (2012) use an empirical Bayesian approach to analyze cloud fractions using a data set of size 2,748,620. Furthermore, none of these methods allow for multivariate dependencies between different geographic regions and time points.

Despite the wide availability of high-dimensional areal data sets exhibiting multivariate spatio-temporal dependencies, the literature on modeling multivariate spatio-temporal areal processes is relatively recent by comparison. For example, various multivariate space–time conditional autoregressive (CAR) models have been proposed by Carlin and Banerjee (2003), Congdon (2002), Pettitt, Weir and

Hart (2002), Zhu, Eickhoff and Yan (2005), Daniels, Zhou and Zou (2006), and Tzala and Best (2008), among others. However, these methodologies cannot efficiently model high-dimensional data sets. Additionally, these approaches impose separability and various independence assumptions, which are not appropriate for many settings, as these models fail to capture important interactions and dependencies between different variables, regions, and times [Stein (2005)]. Hence, we introduce the multivariate spatio-temporal mixed effects model (MSTM) to analyze high-dimensional multivariate data sets that vary over different geographic regions and time points.

The MSTM is built upon the first order linear dynamic spatio-temporal model (DSTM) [Cressie and Wikle (2011)]. To date, no DSTM has been proposed to analyze multivariate high-dimensional areal data, and, as a result, the components of the MSTM require significant methodological development. Specifically, we introduce novel classes of multivariate spatio-temporal basis functions, propagator matrices, and parameter models to be used within the MSTM.

The components of the MSTM can be specified to have a computationally advantageous reduced rank structure [e.g., see Wikle (2010)], which allows us to analyze high-dimensional areal data (e.g., QWIs from the LEHD program). This reduced rank structure is achieved, in part, by extending various aspects of the model suggested by Hughes and Haran (2013) from the univariate spatial-only setting to the multivariate spatio-temporal setting. Specifically, we extend the Moran's I (MI) basis functions to the multivariate spatio-temporal setting [for the spatial-only case see Griffith (2000, 2002, 2004), Griffith and Tiefelsdorf (2007), Hughes and Haran (2013), Porter, Wikle and Holan (2015)]. Further, we propose a novel propagator (or transition) matrix for the first-order vector autoregressive—VAR(1)—model, which we call the MI propagator matrix. In this context, the propagator matrix of the VAR(1) model is specified to have a desirable nonconfounding property, which is similar to the specification of the multivariate spatio-temporal MI basis functions.

We also propose an extension of the spatial random effects covariance parameter model used in Hughes and Haran (2013) and Porter, Holan and Wikle (2015), which we call the MI prior. Here, we interpret the MI prior as a rescaling of the covariance matrix that is specified to be close (in Frobenius norm) to a "target precision" matrix. This parameterization significantly reduces the dimensionality of the parameter space, thereby reducing the computational burden associated with fully Bayesian inference in high-dimensional spatio-temporal settings. Furthermore, this target precision matrix can be sensibly chosen based on knowledge of the underlying spatial process.

In addition to modeling QWIs from the LEHD, the MSTM can be used to effectively address numerous statistical modeling and analysis problems in the context of multivariate spatio-temporal areal data. For example, besides analyzing high-dimensional data, the MSTM can also be used to model nonseparable and nonstationary covariances, and to combine data from multiple repeated surveys. Although we mainly focus on modeling high-dimensional multivariate spatio-temporal areal

data (e.g., QWIs from the LEHD), the MSTM is tremendously flexible and can be readily adapted to other settings.

The remainder of this article is organized as follows. In Section 2 we introduce the LEHD-QWI data set and further describe the methodological challenges that we consider. Next, in Section 3 we provide mathematical foundations for the MSTM. Then, in Section 4 we introduce the multivariate spatio-temporal MI basis functions, the MI propagator matrix, and the parameter model for the covariance matrix of the random effects term. Section 5 provides an empirical study that is used to evaluate the effectiveness of the MSTM in recovering the unobserved latent process ("true" underlying values). Additionally, in Section 5 we use the MSTM to jointly analyze all 7,530,037 QWIs obtained from the US Census Bureau's LEHD program. Finally, Section 6 contains discussion. For convenience of exposition, proofs of the technical results and details surrounding the MCMC algorithm are left to an Appendix.

2. LEHD—Quarterly Workforce Indicators. The LEHD program provides public access QWIs on several earnings variables for each quarter of the year over various geographies of the US (http://www.census.gov/). For a comprehensive description regarding the creation of QWIs, see Abowd et al. (2009). Here, we consider quarterly measures of average monthly income for individuals with steady jobs. A subset of this data set representing QWIs for 2970 US counties for women in the education industry during the third quarter of 2006 is displayed in Figure 1. However, the QWIs are much more extensive. Specifically, the quarterly average monthly income for individuals that have a steady job is available over 92 quarters (ranging from 1990 to 2013), all of the 3145 US counties, by each gender, and by 20 different industries. This results in the aforementioned data set having 7,530,037 observations—which we model jointly.

Monthly Income For Women (Education Industry, 3rd Quarter of 2006)

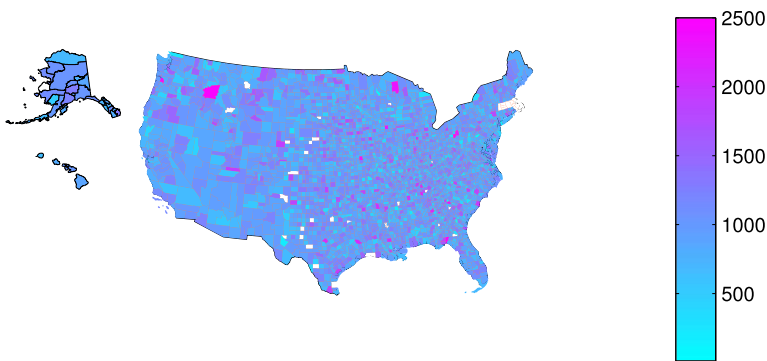FIG. 1. *We present the QWI for quarterly average monthly income* (*US dollars*) *for* 2970 *US counties, for women, for the education industry, and for the third quarter of* 2006. *The white areas indicate QWIs that are not made available by LEHD.*

The high-dimensional nature of QWIs and expansive coverage (e.g., quarterly average monthly incomes) allows economists and other subject matter researchers to study differences in key US economic variables over many regions and times. Consequently, QWIs have had a significant impact on the economics literature; for example, see Davis et al. (2006), Thompson (2009), Dube, Lester and Reich (2013), Allegretto et al. (2013), among others. This demand for QWIs shows a clear need for developing statistical methodology that can be used to analyze such high-dimensional data sets. The current statistical approaches available cannot capitalize on the full utility of the QWIs. For example, Abowd, Schneider and Vilhuber (2013) limit the spatial and temporal scope of their analysis, which allows them to efficiently analyze only a portion of the QWIs.

The complexity of the QWIs is further exacerbated by missing values; by "missing" we mean that the QWI is not provided by the LEHD program. Consider the (quarterly) average monthly income example, the total gender/industry/space/time combinations results in $2 \times 20 \times 3145 \times 92 = 11{,}573{,}600$ possible QWIs. Hence, roughly 35% of the QWIs are missing. This leads to a total of $11{,}573{,}600^2$ pairwise covariances that require modeling using random effects. Nevertheless, allowing for multivariate spatio-temporal covariances is extremely important from the perspective of predicting (imputing) missing QWIs.

As an example, in Figure 1, one might expect the quarterly average monthly income for men to be associated with the value for quarterly average monthly income for women. Likewise, nearby observations in space and time are often similar in value [Cressie and Wikle (2011)]. If no multivariate spatio-temporal dependencies are present in the data, then one can not borrow strength among "similar" variables and "nearby" observations to improve the precision of the estimated QWIs. An exploratory analysis, based on the empirical covariance matrices computed from the log QWIs (not shown), indicates that the QWIs are indeed correlated across different variables, regions, and times. Consequently, this suggests that a statistical model that allows for multivariate spatio-temporal dependence can be efficiently utilized to predict (impute) QWIs.

3. The multivariate spatio-temporal mixed effects model. The DSTM framework is a well-established modeling approach used to analyze data referenced over space and time. This approach is extremely flexible since it allows one to define how a group of spatial regions temporally evolve [e.g., see Cressie and Wikle (2011), page 13], as opposed to defining the temporal evolution of a process at each geographic region of interest. The MSTM represents a novel extension of the DSTM to the multivariate areal data setting, where we now allow groups of spatially referenced variables to evolve over time. Thus, in Sections 3.1 and 3.2 we introduce the MSTM in terms of the familiar "data model" and "process model" DSTM terminology [Cressie and Wikle (2011)].

3.1. *The MSTM data model.* The data model for the MSTM is defined as

$$Z_t^{(\ell)}(A) = Y_t^{(\ell)}(A) + \varepsilon_t^{(\ell)}(A);$$

(1)

$$\ell = 1, \ldots, L, t = T_L^{(\ell)}, \ldots, T_U^{(\ell)}, A \in D_{\mathrm{P},t}^{(\ell)},$$

where $\{Z_t^{(\ell)}(\cdot)\}$ represents multivariate spatio-temporal areal data. The components of (1) are defined and elaborated as follows:

1. The subscript "t" denotes discrete time, and the superscript "ℓ" indexes different variables of interest (e.g., the QWI for women in the education industry). There are a total of L variables of interest (i.e., $\ell = 1, \ldots, L$) and we allow for a different number of observed time points for each of the L variables of interest (i.e., for variable ℓ, $t = T_L^{(\ell)}, \ldots, T_U^{(\ell)}$).
2. We require $T_L^{(\ell)}, \ldots, T_U^{(\ell)}$ to be on the same temporal scale (e.g., quarterly) for each ℓ, $T_L^{(\ell)} \leq T_U^{(\ell)}$, $\min(T_L^{(\ell)}) = 1$, and $\max(T_U^{(\ell)}) = T \geq 1$.
3. The set A represents a generic areal unit. For example, a given set A might represent a state, county, or a census tract. Denote the collection of all $n_t^{(\ell)}$ observed areal units with the set $D_{\mathrm{O},t}^{(\ell)} \equiv \{A_{t,i}^{(\ell)} : i = 1, \ldots, n_t^{(\ell)}\}$; $\ell = 1, \ldots, L$. The observed data locations are different from the prediction locations $D_{\mathrm{P},t}^{(\ell)} \equiv \{A_{t,j}^{(\ell)} : j = 1, \ldots, N_t^{(\ell)}\}$, that is, we consider predicting on a spatial support that may be different from $\{D_{\mathrm{O},t}^{(\ell)}\}$ (e.g., the counties with missing QWIs are not included in $\{D_{\mathrm{O},t}^{(\ell)}\}$, but are included in $\{D_{\mathrm{P},t}^{(\ell)}\}$). Additionally, denote the number of prediction locations at time t as $N_t = \sum_{\ell=1}^{L} N_t^{(\ell)}$ and the total number of prediction locations as $N \equiv \sum_{t=1}^{T} N_t$. In a similar manner, the number of observed locations at time t and total number of observations are given by $n_t = \sum_{\ell=1}^{L} n_t^{(\ell)}$ and $n \equiv \sum_{t=1}^{T} n_t$, respectively.
4. The random process $Y_t^{(\ell)}(\cdot)$ represents the ℓth variable of interest at time t. For example, $Y_t^{(\ell)}(\cdot)$ might represent the quarterly average monthly income for women in the education industry at time t. The stochastic properties of $\{Y_t^{(\ell)}(\cdot)\}$ are defined in Section 3.2. Latent processes like $\{Y_t^{(\ell)}(\cdot)\}$ have been used to incorporate spatio-temporal dependencies [e.g., see Cressie and Wikle (2011)], which we modify to the multivariate spatio-temporal areal data setting.
5. It is assumed that $\varepsilon_t^{(\ell)}(\cdot)$ is a white-noise Gaussian process with mean zero and unknown variance $\mathrm{var}\{\varepsilon_t^{(\ell)}(\cdot)\} = v_t^{(\ell)}(\cdot)$ for $\ell = 1, \ldots, L$, and $t = T_L^{(\ell)}, \ldots, T_U^{(\ell)}$. The presence of $\{\varepsilon_t^{(\ell)}(\cdot)\}$ in (1) allows us to take into account that we do not perfectly observe $\{Y_t^{(\ell)}(\cdot)\}$, and instead observe a noisy version $\{Z_t^{(\ell)}(\cdot)\}$. In many settings, there is information that we can use to define $\{\varepsilon_t^{(\ell)}(\cdot)\}$ (e.g., information provided by the statistical agency). If one

does not account for this extra source of variability, then the total variability of the process $\{Y_t^{(\ell)}(\cdot)\}$ may be underestimated. For example, Finley et al. (2009) show that if one ignores white-noise error in a Gaussian linear model, then one underestimates the total variability of the latent process of interest.

3.2. *The MSTM process model.* The process model for MSTM is defined as

(2)
$$Y_t^{(\ell)}(A) = \mu_t^{(\ell)}(A) + \mathbf{S}_t^{(\ell)}(A)'\boldsymbol{\eta}_t + \xi_t^{(\ell)}(A);$$
$$\ell = 1, \dots, L, t = T_L^{(\ell)}, \dots, T_U^{(\ell)}, A \in D_{\mathrm{P},t}^{(\ell)}.$$

In (2), $Y_t^{(\ell)}(\cdot)$ represents the ℓth spatial random process of interest at time t, which is modeled by three terms on the right-hand side of (2). The first term [i.e., $\{\mu_t^{(\ell)}(\cdot)\}$] is a fixed effect, which is unknown, and requires estimation. We set $\mu_t^{(\ell)}(\cdot) \equiv \mathbf{x}_t^{(\ell)}(\cdot)'\boldsymbol{\beta}_t$, where $\mathbf{x}_t^{(\ell)}$ is a known p-dimensional vector of covariates and $\boldsymbol{\beta}_t \in \mathbb{R}^p$ is an associated unknown parameter vector; $\ell = 1, \dots, L$ and $t = 1, \dots, T$. In general, we allow both $\mathbf{x}_t^{(\ell)}$ and $\boldsymbol{\beta}_t$ to change over time; however, in practice, one must assess whether or not this is appropriate for a given application. For the QWI example we specify $\mathbf{x}_t^{(\ell)}$ and $\boldsymbol{\beta}_t$ to be constant over time.

The second term on the right-hand side of (2) [i.e., $\{\mathbf{S}_t^{(\ell)}(\cdot)'\boldsymbol{\eta}_t\}$] represents multivariate spatio-temporal dependencies. The r-dimensional vectors of multivariate spatio-temporal basis functions $\mathbf{S}_t^{(\ell)}(\cdot) \equiv (S_{t,1}^{(\ell)}(\cdot), \dots, S_{t,r}^{(\ell)}(\cdot))'$ are prespecified for each $t = 1, \dots, T$ and $\ell = 1, \dots, L$, and in Section 4.1 we propose a new class of multivariate spatio-temporal basis functions to use in (2). The r-dimensional random vector $\boldsymbol{\eta}_t$ is assumed to follow a spatio-temporal VAR(1) model [Cressie and Wikle (2011), Chapter 7]

(3)
$$\boldsymbol{\eta}_t = \mathbf{M}_t \boldsymbol{\eta}_{t-1} + \mathbf{u}_t; \qquad t = 2, 3, \dots, T,$$

where for all t the r-dimensional random vector $\boldsymbol{\eta}_t$ is Gaussian with mean zero and has an unknown $r \times r$ covariance matrix $\mathbf{K}_t$; $\mathbf{M}_t$ is a $r \times r$ known propagator matrix (see discussion below); and $\mathbf{u}_t$ is an r-dimensional Gaussian random vector with mean zero and unknown $r \times r$ covariance matrix $\mathbf{W}_t$ and is independent of $\boldsymbol{\eta}_{t-1}$.

First order vector autoregressive models may offer more realistic structure with regards to interactions across space and time. This is a feature that cannot be included in the alternative modeling approaches discussed in Section 1. Additionally, the (temporal) VAR(1) model has been shown to perform well (empirically) in terms of both estimation and prediction for federal data repeated over time [Jones (1980), Bell and Hillmer (1990), Feder (2001)].

The r-dimensional random vectors $\{\boldsymbol{\eta}_t\}$ are not only used to model temporal dependencies in $\{Y_t^{(\ell)}(\cdot)\}$, but are also used to model multivariate dependencies.

Notice that the random effect term $\boldsymbol{\eta}_t$ is common across all L processes. Allowing for a common random effect term between different processes is a straightforward way to induce dependence [Cressie and Wikle (2011), Chapter 7.4]. This strategy has been previously used in the univariate spatial and multivariate spatial settings [e.g., see Royle et al. (1999), Finley et al. (2009), and Banerjee et al. (2010)] and has been extended here.

Finally, the third term on the right-hand side of (2) [i.e., $\{\xi_t^{(\ell)}(\cdot)\}$] represents fine-scale variability and is assumed to be Gaussian white noise with mean zero and unknown variance $\{\sigma_{\xi,t}^2\}$. In general, $\{\xi_t^{(\ell)}(\cdot)\}$ represents the leftover variability not accounted for by $\{\mathbf{S}_t^{(\ell)}(\cdot)'\boldsymbol{\eta}_t\}$. One might consider modeling spatial covariances in $\{\xi_t^{(\ell)}(\cdot)\}$. Minor adjustments to our methodology could be used to incorporate, for example, a CAR model [Banerjee, Carlin and Gelfand (2004), Chapter 3], tapered covariances [Cressie (1993), page 108], or block diagonal covariances [Stein (2014)] in $\{\xi_t^{(\ell)}(\cdot)\}$.

4. Multivariate spatio-temporal mixed effects model specifications.

Many specifications of the MSTM require methodological development before one directly can apply it to the QWIs. In particular, we need to specify the multivariate spatio-temporal basis functions $\{\mathbf{S}_t^{\ell}(\cdot)\}$, the propagator matrices $\{\mathbf{M}_t\}$, and the parameter models for $\{\mathbf{K}_t\}$ and $\{\mathbf{W}_t\}$. These contributions are detailed in Sections 4.1, 4.2, and 4.3, respectively.

4.1. *Moran's I basis functions.* In principle, the r-dimensional vector $\mathbf{S}_t^{(\ell)}(\cdot)$ can belong to any class of spatial basis functions; however, we use Moran's I (MI) basis functions, since they have many properties that are needed to accurately and efficiently model QWIs. In particular, the MI basis functions can be used to model areal data in a reduced dimensional space (i.e., $r \ll n$). This feature allows for fast computation of the distribution of $\{\boldsymbol{\eta}_t\}$, which can become computationally expensive for large r. This will be especially useful for analyzing the QWIs in Section 5.3, which consists of 7,530,037 observations. Additionally, the MI basis functions allow for nonstationarity in space, which is a realistic property for modeling QWIs (see Section 2 for a discussion).

A defining (and mathematically desirable) property of the MI basis functions is that they guarantee there are no issues with confounding between fixed and random effects. This property of removing any confounding frees us to consider inferential questions in addition to multivariate spatio-temporal prediction. For example, the QWIs can be used to investigate the degree of gender inequality in the US by comparing the mean (i.e., $\mu_t^{(\ell)}$) average monthly income for men and women, respectively.

Thus, to derive MI basis functions to use for QWIs, we extend this defining property to the multivariate spatio-temporal setting. Here, the derivation starts with the MI operator. Recall that the MI statistic is a measure of association, which

equals to a weighted sums of squares where the weights are called the MI operator [see Hughes and Haran (2013)]. At time t the MI operator is explicitly defined as

$$\mathbf{G}(\mathbf{X}_t, \mathbf{A}_t) \equiv (\mathbf{I}_{N_t} - \mathbf{X}_t(\mathbf{X}_t'\mathbf{X}_t)^{-1}\mathbf{X}_t')\mathbf{A}_t(\mathbf{I}_{N_t} - \mathbf{X}_t(\mathbf{X}_t'\mathbf{X}_t)^{-1}\mathbf{X}_t');$$

(4)
$$t = 1, \ldots, T,$$

where the $N_t \times p$ matrix $\mathbf{X}_t \equiv (\mathbf{x}_t^{(\ell)}(A) : \ell = 1, \ldots, L, A \in D_{\mathrm{P},t}^{(\ell)})'$, $\mathbf{I}_{N_t}$ is an $N_t \times N_t$ identity matrix, and $\mathbf{A}_t$ is the $N_t \times N_t$ adjacency matrix corresponding to the edges formed by $\{D_{\mathrm{P},t}^{(\ell)} : \ell = 1, \ldots, L\}$. Notice that the MI operator in (4) defines a column space that is orthogonal to $\mathbf{X}_t$. This can be used to ensure nonconfounding between $\boldsymbol{\beta}_t$ and $\boldsymbol{\eta}_t$. Specifically, from the spectral representation $\mathbf{G}(\mathbf{X}_t, \mathbf{A}_t) = \boldsymbol{\Phi}_{X,G,t}\boldsymbol{\Lambda}_{X,G,t}\boldsymbol{\Phi}_{X,G,t}'$, we denote the $N_t \times r$ real matrix formed from the first r columns of $\boldsymbol{\Phi}_{X,G,t}$ as $\mathbf{S}_{X,t}$. Additionally, we set the row of $\mathbf{S}_{X,t}$ that corresponds to variable ℓ and areal unit A equal to $\mathbf{S}_t^{(\ell)}(A)$. Thus, by definition, for each t the $N_t \times p$ matrix of covariates $\mathbf{X}_t$ is linearly independent of the columns of the $N_t \times r$ matrix of basis functions $\mathbf{S}_{X,t}$ and, hence, there are no issues with confounding between $\boldsymbol{\beta}_t$ and $\boldsymbol{\eta}_t$.

It is important to emphasize that the orthogonalization of $\mathbf{X}_t$ to obtain $\mathbf{S}_{X,t}$ is done over the support of the entire spatial region (i.e., $D_{\mathrm{P},t}$), which removes confounded random effects at any prediction location of interest. In principal, one might use an orthogonalization over a subset, say $D \subset D_{\mathrm{P},t}$, and use a different class of basis functions to define $\mathbf{S}_t^{(\ell)}(\cdot)$ at prediction locations outside D. However, in this case prediction locations outside D may suffer from problems with confounding and, hence, inference on the underlying mean $\mu_t^{(\ell)}(\cdot)$ may be incorrect.

4.2. *Moran's I propagator matrix.* The problem of confounding provides motivation for the definition of the MI basis functions $\{\mathbf{S}_{X,t}^{(\ell)}(\cdot)\}$. In a similar manner, the problem of confounding manifests in a spatio-temporal VAR(1) model and can be addressed through careful specification of $\{\mathbf{M}_t\}$. To see this, substitute (3) into (2) to obtain

$$(5) \qquad \mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta}_t + \mathbf{S}_{X,t}\mathbf{M}_t\boldsymbol{\eta}_{t-1} + \mathbf{S}_{X,t}\mathbf{u}_t + \boldsymbol{\xi}_t; \qquad t = 2, \ldots, T,$$

where $\mathbf{y}_t \equiv (Y_t^{(\ell)}(A) : \ell = 1, \ldots, L, A \in D_{\mathrm{P},t}^{(\ell)})'$ and $\boldsymbol{\xi}_t \equiv (\xi_t^{(\ell)}(A) : \ell = 1, \ldots, L, A \in D_{\mathrm{P},t}^{(\ell)})'$ are N_t-dimensional latent random vectors. The specification of $\{\mathbf{S}_{X,t}\}$ using MI basis functions implies that there are no issues with confounding between $\{\boldsymbol{\beta}_t\}$ and $\{\mathbf{u}_t\}$; however, depending on our choice for $\{\mathbf{M}_t\}$, there might be issues with confounding between $\boldsymbol{\eta}_{t-1}$ and the $(p + r)$-dimensional random vector $\boldsymbol{\zeta}_t \equiv (\boldsymbol{\beta}_t', \mathbf{u}_t')'$; $t = 2, \ldots, T$ [although the VAR(1) model assumes $\mathbf{u}_t$ is independent of $\boldsymbol{\eta}_{t-1}$]. Then, rewriting (5), we get

$$(6) \qquad \mathbf{S}_{X,t}'(\mathbf{y}_t - \boldsymbol{\xi}_t) = \mathbf{B}_t\boldsymbol{\zeta}_t + \mathbf{M}_t\boldsymbol{\eta}_{t-1}; \qquad t = 2, \ldots, T,$$

where the $r \times (p + r)$ matrix $\mathbf{B}_t \equiv (\mathbf{S}'_{X,t}\mathbf{X}_t, \mathbf{I})$. The representation in (6) gives rise to what we call the MI propagator matrix, which is defined in an analogous manner to the MI basis functions. Using the spectral representation of $\mathbf{G}(\mathbf{B}_t, \mathbf{I}_r) = \mathbf{\Phi}_{G,B,t}\mathbf{\Lambda}_{G,B,t}\mathbf{\Phi}'_{G,B,t}$, we set the $r \times r$ real matrix $\mathbf{M}_t$ equal to the first r columns of $\mathbf{\Phi}_{G,B,t}$ for each t, which is denoted with $\mathbf{M}_{B,t}$.

Notice that there are no restrictions on $\{\mathbf{M}_{B,t}\}$ to mathematically guarantee that $\mathbf{M}_{B,t}$ does not become "explosive" as t increases. Thus, one should investigate whether or not this is the case when using this model for "long-lead" forecasting. One should also be aware that we do not treat $\{\mathbf{M}_t\}$ as an unknown parameter matrix to be estimated. Instead, we chose a specific form for $\{\mathbf{M}_t\}$, namely, $\{\mathbf{M}_{B,t}\}$, that avoids confounding between $\{\mathbf{\eta}_t\}$ and $\{\mathbf{\zeta}_t\}$. As a result, the final form of $\{\mathbf{M}_{B,t}\}$ might not be spatially interpretable. This issue is addressed in Section 4.3, where constraints are added to the parameter model so that $\text{cov}(\mathbf{\eta}_t) = \mathbf{M}_{B,t}\mathbf{K}_{t-1}\mathbf{M}'_{B,t} + \mathbf{W}_t$ is spatially interpretable. Nevertheless, it is a huge advantage in spatio-temporal modeling to have a known propagator matrix, as a prominent historical challenge with such models is addressing the curse of dimensionality in estimating realistic propagators [Cressie and Wikle (2011), Chapter 7].

4.3. *Parameter models.* Methods for analyzing high-dimensional data (like the QWIs) seek to remove ineffectual or redundant information [for a more in-depth discussion see Sun and Li (2012)]. In Sections 4.1 and 4.2 we impose a reduced rank structure and a nonconfounding property and, as a result, remove information on high frequencies and confounded random effects, respectively. Thus, we specify $\{\mathbf{K}_t\}$ and $\{\mathbf{W}_t\}$ in a manner that offsets these needed computational compromises.

As an example, consider the case where we do not remove confounded random effects. Let $\mathbf{P}_{X,t} \equiv \mathbf{X}_t(\mathbf{X}'_t\mathbf{X}_t)^{-1}\mathbf{X}_t$ and the column space of $\mathbf{P}_{X,t}$ be denoted as $\mathcal{C}(\mathbf{P}_{X,t})$. Rewrite (2) and let $\mathbf{S}_t = [\mathbf{H}_{X,t}, \mathbf{L}_{X,t}]$ and $\mathbf{\eta}_t \equiv (\mathbf{\kappa}'_{X,t}, \mathbf{\delta}'_{X,t})'$ so that

$$(7) \qquad \mathbf{y}_t = \mathbf{X}_t\mathbf{\beta}_t + \mathbf{H}_{X,t}\mathbf{\kappa}_{X,t} + \mathbf{L}_{X,t}\mathbf{\delta}_{X,t} + \mathbf{\xi}_t; \qquad t = 2, \ldots, T.$$

Here, the $N_t \times h$ matrix $\mathbf{H}_{X,t} \in \mathcal{C}(\mathbf{P}_{X,t})^{\perp}$, the $N_t \times l$ matrix $\mathbf{L}_{X,t} \in \mathcal{C}(\mathbf{P}_{X,t})$, h and l are nonnegative integers, $\mathbf{\kappa}_{X,t}$ is a h-dimensional Gaussian random vector, and $\mathbf{\delta}_{X,t}$ is a l-dimensional Gaussian random vector; $t = 2, \ldots, T$. The decomposition in (7) is the space–time analogue of the decomposition used for discussion in Reich, Hodges and Zadnik (2006) and Hughes and Haran (2013). The use of MI basis functions is equivalent to setting h equal to r, $\mathbf{H}_{X,t} = \mathbf{S}_{X,t}$, and $\mathbf{L}_{X,t}$ equal to a $n_t \times l$ matrix of zeros for each t. As a result, the model based on MI basis functions ignores the variability due to $\{\mathbf{\delta}_{X,t}\}$ because it is confounded with $\{\mathbf{\beta}_t\}$. In a similar manner, one can argue that both the reduced rank structure of the MI basis functions and the MI propagator matrix may also ignore other sources of variability.

To address this concern, we consider specifying $\{\mathbf{K}_t\}$ as positive semi-definite matrices that are "close" to target precision matrices (denoted with $\mathbf{P}_t$ for $t = 1, \ldots, T$) that do not ignore these sources of variability; critically, the use of a target precision matrix allows us to reduce the parameter space in a manner that respects the true variability of the process. Specifically, let $\mathbf{K}_t = \sigma_K^2 \mathbf{K}_t^*(\mathbf{P}_t)$, where $\sigma_K^2 > 0$ is unknown and

$$(8) \qquad \mathbf{K}_t^*(\mathbf{P}_t) = \underset{\mathbf{C}}{\arg\min}\{\|\mathbf{P}_t - \mathbf{S}_{X,t}\mathbf{C}^{-1}\mathbf{S}'_{X,t}\|_{\mathrm{F}}^2\}; \qquad t = 1, \ldots, T.$$

Here, $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. In (8), we minimize the Frobenius norm across the space of positive semi-definite matrices. A computable expression of $\mathbf{K}_t^*(\mathbf{P}_t)$ in (8) can be found in Appendix A.

Processes with precision $\mathbf{P}_t$ do not ignore sources of variability like $\boldsymbol{\delta}_{X,t}$ in (7), since $\mathbf{P}_t$ has principal components in $\mathcal{C}(\mathbf{X}_t)$ and principal components associated with high frequencies. Hence, to mitigate the effect of removing certain principal components when defining $\mathbf{S}_t^{(\ell)}(\cdot)$, we specify the $r \times r$ matrix $\mathbf{K}_t$ to be as close as possible [in terms of the Frobenius norm in (8)] to something that has these principal components, namely, the $n_t \times n_t$ matrix $\mathbf{P}_t$. That is, we rescale the total variability of our prior covariance to account for variability ignored for reasons of computation and confounding.

There are many choices for the "target precision" matrices $\{\mathbf{P}_t\}$ in (8). For example, one might use a CAR model and let $\mathbf{P}_t = \mathbf{Q}_t$, where recall $\mathbf{Q}_t = \mathbf{I}_{N_t} - \mathbf{A}_t$; $t = 1, \ldots, T$. This allows one to incorporate neighborhood information into the priors for $\{\mathbf{K}_t\}$. In the case where the areal units are small and regularly spaced, one might consider one of the many spatio-temporal covariance functions that are available [e.g., see Gneiting (1999), Cressie and Huang (1999), and Stein (2005)]. Alternatively, an empirical Bayesian approach might be considered and an estimated precision (or covariance) matrix might be used [e.g., see Sampson and Guttorp (1992)].

The spatial-only case provides additional motivation for the approach in (8). That is, when $T = L = 1$ and $\mathbf{P}_1 = \mathbf{Q}_1$, the prior specification in (8) yields the MI prior introduced in Hughes and Haran (2013). This motivating special case is formally stated and shown in Appendix A.

With both $\{\mathbf{K}_t\}$ and $\{\mathbf{M}_t\}$ specified we can solve for $\{\mathbf{W}_t\}$, that is, using the VAR(1) model

$$(9) \qquad \mathbf{W}_t = \mathbf{K}_t - \mathbf{M}_{B,t}\mathbf{K}_{t-1}\mathbf{M}'_{B,t} \equiv \sigma_K^2 \mathbf{W}_t^*; \qquad t = 2, \ldots, T.$$

In (9), the $r \times r$ matrix $\mathbf{W}_t^* = \mathbf{K}_t^* - \mathbf{M}_{B,t}\mathbf{K}_{t-1}^*\mathbf{M}'_{B,t}$; $t = 2, \ldots, T$. It is important to note that the $r \times r$ matrices in the set $\{\mathbf{W}_t^*\}$ may not be necessarily positive semi-definite. If $\mathbf{W}_t^*$ is not positive semi-definite for some t, then we suggest using the best positive approximation. This is similar to "lifting" adjustments suggested by Kang, Cressie and Shi (2010) in the spatio-temporal setting.

The prior distributions for the remaining parameters are specified so that conjugacy can be used to obtain exact expressions for the full conditionals within a Gibbs sampling algorithm. Specifically, we choose a Gaussian distribution for $\{\boldsymbol{\beta}_t\}$ and inverse gamma (IG) for σ_K^2 and $\{\sigma_{\xi,t}^2\}$. In many cases the statistical agency will provide values for $\{v_t^{(\ell)}(A)\}$ and, thus, no model is required for $\{v_t^{(\ell)}(A)\}$ in this setting. For our motivating QWI example, the LEHD program provides imputation variances for QWIs (http://download.vrdc.cornell.edu/qwipu.experimental/qwiv/beta1/). Imputation variances for QWIs are not available for each county/quarter/industry/gender combination, which is the multivariate spatio-temporal support of the data in Section 2. Thus, we use an IG prior based on the imputation variances that are available. See Appendix B for details regarding the MCMC algorithm, a complete summary of our statistical model, and a discussion on alternative model specifications for related settings.

5. Analysis of quarterly workforce indicators using the MSTM.

In this section we use the MSTM to analyze quarterly average monthly income. In particular, our analysis has two primary goals. The first goal is to demonstrate that the MSTM can reasonably reproduce latent multivariate spatio-temporal fields for the QWI setting. To do this, we perform an "empirical study." Specifically, we perturb a subset of the log quarterly average monthly income (log QWIs), introduced in Section 2, then we test whether or not we can recover the log QWIs using the perturbed version. (Notice that the symmetrizing log transformation is used so that the Gaussian assumptions from Section 3 are met.) An empirical study such as this differs from a traditional simulation study since the emphasis is on illustrating that the MSTM can reproduce values similar to quarterly average monthly income. Therefore, in Section 5.1 we introduce our empirical study design and in Section 5.2 we provide the results of our empirical study.

Our second goal in this section is to establish that the MSTM can be efficiently used to jointly model high-dimensional areal data (see Section 2 for a discussion). The methodological development in Sections 3 and 4 are motivated by striking a balance between modeling realistic multivariate spatio-temporal dependencies and allowing for the possibility of extremely high-dimensional data sets. As such, in Section 5.3 we jointly analyze all 7,530,037 quarterly average monthly income estimates provided by the LEHD program.

For Sections 5.1 through 5.3, the Gibbs sampler, provided in Appendix B, was run for 10,000 iterations with a burn-in of 1000 iterations. Convergence of the Markov chain Monte Carlo algorithm was assessed visually using trace plots of the sample chains, with no lack of convergence detected. Additionally, the batch means estimate of the Monte Carlo error (with batch size 50) [e.g., see Jones et al. (2006), Roberts (1996)] and the Gelman–Rubin diagnostic (computed using three chains) [e.g., see Gelman and Rubin (1992)] did not suggest lack of convergence.

5.1. *Empirical study design.* Abowd et al. (2009) provide a study to assess the quality of the QWIs. Thus, for consistency within the literature we adopt a study design similar to the one used in Section 5.7.2 of Abowd et al. (2009). Specifically, we restrict the data to $t = 4, \ldots, 55$ (quarters between 1991 and 2003), $\ell = 1, 2$ (which represents women and men in the education industry, respectively), and the prediction locations equal the counties in Minnesota that have available QWIs (i.e., $D_{\mathrm{P},t}^{(\ell)} \equiv D_{\mathrm{MN},t}^{(\ell)}$). The scope of this empirical study is smaller than the entire data set introduced in Section 2, since in this section we are primarily interested in showing that the MSTM can recover latent multivariate spatio-temporal fields similar to the quarter average monthly income. See Section 5.3 for a demonstration of using the MSTM to efficiently jointly model the entire 7,530,037 QWIs.

The perturbed version of the log quarterly average monthly income is explicitly written as

$$(10) \quad R_t^{(\ell)}(A) = Z_t^{(\ell)}(A) + \varepsilon_t^{(\ell)}(A); \qquad t = 4, \ldots, 55, \ell = 1, 2, A \in D_{\mathrm{MN},t}^{(\ell)},$$

where $D_{\mathrm{MN},t}^{(\ell)}$ is the set of counties in Minnesota (MN) that have available quarterly average monthly income estimates, $\{R_t^{(\ell)}(A)\}$ represents the perturbed version of the log quarterly average monthly income [log QWIs; denoted by $\{Z_t^{(\ell)}(\cdot)\}$], and the set $\{\varepsilon_t^{(\ell)}(A) : t = 4, \ldots, 55, \ell = 1, 2, A \in D_{\mathrm{O},t}^{(\ell)}\}$ consists of i.i.d. normal random variables with mean zero and variance σ_ε^2. In practice, the quarterly average monthly income estimates are publicly available and are, hence, observed. Nevertheless, for the purposes of this empirical study we will act as if the QWIs are an unobserved multivariate spatio-temporal field to be estimated, and treat $\{R_t^{(\ell)}\}$ as the data process and $\{Z_t^{(\ell)}(\cdot)\}$ as the latent process.

We randomly select 65% of the areal units in $D_{\mathrm{MN},t}^{(\ell)}$ to be "observed," which we denote with the set $D_{\mathrm{MN},\mathrm{O},t}^{(\ell)}$. Thus, for this example, $D_{\mathrm{P},t}$ (given by $D_{\mathrm{MN},t}^{(\ell)}$) and $D_{\mathrm{O},t}$ (given by $D_{\mathrm{MN},\mathrm{O},t}^{(\ell)}$) are not the same. Recall from Section 2 that this choice reflects the amount of observed data present in the entire QWI data set, where 65% of the QWIs are observed. However, it is important to note that the "missing QWI" structure of the data set in Section 2 is different from what we use in this empirical study, since we do not incorporate missing QWIs patterns that occur due to a state's failure to sign a MOU. Recall that if a state does not sign a MOU for a particular year, then the entire state is missing for that year. However, our choice to randomly select 65% of the areal units within $D_{\mathrm{MN},t}^{(\ell)}$ to be "observed" is sufficient for our purposes.

The value for the perturbation variance σ_ε^2 is chosen relative to the variability of the log quarterly average monthly income. The variance of the log quarterly average monthly income, within our study region, is given by $\mathrm{var}\{Z_t^{(\ell)}(A)\} = 0.24$. Thus, we specify the perturbations $\{\varepsilon_t^{(\ell)}(A) : A \in D_{\mathrm{O},t}^{(\ell)}\}$ to have variance $\sigma_\varepsilon^2 \equiv 0.24$. This yields a signal-to-noise ratio of 1, which can be interpreted as a small
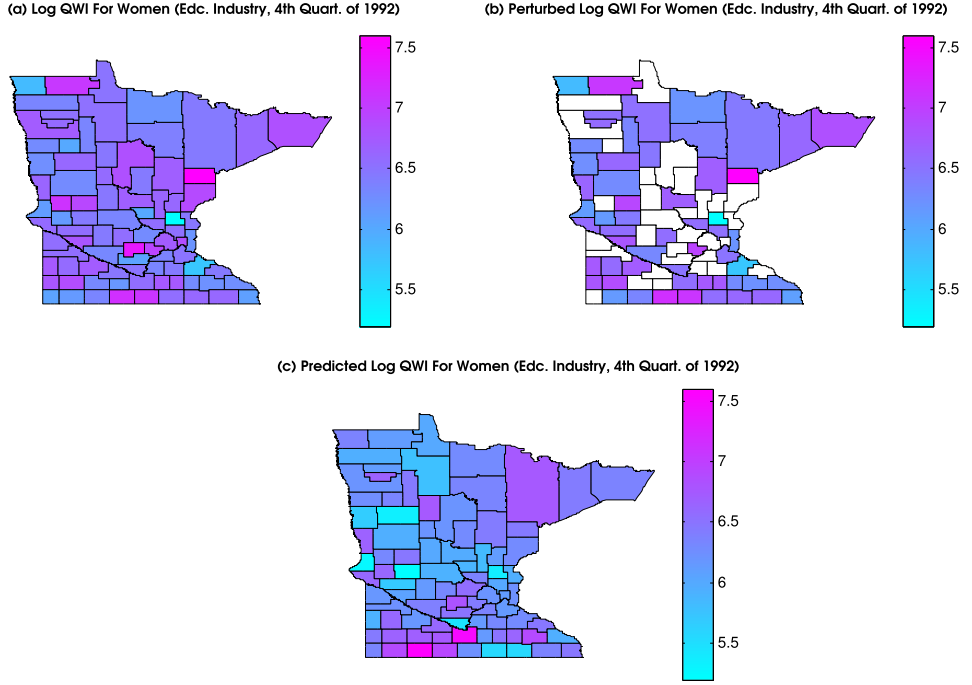
(a) Log QWI For Women (Edc. Industry, 4th Quart. of 1992)   **(b) Perturbed Log QWI For Women (Edc. Industry, 4th Quart. of 1992)**

(c) Predicted Log QWI For Women (Edc. Industry, 4th Quart. of 1992)

FIG. 2.   (a) *Map of the log quarterly average monthly income for women in the education industry* [*i.e.*, $\{Z_8^{(1)}(A)\}$]. *These values correspond to log quarterly average monthly income for women in the education industry, for counties in Minnesota, and for the* 4*th quarter in* 1992. *For comparison, a map of the perturbed log quarterly average monthly income for women in the education industry* [*i.e.*, $\{W_8^{(1)}(A)\}$] *is given in* (b). *The white areas indicate missing regions. In* (c) *we provide the predictions* $\{\widehat{Z}_8^{(1)}(A)\}$ *that are computed using MSTM and the perturbed data* $\{R_t^{(\ell)}(A)\}$ *from equation* (10).

signal-to-noise ratio. We argue that this choice is conservative, since small signal-to-noise ratios traditionally make prediction of a latent process difficult [Aldworth and Cressie (1999)].

We end this section with an example of analyzing a single realization of $\{R_t^{(\ell)}(A) : t = 4, \ldots, 55, \ell = 1, 2, A \in D_{\mathrm{MN,O},t}^{(\ell)}\}$. Consider the selected maps of the log quarterly average monthly income and the perturbed log average monthly income in Figure 2(a) and (b), respectively. Figure 2 visually depicts the difficulty of predicting a latent random field, as the number of "missing" QWIs is rather large and the signal-to-noise ratio is visibly small.

To use the MSTM to predict $\{Z_t^{(\ell)}\}$ from $\{R_t^{(\ell)}\}$, we need to specify the target precision matrix, the covariates, and the number of MI basis functions. Set the target precision matrix equal to $\{\mathbf{Q}_t\}$ as previously described below (8). Let $\mathbf{x}_t^{(\ell)}(A) \equiv 1$, where $g = 1, 2$ indexes men and women, respectively. Also, for illustration let $r = 30$, which is roughly 50% of the available MI basis functions at each time point t. In a sensitivity study (not shown), we see that the MSTM is relatively
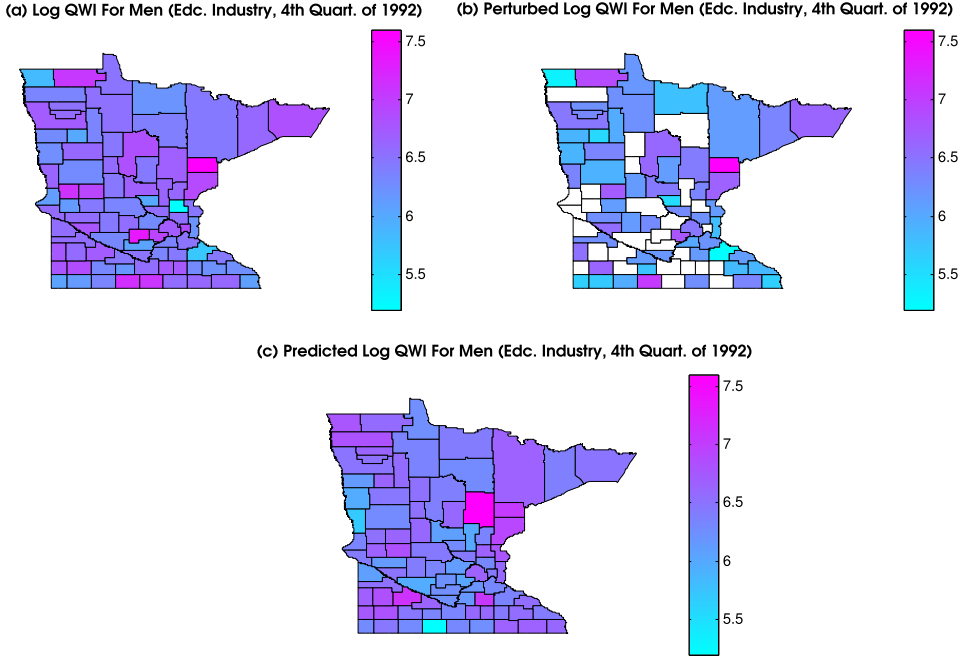
(a) Log QWI For Men (Edc. Industry, 4th Quart. of 1992)

(b) Perturbed Log QWI For Men (Edc. Industry, 4th Quart. of 1992)

(c) Predicted Log QWI For Men (Edc. Industry, 4th Quart. of 1992)

FIG. 3. (a) *Map of the log quarterly average monthly income for men in the education industry* [*i.e.*, $\{Z_8^{(2)}(A)\}$]. *These values correspond to log quarterly average monthly income for men in the education industry, for counties in Minnesota, and for the* 4*th quarter in* 1992. *For comparison, a map of the perturbed log quarterly average monthly income for men in the education industry* [*i.e.*, $\{W_8^{(2)}(A)\}$] *is given in* (b). *The white areas indicate missing regions. In* (c) *we provide the predictions* $\{\widehat{Z}_8^{(2)}(A)\}$ *that are computed using MSTM and the perturbed data* $\{R_t^{(\ell)}(A)\}$ *from equation* (10).

robust to changes to larger values of r. In general, for the purposes of prediction, large values of r are preferable; however, a carefully selected reduced rank set of basis functions can produce as good or better predictions than those based on the full set of basis functions [Bradley, Cressie and Shi (2011, 2014, 2015)]. Using the MSTM with these specifications, we predict $Z_t^{(\ell)}$ using the perturbed values $R_t^{(\ell)}$. In Figure 2(c) we present $\{\widehat{Z}_8^{(1)}(A) : A \in D_{O,8}^{(1)}\}$. In general, we let $\widehat{Z}_t^{(\ell)}$ denote the MSTM predictions based on $\{R_t^{(\ell)}(A) : t = 4, \ldots, 55, \ell = 1, 2, A \in D_{\text{MN},O,t}^{(\ell)}\}$. Similar conclusions are drawn from Figure 3, which provides results for men.

The performance of our predictions are further corroborated by the results presented in Figure 4(a) and (b), where we map the percent relative difference (PRD) between the predicted log quarterly average monthly income and the actual log quarterly average monthly income. That is, the values plotted in Figure 4(a) and (b) are given by

$$(11) \qquad \text{abs}\left\{\frac{\widehat{Z}_8^{(\ell)}(A) - Z_8^{(\ell)}(A)}{Z_8^{(\ell)}(A)}\right\} \times 100\%; \qquad \ell = 1, 2, A \in D_{O,8}^{(\ell)}.$$
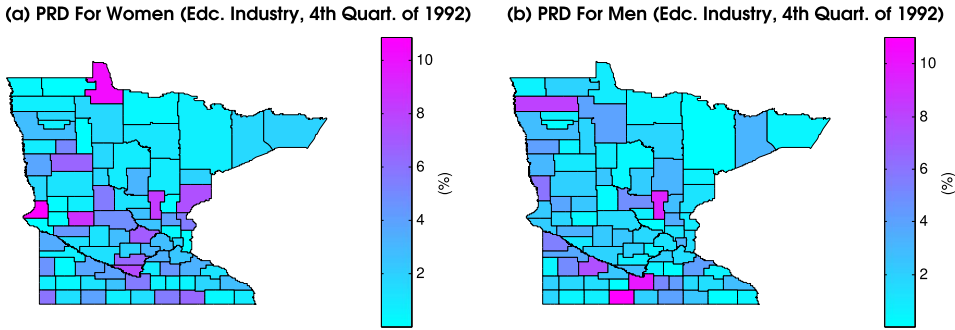
FIG. 4.   *In* (a) *and* (b), *the percent relative difference* (*PRD*) *in* (11) *of the predicted log quarterly monthly average income of women and men within the education industry during the fourth quarter of* 1992.

Additionally, the median PRD across all variables, regions, and time points is 4.87%. Hence, for this example the difference between the predicted and actual log quarterly average monthly income is small relative to the scale of the log quarterly average monthly income. Thus, we appear to be efficiently reproducing the unobserved latent field (as measured by PRD) using the MSTM.

5.2. *Empirical study of multiple replicates*.   There have been no statistical methods used to obtain QWI estimates and measures of precision at missing regions. Thus, in this section we evaluate the performance of $\{\widehat{Z}_t^{(\ell)}\}$ at both observed and missing regions over multiple replicates.

The MSTM from Section 3 is currently the only stochastic modeling approach available to jointly model high-dimensional multivariate spatio-temporal areal data. Since there are no viable alternative methods available, we first assess the quality of the predictions relative to the scale of the data [e.g., see equation (11)]. Specifically, consider the median percent relative difference (MPRD) given by

$$\text{MPRD} \equiv \text{median}\left\{\text{abs}\left[\frac{\widehat{Z}_t^{(\ell)}(A) - Z_t^{(\ell)}(A)}{Z_t^{(\ell)}(A)}\right] \times 100 : \right.$$

(12)

$$\left. t = 4, \ldots, 55, \ell = 1, 2, A \in D_{\text{O},t}^{(\ell)}\right\}.$$

If MPRD in (12) is "close" to zero for a given replicate of the field $\{R_t^{(\ell)}(A) : t = 4, \ldots, 55, \ell = 1, 2, A \in D_{\text{MN,O},t}^{(\ell)}\}$, then the predictions are considered close (relative to the scale of the data) to the log quarterly average monthly income. In Figure 5(a) we provide boxplots [over 50 independent replicates of $\{R_t^{(\ell)}\}$] of MPRD evaluated at observed and missing regions, respectively. Here, we see that the MPRD is larger at missing regions as expected. However, the values of the MPRD are consistently small for both observed and missing regions: the medians
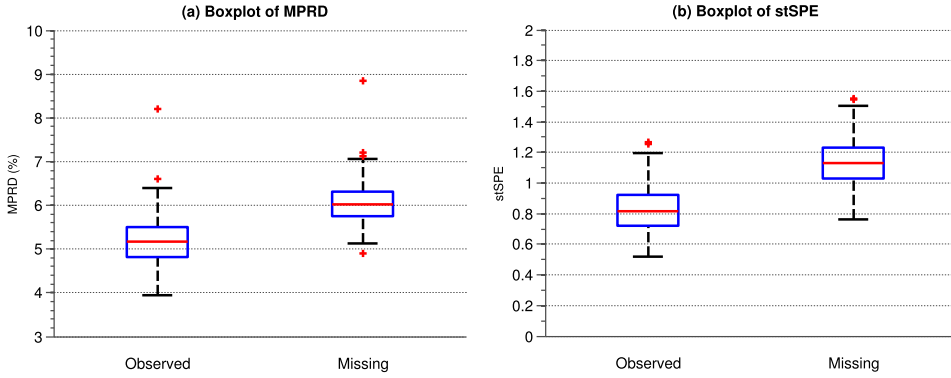
FIG. 5. (a) *Boxplots of the MPRD in* (12), *using the* 50 *replicates of the spatial field* $\{R_t^{(\ell)}(A) : t = 4, \ldots, 55, \ell = 1, 2, A \in D_{\mathrm{MO},\mathrm{O},t}^{(\ell)}\}$, *for observed and missing respectively.* (b) *Boxplots of the stSPE in* (13), *using the* 50 *replicates of the spatial field* $\{R_t^{(\ell)}(A) : t = 4, \ldots, 55, \ell = 1, 2, A \in D_{\mathrm{MO},\mathrm{O},t}^{(\ell)}\}$, *for observed and missing respectively. In both* (a) *and* (b) *we do not plot outliers for the purposes of visualization.*

are given by 5.17% and 6.02% for observed and missing regions, respectively; and the interquartile ranges are given by 0.6915 and 0.5470 for observed and missing regions, respectively. Thus, the MPRD shows that we are obtaining predictions that are close (relative to the scale of the log QWIs) to the log quarterly average monthly income.

Another metric that one might use to validate our conclusions from Figure 5(a) is the standardized squared prediction error (stSPE)

$$\text{(13)} \quad \begin{aligned} \text{stSPE} = \text{average}\big\{\big(\widehat{Z}_t^{(\ell)}(A) - Z_t^{(\ell)}(A)\big)^2 : \\ t = 4, \ldots, 55, \ell = 1, 2, A \in D_{\mathrm{O},t}^{(\ell)}\big\}/\sigma_\varepsilon^2. \end{aligned}$$

If stSPE in (13) is "close" to zero for a given replicate of the field $\{R_t^{(\ell)}(A) : t = 4, \ldots, 55, \ell = 1, 2, A \in D_{\mathrm{MN},\mathrm{O},t}^{(\ell)}\}$, then the predictions are considered close to the log quarterly average monthly income. Also notice that the stSPE in (13) is normalized by σ_ε^2; consequently, we can compare the squared error of our predictions relative to the perturbation variances. This is especially noteworthy for predictions at missing regions, which have no signal in the original perturbed data set.

In Figure 5(b) we provide boxplots [over 50 independent replicates of $\{R_t^{(\ell)}\}$] stSPE evaluated at observed and missing regions, respectively. Here, we see that the MSPE is larger at missing regions as expected. However, the values of the stSPE at observed (missing) regions are consistently smaller (close) than 1: the medians are given by 0.8154 and 1.1293 for observed and missing regions, respectively; and the interquartile ranges are given by 0.1994 and 0.1990 for observed and missing regions, respectively. Thus, the stSPE shows that the error in our pre-

dictions at observed (missing) regions are smaller than (similar to) the perturbation error (i.e., σ_ε^2).

Notice that the stSPE is roughly 0.1293 above 1 at missing locations and 0.1846 below 1 at observed locations; thus, the relative differences from 1 are similar in the two situations. This may be problematic if there are more missing values than observed. However, note that this is not the case for the LEHD data set, which has roughly 65% of the prediction locations observed.

5.3. *Predicting quarterly average monthly income.* We demonstrate the use of MSTM using a high-dimensional multivariate spatio-temporal data set made up of quarterly average monthly income obtained from the LEHD program. In particular, we consider all 7,530,037 observations introduced in Section 2. These values are available over the entire US, which we jointly analyze using the MSTM. We present a subset of this data set in Figure 6(a) and (b). We see that the quarterly average monthly income is relatively constant across each county of the state of Missouri and that men tend to have higher quarterly average monthly income than women. This pattern is consistent across the different spatial locations, industries, and time points.

The primary goals of our analysis in this section is to estimate the quarterly average monthly income, investigate potential gender inequality in the US, and determine whether or not it is computationally feasible to use the MSTM for a data set of this size. Preliminary analyses indicate that the log quarterly average monthly income is roughly Gaussian. Since we assume that the underlying data is Gaussian, we treat the log of the average income as $\{Z_t^{(\ell)}(\cdot)\}$ in (1).

For illustration, we make the following specifications. Set the target precision matrix equal to $\{\mathbf{Q}_t\}$ as previously described below (8). Let $\mathbf{x}_t^{(\ell)}(A) \equiv (1, I(\ell = 1), \ldots, I(\ell = 39), I(g = 1) \times I(\ell = 1), \ldots, I(g = 1) \times I(\ell = 39))'$, where $g = 1, 2$ indexes men and women, respectively, and recall $I(\cdot)$ is the indicator function. Also, following the MSTM specifications from our empirical study, we let $r = 30$, which is roughly 50% of the available MI basis functions at each time point t. Using the MSTM with these specifications, we predict $L \times T = 40 \times 92 = 3680$ different spatial fields. The CPU time required to compute these predictions is approximately 1.2 days, with all of our computations performed in Matlab (Version 8.0) on a dual 10 core 2.8 GHz Intel Xeon E5-2680 v2 processor, with 256 GB of RAM. Of course, additional efforts in efficient programming may result in faster computing; however, these results indicate that it is computationally practical to use the MSTM to analyze massive data.

Although we modeled the entire US simultaneously, for illustration, we present maps of predicted monthly income for the state of Missouri, for each gender, for the education industry, and for the 92-nd quarter [Figure 6(c) and (d)]. The prediction maps are essentially constant over the state of Missouri, where women tend to
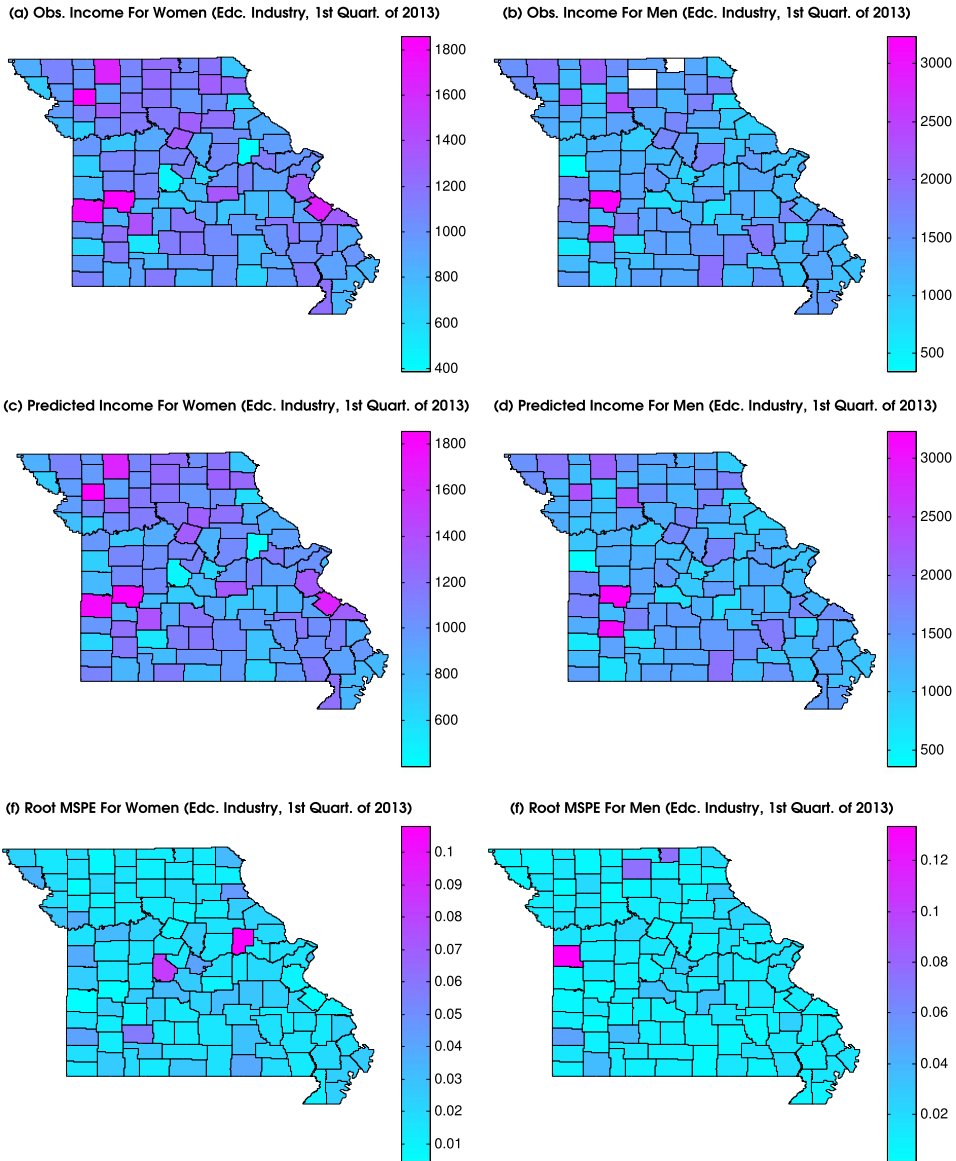
FIG. 6.   (a) *and* (b) *present the QWI for quarterly average monthly income* (*US dollars*) *for the state of Missouri, for each gender, for the education industry, and for the first quarter of* 2013. *LEHD does not provide estimates at every county in the US at every quarter; these counties are shaded white.* (c)–(f) *present the corresponding maps* (*for the state of Missouri, for each gender, for the education industry, and for quarter* 92) *of predicted monthly income* (*US dollars*) *and their respective posterior square root MSPE* (*on the log-scale*). *Notice that the color scales are different for women and men and that the root MSPEs are computed on the log-scale. White areas indicate missing regions.*

have a predicted monthly income of slightly less than 1200 dollars and men consistently have a predicted monthly income of about 1800 dollars. As observed in Figure 6(a) and (b), there is a clear pattern where men have higher predicted monthly income than women. These predictions appear reasonable since the maps of the root MSPE (on the log scale), in Figure 6(e) and (f), indicate we are obtaining precise predictions on the log-scale. Additionally, upon comparison of Figure 6(a) and (b) to Figure 6(c) and (d), we see that the predictions reflect the same general pattern in the data. These results are similar across the different states, industries, and time points.

To further corroborate the patterns in the MSTM predictions, we fit a separate univariate spatial model from Hughes and Haran (2013). Specifically, we fit the univariate spatial model from Hughes and Haran (2013) to the data in Figure 6(a) and (b) with $r = 62$ basis functions (100% of the available basis functions) and obtain the prediction maps (not shown). Notably, the predictions are also fairly constant around 1200 and 1800 dollars. Moreover, the MSPE of the Hughes and Haran (2013) predictions (summed over all US counties) is 4.09 times larger than the MSPE of the predictions from the MSTM summed over all US counties. This may be due, in part, to the fact that the model in Hughes and Haran (2013) does not incorporate multivariate and serial (temporal) dependencies.

The large difference in average monthly income between men and women can be further investigated by comparing the means [i.e., $\boldsymbol{\mu}_t^{(\ell)}(\cdot)$] for men and women, respectively. [Recall from Section 4.1 that we can perform inference on $\boldsymbol{\mu}_t^{(\ell)}(\cdot)$ because we impose a nonconfounding property between $\boldsymbol{\mu}_t^{(\ell)}(\cdot)$ and $\mathbf{S}_t^{(\ell)}(\cdot)'\boldsymbol{\eta}_t$.] Now, let $m_1, \ldots, m_{20}$ indicate industry 1 through 20 for men, and $w_1, \ldots, w_{20}$ for women. Then, for a given A consider the contrast given by $\sum_{k=1}^{20} \mu_{92}^{(m_k)}(A) - \sum_{k=1}^{20} \mu_{92}^{(w_k)}(A)$, which is interpreted as an average difference between the income of men and women over the 20 industries. Hence, this contrast is a global (across industries) measure of income gender differences at the most current time point (notice $t = 92$). A positive (negative) value indicates that men (women) tend to have larger incomes. In Figure 7(a) and (b) we plot the posterior mean and variance of this contrast by state. Here, we see that for the first quarter of 2013, gender inequality is similar across each state (with men consistently having larger quarterly incomes), with the largest disparity occurring in Arizona.

Figure 7(a) and (b) give a sense of the spatial patterns of the between-gender income differences for the first quarter of 2013. We can also investigate the temporal and between-industry patterns in a similar manner. In particular, in Figure 8(a) we plot $\sum_{k,A} \boldsymbol{\mu}_t^{(m_k)}(A)$ and $\sum_{k,A} \boldsymbol{\mu}_t^{(w_k)}(A)$ by quarter (i.e., t). Here, we see that the differences between the genders appears to be constant from 1990 to 2013. Likewise, in Figure 8(b) we identify between industry differences by plotting the posterior mean of $\sum_{t,A} \boldsymbol{\mu}_t^{(m_k)}(A)$ and $\sum_{t,A} \boldsymbol{\mu}_t^{(w_k)}(A)$ by industry (i.e., k). Here, we
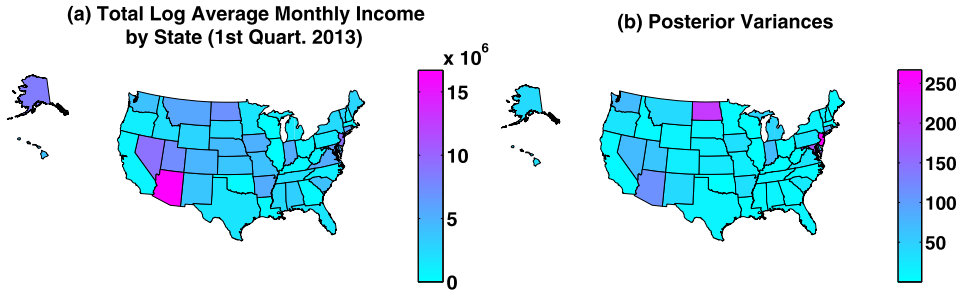
FIG. 7. *Plots of contrasts of* $\boldsymbol{\mu}_t^{(\ell)}(A)$ *(referred to as the log average monthly income). In* (a) *and* (b), *we plot the posterior mean and variance of* $\sum \boldsymbol{\mu}_{92}^{(m_k)}(A) - \sum \boldsymbol{\mu}_{92}^{(w_k)}(A)$ *by state, where the sum aggregates counties to states and is also indexed over the industries.*

observe that gender inequality appears present in each industry, with men consistently having larger mean average monthly income. That is, the posterior mean of $\sum_{t,A} \boldsymbol{\mu}_t^{(m_k)}(A)$ and the values within 95% (pointwise) credible intervals are larger than that for women. Furthermore, we see that the largest difference between log average monthly income occurs in the finance and insurance industries, which also appear to be the most lucrative industries for men.

It should be noted that, despite the inherent computational issues, having an abundance of data has distinct advantages. For example, notice in Figure 6(b) that LEHD does not release data at two counties of Missouri for men in the education industry during quarter 92. Although these values are missing for this variable and time point, LEHD releases QWIs at these two counties (for men in the education industry) for 43 different quarters. Hence, with the observed values from 43 different spatial fields, we reduce the variability of predictions at the two missing counties during the 92nd quarter [compare Figure 6(b) to (f)]. This is particularly useful for the setting when a states does not sign a MOU and, hence, LEHD does not provide estimates here.

6. Discussion. We have introduced fully Bayesian methodology to analyze areal data sets with multivariate spatio-temporal dependencies. In particular, we introduce the multivariate spatio-temporal mixed effects model (MSTM). To date, little has been proposed to model areal data that exhibit multivariate spatio-temporal dependencies. Furthermore, the available alternatives [see Carlin and Banerjee (2003) and Daniels, Zhou and Zou (2006)] do not allow for certain complexities in cross-covariances and fail to accommodate high-dimensional data sets. Hence, the MSTM provides an important addition to the multivariate spatio-temporal literature.

The MSTM was motivated by the Longitudinal Employer-Household Dynamics (LEHD) program's quarterly workforce indicators (QWI) [Abowd et al. (2009)]. In particular, the QWIs are extremely high-dimensional and exhibit complex multivariate spatio-temporal dependencies. Thus, extensive methodological contribu-
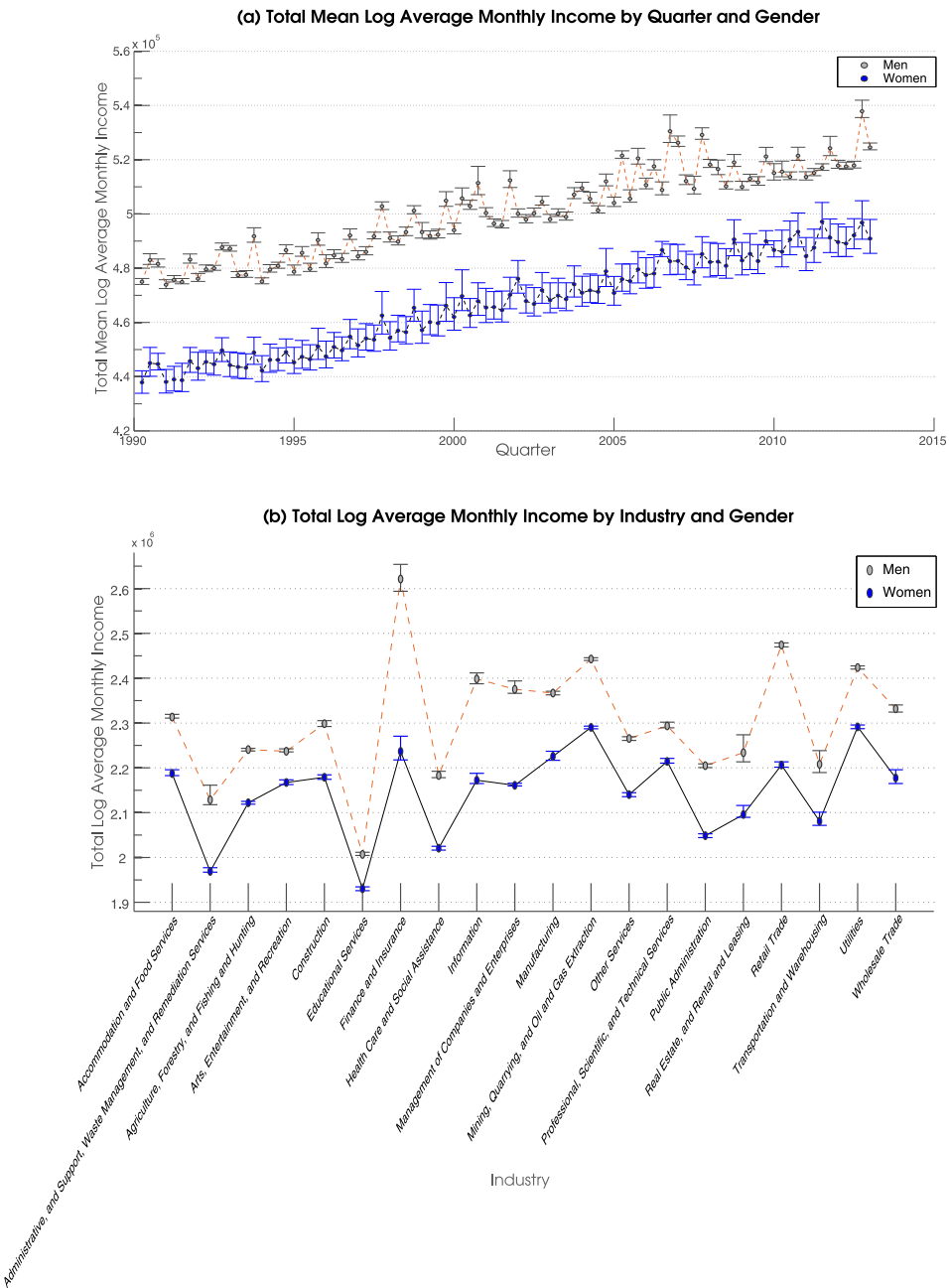
(a) Total Mean Log Average Monthly Income by Quarter and Gender

(b) Total Log Average Monthly Income by Industry and Gender

FIG. 8.   *Plots of contrasts of $\boldsymbol{\mu}_t^{(\ell)}(A)$* (*referred to as the log average monthly income*). *In* (a), *we plot the posterior mean of $\sum_{k,A} \boldsymbol{\mu}_t^{(m_k)}(A)$ and $\sum_{k,A} \boldsymbol{\mu}_t^{(w_k)}(A)$ by quarter. In* (b), *we plot the posterior mean of $\sum_{t,A} \boldsymbol{\mu}_t^{(m_k)}(A)$ and $\sum_{t,A} \boldsymbol{\mu}_t^{(w_k)}(A)$ by industry. In both* (a) *and* (b) *a 95% credible interval is given by horizontal line segments, and for comparison a line is superimposed connecting the intervals associated with males and females, respectively.*

tions, leading to the MSTM, were necessary in order to realistically, jointly model the QWIs' complex multivariate spatio-temporal dependence structure and to allow for the possibility of remarkably high-dimensional areal data.

We conducted an extensive empirical study to demonstrate that the MSTM works extremely well for predicting the QWI, quarterly average monthly income. Specifically, we perturb the log quarterly average monthly income, then predictions of the log quarterly average monthly income are made using the perturbed values and comparisons are made between the predicted and the actual log quarterly average monthly income. The results illustrate that we are consistently recovering the unobserved latent field using the MSTM at both observed and missing regions. This is particularly noteworthy, since there are no other methods that have been used to estimate QWIs at missing regions. In fact, because we borrow strength over different variables, space, and time, we can also predict values for entire states when the values are missing for reasons of an unsigned MOU.

The exceptional effectiveness of our approach is further illustrated through a joint analysis of all the available quarterly average monthly income estimates. This data set, comprised of 7,530,037 observations, is used to predict 3680 different spatial fields consisting of all the counties in the US. The recorded CPU time for this example was 1.2 days, which clearly indicates that it is practical to use the MSTM in high-dimensional data contexts.

In this article, we have found that incorporating different variables, space, and time into an analysis is beneficial for two reasons. First, one can leverage information from nearby (in space and time) observations and related variables to improve predictions and, second, there are inferential questions that are unique to multivariate spatio-temporal processes. For example, in Section 5.3 it was of interest to determine where, when, and what industry had the largest disparity between the average quarterly income of men and women. Here, we found that these differences have been relatively constant over the last two decades, are currently fairly constant over each state, and are the highest within the finance and insurance industries.

Although our emphasis was on analyzing QWIs, our modeling framework allows the MSTM to be applied to a wide array of data sets. For example, the MSTM employs a reduced rank approach to allow for massive multivariate spatio-temporal data sets. Additionally, the MSTM allows for nonstationary and nonseparable multivariate spatio-temporal dependencies. This is achieved, in part, through a novel propagator matrix for a first-order vector autoregressive [VAR(1)] model, which we call the MI propagator matrix. This propagator matrix is an extension of the MI basis function [Griffith (2000, 2002, 2004), Griffith and Tiefelsdorf (2007), Hughes and Haran (2013), Porter, Wikle and Holan (2015)] from the spatial-only setting to the multivariate spatio-temporal setting. We motivate both the MI basis function and the MI propagator matrix as an approximation to a target precision matrix, that allows for both computationally efficient statistical inference and non-confounding regression parameters.

Our model specification also allows for knowledge of the underlying spatial process to be incorporated into the MSTM. Specifically, we propose an extension of the MI prior to the spatio-temporal case. This extension forces the covariance matrix of the random effect to be close (in Frobenius norm) to a "target precision" matrix, which can be chosen based on knowledge of the underlying spatial process. Importantly, this contribution has broader implications, in terms of reducing a parameter space, for defining informative parameter models for high-dimensional spatio-temporal processes.

There are many opportunities for future research. For example, there are many QWIs available that are recorded as counts, which do not satisfy the Gaussian assumption even after a transformation. Thus, the MSTM could be extended to the Poisson data setting. The parameter model introduced in Section 4.3 is also of independent interest. In our applications, we let $\{\mathbf{Q}_t\}$ be the target precision. However, one could conceive of various different "target precisions" built from deterministic models (e.g., for atmospheric variables). Another avenue for future research is to extend the MI propagator matrix, beyond the VAR(1) specification. In fact, this strategy could be easily used for many subject matter domains for other time series models.

APPENDIX A: TECHNICAL RESULTS

PROPOSITION 1. *Let $\mathbf{\Phi}_k$ be a generic $n \times r$ real matrix such that $\mathbf{\Phi}_k'\mathbf{\Phi}_k = \mathbf{I}_r$, $\mathbf{C}$ be a generic $r \times r$ positive definite matrix, $\mathbf{P}_k$ be a generic $n \times n$ positive definite matrix, and let $k = 1, \ldots, K$. Then, the value of $\mathbf{C}$ that minimizes $\sum_{k=1}^{K} \|\mathbf{P}_k - \mathbf{\Phi}_k\mathbf{C}^{-1}\mathbf{\Phi}_k'\|_F^2$ within the space of positive semi-definite covariances is given by*

$$(A.1) \qquad \mathbf{C}^* = \left\{ \mathcal{A}^+\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{\Phi}_k'\mathbf{P}_k\mathbf{\Phi}_k\right)\right\}^{-1},$$

where $\mathcal{A}^+(\mathbf{R})$ is the best positive approximate [Higham (1988)] of a real square matrix $\mathbf{R}$. Similarly, the value of $\mathbf{C}$ that minimizes $\sum_{k=1}^{K} \|\mathbf{P}_k - \mathbf{\Phi}_k\mathbf{C}\mathbf{\Phi}_k'\|_F^2$ within the space of positive semi-definite covariances is given by

$$(A.2) \qquad \mathcal{A}^+\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{\Phi}_k'\mathbf{P}_k\mathbf{\Phi}_k\right).$$

PROOF. By definition of the Frobenius norm,

$$\sum_{k=1}^{K}\|\mathbf{P}_k - \mathbf{\Phi}_k\mathbf{C}^{-1}\mathbf{\Phi}_k'\|_F^2$$

$$= \sum_{k=1}^{K}\text{trace}\{(\mathbf{P}_k - \mathbf{\Phi}_k\mathbf{C}^{-1}\mathbf{\Phi}_k')'(\mathbf{P}_k - \mathbf{\Phi}_k\mathbf{C}^{-1}\mathbf{\Phi}_k')\}$$

$$(A.3) \qquad = \sum_{k=1}^{K} \{\operatorname{trace}(\mathbf{P}_k' \mathbf{P}_k) - 2 \times \operatorname{trace}(\mathbf{\Phi}_k' \mathbf{P}_k' \mathbf{\Phi}_k \mathbf{C}^{-1}) + \operatorname{trace}(\mathbf{C}^{-2})\}$$

$$= \sum_{k=1}^{K} \operatorname{trace}(\mathbf{P}_k' \mathbf{P}_k) - K \times \operatorname{trace}\left\{\left(\frac{1}{K}\sum_{k=1}^{K} \mathbf{\Phi}_k' \mathbf{P}_k \mathbf{\Phi}_k\right)^2\right\}$$

$$+ K \times \left\| \mathbf{C}^{-1} - \frac{1}{K}\sum_{k=1}^{K} \mathbf{\Phi}_k' \mathbf{P}_k \mathbf{\Phi}_k \right\|_F^2.$$

It follows from Theorem 2.1 of Higham (1988) that the minimum of (A.3) is given by equation (A.1) in the main document. In a similar manner, if one substitutes $\mathbf{C}$ for $\mathbf{C}^{-1}$ in (A.3), then we obtain the result in equation (A.2) in the main document. $\square$

PROPOSITION 2. *Let $\mathbf{S}_{X,1}$ be the MI propagator matrix and $\mathbf{C}$ be a generic $r \times r$ positive definite matrix. Then, the value of $\mathbf{C}$ that minimizes $\|\mathbf{Q}_1 - \mathbf{S}_{X,1}\mathbf{C}\mathbf{S}_{X,1}'\|_F^2$ within the space of positive semi-definite covariances is given by*

$$(A.4) \qquad\qquad \mathbf{C}^* = \mathcal{A}^+(\mathbf{S}_{X,1}' \mathbf{Q}_1 \mathbf{S}_{X,1}).$$

PROOF. The proof of Proposition 2 follows immediately from Proposition 1. Specifically, let $K = 1$, $\mathbf{\Phi}_1 = \mathbf{S}_{X,1}$, and $\mathbf{P}_1 = \mathbf{Q}_1$. Then, apply Proposition 1. If $\mathbf{S}_{X,1}' \mathbf{Q}_1 \mathbf{S}_{X,1}$ is positive definite, then (17) leads to the prior specification in Hughes and Haran (2013). Porter, Holan and Wikle (2015) show that $\mathbf{S}_{X,1}' \mathbf{Q}_1 \mathbf{S}_{X,1}$ is positive definite as long as an intercept is included in the definition of $\mathbf{X}_1$. $\square$

APPENDIX B: FULL CONDITIONAL DISTRIBUTIONS

The model that we use for multivariate spatio-temporal data is given by

Data model: $Z_t^{(\ell)}(A) | \boldsymbol{\beta}_t, \boldsymbol{\eta}_t, \xi_t^{(\ell)}(\cdot)$

$$\overset{\text{ind}}{\sim} \operatorname{Normal}(\mathbf{x}_t^{(\ell)}(A)' \boldsymbol{\beta}_t + \mathbf{S}_{X,t}^{(\ell)}(A)' \boldsymbol{\eta}_t + \xi_t^{(\ell)}, \upsilon_t^{(\ell)}(A));$$

Process model 1: $\boldsymbol{\eta}_t | \boldsymbol{\eta}_{t-1}, \mathbf{M}_{B,t}, \mathbf{W}_t \sim \operatorname{Gaussian}(\mathbf{M}_{B,t} \boldsymbol{\eta}_{t-1}, \mathbf{W}_t)$;

Process model 2: $\boldsymbol{\eta}_1 | \mathbf{K}_1 \sim \operatorname{Gaussian}(\mathbf{0}, \mathbf{K}_1)$;

Process model 3: $\xi_t^{(\ell)}(\cdot) | \sigma_{\xi,t}^2 \overset{\text{ind}}{\sim} \operatorname{Normal}(0, \sigma_{\xi,t}^2)$;

Parameter model 1: $\delta^{(k)} \sim \operatorname{IG}(\alpha_\upsilon, \beta_\upsilon)$;

Parameter model 2: $\boldsymbol{\beta}_t \sim \operatorname{Gaussian}(\boldsymbol{\mu}_\beta, \sigma_\beta^2 \mathbf{I}_p)$;

Parameter model 3: $\sigma_{\xi,t}^2 \sim \operatorname{IG}(\alpha_\xi, \beta_\xi)$;

Parameter model 4: $\sigma_K^2 \sim \text{IG}(\alpha_K, \beta_K)$;

$$\ell = 1, \ldots, L, t = T_L^{(\ell)}, \ldots, T_U^{(\ell)}, k = 1, 2, A \in D_{O,t}^{(\ell)},$$

where $\sigma_\beta^2 > 0$, $\alpha_v > 0$, $\alpha_\xi > 0$, $\alpha_K > 0$, $\beta_v > 0$, $\beta_\xi > 0$, and $\beta_K > 0$. In Sections 5 and 6 the prior mean of $\boldsymbol{\mu}_\beta$ is set equal to a p-dimensional zero vector, and the corresponding variance σ_β^2 is set equal to 10^{15} so that the prior on $\{\boldsymbol{\beta}_t\}$ is vague. In Sections 5 and 6 we also specify α_ξ, α_K, β_v, β_ξ, and β_K so that the prior distributions of $\sigma_{\xi,t}^2$ and $\sigma_{K,t}^2$ are vague. Specifically, we let $\alpha_\xi = \alpha_K = 2$, and $\beta_v = \beta_\xi = \beta_K = 1$; here, the IG(2, 1) prior is interpreted as vague since it has infinite variance.

We now specify the full conditional distributions for the process variables [i.e., $\{\boldsymbol{\eta}_t\}$ and $\{\xi_t^{(\ell)}(\cdot)\}$] and the parameters [i.e., $\{v_t^{(\ell)}(\cdot)\}$, $\{\boldsymbol{\beta}_t\}$, $\{\sigma_{\xi,t}^2\}$, and σ_K^2].

Full conditionals for process variables. Let the n_t-dimensional random vectors $\mathbf{z}_t \equiv (Z_t^{(\ell)}(A) : \ell = 1, \ldots, L, A \in D_{O,t}^{(\ell)})'$, $\boldsymbol{\xi}_t \equiv (\xi_t^{(\ell)}(A) : \ell = 1, \ldots, L, A \in D_{O,t}^{(\ell)})'$, and the $n_t \times p$ matrix $\mathbf{X}_t \equiv (\mathbf{x}_t^{(\ell)}(A) : \ell = 1, \ldots, L, A \in D_{O,t}^{(\ell)})'$; $t = 1, \ldots, T$. Then, we update the full conditional for $\boldsymbol{\eta}_{1:T} \equiv (\boldsymbol{\eta}_t' : t = 1, \ldots, T)'$ at each iteration of the Gibbs sampler using the Kalman smoother. We accomplish this by performing the following steps:

1. Find the Kalman filter using the shifted measurements $\{\widetilde{\mathbf{z}}_t : \widetilde{\mathbf{z}}_t = \mathbf{z}_t - \mathbf{X}_t \boldsymbol{\beta}_t - \boldsymbol{\xi}_t\}$ [Carter and Kohn (1994), Cressie and Wikle (2011), Frühwirth-Schnatter (1994), Shumway and Stoffer (2006)]. That is, for $t = 1, \ldots, T$ compute

(a) $\boldsymbol{\eta}_{t|t}^{[j]} \equiv E(\boldsymbol{\eta}_t | \widetilde{\mathbf{z}}_{1:t}, \boldsymbol{\theta}_t^{[j]})$,

(b) $\boldsymbol{\eta}_{t|(t-1)}^{[j]} \equiv E(\boldsymbol{\eta}_t | \widetilde{\mathbf{z}}_{1:(t-1)}, \boldsymbol{\theta}_t^{[j]})$,

(c) $\mathbf{P}_{t|t}^{[j]} \equiv \text{cov}(\boldsymbol{\eta}_t | \widetilde{\mathbf{z}}_{1:t}, \boldsymbol{\theta}_t^{[j]})$,

(d) $\mathbf{P}_{t|(t-1)}^{[j]} \equiv \text{cov}(\boldsymbol{\eta}_t | \widetilde{\mathbf{z}}_{1:(t-1)}, \boldsymbol{\theta}_t^{[j]})$,

where $\mathbf{P}_{1|1}^{[j]} = (\sigma_K^{[j]})^2 \mathbf{K}^*$ and $\boldsymbol{\theta}_t^{[j]}$ represents the jth MCMC draw of $\boldsymbol{\theta}_t$ and σ_K^2, respectively.

2. Sample $\boldsymbol{\eta}_T^{[j+1]} \sim \text{Gaussian}(\boldsymbol{\eta}_{T|T}^{[j]}, \mathbf{P}_{T|T}^{[j]})$.

3. For $t = T - 1, T - 2, \ldots, 1$ sample

$$\boldsymbol{\eta}_t^{[j+1]} \sim \text{Gaussian}(\boldsymbol{\eta}_{t|t}^{[j]} + \mathbf{J}_t^{[j]}(\boldsymbol{\eta}_{t+1}^{[j]} - \boldsymbol{\eta}_{t+1|t}^{[j]}), \mathbf{P}_{t|t}^{[j]} - \mathbf{J}_t^{[j]} \mathbf{P}_{t+1|t}^{[j]} (\mathbf{J}_t^{[j]})'),$$

where $\mathbf{J}_t^{[j]} \equiv \mathbf{P}_{t|t}^{[j]} \mathbf{M}_t' (\mathbf{P}_{t+1|t}^{[j]})^{-1}$.

Notice that within each MCMC iteration we need to compute the Kalman filter and Kalman smoothing equations. This adds more motivation for reduced rank

modeling, that is, if r is large (i.e., if r is close in value to n), this step is not computationally feasible.

The full conditional for the remaining process variable $\{\xi_t^{(\ell)}(\cdot)\}$ can also be computed efficiently [Ravishanker and Dey (2002)]. The full conditional for $\{\xi_t^{(\ell)}(\cdot)\}$ is given by $\boldsymbol{\xi}_t \sim \text{Gaussian}(\boldsymbol{\mu}_{\xi,t}^*, \boldsymbol{\Sigma}_{\xi,t}^*)$, where $\boldsymbol{\Sigma}_{\xi,t}^* \equiv (\mathbf{V}_t^{-1} + \sigma_\xi^{-2}\mathbf{I}_{N_t})^{-1}$, $\boldsymbol{\mu}_{\xi,t}^* \equiv \boldsymbol{\Sigma}_\xi^* \times \mathbf{V}_t^{-1} \times (\mathbf{z}_t - \mathbf{X}_t\boldsymbol{\beta}_t - \mathbf{S}_t\boldsymbol{\eta}_t)$, $\mathbf{V}_t \equiv \text{diag}(v_t^{(\ell)}(A) : \ell = 1, \ldots, L, A \in D_{\mathrm{O},t}^{(\ell)})$, and $\mathbf{S}_t \equiv (\mathbf{S}_t^{(\ell)}(A) : \ell = 1, \ldots, L, A \in D_{\mathrm{O},t}^{(\ell)})'$; $t = 1, \ldots, T$.

Full conditionals for the parameters. Similar to the full conditional for $\{\xi_t^{(\ell)}(\cdot)\}$ [Ravishanker and Dey (2002)], we also have the following full conditional for $\boldsymbol{\beta}_t$: $\boldsymbol{\beta}_t \sim \text{Gaussian}(\boldsymbol{\mu}_{\beta,t}^*, \boldsymbol{\Sigma}_{\beta,t}^*)$, where $\boldsymbol{\Sigma}_{\beta,t}^* \equiv (\mathbf{X}_t'\mathbf{V}_t^{-1}\mathbf{X}_t + \sigma_\beta^{-2}\mathbf{I}_p)^{-1}$, and $\boldsymbol{\mu}_{\beta,t}^* \equiv \boldsymbol{\Sigma}_\beta^* \times \mathbf{X}_t'\mathbf{V}_t^{-1}(\mathbf{z}_t - \boldsymbol{\xi}_t - \mathbf{S}_t\boldsymbol{\eta}_t)$; $t = 1, \ldots, T$. The exact form of the full conditionals for σ_K^2 and $\{\sigma_{\xi,t}^2\}$ can also be found in a straightforward manner. It follows that the full conditionals for σ_K^2 and $\sigma_{\xi,t}^2$ are $\text{IG}(\text{Tr}/2 + 2, 1 + \boldsymbol{\eta}_1'\mathbf{K}_1^{*-1}\boldsymbol{\eta}_1/2 + \sum_{t=2}^T (\boldsymbol{\eta}_t - \mathbf{M}_t\boldsymbol{\eta}_{t-1})'\mathbf{W}_t^{*-1}(\boldsymbol{\eta}_t - \mathbf{M}_t\boldsymbol{\eta}_{t-1})/2)$ and $\text{IG}(n/2 + 2, 1 + \boldsymbol{\xi}_t'\boldsymbol{\xi}_t/2)$ (for $t = 1, \ldots, T$), respectively.

Imputation variances for QWIs are not currently available for each county/quarter/industry/gender combination, which is the multivariate spatio-temporal support of the data in Section 2. Thus, we specify a prior distribution for $\{v_t^{(\ell)}(A)\}$ that capitalizes on the available information, namely, imputation variances defined for QWIs given at each county/quarter/industry combination. Denote these imputation variances with $\tilde{v}_t^{(m)}(\cdot)$, where $m = 1, \ldots, 20$ and $t = 1, \ldots, 92$. This leads us to our prior for $\{v_t^{(\ell)}(A)\}$ given by

$$v_t^{(\ell)}(A) = \begin{cases} \tilde{v}_t^{(\ell)}(A)\delta^{(1)}/\exp\{2Z_t^{(\ell)}(A)\}, \\ \quad \text{if } \ell = 1, \ldots, 20, \\ \tilde{v}_t^{(\ell-20)}(A)\delta^{(2)}/\exp\{2Z_t^{(\ell)}(A)\}, \\ \quad \text{if } \ell = 21, \ldots, 40; t = 1, \ldots, 92, A \in D_{\mathrm{O},t}^{(\ell)}, \end{cases}$$

where $\delta^{(k)} > 0$ for $k = 1, 2$, and we let $\ell = 1, \ldots, 20$ indicate men in each of the 20 industries and $\ell = 21, \ldots, 40$ indicate women in each of the 20 industries, respectively. We divide by $\exp\{2Z_t^{(\ell)}(A)\}$ to transform $\tilde{v}_t^{(\ell)}$ to the log-scale; specifically, we use the delta method [see Oehlert (1992), among others] to transform the variances to the log-scale. Thus, our model for the variances $\{v_t^{(\ell)}(A)\}$ is a simple reweighting (by weights in $\{\delta^{(k)}\}$) of the imputation variances (on the log-scale) obtained from the LEHD program. We note that our predictions are relatively robust to this specification.

In the empirical study in Sections 5.1 and 5.2, we use the known value of $v_t^{(\ell)}(A)$ and, hence, no distribution was placed on $\delta^{(1)}$ and $\delta^{(2)}$. In many cases this is reasonable since the statistical agency provides values for $v_t^{(\ell)}(A)$.

In Section 5.3 we let $\delta^{(k)} \sim \text{IG}(1, 2)$; $k = 1, 2$. Now, let $\ell = 1, \ldots, 20$ indicate the spatial fields corresponding to each of the 20 industries for men and $\ell = 21, \ldots, 40$ indicate the spatial fields corresponding to each of the 20 industries for women. The full conditionals for $\delta^{(1)}$ and $\delta^{(2)}$ are $\text{IG}(M/2 + 2, 1 + \sum_{\ell=1}^{20} \sum_{t=1}^{92} \sum_{A \in D_{O,t}^{(\ell)}} (Z_t^{(\ell)}(A) - \mathbf{x}_t^{(\ell)}(A)'\boldsymbol{\beta}_t - \mathbf{S}_{X,t}^{(\ell)}(A)'\boldsymbol{\eta}_t - \xi_t^{(\ell)})^2 / 2\tilde{v}_t^{(\ell)}(A))$ and $\text{IG}(F/2 + 2, 1 + \sum_{\ell=21}^{40} \sum_{t=1}^{92} \sum_{A \in D_{O,t}^{(\ell)}} (Z_t^{(\ell)}(A) - \mathbf{x}_t^{(\ell)}(A)'\boldsymbol{\beta}_t - \mathbf{S}_{X,t}^{(\ell)}(A)'\boldsymbol{\eta}_t - \xi_t^{(\ell)})^2 / 2\tilde{v}_t^{(\ell)}(A))$, where $M \equiv \sum_{\ell=1}^{20} \sum_{t=1}^{92} n_t^{(\ell)}$ and $F \equiv \sum_{\ell=21}^{40} \sum_{t=1}^{92} n_t^{(\ell)}$.

In some settings, survey error variances are not provided. The case of unknown survey variance leads to interesting and difficult modeling questions. In particular, when $\text{var}(\varepsilon_t^{(\ell)}) = v_t^{(\ell)}(\cdot)$ is unknown, there may be issues with identifiability between $\sigma_{\xi,t}^2$ and $v_t^{(\ell)}(\cdot)$ when $v_t^{(\ell)}(\cdot)$ is roughly constant across variables and locations [see Bradley, Cressie and Shi (2015), for a discussion]. To avoid this issue of identifiability, one might combine $\xi(\cdot)$ and $\varepsilon_t^{(\ell)}(\cdot)$, and then estimate the sums $\xi(\cdot) + \varepsilon_t^{(\ell)}(\cdot)$ and $v_t^{(\ell)}(\cdot) + \sigma_{\xi,t}^2$, respectively. In the environmental context, others have addressed this identifiability problem by avoiding the use of likelihoods and adopting a moment-based approach to estimate $v_t^{(\ell)}(\cdot)$; specifically, see Kang, Cressie and Shi (2010) and Katzfuss and Cressie (2012) for the definition of a variogram-extrapolation technique to estimate $v_t^{(\ell)}(\cdot)$ and Kang, Cressie and Shi (2010) for a method of moments estimator.

REFERENCES

ABOWD, J., SCHNEIDER, M. and VILHUBER, L. (2013). Differential privacy applications to Bayesian and linear mixed model estimation. *Journal of Privacy and Confidentiality* **5** 73–105.

ABOWD, J., STEPHENS, B., VILHUBER, L., ANDERSSON, F., MCKINNEY, K., ROEMER, M. and WOODCOCK, S. (2009). The LEHD infrastructure files and the creation of the Quarterly Workforce Indicators. In *Producer Dynamics*: *New Evidence from Micro Data* (T. Dunne, J. Jensen and M. Roberts, eds.) 149–230. Univ. Chicago Press, Chicago.

ALDWORTH, J. and CRESSIE, N. (1999). Sampling designs and prediction methods for Gaussian spatial processes. In *Multivariate Analysis*, *Design of Experiments*, *and Survey Sampling. Statist. Textbooks Monogr.* **159** 1–54. Dekker, New York. MR1719054

ALLEGRETTO, S., DUBE, A., REICH, M. and ZIPPERER, B. (2013). Credible research designs for minimum wage studies. Working paper series 1–63, Institute for Research on Labor and Employment.

BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall, London, UK.

BANERJEE, S., GELFAND, A. E., FINLEY, A. O. and SANG, H. (2008). Gaussian predictive process models for large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 825–848. MR2523906

BANERJEE, S., FINLEY, A. O., WALDMANN, P. and ERICSSON, T. (2010). Hierarchical spatial process models for multiple traits in large genetic trials. *J. Amer. Statist. Assoc.* **105** 506–521. MR2724841

BELL, W. and HILLMER, S. (1990). The time series approach to estimation for repeated surveys. *Surv. Methodol.* **16** 195–215.

BRADLEY, J. R., CRESSIE, N. and SHI, T. (2011). Selection of rank and basis functions in the spatial random effects model. In *Proceedings of the* 2011 *Joint Statistical Meetings* 3393–3406. American Statistical Association, Alexandria, VA.

BRADLEY, J. R., CRESSIE, N. and SHI, T. (2014). A comparison of spatial predictors when datasets could be very large. Preprint. Available at arXiv:1410.7748.

BRADLEY, J. R., CRESSIE, N. and SHI, T. (2015). Comparing and selecting spatial predictors using local criteria. *TEST* **24** 1–28 (Rejoinder, pp. 54–60). MR3314567

CARLIN, B. P. and BANERJEE, S. (2003). Hierarchical multivariate CAR models for spatio-temporally correlated survival data. In *Bayesian Statistics*, 7 (*Tenerife*, 2002) 45–63. Oxford Univ. Press, New York. MR2003166

CARTER, C. K. and KOHN, R. (1994). On Gibbs sampling for state space models. *Biometrika* **81** 541–553. MR1311096

CONGDON, P. (2002). A multivariate model for spatio-temporal health outcomes with an application to suicide mortality. *Geogr. Anal.* **36** 235–258.

CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*, rev. ed. Wiley, New York. MR1239641

CRESSIE, N. and HUANG, H.-C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *J. Amer. Statist. Assoc.* **94** 1330–1340. MR1731494

CRESSIE, N. and JOHANNESSON, G. (2008). Fixed rank kriging for very large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 209–226. MR2412639

CRESSIE, N. and WIKLE, C. K. (2011). *Statistics for Spatio-Temporal Data*. Wiley, Hoboken, NJ. MR2848400

DANIELS, M. J., ZHOU, Z. and ZOU, H. (2006). Conditionally specified space-time models for multivariate processes. *J. Comput. Graph. Statist.* **15** 157–177. MR2269367

DAVIS, E., FREEDMAN, M., LANE, J., McCALL, B., NESTORIAK, N. and PARK, T. (2006). Supermarket human resource practices and competition from mass merchandisers. *Am. J. Agric. Econ.* **88** 1289–1295.

DUBE, A., LESTER, T. and REICH, M. (2013). Minimum wage, labor market flows, job turnover, search frictions, monopsony, unemployment. Working paper series 1–63, Institute for Research on Labor and Employment.

FEDER, M. (2001). Time series analysis of repeated surveys: The state-space approach. *Stat. Neerl.* **55** 182–199. MR1862486

FINLEY, A. O., SANG, H., BANERJEE, S. and GELFAND, A. E. (2009). Improving the performance of predictive process modeling for large datasets. *Comput. Statist. Data Anal.* **53** 2873–2884. MR2667597

FRÜHWIRTH-SCHNATTER, S. (1994). Data augmentation and dynamic linear models. *J. Time Series Anal.* **15** 183–202. MR1263889

GELMAN, A. and RUBIN, D. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 473–511.

GNEITING, T. (1999). Correlation functions for atmospheric data analysis. *Q. J. R. Meteorol. Soc.* **125** 2449–2464.

GRIFFITH, D. (2000). A linear regression solution to the spatial autocorrelation problem. *J. Geogr. Syst.* **2** 141–156.

GRIFFITH, D. A. (2002). A spatial filtering specification for the auto-Poisson model. *Statist. Probab. Lett.* **58** 245–251. MR1920751

GRIFFITH, D. (2004). A spatial filtering specification for the auto-logistic model. *Environ. Plann. A* **36** 1791–1811.

GRIFFITH, D. and TIEFELSDORF, M. (2007). Semiparametric filtering of spatial autocorrelation: The eigenvector approach. *Environ. Plann. A* **39** 1193–1221.

HIGHAM, N. J. (1988). Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra Appl.* **103** 103–118. MR0943997

HUGHES, J. and HARAN, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 139–159. MR3008275

JONES, R. G. (1980). Best linear unbiased estimators for repeated surveys. *J. Roy. Statist. Soc. Ser. B* **42** 221–226. MR0583360

JONES, G. L., HARAN, M., CAFFO, B. S. and NEATH, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *J. Amer. Statist. Assoc.* **101** 1537–1547. MR2279478

KANG, E. L., CRESSIE, N. and SHI, T. (2010). Using temporal variability to improve spatial mapping with application to satellite data. *Canad. J. Statist.* **38** 271–289. MR2682762

KATZFUSS, M. and CRESSIE, N. (2012). Bayesian hierarchical spatio-temporal smoothing for very large datasets. *Environmetrics* **23** 94–107. MR2873787

LINDGREN, F., RUE, H. and LINDSTRÖM, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 423–498. MR2853727

OEHLERT, G. W. (1992). A note on the delta method. *Amer. Statist.* **46** 27–29. MR1149146

PETTITT, A. N., WEIR, I. S. and HART, A. G. (2002). A conditional autoregressive Gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. *Stat. Comput.* **12** 353–367. MR1951708

PORTER, A. T., HOLAN, S. H. and WIKLE, C. K. (2015). Bayesian semiparametric hierarchical empirical likelihood spatial models. *J. Statist. Plann. Inference* **165** 78–90. MR3350260

PORTER, A. T., WIKLE, C. K. and HOLAN, S. H. (2015). Small area estimation via multivariate Fay-Herriot models with latent spatial dependence. *Aust. N. Z. J. Stat.* **57** 15–29. MR3335325

RAVISHANKER, N. and DEY, D. K. (2002). *A First Course in Linear Model Theory*. Chapman & Hall/CRC, Boca Raton, FL.

REICH, B. J., HODGES, J. S. and ZADNIK, V. (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics* **62** 1197–1206. MR2307445

ROBERTS, G. O. (1996). Markov chain concepts related to sampling algorithms. In *Markov Chain Monte Carlo in Practice* (W. Gilks, S. Richardson and D. Spiegelhalter, eds.). *Interdiscip. Statist.* 45–57. Chapman & Hall, London. MR1397967

ROYLE, J., BERLINER, M., WIKLE, C. and MILLIFF, R. (1999). A hierarchical spatial model for constructing wind fields from scatterometer data in the Labrador sea. In *Case Studies in Bayesian Statistics* (R. Kass, B. Carlin, A. Carriquiry, A. Gelman, I. Verdinelli and M. West, eds.) 367–382. Springer, New York.

SAMPSON, P. and GUTTORP, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *J. Amer. Statist. Assoc.* **87** 108–119.

SENGUPTA, A., CRESSIE, N., FREY, R. and KAHN, B. (2012). Statistical modeling of MODIS cloud data using the spatial random effects model. In *Proceedings of the Joint Statistical Meetings* 3111–3123. American Statistical Association, Alexandria, VA.

SHUMWAY, R. H. and STOFFER, D. S. (2006). *Time Series Analysis and Its Applications: With R Examples*, 2nd ed. Springer, New York. MR2228626

STEIN, M. L. (2005). Space-time covariance functions. *J. Amer. Statist. Assoc.* **100** 310–321. MR2156840

STEIN, M. L. (2014). Limitations on low rank approximations for covariance matrices of spatial data. *Spat. Stat.* **8** 1–19. MR3326818

SUN, Y. and LI, B. (2012). Geostatistics for large datasets. In *Space-Time Processes and Challenges Related to Environmental Problems* (E. Porcu, J. M. Montero and M. Schlather, eds.) 55–77. Springer, Berlin.

THOMPSON, J. (2009). Using local labor market data to re-examine the employment effects of the minimum wage. *Ind. Labor Relat. Rev.* **63** 343–366.

TZALA, E. and BEST, N. (2008). Bayesian latent variable modelling of multivariate spatio-temporal variation in cancer mortality. *Stat. Methods Med. Res.* **17** 97–118. MR2420192

WIKLE, C. K. (2010). Low-rank representations for spatial processes. In *Handbook of Spatial Statistics* 107–118. CRC Press, Boca Raton, FL. MR2730946

ZHU, J., EICKHOFF, J. C. and YAN, P. (2005). Generalized linear latent variable models for repeated measures of spatially correlated multivariate data. *Biometrics* **61** 674–683. MR2196155

DEPARTMENT OF STATISTICS
UNIVERSITY OF MISSOURI
146 MIDDLEBUSH HALL
COLUMBIA, MISSOURI 65211
USA
E-MAIL: bradleyjr@missouri.edu