Modeling Dependence in Spatio-Temporal Econometrics

Noel Cressie and Christopher K. Wikle

Abstract This chapter is concerned with lattice data that have a temporal label as well as a spatial label, where these spatio-temporal data appear in the "space-time cube" as a time series of spatial lattice (regular or irregular) processes. The spatio-temporal autoregressive (STAR) models have traditionally been used to model such data but, importantly, one should include a component of variation that models instantaneous spatial dependence as well. That is, the STAR model should include the spatial autoregressive (SAR) model as a subcomponent, for which we give a generic form. Perhaps more importantly, we illustrate how noisy and missing data can be accounted for by using the STAR-like models as process models, alongside a data model and potentially a parameter model, in a hierarchical statistical model (HM).

1 Introduction

Spatial Econometrics has its origins in the statistical modeling of data that are labeled with a spatial (regular or irregular) lattice and, hence, they fall under Tobler's first law of geography (everything is related to everything else, but near things are more related than distant things; Tobler, 1970). Spatial-econometric models were inspired by the autoregressive (AR) statistical models found in time series analysis, where the data are temporally labeled and things in the recent past are more related than things in the distant past. The area of study known as Econometrics has these AR (combined with moving-average) models at its core. Spatial Econometrics has mimicked Econometrics with spatial autoregressive (SAR) models at its core [e.g., Anselin, 1988, Arbia, 2006]. In this chapter, we consider the spatial and the temporal

Noel Cressie
University of Wollongong, Australia, e-mail: ncressie@uow.edu.au

Christopher K. Wikle
University of Missouri, USA, e-mail: wiklec@missouri.edu

aspects together and give spatio-temporal-econometric models based on spatio-temporal autoregressive moving average (STARMA) models that can be fitted to spatio-temporal data.

One might think that SAR models have very similar statistical properties to those of AR models. However, the time dimension is ordered whereas the spatial dimension is not (unless one-dimensional space provides the spatial labels or a partial order is imposed on a high-dimensional space; see e.g., Tjostheim, 1978, Cressie and Davidson, 1998). While the SAR models of Spatial Econometrics can be defined analogously to the (temporal) AR models of Econometrics, some of their spatial-statistical-dependence properties are quite different from those of their temporal counterparts. Furthermore, the notion of filtering out noise due to measurement error, which is common in signal processing, has not been given the emphasis it deserves in Spatial Econometrics. These and other issues will be discussed and extended to spatio-temporal-econometric models.

Consider now data with both a spatial label and a temporal label, where these spatio-temporal data appear in the "space-time cube," as a time series of spatial lattice processes. The spatio-temporal autoregressive (STAR) models have traditionally been used but, importantly, one should include a component of variation that models instantaneous spatial dependence as well. That is, the STAR model should include the SAR model as a sub-component. In this chapter, we give the generic form for such a spatio-temporal model. We also consider the fundamental problem of how to handle measurement error in the data as well as missing data, by introducing a data model along with the STAR model, which defines a hierarchical statistical model (HM).

This chapter is organized as follows. Section 2 motivates why space and time are important factors in any scientific investigation and why modeling statistical dependence is key when making inferences from spatio-temporal data. Section 3 develops the core statistical models of Spatio-Temporal Econometrics. Section 4 looks back at the evolution of Spatial Econometrics and notes how some key advances in spatial-statistical modeling have been slow to take hold. Section 5 returns to the core spatio-temporal econometric models presented in Section 3 and gives a modern, HM approach to modeling spatio-temporal data on regular or irregular spatial lattices. Some general remarks are given in Section 6, and a brief technical appendix concludes the chapter.

2 Spatio-temporal statistics

Spatio-temporal data were essential to the nomadic tribes of early civilizations, who used them to return to seasonal hunting grounds. On a bigger scale, data sets on weather, geology, plants, animals, and indigenous people were collected by early explorers seeking to map and exploit new lands. In a sense, we are all analyzers of spatial and temporal data. As we plan our futures (economically, socially, educationally, etc.), we must take into account the present and seek guidance from

the past. As we look at a map to plan a trip, we are letting its spatial abstraction guide us.

There is an important statistical characteristic of spatio-temporal data that is almost ubiquitous, namely that nearby (in space and time) observations tend to be more alike than those that are far apart. A simple, often-effective forecast of tomorrow's weather is to use today's observed weather. This "persistence" forecast is based on observing large autocorrelations between successive days. Such dependence behavior in "nearby" temporal data is also seen in "nearby" spatial data, such as in studies of the environment. Statistics for spatio-temporal data is challenging due to this dependence in time and space. One fundamental scientific problem that arises is understanding the evolution of spatial processes over time (e.g., the evolution of sea-ice coverage in the Arctic; sea surface temperature and the El Niño phenomenon; and time trends of precipitation in agricultural regions). Proper inference to determine if evolutionary components (natural or anthropogenic) are real requires a spatio-temporal *statistical* methodology.

The scientific method involves observation, inspiration, hypothesis-generation, experimentation (to support or refute the current scientific hypothesis), inference, more inspiration, more hypothesis-generation, and so forth. In a sense, everything begins with observation, but it is quickly apparent to a scientist that unless data are obtained in a more-or-less controlled manner (i.e., according to an experimental design), scientifically defensible inference can be challenging. Understanding the role of dependencies when the data are spatial or temporal or both, provides an important perspective when working with experimental data versus observational data.

It is our belief that statistical models used for describing temporal variability in space should represent the variability dynamically. Models used in Physics, Chemistry, Biology, Economics, etc., do this all the time with difference equations and differential equations to express the dynamical evolutionary mechanisms. Why should this change when the models become statistical? Perhaps it is because there is often an alternative framework, for example a model based on correlations, that describes the spatio-temporal dependence. However, this descriptive approach does not directly involve evolutionary mechanisms and, as a consequence, it can push scientific understanding of the Physics/Chemistry/Biology/Economics/etc. into the background. There is in fact a way to have both, in the form of a scientific-statistical model that recognizes the dynamical scientific aspects of the phenomenon, with its uncertainties expressed through statistical models. Obviously, descriptive (correlational) statistical models have a role to play when little is known about the etiology of the phenomenon; however, when possible, we believe that one should use a dynamical statistical approach to model spatio-temporal data, such as the models given in Section 3.2.

2.1 Uncertainty and Data

Central to the observation, summarization, and inference (including prediction) of spatio-temporal processes are *data*. All data come bundled with error. In particular, along with the obvious errors associated with measuring, manipulating, and archiving, there are other errors, such as discrete spatial and temporal sampling of an inherently continuous system. Consequently, there are always scales of variability that are unresolvable and that will further "contaminate" the data. For example, in Atmospheric Science, this is considered as a form of "turbulence," and it corresponds to the well known aliasing problem in time series analysis [e.g., Chatfield, 1989, p. 126] and the micro-scale component of the "nugget effect" in geostatistics [e.g., Cressie, 1993, p. 59].

Furthermore, spatio-temporal data are rarely sampled at spatial or temporal locations that are optimal for the analysis of a specific scientific problem. For instance, in environmental studies there is often a bias in data-coverage towards areas where population density is large and, within a given area, the coverage is often limited by cost. Thus, the location of a measuring site and its temporal sampling frequency may have very little to do with the underlying scientific mechanisms. A scientific study should include the *design* of data locations and sampling frequencies when framing questions, when choosing statistical-analysis techniques, and when interpreting results. This task is complicated since the data are nearly always statistically dependent in space and time, and hence most of the traditional statistical methods taught in introductory statistics courses (which assume independent and identically distributed, or *iid*, errors) do not apply or have to be modified.

2.2 Uncertainty and Models

Science attempts to *explain* the world in which we live, but that world is very complex. A model is a simplification of some well chosen aspects of the world, where the level of complexity often depends on the question being asked. Pragmatically, the goal of a model is to predict and, at the same time, scientists want to incorporate their understanding of how the world works, into their models. For example, the motion of a pendulum can be modeled using Newton's second law and the simple gravity pendulum that ignores the effect of friction and air resistance. The model predicts future locations of the pendulum quite well, with smaller-order modifications needed when the pendulum is used for precise time-keeping. Models that are scientifically meaningful, that predict well, and that are conceptually simple are generally preferred. However, an injudicious application of Occam's razor (or "the law of parsimony") might elevate simplicity over the other two criteria. For example, a statistical model based on correlational associations might be simpler than a model based on scientific theory.

The way to bridge this divide is to focus on what is more-or-less-certain in the scientific theory, and use *scientific-statistical* relationships to characterize it. In other

words, we suggest that the uncertainties in the models be expressed probabilistically. As the data become more expansive, it is natural that they might suggest a more complex model. Clearly, there is a balance to be struck between too much simplicity, so failing to recognize an important signal in the data, and too much complexity, which results in a non-existent signal being "discovered." The research area known as *model choice* uses various criteria (e.g., AIC, DIC) to achieve this balance [e.g., Wikle et al., 2019, pp. 284-287].

2.3 Conditional probabilities in a hierarchical statistical model (HM)

There is a very general way to express uncertainties coming from different sources, through an approach known as hierarchical statistical modeling. There are data Z that measure Y (with measurement uncertainty), there is the scientific process Y (with less or more uncertainty), and there are parameters θ (unknown, not certain) that control the conditional probability distribution of Z given Y, and the probability distribution of Y. In this chapter, the quantities in which we are interested are random vectors and random variables.

The following conditional probabilities are the basic building blocks of a hierarchical statistical model (HM):

Data model: $[Z|Y,\theta]$

Process model: $[Y|\theta]$

where, using generic random quantities A and B, $[A]$ denotes the marginal distribution of A, $[A,B]$ denotes the joint distribution of A and B, and $[A|B]$ denotes the conditional distribution of A given B. Now the joint distribution of Z and Y can be decomposed as follows. From the equation $[A,B] = [A|B][B]$, we have

$$[Z,Y|\theta] = [Z|Y,\theta][Y|\theta], \tag{1}$$

which is simply a product of the data model and the process model.

In the HM above, it is assumed that θ is fixed (not random), and that all probability distributions are conditional on the fixed values of the parameters. Inference on Y depends on the following distribution (sometimes called the *predictive distribution*), obtained from Bayes' Theorem:

$$[Y|Z,\theta] = \frac{[Z|Y,\theta][Y|\theta]}{[Z|\theta]}, \tag{2}$$

where the normalizing "constant," $[Z|\theta] = \int [Z|Y,\theta][Y|\theta]dY$, ensures that the total probability of the predictive distribution is 1.

What can be done about θ? A Bayesian approach would augment the HM with a *parameter model*, $[\theta]$, which is usually called the *prior*. In this chapter, we want to make an important point, that Spatial (and Spatio-Temporal) Econometrics does *not* need to adopt a Bayesian approach to use Bayes' Theorem, given by (2), and to exploit the power of an HM. Henceforth in this chapter, we adopt a non-Bayesian

approach and assume that θ is fixed but unknown, with some closing remarks about this given in Section 6.

In practice, θ is often specified using an estimate, in which case (2) is replaced with $[Y|Z,\widehat{\theta}]$, where $\widehat{\theta}$ is an estimate of θ (i.e., depends on the data Z). It is also possible that θ is estimated from an independent study or is simply an educated guess. It is this "empirical" step of "plugging in $\widehat{\theta}$" that we shall adopt in this chapter. A fully Bayesian HM can be found in, for example, Wikle et al. [2019, pp. 168-170].

2.4 "Classical" Statistical Modeling

Here we use "classical" as an adjective for both frequentist and Bayesian modeling. The HM introduces data Z, process Y, and parameters θ; however, the "classical" model found in the work of Fisher [e.g., Fisher, 1935] has only data Z and parameters θ, as does the "classical" model of Bayes and many who followed him [e.g., Press, 1989]. Classical frequentists base their inferences on the *likelihood*, $[Z|\theta]$. Classical Bayesians base their inferences on the *posterior distribution*, $[\theta|Z]$, which requires both a likelihood $[Z|\theta]$ and a prior $[\theta]$ to be specified. Both approaches miss the fundamental importance of modeling the *latent process* Y, where the Physics/Chemistry/Biology/Economics/etc. typically resides.

To be sure, Statistics has played and continues to play an important role in Science, but often using simple, introductory-textbook approaches based on correlation and regression. Without Y being made explicit in statistical models, Science has often chosen its own path to statistical inference. Scientists know that parameters θ are important; these might be starting values, or boundary conditions, or diffusion constants, and so forth. In what follows, we give a deliberately simplistic description of how a traditional scientist might use Statistics in her/his research, although we note that in some disciplines this is changing fast. It is our hope that this modern way of building statistical-dependence models will happen in Spatial Econometrics and in Spatio-Temporal Econometrics (presented in Sections 4 and 5 of this chapter).

Scientific experiments produce data Z, and variability in the data is generally recognized by scientists. One approach to support, refine, or refute a scientific theory has been to "smooth" the data first. Consider the smoother f, and write

$$\widetilde{Y} = f(Z).$$

The scientist might then assume that any (random) variability has been removed and that $\widetilde{Y}$ can now be treated as the true process with no uncertainty. A less extreme viewpoint would be to consider that $\widetilde{Y}$ is "close to" the true process Y. In that case, the scientist might fit a model for Y using the "data" $\widetilde{Y}$. If the model for Y is $[Y|\theta_P]$, namely a process model with parameters θ_P that are a subset of θ, the scientist might use classical Statistics to fit $[Y|\theta_P]$ to $\widetilde{Y}$. While the approach just described can be effective when the "signal" is strong, it also has the potential to declare the presence of a signal when it may simply be the result of chance fluctuations.

Given the data are to be smoothed, it should be recognized that they are often a combination of raw observations and algorithmic manipulations. The statistical scientist might write instead,

$$\widetilde{Z} = f(Z), \tag{3}$$

where the notation $\widetilde{Z}$ in (3) is deliberate and suggests an important difference between the two ways to think about $f(Z)$.

An HM can be fitted using the data $\widetilde{Z}$, where the data model, $[\widetilde{Z}|Y, \theta]$, recognizes any remaining uncertainty in $\widetilde{Z}$ after smoothing. Inference on the process Y is based on the predictive distribution obtained from (2):

$$[Y|\widetilde{Z}, \theta] \propto [\widetilde{Z}|Y, \theta][Y|\theta], \tag{4}$$

where "$\propto$" means "is proportional to." By writing the data manipulation and pre-processing according to (3), we have a coherent way to decompose the variability in $\widetilde{Z}$ through (4). (Bayesian statisticians would then specify a prior distribution $[\theta]$, but the ultimate goal of inference on Y and θ remains unchanged.)

While the picture painted above is simplistic, it does illustrate that scientific interest is in Y. If a classical frequentist statistician were to include the scientific model $[Y|\theta]$ in the analysis, it should be done in the calculation of the marginal model,

$$[\widetilde{Z}|\theta] = \int [\widetilde{Z}|Y, \theta][Y|\theta]dY\,.$$

That is, the classical frequentist who bases inference on the likelihood should recognise Y and then integrate it out. However, if there is no such recognition in the first place, the model chosen to be fitted, $[\widetilde{Z}|\theta]$, may be difficult to interpret scientifically or, worse yet, may be inappropriately interpreted.

The classical Bayesian is also compromised; inclusion of the scientific model $[Y|\theta]$ yields the posterior distribution of θ,

$$[\theta|\widetilde{Z}] \propto \int [\widetilde{Z}|Y, \theta][Y|\theta]dY \times [\theta]\,.$$

This has the same potential for misinterpretation, if the Bayesian modeler tries to model directly $[\widetilde{Z}|\theta]$ and uses it in $[\widetilde{Z}|\theta] \times [\theta]$.

Spatial Econometrics has a tradition of fitting data directly to process models, and hence from the HM perspective it leaves the data model out of its formalism. As a result, variability due to measurement error is confounded with process-model error. That is, Spatial Econometrics has traditionally taken the classical-frequentist approach to inference. In the next section, we concentrate on *process models* for processes indexed by both space and time and, in Sections 4 and 5, we return to the HM where the data model is formulated along with the process model (and θ is estimated).

3 Spatio-temporal-econometric modeling

There are a number of ways to express statistically that "things" nearby (in space and time) are more related than distant "things." In this section, we illustrate the fundamental difference between space and time with a simple example, and then we show how dynamical spatio-temporal-econometric models can be built that capture the best features of Spatial Econometrics and multivariate time series analysis. In what follows, we let $Y_t(\mathbf{s})$ denote a random variable at spatial location $\mathbf{s}$ and time t, and then we allow $\mathbf{s}$ and t to vary over a spatio-temporal domain of interest.

3.1 Spatial Description and Temporal Dynamics: A Simple Example

The best way to compare space and time in our statistical context is to consider a simple example, where the spatial domain $D_s \equiv \{s_0, s_0 + \Delta, \ldots, s_0 + 99\Delta\}$ is defined in one dimension, and the temporal domain $D_t \equiv \{0, 1, 2, \ldots\}$ is defined on the nonnegative integers. Then let $\{Y_t(s) : s \in D_s, t \in D_t\}$ be a spatio-temporal process of interest; recall that in the space-time cube, fixing $t = t_0$ yields a spatial process and fixing $s = s_0$ yields a time series.

Define the spatial process at the fixed time point t_0 to be the 100-dimensional vector,

$$\mathbf{Y}_{t_0} \equiv (Y_{t_0}(s_0), \ldots, Y_{t_0}(s_0 + 99\Delta))',$$

and define the time series at fixed spatial location s_0 to be the (different) 100-dimensional vector,

$$\mathbf{Y}(s_0) \equiv (Y_{t_0}(s_0), \ldots, Y_{t_0+99}(s_0))'.$$

For illustrative purposes, the dimension of these vectors were arbitrarily chosen to be 100. By comparing spatial statistical models for $\mathbf{Y}_{t_0}$ and time series models for $\mathbf{Y}(s_0)$, we can see to what extent space is modeled differently from time. Note that we deliberately chose the dimensions of the vectors to be the same to make the comparison easier, but they need not be.

Let us consider the vector $\mathbf{Y}_{t_0}$. A simple departure from independence for a *spatial process* is nearest-neighbor dependence expressed through conditional distributions. Let $\text{Gau}(\mu, \sigma^2)$ denote a Gaussian distribution with mean μ and variance σ^2. Assume, for $i = 1, \ldots, 98$, the Gaussian (conditional) distribution,

$$\begin{aligned} &Y_{t_0}(s_i)|\{Y_{t_0}(s_j) : j = 0, \ldots, 99 \text{ and } j \neq i\} \\ &\quad \sim \text{Gau}((\phi_{t_0}/(1 + \phi_{t_0}^2))\{Y_{t_0}(s_{i-1}) + Y_{t_0}(s_{i+1})\}, \sigma_{t_0}^2/(1 + \phi_{t_0}^2)), \end{aligned} \tag{5}$$

where $s_i \equiv s_0 + i\Delta$; $i = 0, \ldots, 99$. On the edges of the transect, assume

$$\begin{aligned} Y_{t_0}(s_0)|\{Y_{t_0}(s_j) : j = 1, \ldots, 99\} &\sim \text{Gau}(\phi_{t_0} Y_{t_0}(s_1), \sigma_{t_0}^2), \\ Y_{t_0}(s_{99})|\{Y_{t_0}(s_j) : j = 0, \ldots, 98\} &\sim \text{Gau}(\phi_{t_0} Y_{t_0}(s_{98}), \sigma_{t_0}^2). \end{aligned}$$

In (5), assume that the *spatial-dependence parameter*, ϕ_{t_0}, satisfies $|\phi_{t_0}| \leq 1$. Based on these assumptions, it can be shown that $E(\mathbf{Y}_{t_0}) = \mathbf{0}$, and the correlation between nearest neighbors is

$$\text{corr}(Y_{t_0}(s_i), Y_{t_0}(s_{i-1})) = \phi_{t_0}; \quad i = 1, \ldots, 99. \tag{6}$$

The process given by (6) is *descriptive* in that it is given simply in terms of correlation.

Let us now consider the vector $\mathbf{Y}(s_0)$. A simple departure from independence for a *time series* is a first-order autoregressive process. Assume that

$$Y_t(s_0) = \phi(s_0)Y_{t-1}(s_0) + \delta_t; \quad t = t_0 + 1, \ldots, t_0 + 99, \tag{7}$$

where δ_t is independent of $Y_{t-1}(s_0)$, and the elements of $\{\delta_t\}$ are *iid* as $\text{Gau}(0, \sigma^2_\delta(s_0))$, for $t = t_0, t_0 + 1, \ldots, t_0 + 99$. To initialize the process, assume

$$Y_{t_0}(s_0) \sim \text{Gau}(0, \sigma^2_\delta(s_0)/(1 - \phi(s_0)^2)),$$

which is a deliberate choice, as is assuming that the *temporal-dependence parameter* $\phi(s_0)$ satisfies $|\phi(s_0)| < 1$. Based on these assumptions, it can be shown that $E(\mathbf{Y}(s_0)) = \mathbf{0}$, $\text{var}(Y_t(s_0))$ does not depend on t, and the correlation between two adjacent time points is:

$$\text{corr}(Y_{t-1}(s_0), Y_t(s_0)) = \phi(s_0); \quad t = t_0 + 1, \ldots, t_0 + 99. \tag{8}$$

The dependence in the process given by (7) is *dynamical* in that it shows how current values are related mechanistically to past values. More generally, the dependence of current values on past values can be expressed probabilistically, and (7) has an equivalent probabilistic expression in terms of the conditional probability of $Y_t(s_0)$ given past values:

$$Y_t(s_0)|Y_{t-1}(s_0), \ldots, Y_{t_0}(s_0) \sim \text{Gau}(\phi(s_0)Y_{t-1}(s_0), \sigma^2_\delta(s_0)).$$

Such time series models are sometimes referred to as causal.

Let us compare and contrast the spatial process (5) and the time series (7). Both are Gaussian with mean zero. From (6) and (8), we see that if $\phi_{t_0} = \phi(s_0)$, they imply the *same* correlation between adjacent random variables. In fact, because of the Gaussian assumption, if the temporal-dependence and the spatial-dependence parameters are equal, the processes are probabilistically identical! However, the spatial process (5) looks east and west for dependence, in contrast to the time series (7), which is causal and looks to the past. This example has a cautionary aspect. Clearly, a description of the properties of spatial or temporal statistical dependence of the model through just moments or even through joint probability distributions, can completely miss the genesis of the statistical dependence, such as the dynamical structure given by (7).

Now, when it comes to considering space and time together in $\{Y_t(\mathbf{s})\}$, we believe that (whenever possible) the temporal dependence should be expressed dynamically, based on Physical/Chemical/Biological/Economic/etc. considerations, since here the

etiology of the phenomenon is clearest. In a contribution to the Statistics literature that was well ahead of its time, Hotelling [1927] gave various statistical analyses based on dynamical models from stochastic differential equations (albeit only for the temporal dimension).

This *dynamical* approach to spatio-temporal statistical modeling contrasts to that of some others, where time is treated as an extra (although different) dimension. In that case, *descriptive* expressions of spatial dependencies through covariance functions are modified to account for the additional temporal dimension. We call this expression descriptive because usually it is not accompanied by an explanation of why the temporal dependence is present.

3.2 Time series of spatial processes

In Spatio-Temporal Econometrics, a generic spatio-temporal process Y is

$$\{Y_t(\mathbf{s}_i) : i = 1, \ldots, n; t = 0, 1, \ldots\}$$

and, for the moment, we can imagine that $Y_t(\cdot)$ is observed at every one of the n spatial locations for all t. We write the spatial process at time t as the vector,

$$\mathbf{Y}_t \equiv (Y_t(\mathbf{s}_1), \ldots, Y_t(\mathbf{s}_n))'; \quad t = 0, 1, \ldots .$$

Hence, the original spatio-temporal process can be written as the *multivariate time series*,

$$\mathbf{Y}_0, \mathbf{Y}_1, \ldots .$$

In Spatial Econometrics, the spatial statistical modeling of an individual $\mathbf{Y}_t$ has been largely based on SAR models (see below), although CAR models are equally appropriate [e.g., Allcroft and Glasbey, 2003].

The vector notation enables us to express the Markov property for $\{\mathbf{Y}_t\}$ succinctly as,

$$[\mathbf{Y}_t|\mathbf{Y}_0, \ldots, \mathbf{Y}_{t-1}] = [\mathbf{Y}_t|\mathbf{Y}_{t-1}]; \quad t = 1, 2, \ldots .$$

An example of a process satisfying the Markov property is the VAR(1) model of dimension n:

$$\mathbf{Y}_t = \mathbf{M}\mathbf{Y}_{t-1} + \boldsymbol{\eta}_t \tag{9}$$

where, in its full generality, $\mathbf{M}$ has n^2 parameters, and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}} \equiv \text{var}(\boldsymbol{\eta}_t)$ has $O(n^2)$ parameters. However, the spatial context can be used to reduce the number of parameters drastically.

For example, suppose we assume that the (i, j)th entry of $\mathbf{M}$ equals 0, unless $\|\mathbf{s}_i - \mathbf{s}_j\| \leq h$, for a given $h > 0$. Then the current value at $\mathbf{s}_i$ is related to those immediate-past values at $\mathbf{s}_i$ *and nearby values* at $\mathbf{s}_j$ (within a radius of h). Thus, rather than $\mathbf{M}$ being made up of n^2 parameters, the parameter space can be made $O(n)$ by making $\mathbf{M}$ sparse through spatial proximities of the n locations. A similar

modeling strategy that allows further reduction in the size of the parameter space would choose $\mathbf{\Sigma}_\eta$ to be sparse (a geostatistical-type spatial model) or $\mathbf{\Sigma}_\eta^{-1}$ to be sparse (a lattice-type spatial model).

The VAR(1) model is a special case of the spatio-temporal autoregressive moving-average (STARMA) models. It is generally true that for these and other multivariate time series, the number of parameters can be enormous, and an important skill of the modeler is to reduce drastically the size of the parameter space. We believe that this is best achieved through recognizing and preserving any known spatio-temporal interactions in the underlying process $\{Y_t(\mathbf{s})\}$.

3.3 Space-time autoregressive moving-average (STARMA) models

We could look for even more generality than a VAR(1) model in the temporal domain, by assuming higher orders of autoregression as well as a moving-average type of dependence. Define the *spatio-temporal autoregressive moving average (STARMA)* models (Ali, 1979; Pfeifer and Deutsch, 1980; and Cressie, 1993, p. 450) as

$$\mathbf{Y}_t = \sum_{k=0}^{p} \left(\sum_{j=1}^{\lambda_k} f_{kj} \mathbf{U}_{kj} \right) \mathbf{Y}_{t-k} + \sum_{l=0}^{q} \left(\sum_{j=1}^{\mu_l} g_{lj} \mathbf{V}_{lj} \right) \boldsymbol{\omega}_{t-l}\,; \quad t = 0, 1, \ldots,$$

where $\{\mathbf{U}_{kj}\}$ and $\{\mathbf{V}_{lj}\}$ are known weight matrices; p and q are the orders of the autoregressive part and the moving-average part, respectively; $\{f_{kj}\}$ and $\{g_{lj}\}$ are parameters of the model; $\{\boldsymbol{\omega}_t\}$ are *iid* random vectors with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}_\omega$; and the index j is used to denote substructures. These are core models in Spatio-Temporal Econometrics.

Under reparameterization, we obtain

$$\mathbf{Y}_t = \sum_{k=0}^{p} \mathbf{B}_k \mathbf{Y}_{t-k} + \sum_{l=0}^{q} \mathbf{E}_l \boldsymbol{\omega}_{t-l} \tag{10}$$

where, without loss of generality, we henceforth put $\mathbf{\Sigma}_\omega = \sigma_\omega^2 \mathbf{I}$ and, for identifiability reasons, $\mathbf{B}_0$ has zero entries down the diagonal. It is important to note that the index k in (10) starts at $k = 0$; the matrix $\mathbf{B}_0$ models instantaneous spatial dependence in the same way that spatial dependence is modeled in a SAR model. As for the SAR model, we assume that $(\mathbf{I} - \mathbf{B}_0)$ is invertible.

The number of parameters in (10) is still very large. Consider several simple cases. First, $p = 0$ and $q = 0$ results in a time series of purely spatial processes without any temporal dependence linking them:

$$\mathbf{Y}_t = \mathbf{B}_0 \mathbf{Y}_t + \mathbf{E}_0 \boldsymbol{\omega}_t\,; \quad t = 0, 1, \ldots\,.$$

To see this clearly, rewrite the expression above:

$$\mathbf{Y}_t = (\mathbf{I} - \mathbf{B}_0)^{-1}\mathbf{E}_0\boldsymbol{\omega}_t\,; \quad t = 0, 1, \ldots$$

and, since $\{\boldsymbol{\omega}_0, \boldsymbol{\omega}_1, \ldots\}$ are mutually independent, we see that the time series $\{\mathbf{Y}_t\}$ defined just above has no temporal dependence. When $\mathbf{E}_0 = \mathbf{I}$, the multivariate time series consists of *iid* mean-zero SARs.

The second case is $p = 1$ and $q = 0$, and recall that $\mathbf{B}_0$ has all-zero diagonal entries. Then,

$$\mathbf{Y}_t = \mathbf{B}_0\mathbf{Y}_t + \mathbf{B}_1\mathbf{Y}_{t-1} + \mathbf{E}_0\boldsymbol{\omega}_t\,; \quad t = 0, 1, \ldots\,.$$

Given $\mathbf{Y}_{t-1}$, the vector $\mathbf{Y}_t$ has spatial statistical dependence that is expressed in the form of a SAR model. From Cressie (1993, p. 409), a SAR can be written as a CAR, which is a Markov random field with simple *conditional* probability dependencies. The equation just above can be written equivalently as,

$$\mathbf{Y}_t = (\mathbf{I} - \mathbf{B}_0)^{-1}\mathbf{B}_1\mathbf{Y}_{t-1} + (\mathbf{I} - \mathbf{B}_0)^{-1}\mathbf{E}_0\boldsymbol{\omega}_t \equiv \mathbf{M}\mathbf{Y}_{t-1} + \boldsymbol{\eta}_t\,,$$

where $\mathbf{M} \equiv (\mathbf{I}-\mathbf{B}_0)^{-1}\mathbf{B}_1$ and $\{\boldsymbol{\eta}_t\}$ are *iid* with mean zero and $\mathrm{var}(\boldsymbol{\eta}_t) = \boldsymbol{\Sigma}_\eta = \sigma^2_\omega(\mathbf{I}-\mathbf{B}_0)^{-1}\mathbf{E}_0\mathbf{E}_0'(\mathbf{I}-\mathbf{B}_0')^{-1}$. This is a VAR(1) model, and recall that the matrix $\mathbf{B}_0$ represents "instantaneous" spatial dependence. Notice that if we multiply out $(\mathbf{I} - \mathbf{B}_0)^{-1}\mathbf{B}_1$, where $(\mathbf{I} - \mathbf{B}_0)$ is sparse, we obtain a propagator matrix $\mathbf{M} = (\mathbf{I} - \mathbf{B}_0)^{-1}\mathbf{B}_1$ that is generally *not* sparse.

Another way to achieve a VAR(1) model is the third case, $p = 1$, $q = 0$, *and* $\mathbf{B}_0 \equiv \mathbf{0}$. Then,

$$\mathbf{Y}_t = \mathbf{B}_1\mathbf{Y}_{t-1} + \mathbf{E}_0\boldsymbol{\omega}_t\,; \quad t = 0, 1, \ldots\,,$$

which is equivalent to

$$\mathbf{Y}_t = \mathbf{M}\mathbf{Y}_{t-1} + \boldsymbol{\eta}_t\,,$$

where now $\mathbf{M} \equiv \mathbf{B}_1$, and $\{\boldsymbol{\eta}_t\}$ are *iid* with mean zero and $\mathrm{var}(\boldsymbol{\eta}_t) = \boldsymbol{\Sigma}_\eta = \sigma^2_\omega\mathbf{E}_0\mathbf{E}_0'$.

There are clearly a number of different ways to arrive at the same type of model. The difference between them lies in their parameterizations. One way to think of $\mathbf{B}_0$ is that it captures the variability at time steps much smaller than the unit of time specified for the autoregression. Small-temporal-scale dynamics, which may be important and unwise to ignore, are collected together into the matrix $\mathbf{B}_0$ that models *instantaneous spatial dependence* [Cressie, 1993, p. 450; LeSage and Pace, 2009, Section 2.1]. Thus, this instantaneous spatial dependence is in fact an approximation of dynamical structure running at time scales much shorter than the unit of time in the autoregression.

4 Spatial-econometric modeling

We saw in Section 3.1 that a spatial Gaussian process in one-dimensional space that is described through its covariance function, can be probabilistically equivalent to a corresponding temporal process (i.e., a time series) that is modeled dynami-

cally through an autoregressive mechanism. Then in Section 3.3, we generalized the autoregressive model by collecting all the spatial-process values into a vector, resulting in a very flexible class of multivariate dynamical models for spatio-temporal processes.

Spatial Econometrics grew out of seeing how dependence was modeled in time in Econometrics. This was achieved through Box-Jenkins ARIMA modeling [Box and Jenkins, 1970] and the use of "backshift" operators, and then by applying the same idea with "spatial-shift" matrices to generate dependence in space [Paelinck and Klaasen, 1979]. For example, the mean-zero AR(1) model for the time series $\{Y_t\}$ is defined as, $Y_t = \phi Y_{t-1} + \delta_t$, where Y_{t-1} is independent of δ_t, and $\{\delta_1, \delta_2, \ldots\}$ are *iid* with $E(\delta_t) = 0$ and $\text{var}(\delta_t) = \sigma_\delta^2$ (see eq. (7)). This equation can be written equivalently in terms of the backshift operator B as:

$$Y_t = \phi B Y_t + \delta_t \,. \tag{11}$$

At the core of Spatial Econometrics are models for $\mathbf{Y} \equiv (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))'$ that mechanistically connect $Y(\mathbf{s}_i)$ to its "neighbors": Replace ϕB in (11) with the square-matrix operator $\mathbf{B}_0$ whose diagonal elements are defined to be zero, and any off-diagonal element that is zero indicates a lack of spatial "connection" between the two corresponding locations. The resulting SAR model is,

$$\mathbf{Y} = \mathbf{B}_0 \mathbf{Y} + \boldsymbol{\omega}\,, \tag{12}$$

where $E(\boldsymbol{\omega}) = \mathbf{0}$ and $\text{var}(\boldsymbol{\omega}) = \sigma_\omega^2 \mathbf{I}$, which was introduced in Section 3.3 as a way to capture instantaneous spatial dependence in a mean-zero spatio-temporal process.

If we write $\mathbf{B}_0 = \phi \mathbf{B}$, where $\mathbf{B}$ is the square matrix (b_{ij}), then the generalization from "time" in (11) to "space" in (12) looks beguilingly straightforward. However, these are mathematical relationships, and nothing has been said yet about the statistical dependence between $\mathbf{B}_0\mathbf{Y}$ and $\boldsymbol{\omega}$ in (12). Recall that in the AR(1) process given by (11), $Y_{t-1} (= BY_t)$ and δ_t are independent. In the SAR process given by (12), $\mathbf{B}_0\mathbf{Y} = \mathbf{B}_0(\mathbf{I} - \mathbf{B}_0)^{-1}\boldsymbol{\omega}$, and hence $\text{cov}(\mathbf{B}_0\mathbf{Y}, \boldsymbol{\omega}) = \mathbf{B}_0(\mathbf{I} - \mathbf{B}_0)^{-1}\text{var}(\boldsymbol{\omega}) = \sigma_\omega^2 \mathbf{B}_0(\mathbf{I} - \mathbf{B}_0)^{-1}$, which shows that $\mathbf{B}_0\mathbf{Y}$ and $\boldsymbol{\omega}$ are statistically dependent.

This latter property means one has to be very careful when interpreting the SAR model. It has been misinterpreted as being causal in Spatial Econometrics; Gibbons and Overman [2012] address this mistake directly, and the presence of non-zero covariances between the autoregressive part, $\mathbf{B}_0\mathbf{Y}$, and the error, $\boldsymbol{\omega}$, is a manifestation of the fundamentally different structure of the SAR model and the AR model, which *is* causal.

For $\mathbf{B}_0 = \phi\mathbf{B}$, (12) can be written as

$$Y(\mathbf{s}_i) = \phi \sum_{j=1}^{n} b_{ij} Y(\mathbf{s}_j) + \omega(\mathbf{s}_i)\,; \quad i = 1, \ldots, n\,,$$

where recall $b_{ii} = 0$. In a naive cross-validation exercise, $Y(\mathbf{s}_i)$ would be deleted and then predicted with $\widehat{Y}(\mathbf{s}_i) \equiv \phi \sum_{j=1}^{n} b_{ij} Y(\mathbf{s}_j)$; then $\widehat{Y}(\mathbf{s}_i)$ would be compared to

$Y(\mathbf{s}_i)$ via, say, $(\widehat{Y}(\mathbf{s}_i) - Y(\mathbf{s}_i))^2$. However, this $\widehat{Y}(\mathbf{s}_i)$ is an inferior predictor of $Y(\mathbf{s}_i)$, since the optimal cross-validation predictor of $Y(\mathbf{s}_i)$ is,

$$Y^*(\mathbf{s}_i) \equiv E(Y(\mathbf{s}_i)|\mathbf{Y}_{-i}),$$

for $\mathbf{Y}_{-i}$ the $(n-1)$-dimensional vector with $Y(\mathbf{s}_i)$ removed from $\mathbf{Y}$. From the Lemma given in the Appendix, $Y^*(\mathbf{s}_i)$ can be derived analytically from the full $n \times n$ covariance matrix, $\text{var}(\mathbf{Y}) = \sigma^2_\omega\{(\mathbf{I} - \phi\mathbf{B})(\mathbf{I} - \phi\mathbf{B}')\}^{-1}$, and it is different from $\widehat{Y}(\mathbf{s}_i)$.

Note that while a derivation of $Y^*(\mathbf{s}_i)$, albeit straightforward and resulting in a closed-form expression, is necessary for the SAR model, $Y^*(\mathbf{s}_i)$ is immediately available from the CAR (conditional autoregressive) model, although this model is used much less frequently in Spatial Econometrics. (For readers interested in the relationships between SAR and CAR models, see Cressie, 1993, p. 408-410, and Ver Hoef et al., 2018.)

Another caution with the use of SAR models in Spatial Econometrics comes with how they are specified when the spatial process $\mathbf{Y}$ does *not* have mean zero. One should take guidance from how the time series model (11) would be modified to handle, say, the regression, $E(Y_t) = \mathbf{x}_t'\boldsymbol{\beta}$. The time series model,

$$Y_t - \mathbf{x}_t'\boldsymbol{\beta} = \phi \cdot (Y_{t-1} - \mathbf{x}_{t-1}'\boldsymbol{\beta}) + \delta_t, \tag{13}$$

is an AR(1) process that preserves the mean structure, $E(Y_t) = \mathbf{x}_t'\boldsymbol{\beta}$. For reasons that are not clear, the Spatial-Econometrics literature (e.g., Anselin, 1988) shows a preference to include the regression term, $\mathbf{X}\boldsymbol{\beta}$, and the spatial-dependence operator $\mathbf{B}_0$ in its core model as follows:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{B}_0\mathbf{Y} + \boldsymbol{\omega}, \tag{14}$$

where $\boldsymbol{\omega} \equiv (\omega(\mathbf{s}_1), \ldots, \omega(\mathbf{s}_n))'$ represents model error with $E(\boldsymbol{\omega}) = \mathbf{0}$.

As a consequence of (14), $E(\mathbf{Y}) = (\mathbf{I}-\mathbf{B}_0)^{-1}\mathbf{X}\boldsymbol{\beta}$, which results in the confounding of large-scale regression effects $\boldsymbol{\beta}$ with small-scale spatial-dependence effects $\mathbf{B}_0$. This can be avoided by taking a cue from the time series model (13). That is, to generalize the SAR model to include regression, we write

$$(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{B}_0(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\omega}.$$

Now $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ and $\text{var}(\mathbf{Y}) = \sigma^2_\omega\{(\mathbf{I} - \mathbf{B}_0)(\mathbf{I} - \mathbf{B}_0')\}^{-1}$, and hence $\boldsymbol{\beta}$ appears only in $E(\mathbf{Y})$, and $\mathbf{B}_0$ appears only in $\text{var}(\mathbf{Y})$. There is an equivalent way to write this model, namely

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}, \text{ and } \mathbf{U} = \mathbf{B}_0\mathbf{U} + \boldsymbol{\omega}, \tag{15}$$

which does appear in the more recent Spatial-Econometrics literature and is called a spatial error model [e.g., LeSage and Pace, 2009, Section 2.3]. Our point is that software based on the model (14) should not be used when fitting spatial statistical models to spatial data with covariates $\mathbf{X}$, due to the confounding of large-scale and small-scale effects and the consequent misinterpretation of a fitted model (14).

More generally, confounding between fixed effects and spatial random effects has become an important topic in the spatial-statistics literature [e.g., Reich et al., 2006, Paciorek, 2010, Hodges and Reich, 2010, Hughes and Haran, 2013, Hanks et al., 2015]. There is still some uncertainty as to what extent these models are able to account for confounding; appropriate mitigation approaches depend on the underlying dependence structure of the random effects, the extent to which covariates are known, and the spatial support [Hanks et al., 2015]. Geographers and spatial econometricians have been aware of spatial confounding for some time in the context of areal data, and they have provided "Moran's I" eigenvector approaches that make the spatial random effects orthogonal to the fixed effects [e.g., Griffith, 2000, 2003]. Spatial statisticians have also considered Moran's I basis functions and extensions in this context [Hughes and Haran, 2013, Bradley et al., 2015]. However, it is unclear how to force random effects to be in the space orthogonal to the fixed effects if the fixed effects have continuous support as they do in geostatistical models [Hanks et al., 2015]. More recently, Bradley et al. [2020] considered confounding between the spatial process and the error process and showed that accounting for dependence between these two processes can improve prediction accuracy.

In Section 2.3, we made the point that observations (Z) on a process are different from the values of the process itself (Y). This is typically due to measurement error ("noisiness"), and it can also be due to gaps in the observations ("missingness"). This can be captured in a spatial-statistical model by writing,

$$Z(\mathbf{s}_i) = Y(\mathbf{s}_i) + \varepsilon(\mathbf{s}_i)\,; \quad \mathbf{s}_i \in D^* \subset \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}\,. \tag{16}$$

In (16), $Z(\mathbf{s}_i)$ is an observation at spatial location $\mathbf{s}_i$ in D^*; locations not in D^* are considered as missing; and $\varepsilon(\cdot)$ is an independent measurement-error process with $\text{var}(\varepsilon(\mathbf{s}_i)) = \sigma^2_\varepsilon > 0$. Goulard et al. [2017] consider spatial-econometric models for missing data, but they do not recognize that the measurement-error component of variation $\varepsilon(\cdot)$ is different from the model-error component of variation $\omega(\cdot)$.

In the general case of non-zero mean due to regression effects, (15) is the *process model* that represents all components of $\mathbf{Y} \equiv (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))'$, even though some might not be observed, and (16) is the *data model* for data $\{Z(\mathbf{s}_i) : \mathbf{s}_i \in D^*\}$ that are observed. That is, in terms of the HM presented in Section 2.3, (16) defines $[Z|Y]$ and (15) defines $[Y]$, where dependence of these two models on parameters $\theta \equiv \{\boldsymbol{\beta}, \sigma^2_\omega, \sigma^2_\varepsilon\}$ has been dropped from the notation for ease of exposition. Specifically, the HM is:

Data model: $Z(\mathbf{s}_i)|Y(\mathbf{s}_i) \sim \text{Gau}(Y(\mathbf{s}_i), \sigma^2_\varepsilon)$, and define $\mathbf{Z} \equiv (Z(\mathbf{s}) : \mathbf{s} \in D^*)'$.

Process model: $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{B}_0(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\omega}$, where $\boldsymbol{\omega} \sim \text{Gau}(\mathbf{0}, \sigma^2_\omega \mathbf{I})$.

The data model and the process model together allow calculation of the predictive distribution: Since $[\mathbf{Z}|\mathbf{Y}]$ is Gaussian and $[\mathbf{Y}]$ is Gaussian, so too is the joint distribution $[\mathbf{Y}, \mathbf{Z}]$ the marginal distribution $[\mathbf{Z}]$, and most importantly the predictive distribution $[\mathbf{Y}|\mathbf{Z}]$. Hence, the key calculations for inference on $\mathbf{Y}$ from the "imperfect" data $\mathbf{Z}$ are the conditional moments,

$$E(Y(\mathbf{s}_i)|\mathbf{Z})\,,\ \text{var}(Y(\mathbf{s}_i)|\mathbf{Z})\,,\ \text{and}\ \text{cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_j)|\mathbf{Z})\,,\ \text{for}\ i, j = 1, \ldots, n\,,$$

and recall that there are locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \backslash D^*$ at which there are no observations. These are known in closed form from Bayes' Theorem given by (2), where the distributions in the numerator of (2) are obtained from (15) and (16), and likelihood-based estimates of θ are used in place of θ. No time-consuming iterative algorithms are needed to calculate them; see the Lemma in the Appendix. The one bottleneck may be fast computation of $\text{var}(\mathbf{Z})^{-1}$ when the data set $\mathbf{Z}$ is very large; see Burden et al. [2015] for a reduced-rank approach to this problem and a comparison to the Spatial-Econometrics literature where fast computation of $\text{var}(\mathbf{Y})^{-1}$ is the focus.

The lessons learned from this section are first to *de-trend* the spatio-temporal data using covariates and then to *use HMs* to capture the imperfections of noisy and missing data. The next section will apply these lessons to the spatio-temporal setting given in Section 3.

5 Modern spatio-temporal-econometric hierarchical models

All the ideas and methodology that are needed have been presented in the preceding sections. It is simply a matter of tying them together now in a series of steps that bears a resemblance to pseudocode for algorithmic development.

Recall that the generic spatio-temporal data are Z, the generic underlying process being measured is Y, which represents the whole process $\{Y_t(\mathbf{s})\}$, and the generic parameters are θ. Due to incomplete data ("missingness"), Z will be of smaller dimension than Y, and the presence of measurement error ("noise") results in the conditional distribution,

$$Z_t(\mathbf{s})|Y, \sigma^2_\varepsilon \sim \text{Gau}(Y_t(\mathbf{s}), \sigma^2_\varepsilon),$$

provided an observation occurs at location $\mathbf{s}$ and time t in the spatio-temporal domain of interest $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \times \{0, 1, \ldots, T\}$.

The building blocks of dynamical models in Spatio-Temporal Econometrics are given below in a sequence of eight steps:

1. $[Z|Y, \theta] = \prod_{D^*}[Z_t(\mathbf{s})|Y, \sigma^2_\varepsilon]$, for D^* the set of all *spatio-temporal data locations*, is Gaussian.
2. $[Y|\theta]$ is a (high-dimensional) Gaussian distribution; see, for example, (10) or its modification that includes regression:

$$(\mathbf{Y}_t - \mathbf{X}_t\boldsymbol{\beta}) = \sum_{k=0}^{p} \mathbf{B}_k(\mathbf{Y}_{t-k} - \mathbf{X}_{t-k}\boldsymbol{\beta}) + \sum_{l=0}^{q} \mathbf{E}_l\boldsymbol{\omega}_{t-l},$$

   for $t = p, p+1, \ldots, T$.
3. $[Z, Y|\theta] = [Z|Y, \theta][Y|\theta]$ is Gaussian (since 1. and 2. are Gaussian).
4. $L(\theta) \equiv [Z|\theta] = \int [Z, Y|\theta]\, dY$; θ includes σ^2_ε, $\boldsymbol{\beta}$, the spatio-temporal-variation parameters in $\{\mathbf{B}_k\}$ and $\{\mathbf{E}_l\}$, and $\{\text{var}(\boldsymbol{\omega}_{t-l})\}$. Recall that $[Z|\theta]$ is Gaussian.

5. Estimate θ with $\hat{\theta} = \arg\sup_{\theta} L(\theta)$, the maximum likelihood estimator.
6. $[Y|Z,\hat{\theta}] = [Z|Y,\hat{\theta}][Y|\hat{\theta}]/[Z|\hat{\theta}]$ is a Gaussian distribution called the (empirical) predictive distribution.
7. $E(Y|Z,\hat{\theta})$ and $\text{var}(Y|Z,\hat{\theta})$ characterize the predictive distribution; both can be calculated straightforwardly in closed form, using the Lemma in the Appendix.
8. Estimation and prediction: Report and interpret $\hat{\theta}$ and its uncertainties (estimation). Make a choropleth map of $E(Y|Z,\hat{\theta})$, which is the HM's spatio-temporal predictor of Y (prediction). Make a second choropleth map of $(\text{diag}(\text{var}(Y|Z,\hat{\theta})))^{1/2}$, which uses the HM to quantify the uncertainty in the first map.

These are the basic steps taken to fit the dynamical spatio-temporal models given in Chapter 5 of Wikle et al. [2019]: There, Sections 5.2 and 5.3 are the most relevant to the development given in this chapter.

6 Concluding remarks

We would like to expression our best wishes to Christine (Thomas-Agnan) on the occasion of her 65-th birthday. She has been a gracious host and an engaging co-author during several long-terms visits by the first author to Université Toulouse 1 Capitole.

Our approach to the problem of "scientific understanding in the presence of uncertainty" takes a probabilistic viewpoint, which allows us to build useful spatio-temporal statistical models and make scientific inferences for various spatial and temporal scales. Accounting for the uncertainty enables us to look for possible associations within and between variables in the underlying scientific process, with the potential for finding mechanisms that extend, modify, or even disprove a scientific theory. The dynamical spatio-temporal-econometric models described in this chapter are an important subset of a much larger class of dynamical HMs for the twenty-first century [Wikle et al., 2019]. We have concentrated on HMs where the parameters θ are estimated from the data, which are called empirical HMs. Bayesian HMs arise when a prior, $[\theta]$, is assigned to the unknown parameters θ. In many cases, the predictive moments, $E(Y|Z)$ and $\text{var}(Y|Z)$, from the Bayesian HM are not available in closed form. Then sampling from the predictive distribution, $[Y|Z]$, is a way to solve this problem (e.g., using MCMC).

There are many challenges associated with building HMs and then carrying out valid inferences. A broad perspective is that there is subjectivity involved with the specification of *all* model components, specifically here the data model and the process model. However, it is not always clear what "subjective" means in this context. For example, it might be "subjective" to use deterministic relationships to motivate a stochastic model, such as for tropical winds [e.g., Wikle et al., 2001], yet the science upon which such a model is based comes from Newton's laws of motion. Thus, we believe that it is not helpful to try to classify probability distributions that determine the statistical model, as subjective or objective. It would be better to ask

about the sensitivity of inferences to model choices and whether such choices make sense scientifically.

Given that a modeler brings so much information to the table when developing models, the conditional-probability framework presented earlier can be used to recognize that this information, say I, is part of what is involved in the conditioning. For the HM, we have

$$[Y|Z,\theta,I] \propto [Z|Y,\theta,I][Y|\theta,I]\,.$$

A major challenge in this paradigm is, to the extent possible, acknowledgement of the importance of this information, I. It is often the case that a team of researchers at the table has a collective "I" that is better quantified and more appropriate than any individual's "I."

In the HM approach, there are certainly cases where models have to be simplified due to practical concerns. Perhaps the computational issues in a given formulation are limiting, which usually leads to a modification of the model. Such practical concerns apply to all statistical inferences in complicated modeling scenarios. This tension between the model you want and the model with which you can compute is healthy, and in modern statistical computing it has led to algorithms that only *approximate* valid inferences. However, user beware! Approximations to approximations can lead to a serious propagation of errors.

Data hold so much potential, but unless they can be organized into a database they are an entropic collection of digits or bits. With the ability in a database to structure, search, filter, query, visualize, and summarize, the data begin to contain *information*. Some of this information comes from judicious use of statistics (i.e., summaries). Then, in going from information to *knowledge*, Science (and, with it, Statistical Science) takes over. Statistical Science makes contributions at all levels of the data-information-knowledge pyramid, but it has often stopped short of the summit where knowledge is used to determine policy. At the interface between Science, Statistics, and Policy, there is an enormous need for decision-making in the presence of uncertainty.

Finally, it is the responsibility of the research team to temper the tendency to fit ever-more-complicated models, and to use model-selection criteria (e.g., AIC, BIC, DIC, etc.) that concentrate on the twin pillars of predictability and parsimony [e.g., Spiegelhalter et al., 2002, Wikle et al., 2019]. But these criteria do not address the third pillar, namely scientific interpretability (i.e., knowledge). Our approach to spatio-temporal-econometric modeling is to use the hierarchical-modeling paradigm and, where possible, choose statistical models based on this third pillar, while not ignoring the other two.

Acknowledgements

Material from Chapters 1, 2, and 6 of Cressie and Wikle [2011] is used with permission from the publisher: Copyright ©2011 by John Wiley & Sons, Inc. All rights reserved. Cressie's research was supported by Australian Research Council Discovery Project DP190100180. Wikle's research was supported by U.S. National Science Foundation grants SES-1853096 and DMS-1811745.

Appendix

Throughout the chapter, we have referred to the predictive distribution $[\mathbf{Y}|\mathbf{Z}]$ that arises from a joint Gaussian distribution, $[\mathbf{Y}, \mathbf{Z}]$. Specifically, we have claimed that $[\mathbf{Y}|\mathbf{Z}]$ is Gaussian and the first two moments can be obtained analytically without resort to iteration, simulation, or approximation. This claim is due to the following lemma from multivariate analysis [e.g., Rencher and Christensen, 2012, p. 97].

Lemma:
Consider the Gaussian random vector, $\mathbf{U} \equiv (\mathbf{U}_1', \mathbf{U}_2')'$, and its first two moments:

$$E(\mathbf{U}) \equiv \boldsymbol{\mu} \equiv (\boldsymbol{\mu}_1', \boldsymbol{\mu}_2')', \text{ and } \text{var}(\mathbf{U}) \equiv \boldsymbol{\Sigma} \equiv \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

Then the conditional distribution, $[\mathbf{U}_1|\mathbf{U}_2]$ is also Gaussian with mean vector,

$$E(\mathbf{U}_1|\mathbf{U}_2) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{U}_2 - \boldsymbol{\mu}_2)$$

and variance-covariance matrix,

$$\text{var}(\mathbf{U}_1|\mathbf{U}_2) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\,.$$

References

D. J. Allcroft and C. A. Glasbey. A latent Gaussian Markov random-field model for spatio-temporal rainfall disaggregation. *Journal of the Royal Statistical Society, Series C*, 52:487–498, 2003.

L. Anselin. *Spatial Econometrics: Methods and Models*. Kluwer, Dordrecht, NL, 1988.

G. Arbia. *Spatial Econometrics: Statistical Foundations and Applications to Regional Convergence*. Springer, Berlin, DE, 2006.

G. E. P. Box and G. M. Jenkins. *Time Series Analysis, Forecasting, and Control*. Holden-Day, Oakland, CA, 1970.

J. R. Bradley, S. H. Holan, and C. K. Wikle. Multivariate spatio-temporal models for high-dimensional areal data with application to longitudinal employer-household dynamics. *Annals of Applied Statistics*, 9:1761–1791, 2015.

J. R. Bradley, C. K. Wikle, and S. H. Holan. Hierarchical models for spatial data with errors that are correlated with the latent process. *Statistica Sinica*, 30:81–109, 2020.

S. Burden, N. Cressie, and D. G. Steel. The SAR model for very large datasets: A reduced-rank approach. *Econometrics*, 3:317–338, 2015.

C. Chatfield. *The Analysis of Time Series: An Introduction, 4th edn*. Chapman and Hall, London, UK, 1989.

N. Cressie. *Statistics for Spatial Data, rev. edn*. John Wiley & Sons, New York, NY, 1993.

N. Cressie and J. L. Davidson. Image analysis with partially ordered Markov models. *Computational Statistics and Data Analysis*, 29:1–26, 1998.

N. Cressie and C. K. Wikle. *Statistics for Spatio-Temporal Data*. John Wiley & Sons, Hoboken, NJ, 2011.

R. A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, UK, 1935.

S. Gibbons and H. G. Overman. Mostly pointless spatial econometrics. *Journal of Regional Science*, 52:172–191, 2012.

M. Goulard, T. Laurent, and C. Thomas-Agnan. About predictions in spatial autoregressive models: optimal and almost-optimal strategies. *Spatial Economic Analysis*, 12:304–325, 2017.

D. A. Griffith. A linear regression solution to the spatial autocorrelation problem. *Journal of Geographical Systems*, 2(2):141–156, 2000.

D. A. Griffith. *Spatial Autocorrelation and Spatial Filtering: Gaining Understanding Through Theory and Scientific Visualization*. Springer-Verlag, Berlin, DE, 2003.

E. M. Hanks, E. M. Schliep, M. B. Hooten, and J. A. Hoeting. Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification. *Environmetrics*, 26:243–254, 2015.

J. S. Hodges and B. J. Reich. Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, 64:325–334, 2010.

H. Hotelling. Differential equations subject to error and population estimates. *Journal of the American Statistical Association*, 22:283–314, 1927.

J. Hughes and M. Haran. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society, Series B*, 75:139–159, 2013.

J. LeSage and R. L. Pace. *Introduction to Spatial Econometrics*. Chapman Hall/CRC Press, Boca Raton, FL, 2009.

C. J. Paciorek. The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statistical Science*, 25:107, 2010.

J. H. P. Paelinck and L. H. Klaasen. *Spatial Econometrics*. Saxon House, Farnborough, UK, 1979.

S. J. Press. *Bayesian Statistics: Principles, Models, and Applications*. John Wiley & Sons, New York, NY, 1989.

B. J. Reich, J. S. Hodges, and V. Zadnik. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, 62:1197–1206, 2006.

A. C. Rencher and W. F. Christensen. *Methods of Multivariate Analysis, 3rd edn.* John Wiley & Sons, Hoboken, NJ, 2012.

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

D. Tjostheim. Statistical spatial series modelling. *Advances in Applied Probability*, 10:130–154, 1978.

W. R. Tobler. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46:234–240, 1970.

J. Ver Hoef, E. Hanks, and M. Hooten. On the relationship between conditional (CAR) and simultaneous (SAR) autoregressive models. *Spatial Statistics*, 25: 68–85, 2018.

C. K. Wikle, R. F. Milliff, D. Nychka, and L. M. Berliner. Spatiotemporal hierarchical Bayesian modeling: Tropical ocean surface winds. *Journal of the American Statistical Association*, 96:382–397, 2001.

C. K. Wikle, A. Zammit-Mangion, and N. Cressie. *Spatio-Temporal Statistics with R*. Chapman & Hall/CRC Press, Boca Raton, FL, 2019.