Chapter 11

Spatiotemporal modeling

In both theoretical and applied work, spatiotemporal modeling has received dramatically increased attention in the past few years. The reason is easy to see: the proliferation of data sets that are both spatially and temporally indexed, and the attendant need to understand them. For example, in studies of air pollution, we are interested not only in the spatial nature of a pollutant surface, but also in how this surface changes over time. Customarily, ongoing temporal measurements (e.g., hourly, daily, three-day average, etc.) are collected at monitoring sites yielding long-time series of data. Similarly, with climate data we may be interested in spatial patterns of temperature or precipitation at a given time, but also in dynamic patterns in weather. With real estate markets, we might be interested in how the single-family home sales market changes on a quarterly or annual basis. Here an additional wrinkle arises in that we do not observe the *same* locations for each time period; the data are cross-sectional, rather than longitudinal.

Applications with areal unit data are also commonplace. For instance, we may look at annual lung cancer rates by county for a given state over a number of years to judge the effectiveness of a cancer control program. Or we might consider daily asthma hospitalization rates by zip code, over a period of several months.

From a methodological point of view, the introduction of time into spatial modeling brings a substantial increase in the scope of our work, as we must make separate decisions regarding spatial correlation, temporal correlation, and how space and time interact in our data. Such modeling will also carry an obvious associated increase in notational and computational complexity.

As in previous chapters, we make a distinction between the cases where the geographical aspect of the data is at point level versus where it is at areal unit level. Again the former case is typically handled via Gaussian process models, while the latter often uses CAR specifications. A parallel distinction could be drawn for the temporal scale: Is time viewed as continuous (say, over $\Re^+$ or some subinterval thereof) or discrete (hourly, daily, etc.)? In the former case there is a conceptual measurement at each moment t. But in the latter case, we must determine whether each measurement should be interpreted as a block average over some time interval (analogous to block averaging in space), or whether it should be viewed merely as a measurement, e.g., a count attached to an associated time interval (and thus analogous to an areal unit measurement). Relatedly, when time is discretized, are we observing a time series of spatial data, e.g., the same points or areal units in each time period (as would be the case in our climate and pollution examples)? Or are we observing cross-sectional data, where the locations change with time period (as in our real estate setting)? In the case of time series, we could regard the data as a multivariate measurement vector at each location or areal unit. We could then employ multivariate spatial data models as in the previous chapter. With short series, this might be reasonable; with longer series, we would likely want to introduce aspects of usual time series modeling.

The nature and location of missing data is another issue that we have faced before, yet becomes doubly complicated in the spatiotemporal setting. The major goal of our earlier

kriging methods is to impute missing values at locations for which no data have been observed. Now we may encounter time points for which we lack spatial information, locations for which information is lacking for certain (possibly future) time points, or combinations thereof. Some of these combinations will be extrapolations (e.g., predicting future values at locations for which no data have been observed) that are statistically riskier than others (e.g., filling in missing values at locations for which we have data at some times but not others). Here the Bayesian hierarchical approach is particularly useful, since it not only helps organize our thinking about the model, but also fully accounts for all sources of uncertainty, and properly delivers wider confidence intervals for predictions that are "farther" from the observed data (in either space or time).

Our Atlanta data set (Figure 7.2) illustrates the sort of misalignment problem we face in many spatiotemporal settings. Here the number of ozone monitoring stations is small (just 8 or 10), but the amount of data collected from these stations over time (92 summer days for each of three years) is substantial. In this case, under suitable modeling assumptions, we may not only learn about the temporal nature of the data, but also enhance our understanding of the spatial process.

In the next few sections we consider the case of point-level spatial data, so that point-point and point-block realignment can be contemplated as in Section 7.1. We initially focus on relatively simple *separable* forms for the space-time correlation, but also consider more complex forms that do not impose the strong restrictions on space-time interaction that separability implies. We subsequently move on to spatiotemporal modeling for data where the spatial component can only be thought of as areal (block) level.

11.1 General modeling formulation

11.1.1 Preliminary analysis

Before embarking on a general spatiotemporal modeling formulation, consider the case of point-referenced data where time is discretized to customary integer-spaced intervals. We may look at a spatiotemporally indexed datum $Y(\mathbf{s},t)$ in two ways. Writing $Y(\mathbf{s},t) = Y_{\mathbf{s}}(t)$, it is evident that we have a spatially varying time series model. Writing $Y(\mathbf{s},t) = Y_t(\mathbf{s})$, we instead have a temporally varying spatial model.

In fact, with locations $\mathbf{s}_i$, $i = 1,\ldots,n$ and time points $t = 1,\ldots,T$, we can collect the data into Y, an $n \times T$ matrix. Column averages of Y produce a space-averaged time series, while row averages yield a time-averaged spatial realization. In fact, suppose we center each column of Y by the vector of row averages and call the resulting matrix $\widetilde{Y}_{rows}$. Then clearly $\widetilde{Y}_{rows}\mathbf{1}_T = \mathbf{0}$, but also $\frac{1}{T}\widetilde{Y}_{rows}\widetilde{Y}_{rows}^T$ is an $n \times n$ matrix that is the sample spatial covariance matrix. Similarly, suppose we center each row of Y by the vector of column averages and call the resulting matrix $\widetilde{Y}_{cols}$. Now $\mathbf{1}_n^T\widetilde{Y}_{cols} = \mathbf{0}$ and $\frac{1}{n}\widetilde{Y}_{cols}^T\widetilde{Y}_{cols}$ is the $T \times T$ sample autocorrelation matrix.

One could also center Y by the grand mean of the $Y(\mathbf{s},t)$. Indeed, to examine residual spatiotemporal structure, adjusted for the mean, one could fit a suitable OLS regression to the $Y(\mathbf{s},t)$ and examine $\widehat{E}$, the matrix of residuals $\widehat{e}(\mathbf{s},t)$. As above, $\frac{1}{T}\widehat{E}\widehat{E}^T$ is the residual spatial covariance matrix while $\frac{1}{n}\widehat{E}^T\widehat{E}$ is the residual autocorrelation matrix.

We can create the singular value decomposition (Harville, 1997) for any of the foregoing matrices. Using E which would be most natural in practice to consider spatiotemporal structure, we can write

$$E = UDV^T = \sum_{l=1}^{\min(n,T)} d_l \mathbf{u}_l \mathbf{v}_l^T , \tag{11.1}$$

where U is an $n \times n$ orthogonal matrix with columns $\mathbf{u}_l$, V is a $T \times T$ orthogonal matrix with columns $\mathbf{v}_l$, and D is an $n \times T$ matrix of the form $\binom{\Delta}{0}$ where Δ is $T \times T$ diagonal

with diagonal entries d_l, $l = 1, \ldots, T$. Without loss of generality, we can assume the d_l's are arranged in decreasing order of their absolute values. Then, $\mathbf{u}_l \mathbf{v}_l^T$ is referred to as the lth *empirical orthogonal function* (EOF) since $\mathbf{u}_l \mathbf{v}_l^T \perp \mathbf{u}_m \mathbf{v}_m^T, l \neq m$ and $(\mathbf{u}_l \mathbf{v}_l^T)^T \mathbf{u}_l \mathbf{v}_l^T = 1$.

Thinking of $\mathbf{u}_l = (u_l(\mathbf{s}_1), \ldots, u_l(\mathbf{s}_n))^T$ and $\mathbf{v}_l = (v_l(1), \ldots, v_l(T))^T$, the expression in (11.1) represents the observed data as a sum of products of spatial and temporal variables, i.e., $E(\mathbf{s}_i, t) = \sum d_l u_l(\mathbf{s}_i) v_l(t)$. Suppose we approximate E by its first EOF, that is, $E \approx d_1 \mathbf{u}_1 \mathbf{v}_1^T$. Then we are saying that $E(\mathbf{s}_i, t) \approx d_1 u_1(\mathbf{s}_i) v_1(t)$, i.e., the spatiotemporal process can be approximated by a product of a spatial process and a temporal process. If the u_1 and v_1 processes are mean 0 (as they would be for modeling residuals) and independent, this implies a *separable* covariance function for $E(\mathbf{s}, t)$ (see (11.18)).[1] Indeed, if the first term in the sum in (11.1) explains much of the residual matrix E, this is often taken as evidence for specifying a separable model. In any event, it does yield a reduction in dimension, introducing $n + T$ variables to represent E, rather than nT. Adding the second EOF yields the approximation $E(\mathbf{s}_i, t) \approx d_1 u_1(\mathbf{s}_i) v_1(t) + d_2 u_2(\mathbf{s}_i) v_2(t)$, a representation involving only $2(n + T)$ variables, and so on.

Note that, if, say, $T < n$,

$$EE^T = UDD^T U^T = U \begin{pmatrix} \Delta^2 & 0 \\ 0 & 0 \end{pmatrix} U^T = \sum_{l=1}^{T} d_l^2 \mathbf{u}_l \mathbf{u}_l^T ,$$

clarifying the interpretation of the d_l's. (Of course, $EE^T = V^T D^T D V = V^T \Delta^2 V$ as well.) Altogether, when applicable, EOFs provide an exploratory tool for learning about spatial structure and suggesting models, in the spirit of the tools described in Section 2.3. For full inference, however, we require a full spatiotemporal model specification, the subject to which we now turn.

11.1.2 Model formulation

Modeling for spatiotemporal data can be given a fairly general formulation that naturally extends that of Chapter 6. Consider point-referenced locations and continuous time. Let $Y(\mathbf{s}, t)$ denote the measurement at location $\mathbf{s}$ at time t. Extending (6.1), for continuous data assumed to be roughly normally distributed, we can write the general form

$$Y(\mathbf{s}, t) = \mu(\mathbf{s}, t) + e(\mathbf{s}, t) , \tag{11.2}$$

where $\mu(\mathbf{s}, t)$ denotes the mean structure and $e(\mathbf{s}, t)$ denotes the residual. If $\mathbf{x}(\mathbf{s}, t)$ is a vector of covariates associated with $Y(\mathbf{s}, t)$ then we can set $\mu(\mathbf{s}, t) = \mathbf{x}(\mathbf{s}, t)^T \boldsymbol{\beta}(\mathbf{s}, t)$. Note that this form allows spatiotemporally varying coefficients (in the spirit of Section 9.6), which is likely more general than we would want; $\boldsymbol{\beta}(\mathbf{s}, t) = \boldsymbol{\beta}$ is frequently adopted. If t is discretized, $\boldsymbol{\beta}(\mathbf{s}, t) = \boldsymbol{\beta}_t$ might be appropriate if there were enough time points to suggest a temporal change in the coefficient vector. Similarly, setting $\boldsymbol{\beta}(\mathbf{s}, t) = \boldsymbol{\beta}(\mathbf{s})$ yields spatially varying coefficients, again following Section 9.6. Finally, $e(\mathbf{s}, t)$ would typically be rewritten as $w(\mathbf{s}, t) + \epsilon(\mathbf{s}, t)$, where $\epsilon(\mathbf{s}, t)$ is a Gaussian white noise process and $w(\mathbf{s}, t)$ is a mean-zero spatiotemporal process.

We can therefore view (11.2) as a hierarchical model with a conditionally independent first stage given $\{\mu(\mathbf{s}, t)\}$ and $\{w(\mathbf{s}, t)\}$. But then, in the spirit of Section 6.2, we can replace the Gaussian first stage with another first-stage model (say, an exponential family model) and write $Y(\mathbf{s}, t) \sim f(y(\mathbf{s}, t) \mid \mu(\mathbf{s}, t), w(\mathbf{s}, t))$, where

$$f(y(\mathbf{s}, t) \mid \mu(\mathbf{s}, t), w(\mathbf{s}, t)) = h(y(\mathbf{s}, t)) \exp\{\gamma[\eta(\mathbf{s}, t) y(\mathbf{s}, t) - \chi(\eta(\mathbf{s}, t))]\} , \tag{11.3}$$

[1] It is routine to see that this will not be the case if the u_1 process and the v_1 process are not independent.

where γ is a positive dispersion parameter. In (11.3), $g(\eta(\mathbf{s},t)) = \mu(\mathbf{s},t) + w(\mathbf{s},t)$ for some link function g.

For areal unit data with discrete time, let Y_{it} denote the measurement for unit i at time period t. (In some cases we might obtain replications at i or t, e.g., the jth cancer case in county i, or the the jth property sold in school district i.) Analogous to (11.2) we can write

$$Y_{it} = \mu_{it} + e_{it} \; . \tag{11.4}$$

Now $\mu_{it} = \mathbf{x}_{it}^T \boldsymbol{\beta}_t$ (or perhaps just $\boldsymbol{\beta}$), and $e_{it} = w_{it} + \epsilon_{it}$ where the ϵ_{it} are unstructured heterogeneity terms and the w_{it} are spatiotemporal random effects, typically associated with a spatiotemporal CAR specification. Choices for this latter part of the model will be presented in Section 11.7.

Since areal unit data are often non-Gaussian (e.g., sparse counts), again we would view (11.4) as a hierarchical model and replace the first stage Gaussian specification with, say, a Poisson model. We could then write $Y_{it} \sim f(y_{it}|\mu_{it}, w_{it})$, where

$$f(y_{it} \mid \mu_{it}, w_{it}) = h(y_{it}) \exp\{\gamma[\eta_{it} y_{it} - \chi(\eta_{it})]\} \; , \tag{11.5}$$

with γ again a dispersion parameter, and $g(\eta_{it}) = \mu_{it} + w_{it}$ for some suitable link function g. With replications, we obtain Y_{ijt} hence $\mathbf{x}_{ijt}, \mu_{ijt}$, and η_{ijt}. Now we can write $g(\eta_{ijt}) = \mu_{ijt} + w_{ijt} + \epsilon_{ijt}$, enabling separation of spatial and heterogeneity effects.

Returning to the point-referenced data model (11.2), spatiotemporal richness is captured by extending $e(\mathbf{s},t)$ beyond $\epsilon(\mathbf{s},t)$, a white noise process, as noted above. As a result, we need forms for $w(\mathbf{s},t)$. Below, α's denote temporal effects and w's denote spatial effects. Following Gelfand, Ecker, Knight, and Sirmans (2004) with t discretized, consider the following forms for $w(\mathbf{s},t)$:

$$w(\mathbf{s},t) = \alpha(t) + w(\mathbf{s}) \; , \tag{11.6}$$

$$w(\mathbf{s},t) = \alpha_{\mathbf{s}}(t) \; , \tag{11.7}$$

$$\text{and } w(\mathbf{s},t) = w_t(\mathbf{s}) \; . \tag{11.8}$$

The given forms avoid specification of space-time interactions. With regard to (11.6), (11.7), and (11.8), the $\epsilon(\mathbf{s},t)$ are i.i.d. $N(0,\sigma_\epsilon^2)$ and independent of the other processes. This pure error is viewed as a residual adjustment to the spatiotemporal explanation. (One could allow $Var(\epsilon(\mathbf{s},t)) = \sigma_\epsilon^{2(t)}$, i.e., an error variance that changes with time. Modification to the details below is straightforward.)

Expression (11.6) provides an additive form in temporal and spatial effects. In fact, we can also introduce a multiplicative form, $\alpha(t)w(\mathbf{s})$ which would, of course, become additive on the log scale. Expression (11.7) provides temporal evolution at each site; temporal effects are nested within sites. Expression (11.8) provides spatial evolution over time; spatial effects are nested within time. Spatiotemporal modeling beyond (11.6), (11.7), and (11.8) (particularly if t is continuous) necessitates the choice of a specification to connect the space and time scales; this is the topic of Section 11.2.

Next, we consider the components in (11.6), (11.7), and (11.8) in more detail. In (11.6), if t were continuous we could model $\alpha(t)$ as a one-dimensional stationary Gaussian process. In particular, for the set of times, $\{t_1, t_2, ..., t_m\}$, $\boldsymbol{\alpha} = (\alpha(t_1), ..., \alpha(t_m))' \sim N(\mathbf{0}, \sigma_\alpha^2 \Sigma(\phi))$ where $(\Sigma(\phi))_{rs} = Corr(\alpha(t_r), \alpha(t_s)) = \rho(\mid t_r - t_s \mid; \phi)$ for ρ a valid one-dimensional correlation function. A typical choice for ρ would be the exponential, $\exp(-\phi \mid t_r - t_s \mid)$ though other forms, analogous to the spatial forms in Table 2.1 are possible.

With t confined to an indexing set, $t = 1, 2, \ldots T$, we can simply view $\alpha(1), ..., \alpha(T)$ as the coefficients associated with a set of time dummy variables. With this assumption for the $\alpha(t)$'s, suppose in (11.6), $w(\mathbf{s})$ is set to zero, $\boldsymbol{\beta}(t)$ is assumed constant over time and $\mathbf{X}(\mathbf{s},t)$

is assumed constant over t. Then, upon differencing, we find models described in the real estate literature, e.g., the seminal model for repeat property sales given in Bailey, Muth, and Nourse (1963). Also within these assumptions but restoring $\boldsymbol{\beta}$ to $\boldsymbol{\beta}(t)$, we obtain the extension of Knight, Dombrow, and Sirmans (1995). In very recent work (Paci et al., 2013) the multiplicative form is used, with differencing, to implement real-time ozone forecasting.

Alternatively, we might set $\alpha(t+1) = \rho\alpha(t) + \eta(t)$ where $\eta(t)$ are i.i.d. $N(0, \sigma^2_\alpha)$. If $\rho < 1$ we have the familiar stationary $AR(1)$ time series, a special case of the continuous time model of the previous paragraph. If $\rho = 1$ the $\alpha(t)$ follow a random walk. With a finite set of times, time-dependent coefficients are handled analogously to the survival analysis setting (see, e.g., Cox and Oakes, 1984, Ch. 8).

The autoregressive and random walk specifications are naturally extended to provide a model for the $\alpha_{\mathbf{s}}(t)$ in (11.7). That is, we assume $\alpha_{\mathbf{s}}(t+1) = \rho\alpha_{\mathbf{s}}(t) + \eta_{\mathbf{s}}(t)$ where again the $\eta_{\mathbf{s}}(t)$ are all i.i.d. Thus, there is no spatial modeling; rather, we imagine independent conceptual time series at each location. With spatial time series we can fit this model. With cross-sectional data, there is no information in the data about ρ so the likelihood can only identify the stationary variance $\sigma^2_\alpha/(1-\rho^2)$ but not σ^2_α or ρ. The case $\rho < 1$ with $\boldsymbol{\beta}(t)$ constant over time provides the models proposed in Hill, Knight, and Sirmans (1997) and in Hill, Sirmans, and Knight (1999). If $\rho = 1$ with $\boldsymbol{\beta}(t)$ and $\mathbf{X}(\mathbf{s}, t)$ constant over time, upon differencing we obtain the widely used model of Case and Shiller (1989). In application, it will be difficult to learn about the $\alpha_{\mathbf{s}}$ processes with typically one or at most two observations for each $\mathbf{s}$. The $w(\mathbf{s})$ are modeled as a Gaussian process following Section 3.1.

For $w_t(\mathbf{s})$ in (11.8), assuming t restricted to an index set, we can view the $w_t(\mathbf{s})$ as a collection of independent spatial processes. That is, rather than defining a dummy variable at each t, we conceptualize a separate spatial dummy *process* at each t. The components of $\mathbf{w}_t$ correspond to the sites at which measurements were observed in the time interval denoted by t. Thus, we capture the dynamics of location in a very general fashion. In particular, comparison of the respective process parameters reveals the nature of spatial evolution over time.

With a single time dummy variable at each t, assessment of temporal effects would be provided through inference associated with these variables. For example, a plot of the point estimates against time would clarify size and trend for the effects. With distinct spatial processes, how can we see such temporal patterns? A convenient reduction of each spatial process to a univariate random variable is the block average (see Expression (7.1)).

To shed the independence assumption for the $w_t(\mathbf{s})$, we could instead assume that $w_t(\mathbf{s}) = \sum_{j=1}^{t} v_j(\mathbf{s})$ where the $v_j(\mathbf{s})$ are i.i.d. processes, again of one of the foregoing forms. Now, for $t < t^*$, $\mathbf{w}_t$ and $\mathbf{w}_{t^*}$ are not independent but $\mathbf{w}_t$ and $\mathbf{w}_{t^*} - \mathbf{w}_t$ are. This leads us to dynamic spatiotemporal models that are the focus of Section 11.4.

11.1.3 Associated distributional results

We begin by developing the likelihood under model (11.2) using (11.6), (11.7), or (11.8). Assuming t belongs to the set $\{1, 2, ..., T\}$, it is convenient to first obtain the joint distribution for $\mathbf{Y}' = (\mathbf{Y}'_1, ..., \mathbf{Y}'_T)$ where $\mathbf{Y}'_t = (Y(\mathbf{s}_1, t), ..., Y(\mathbf{s}_n, t))$. That is, each $\mathbf{Y}_t$ is $n \times 1$ and $\mathbf{Y}$ is $Tn \times 1$. This joint distribution will be multivariate normal. Thus, the joint distribution for the observed $Y(\mathbf{s}, t)$ requires only pulling off the appropriate entries from the mean vector and appropriate rows and columns from the covariance matrix. This simplifies the computational bookkeeping, though care is still required.

In the constant $\boldsymbol{\beta}$ case, associate with $\mathbf{Y}_t$ the matrix X_t whose ith row is $\mathbf{X}(\mathbf{s}_i, t)'$. Let $\boldsymbol{\mu}_t = X_t\boldsymbol{\beta}$ and $\boldsymbol{\mu}' = (\mu'_1, ..., \mu'_T)$. In the time-dependent parameter case we merely set $\boldsymbol{\mu}_t = X_t\boldsymbol{\beta}(t)$.

Under (11.6), let $\boldsymbol{\alpha}' = (\alpha(1), \ldots, \alpha(T))$, $\mathbf{w}' = (\mathbf{w}(\mathbf{s}_1), \ldots, \mathbf{w}(\mathbf{s}_n))$ and $\boldsymbol{\epsilon}' = (\epsilon(\mathbf{s}_1, 1), \epsilon(\mathbf{s}_1, 2), \ldots, \epsilon(\mathbf{s}_n, T)$. Then,

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\alpha} \otimes \mathbf{1}_{n\times 1} + \mathbf{1}_{T\times 1} \otimes \mathbf{w} + \boldsymbol{\epsilon} \tag{11.9}$$

where $\otimes$ denotes the Kronecker product. Hence, given $\boldsymbol{\beta}$ along with the temporal and spatial effects,

$$\mathbf{Y} \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{w}, \sigma^2_\epsilon \sim N(\boldsymbol{\mu} + \boldsymbol{\alpha} \otimes \mathbf{1}_{n\times 1} + \mathbf{1}_{T\times 1} \otimes \mathbf{w}\ ,\ \sigma^2_\epsilon I_{Tn\times Tn})\ . \tag{11.10}$$

Let $\mathbf{w} \sim N(\mathbf{0}, \sigma^2_w H(\delta))$. Suppose the $\alpha(t)$ follow an $AR(1)$ model, so that $\boldsymbol{\alpha} \sim N(\mathbf{0}, \sigma^2_\alpha A(\rho))$ where $(A(\rho))_{ij} = \rho^{|i-j|}/(1-\rho^2)$. Hence, if $\boldsymbol{\alpha}, \mathbf{w}$ and $\boldsymbol{\epsilon}$ are independent, marginalizing over $\boldsymbol{\alpha}$ and $\mathbf{w}$, i.e., integrating (11.10) with regard to the prior distribution of $\boldsymbol{\alpha}$ and $\mathbf{w}$, we obtain

$$\begin{aligned}&\mathbf{Y} \mid \boldsymbol{\beta}, \sigma^2_\epsilon, \sigma^2_\alpha, \rho, \sigma^2_w, \delta \\ &\sim N\left(\boldsymbol{\mu}\ ,\ \sigma^2_\alpha A(\rho) \otimes \mathbf{1}_{n\times 1}\mathbf{1}'_{n\times 1} + \sigma^2_w \mathbf{1}_{T\times 1}\mathbf{1}'_{T\times 1} \otimes H(\delta) + \sigma^2_\epsilon I_{Tn\times Tn}\right)\ .\end{aligned} \tag{11.11}$$

If the $\alpha(t)$ are coefficients associated with dummy variables (now $\boldsymbol{\beta}$ does not contain an intercept) we only marginalize over $\mathbf{w}$ to obtain

$$\begin{aligned}&\mathbf{Y} \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2_\epsilon, \sigma^2_w, \delta \\ &\quad\sim N\left(\boldsymbol{\mu} + \boldsymbol{\alpha} \otimes \mathbf{1}_{n\times 1}\ ,\ \sigma^2_w \mathbf{1}_{T\times 1}\mathbf{1}'_{T\times 1} \otimes H(\delta) + \sigma^2_\epsilon I_{Tn\times Tn}\right)\ .\end{aligned} \tag{11.12}$$

The likelihood resulting from (11.10) arises as a product of independent normal densities by virtue of the conditional independence. This can facilitate model fitting but at the expense of a very high-dimensional posterior distribution. Marginalizing to (11.11) or (11.12) results in a much lower-dimensional posterior. Note, however, that while the distributions in (11.11) and (11.12) can be determined, evaluating the likelihood (joint density) requires evaluation of a high-dimensional quadratic form and determinant calculation.

Turning to (11.7), if $\boldsymbol{\alpha}'(t) = (\alpha_{s_1}(t), \ldots, \alpha_{s_n}(t))$ and now we also define $\boldsymbol{\alpha}' = (\boldsymbol{\alpha}'(1), \ldots, \boldsymbol{\alpha}'(T))$ with $\boldsymbol{\epsilon}$ as above, then

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\alpha} + \boldsymbol{\epsilon}\ .$$

Now

$$\mathbf{Y} \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2_\epsilon \sim N\left(\boldsymbol{\mu} + \boldsymbol{\alpha}\ ,\ \sigma^2_\epsilon I_{Tn\times Tn}\right)\ .$$

If the $\alpha_{s_i}(t)$ follow an $AR(1)$ model independently across i, then marginalizing over $\boldsymbol{\alpha}$,

$$\mathbf{Y} \mid \boldsymbol{\beta}, \sigma^2_\epsilon, \sigma^2_\alpha, \rho \sim N(\boldsymbol{\mu}\ ,\ A(\rho) \otimes I_{Tn\times Tn} + \sigma^2_\epsilon I_{Tn\times Tn}). \tag{11.13}$$

For (11.8), let $\mathbf{w}'_t = (w_t(\mathbf{s}_1), \ldots, w_t(\mathbf{s}_n))$ and $\mathbf{w}' = (\mathbf{w}'_1, \ldots, \mathbf{w}'_T)$. Then with $\boldsymbol{\epsilon}$ as above,

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{w} + \boldsymbol{\epsilon} \tag{11.14}$$

and

$$\mathbf{Y} \mid \boldsymbol{\beta}, \mathbf{w}, \sigma^2_\epsilon \sim N\left(\boldsymbol{\mu} + \mathbf{w}\ ,\ \sigma^2_\epsilon I_{Tn\times Tn}\right)\ . \tag{11.15}$$

If $\mathbf{w}_t \sim N(\mathbf{0}, \sigma^{2(t)}_w H(\delta^{(t)}))$ independently for $t = 1, \ldots, T$, then, marginalizing over $\mathbf{w}$,

$$\mathbf{Y} \mid \boldsymbol{\beta}, \sigma^2_\epsilon, \boldsymbol{\sigma}^2_w, \boldsymbol{\delta} \sim N(\boldsymbol{\mu}\ ,\ D(\boldsymbol{\sigma}^2_w, \boldsymbol{\delta}) + \sigma^2_\epsilon I_{Tn\times Tn})\ , \tag{11.16}$$

where $\boldsymbol{\sigma}^{2\prime}_w = (\sigma^{2(1)}_w, \ldots, \sigma^{2(T)}_w)$, $\boldsymbol{\delta}' = (\delta^{(1)}, \ldots, \delta^{(T)})$, and $D(\boldsymbol{\sigma}^2_w, \boldsymbol{\delta})$ is block diagonal with the tth block being $\sigma^{2(t)}_w(H(\delta^{(t)}))$. Because D is block diagonal, likelihood evaluation associated with (11.16) is less of an issue than for (11.11) and (11.12).

We note that with either (11.7) or (11.8), $e(\mathbf{s}, t)$ is comprised of two sources of error that the data cannot directly separate. However, by incorporating a stochastic assumption on the $\alpha_{\mathbf{s}}(t)$ or on the $w_t(\mathbf{s})$, we can learn about the processes that guide the error components, as (11.13) and (11.16) reveal.

11.1.4 Prediction and forecasting

We now turn to forecasting under (11.2) with models (11.6), (11.7), or (11.8). Such forecasting involves prediction at location $\mathbf{s}_0$ and time t_0, i.e., of $Y(\mathbf{s}_0, t_0)$. Here $\mathbf{s}_0$ may correspond to an already observed location, perhaps to a new location. However, typically $t_0 > T$ is of interest. Such prediction requires specification of an associated vector of characteristics $\mathbf{X}(\mathbf{s}_0, t_0)$. Also, prediction for $t_0 > T$ is available in the fixed coefficients case. For the time-varying coefficients case, we would need to specify a temporal model for $\boldsymbol{\beta}(t)$.

In general, within the Bayesian framework, prediction at $(\mathbf{s}_0, t_0)$ follows from the posterior predictive distribution of $f(Y(\mathbf{s}_0, t_0) \mid \mathbf{Y})$ where $\mathbf{Y}$ denotes the observed data vector. Assuming $\mathbf{s}_0$ and t_0 are new, and for illustration, taking the form in (11.6),

$$f(Y(\mathbf{s}_0, t_0) \mid \mathbf{Y}) = \int f(Y(\mathbf{s}_0, t_0) \mid \boldsymbol{\beta}, \sigma_\epsilon^2, \alpha(t_0), w(\mathbf{s}_0)) \times dF(\boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{w}, \sigma_\epsilon^2, \sigma_\alpha^2, \rho, \sigma_w^2, \delta, \alpha(t_0), w(\mathbf{s}_0) \mid \mathbf{Y}) \; . \tag{11.17}$$

Using (11.17), given a random draw $(\boldsymbol{\beta}^*, \sigma_\epsilon^{2*}, \alpha(t_0)^*, w(\mathbf{s}_0)^*)$ from the posterior $f(\boldsymbol{\beta}, \sigma_\epsilon^2, \alpha(t_0), w(\mathbf{s}_0) \mid \mathbf{Y})$, if we draw $Y^*(\mathbf{s}_0, t_0)$ from $N(X'(\mathbf{s}_0, t_0)\boldsymbol{\beta}^* + \alpha(t_0)^* + w(\mathbf{s}_0)^*, \sigma_e^{2*})$, marginally, $Y^*(\mathbf{s}_0, t_0) \sim f(Y(\mathbf{s}_0, t_0) \mid \mathbf{Y})$.

Using sampling-based model fitting and working with (11.10), we obtain samples $(\boldsymbol{\beta}^*, \sigma_\epsilon^{2*}, \sigma_\alpha^{2*}, \rho^*, \sigma_w^{2*}, \delta^*, \boldsymbol{\alpha}^*, \mathbf{w}^*)$ from the posterior distribution, $p(\boldsymbol{\beta}, \sigma_\epsilon^2, \sigma_\alpha^2, \rho, \sigma_w^2, \delta, \boldsymbol{\alpha}, \mathbf{w} \mid \mathbf{Y})$. But $f(\boldsymbol{\beta}, \sigma_\epsilon^2, \sigma_\alpha^2, \rho, \sigma_w^2, \delta, \boldsymbol{\alpha}, \mathbf{w}, \alpha(t_0), w(\mathbf{s}_0) \mid \mathbf{Y}) = f(\alpha(t_0) \mid \boldsymbol{\alpha}, \sigma_\alpha^2, \rho) \cdot f(w(\mathbf{s}_0) \mid \mathbf{w}, \sigma_w^2, \delta) \cdot f(\boldsymbol{\beta}, \sigma_\epsilon^2, \sigma_\alpha^2, \rho, \sigma_w^2, \delta, \boldsymbol{\alpha}, \mathbf{w} \mid \mathbf{Y})$. If, e.g., $t_0 = T + 1$, and $\alpha(t)$ is modeled as a time series, $f(\alpha(T+1) \mid \boldsymbol{\alpha}, \sigma_\alpha^2, \rho)$ is $N(\rho\alpha(T), \sigma_\alpha^2)$. If the $\alpha(t)$ are coefficients associated with dummy variables, setting $\alpha(T+1) = \alpha(T)$ is, arguably, the best one can do. The joint distribution of $\mathbf{w}$ and $w(\mathbf{s}_0)$ is a multivariate normal from which $f(w(\mathbf{s}_0) \mid \mathbf{w}, \sigma_w^2, \delta)$ is a univariate normal. So if $\alpha(t_0)^* \sim f(\alpha(t_0) \mid \boldsymbol{\alpha}^*, \sigma_\alpha^{2*}, \rho^*)$ and $w(\mathbf{s}_0)^* \sim f(w(\mathbf{s}_0) \mid \mathbf{w}^*, \sigma_w^{2*}, \delta^*)$, along with $\boldsymbol{\beta}^*$ and σ_ϵ^{2*} we obtain a draw from $f(\boldsymbol{\beta}, \sigma_\epsilon^2, \alpha(t_0), w(\mathbf{s}_0) \mid \mathbf{Y})$. (If $t_0 \epsilon \{1, 2, ..., T\}$, $\alpha(t_0)$ is a component of $\boldsymbol{\alpha}$, then $\alpha(t_0)^*$ is a component of $\boldsymbol{\alpha}^*$. If $\mathbf{s}_0$ is one of the $\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n$, $w^*(\mathbf{s}_0)$ is a component of $\mathbf{w}^*$.) Alternatively, one can work with (11.13). Now, having marginalized over $\boldsymbol{\alpha}$ and $\mathbf{w}$, $Y(\mathbf{s}, t)$ and $\mathbf{Y}$ are no longer independent. They have a multivariate normal distribution from which $f(Y(\mathbf{s}, t) \mid \mathbf{Y}, \boldsymbol{\beta}, \sigma_\epsilon^2, \sigma_\alpha^2, \rho, \sigma_w^2, \delta)$ must be obtained. Note that for multiple predictions, $w(\mathbf{s}_0)$ is replaced by a vector, say, $\mathbf{w}_0$. Now $f(\mathbf{w}_0 | \mathbf{w}, \sigma_w^2, \delta)$ is a multivariate normal distribution. No additional complications arise.

Example 11.1 *(Baton Rouge home sales).* We present a portion of the data analysis developed in Gelfand et al. (2004) for sales of single-family homes drawn from two regions in the city of Baton Rouge, LA. The two areas are known as Sherwood Forest and Highland Road. These regions are approximately the same size and have similar levels of transaction activity; they differ chiefly in the range of neighborhood characteristics and house amenities found within. Sherwood Forest is a large, fairly homogeneous neighborhood located east, southeast of downtown Baton Rouge. Highland Road, on the other hand, is a major thoroughfare connecting downtown with the residential area to the southeast. Rather than being one homogeneous neighborhood, the Highland Road area consists, instead, of heterogeneous subdivisions. Employing two regions makes a local isotropy assumption more comfortable and allows investigation of possibly differing time effects and location dynamics.

For these regions, a subsample of all homes sold only once during the period 1985 through 1995 (single-sale transactions) and a second subsample of homes sold more than once (repeat-sale transactions) were drawn. These two samples can be studied separately to assess whether the population of single-sale houses differs from that of repeat-sale houses. The sample sizes are provided by year in Table 11.1. The location of each property is defined by its latitude and longitude coordinates, rescaled to UTM projection. In addition, a variety of house characteristics, to control for physical differences among the properties, are recorded at the time of sale. We use age, living area, other area (e.g., patios, garages, and

	Highland		Sherwood	
Year	Repeat	Single	Repeat	Single
1985	25	40	32	29
1986	20	35	32	39
1987	27	32	27	37
1988	16	26	20	34
1989	21	25	24	35
1990	42	29	27	37
1991	29	30	25	31
1992	33	38	39	27
1993	24	40	31	40
1994	26	35	20	34
1995	26	35	21	32
Total	289	365	298	375

Table 11.1 *Sample size by region, type of sale, and year.*

	Highland		Sherwood	
Variable	Repeat	Single	Repeat	Single
Age	11.10	12.49	14.21	14.75
	(8.15)	(11.37)	(8.32)	(10.16)
Bathrooms	2.18	2.16	2.05	2.02
	(0.46)	(0.56)	(0.36)	(0.40)
Living area	2265.4	2075.8	1996.0	1941.5
	(642.9)	(718.9)	(566.8)	(616.2)
Other area	815.1	706.0	726.0	670.6
	(337.7)	(363.6)	(258.1)	(289.2)

Table 11.2 *Mean (standard deviation) for house characteristics by region and type of sale.*

carports) and number of bathrooms as covariates in our analysis. Summary statistics for these attributes appear in Table 11.2. We see that the homes in the Highland Road area are somewhat newer and slightly larger than those in the Sherwood area. The greater heterogeneity of the Highland Road homes is borne out by the almost uniformly higher standard deviations for each covariate. In fact, we have more than 20 house characteristics in our data set, but elaborating the mean with additional features provides little improvement in R^2 and introduces multicollinearity problems. So, we confine ourselves to the four explanatory variables above and turn to spatial modeling to explain a portion of the remaining variability. Empirical semivariograms (2.9) offer evidence of spatial association, after adjusting for house characteristics.

We describe the results of fitting the model with mean $\mu(\mathbf{s}) = \mathbf{x}(\mathbf{s})^T\boldsymbol{\beta}$ and the error structure in (11.8). This is also the preferred model using the predictive model choice approach of Gelfand and Ghosh (5.14); we omit details. Fixed coefficients were justified by the shortness of the observation period. Again, an exponential isotropic correlation function was adopted.

Variable	Repeat	Single
Highland region:		
intercept (β_0)	11.63 (11.59, 11.66)	11.45 (11.40, 11.50)
age (β_1)	–0.04 (–0.07, –0.02)	–0.08 (–0.11, –0.06)
bathrooms (β_2)	0.02 (–0.01, 0.04)	0.02 (–0.01, 0.05)
living area (β_3)	0.28 (0.25, 0.31)	0.33 (0.29, 0.37)
other area (β_4)	0.08 (0.06, 0.11)	0.07 (0.04, 0.09)
Sherwood region:		
intercept (β_0)	11.33 (11.30, 11.36)	11.30 (11.27, 11.34)
age (β_1)	–0.06 (–0.07, –0.04)	–0.05 (–0.07, –0.03)
bathrooms (β_2)	0.05 (0.03, 0.07)	0.00 (–0.02, 0.02)
living area (β_3)	0.19 (0.17, 0.21)	0.22 (0.19, 0.24)
other area (β_4)	0.02 (0.01, 0.04)	0.06 (0.04, 0.08)

Table 11.3 *Parameter estimates (median and 95% interval estimates) for house characteristics.*

To complete the Bayesian specification, we adopt rather noninformative priors in order to resemble a likelihood/least squares analysis. In particular, we assume a flat prior on the regression parameter $\boldsymbol{\beta}$ and inverse gamma (a, b) priors for σ_ϵ^2, $\sigma_w^{2(t)}$ and $\delta^{(t)}$, $t = 1, \ldots T$. The shape parameter for these inverse gamma priors was fixed at 2, implying an infinite prior variance. We choose the inverse gamma scale parameter for all $\delta^{(t)}$'s to be equal, i.e., $b_{\delta^{(1)}} = b_{\delta^{(2)}} = \ldots = b_{\delta^{(T)}} = b_\delta$, say, and likewise for $\sigma_w^{2(t)}$. Furthermore, we set $b_{\sigma_\epsilon} = b_{\sigma_w^2}$ reflecting uncertain prior contribution from the nugget to the sill. Finally, the exact values of b_{σ_ϵ}, $b_{\sigma_w^2}$ and b_δ vary between region and type of sale reflecting different prior beliefs about these characteristics.

Inference for the house characteristic coefficients is provided in Table 11.3 (point and 95% interval estimates). Age, living area, and other area are significant in all cases; number of bathrooms is significant only in Sherwood repeat sales. Significance of living area is much stronger in Highland than in Sherwood. The Highland sample is composed of homes from several heterogeneous neighborhoods. As such, living area not only measures differences in house size, but may also serve as a partial proxy for construction quality and for neighborhood location within the sample. The greater homogeneity of homes in Sherwood implies less variability in living area (as seen in Table 11.2) and reduces the importance of these variables in explaining house price.

Turning to the error structure, the parameters of interest for each region are the $\sigma_w^{2(t)}$, the $\delta^{(t)}$, and σ_e^2. The sill at time t is $Var(Y(s,t)) = \sigma_w^{2(t)} + \sigma_e^2$. Figure 11.1 plots the posterior medians of these sills. We see considerable difference in variability over the groups and over time, providing support for distinct spatial models at each t. Variability is highest for Highland single sales, lowest for Sherwood repeats. The additional insight is the effect of time. Variability is generally increasing over time.

We can obtain posterior median and interval estimates for $\sigma_w^{2(t)}/(\sigma_e^2 + \sigma_w^{2(t)})$, the proportion of spatial variance to total. The strength of the spatial story is considerable; 40 to 80% of the variability is spatial.

In Figure 11.2 we provide point and interval estimates for the range. The ranges for the repeat sales are quite similar for the two regions, showing some tendency to increase in the later years of observation. By contrast, the range for the Highland single sales is much different from that for Sherwood. It is typically greater and much more variable. The latter again is a refection of the high variability in the single-sale home prices in Highland. The resulting posteriors are more dispersed.

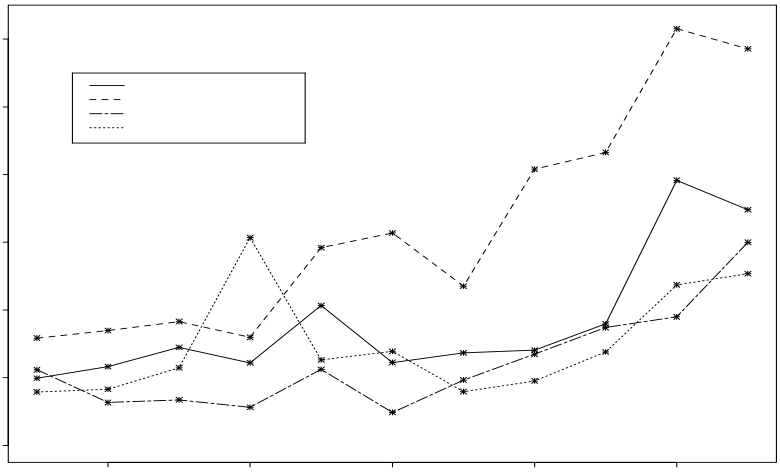

Figure 11.1 *Posterior median sill by year.*

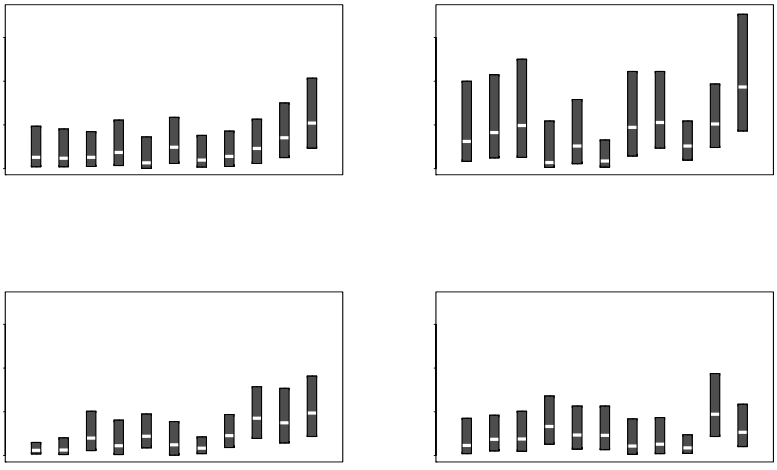

Figure 11.2 *Posterior median and 95% interval estimates for the range by year for (a) Highland repeat sales, (b) Highland single sales, (c) Sherwood repeat sales, and (d) Sherwood single sales.*

Finally, in Figure 11.3, we present the posterior distribution of the block averages, mentioned at the end of Subsection 11.1.2, for each of the four analyses. Again, these block averages are viewed as analogues of more familiar time dummy variables. Time effects are evident. In all cases, we witness somewhat of a decline in magnitude in the 1980s and an increasing trend in the 1990s.

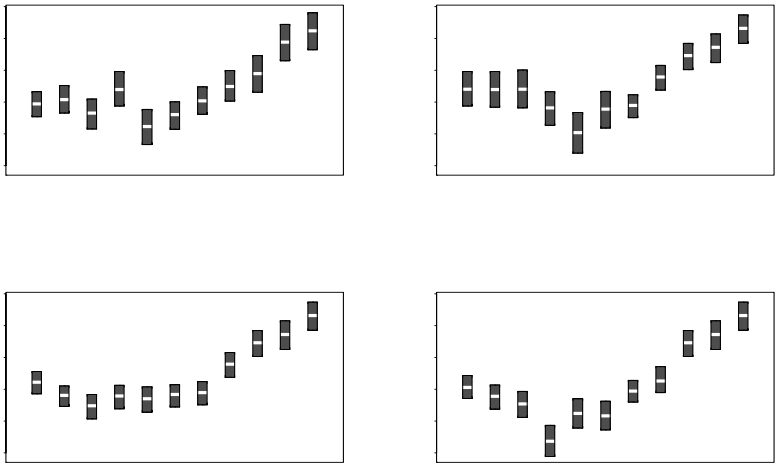

Figure 11.3 *Posterior median and 95% interval estimates for the block averages by year for (a) Highland repeat sales, (b) Highland single sales, (c) Sherwood repeat sales, and (d) Sherwood single sales.*

11.2 Point-level modeling with continuous time

Suppose now that $\mathbf{s} \in \Re^2$ and $t \in \Re^+$ and we seek to define a spatiotemporal process $Y(\mathbf{s}, t)$. As in Subsection 3.1 we have to provide a joint distribution for an uncountable number of random variables. Again, we do this through arbitrary finite dimensional distributions. Confining ourselves to the Gaussian case, we only need to specify a valid spatiotemporal covariance function. Here, "valid" means that for any set of locations and any set of time points, the covariance matrix for the resulting set of random variables is positive definite. An important point here is that it is not sensible to combine $\mathbf{s}$ and t and propose a valid correlation function on $\Re^3$. This is because distance in space has nothing to do with "distance" on the time scale.

As a result, a stationary spatiotemporal covariance specification is assumed to take the form $cov(Y(\mathbf{s}, t), Y(\mathbf{s}', t')) = c(\mathbf{s} - \mathbf{s}', t - t')$. An isotropic form sets $cov(Y(\mathbf{s}, t), Y(\mathbf{s}', t')) = c(||\mathbf{s} - \mathbf{s}'||, |t - t'|)$. A frequently used choice is the *separable* form

$$cov(Y(\mathbf{s}, t), Y(\mathbf{s}', t')) = \sigma^2 \rho^{(1)}(\mathbf{s} - \mathbf{s}'; \boldsymbol{\phi})\, \rho^{(2)}(t - t'; \boldsymbol{\psi})\,, \tag{11.18}$$

where $\rho^{(1)}$ is a valid two-dimensional correlation function and $\rho^{(2)}$ is a valid one-dimensional correlation function. Expression (11.18) shows that dependence attenuates in a multiplicative manner across space and time. Forms such as (11.18) have a history in spatiotemporal modeling; see, e.g., Mardia and Goodall (1993) and references therein.

Why is (11.18) valid? For locations $\mathbf{s}_1, \ldots, \mathbf{s}_I$ and times $t_1, \ldots, t_J$, collecting the variables a vector $\mathbf{Y}_s^T = (\mathbf{Y}^T(\mathbf{s}_1), \ldots, \mathbf{Y}^T(\mathbf{s}_I))$ where $\mathbf{Y}(\mathbf{s}_i) = (Y(\mathbf{s}_i, t_1), \ldots, Y(\mathbf{s}_i, t_J))^T$, the covariance matrix of $\mathbf{Y}_s$ is

$$\Sigma_{\mathbf{Y}_s}(\sigma^2, \boldsymbol{\phi}, \boldsymbol{\psi}) = \sigma^2 H_s(\boldsymbol{\phi}) \otimes H_t(\boldsymbol{\psi})\,, \tag{11.19}$$

where "$\otimes$" again denotes the Kronecker product. In (11.19), $H_s(\boldsymbol{\phi})$ is $I \times I$ with $(H_s(\boldsymbol{\phi}))_{ii'} = \rho^{(1)}(\mathbf{s}_i - \mathbf{s}_i'; \boldsymbol{\theta})$, and $H_t(\boldsymbol{\psi})$ is $J \times J$ with $(H_t(\boldsymbol{\psi}))_{jj'} = \rho^{(2)}(t_j - t_{j'}; \boldsymbol{\psi})$. Expression (11.19) clarifies that $\Sigma_{\mathbf{Y}_s}$ is positive definite, following the argument below (9.11). So, $\mathbf{Y}_s$ will

be IJ-dimensional multivariate normal with, in obvious notation, mean vector $\boldsymbol{\mu}_s(\boldsymbol{\beta})$ and covariance matrix (11.19).

Given a prior for $\boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}$, and $\boldsymbol{\psi}$, the Bayesian model is completely specified. Simulation-based model fitting can be carried out similarly to the static spatial case by noting the following. The log-likelihood arising from $\mathbf{Y}_s$ is

$$-\tfrac{1}{2}\log\left|\sigma^2 H_s(\boldsymbol{\phi})\otimes H_t(\boldsymbol{\psi})\right| \\ -\tfrac{1}{2\sigma^2}(\mathbf{Y}_s-\boldsymbol{\mu}_s(\boldsymbol{\beta}))^T(H_s(\boldsymbol{\phi})\otimes H_t(\boldsymbol{\psi}))^{-1}(\mathbf{Y}_s-\boldsymbol{\mu}_s(\boldsymbol{\beta}))\ .$$

But in fact $\left|\sigma^2 H_s(\boldsymbol{\phi})\otimes H_t(\boldsymbol{\psi})\right| = (\sigma^2)^{IJ}\left|H_s(\boldsymbol{\phi})\right|^J\left|H_t(\boldsymbol{\psi})\right|^I$ and $(H_s(\boldsymbol{\phi})\otimes H_t(\boldsymbol{\psi}))^{-1} = H_s^{-1}(\boldsymbol{\phi})\otimes H_t^{-1}(\boldsymbol{\psi})$ by properties of Kronecker products. In other words, even though (11.19) is $IJ\times IJ$, we need only the determinant and inverse for an $I\times I$ and a $J\times J$ matrix, expediting likelihood evaluation and hence Gibbs sampling.

With regard to prediction, first consider new locations $\mathbf{s}'_1,...,\mathbf{s}'_K$ with interest in inference for $Y(\mathbf{s}'_k,t_j)$. As with the observed data, we collect the $Y(\mathbf{s}'_k,t_j)$ into vectors $\mathbf{Y}(\mathbf{s}'_k)$, and the $\mathbf{Y}(\mathbf{s}'_k)$ into a single $KJ\times 1$ vector $\mathbf{Y}_{s'}$. Even though we may not necessarily be interested in every component of $\mathbf{Y}_{s'}$, the simplifying forms that follow suggest that, with regard to programming, it may be easiest to simulate draws from the entire predictive distribution $f(\mathbf{Y}_{s'}\mid\mathbf{Y}_s)$ and then retain only the desired components.

Since $f(\mathbf{Y}_{s'}|\mathbf{Y}_s)$ has a form analogous to (7.3), given posterior samples $(\boldsymbol{\beta}^*_g,\sigma^{2*}_g,\boldsymbol{\phi}^*_g,\boldsymbol{\psi}^*_g)$, we draw $\mathbf{Y}^*_{s',g}$ from $f(\mathbf{Y}_{s'}\mid\mathbf{Y}_s,\boldsymbol{\beta}^*_g,\sigma^{2*}_g,\boldsymbol{\phi}^*_g,\boldsymbol{\psi}^*_g)$, $g=1,\ldots,G$. Analogous to (7.4),

$$f\left(\begin{pmatrix}\mathbf{Y}_s\\ \mathbf{Y}_{s'}\end{pmatrix}\middle|\,\boldsymbol{\beta},\sigma^2,\boldsymbol{\phi},\boldsymbol{\psi}\right) = N\left(\begin{pmatrix}\boldsymbol{\mu}_s(\boldsymbol{\beta})\\ \boldsymbol{\mu}_{s'}(\boldsymbol{\beta})\end{pmatrix},\ \Sigma_{\mathbf{Y}_s,\mathbf{Y}_{s'}}\right) \tag{11.20}$$

where

$$\Sigma_{\mathbf{Y}_s,\mathbf{Y}_{s'}} = \sigma^2\begin{pmatrix}H_s(\boldsymbol{\phi})\otimes H_t(\boldsymbol{\psi}) & H_{s,s'}(\boldsymbol{\phi})\otimes H_t(\boldsymbol{\psi})\\ H^T_{s,s'}(\boldsymbol{\phi})\otimes H_t(\boldsymbol{\psi}) & H_{s'}(\boldsymbol{\phi})\otimes H_t(\boldsymbol{\psi})\end{pmatrix},$$

with obvious definitions for $H_{s'}(\boldsymbol{\phi})$ and $H_{s,s'}(\boldsymbol{\phi})$. But then the conditional distribution $\mathbf{Y}_{s'}\mid\mathbf{Y}_s,\boldsymbol{\beta},\sigma^2,\boldsymbol{\phi},\boldsymbol{\psi}$ is also normal, with mean

$$\begin{aligned}&\boldsymbol{\mu}_{s'}(\boldsymbol{\beta})+(H^T_{s,s'}(\boldsymbol{\phi})\otimes H_t(\boldsymbol{\phi}))(H_s(\boldsymbol{\phi})\otimes H_t(\boldsymbol{\psi}))^{-1}(Y_s-\boldsymbol{\mu}_s(\boldsymbol{\beta}))\\ &\quad=\boldsymbol{\mu}_{s'}(\boldsymbol{\beta})+(H^T_{s,s'}(\boldsymbol{\phi})H_s^{-1}(\boldsymbol{\phi})\otimes I_{J\times J})(\mathbf{Y}_s-\boldsymbol{\mu}_s(\boldsymbol{\beta}))\ ,\end{aligned} \tag{11.21}$$

and covariance matrix

$$\begin{aligned}&H_{s'}(\boldsymbol{\phi})\otimes H_t(\boldsymbol{\psi})\\ &\quad-(H^T_{s,s'}\otimes H_t(\boldsymbol{\psi}))(H_s(\boldsymbol{\phi})\otimes H_t(\boldsymbol{\psi}))^{-1}(H_{s,s'}(\boldsymbol{\phi})\otimes H_t(\boldsymbol{\psi}))\\ &\qquad=(H_{s'}(\boldsymbol{\phi})-H^T_{s,s'}(\boldsymbol{\phi})H_s^{-1}(\boldsymbol{\phi})H_{s,s}(\boldsymbol{\phi}))\otimes H_t(\boldsymbol{\psi})\ ,\end{aligned} \tag{11.22}$$

using standard properties of Kronecker products.

In (11.21), time disappears apart from $\boldsymbol{\mu}_{s'}(\boldsymbol{\beta})$, while in (11.22), time "factors out" of the conditioning. Sampling from this normal distribution usually employs the inverse square root of the conditional covariance matrix, but conveniently, this is

$$(H_{s'}(\boldsymbol{\phi})-H^T_{s,s'}(\boldsymbol{\phi})H_s^{-1}(\boldsymbol{\phi})H_{s,s'}(\boldsymbol{\phi}))^{-\frac{1}{2}}\otimes H_t^{-\frac{1}{2}}(\boldsymbol{\psi})\ ,$$

so the only work required beyond that in (7.5) is obtaining $H_t^{-\frac{1}{2}}(\boldsymbol{\psi})$, since $H_t^{-1}(\boldsymbol{\psi})$ will already have been obtained in evaluating the likelihood, following the discussion above.

For prediction not for points but for areal units (blocks) $B_1,\ldots,B_K$, we would set $\mathbf{Y}^T(B_k)=(Y(B_k,t_1),\ldots,Y(B_k,t_J))$ and then further set $\mathbf{Y}^T_B=(\mathbf{Y}^T(B_1),\ldots,\mathbf{Y}^T(B_K))$. Analogous to (7.6) we seek to sample $f(\mathbf{Y}_B\mid\mathbf{Y}_s)$, so we require $f(\mathbf{Y}_B\mid$

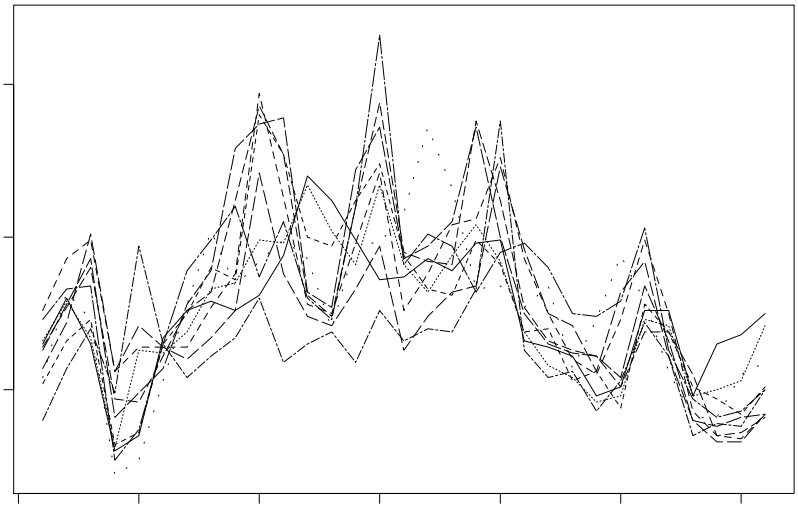

Figure 11.4 *Observed 1-hour maximum ozone measurement by day, July 1995, 10 Atlanta monitoring sites.*

$\mathbf{Y}_s, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \boldsymbol{\psi}$). Analogous to (11.20), this can be derived from the joint distribution $f((\mathbf{Y}_s, \mathbf{Y}_B)^T \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \boldsymbol{\psi})$, which is

$$N\left(\begin{pmatrix} \boldsymbol{\mu}_s(\boldsymbol{\beta}) \\ \boldsymbol{\mu}_B(\boldsymbol{\beta}) \end{pmatrix}, \sigma^2 \begin{pmatrix} H_s(\boldsymbol{\phi}) \otimes H_t(\boldsymbol{\psi}) & H_{s,B}(\boldsymbol{\phi}) \otimes H_t(\boldsymbol{\psi}) \\ H_{s,B}^T(\boldsymbol{\phi}) \otimes H_t(\boldsymbol{\psi}) & H_B(\boldsymbol{\phi}) \otimes H_t(\boldsymbol{\psi}) \end{pmatrix}\right),$$

with $\boldsymbol{\mu}_B(\boldsymbol{\beta})$, $H_B(\boldsymbol{\phi})$, and $H_{s,B}(\boldsymbol{\phi})$ defined as in Section 7.1.2. Thus the distribution $f(\mathbf{Y}_B|\mathbf{Y}_s, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \boldsymbol{\psi})$ is again normal with mean and covariance matrix as given in (11.21) and (11.22), but with $\boldsymbol{\mu}_B(\boldsymbol{\beta})$ replacing $\boldsymbol{\mu}_{s'}(\boldsymbol{\beta})$, $H_B(\boldsymbol{\phi})$ replacing $H_{s'}(\boldsymbol{\phi})$, and $H_{s,B}(\boldsymbol{\phi})$ replacing $H_{s,s'}(\boldsymbol{\phi})$. Using the same Monte Carlo integrations as proposed in Section 7.1.2 leads to sampling the resultant $\widehat{f}(\mathbf{Y}_B \mid \mathbf{Y}_s, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \boldsymbol{\psi})$, and the same technical justification applies.

If we started with block data, $Y(B_i, t_j)$, then following (7.11) and (11.19),

$$f(\mathbf{Y}_B \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \boldsymbol{\psi}) = N(\mu_B(\boldsymbol{\beta}), \sigma^2(H_B(\boldsymbol{\phi}) \otimes H_t(\boldsymbol{\psi}))). \tag{11.23}$$

Given (11.23), the path for prediction at new points or at new blocks is clear, following the above and the end of Section 7.1.2; we omit the details.

Note that the association structure in (11.18) allows *forecasting* of the spatial process at time t_{J+1}. This can be done at observed or unobserved points or blocks following the foregoing development. To retain the above simplifying forms, we would first simulate the variables at t_{J+1} associated with observed points or blocks (with no change of support). We would then revise $H_t(\boldsymbol{\phi})$ to be $(J+1) \times (J+1)$ before proceeding as above.

Example 11.2 To illustrate the methods above, we use a spatiotemporal version of the Atlanta ozone data set. As mentioned in Section 7.1, we actually have ozone measurements at the 10 fixed monitoring stations shown in Figure 1.3 over the 92 summer days in 1995. Figure 11.4 shows the daily 1-hour maximum ozone reading for the sites during July of this same year. There are several sharp peaks, but little evidence of a weekly (seven-day) period in the data. The mean structure appears reasonably constant in space, with the ordering

	Spatial only		Spatiotemporal	
	Point	95% Interval	Point	95% Interval
Point A	.125	(.040, .334)	.139	(.111, .169)
Point B	.116	(.031, .393)	.131	(.098, .169)
Zip 30317 (east-central)	.130	(.055, .270)	.138	(.121, .155)
Zip 30344 (south-central)	.123	(.055, .270)	.135	(.112, .161)
Zip 30350 (north)	.112	(.040, .283)	.109	(.084, .140)

Table 11.4 *Posterior medians and 95% equal-tail credible intervals for ozone levels at two points, and for average ozone levels over three blocks (zip codes), purely spatial model versus spatiotemporal model, Atlanta ozone data for July 15, 1995.*

of the site measurements changing dramatically for different days. Moreover, with only 10 "design points" in the metro area, any spatial trend surface we fit would be quite speculative over much of the study region (e.g., the northwest and southwest metro; see Figure 1.3). The temporal evolution of the series is not inconsistent with a constant mean autoregressive error model; indeed, the lag 1 sample autocorrelation varies between .27 and .73 over the 10 sites, strongly suggesting the need for a model accounting for both spatial and temporal correlations.

We thus fit our spatiotemporal model with mean $\boldsymbol{\mu}(\mathbf{s}, t; \boldsymbol{\beta}) = \mu$, but with spatial and temporal correlation functions $\rho^{(1)}(\mathbf{s}_i - \mathbf{s}_{i'}; \phi) = e^{-\phi \|\mathbf{s}_i - \mathbf{s}_{i'}\|}$ and $\rho^{(2)}(t_j - t_{j'}; \psi) = \psi^{|j-j'|}$. Hence our model has four parameters: we use a flat prior for μ, an $IG(3, 0.5)$ prior for σ^2, a $G(0.003, 100)$ prior for ϕ, and a $U(0, 1)$ prior for ψ (thus eliminating the implausible possibility of *negative* autocorrelation in our data, but favoring no positive value over any other). To facilitate our Gibbs-Metropolis approach, we transform to $\theta = \log \phi$ and $\lambda = \log(\psi/(1-\psi))$, and subsequently use Gaussian proposals on these transformed parameters.

Running 3 parallel chains of 10,000 iterations each, sample traces (not shown) again indicate virtually immediate convergence of our algorithm. Posterior medians and 95% equal-tail credible intervals for the four parameters are as follows: for μ, 0.068 and (0.057, 0.080); for σ^2, 0.11 and (0.08, 0.17); for ϕ, 0.06 and (0.03, 0.08); and for ψ, 0.42 and (0.31, 0.52). The rather large value of ψ confirms the strong temporal autocorrelation suspected in the daily ozone readings.

Comparison of the posteriors for σ^2 and ϕ with those obtained for the static spatial model in Example 7.1 is not sensible, since these parameters have different meanings in the two models. Instead, we make this comparison in the context of point-point and point-block prediction. Table 11.4 provides posterior predictive summaries for the ozone concentrations for July 15, 1995, at points A and B (see Figure 1.3), as well as for the block averages over three selected Atlanta city zips: 30317, an east-central city zip very near to two monitoring sites; 30344, the south-central zip containing the points A and B; and 30350, the northernmost city zip. Results are shown for both the spatiotemporal model of this subsection and for the static spatial model previously fit in Example 7.1. Note that all the posterior medians are a bit higher under the spatiotemporal model, except for that for the northern zip, which remains low. Also note the significant increase in precision afforded by this model, which makes use of the data from all 31 days in July 1995, instead of only that from July 15. Figure 11.5 shows the estimated posteriors giving rise to the first and last rows in Table 11.4 (i.e., corresponding to the the July 15, 1995, ozone levels at point A and the block average over the northernmost city zip, 30350). The Bayesian approach's ability to reflect differing amounts of predictive uncertainty for the two models is clearly evident.

Finally, Figure 11.6 plots the posterior medians and upper and lower .025 quantiles produced by the spatiotemporal model by day for the ozone concentration at point A, as well as those for the block average in zip 30350. Note that the overall temporal pattern

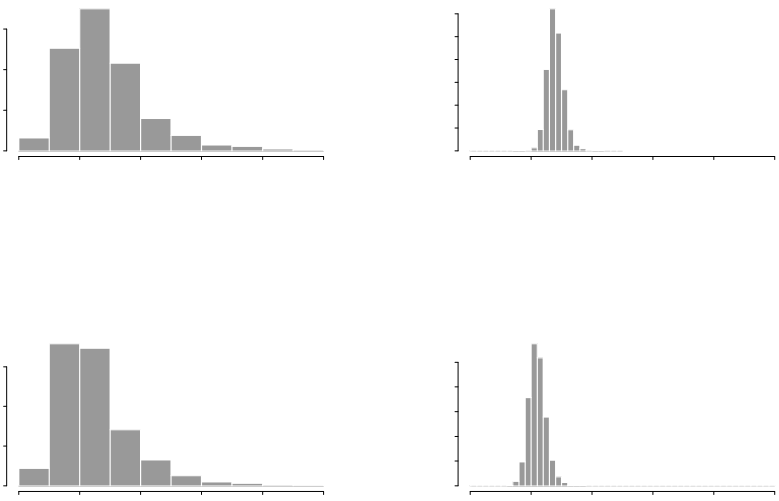

Figure 11.5 *Posterior predictive distributions for ozone concentration at point A and the block average over zip 30350, purely spatial model versus spatiotemporal model, Atlanta ozone data for July 15, 1995.*

is quite similar to that for the data shown in Figure 11.4. Since point A is rather nearer to several data observation points, the confidence bands associated with it are often a bit narrower than those for the northern zip, but this pattern is not perfectly consistent over time. Also note that the relative positions of the bands for July 15 are consistent with the data pattern for this day seen in Figure 1.3, when downtown ozone exposures were higher than those in the northern metro. Finally, the day-to-day variability in the predicted series is substantially larger than the predictive variability associated with any given day.

11.3 Nonseparable spatiotemporal models ⋆

The separable form for the spatiotemporal covariance function in (11.18) is convenient for computation and offers attractive interpretation. However, its form limits the nature of space-time interaction. Additive forms, arising from $w(\mathbf{s},t) = w(\mathbf{s}) + \alpha(t)$ with $w(\mathbf{s})$ and $\alpha(t)$ independent may be even more unsatisfying.

A simple way to extend (11.18) is through *mixing*. For instance, suppose $w(\mathbf{s},t) = w_1(\mathbf{s},t) + w_2(\mathbf{s},t)$ with w_1 and w_2 independent processes, each with a separable spatiotemporal covariance function, say $c_\ell(\mathbf{s}-\mathbf{s}', t-t) = \sigma_\ell^2 \rho_\ell^{(1)}(\mathbf{s}-\mathbf{s}')\rho_\ell^{(2)}(t-t')$, $\ell = 1,2$. Then the covariance function for $w(\mathbf{s},t)$ is evidently the sum and is not separable. Building covariance functions in this way is easy to interpret but yields an explosion of parameters with finite mixing. Continuous parametric mixing, e.g.,

$$c(\mathbf{s}-\mathbf{s}', t-t) = \sigma^2 \int \rho^{(1)}(\mathbf{s}-\mathbf{s}', \boldsymbol{\phi})\rho^{(2)}(t-t', \boldsymbol{\psi})\, G_{\boldsymbol{\gamma}}(d\boldsymbol{\phi}, d\boldsymbol{\psi}) \,, \tag{11.24}$$

yields a function that depends only on σ^2 and $\boldsymbol{\gamma}$. Extensions of these ideas are developed

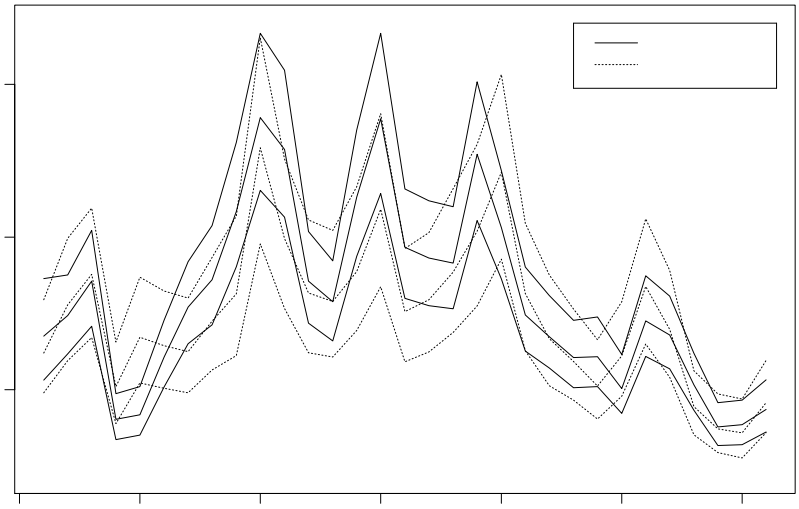

Figure 11.6 *Posterior medians and upper and lower .025 quantiles for the predicted 1-hour maximum ozone concentration by day, July 1995; solid lines, point A; dotted lines, block average over zip 30350 (northernmost Atlanta city zip).*

in De Iaco et al. (2002). However, these forms have not received much attention in the literature to date.

Cressie and Huang (1999) introduce a flexible class of nonseparable stationary covariance functions that allow for space-time interaction. However, they work in the spectral domain and require that $c(\mathbf{s}-\mathbf{s}', t-t')$ can be computed explicitly, i.e., the Fourier inversion can be obtained in closed-form. Unfortunately this occurs only in very special cases. Recent work by Gneiting (2002) adopts a similar approach but obtains very general classes of valid space-time models that do not rely on closed form Fourier inversions. One simple example is the class $c(\mathbf{s}-\mathbf{s}', t-t') = \sigma^2(|t-t'|+1)^{-1}\exp(-||\mathbf{s}-\mathbf{s}'||(|t-t'|+1)^{-\beta/2})$. Here, β is a space-time interaction parameter; $\beta = 0$ provides a separable specification.

Stein (2005) also works in the spectral domain, providing a class of spectral densities whose resulting spatiotemporal covariance function is nonseparable with flexible analytic behavior. These spectral densities extend the Matérn form; see (3.5) or the discussion below on Equation (A.2) in Appendix A. In particular, the spectral density is

$$\widehat{c}(\mathbf{w}, v) \propto [c_1(\alpha_1^2 + ||\mathbf{w}||^2)^{\alpha_1} + c_2(\alpha_2 + v^2)^{\alpha_2}]^{-v} .$$

Unfortunately, the associated covariance function cannot be computed explicitly; fast Fourier transforms (see Appendix Section A.1) offer the best computational prospects. Also, unlike Gneiting's class, separability does not arise as a special or limiting case. For further discussion of the above, see Chapter 23 of Gelfand et al. (2010). We also mention related work using "blurring" discussed in Brown, Kåresen, Roberts, and Tonellato (2000).

11.4 Dynamic spatiotemporal models ⋆

In this section we follow the approach taken in Banerjee, Gamerman, and Gelfand, (2003), viewing the data as arising from a time series of spatial processes. In particular, we work in

the setting of dynamic models (West and Harrison, 1997), describing the temporal evolution in a latent space. We achieve a class of dynamic models for spatiotemporal data.

Here, there is a growing literature. Non-Bayesian approaches include Huang and Cressie (1996), Wikle and Cressie (1999), and Mardia et al. (1998). Bayesian approaches include Tonellato (1997), Sanso and Guenni (1999), Stroud et al. (2001), and Huerta et al. (2003). The paper by Stroud et al. (2001) is attractive in being applicable to any data set that is continuous in space and discrete in time and allows straightforward computation using Kalman filtering.

11.4.1 Brief review of dynamic linear models

Dynamic linear models, often referred to as state-space models in the time-series literature, offer a versatile framework for fitting several time-varying models (West and Harrison, 1997). We briefly outline the general dynamic linear modeling framework. Thus, let $\mathbf{Y}_t$ be a $m \times 1$ vector of observables at time t. $\mathbf{Y}_t$ is related to a $p \times 1$ vector, $\boldsymbol{\theta}_t$, called the state vector, through a *measurement equation*. In general, the elements of $\boldsymbol{\theta}_t$ are not observable, but are generated by a first-order Markovian process, resulting in a *transition equation*. Therefore, we can describe the above framework as

$$\begin{aligned} \mathbf{Y}_t &= F_t\boldsymbol{\theta}_t + \boldsymbol{\epsilon}_t,\ \boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \Sigma_t^{\epsilon}). \\ \boldsymbol{\theta}_t &= G_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\eta}_t,\ \boldsymbol{\eta}_t \sim N(\mathbf{0}, \Sigma_t^{\eta}) \ , \end{aligned}$$

where F_t and G_t are $m \times p$ and $p \times p$ matrices, respectively. The first equation is the measurement equation, where $\boldsymbol{\epsilon}_t$ is a $m \times 1$ vector of serially uncorrelated Gaussian variables with mean $\mathbf{0}$ and an $m \times m$ covariance matrix, Σ_t^{ϵ}. The second equation is the transition equation with $\boldsymbol{\eta}_t$ being a $p \times 1$ vector of serially uncorrelated zero-centered Gaussian disturbances and Σ_t^{η} the corresponding $p \times p$ covariance matrix. Note that under (11.25), the association structure can be computed explicitly across time, e.g., $Cov(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}) = G_t Var(\boldsymbol{\theta}_{t-1})$ and $Cov(\mathbf{Y}_t, \mathbf{Y}_{t-1}) = F_t G_t Var(\boldsymbol{\theta}_{t-1}) F_t^T$.

F_t (in the measurement equation) and G_t (in the transition equation) are referred to as *system matrices* that may change over time. F_t and G_t may involve unknown parameters but, given the parameters, temporal evolution is in a predetermined manner. The matrix F_t is usually specified by the design of the problem at hand, while G_t is specified through modeling assumptions; for example, $G_t = I_p$, the $p \times p$ identity matrix would provide a random walk for $\boldsymbol{\theta}_t$. Regardless, the system is linear, and for any time point t, $\mathbf{Y}_t$ can be expressed as a linear combination of the present $\boldsymbol{\epsilon}_t$ and the present and past $\boldsymbol{\eta}_t$'s.

11.4.2 Formulation for spatiotemporal models

In this section we adapt the above dynamic modeling framework to univariate spatiotemporal models with spatially varying coefficients. For this we consider a collection of sites $S = \{\mathbf{s}_1, ..., \mathbf{s}_{N_s}\}$, and time-points $T = \{t_1, ..., t_{N_t}\}$, yielding observations $Y(\mathbf{s}, t)$, and covariate vectors $\mathbf{x}(\mathbf{s}, t)$, for every $(\mathbf{s}, t) \in S \times T$.

The response, $Y(\mathbf{s}, t)$, is first modeled through a measurement equation, which incorporates the measurement error, $\epsilon(\mathbf{s}, t)$, as serially and spatially uncorrelated zero-centered Gaussian disturbances. The transition equation now involves the regression parameters (slopes) of the covariates. The slope vector, say $\tilde{\boldsymbol{\beta}}(\mathbf{s}, t)$, is decomposed into a purely temporal component, $\boldsymbol{\beta}_t$, and a spatiotemporal component, $\boldsymbol{\beta}(\mathbf{s}, t)$. Both these are generated through transition equations, capturing their Markovian dependence in time. While the transition equation of the purely temporal component is as in usual state-space modeling, the spatiotemporal component is generated by a multivariate Gaussian spatial process.

Thus, we may write the spatiotemporal modeling framework as

$$\begin{aligned} Y(\mathbf{s},t) &= \mu(\mathbf{s},t) + \epsilon(\mathbf{s},t);\ \epsilon(\mathbf{s},t) \overset{ind}{\sim} N\left(0,\sigma^2\right), && (11.25)\\ \mu(\mathbf{s},t) &= \mathbf{x}^T(\mathbf{s},t)\,\tilde{\boldsymbol{\beta}}(\mathbf{s},t), \\ \tilde{\boldsymbol{\beta}}(\mathbf{s},t) &= \boldsymbol{\beta}_t + \boldsymbol{\beta}(\mathbf{s},t), && (11.26)\\ \boldsymbol{\beta}_t &= \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t,\ \boldsymbol{\eta}_t \overset{ind}{\sim} N_p\left(\mathbf{0},\Sigma_{\boldsymbol{\eta}}\right), \\ \text{and } \boldsymbol{\beta}(\mathbf{s},t) &= \boldsymbol{\beta}(\mathbf{s},t-1) + \mathbf{w}(\mathbf{s},t)\ . \end{aligned}$$

In (11.26), we introduce a linear model of coregionalization (Section 9.5) for $\mathbf{w}(\mathbf{s},t)$, i.e., $\mathbf{w}(\mathbf{s},t) = A\mathbf{v}(\mathbf{s},t)$, with $\mathbf{v}(\mathbf{s},t) = (v_1(\mathbf{s},t),...,v_p(\mathbf{s},t))^T$, yielding $\Sigma_w = AA^T$. The $v_l(\mathbf{s},t)$ are serially independent replications of a Gaussian process with unit variance and correlation function $\rho_l(\cdot\,;\phi_l)$, henceforth denoted by $GP(0,\rho_l(\cdot\,;\phi_l))$, for $l = 1,\ldots,p$ and independent across l. In the current context, we assume that A does not depend upon $(\mathbf{s},t)$. Nevertheless, this still allows flexible modeling for the spatial covariance structure, as we discuss below.

Moreover, allowing a spatially varying coefficient $\boldsymbol{\beta}(\mathbf{s},t)$ to be associated with $\mathbf{x}(\mathbf{s},t)$ provides an arbitrarily rich explanatory relationship for the x's with regard to the Y's (see Section 9.6 in this regard). By comparison, in Stroud et al. (2001), at a given t, a locally weighted mixture of linear regressions is proposed and only the purely temporal component of $\tilde{\boldsymbol{\beta}}(\mathbf{s},t)$ is used. Such a specification requires both number of basis functions and number of mixture components.

Returning to our specification, note that if $v_l(\cdot,t) \overset{ind}{\sim} GP(0,\rho(\cdot;\phi))$, we have the intrinsic or separable model for $w(\mathbf{s},t)$. Allowing different correlation functions and decay parameters for the $v_l(\mathbf{s},t)$, i.e., $v_l(\cdot,t) \overset{ind}{\sim} GP(0,\rho_l(\cdot;\phi_l))$ yields the linear model of coregionalization (Section 9.5).

Following Section 11.4.1, we can compute the general association structure for the Y's under (11.25) and (11.26). For instance, we have the result that

$$\mathrm{Cov}(Y(\mathbf{s},t),Y(\mathbf{s}',t-1)) = \mathbf{x}^T(\mathbf{s},t)\,\Sigma_{\tilde{\boldsymbol{\beta}}(\mathbf{s},t),\tilde{\boldsymbol{\beta}}(\mathbf{s}',t-1)}\mathbf{x}(\mathbf{s},t-1)\ ,$$

where $\Sigma_{\tilde{\boldsymbol{\beta}}(\mathbf{s},t),\tilde{\boldsymbol{\beta}}(\mathbf{s}',t-1)} = (t-1)\left(\Sigma_{\boldsymbol{\eta}} + \sum_{l=1}^{p}\rho_l(\mathbf{s}-\mathbf{s}';\phi_l)\,\mathbf{a}_l\mathbf{a}_l^T\right)$. Furthermore,

$$\mathrm{Var}(Y(\mathbf{s},t)) = \mathbf{x}^T(\mathbf{s},t)\,t\left[\Sigma_{\boldsymbol{\eta}} + AA^T\right]\mathbf{x}(\mathbf{s},t)$$

with the result that $\mathrm{Corr}(Y(\mathbf{s},t),Y(\mathbf{s}',t-1)) = O(1)$ as $t\to\infty$.

A Bayesian hierarchical model for (11.25) and (11.26) may be completed by prior specifications such as

$$\begin{aligned} \boldsymbol{\beta}_0 &\sim N(\mathbf{m}_0,C_0) \text{ and } \boldsymbol{\beta}(\cdot,0)\equiv 0. && (11.27)\\ \Sigma_{\boldsymbol{\eta}} &\sim IW\left(a_{\boldsymbol{\eta}},B_\eta\right),\ \Sigma_{\mathbf{w}} \sim IW(a_{\mathbf{w}},B_{\mathbf{w}}) \text{ and } \sigma_\epsilon^2 \sim IG(a_\epsilon,b_\epsilon), \\ \mathbf{m}_0 &\sim N(\mathbf{0},\Sigma_0);\ \Sigma_0 = 10^5\times I_p\ , \end{aligned}$$

where B_η and $B_{\mathbf{w}}$ are $p\times p$ precision (hyperparameter) matrices for the inverted Wishart distribution.

Consider now data, in the form $(Y(\mathbf{s}_i,t_j))$ with $i = 1,2,...,N_s$ and $j = 1,2,...,N_t$. Let us collect, for each time point, the observations on all the sites. That is, we form, $\mathbf{Y}_t = (Y(\mathbf{s}_1,t),...,Y(\mathbf{s}_{N_s},t))^T$ and the $N_s\times N_s p$ block diagonal matrix $F_t = \left(\mathbf{x}^T(\mathbf{s}_1,t),\mathbf{x}^T(\mathbf{s}_2,t),...,\mathbf{x}^T(\mathbf{s}_N,t)\right)$ for $t = t_1,...,t_{N_t}$. ^{T}. Analogously we form the $N_s p\times 1$ vector $\boldsymbol{\theta}_t = \mathbf{1}_{N_s}\otimes\boldsymbol{\beta}_t + \boldsymbol{\beta}_t^*$, where $\boldsymbol{\beta}_t^* = (\boldsymbol{\beta}(\mathbf{s}_1,t),\ldots,\boldsymbol{\beta}(\mathbf{s}_{N_s},t))^T$, $\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t$, $\boldsymbol{\eta}_t \overset{ind}{\sim}$

$N_p\left(\mathbf{0}, \Sigma_{\boldsymbol{\eta}}\right)$; and, with $\mathbf{w}_t = \left(\mathbf{w}^T\left(\mathbf{s}_1, t\right), \ldots, \mathbf{w}^T\left(\mathbf{s}_{N_s}, t\right)\right)^T$,

$$\boldsymbol{\beta}_t^* = \boldsymbol{\beta}_{t-1}^* + \mathbf{w}_t,\ \mathbf{w}_t \overset{ind}{\sim} N\left(\mathbf{0}, \sum_{l=1}^{p}\left(R_l\left(\phi_l\right) \otimes \Sigma_{\mathbf{w},l}\right)\right) ,$$

where $[R_l(\phi_l)]_{ij} = \rho_l(\mathbf{s}_i - \mathbf{s}_j; \phi_l)$ is the correlation matrix for $v_l(\cdot, t)$. We then write the data equation for a dynamic spatial model as

$$\mathbf{Y}_t = F_t\boldsymbol{\theta}_t + \boldsymbol{\epsilon}_t;\ t = 1, \ldots, N_t;\ \boldsymbol{\epsilon}_t \sim N\left(\mathbf{0}, \sigma_\epsilon^2 I_{N_s}\right) .$$

With the prior specifications in (11.27), we can design a Gibbs sampler with Gaussian full conditionals for the temporal coefficients $\{\boldsymbol{\beta}_t\}$, the spatiotemporal coefficients $\{\boldsymbol{\beta}_t^*\}$, inverted Wishart for $\Sigma_{\boldsymbol{\eta}}$, and Metropolis steps for ϕ and the elements of $\Sigma_{\mathbf{w},l}$. Updating of $\Sigma_{\mathbf{w}} = \sum_{l=1}^{p}\Sigma_{\mathbf{w},l}$ is most efficiently done by reparametrizing the model in terms of the matrix square root of $\Sigma_{\mathbf{w}}$, say, A, and updating the elements of the lower triangular matrix A. To be precise, consider the full conditional distribution,

$$f\left(\Sigma_{\mathbf{w}}|\boldsymbol{\gamma}, \phi_1, \phi_2\right) \propto f\left(\Sigma_{\mathbf{w}}|a_\gamma, B_\gamma\right) \frac{1}{\left|\sum_{l=1}^{p} R_l(\phi_l)\otimes\Sigma_{\mathbf{w},l}\right|} \times \exp\left(-\tfrac{1}{2}\boldsymbol{\beta}^{*T}\left(J^{-1} \otimes \left(\sum_{l=1}^{p} R_l\left(\phi_l\right) \otimes \Sigma_{\mathbf{w},l}\right)^{-1}\right)\boldsymbol{\beta}^*\right) .$$

The one-to-one relationship between elements of $\Sigma_{\mathbf{w}}$ and the Cholesky square root A is well known (see, e.g., Harville, 1997, p. 235). So, we reparametrize the above full conditional as

$$f\left(A|\boldsymbol{\gamma}, \phi_1, \phi_2\right) \propto f\left(h\left(A\right)|a_\gamma, B_\gamma\right)\left|\frac{\partial h}{\partial a_{ij}}\right| \frac{1}{\left|\sum_{l=1}^{p} R_l(\phi_l)\otimes\left(\mathbf{a}_l\mathbf{a}_l^T\right)\right|} \times \exp\left(-\tfrac{1}{2}\boldsymbol{\beta}^{*T}\left(J^{-1} \otimes \left(\sum_{l=1}^{p} R_l\left(\phi_l\right) \otimes \left(\mathbf{a}_l\mathbf{a}_l^T\right)\right)^{-1}\right)\boldsymbol{\beta}^*\right) .$$

Here, h is the function taking the elements of A, say, a_{ij}, to those of the symmetric positive definite matrix $\Sigma_{\mathbf{w}}$. In the 2×2 case we have

$$h\left(a_{11}, a_{21}, a_{22}\right) = \left(a_{11}^2, a_{11}a_{21}, a_{21}^2 + a_{22}^2\right) ,$$

and the Jacobian is $4a_{11}^2a_{22}$. Now, the elements of A are updated with univariate random-walk Metropolis proposals: lognormal or gamma for a_{11} and a_{22}, and normal for a_{21}. Additional computational burden is created, since now the likelihood needs to be computed for each of the three updates, but the chains are much better tuned (by controlling the scale of the univariate proposals) to move around the parameter space, thereby leading to better convergence behavior.

Example 11.3 (*Modeling temperature given precipitation*). Our spatial domain, shown in Figure 11.7 along with elevation contours (in 100-m units), provides a sample of 50 locations (indicated by "+") in the state of Colorado. Each site provides information on monthly maximum temperature, and monthly mean precipitation. We denote the temperature summary in location $\mathbf{s}$ at time t, by $Y\left(\mathbf{s}, t\right)$, and the precipitation by $x\left(\mathbf{s}, t\right)$. Forming a covariate vector $\mathbf{x}^T\left(\mathbf{s}, t\right) = \left(1, x\left(\mathbf{s}, t\right)\right)$, we analyze the data using a coregionalized dynamic model, as outlined in Subsection 11.4.2. As a result, we have an intercept process $\tilde{\beta}_0\left(\mathbf{s}, t\right)$ and a slope process $\tilde{\beta}_1\left(\mathbf{s}, t\right)$, and the two processes are dependent.

Figure 11.8 displays the time-varying intercepts and slopes (coefficient of precipitation). As expected, the intercept is higher in the summer months and lower in the winter months, highest in July, lowest in December. In fact, the gradual increase from January to July, and the subsequent decrease toward December is evident from the plot. Precipitation seems to have a negative impact on temperature, although this seems to be significant only in the months of January, March, May, June, November, and December, i.e., seasonal pattern is retrieved although no such structure is imposed.

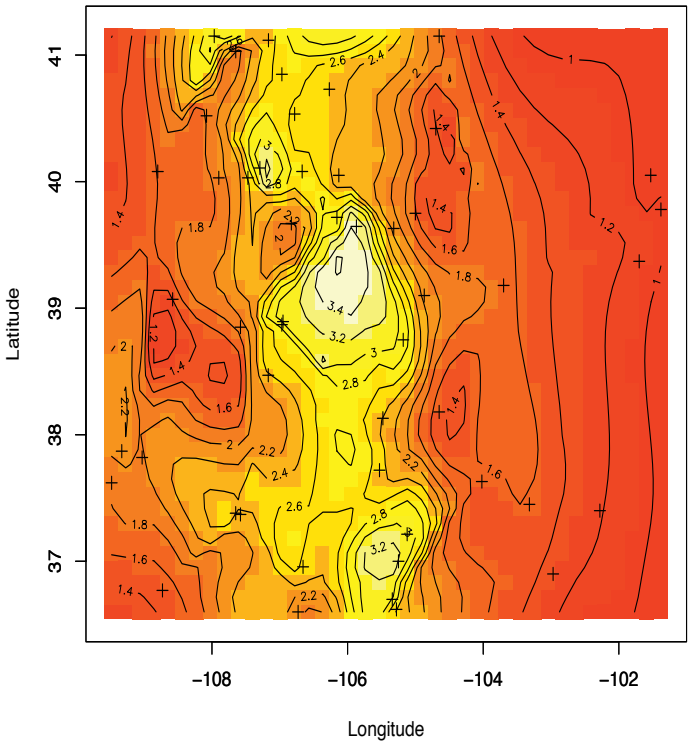

Figure 11.7 *Map of the region in Colorado that forms the spatial domain. The data for the illustrations come from 50 locations, marked by "+" signs in this region.*

Table 11.5 displays the credible intervals for elements of the $\Sigma_{\boldsymbol{\eta}}$ matrix. Rows 1 and 2 show the medians and credible intervals for the respective *variances*; while Row 3 shows the *correlation*. The corresponding results for the elements of $\Sigma_{\mathbf{w}}$ are given in Table 11.6. A significant negative correlation is seen between the intercept and the slope processes, justifying our use of dependent processes. Next, in Table 11.7, we provide the measurement error variances for temperature along with the estimates of the spatial correlation parameters for the intercept and slope process. Also presented are the ranges implied by ϕ_1 and ϕ_2 for the marginal intercept process, $w_1(\mathbf{s})$, and the marginal slope process, $w_2(\mathbf{s})$. The first range is computed by solving for the distance d, $\rho(\phi_1, d) = 0.05$, while the second range is obtained by solving $\left(a_{21}^2 \exp(-\phi_1 d) + a_{22}^2 \exp(-\phi_2 d)\right) / \left(a_{21}^2 + a_{22}^2\right) = 0.05$. The ranges are presented in units of 100 km with the maximum observed distance between our sites being approximately 742 km.

Finally, Figure 11.9 displays the time-sliced image-contour plots for the slope process; similar figures can be drawn for the intercept process. For both processes, the spatial variation is better captured in the central and western edges of the domain. In Figure 11.9, all the months display broadly similar spatial patterns, with denser contour variations toward the west than the east. However, the spatial pattern does seem to be more pronounced in the months with more extreme weather, namely in the winter months of November through January and the summer months of June through August.

11.4.3 Spatiotemporal data

A natural extension of the modeling of the previous sections is to the case where we have data correlated at spatial locations across time. If, as in Section 11.2, we assume that time

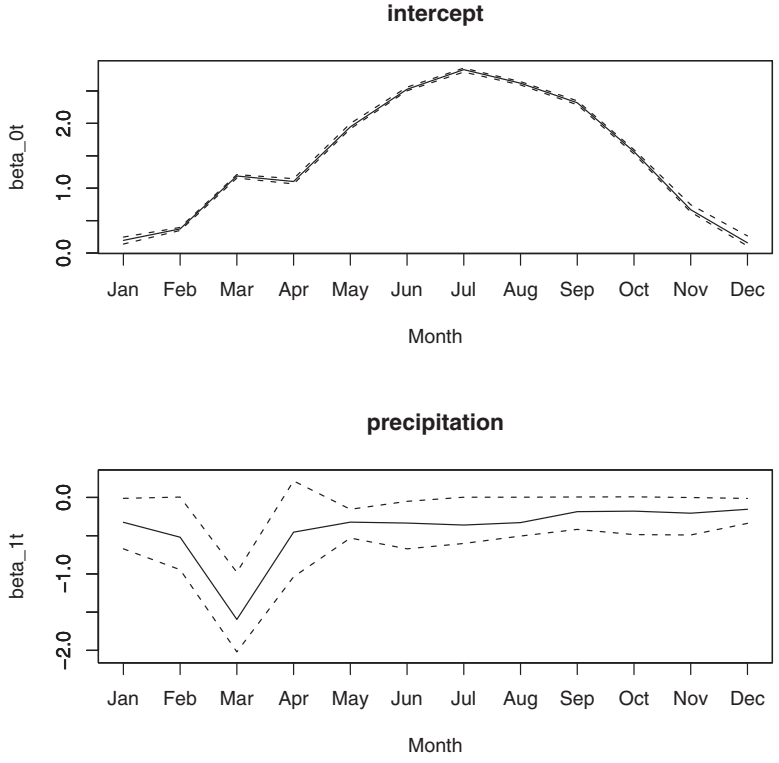

Figure 11.8 *Posterior distributions for the time-varying parameters in the temperature given precipitation example. The top graph corresponds to the intercept, while the lower one is the coefficient of precipitation. Solid lines represent the medians while the dashed lines correspond to the upper and lower credible intervals.*

Σ_η	Median (2.5%, 97.5%)
$\Sigma_\eta[1,1]$	0.296 (0.130, 0.621)
$\Sigma_\eta[2,2]$	0.786 (0.198, 1.952)
$\Sigma_\eta[1,2]/\sqrt{\Sigma_\eta[1,1]\Sigma_\eta[2,2]}$	–0.562 (–0.807, –0.137)

Table 11.5 *Estimates of the variances and correlation from Σ_η, dynamic spatiotemporal modeling example.*

is discretized to a finite set of equally spaced points on a scale, we can conceptualize a time series of spatial processes that are observed only at the spatial locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$.

Adopting a general notation that parallels (9.54), let

$$Y(\mathbf{s},t) = \mathbf{X}^T(\mathbf{s},t)\,\tilde{\boldsymbol{\beta}}(\mathbf{s},t) + \epsilon(\mathbf{s},t)\ ,\quad t = 1,2,\ldots,M\ . \tag{11.28}$$

That is, we introduce spatiotemporally varying intercepts and spatiotemporally varying slopes. Alternatively, if we write $\tilde{\boldsymbol{\beta}}(\mathbf{s},t) = \boldsymbol{\beta}(\mathbf{s},t) + \boldsymbol{\mu}_\beta$, we are partitioning the total error into $p+1$ spatiotemporal intercept pieces including $\epsilon(\mathbf{s},t)$, each with an obvious interpretation. So we continue to assume that $\epsilon(\mathbf{s},t) \overset{iid}{\sim} N\left(0,\tau^2\right)$, but need to specify a model for $\tilde{\boldsymbol{\beta}}(\mathbf{s},t)$. Regardless, (11.28) defines a nonstationary process having moments $E(Y(\mathbf{s},t)) = \mathbf{X}^T(\mathbf{s},t)\tilde{\boldsymbol{\beta}}(s,t)$, $Var(Y(\mathbf{s},t)) = \mathbf{X}^T(\mathbf{s},t)\Sigma_{\tilde{\boldsymbol{\beta}}_{(,t)}}\mathbf{X}(\mathbf{s},t) + \tau^2$, and $Cov(Y(\mathbf{s},t), Y(\mathbf{s}',t')) = \mathbf{X}^T(\mathbf{s},t)\Sigma_{\tilde{\beta}(\mathbf{s},t),\tilde{\beta}(\mathbf{s}',t')}\mathbf{X}(\mathbf{s}',t')$.

$\Sigma_{\mathbf{w}}$	Median (2.5%, 97.5%)
$\Sigma_{\mathbf{w}}[1,1]$	0.017 (0.016, 0.019)
$\Sigma_{\mathbf{w}}[2,2]$	0.026 (0.0065, 0.108)
$\Sigma_{\mathbf{w}}[1,2]/\sqrt{\Sigma_{\mathbf{w}}[1,1]\,\Sigma_{\mathbf{w}}[2,2]}$	–0.704 (–0.843, –0.545)

Table 11.6 *Estimates of the variances and correlation from $\Sigma_{\mathbf{w}}$, dynamic spatiotemporal modeling example.*

Parameters	Median (2.5%, 97.5%)
σ^2_ϵ	0.134 (0.106, 0.185)
ϕ_1	1.09 (0.58, 2.04)
ϕ_2	0.58 (0.37, 1.97)
Range for intercept process	2.75 (1.47, 5.17)
Range for slope process	4.68 (1.60, 6.21)

Table 11.7 *Nugget effects and spatial correlation parameters, dynamic spatiotemporal modeling example.*

Section 11.4 handled (11.28) using a dynamic model. Here we consider four alternative specifications for $\boldsymbol{\beta}(\mathbf{s},t)$. Paralleling the customary assumption from longitudinal data modeling (where the time series are usually short), we could set

- **Model 1:** $\boldsymbol{\beta}(\mathbf{s},t)=\boldsymbol{\beta}(\mathbf{s})$, where $\boldsymbol{\beta}(\mathbf{s})$ is modeled as in the previous sections. This model can be viewed as a locally linear growth curve model.
- **Model 2:** $\boldsymbol{\beta}(\mathbf{s},t)=\boldsymbol{\beta}(\mathbf{s})+\boldsymbol{\alpha}(t)$, where $\boldsymbol{\beta}(\mathbf{s})$ is again as in Model 1. In modeling $\boldsymbol{\alpha}(t)$, two possibilities are (i) treat the $\alpha_k(t)$ as time dummy variables, taking this set of pM variables to be *a priori* independent and identically distributed; and (ii) model the $\boldsymbol{\alpha}(t)$ as a random walk or autoregressive process. The components could be assumed independent across k, but for greater generality, we take them to be dependent, using a separable form that replaces $\mathbf{s}$ with t and takes ρ to be a valid correlation function in just one dimension.
- **Model 3:** $\boldsymbol{\beta}(\mathbf{s},t)=\boldsymbol{\beta}^{(t)}(\mathbf{s})$, i.e., we have spatially varying coefficient processes nested within time. This model is an analogue of the nested effects areal unit specification in Waller et al. (1997); see also Gelfand, Ecker et al. (2003). The processes are assumed independent across t (essentially dummy time processes) and permit temporal evolution of the coefficient process. Following Subsection 9.6.2, the process $\boldsymbol{\beta}^{(t)}(\mathbf{s})$ would be mean-zero, second-order stationary Gaussian with cross-covariance specification at time t, $C^{(t)}(\mathbf{s},\mathbf{s}')$ where $\left(C^{(t)}(\mathbf{s},\mathbf{s}')\right)_{lm}=\rho\left(\mathbf{s}-\mathbf{s}';\phi^{(t)}\right)\tau^{(t)}_{lm}$. We have specified Model 3 with a common $\boldsymbol{\mu}_\beta$ across time. This enables some comparability with the other models we have proposed. However, we can increase flexibility by replacing $\boldsymbol{\mu}_\beta$ with $\boldsymbol{\mu}^{(t)}_\beta$.
- **Model 4:** For $\rho^{(1)}$ a valid two-dimensional correlation function, $\rho^{(2)}$ a valid one-dimensional choice, and T positive definite symmetric, $\boldsymbol{\beta}(\mathbf{s},t)$ such that $\Sigma_{[\boldsymbol{\beta}(\mathbf{s},t),\boldsymbol{\beta}(\mathbf{s}',t')]}=\rho^{(1)}(\mathbf{s}-\mathbf{s}';\phi)\,\rho^{(2)}(t-t';\gamma)\,T$. This model proposes a separable covariance specification in space and time, as in Section 11.2. Here $\rho^{(1)}$ obtains spatial association as in earlier subsections that is attenuated across time by $\rho^{(2)}$. The resulting covariance matrix for the full vector $\boldsymbol{\beta}$, blocked by site and time within site has the convenient form $H_2(\gamma)\otimes H_1(\phi)\otimes T$.

In each of the above models we can marginalize over $\boldsymbol{\beta}(\mathbf{s},t)$ as we did earlier in this section. Depending upon the model it may be more computationally convenient to block the data by site or by time. We omit the details and notice only that, with n sites and T

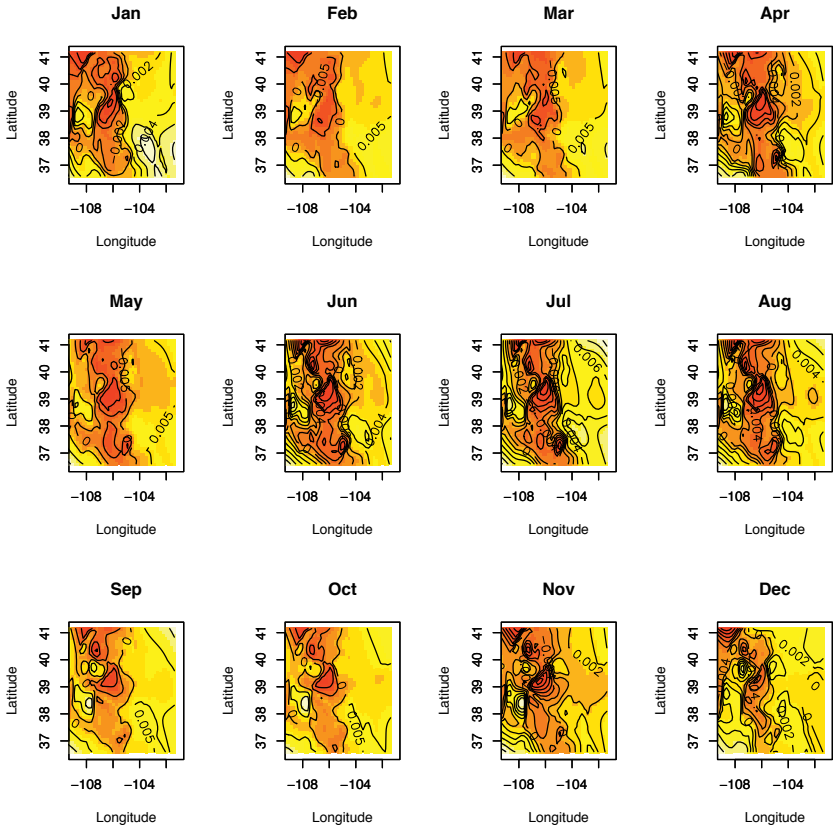

Figure 11.9 *Time-sliced image-contour plots displaying the posterior mean surface of the spatial residuals corresponding to the slope process in the temperature given precipitation model.*

Model	Independent process			Dependent process		
	G	P	D_∞	G	P	D_∞
1	88.58	56.15	144.73	54.54	29.11	83.65
2a	77.79	50.65	128.44	47.92	26.95	74.87
2b	74.68	50.38	125.06	43.38	29.10	72.48
3a	59.46	48.55	108.01	43.74	20.63	64.37
3b	57.09	48.41	105.50	42.35	21.04	63.39
4	53.55	52.98	106.53	37.84	26.47	64.31

Table 11.8 *Model choice criteria for various spatiotemporal process models.*

time points, the resulting likelihood will involve the determinant and inverse of an $nT \times nT$ matrix (typically a large matrix; see Appendix Section A.2).

Note that all of the foregoing modeling can be applied to the case of cross-sectional data where the set of observed locations varies with t. This is the case, for instance, with our real estate data. We only observe a selling price at the time of a transaction. With n_t locations in year t, the likelihood for all but Model 3 will involve a $\sum n_t \times \sum n_t$ matrix.

Example 11.4 *(Baton Rouge housing prices (contd.))*. We now turn to the dynamic version of Baton Rouge dataset presented in Example 9.4. From the Baton Rouge database we drew a sample of 120 transactions at distinct spatial locations for the years 1989, 1990, 1991, and

1992. We compare Models 1–4. In particular, we have two versions of Model 2; 2a has the $\boldsymbol{\alpha}(t)$ as four i.i.d. time dummies, while 2b uses the multivariate temporal process model for $\boldsymbol{\alpha}(t)$. We also have two versions of Model 3; 3a has a common $\boldsymbol{\mu}_\beta$ across t, while 3b uses $\boldsymbol{\mu}_\beta^{(t)}$. In all cases the five-dimensional spatially varying coefficient model for $\boldsymbol{\beta}$'s was employed. Table 11.8 shows the results. Model 3, where space is nested within time, turns out to be the best with Model 4 following closely behind. We omit the posterior inference summary for Model 3b, noting only that the overall coefficients $\left(\boldsymbol{\mu}_\beta^{(t)}\right)$ do not change much over time. However, there is some indication that spatial range is changing over time.

11.5 Fitting dynamic spatiotemporal models using `spBayes`

`spBayes` offers a relatively simple, but rather flexible, univariate version of the dynamic models discussed in the preceding section. Suppose, $y_t(\mathbf{s})$ denotes the observation at location $\mathbf{s}$ and time t. We model $y_t(\mathbf{s})$ through a *measurement equation* that provides a regression specification with a space-time varying intercept and serially and spatially uncorrelated zero-centered Gaussian disturbances as measurement error $\epsilon_t(\mathbf{s})$. Next a *transition equation* introduces a $p \times 1$ coefficient vector, say, $\boldsymbol{\beta}_t$, which is a purely temporal component (i.e., time-varying regression parameters), and a spatio-temporal component $u_t(\mathbf{s})$. Both these are generated through transition equations, capturing their Markovian dependence in time. While the transition equation of the purely temporal component is akin to usual state-space modeling, the spatio-temporal component is generated using Gaussian spatial processes. The overall model, for $t = 1, 2, \ldots, N_t$, is written as

$$\begin{aligned} y_t(\mathbf{s}) &= \mathbf{x}_t(\mathbf{s})^\top \boldsymbol{\beta}_t + u_t(\mathbf{s}) + \epsilon_t(\mathbf{s}), \quad \epsilon_t(\mathbf{s}) \overset{ind.}{\sim} N(0, \tau_t^2) \; ; \\ \boldsymbol{\beta}_t &= \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \overset{i.i.d.}{\sim} N(0, \boldsymbol{\Sigma}_\eta) \; ; \\ u_t(\mathbf{s}) &= u_{t-1}(\mathbf{s}) + w_t(\mathbf{s}), \quad w_t(\mathbf{s}) \overset{ind.}{\sim} GP\left(\mathbf{0}, C_t(\cdot, \boldsymbol{\theta}_t)\right) \; , \end{aligned} \tag{11.29}$$

where the abbreviations *ind.* and *i.i.d* are *independent* and *independent and identically distributed*, respectively. Here $\mathbf{x}_t(\mathbf{s})$ is a $p \times 1$ vector of predictors and $\boldsymbol{\beta}_t$ is a $p \times 1$ vector of coefficients. In addition to an intercept, $\mathbf{x}_t(\mathbf{s})$ can include location specific variables useful for explaining the variability in $y_t(\mathbf{s})$. The $GP(\mathbf{0}, C_t(\cdot, \boldsymbol{\theta}_t))$ denotes a spatial Gaussian process with covariance function $C_t(\cdot; \boldsymbol{\theta}_t)$. We customarily specify $C_t(\mathbf{s}_1, \mathbf{s}_2; \boldsymbol{\theta}_t) = \sigma_t^2 \rho(\mathbf{s}_1, \mathbf{s}_2; \phi_t)$, where $\boldsymbol{\theta}_t = \{\sigma_t^2, \phi_t\}$ and $\rho(\cdot; \phi)$ is a *correlation function* with ϕ controlling the correlation decay and σ_t^2 represents the spatial variance component. We further assume $\boldsymbol{\beta}_0 \sim N(\mathbf{m}_0, \boldsymbol{\Sigma}_0)$ and $u_0(\mathbf{s}) \equiv 0$, which completes the prior specifications leading to a well-identified Bayesian hierarchical model with reasonable dependence structures. In practice, estimation of model parameters are usually very robust to these hyper-prior specifications. Also note that (11.29) reduces to a simple spatial regression model for $t = 1$.

We consider settings where the inferential interest lies in spatial prediction or interpolation over a region for a set of discrete time points. We also assume that the same locations are monitored for each time point resulting in a space-time matrix whose rows index the locations and columns index the time points, i.e., the (i, j)-th element is $y_j(\mathbf{s}_i)$. Our algorithm will accommodate the situation where some cells of the space-time data matrix may have missing observations, as is common in monitoring environmental variables.

The dynamic model (11.29) and a computationally efficient low-rank version using the *predictive process* (see Section 12.4) are implemented in the `spDynLM` function. Here we illustrate the full rank dynamic model using an ozone monitoring dataset that was previously analyzed by Sahu and Bakar (2012). This is a relatively small dataset and does not require dimension reduction.

Figure 11.10 *Open and filled circle symbols indicate the location of 28 ozone monitoring stations across New York State. Filled circle symbols identify those stations that have half of the daily ozone measurements withheld to assess model predictive performance.*

The dataset comprises 28 Environmental Protection Agency monitoring stations that recorded ozone from July 1 to August 31, 2006. The outcome is daily 8-hour maximum average ozone concentrations (parts per billion; O3.8HRMAX), and predictors include maximum temperature (Celsius; cMAXTMP), wind speed (knots; WDSP), and relative humidity (RM). Of the 1,736 possible observations, i.e., n=28 locations times N_t=62 daily O3.8HRMAX measurements, 114 are missing. In this illustrative analysis we use the predictors cMAXTMP, WDSP, and RM as well as the spatially and temporally structured residuals to predict missing O3.8HRMAX values. To gain a better sense of the dynamic model's predictive performance, we withheld half of the observations from the records of three stations for subsequent validation. Figure 11.10 shows the monitoring station locations and identifies those stations where data were withheld.

The first `spDynLM` function argument is a list of N_t symbolic model statements representing the regression within each time step. This can be easily assembled using the `lapply` function as shown in the code below. Here too, we define the station coordinates as well as starting, tuning, and prior distributions for the model parameters. Exploratory data analysis using time step specific variograms can be helpful for defining starting values and prior support for parameters in $\boldsymbol{\theta}_t$ and τ_t^2. To avoid cluttering the code, we specify the same prior for the ϕ_t's, σ_t^2's, and τ_t^2's. As in the other `spBayes` model functions, one can choose among several popular spatial correlation functions including the exponential, spherical, Gaussian and Matérn. The exponential correlation function is specified in the `spDynLM` call below. Unlike other model functions described in the preceding sections, the `spDynLM` function will accept `NA` $y_t(\mathbf{s})$ values. The sampler will provide posterior predictive samples for these missing values. If the `get.fitted` argument is `TRUE` then these posterior predictive samples are saved along with posterior *fitted* values for locations where the outcomes are observed.

```
> mods <- lapply(paste("O3.8HRMAX.",1:N.t, "~cMAXTMP.",1:N.t,
```

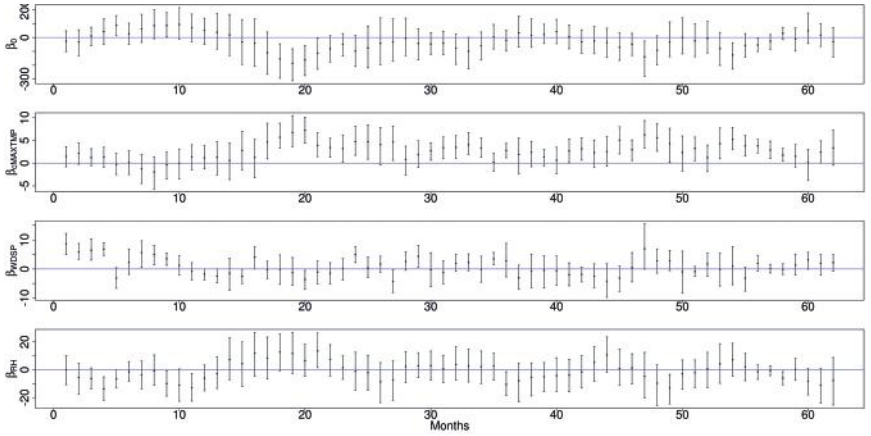

Figure 11.11 *Posterior distribution medians and 95% credible intervals for model intercept and predictors.*

```
+                       "+WDSP.",1:N.t, "+RH.",1:N.t, sep=""),
+                        as.formula)
> p <- 4 ##number of predictors
> coords <- NYOzone.dat[,c("X.UTM","Y.UTM")]/1000
> max.d <- max(iDist(coords))
> starting <- list("beta"=rep(0,N.t*p),
+                  "phi"=rep(3/(0.5*max.d),N.t),
+                  "sigma.sq"=rep(2,N.t), "tau.sq"=rep(1,N.t),
+                  "sigma.eta"=diag(rep(0.01, p)))
> tuning <- list("phi"=rep(2, N.t))
> priors <- list("beta.0.Norm"=list(rep(0,p), diag(100000,p)),
+                "phi.Unif"=list(rep(3/(0.9*max.d), N.t),
+                                rep(3/(0.05*max.d), N.t)),
+                "sigma.sq.IG"=list(rep(2,N.t), rep(25,N.t)),
+                "tau.sq.IG"=list(rep(2,N.t), rep(25,N.t)),
+                "sigma.eta.IW"=list(2, diag(0.001,p)))
> n.samples <- 5000
> m.i <- spDynLM(mods, data=NYOzone.dat,
+                coords=as.matrix(coords), starting=starting,
+                tuning=tuning, priors=priors, get.fitted=TRUE,
+                cov.model="exponential", n.samples=n.samples,
+                n.report=2500)
```

Time series plots of parameters' posterior summary statistics are often useful for exploring the temporal evolution of the parameters. In the case of the regression coefficients, these plots describe the time-varying trend in the outcome and impact of covariates. For example, the sinusoidal pattern in the model intercept, β_0, seen in Figure 11.11, correlates strongly with both cMAXTMP, RM, and to a lesser degree with WDSP. With only a maximum of 28 observations within each time step, there is not much information to inform estimates of $\boldsymbol{\theta}$. As seen in Figure 11.12, this paucity of information is reflected in the imprecise CI's for the ϕ's and small deviations from the priors on σ^2 and τ^2. There are, however, noticeable trends in the variance components over time.

Figure 11.13 shows the observed and predicted values for the three stations used for validation. Here, open circle symbols indicate those observations used for parameter

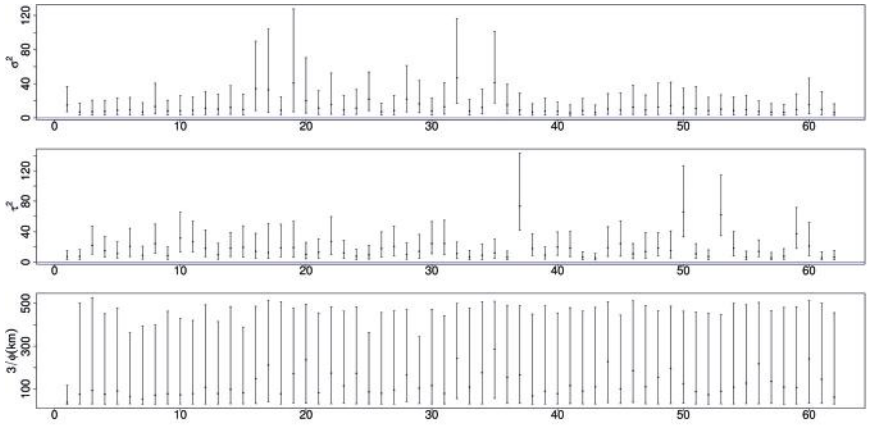

Figure 11.12 *Posterior distribution medians and 95% credible intervals for* $\boldsymbol{\theta}$ *and* τ^2.

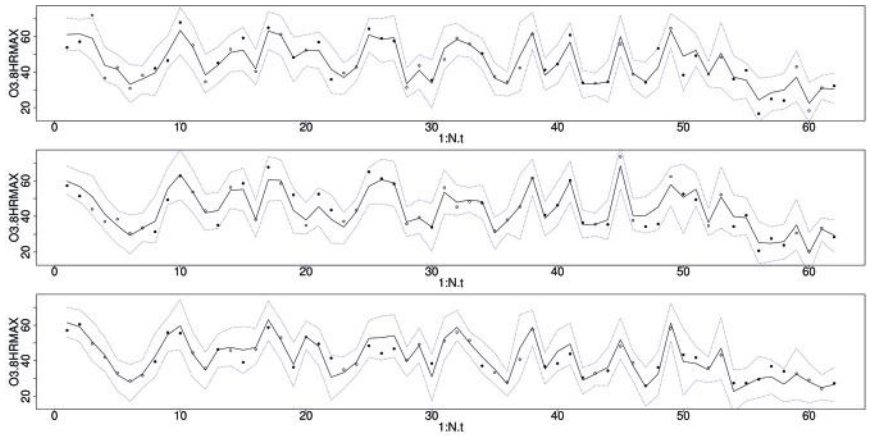

Figure 11.13 *Posterior predicted distribution medians and 95% credible intervals, solid and dashed lines, respectively, for three stations. Open circle symbols indicate those observations use for model parameter estimation and filled circle symbols indicate those observations withheld for validation.*

estimation and filled circles identify holdout observations. The posterior predicted median and 95% CI's are overlaid using solid and dashed lines, respectively. Three of the 36 holdout measurements fell outside of their 95% predicted CI, a ∼92% coverage rate. As noted in Sahu and Bakar (2012), there is a noticeable reduction in ozone levels in the last two weeks in August.

11.6 Geostatistical space-time modeling driven by differential equations

The objective of this section is to consider space-time modeling in the context of stochastic differential equations, in particular, using stochastic diffusion processes. Such processes arise frequently in application. For example, we find various environmental diffusions: for emerging diseases such as avian or H1N1 flu, for the progression of invasive species, and for transformation of the landscape. In Section 8.8.3, we find a diffusion model to describe a space-time point pattern, urban development with regard to single family homes. Here, we consider spread in space and time in the geostatistical setting, with associated uncertainty, with potential explanatory covariates. Our starting point is a deterministic integro-difference

equation or partial differential equation. Many of the ideas in this section arise from the work of Wikle and colleagues. See, e.g., Wikle and Hooten (2006, 2010) and references therein. In general, differential equations that have analytical solutions are too simple to capture what we seek in practice. So, to accommodate more flexible forms, we adopt a strategy which carries out discretization in time. In fact, it may be argued that we should begin with a temporally discretized version, incorporating the features we seek, rather than attempting to frame a particular SDE.

We continue to work within our hierarchical paradigm,

$$[data|process, parameters][process|parameters][parameters]$$

We continue to work within the Bayesian framework, employing structured dependence in space and time. So, model fitting using MCMC will be challenging and we will typically have to resort to dimension reduction techniques (see Chapter 12).

More precisely, our discretization envisions continuous space with discrete time, i.e., $w_t(\mathbf{s})$. Without loss of generality we can take time to be $t \in \{1, 2, ..., T\}$. The customary terminology here refers to $w_t(\mathbf{s})$ as a *dynamical* process. In fact, we simplify to a first order Markov process, i.e., for the finite set of locations $\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n$, let $\mathbf{w}_t = (w_t(\mathbf{s}_1), w_t(\mathbf{s}_2), ..., w_t(\mathbf{s}_n))^T$. Then $[\mathbf{w}_t|\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_{t-1}] = [\mathbf{w}_t \mid \mathbf{w}_{t-1}]$. For example, a linear update would be

$$\mathbf{w}_t = H\mathbf{w}_{t-1} + \boldsymbol{\eta}_t \tag{11.30}$$

where $\eta_t(\mathbf{s})$ incorporates spatial structure. In the literature, this is referred to as a vector AR(1) model and H is called the propagator matrix. The modeling challenge is the specification of H.

We can look at several cases. For instance, consider $H = I$. Note that this form will not provide stationary behavior but, more importantly, there is no interaction across space and time and so this choice would not be realistic for most dynamic processes of interest. Next, consider $H = \text{Diag}(h)$ where $\text{Diag}(h)$ has diagonal elements $0 < h_i < 1$. Now, we achieve stationarity but still we have no space-time interaction. A more general form is an integro-difference equation (IDE),

$$w_t(\mathbf{s}) = \int h(\mathbf{s}, \mathbf{r}; \phi) w_{t-1}(\mathbf{r}) d\mathbf{r} + \eta_t(\mathbf{s}). \tag{11.31}$$

We see that (11.31) does enable the dynamics we seek. In particular, in (11.31), h is a "redistribution kernel," providing redistribution in space which determines the rate of *diffusion* and the *advection*. If we require $w > 0$, then we could work with $\log w_t(\mathbf{s}) = \log(\int h(\mathbf{s}, \mathbf{r}; \phi) w_{t-1}(\mathbf{r}) d\mathbf{r}) + \eta_t(\mathbf{s})$. Again, we resort to discretization in order to supply the H matrix. In this regard, we might begin with forms for h in (11.31); would we want a stationary choice, $h(\mathbf{s}, \mathbf{r}; \phi)$ or would a time-dependent choice, $h_t(\mathbf{s}, \mathbf{r}; \phi)$ be more appropriate? Might ϕ depend upon $\mathbf{r}$?

Recall the linear partial differential equation (PDE), $\frac{dw(\mathbf{s},t)}{dt} = h(\mathbf{s})w(\mathbf{s}, t)$. Applying finite differencing yields $w(\mathbf{s}, t + \Delta t) - w(\mathbf{s}, t) = h(\mathbf{s})w(\mathbf{s}, t)\Delta t$, i.e., $w(\mathbf{s}, t + 1) \approx \tilde{h}(\mathbf{s})w(\mathbf{s}, t)$. We see that the linear PDE suffers the same problems as we noted above. There is no space-time interaction; there is no redistribution over space. So, we need more general PDE's which, as noted above, can motivate an IDE, can illuminate the choice of H.

In this regard, it may be useful to note the "forward" vs. "backward" perspective associated with an IDE. In one sense, we can think of (11.31) as moving forward in time, taking us from a current "state," $w_t(\mathbf{s}), \mathbf{s} \in D$ to a new state, $w_{t+1}(\mathbf{s}), \mathbf{s} \in D$. In another sense, we can look backwards, thinking of (11.31) as clarifying how the "state" $w_t(\mathbf{r}), \mathbf{r} \in D$ contributed to give us the current state, $w_{t+1}(\mathbf{s}), \mathbf{s} \in D$. Depending upon the process we are modeling, specification of h may emerge more naturally under one perspective rather

than the other. Also, IDE's can be specified directly without using PDE's. That is, $h(\mathbf{s}, \mathbf{r})$ can be developed using process-based assumptions, e.g., as a sum of a survival/growth term plus a birth/replenishment term as in Ghosh et al. (2012).

Now, consider a diffusion in one dimension. Fick's Law of diffusion (Fick, 1855) asserts that the diffusive flux from *high* concentration to *low* is $-\delta \frac{\partial w(x,t)}{\partial x}$ with δ being the diffusion coefficient. More flexible dynamics arise with a location varying diffusion coefficient $\delta(x)$ supplying a location varying diffusion. The associated diffusion equation is $\partial w/\partial t = -\partial \mathtt{flux}/\partial x$, i.e., $\frac{\partial w(x,t)}{\partial t} = \frac{\partial}{\partial x}(\delta(x)\frac{\partial w(x,t)}{\partial x})$ Applying the chain rule, the one-dimensional diffusion equation is

$$\frac{\partial w(x,t)}{\partial t} = \delta^{'}(x)\frac{\partial w(x,t)}{\partial x} + \delta(x)\frac{\partial^2 w(x,t)}{\partial x^2}. \tag{11.32}$$

Moving to two-dimensional space, writing $\mathbf{s} = (x, y)$, the diffusive flux is $-\delta(x, y)\nabla w(x, y, t)$ where $\nabla w(x, y, t)$ is the concentration gradient at time t. Now, the resulting diffusion PDE is

$$\frac{\partial w(x,y,t)}{\partial t} = \frac{\partial}{\partial x}(\delta(x,y)\frac{\partial w(x,y,t)}{\partial x}) + \frac{\partial}{\partial y}(\delta(x,y)\frac{\partial w(x,y,t)}{\partial y}). \tag{11.33}$$

We can complete the chain rule calculation to explicitly obtain the spatial diffusion equation, noting that it will involve the second order partial derivatives, $\frac{\partial^2 w}{\partial x^2}$ and $\frac{\partial^2 w}{\partial y^2}$. Now, suppose we introduce Δt, Δx, Δy and replace ∂'s with finite differences (first forward and second order centered). The resulting expressions are elaborate and messy but are developed in careful detail in Hooten and Wikle (2007). The critical point is that, after the smoke clears, we obtain the propagator matrix H to insert into $\mathbf{w}_{t+\Delta t} = H\mathbf{w}_t$. Again, we would add independent spatial noise at each time point, $\boldsymbol{\eta}_t$. Evidently, we are back to our earlier redistribution form in (11.6).

Hooten and Wikle (2007) illustrate with data from the U.S. Breeding Bird Survey, focusing on the Eurasian collared dove. For the years 1986–2003, the data consist of recorded bird counts by sight (for three minutes) over a collection of routes each roughly of length 40kms with 50 stops per route. The counts are attached to grid boxes i in year t and are given a Poisson specification reflecting the number of visits to the site in a given year and a model for the associated expected counts (intensities, see Chapter 8) per visit. The H matrix is a function of the vector of local diffusion coefficients. Since the number of sites is large, dimension reduction is applied to the vector of w's (see Chapter 12). We do not offer further detail here, encouraging the reader to consul the Hooten and Wikle (2007) paper. Instead we present a different geostatistical example below.

The foregoing dynamics redistribute the existing population spatially over time. However, in many situations it would be the case that there is growth or decline in the population. For instance, with housing stock, while some new homes are built, others are torn down. With species populations, change in population size is *density dependent*; competition may encourage or discourage population growth. Hence, we might attempt to add a growth rate to the model. An illustrative choice, which we employ below, is the logistic differential equation (see, e.g., Kot, 2001),

$$\frac{\partial w(\mathbf{s},t)}{\partial t} = rw(\mathbf{s},t)\left(1 - \frac{w(\mathbf{s},t)}{K}\right). \tag{11.34}$$

Here, r is the growth rate and K is the carrying capacity. In practice, we would imagine a spatially varying growth rate, $r(\mathbf{s})$ and perhaps even a spatially varying capacity, $K(\mathbf{s})$.

Turning to more general structures, suppose $w(\mathbf{s}, t)$ is a mean (second stage) specification for a space-time geostatistical model or GLM (or perhaps an intensity for a space-time point pattern, as in Section 10.3.3). A general deterministic diffusion PDE for $w(\mathbf{s}, t)$ looks like

$$\frac{\partial w(\mathbf{s},t)}{\partial t} = a(w(\mathbf{s},t), v(\mathbf{s},t); \theta) \tag{11.35}$$

where $v(\mathbf{s},t)$ includes other potential variables; in its simplest form, $v(\mathbf{s},t)=t$. Furthermore, we might extend θ to $\theta(\mathbf{s})$ or perhaps to $\theta(\mathbf{s},t)$.

To think about adding uncertainty, ignore location $\mathbf{s}$ for the moment and consider a usual nonlinear differential equation, $d\mu(t)=a(\mu(t),t,\theta)dt$ with $\mu(0)=\mu_0$. A simple way to add stochasticity is to make θ random. However, this imposes the likely unreasonable assumption that the functional form of the equation is *true*. Instead, we might assume $d\mu(t)=a(\mu(t),t,\theta)dt+b(\mu(t),t,\theta)dZ(t)$ where $Z(t)$ is variance 1 Brownian motion over R^1 (Section 3.2) with a and b the "drift" and "volatility," respectively. Now we obtain a *stochastic* differential equation (SDE) in which we would still assume θ to be random. A bit more generality is achieved with the form, $d\mu(t)=a(\mu(t),t,\theta(t))dt$ where $\theta(t)$ is driven by an SDE, $d\theta(t)=g(\theta(t),t,\beta)dt+h(\theta(t),\sigma)dZ(t)$ where, again, $Z(t)$ is variance 1 Brownian motion.

Now, we add space. We write $d\mu(\mathbf{s},t)=a(\mu(\mathbf{s},t),t,\theta(\mathbf{s}))dt$ with $\mu(\mathbf{s},0))=\mu_0(\mathbf{s})$, a partial differential equation (PDE). We add randomness through $\theta(\mathbf{s})$, a process realization, resulting in a stochastic process of differential equations. Again, we don't believe the form of the PDE is true. So, we write

$$d\mu(\mathbf{s},t)=a(\mu(\mathbf{s},t),t,\theta(\mathbf{s}))dt+b(\mu(\mathbf{s},t),t,\theta(\mathbf{s}))dZ(\mathbf{s},t)\,, \tag{11.36}$$

where we need to extend our modeling of Brownian motion to $Z(\mathbf{s},t)$. For a fixed finite set of spatial locations we customarily assume independent Brownian motion at each location, allowing the $\theta(\mathbf{s})$ process to provide the spatial dependence. Again, we can work with the more general form, $d\mu(\mathbf{s},t)=a(\mu(\mathbf{s},t),t,\theta(\mathbf{s},t))dt$ where, simplifying the volatility, $d\theta(\mathbf{s},t)=g(\theta(\mathbf{s},t))dt+bdZ(\mathbf{s},t)$. When $d\theta(\mathbf{s},t)=\alpha(\theta(\mathbf{s},t)-\theta(\mathbf{s}))dt+bdZ(\mathbf{s},t)$ with $\theta(\mathbf{s})$ a process realization as above, we have characterized $\theta(\mathbf{s},t)$ through an infinite dimensional SDE and it can be shown that the associated covariance function of the space time process is separable.

The usual space-time "geostatistics" setting with observations at locations and times in the foregoing context takes the form $Y(\mathbf{s},t)=w(\mathbf{s},t)+\epsilon(\mathbf{s},t)$ where now, $w(\mathbf{s},t)$ is modeled through a stochastic PDE. In particular, the logistic equation now extended to space and time becomes

$$\frac{\partial w(\mathbf{s},t)}{\partial t}=r(\mathbf{s},t)w(\mathbf{s},t)\left(1-\frac{w(\mathbf{s},t)}{K(\mathbf{s})}\right). \tag{11.37}$$

Again, with time discretized to intervals Δt, indexed as $t_j, j=0,1,2,...J$ and locations $\mathbf{s}_i$, our data takes the form $\{Y(\mathbf{s}_i,t_j)\}$ with resulting dynamic model $Y(\mathbf{s}_i,t_j)=w(\mathbf{s}_i,t_j)+\varepsilon(\mathbf{s}_i,t_j)$. Using Euler's approximation yields the difference equation:

$$\Delta w(\mathbf{s},t_j)=r(\mathbf{s},t_j)w(\mathbf{s},t_{j-1})\left[1-\frac{w(\mathbf{s},t_{j-1})}{K(\mathbf{s})}\right]\Delta t, \tag{11.38}$$

$$w(\mathbf{s},t_j)\approx w(\mathbf{s},0)+\sum_{l=1}^{j}\Delta w(\mathbf{s},t)\,. \tag{11.39}$$

For $r(\mathbf{s},t)$, we adopt the model given above for $\theta(\mathbf{s},t)$, i.e., $dr(\mathbf{s},t)=\alpha(r(\mathbf{s},t)-r(\mathbf{s}))dt+bdZ(\mathbf{s},t)$. For the initial positive $w(\mathbf{s},0)$ and $K(\mathbf{s})$ we can use log-Gaussian spatial processes with regression forms for the means:

$$\log w(\mathbf{s},0)=\mu_w(X_w(\mathbf{s}),\beta_w)+\eta_w(\mathbf{s})$$

with $\eta_w(\mathbf{s})\sim \mathrm{GP}\left(0,\sigma_w^2\rho_w(\mathbf{s}-\mathbf{s}';\phi_w)\right)$ and

$$\log K(\mathbf{s})=\mu_K(X_K(\mathbf{s}),\beta_K)+\eta_K(\mathbf{s})$$

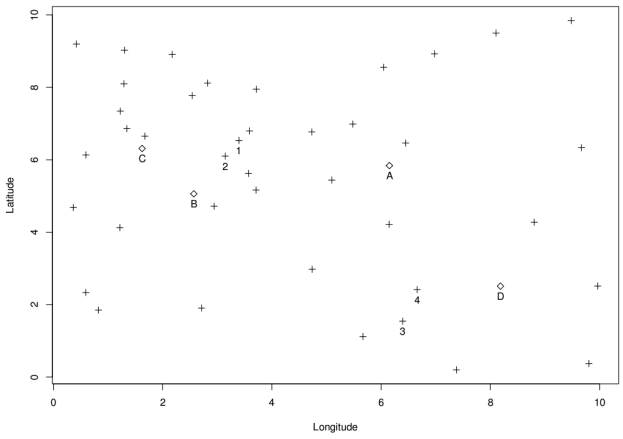

Figure 11.14 *Simulated data at four selected locations marked as 1, 2, 3, and 4 are shown as small circles. Hold-out data at four randomly chosen locations for out-of-sample validation is shown in diamond shape and marked as A, B, C and D.*

with $\eta_K(\mathbf{s}) \sim \text{GP}\left(0, \sigma_K^2 \rho_K(\mathbf{s}-\mathbf{s}'; \phi_K)\right)$. In fact, to simplify here, we assume $K(s)$ known and set to 1, yielding interpretation of w on a *percent* scale with K as 100% or full capacity. We add similar modeling for $\mu_r(\mathbf{s})$. Altogether, we have specified a hierarchical model. After specifying priors, we fit with MCMC cumulatively in w over time. We consider the usual prediction questions: (i) interpolating the past at new locations and (ii) forecasting the future at current and new locations

For our specific example, we use a 10×10 study region introducing 44 locations over 30 time periods with four sites retained as holdout for validation (see Figure 11.14). To illustrate with a different choice of covariance function, we adopted the Matérn with smoothness parameter $\nu = 3/2$ for both $w_0(\mathbf{s})$ and $r(\mathbf{s})$. Hence, the resulting space time covariance function for $r(\mathbf{s}, t)$ becomes

$$\sigma_r^2 \exp\left(-\alpha\left|t_{j_1} - t_{j_2}\right|\right)\left(1 + \phi_r\left|\left|\mathbf{s}_{i_1} - \mathbf{s}_{i_2}\right|\right|\right) \exp\left(-\phi_r\left|\mathbf{s}_{i_1} - \mathbf{s}_{i_2}\right|\right) \ .$$

We use the simulated $\mathbf{r}$ and $\mathbf{w}_0$ with the transition equation recursively to obtain $\boldsymbol{\Delta} w_j$ and $\mathbf{w}_j$ for each of the 30 periods. The observed data are sampled as mutually independent given $\mathbf{w}_j$ with the random noise $\boldsymbol{\varepsilon}_j$. The data at four selected locations marked as 1, 2, 3, and 4 in Figure 11.14 are shown as small circles. We leave out the data at four randomly chosen locations (shown in diamond shape and marked as A, B, C and D in Figure 11.14) for spatial prediction and out-of-sample validation for our model.

We fit the foregoing model to the data at the remaining 40 locations (hence a 40×30 spatiotemporal data set). We use very vague priors for the constant means: $\pi(\mu_w) \sim N\left(0, 10^8\right)$ and $\pi(\mu_r) \sim N\left(0, 10^8\right)$. We use conjugate gamma priors for the precision parameters of r and w_0: $\pi\left(1/\sigma_r^2\right) \sim \mathit{Gamma}(1,1)$ and $\pi\left(1/\sigma_w^2\right) \sim \mathit{Gamma}(1,1)$. The positive parameter for the temporal correlation of r also has a vague log-normal prior: $\pi(\alpha) \sim \mathit{log\text{-}N}\left(0, 10^8\right)$. Because the spatial range parameters ϕ_r and ϕ_w are only weakly identified, we only use informative and discrete prior for them. Indeed, we have chosen 20 values (from 0.1 to 2.0) and assume uniform priors over them for both ϕ_r and ϕ_w.

We use the random-walk Metropolis-Hastings algorithm to simulate posterior samples of $\mathbf{r}$ and $\mathbf{w}_0$. We draw the entire vector of $\mathbf{w}_0$ for all 40 locations as a single block in every

Model Parameters	True Value	Posterior Mean	95% Equal-tail Interval
μ_w	-4.2	-4.14	(-4.88, -3.33)
σ_w	1.0	0.91	(0.62, 1.46)
ϕ_w	0.7	0.77	(0.50, 1.20)
σ_ε	0.05	0.049	(0.047, 0.052)
μ_r	0.24	0.24	(0.22, 0.26)
σ_r	0.08	0.088	(0.077, 0.097)
ϕ_r	0.7	0.78	(0.60, 1.10)
α	0.6	0.64	(0.51, 0.98)

Table 11.9 *Parameters and their posterior inference for the simulated example.*

iteration. Because $\mathbf{r}$ is very high-dimensional (a 40×30 matrix concatenated into a vector), we cannot draw the entire matrix of $\mathbf{r}$ as one block and achieve a satisfactory acceptance rate. So, we partition $\mathbf{r}$ into 40 row blocks (location-wise) in every odd-numbered iteration and 30 column blocks (period-wise) in every even numbered iteration. Each block is drawn in one Metropolis step. The posterior samples start to converge after about 30,000 iterations. Given the sampled $\mathbf{r}$ and $\mathbf{w}_0$, the mean parameters μ_r, μ_w and the precision parameters $1/\sigma_r^2$ and $1/\sigma_\Lambda^2$ all have conjugate priors, and therefore their posterior samples are drawn directly. ϕ_r and ϕ_w have discrete priors and therefore are also directly sampled. We use the random-walk Metropolis-Hastings algorithm to draw α.

We obtain 200,000 samples from the algorithm and discard the first 100,000 as burn-in. For the posterior inference, we use 4,000 subsamples from the remaining 100,000 samples, with a thinning equal to 25. The posterior means and 95% equal-tail Bayesian posterior predictive intervals for the model parameters are presented in Table 11.9. Evidently we are recovering the true parameter values very well.

Figure 11.15 displays the posterior mean of the growth curves and 95% Bayesian predictive intervals for the four locations which were used in the fitting (1, 2, 3 and 4), compared with the actual latent growth curve $w(\mathbf{s}, t)$ and observed data. Up to the uncertainty in the model we approximate the actual curves very well. The fitted mean growth curves almost perfectly overlap with the actual simulated growth curves. The empirical coverage of the Bayesian predictive bounds is 93.4%.

Interpolation yields the predictive growth curve for the four hold out locations (A, B, C and D). In Figure 11.15 we display the means of the predicted curves and 95% Bayesian predictive intervals, together with the hold-out data. We can see the spatial prediction captures the patterns of the hold-out data very well. The predicted mean growth curves overlap with the actual simulated growth curves very well except for location D (because location D is rather far from all the observed locations). The empirical coverage of the Bayesian predictive intervals is 95.8%.

Finally, for comparison and in the absence of covariates, we also fit the following customary process realization model with space-time random effects to the simulated data set

$$\mathbf{y}_j = \mu \mathbf{1} + \boldsymbol{\xi}_j + \boldsymbol{\epsilon}_j \; ; \;\; \boldsymbol{\epsilon}_j \sim N\left(0, \sigma_\epsilon^2 I_n\right), \;\; \text{for} \;\; j = 0, 1, \ldots, J \tag{11.40}$$

where the random effects $\boldsymbol{\xi} = [\boldsymbol{\xi}_0, \ldots, \boldsymbol{\xi}_J]$ come from a Gaussian process with a separable spatio-temporal correlation of the form:

$$C_\xi(t - t', s - s') = \sigma_\xi^2 \exp\left(-\alpha_\xi \left|t - t'\right|\right) \left(\phi_\xi \left|s - s'\right|\right)^\nu \kappa_\nu \left(\phi_\xi \left|s - s'\right|\right), \; \nu = \frac{3}{2} \, . \tag{11.41}$$

Comparison of model performance between the SPDE model and this model is done using spatial prediction at the four new locations. In Figure 11.16 we display the means of the predicted curves and 95% Bayesian predictive intervals, together with the hold-out data.

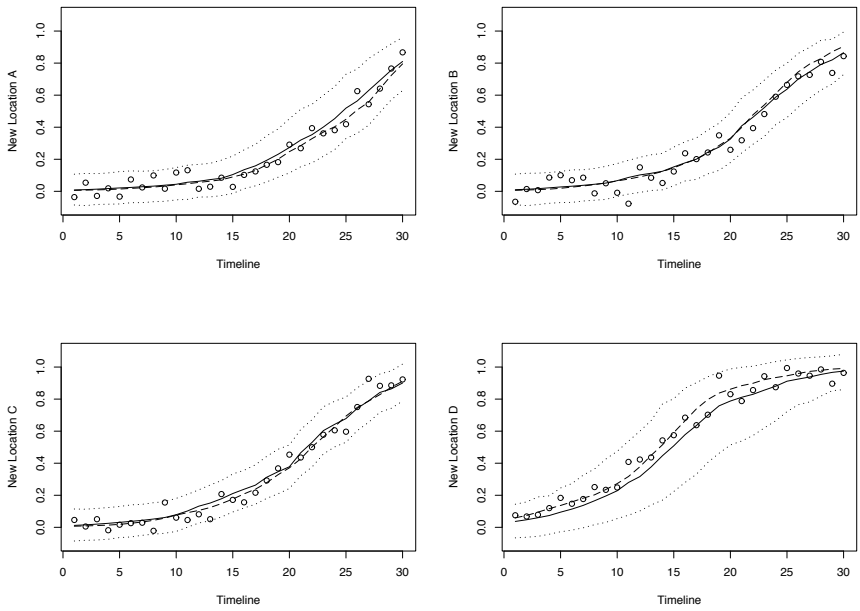

Figure 11.15 *Hold-out space-time geostatistical data at four locations, actual (dashed line) and predicted mean growth curves (solid line) and 95% predictive intervals (dotted line) by our model (11.36) for the simulated data example.*

For the four hold-out sites, the average mean square error for the SPDE model is 1.75×10^{-3} versus 3.34×10^{-3} for the standard model. The average length of the 95% predictive intervals for the SPDE model is 0.29 versus 0.72 for the standard model. It is evident that, when we have an SPDE driving the data, discretized as above, we can learn about it and will do better in terms of prediction than using a standard model.

11.7 Areal unit space-time modeling

We now return to spatiotemporal modeling for areal unit data, following the discussion of Equations (11.4) and (11.5) in Section 11.1. Recall that we have briefly discussed general space-time SAR modeling in Section 4.4.2. Here, we focus on the spatiotemporal disease mapping setting.

11.7.1 Aligned data

In the aligned data case, matters are relatively straightforward. Consider for example the spatiotemporal extension of the standard disease mapping setting described in Section 6.4.1. Here we would have $Y_{i\ell t}$ and $E_{i\ell t}$, the observed and expected disease counts in county i and demographic subgroup ℓ (race, gender, etc.) during time period t (without loss of generality we let t correspond to years in what follows). Again the issue of whether the $E_{i\ell t}$ are internally or externally standardized arises; in the more common former case we would use $n_{i\ell t}$, the number of persons at risk in county i during year t, to compute $E_{i\ell t} = n_{i\ell t}(\sum_{i\ell t} Y_{i\ell t}/\sum_{i\ell t} n_{i\ell t})$. That is, $E_{i\ell t}$ is the number of cases we would expect if the grand disease rate (all regions, subgroups, and years) were in operation throughout. The extension

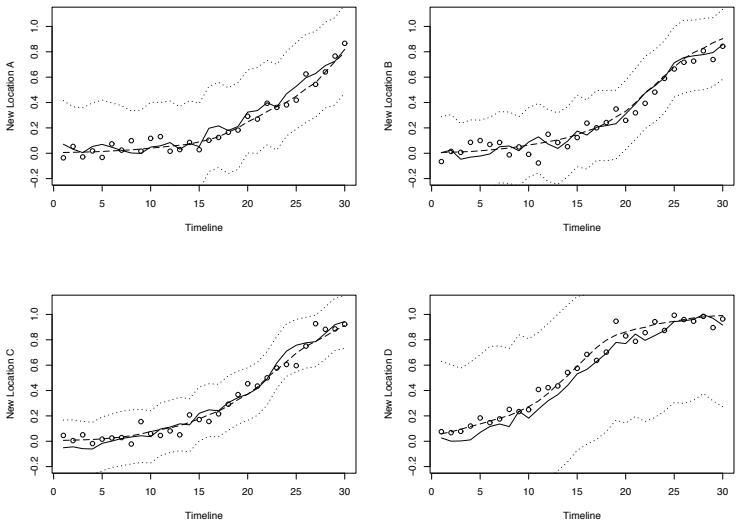

Figure 11.16 *Hold-out space-time geostatistical data at four locations, actual (dashed line) and predicted mean growth curves (solid line) and 95% predictive intervals (dotted line) by the benchmark model (11.40) for the simulated data example.*

of the basic Section 6.4.1 Poisson regression model is then

$$Y_{i\ell t} \mid \mu_{i\ell t} \overset{ind}{\sim} Po\left(E_{i\ell t}\, e^{\mu_{i\ell t}}\right) ,$$

where $\mu_{i\ell t}$ is the log-relative risk of disease for region i, subgroup ℓ, and year t.

It now remains to specify the main effect and interaction components of $\mu_{i\ell t}$, and corresponding prior distributions. First the main effect for the demographic subgroups can be taken to have ordinary linear regression structure, i.e., $\varepsilon_\ell = \mathbf{x}'_\ell \boldsymbol{\beta}$, with a flat prior for $\boldsymbol{\beta}$. Next, the main effects for time (say, δ_t) can be assigned flat priors (if we wish to view them as fixed effects, i.e., temporal dummy variables), or an $AR(1)$ specification (if we wish them to reflect temporal autocorrelation). In some cases an even simpler structure (say, $\delta_t = \gamma t$) may be appropriate.

Finally, the main effects for space are similar to those assumed in the nontemporal case. Specifically, we might let

$$\psi_i = \mathbf{z}'_i \boldsymbol{\omega} + \theta_i + \phi_i ,$$

where $\boldsymbol{\omega}$ has a flat prior, the θ_i capture *heterogeneity* among the regions via the i.i.d. specification,

$$\theta_i \overset{iid}{\sim} N(0\,,\, 1/\tau) ,$$

and the ϕ_i capture regional *clustering* via the CAR prior,

$$\phi_i \mid \phi_{j \neq i} \sim N(\bar{\phi}_i\,,\, 1/(\lambda m_i)) .$$

As usual, m_i is the number of neighbors of region i, and $\bar{\phi}_i = m_i^{-1} \Sigma_{j \in \partial_i}\, \phi_j$.

Turning to spatiotemporal interactions, suppose for the moment that demographic effects are not affected by region and year. Consider then the *nested* model,

$$\theta_{it} \overset{iid}{\sim} N(0\,,\, 1/\tau_t) \;\text{ and }\; \phi_{it} \sim CAR(\lambda_t) , \tag{11.42}$$

where $\tau_t \overset{iid}{\sim} G(a,b)$ and $\lambda_t \overset{iid}{\sim} G(c,d)$. Provided these hyperpriors are not too informative, this allows "shrinkage" of the year-specific effects toward their grand mean, and in a way that allows the data to determine the amount of shrinkage.

Thus our most general model for $\mu_{i\ell t}$ is

$$\mu_{i\ell t} = \mathbf{x}'_\ell \boldsymbol{\beta} + \delta_t + \mathbf{z}'_i \boldsymbol{\omega} + \theta_{it} + \phi_{it} \ ,$$

with corresponding joint posterior distribution proportional to

$$L(\boldsymbol{\beta}, \delta, \boldsymbol{\omega}, \boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{y}) p(\delta) p(\boldsymbol{\theta}|\tau) p(\boldsymbol{\phi}|\lambda) p(\tau) p(\lambda) \ .$$

Computation via univariate Metropolis and Gibbs updating steps is relatively straightforward (and readily available in this aligned data setting in the `WinBUGS` language). However, convergence can be rather slow due to the weak identifiability of the joint parameter space. As a possible remedy, consider the the simple space-only case again for a moment. We may transform from $(\boldsymbol{\theta}, \boldsymbol{\phi})$ to $(\boldsymbol{\theta}, \boldsymbol{\eta})$ where $\eta_i = \theta_i + \phi_i$. Then $p(\boldsymbol{\theta}, \boldsymbol{\eta}|\mathbf{y}) \propto L(\boldsymbol{\eta}; \mathbf{y}) p(\boldsymbol{\theta}) p(\boldsymbol{\eta} - \boldsymbol{\theta})$, so that

$$p(\eta_i \mid \eta_{j \neq i}, \boldsymbol{\theta}, \mathbf{y}) \propto L(\eta_i; y_i) \, p(\eta_i - \theta_i \mid \{\eta_j - \theta_j\}_{j \neq i})$$

and

$$p(\theta_i \mid \theta_{j \neq i}, \boldsymbol{\eta}, \mathbf{y}) \propto p(\theta_i) \, p(\eta_i - \theta_i \mid \{\eta_j - \theta_j\}_{j \neq i}) \ .$$

This simple transformation improves matters since each η_i full conditional is now well identified by the data point Y_i, while the weakly identified (indeed, "Bayesianly unidentified") θ_i now emerges in closed form as a normal distribution (since the nonconjugate Poisson likelihood no longer appears).

Example 11.5 The study of the trend of risk for a given disease in space and time may provide important clues in exploring underlying causes of the disease and helping to develop environmental health policy. Waller, Carlin, Xia, and Gelfand (1997) consider the following data set on lung cancer mortality in Ohio. Here Y_{ijkt} is the number of lung cancer deaths in county i during year t for gender j and race k in the state of Ohio. The data are recorded for $J = 2$ genders (male and female, indexed by s_j) and $K = 2$ races (white and nonwhite, indexed by r_k) for each of the $I = 88$ Ohio counties over $T = 21$ years (1968–1988).

We adopt the model,

$$\mu_{ijkt} = s_j \alpha + r_k \beta + s_j r_k \xi + \theta_{it} + \phi_{it} \ , \tag{11.43}$$

where $s_j = 1$ if $j = 2$ (female) and 0 otherwise, and $r_k = 1$ if $k = 2$ (nonwhite) and 0 otherwise. That is, there is one subgroup (white males) for which there is no contribution to the mean structure (11.52). For our prior specification, we select

$$\begin{gathered} \theta_{it} \overset{ind}{\sim} N\left(0 \, , \, \tfrac{1}{\tau_t}\right) \quad \text{and} \quad \phi_{it} \sim CAR(\lambda_t) \ ; \\ \alpha, \beta, \xi \sim \text{flat} \ ; \\ \tau_t \overset{iid}{\sim} G(1, 100) \quad \text{and} \quad \lambda_t \overset{iid}{\sim} G(1, 7) \ , \end{gathered}$$

where the relative sizes of the hyperparameters in these two gamma distributions were selected following guidance given in Bernardinelli et al. (1995); see also Best et al. (1999) and Eberly and Carlin (2000).

Regarding implementation, five parallel, initially overdispersed MCMC chains were run for 500 iterations. Graphical monitoring of the chains for a representative subset of the parameters, along with sample autocorrelations and Gelman and Rubin (1992) diagnostics, indicated an acceptable degree of convergence by around the 100th iteration.

Demographic subgroup	Contribution to ε_{jk}	Fitted relative risk
White males	0	1
White females	α	0.34
Nonwhite males	β	1.02
Nonwhite females	$\alpha + \beta + \xi$	0.28

Table 11.10 *Fitted relative risks, four sociodemographic subgroups in the Ohio lung cancer data.*

Histograms of the sampled values showed θ_{it} distributions centered near 0 in most cases, but ϕ_{it} distributions typically removed from 0, suggesting that the heterogeneity effects are not really needed in this model. Plots of $E(\tau_t|\mathbf{y})$ and $E(\lambda_t|\mathbf{y})$ versus t suggest increasing clustering and slightly increasing heterogeneity over time. The former might be the result of flight from the cities to suburban "collar counties" over time, while the latter is likely due to the elevated mean levels over time (for the Poisson, the variance increases with the mean).

Fitted relative risks obtained by Waller et al. (1997) for the four main demographic subgroups are shown in Table 11.10. The counterintuitively positive fitted value for nonwhite females may be an artifact of the failure of this analysis to age-standardize the rates prior to modeling (or at least to incorporate age group as another demographic component in the model). To remedy this, consider the following revised and enhanced model, described by Xia and Carlin (1998), where we assume that

$$Y^*_{ijkt} \sim Poisson(E_{ijkt} \exp(\mu_{ijkt})) \,, \tag{11.44}$$

where again Y^*_{ijkt} denotes the observed age-adjusted deaths in county i for sex j, race k, and year t, and E_{ijkt} are the expected death counts. We also incorporate an ecological level smoking behavior covariate into our log-relative risk model, namely,

$$\mu_{ijkt} = \mu + s_j\alpha + r_k\beta + s_j r_k \xi + p_i\rho + \gamma t + \phi_{it} \,, \tag{11.45}$$

where p_i is the true smoking proportion in county i, γ represents the fixed time effect, and the ϕ_{it} capture the random spatial effects over time, wherein clustering effects are nested within time. That is, writing $\boldsymbol{\phi}_t = (\phi_{1t}, \ldots, \phi_{It})'$, we let $\boldsymbol{\phi}_t \sim CAR(\lambda_t)$ where $\lambda_t \stackrel{iid}{\sim} G(c,d)$. We assume that the sociodemographic covariates (sex and race) do not interact with time or space. Following the approach of Bernardinelli, Pascutto et al. (1997), we introduce both sampling error and spatial correlation into the smoking covariate. Let

$$q_i \,|\, p_i \sim N(p_i, \sigma_q^2),\ i = 1, \ldots, I, \ \text{ and} \tag{11.46}$$

$$\mathbf{p} \sim CAR(\lambda_p) \iff p_i \,|\, p_{j\neq i} \sim N(\mu_{p_i}, \sigma^2_{p_i}),\ i = 1, \ldots, I \,, \tag{11.47}$$

where q_i is the current smoking proportion observed in a sample survey of county i (an imperfect measurement of p_i), $\mu_{p_i} = \sum_{j\neq i} w_{ij} p_j / \sum_{j\neq i} w_{ij}$, and $\sigma^2_{p_i} = (\lambda_p \sum_{j\neq i} w_{ij})^{-1}$. Note that the amount of smoothing in the two CAR priors above may differ, since the smoothing is controlled by different parameters λ_ϕ and λ_p. Like λ_ϕ, λ_p is also assigned a gamma hyperprior, namely, a $G(e,f)$.

We ran 5 independent chains using our Gibbs-Metropolis algorithm for 2200 iterations each; plots suggested discarding the first 200 samples as an adequate burn-in period. We obtained the 95% posterior credible sets [–1.14, –0.98], [0.07, 0.28], and [–0.37, –0.01] for α, β, and ξ, respectively. Note that all 3 fixed effects are significantly different from 0, in contrast to our Table 11.10 results, which failed to uncover a main effect for race. The corresponding point estimates are translated into the fitted relative risks for the four sociodemographic

Demographic subgroup	Contribution to ε_{jk}	Fitted log-relative risk	Fitted relative risk
White males	0	0	1
White females	α	–1.06	0.35
Nonwhite males	β	0.18	1.20
Nonwhite females	$\alpha + \beta + \xi$	–1.07	0.34

Table 11.11 *Fitted relative risks, four sociodemographic subgroups in the Ohio lung cancer data*

subgroups in Table 11.11. Nonwhite males experience the highest risk, followed by white males, with females of both races having much lower risks.

11.7.2 Misalignment across years

In this subsection we develop a spatiotemporal model to accommodate the situation of Figure 11.17, wherein the response variable and the covariate are spatially aligned within any given timepoint, but not across timepoints (due to periodic changes in the regional grid). Assuming that the observed disease count Y_{it} for zip i in year t is conditionally independent of the other zip-level disease counts given the covariate values, we have the model,

$$Y_{it} \mid \mu_{it} \overset{ind}{\sim} Po(E_{it} \exp(\mu_{it})),\ i = 1, \ldots, I_t,\ t = 1, \ldots, T,$$

where the expected count for zip i in year t, E_{it}, is proportional to the population count. In our case, we set $E_{it} = Rn_{it}$, where n_{it} is the population count in zip i at year t and $R = (\sum_{it} Y_{it})/(\sum_{it} n_{it})$, the grand asthma hospitalization rate (i.e., the expected counts assume homogeneity of disease rates across all zips and years). The log-relative risk is modeled as

$$\mu_{it} = x_{it}\beta_t + \delta_t + \theta_{it} + \phi_{it}, \tag{11.48}$$

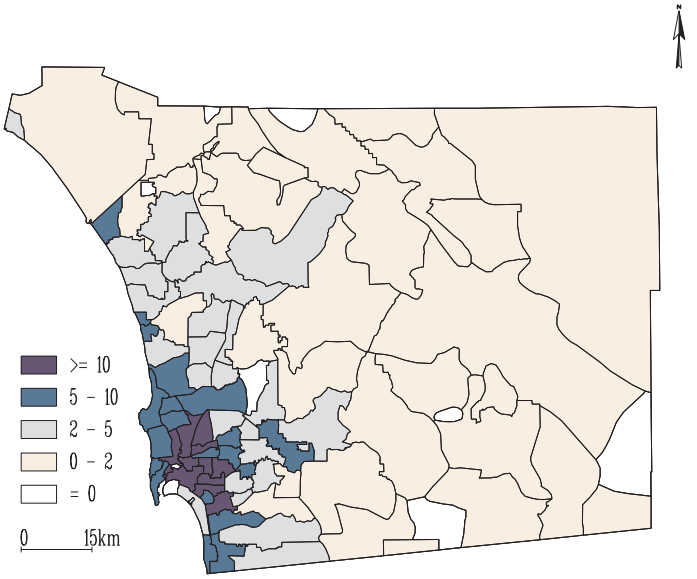

Figure 11.17 *Traffic density (average vehicles per km of major roadway) in thousands by zip code for 1983, San Diego County.*

where x_{it} is the zip-level exposure covariate (traffic density) depicted for 1983 in Figure 11.17, β_t is the corresponding main effect, δ_t is an overall intercept for year t, and θ_{it} and ϕ_{it} are zip- and year-specific heterogeneity and clustering random effects, analogous to those described in Section 11.7.1. The changes in the zip grid over time cloud the interpretation of these random effects (e.g., a particular region may be indexed by different i in different years), but this does not affect the interpretation of the main effects β_t and δ_t; it is simply the analogue of unbalanced data in a longitudinal setting. In the spatiotemporal case, the distributions on these effects become

$$\boldsymbol{\theta}_t \overset{ind}{\sim} N\left(0, \frac{1}{\tau_t} I\right) \text{ and } \boldsymbol{\phi}_t \overset{ind}{\sim} CAR(\lambda_t)\,, \tag{11.49}$$

where $\boldsymbol{\theta}_t = (\theta_1, \ldots, \theta_{I_t})'$, $\boldsymbol{\phi}_t = (\phi_1, \ldots, \phi_{I_t})'$, and we encourage similarity among these effects across years by assuming $\tau_t \overset{iid}{\sim} G(a, b)$ and $\lambda_t \overset{iid}{\sim} G(c, d)$, where G again denotes the gamma distribution. Placing flat (uniform) priors on the main effects β_t and δ_t completes the model specification. Note that the constraints $\sum_i \phi_{it} = 0$, $t = 1, \ldots, T$ must be added to identify the year effects δ_t, due to the location invariance of the CAR prior.

Example 11.6 Asthma is the most common chronic disease diagnosis for children in the U.S. (National Center for Environmental Health, 1996). A large number of studies have shown a correlation between known products and byproducts of auto exhaust (such as ozone, nitrogen dioxide, and particulate matter) and pediatric asthma ER visits or hospitalizations. Several studies (e.g., Tolbert et al., 2000; Zidek et al., 1998; Best et al., 2000) have used hierarchical Bayesian methods in such investigations. An approach taken by some authors is to use proximity to major roadways (or some more refined measure of closeness to automobile traffic) as an omnibus measure of exposure to various asthma-inducing pollutants. We too adopt this approach and use the phrase "exposure" in what follows, even though in fact our traffic measures are really surrogates for the true exposure.

Our data set arises from San Diego County, CA, the region pictured in Figure 11.17. The city of San Diego is located near the southwestern corner of the map; the map's western boundary is the Pacific Ocean, while Mexico forms its southern boundary. The subregions pictured are the zip codes as defined in 1983; as mentioned earlier this grid changes over time. Specifically, during the course of our eight-year (1983–1990) study period, the zip code boundaries changed four times: in 1984, 1987, 1988, and 1990.

The components of our data set are as follows. First, for a given year, we have the number of discharges from hospitalizations due to asthma for children aged 14 and younger by zip code (California Office of Statewide Health Planning and Development, 1997). The primary diagnosis was asthma based on the International Classification of Diseases, code 493 (U.S. Department of Health and Human Services, 1989). Assuming that patient records accurately report the correct zip code of residence, these data can be thought of as error-free.

Second, we have zip-level population estimates (numbers of residents aged 14 and younger) for each of these years, as computed by Scalf and English (1996). These estimates were obtained in `ARC/INFO` using the following process. First, a land-use covariate was used to assist in a linear interpolation between the 1980 and 1990 U.S. Census figures, to obtain estimates at the census block group level. Digitized hard-copy U.S. Postal Service maps or suitably modified street network files provided by the San Diego Association of Governments (SANDAG) were then used to reallocate these counts to the zip code grid for the year in question. To do this, the GIS first created a subregional grid by intersecting the block group and zip code grids. The block group population totals were allocated to the subregions per a combination of subregional area and population density (the latter again based on the land-use covariate). Finally, these imputed subregional counts were reaggregated to the zip grid. While there are several possible sources of uncertainty in these

calculations, we ignore them in our initial round of modeling, assuming these population counts to be fixed and known.

Finally, for each of the major roads in San Diego County, we have mean yearly traffic counts on each road segment in our map. Here "major" roads are defined by SANDAG to include interstate highways or equivalent, major highways, access or minor highways, and arterial or collector routes. The sum of these numbers within a given zip divided by the total length of its major roads provides an aggregate measure of traffic exposure for the zip. These zip-level *traffic densities* are plotted for 1983 in Figure 11.17; this is the exposure measure we use in the following text.

We set $a = 1$, $b = 10$ (i.e., the τ_t have prior mean and standard deviation both equal to 10) and $c = 0.1$, $d = 10$ (i.e., the λ_t have prior mean 1, standard deviation $\sqrt{10}$). These are fairly vague priors designed to let the data dominate the allocation of excess spatial variability to heterogeneity and clustering. (As mentioned near Equation (6.26), simply setting these two priors equal to each other would not achieve this, since the prior for the θ_{it} is specified *marginally*, while that for the ϕ_{it} is specified *conditionally* given the neighboring ϕ_{jt}.) Our MCMC implementation ran 3 parallel sampling chains for 5000 iterations each, and discarded the first 500 iterations as preconvergence "burn-in."

Plots of the posterior medians and 95% equal-tail Bayesian confidence intervals for β_t (not shown) makes clear that, with the exception of that for 1986, all of the β_t's are significantly greater than 0. Hence, the traffic exposure covariate in Figure 11.17 is positively associated with increased pediatric asthma hospitalization in seven of the eight years of our study. To interpret these posterior summaries, recall that their values are on the *log*-relative risk scale. Thus a zip having a 1983 traffic density of 10,000 cars per km of roadway would have median relative risk $e^{10(.065)} = 1.92$ times higher than a zip with essentially no traffic exposure, with a corresponding 95% confidence interval of $(e^{10(.000)}, e^{10(.120)}) = (1.00, 3.32)$. There also appears to be a slight weakening of the traffic-asthma association over time.

Figure 11.18 provides `ARC/INFO` maps of the crude and fitted asthma rates (per thousand) in each of the zips for 1983. The crude rates are of course given by $r_{it} = Y_{it}/n_{it}$, while the fitted rates are given by $R\exp(\hat{\mu}_{it})$, where R is again the grand asthma rate across all zips and years and $\hat{\mu}_{it}$ is obtained by plugging in the estimated posterior means for the various components in Equation (11.48). The figure clearly shows the characteristic Bayesian shrinkage of the crude rates toward the grand rate. In particular, no zip is now assigned a rate of exactly zero, and the rather high rates in the thinly populated eastern part of the map have been substantially reduced. However, the high observed rates in urban San Diego continue to be high, as the method properly recognizes the much higher sample sizes in these zips. There also appears to be some tendency for clusters of similar crude rates to be preserved, the probable outcome of the CAR portion of our model.

11.7.3 Nested misalignment both within and across years

In this subsection we extend our spatiotemporal model to accommodate the situation of Figure 11.19, wherein the covariate is available on a grid that is a refinement of the grid for which the response variable is available (i.e., nested misalignment within years, as well as misalignment across years). Letting the subscript j index the subregions (which we also refer to as *atoms*) of zip i, our model now becomes

$$Y_{ijt} \mid \mu_{ijt} \sim Po(E_{ijt}\exp(\mu_{ijt})),\ i = 1, \ldots, I_t,\ j = 1, \ldots, J_{it},\ t = 1, \ldots, T,$$

where the expected counts E_{ijt} are now Rn_{ijt}, with the grand rate R as before. The population of atom ijt is not known, and so we determine it by areal interpolation as $n_{ijt} = n_{it}(\text{area of atom } ijt)/(\text{area of zip } it)$. The log-relative risk in atom ijt is then modeled as

$$\mu_{ijt} = x_{ijt}\beta_t + \delta_t + \theta_{it} + \phi_{it}, \tag{11.50}$$

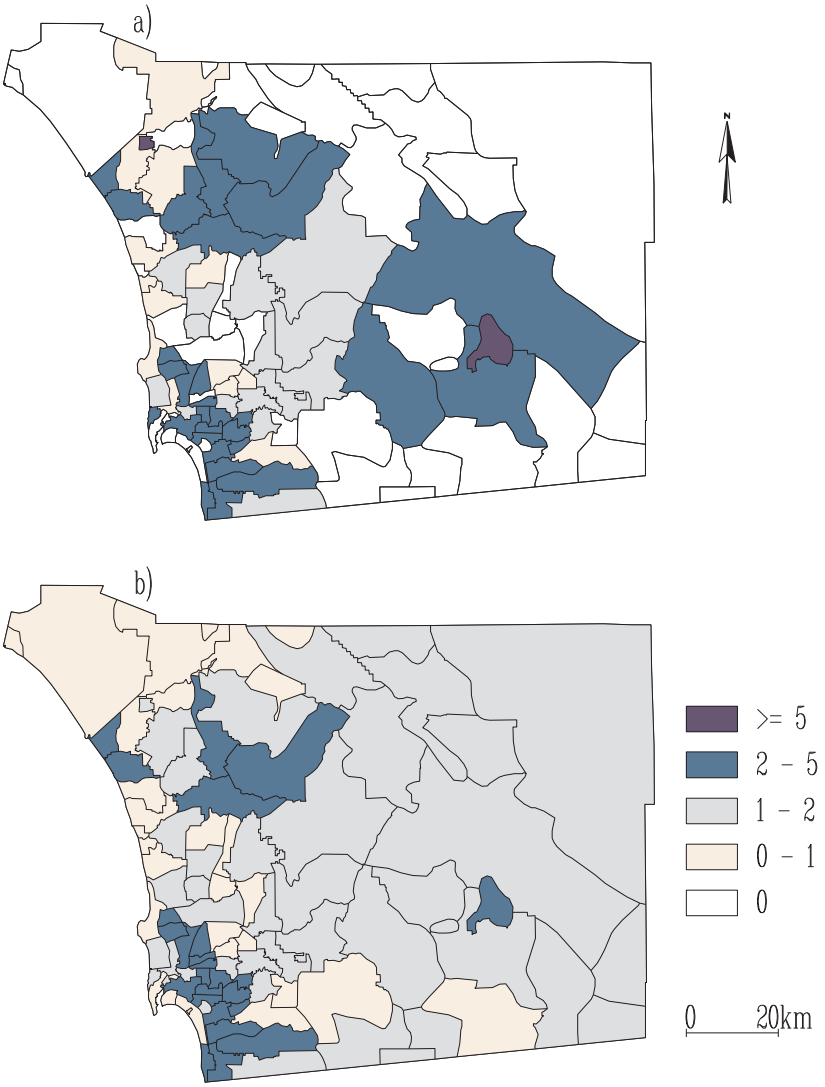

Figure 11.18 *Pediatric asthma hospitalization rate (per thousand children) by zip code for 1983, San Diego County: (a) crude rate, (b) temporally misaligned model fitted rate.*

where x_{ijt} is now the atom-level exposure covariate (depicted for 1983 in Figure 11.19), but β_t, δ_t, θ_{it} and ϕ_{it} are as before. Thus our prior specification is exactly that of the previous subsection; priors for the $\boldsymbol{\theta}_t$ and $\boldsymbol{\phi}_t$ as given in Equation (11.49), exchangeable gamma hyperpriors for the τ_t and λ_t with $a = 1$, $b = 10$, $c = 0.1$, and $d = 10$, and flat priors for the main effects β_t and δ_t.

Since only the zip-level hospitalization totals Y_{it} (and not the atom-level totals Y_{ijt}) are observed, we use the additivity of conditionally independent Poisson distributions to obtain

$$Y_{it} \mid \beta_t, \delta_t, \theta_{it}, \phi_{it} \sim Po\left(\sum_{j=1}^{J_{it}} E_{ijt} \exp(\mu_{ijt})\right), \; i = 1, \ldots, I_t, \; t = 1, \ldots, T. \tag{11.51}$$

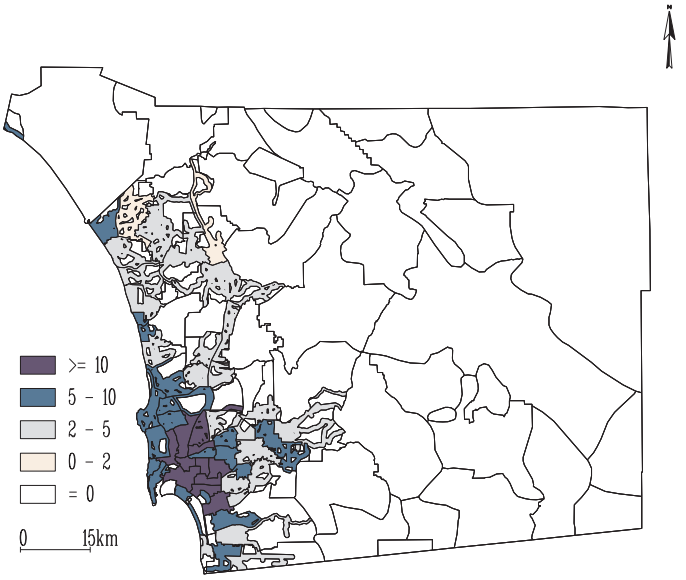

Figure 11.19 *Adjusted traffic density (average vehicles per km of major roadway) in thousands by zip code subregion for 1983, San Diego County.*

Using Expression (11.51), we can obtain the full Bayesian model specification for the observed data as

$$\left[\prod_{t=1}^{T}\prod_{i=1}^{I_t} p\left(y_{it}|\beta_t,\delta_t,\theta_{it},\phi_{it}\right)\right]\left[\prod_{t=1}^{T} p(\boldsymbol{\theta}_t|\tau_t)p(\boldsymbol{\phi}_t|\lambda_t)p(\tau_t)p(\lambda_t)\right] \tag{11.52}$$

As in the previous section, only the τ_t and λ_t parameters may be updated via ordinary Gibbs steps, with Metropolis steps required for the rest.

Note that model specification (11.52) makes use of the atom-level covariate values x_{ijt}, but only the zip-level hospitalization counts Y_{it}. Of course, we might well be interested in *imputing* the values of the missing subregional counts Y_{ijt}, whose full conditional distribution is multinomial, namely,

$$(Y_{i1t},\ldots,Y_{iJ_it}) \,|\, Y_{it},\beta_t,\delta_t,\theta_{it},\phi_{it} \sim Mult(Y_{it},\{q_{ijt}\}) \,, \tag{11.53}$$

$$\text{where} \quad q_{ijt} = \frac{E_{ijt}e^{\mu_{ijt}}}{\sum_{j=1}^{J_{it}} E_{ijt}e^{\mu_{ijt}}} \,.$$

Since this is a purely predictive calculation, Y_{ijt} values need not be drawn as part of the MCMC sampling order, but instead at the very end, conditional on the post-convergence samples.

Zhu, Carlin, English, and Scalf (2000) use Figure 11.19 to refine the definition of exposure used in Example 11.6 by subdividing each zip into subregions based on whether or not they are closer than 500 m to a major road. This process involves creating "buffers" around each road and subsequently overlaying them in a GIS, and has been previously used in several studies of vehicle emissions. This definition leads to some urban zips becoming "entirely exposed," as they contain no point further than 500 m from a major road; these are roughly the zips with the darkest shading in Figure 11.17 (i.e., those having traffic densities greater than 10,000 cars per year per km of major roadway). Analogously, many zips in the thinly populated eastern part of the county contained at most one major road, suggestive of little

or no traffic exposure. As a result, we (somewhat arbitrarily) defined those zips in the two lightest shadings (i.e., those having traffic densities less than 2,000 cars per year per km of roadway) as being "entirely unexposed." This typically left slightly less than half the zips (47 for the year shown, 1983) in the middle range, having some exposed and some unexposed subregions, as determined by the intersection of the road proximity buffers. These subregions are apparent as the lightly shaded regions in Figure 11.19; the "entirely exposed" regions continue to be those with the darkest shading, while the "entirely unexposed" regions have no shading.

The fitted rates obtained by Zhu et al. (2000) provide a similar overall impression as those in Figure 11.18, except that the newer map is able to show subtle differences within several "partially exposed" regions. These authors also illustrate the interpolation of missing subregional counts Y_{ijt} using Equation (11.53). Analogous to the block-block FMPC imputation in Subsection 7.2, the sampling-based hierarchical Bayesian method produces more realistic estimates of the subregional hospitalization counts, with associated confidence limits emerging as an automatic byproduct.

11.7.4 Nonnested misalignment and regression

In this subsection we consider spatiotemporal *regression* in the misaligned data setting motivated by our Atlanta ozone data set. Recall that the first component of this data set provides ozone measurements X_{itr} at between 8 and 10 fixed monitoring sites i for day t of year r, where $t = 1, \ldots, 92$ (the summer days from June 1 through August 31) and $r = 1, 2, 3$, corresponding to years 1993, 1994, and 1995 For example, Figure 1.3 shows the 8-hour daily maximum ozone measurements (in parts per million) at the 10 monitoring sites for a particular day (July 15, 1995), along with the boundaries of the 162 zip codes in the Atlanta metropolitan area.

A *second* component of this data set (about which we so far have said far less) provides relevant health outcomes, but only at the zip code level. Specifically, for each zip l, day t, and year r, we have the number of pediatric emergency room (ER) visits for asthma, Y_{ltr}, as well as the total number of pediatric ER visits, n_{ltr}. These data come from a historical records-based investigation of pediatric asthma emergency room visits to seven major emergency care centers in the Atlanta metropolitan statistical area during the same three summers. Our main substantive goal is an investigation of the relationship between ozone and pediatric ER visits for asthma in Atlanta, controlling for a range of sociodemographic covariates. Potential covariates (available only as zip-level summaries in our data set) include average age, percent male, percent black, and percent using Medicaid for payment (a crude surrogate for socioeconomic status). Clearly an investigation of the relationship between ozone exposure and pediatric ER visit count cannot be undertaken until the mismatch in the support of the (point-level) predictor and (zip-level) response variables is resolved.

A naive approach would be to average the ozone measurements belonging to a specific zip code, then relate this average ozone measurement to the pediatric asthma ER visit count in this zip. In fact, there are few monitoring sites relative to the number of zip codes; Figure 1.3 shows most of the zip codes contain no sites at all, so that most of the zip-level ER visit count data would be discarded. An alternative would be to aggregate the ER visits over the entire area and model them as a function of the average of the ozone measurements (that is, eliminate the spatial aspect of the data and fit a temporal-only model). Using this idea in a Poisson regression, we obtained a coefficient for ozone of 2.48 with asymptotic standard error 0.71 (i.e., significant positive effect of high ozone on ER visit rates). While this result is generally consistent with our findings, precise comparison is impossible for a number of reasons. First, this approach requires use of data from the entire Atlanta metro area (due to the widely dispersed locations of the monitoring stations), not data from the city only as our approach allows. Second, it does not permit use of available covariates (such

as race and SES) that were spatially but not temporally resolved in our data set. Third, standardizing using expected counts E_i (as in Equation (11.54) below) must be done only over days (not regions), so the effect of including them is now merely to adjust the model's intercept.

We now describe the disease component of our model, and subsequently assemble the full Bayesian hierarchical modeling specification for our spatially misaligned regression. Similar to the model of Subsection 11.7.3, we assume the zip-level asthma ER visit counts, Y_{ltr} for zip l during day t of summer r, follow a Poisson distribution,

$$Y_{ltr} \sim Poisson\left(E_{ltr} \exp(\lambda_{ltr})\right) , \tag{11.54}$$

where the E_{ltr} are expected asthma visit counts, determined via internal standardization as $E_{ltr} = n_{ltr}(\sum_{ltr} Y_{ltr} / \sum_{ltr} n_{ltr})$, where n_{ltr} is the total number of pediatric ER visits in zip code l on day t of year r. Thus E_{ltr} is the number of pediatric ER asthma visits we would expect from the given zip and day if the proportion of such visits relative to the total pediatric ER visit rate was homogeneous across all zips, days, and years. Hence λ_{ltr} in (11.54) can be interpreted as a log-relative risk of asthma among those children visiting the ER in group ltr. Our study design is thus a *proportional admissions model* (Breslow and Day, 1987, pp. 153–155).

We do not take n_{ltr} equal to the total number of children *residing* in zip l on day t of year r, since this standardization would implicitly presume a constant usage of the ER for pediatric asthma management across all zips, which seems unlikely (children from more affluent zips are more likely to have the help of family doctors or specialists in managing their asthma, and so would not need to rely on the ER; see Congdon and Best, 2000, for a solution to the related problem of adjusting for patient referral practices). Note however that this in turn means that our disease (pediatric asthma visits) is not particularly "rare" relative to the total (all pediatric visits). As such, our use of the Poisson distribution in (11.54) should not be thought of as an approximation to a binomial distribution for a rare event, but merely as a convenient and sensible model for a discrete variable.

For the log-relative risks in group ltr, we begin with the model,

$$\lambda_{ltr} = \beta_0 + \beta_1 X_{l,t-1,r} + \sum_{c=1}^{C} \alpha_c Z_{cl} + \sum_{d=1}^{D} \delta_d W_{dt} + \theta_l \ . \tag{11.55}$$

Here, β_0 is an intercept term, and β_1 denotes the effect of ozone exposure $X_{l,t-1,r}$ in zip l during day $t-1$ of year r. Note that we model pediatric asthma ER visit counts as a function of the ozone level on the *previous* day, in keeping with the most common practice in the epidemiological literature (see, e.g., Tolbert et al., 2000). This facilitates next-day predictions for pediatric ER visits given the current day's ozone level, with our Bayesian approach permitting full posterior inference (e.g., 95% prediction limits). However, it also means we have only $(J-1) \times 3 = 273$ days worth of usable data in our sample. Also, $\mathbf{Z}_l = (Z_{1l}, \ldots, Z_{Cl})^T$ is a vector of C zip-level (but not time-varying) sociodemographic covariates with corresponding coefficient vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_C)^T$, and $\mathbf{W}_t = (W_{1t}, \ldots, W_{Dt})^T$ is a vector of D day-level (but not spatially varying) temporal covariates with corresponding coefficient vector $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_D)^T$. Finally, θ_l is a zip-specific random effect designed to capture extra-Poisson variability in the observed ER visitation rates. These random effects may simply be assumed to be exchangeable draws from a $N(0, 1/\tau)$ distribution (thus modeling overall *heterogeneity*), or may instead be assumed to vary spatially using a conditionally autoregressive (CAR) specification.

Of course, model (11.54)–(11.55) is not fittable as stated, since the zip-level previous-day ozone values $X_{l,t-1,r}$ are not observed. Fortunately, we may use the methods of Section 7.1 to perform the necessary point-block realignment. To connect our Equation (11.55) notation

with that used in Section 7.1, let us write $\mathbf{X}_{B,r} \equiv \{X_{l,t-1,r},\ l = 1, \ldots, L, t = 2, \ldots, J\}$ for the unobserved block-level data from year r, and $\mathbf{X}_{s,r} \equiv \{X_{itr},\ i = 1, \ldots, I, t = 1, \ldots, J\}$ for the observed site-level data from year r. Then, from Equations (11.21) and (11.22) and assuming no missing ozone station data for the moment, we can find the conditional predictive distribution $f(\mathbf{X}_{B,r}|\mathbf{X}_{s,r}, \boldsymbol{\gamma}_r, \sigma_r^2, \boldsymbol{\phi}_r, \boldsymbol{\rho}_r)$ for year r. However, for these data some components of the $\mathbf{X}_{s,r}$ will be missing, and thus replaced with imputed values $\mathbf{X}_{s,r}^{(m)}$, $m = 1, \ldots, M$, for some modest number of imputations M (say, $M = 3$). (In a slight abuse of notation here, we assume that any *observed* component of $\mathbf{X}_{s,r}^{(m)}$ is simply set equal to that observed value for all m.)

Thus, the full Bayesian hierarchical model specification is given by

$$\begin{aligned}[\textstyle\prod_r \prod_t \prod_l f(Y_{ltr}|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\theta}, X_{l,t-1,r})]\, p(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\theta}) \\ \times \Big[\textstyle\prod_r f(\mathbf{X}_{B,r}|\mathbf{X}_{s,r}^{(m)}, \boldsymbol{\gamma}_r, \sigma_r^2, \boldsymbol{\phi}_r, \boldsymbol{\rho}_r) \\ \times f(\mathbf{X}_{s,r}^{(m)}|\boldsymbol{\gamma}_r, \sigma_r^2, \boldsymbol{\phi}_r, \boldsymbol{\rho}_r) p(\boldsymbol{\gamma}_r, \sigma_r^2, \boldsymbol{\phi}_r, \boldsymbol{\rho}_r)\Big] \ ,\end{aligned} \tag{11.56}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$, and $\boldsymbol{\gamma}_r, \sigma_r^2, \boldsymbol{\phi}_r$ and $\boldsymbol{\rho}_r$ are year-specific versions of the parameters in (7.6). Note that there is a posterior distribution for each of the M imputations. Model (11.56) assumes the asthma-ozone relationship does not depend on year; the misalignment parameters are year-specific only to permit year-by-year realignment.

Zhu, Carlin, and Gelfand (2003) offer a reanalysis of the Atlanta ozone and asthma data by fitting a version of model (11.55), namely,

$$\lambda_{ltr} = \beta_0 + \beta_1 X_{l,t-1,r}^{*(m,v)} + \alpha_1 Z_{1l} + \alpha_2 Z_{2l} + \delta_1 W_{1t} + \delta_2 W_{2t} + \delta_3 W_{3t} + \delta_4 W_{4t} \ , \tag{11.57}$$

where $X_{l,t-1,r}^{*(m,v)}$ denotes the (m, v)th imputed value for the zip-level estimate of the 8-hour daily maximum ozone measurement on the previous day $(t-1)$. Our zip-specific covariates are Z_{1l} and Z_{2l}, the percent high socioeconomic status and percent black race of those pediatric asthma ER visitors from zip l, respectively. Of the day-specific covariates, W_{1t} indexes day of summer ($W_{1t} = t$ mod 91) and $W_{2t} = W_{1t}^2$, while W_{3t} and W_{4t} are indicator variables for days in 1994 and 1995, respectively (so that 1993 is taken as the reference year). We include both linear and quadratic terms for day of summer in order to capture the rough U-shape in pediatric ER asthma visits, with June and August higher than July.

The analysis of Zhu et al. (2003) is only approximate, in that they run *separate* MCMC algorithms on the portions of the model corresponding to the two lines of model (11.56). In the spirit of the multiple imputation approach to the missing (point-level) ozone observations, they also retain $V = 3$ post-convergence draws from each of our $M = 3$ imputed data sets, resulting in $MV = 9$ zip-level approximately imputed ozone vectors $\mathbf{X}_{B,r}^{*(m,v)}$.

The results of this approach are shown in Table 11.12. The posterior median of β_1 (.7860) is positive, as expected. An increase of .02 ppm in 8-hour maximum ozone concentration (a relatively modest increase, as seen from Figure 1.3) thus corresponds to a fitted relative risk of $\exp(.7860 \times .02) \approx 1.016$, or a 1.6% increase in relative risk of a pediatric asthma ER visit. However, the 95% credible set for β_1 does include 0, meaning that this positive association between ozone level and ER visits is not "Bayesianly significant" at the 0.05 level. Using a more naive approach but data from all 162 zips in the Atlanta metro area, Carlin et al. (1999) estimate the above relative risk as 1.026, marginally significant at the .05 level (that is, the lower limit of the 95% credible set for β_1 was precisely 0).

Regarding the demographic variables, the effects of both percent high SES and percent black emerge as significantly different from 0. The relative risk for a zip made entirely of high SES residents would be slightly more than half that of a comparable all-low SES zip, while a zip with a 100% black population would have a relative risk nearly 1.8 times that of a 100% nonblack zip. As for the temporal variables, day of summer is significantly negative

Parameter	Effect	Posterior median	95% Posterior credible set	Fitted relative risk
β_0	intercept	–0.4815	(–0.5761, –0.3813)	—
β_1	ozone	0.7860	(–0.7921, 2.3867)	1.016†
α_1	high SES	–0.5754	(–0.9839, –0.1644)	0.562
α_2	black	0.5682	(0.3093, 0.8243)	1.765
δ_1	day	–0.0131	(–0.0190, –0.0078)	—
δ_2	day^2	0.00017	(0.0001, 0.0002)	—
δ_3	year 1994	0.1352	(0.0081, 0.2478)	1.145
δ_4	year 1995	0.4969	(0.3932, 0.5962)	1.644

Table 11.12 *Fitted relative risks for the parameters of interest in the Atlanta pediatric asthma ER visit data, full model. (†This is the posterior median relative risk predicted to arise from a .02 ppm increase in ozone.)*

and its square is significantly positive, confirming the U-shape of asthma relative risks over a given summer. Both year 1994 and year 1995 show higher relative risk compared with year 1993, with estimated increases in relative risk of about 15% and 64%, respectively.

11.8 Areal-level continuous time modeling

Quick, Banerjee and Carlin (2013) address the less common setting where space is discrete and time is continuous. This can be envisioned in situations where a collection of N_s *spatially associated* functions of time over N_s regions are posited. Put another way, functions arising from neighboring regions are believed to resemble each other. The functional data analysis literature (Ramsay and Silverman, 1997, and references therein) deals almost exclusively with kernel smoothers and roughness-penalty type (spline) models; recent discrete-space, continuous time examples using spline-based methods include the works by MacNab and Gustafson (2007) and Ugarte et al. (2010). Baladandayuthapani et al. (2008) consider spatially correlated functional data modeling for point-referenced data by treating space as continuous. Delicado et al. (2010) provide a review of and point out that spatially associated functional modeling of time has received little attention, especially for regionally aggregated data.

Quick et al. (2013) propose a class of Bayesian space-time models based upon a dynamic MRF that evolves continuously over time. This accommodates spatial processes that are posited to be spatially indexed over a geographical map with a well-defined system of neighbors. Rather than modeling time using simple parametric forms, as is often done in longitudinal contexts, these authors employ a stochastic process, enhancing the model's adaptability to the data.

The benefits of using a continuous-time model over a discrete-time model are pronounced when investigators (e.g. public health officials) seek to understand the local effects of temporal impact at a resolution finer than that at which the data were sampled. For instance, despite collecting data monthly, there may be interest in interpolating over a particular week or even at a given day of that month. Dynamic space-time models that treat time discretely can offer statistically legitimate inference only at the level of the data. In additin, the modeling also allows us to subsequently carry out inference on temporal gradients; that is, the rate of change of the underlying process over time (see Chapter 13 for inference on spatial gradients). Quick et al. (2013) show how such inference can be carried out in fully model-based fashion using exact posterior predictive distributions for the gradients at any arbitrary time point.

11.8.1 Areally referenced temporal processes

Here we provide a brief overview of the approach proposed by Quick et al. (2013). Consider a map of a geographical region comprising N_s regions that are delineated by well-defined boundaries, and let $Y_i(t)$ be the outcome arising from region i at time t. For every region i, we believe that $Y_i(t)$ exists, at least conceptually, at every time point. However, the observations are collected not continuously but at discrete time points, say $\mathcal{T} = \{t_1, t_2, \ldots, t_{N_t}\}$. For simplicity, let us assume that the data comes from the same set of time points in $\mathcal{T}$ for each region. This is not strictly necessary for the ensuing development, but will facilitate the notation.

A spatial random effect model that treats space as continuous and time as discrete assumes that

$$Y_i(t) = \mu_i(t) + Z_i(t) + \epsilon_i(t), \quad \epsilon_i(t) \overset{ind}{\sim} N(0, \tau_i^2) \text{ for } i = 1, 2, \ldots, N_s \,, \tag{11.58}$$

where $\mu_i(t)$ captures large scale variation or trends, for example using a regression model, and $Z_i(t)$ is an underlying areally-referenced stochastic process over time that captures smaller-scale variations in the time scale while also accommodating spatial associations. Each region also has its own variance component, τ_i^2, which captures residual variation not captured by the other components.

The process $Z_i(t)$ specifies the probability distribution of correlated space-time random effects while treating space as discrete and time as continuous. We seek a specification that will allow temporal processes from neighboring regions to be more alike than from non-neighbors. As regards spatial associations, we will respect the discreteness inherent in the aggregated outcome. Rather than model an underlying response surface continuously over the region of interest, we want to treat the $Z_i(t)$'s as functions of time that are smoothed across neighbors.

The neighborhood structure arises from a discrete topology comprising a list of neighbors for each region. This is described using an $N_s \times N_s$ adjacency matrix $W = \{w_{ij}\}$, where $w_{ij} = 0$ if regions i and j are not neighbors and $w_{ij} = c \neq 0$ when regions i and j are neighbors, denoted by $i \sim j$. By convention, the diagonal elements of W are all zero. To account for spatial association in the $Z_i(t)$'s, a temporally evolving MRF for the areal units at any arbitrary time point t specifies the full conditional distribution for $Z_i(t)$ as depending only upon the neighbors of region i,

$$p(Z_i(t) \,|\, \{Z_{j \neq i}(t)\}) \sim N\left(\sum_{j \sim i} \alpha \frac{w_{ij}}{w_{i+}} Z_j(t), \frac{\sigma^2}{w_{i+}}\right),$$

where $w_{i+} = \sum_{j \sim i} w_{ij}$, $\sigma^2 > 0$, and α is a propriety parameter described below. This means that the $N_s \times 1$ vector $\mathbf{Z}(t) = (Z_1(t), Z_2(t), \ldots, Z_{N_s}(t))^T$ follows a multivariate normal distribution with zero mean and a precision matrix $\frac{1}{\sigma^2}(D - \alpha W)$, where D is a diagonal matrix with w_{i+} as its i-th diagonal elements. The precision matrix is invertible as long as $\alpha \in (1/\lambda_{(1)}, 1/\lambda_{(n)})$, where $\lambda_{(1)}$ (which can be shown to be negative) and $\lambda_{(n)}$ (which can be shown to be 1) are the smallest (i.e., most negative) and largest eigenvalues of $D^{-1/2} W D^{-1/2}$, respectively, and this yields a proper distribution for $\mathbf{Z}(t)$ at each timepoint t.

The MRF in (11.59) does not allow temporal dependence; the $\mathbf{Z}(t)$'s are independently and identically distributed as $N\left(\mathbf{0}, \sigma^2(D - \alpha W)^{-1}\right)$. We could allow time-varying parameters σ_t^2 and α_t so that $\mathbf{Z}(t) \overset{ind}{\sim} N\left(\mathbf{0}, \sigma_t^2(D - \alpha_t W)^{-1}\right)$ for every t. If time were treated discretely, then we could envision dynamic autoregressive priors for these time-varying parameters, or some transformations thereof. However, there are two reasons why we do not

pursue this further. First, we do not consider time as discrete because that would preclude inference on temporal gradients, which, as we have mentioned, is a major objective here. Second, time-varying hyperparameters, especially the α_t's, in MRF models are usually weakly identified by the data; they permit very little prior-to-posterior learning and often lead to over-parametrized models that impair predictive performance over time.

Quick et al. (2013) prefer to jointly build spatialtemporal associations into the model using a multivariate process specification for $\mathbf{Z}(t)$. A highly flexible and computationally tractable option is to assume that $\mathbf{Z}(t)$ is a zero-centered multivariate Gaussian process, $GP(\mathbf{0}, K_Z(\cdot,\cdot))$, where the matrix-valued covariance function (e.g., "*cross-covariance* matrix function," Cressie, 1993) $K_Z(t,u) = \text{cov}\{\mathbf{Z}(t), \mathbf{Z}(u)\}$ is defined to be the $N_s \times N_s$ matrix with (i,j)-th entry $\text{cov}\{Z_i(t), Z_j(u)\}$ for any $(t,u) \in \Re^+ \times \Re^+$. Thus, for any two positive real numbers t and u, $K_Z(t,u)$ is an $N_s \times N_s$ matrix with (i,j)-th element given by the covariance between $Z_i(t)$ and $Z_j(u)$. These multivariate processes are *stationary* when the covariances are functions of the separation between the time-points, in which case we write $K_Z(t,u) = K_Z(\Delta)$, and *fully symmetric* when $K_Z(t,u) = K_Z(|\Delta|)$, where $\Delta = t - u$.

To ensure valid joint distributions for process realizations, we use a constructive approach similar to that used in *linear models of coregionalization* (LMC) and, more generally, belonging to the class of multivariate latent process models. We assume that $\mathbf{Z}(t)$ arises as a (possibly temporally-varying) linear transformation $\mathbf{Z}(t) = A(t)\mathbf{v}(t)$ of a simpler process $\mathbf{v}(t) = (v_1(t), v_2(t), \ldots, v_{N_s}(t))^T$, where the $v_i(t)$'s are univariate temporal processes, independent of each other, and with unit variances. This differs from the conventional LMC approach based on *spatial* processes, which treats space as continuous. The matrix-valued covariance function for $\mathbf{v}(t)$, say, $K_{\mathbf{v}}(t,u)$, thus has a simple diagonal form and $K_Z(t,u) = A(t)K_{\mathbf{v}}(t,u)A(u)^T$. The dispersion matrix for $\mathbf{Z}$ is $\Sigma_Z = \mathcal{A}\Sigma_{\mathbf{v}}\mathcal{A}^T$, where $\mathcal{A}$ is a block-diagonal matrix with $A(t_j)$'s as blocks, and $\Sigma_{\mathbf{v}}$ is the dispersion matrix constructed from $K_{\mathbf{v}}(t,u)$. Constructing simple valid matrix-valued covariance functions for $\mathbf{v}(t)$ automatically ensures valid probability models for $\mathbf{Z}(t)$. Also note that for $t = u$, $K_{\mathbf{v}}(t,t)$ is the identity matrix so that $K_Z(t,t) = A(t)A(t)^T$ and $A(t)$ is a square-root (e.g. obtained from the triangular Cholesky factorization) of the matrix-valued covariance function at time t.

The above framework subsumes several simpler and more intuitive specifications. One particular specification that we pursue here assumes that each $v_i(t)$ follows a stationary Gaussian Process $GP(0, \rho(\cdot,\cdot;\boldsymbol{\phi}))$, where $\rho(\cdot,\cdot;\phi)$ is a positive definite correlation function parametrized by $\boldsymbol{\phi}$ (e.g. Stein, 1999), so that $\text{cov}(v_i(t), v_i(u)) = \rho(t,u;\phi)$ for every $i = 1, 2, \ldots, N_s$ for all non-negative real numbers t and u. Since the $v_i(t)$ are independent across i, $\text{cov}\{v_i(t), v_j(u)\} = 0$ for $i \neq j$.

The matrix-valued covariance function for $\mathbf{Z}(t)$ becomes $K_Z(t,u) = \rho(t,u;\boldsymbol{\phi})A(t)A(u)^T$. If we further assume that $A(t) = A$ is constant over time, then the process $\mathbf{Z}(t)$ is stationary if and only if $\mathbf{v}(t)$ is stationary. Further, we obtain a *separable* specification, so that $K_Z(t,u) = \rho(t,u;\boldsymbol{\phi})AA^T$. Letting A be some square-root (e.g., Cholesky) of the $N_s \times N_s$ dispersion matrix $\sigma^2(D - \alpha W)^{-1}$ and $R(\boldsymbol{\phi})$ be the $N_t \times N_t$ temporal correlation matrix having (i,j)-th element $\rho(t_i, t_j;\boldsymbol{\phi})$ yields

$$K_Z(t,u) = \sigma^2\rho(t,u;\boldsymbol{\phi})(D - \alpha W)^{-1} \text{ and } \Sigma_Z = R(\boldsymbol{\phi}) \otimes \sigma^2(D - \alpha W)^{-1}\,. \tag{11.59}$$

It is straightforward to show that the marginal distribution from this constructive approach for each $\mathbf{Z}(t_i)$ is $N\left(\mathbf{0}, \sigma^2(D - \alpha W)^{-1}\right)$, the same marginal distribution as the temporally independent MRF specification in (11.59). Therefore, our constructive approach ensures a valid space-time process, where associations in space are modeled discretely using a MRF, and those in time through a continuous Gaussian process.

This separable specification is easily interpretable because it factorizes the dispersion into a spatial association component (areal) and a temporal component. Another significant practical advantage is its computational feasibility. Estimating more general space-time

models usually entails matrix factorizations with $O(N_s^3 N_t^3)$ computational complexity. The separable specification allows us to reduce this complexity substantially by avoiding factorizations of $N_s N_t \times N_s N_t$ matrices. One could design algorithms to work with matrices whose dimension is the smaller of N_s and N_t, thereby accruing massive computational gains. More general models using this approach are introduced and discussed in the online supplement (Quick et al., 2013), but since they do not offer anything new in terms of temporal gradients, we do not pursue them further.

11.8.2 Hierarchical modeling

Following Quick et al. (2013), we build a hierarchical modeling framework using the likelihood from our spatial random effects model in (11.58) and the distributions emerging from the temporal Gaussian process discussed in Section 11.8.1. The mean $\mu_i(t)$ in (11.58) is often indexed by a parameter vector $\boldsymbol{\beta}$, for example a linear regression with regressors indexed by space and time so that $\mu_i(t;\boldsymbol{\beta}) = \mathbf{x}_i(t)^T\boldsymbol{\beta}$.

The posterior distribution is

$$\begin{aligned} p(\boldsymbol{\theta}, \mathbf{Z} \mid \mathbf{Y}) \quad \propto \quad & p(\boldsymbol{\phi}) \times IG(\sigma^2 \mid a_\sigma, b_\sigma) \times \left(\prod_{i=1}^{M} IG(\tau_i^2 \mid a_\tau, b_\tau) \right) \\ & \times N(\boldsymbol{\beta} \mid \mu_\beta, \Sigma_\beta) \times Beta(\alpha \mid a_\alpha, b_\alpha) \\ & \times N(\mathbf{Z} \mid \mathbf{0}, R(\phi) \otimes \sigma^2 (D - \alpha W)^{-1}) \\ & \times \prod_{j=1}^{N_t} \prod_{i=1}^{N_s} N(Y_i(t_j) \mid \mathbf{x}_i(t_j)^T\boldsymbol{\beta} + Z_i(t_j), \tau_i^2) \, , \end{aligned} \tag{11.60}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\phi}, \alpha, \sigma^2, \boldsymbol{\beta}, \tau_1^2, \tau_2^2, \ldots, \tau_{N_s}^2\}$ and $\mathbf{Y}$ is the vector of observed outcomes defined analogous to $\mathbf{Z}$. The parametrizations for the standard densities are as in Carlin and Louis (2008). We assume all the other hyperparameters in (11.60) are known.

Recall the separable matrix-valued covariance function in (11.59). The correlation function $\rho(\cdot;\boldsymbol{\phi})$ determines process smoothness and we choose it to be a fully symmetric Matérn correlation function. Markov chain Monte Carlo (MCMC) can be used to evaluate the joint posterior in (11.60), using Metropolis steps for updating $\boldsymbol{\phi}$ and Gibbs steps for all other parameters; details are available in the supplemental article (Quick et al., 2013). Sampling-based Bayesian inference seamlessly delivers inference on the residual spatial effects. Specifically, if t_0 is an arbitrary unobserved timepoint, then, for any region i, we sample from the posterior predictive distribution

$$p(Z_i(t_0) \mid \mathbf{Y}) = \int p(Z_i(t_0) \mid \mathbf{Z}, \boldsymbol{\theta}) \, p(\boldsymbol{\theta}, \mathbf{Z} \mid \mathbf{Y}) d\boldsymbol{\theta} \, d\mathbf{Z} \, .$$

This is achieved using *composition sampling*: for each sampled value of $\{\boldsymbol{\theta}, \mathbf{Z}\}$, we draw $Z_i(t_0)$, one for one, from $p(Z_i(t_0) \mid \mathbf{Z}, \boldsymbol{\theta})$, which is Gaussian. Also, our sampler easily adapts to situations where $Y_i(t)$ is missing (or not monitored) for some of the time points in region i. We simply treat such variables as missing values and update them, from their associated full conditional distributions, which of course are $N(\mathbf{x}_i(t)^T\boldsymbol{\beta} + Z_i(t), \tau_i^2)$. We assume that all predictors in $\mathbf{x}_i(t)$ will be available in the space-time data matrix, so this temporal interpolation step for missing outcomes is straightforward and inexpensive.

Model checking is facilitated by simulating *independent* replicates for each observed outcome: for each region i and observed timepoint t_j, we sample from $p(Y_{rep,i}(t_j) \mid \mathbf{Y})$, which is equal to

$$\int N(Y_{rep,i}(t_j) \mid \mathbf{x}_i(t_j)^T\boldsymbol{\beta} + Z_i(t_j), \tau_i^2) \, p(\boldsymbol{\beta}, Z_i(t_j), \tau_i^2 \mid \mathbf{Y}) d\boldsymbol{\beta} dZ_i(t_j) d\tau_i^2 \, ,$$

	p_D	DIC*
Simple Linear Regression	79	9,894
Random Intercept and Slope	165	4,347
CAR Model	117	7,302
Areally Referenced Gaussian Process	5,256	0

Table 11.13 *Comparisons between our areally referenced Gaussian process model and the three alternatives. p_D is a measure of model complexity, as it represents the effective number of parameters. Smaller values of DIC indicate a better trade-off between in sample model fit and model complexity. DIC* is standardized relative to the areally referenced Gaussian Process model.*

where $p(\boldsymbol{\beta}, Z_i(t_j), \tau_i^2 \mid \mathbf{Y})$ is the marginal posterior distribution of the unknowns in the likelihood. Sampling from the posterior predictive distribution is straightforward, again, using composition sampling.

Example 11.7 Quick et al. (2013) analyze a dataset consisting of monthly asthma hospitalization rates in the 58 counties of California over an 18-year period. As such, $N_t = 12 * 18 = 216$, and we can simply set $t_j = j = 1, 2, \ldots, N_t$. The covariates in this model include population density, ozone level, the percent of the county under 18, and percent black. Population-based covariates are calculated for each county using the 2000 U.S. Census, so they do not vary temporally. In order to accommodate seasonality in the data, monthly fixed effects are included, using January as a baseline. Thus, after accounting for the monthly fixed effects and the four covariates of interest, $\mathbf{x}_i(t)$ is a 16×1 vector.

We compare the model in (11.29) with three alternative models using the DIC criterion (Spiegelhalter et al. 2002). These models are all still of the form

$$Y_i(t) = \mathbf{x}_i(t)'\boldsymbol{\beta} + Z_i(t) + \epsilon_i(t), \quad \epsilon_i(t) \overset{ind}{\sim} N(0, \tau_i^2) \text{ for } i = 1, 2, \ldots, N_s \,, \tag{11.61}$$

but with different $Z_i(t)$. Our first model is a simple linear regression model which ignores both the spatial and the temporal autocorrelation, i.e., $Z_i(t) = 0 \; \forall \; i, t$. The second model allows for a random intercept and random temporal slope, but ignores the spatial nature of the data, i.e., here $Z_i(t) = \alpha_{0i} + \alpha_{1i}t$, where $\alpha_{ki} \overset{iid}{\sim} N(0, \sigma_k^2)$, for $k = 0, 1$. In this model, to preserve model identifiability, we must remove the global intercept from our design matrix, $\mathbf{x}_i(t)$. Our third model builds upon the second, but introduces spatial autocorrelation by letting $\boldsymbol{\alpha}_k = (\alpha_{k1}, \ldots, \alpha_{kN_s})' \sim CAR(\sigma_k^2), k = 0, 1$. The results of the model comparison can be seen in Table 11.13, which indicates that our Gaussian process model has the lowest DIC value, and is thus the preferred model and the only one we consider henceforth. The surprisingly large p_D for the areally referenced Gaussian process model arises due to the very large size of the dataset (58 counties $\times$ 216 timepoints).

The estimates for our model parameters can be seen in Table 11.14. The coefficients for the monthly covariates indicate decreased hospitalization rates in the summer months, a trend which is consistent with previous findings. The coefficients for population density, percent under 18, and percent black are all significantly positive, also as expected. There is a large range of values for the county-specific residual variance parameters, τ_i^2. Perhaps not surprisingly, the magnitude of these terms seems to be negatively correlated with the population of the given counties, demonstrating the effect a (relatively) small denominator can have when computing and modeling rates. The strong spatial story seen in the maps is reflected by the size of σ^2 compared to the majority of the τ_i^2. There is also relatively strong temporal correlation, with $\phi = 0.9$ corresponding to $\rho(t_i, t_j; \phi) \geq 0.4$ for $|t_j - t_i|$ less than 2 months.

Maps of the yearly (averaged across month) spatiotemporal random effects can be seen in Figure 11.20. Since here we are dealing with the *residual* curve after accounting for a number

Parameter	Median (95% CI)	Parameter	Median (95% CI)
β_0 (Intercept)	9.17 (8.93, 9.42)	β_{11} (August)	-3.58 (-4.02, -3.13)
β_1 (Pop Den)	0.60 (0.49, 0.70)	β_{12} (September)	-1.96 (-2.37, -1.54)
β_3 (% Under 18)	1.24 (1.15, 1.34)	β_{13} (October)	-1.36 (-1.73, -1.00)
β_4 (% Black)	1.12 (1.01, 1.24)	β_{14} (November)	-0.71 (-1.02, -0.42)
β_5 (February)	-0.25 (-0.46, -0.04)	β_{15} (December)	0.63 (0.41, 0.86)
β_6 (March)	-0.21 (-0.48, 0.07)	ϕ	0.90 (0.84, 0.97)
β_7 (April)	-1.47 (-1.81, -1.12)	α	0.77 (0.71, 0.80)
β_8 (May)	-1.17 (-1.53, -0.8)	σ^2	21.52 (20.18, 23.06)
β_9 (June)	-2.79 (-3.21, -2.4)	$\bar{\tau}^2$	3.32 (0.18, 213.16)
β_{10} (July)	-3.78 (-4.21, -3.37)		

Table 11.14 *Parameter estimates for asthma hospitalization data, where estimates for* $\bar{\tau}^2$ *represent the median (95% CI) of the* $\tau_i^2, i = 1, \ldots, N_s = 58$.

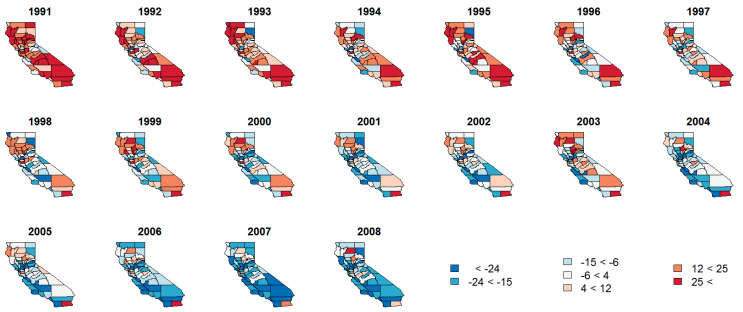

Figure 11.20 *Spatial random effects for asthma hospitalization data, by year*

of mostly non-time-varying covariates, it comes as no surprise that the spatiotemporal random effects capture most of the variability in the model, including the striking decrease in yearly hospitalization rates over the study period. It also appears that our model is providing a better fit to the data in the years surrounding 2000, perhaps indicating that we could improve our fit by allowing our demographic covariates to vary temporally. Our model also appears to be performing well in the central counties, where asthma hospitalization rates remained relatively stable for much of the study period.

11.9 Exercises

1. Suppose $Var(\epsilon(\mathbf{s}, t))$ in (11.6), (11.7), and (11.8) is revised to $\sigma_\epsilon^{2(t)}$.
   (a) Revise expressions (11.11), (11.13), and (11.16), respectively.
   (b) How would these changes affect simulation-based model fitting?
2. The data `www.biostat.umn.edu/~brad/data/ColoradoS-T.dat` contain the maximum monthly temperatures (in tenths of a degree Celcius) for 50 locations over 12 months in 1997. The elevation at each of the 50 sites is also given.
   (a) Treating month as the discrete time unit, temperature as the dependendent variable, and elevation as a covariate, fit the additive space-time model (11.6) to this data. Provide posterior estimates of the important model parameters, and draw image-contour plots for each month.

   (*Hint:* Modify the `WinBUGS` code in Example 6.1 to fit a simple, nested spatiotemporal model. That is, use either the "direct" approach or the `spatial.exp` command to build an exponential kriging model for the data for a given month t with a range

parameter ϕ_t, and then assume these parameters are in turn i.i.d. from (say) a $U(0, 10)$ distribution.)

(b) Compare a few sensible models (changing the prior for the ϕ_t, including/excluding the covariate, etc.) using the DIC tool in `WinBUGS`. How does DIC seem to perform in this setting?

(c) Repeat part (a) assuming the error structures (11.7) and (11.8). Can these models still be fit in `WinBUGS`, or must you now resort to your own `C, Fortran`, or `R` code?

3. Suppose $Y(\mathbf{s}_i, t_j)$, $i = 1, \ldots, n$, $j = 1, \ldots, m$ arise from a mean-zero stationary spatiotemporal process. Let $a_{ii'} = \sum_{j=1}^{m} Y(\mathbf{s}_i, t_j)Y(\mathbf{s}_{i'}, t_j)/m$, let $b_{jj'} = \sum_{i=1}^{n} Y(\mathbf{s}_i, t_j)Y(\mathbf{s}_i, t_{j'})/n$, and let $c_{ii',jj'} = Y(\mathbf{s}_i, t_j)Y(\mathbf{s}_{i'}, t_{j'})$.

(a) Obtain $E(a_{ii'})$, $E(b_{jj'})$, and $E(c_{ii',jj'})$.

(b) Argue that if we plot $c_{ii',jj'}$ versus $a_{ii'} \cdot b_{jj'}$, under a separable covariance structure, we can expect the plotted points to roughly lie along a straight line. (As a result, we might call this a *separability plot*.) What is the slope of this theoretical line?

(c) Create a separability plot for the data in Exercise 2. Was the separability assumption there justified?

4. Consider again the data and model of Example 11.5, the former located at `www.biostat.umn.edu/~brad/data2.html`. Fit the Poisson spatiotemporal disease mapping model (11.44), but where we discard the smoking covariate, and also reverse the gender scores ($s_j = 1$ if male, 0 if female) so that the log-relative risk (11.45) is reparametrized as

$$\mu_{ijkt} = \mu + s_j\alpha + r_k\beta + s_j r_k(\xi - \alpha - \beta) + \gamma t + \phi_{it} \ .$$

Under this model, β now unequivocally captures the difference in log-relative risk between white and nonwhite females.

(a) Use either `WinBUGS` or your own `R, C++`, or `Fortran` code to find point and 95% interval estimates of β. Is there any real difference between the two female groups?

(b) Use either the mapping tool within `WinBUGS` or your own `ArcView` or other GIS code to map the fitted median nonwhite female lung cancer death rates per 1000 population for the years 1968, 1978, and 1988. Interpret your results. Is a temporal trend apparent?

5. In the following, let C_1 be a valid two-dimensional isotropic covariance function and let C_2 be a valid one-dimensional isotropic covariance function. Let $C_A(\mathbf{s}, t) = C_1(\mathbf{s}) + C_2(t)$ and $C_M(\mathbf{s}, t) = C_1(\mathbf{s})C_2(t)$. C_A is referred to as an *additive* (or *linear*) space-time covariance function, while C_M is referred to as a *multiplicative* space-time covariance function.

(a) Why are C_A and C_M valid?

(b) Comment on the behavior of C_A and C_M as $||\mathbf{s} - \mathbf{s}', t - t'|| \to 0$ (local limit), and as $||\mathbf{s} - \mathbf{s}', t - t'|| \to \infty$ (global limit).

6. Suppose we observe a constant mean space-time process, $Y(\mathbf{s}, t)$ at equally spaced time points over a regular lattice. How might we obtain simple sample estimates for the covariance functions, $C(\mathbf{s}, t)$, $C(\mathbf{s})$, and $C(t)$? How might we use these to do some exploratory data analysis with regard to separability of the covariance function?

7. Suppose a simple dynamic model $Y_t(\mathbf{s}) = \gamma Y_{t-1}(\mathbf{s}) + \eta_t(\mathbf{s})$ where $|\gamma| < 1$ and the $\eta_t(\mathbf{s})$ are independent and identically distributed stationary mean 0 Gaussian processes for $\mathbf{s} \in D$. Show that the $Y_t(\mathbf{s})$ has a separable covariance function in space and time. Show that this is not the case if γ depends upon $\mathbf{s}$. More generally, consider the dynamical process with integro-difference equation $Y_t(\mathbf{s}) = \int h(\mathbf{s} - \mathbf{s}')Y_{t-1}(\mathbf{s}')d\mathbf{s}' + \eta_t(\mathbf{s})$ with $\eta(\mathbf{s})$ as above. Show that the process is not separable unless $h(\mathbf{0}) = \gamma \neq 0$ and $h(\mathbf{u}) = 0$ for almost all $\mathbf{u} \neq 0$.