Spatial Statistics

Noel Cressie   Matthew T. Moores

Centre for Environmental Informatics
National Institute for Applied Statistics Research Australia (NIASRA)
School of Mathematics & Applied Statistics
University of Wollongong, NSW 2522, Australia

May 18, 2021

Abstract

Spatial statistics is an area of study devoted to the statistical analysis of data that have a spatial label associated with them. Geographers often refer to the "location information" associated with the "attribute information," whose study defines a research area called "spatial analysis." Many of the ways to manipulate spatial data are driven by algorithms with no uncertainty quantification associated with them. When a spatial analysis is statistical, that is, it incorporates uncertainty quantification, it falls in the research area called spatial statistics. The primary feature of spatial statistical models is that nearby attribute values are more statistically dependent than distant attribute values; this is a paraphrasing of what is sometimes called the First Law of Geography (Tobler, 1970).

1   Introduction

Spatial statistics provides a probabilistic framework for giving answers to those scientific questions where spatial-location information is present in the data, and that information is relevant to the questions being asked. The role of probability theory in (spatial) statistics is to model the uncertainty, both in the scientific theory behind the question, and in the (spatial) data coming from measurements of the (spatial) process that is a representation of the scientific theory.

In spatial statistics, uncertainty in the scientific theory is expressed probabilistically through a spatial stochastic process, which can be written most generally as:

$$\{Y(\mathbf{s}) : \mathbf{s} \in \mathscr{D}\}, \tag{1}$$

where $Y(\mathbf{s})$ is the random attribute value at location $\mathbf{s}$, and $\mathscr{D}$ is a subset of a d-dimensional space, here Euclidean space $\mathbb{R}^d$, that indexes all possible spatial locations of interest. Contained within $\mathscr{D}$ is a (possibly random) set D that indexes those parts of $\mathscr{D}$ relevant to the scientific study. We shall see below that D can have different set properties, depending upon whether the spatial process is a geostatistical process, a lattice process, or a point process.

It is convenient to express the joint probability model defined by random $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$ and random D in the following shorthand, $[Y, D]$, which we refer to as the spatial process model. Now,

$$[Y, D] = [Y \mid D][D], \tag{2}$$

where for generic random quantities A and B, their joint probability measure is denoted by $[A, B]$; the conditional probability measure of A given B is denoted by $[A \mid B]$; and the marginal probability measure of B is denoted by $[B]$. In this review of spatial statistics, expression (2) formalizes the general definition of a spatial statistical model given in Cressie (1993, Section 1.1).

The model (2) covers the three principal spatial statistical areas according to three different assumptions about $[D]$, which leads to three different types of spatial stochastic process, $[Y \mid D]$; these are described further in the next section, titled "Spatial Process Models." Spatial statistics has, in the past, classified its methodology according to the types of spatial data, denoted here as $\mathbf{Z}$ (e.g., Ripley, 1981; Upton and Fingleton, 1985; and Cressie, 1993), rather than the types of spatial processes Y that underlie the spatial data.

In this review, we classify our spatial-statistical modeling choices according to the *process model* (2). Then the *data model*, namely the distribution of the data $\mathbf{Z}$ given both Y and D in (2), is the straightforward conditional-probability measure,

$$[\mathbf{Z} \mid Y, D]. \tag{3}$$

For example, the spatial data $\mathbf{Z}$ could be the vector $(Z(\mathbf{s}_1), ..., Z(\mathbf{s}_n))'$, of imperfect measurements of Y taken at given spatial locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \subset D$, where the data are assumed to be *conditionally independent*. That is, the data model is

$$[\mathbf{Z} \mid Y, D] = \prod_{i=1}^{n} [Z(\mathbf{s}_i) \mid Y, D]. \tag{4}$$

Notice that while (4) is based on conditional independence, the *marginal* distribution, $[\mathbf{Z} \mid D]$, does *not* exhibit independence: The spatial-statistical dependence in $\mathbf{Z}$, articulated in the First Law of Geography that was discussed in the abstract, is inherited from $[Y \mid D]$ and (4) as follows:

$$[\mathbf{Z} \mid D] = \int [\mathbf{Z} \mid Y, D][Y \mid D] \, dY.$$

Another example is where the randomness is in D but not in Y. If D is a point process (a special case of a random set), then the data $\mathbf{Z} = \{N, \mathbf{s}_1, \ldots, \mathbf{s}_N\}$, where N is the random number of points in the now-bounded region $\mathscr{D}$, and $D = \{\mathbf{s}_1, \ldots, \mathbf{s}_N\}$ are the random locations of the points. If there are measurements (sometimes called "marks") $\{Z(\mathbf{s}_1), ..., Z(\mathbf{s}_N)\}$ associated with the random points in D, these should be included within $\mathbf{Z}$. That is,

$$\mathbf{Z} = \{N, (\mathbf{s}_1, Z(\mathbf{s}_1)), \ldots, (\mathbf{s}_N, Z(\mathbf{s}_N))\}. \tag{5}$$

This description of spatial statistics given by (2) and (3) captures the (known) uncertainty in the scientific problem being addressed, namely *scientific uncertainty* through the spatial process model (2) and *measurement uncertainty* through the data model (3). Together, (2) and (3) define a *hierarchical statistical model*, here for spatial data, although this hierarchical formulation through the conditional probability distributions, $[\mathbf{Z} \mid Y, D], [Y \mid D]$, and $[D]$ for general Y and D, is appropriate throughout all of applied statistics.

It is implicit in (2) and (3) that any parameters θ associated with the process model and the data model are known. We now discuss how to handle parameter uncertainty in the hierarchical statistical model. A Bayesian would put a probability distribution on θ : Let $[\theta]$ denote the parameter model (or prior) that captures parameter uncertainty. Then, using obvious notation, all the uncertainty in the problem is expressed through the joint probability measure,

$$\begin{aligned}
[\mathbf{Z}, Y, D, \theta] &= [\mathbf{Z}, Y, D \mid \theta][\theta] & (6) \\
&= [\mathbf{Z} \mid Y, D, \theta][Y \mid D, \theta][D \mid \theta][\theta]. & (7)
\end{aligned}$$

A *Bayesian hierarchical model* uses the decomposition (7), but there is also an *empirical hierarchical model* that substitutes a point estimate $\hat{\theta}$ of θ into the first factor on the left-hand side of (6), resulting in its being written as,

$$[\mathbf{Z}, Y, D \mid \hat{\theta}] = [\mathbf{Z} \mid Y, D, \hat{\theta}][Y \mid D, \hat{\theta}][D \mid \hat{\theta}]. \tag{8}$$

Finding efficient estimators of θ from the spatial data $\mathbf{Z}$ is an important problem in spatial statistics, but in this review we emphasize the problem of spatial prediction of Y. In what follows, we shall assume that the parameters are either known or have been estimated. Hence, for convenience, we can drop $\hat{\theta}$ in (8) and observe that the uncertainty in the problem is expressed through the joint probability measure,

$$[\mathbf{Z}, Y, D] = [\mathbf{Z} \mid Y, D][Y \mid D][D], \tag{9}$$

and Bayes' Rule can be used to infer the unknowns Y and D through the *predictive distribution*:

$$[Y, D \mid \mathbf{Z}] = \frac{[\mathbf{Z} \mid Y, D][Y \mid D][D]}{[\mathbf{Z}]}, \tag{10}$$

where $[\mathbf{Z}]$ is the normalization constant that ensures that the right-hand side of (10) integrates or sums to 1. If the spatial index set D *is fixed and known* then we can drop D from (10), and Bayes' Rule simplifies to:

$$[Y \mid \mathbf{Z}] = \frac{[\mathbf{Z} \mid Y][Y]}{[\mathbf{Z}]}, \tag{11}$$

which is the *predictive distribution* of Y (when D is fixed and known). It is this expression that is often used in spatial statistics for prediction. For example, the well known *simple kriging predictor* can easily be identified as the predictive mean of (11) under Gaussian distributional assumptions for both (2) and (3) (Cressie and Wikle, 2011, pp. 139-141).

Our review of spatial statistics starts with a presentation in the next section, "Spatial Process Models," of a number of commonly used spatial process models, which includes multivariate models. Following that, the section, "Spatial Discretization," turns attention to discretization of $\mathscr{D} \subset \mathbb{R}^d$, which is an extremely important consideration when actually computing the predictive distribution (10) or (11). The extension of spatial process models to spatio-temporal process models is discussed in the section, "Spatio-Temporal Processes." Finally, in the "Conclusion" section, we briefly discuss important recent research topics in spatial statistics, but due to a lack of space we are unable to present them in full. It will be interesting to see ten years from now, how these topics have evolved.

2   Spatial Process Models

In this section, we set out various ways that the probability distributions $[Y \mid D]$, $[Y]$, and $[D]$, given in Bayes' Rule (10), can be represented in the spatial context. These are not to be confused with $[\mathbf{Z} \mid D]$ and $[\mathbf{Z}]$, the probability distributions of the spatial data. In many parts of the spatial-statistics literature, this confusion is noticeable when researchers build models directly for $[\mathbf{Z} \mid D]$. Taking a hierarchical approach, we capture knowledge of the scientific process starting with the statistical models, $[Y \mid D]$ and $[D]$, and then we model the measurement errors and missing data through $[\mathbf{Z} \mid Y, D]$. Finally, Bayes' Rule (10) allows inference on the unknowns Y and D through the predictive distribution, $[Y, D \mid \mathbf{Z}]$.

We present three types of spatial process models, where their distinction is made according to the index set D of all spatial locations at which the process Y is defined. For a *geostatistical process*, $D = D^G$, which is a known set over which the locations vary continuously and whose area (or volume) is > 0. For a *lattice process*, $D = D^L$, which is a known set whose locations vary discretely and the number of locations is countable; note that the area of D^L is equal to zero. For a *point process*, $D = D^P$, which is a random set made up of random points in $\mathbb{R}^d$.

2.1 Geostatistical Processes

In this section, we assume that the spatial locations D are given by D^G, where D^G is known. Hence D can be dropped from any of the probability distributions in (10), resulting in (11). This allows us to concentrate on Y and, to feature the spatial index, we write Y equivalently as $\{Y(\mathbf{s}) : \mathbf{s} \in D^G\}$. A property of geostatistical processes is that D^G has positive area and hence is uncountable.

Traditionally, a geostatistical process has been specified up to second moments. Starting with the most general specification, we have

$$\mu_Y(\mathbf{s}) \equiv \mathrm{E}(Y(\mathbf{s})); \quad \mathbf{s} \in D^G \tag{12}$$

$$C_Y(\mathbf{s}, \mathbf{u}) \equiv \mathrm{cov}(Y(\mathbf{s}), Y(\mathbf{u})); \quad \mathbf{s}, \mathbf{u} \in D^G. \tag{13}$$

From (12) and (13), an optimal spatial linear predictor $\hat{Y}(\mathbf{s}_0)$ of $Y(\mathbf{s}_0)$, can be obtained that depends on spatial data $\mathbf{Z} \equiv (Z(\mathbf{s}_1), ..., Z(\mathbf{s}_n))'$. This is an n-dimensional vector indexed by the data's n known spatial locations, $D^{G*} \equiv \{\mathbf{s}_1, ..., \mathbf{s}_n\} \subset D^G$. In practice, estimation of the parameters θ that specify completely (12) and (13) can be problematic due to the lack of replicated data, so Matheron (1963) made stationarity assumptions that together are now known as *intrinsic stationarity*. That is, for all $\mathbf{s}, \mathbf{u} \in D^G$, assume

$$E(Y(\mathbf{s})) = \mu_Y^o \tag{14}$$

$$\mathrm{var}(Y(\mathbf{s}) - Y(\mathbf{u})) = 2\gamma_Y^o(\mathbf{s} - \mathbf{u}), \tag{15}$$

where (15) is equal to $C_Y(\mathbf{s}, \mathbf{s}) + C_Y(\mathbf{u}, \mathbf{u}) - 2C_Y(\mathbf{s}, \mathbf{u})$. The quantity $2\gamma_Y^o(\cdot)$ is called the *variogram*, and $\gamma_Y^o(\cdot)$ is called the *semivariogram* (or occasionally the semivariance).

If the assumption in (15) were replaced by

$$\mathrm{cov}(Y(\mathbf{s}), Y(\mathbf{u})) = C_Y^o(\mathbf{s} - \mathbf{u}), \text{ for all } \mathbf{s}, \mathbf{u} \in D^G, \tag{16}$$

then (16) and (14) together are known as *second-order stationarity*. Matheron chose (15) because he could derive optimal-spatial-linear-prediction (i.e., kriging) equations of $Y(\mathbf{s}_0)$ without having to know or estimate μ_Y^o. Here, "optimal" is in reference to a spatial linear predictor $\hat{Y}(\mathbf{s}_0)$ that minimizes the mean-squared prediction error (MSPE),

$$\mathrm{E}\left[\left(\hat{Y}(\mathbf{s}_0) - Y(\mathbf{s}_0)\right)^2\right], \text{ for any } \mathbf{s}_0 \in D^G, \tag{17}$$

where $\hat{Y}(\mathbf{s}_0) \equiv \sum_{i=1}^n \lambda_i Z(\mathbf{s}_i)$. The minimization in (17) is with respect to the coefficients $\{\lambda_i : i = 1, ..., n\}$ subject to the unbiasedness constraint, $\mathrm{E}(\hat{Y}(\mathbf{s}_0)) = \mathrm{E}(Y(\mathbf{s}_0))$, or equivalently subject to the constraint $\sum_{i=1}^n \lambda_i = 1$ on $\{\lambda_i\}$. With optimally chosen $\{\lambda_i\}$, $\hat{Y}(\mathbf{s}_0)$ is known as the *kriging predictor*. Matheron called this approach to spatial prediction *ordinary kriging*, although it is known in other fields as BLUP (Best Linear Unbiased Prediction); Cressie (1990) gave the history of kriging and showed that it could also be referred to descriptively as *spatial BLUP*.
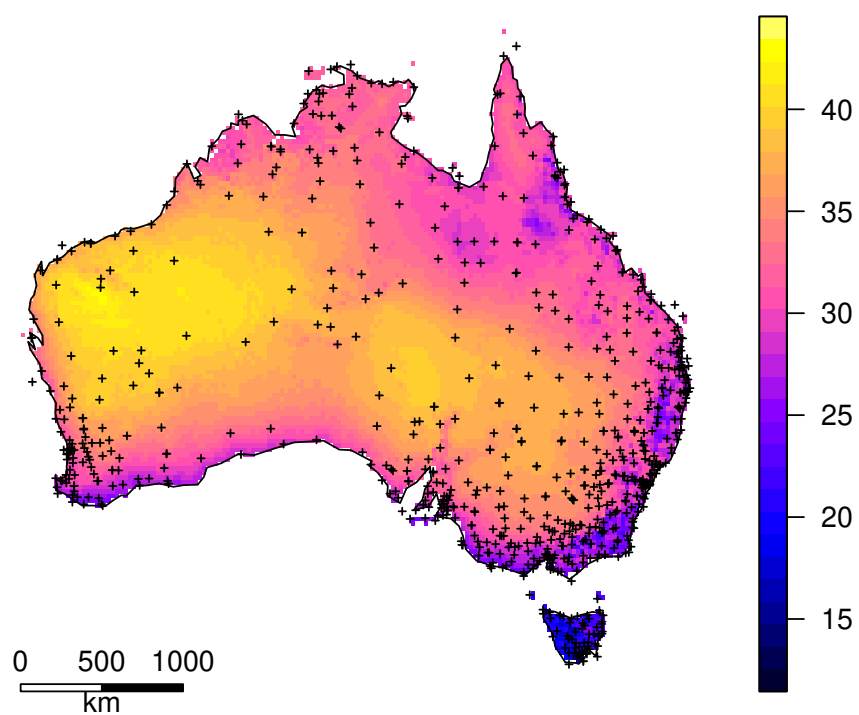
Figure 1: Map of a kriging predictor of Australian temperature in January 2009, superimposed on spatial locations of data.
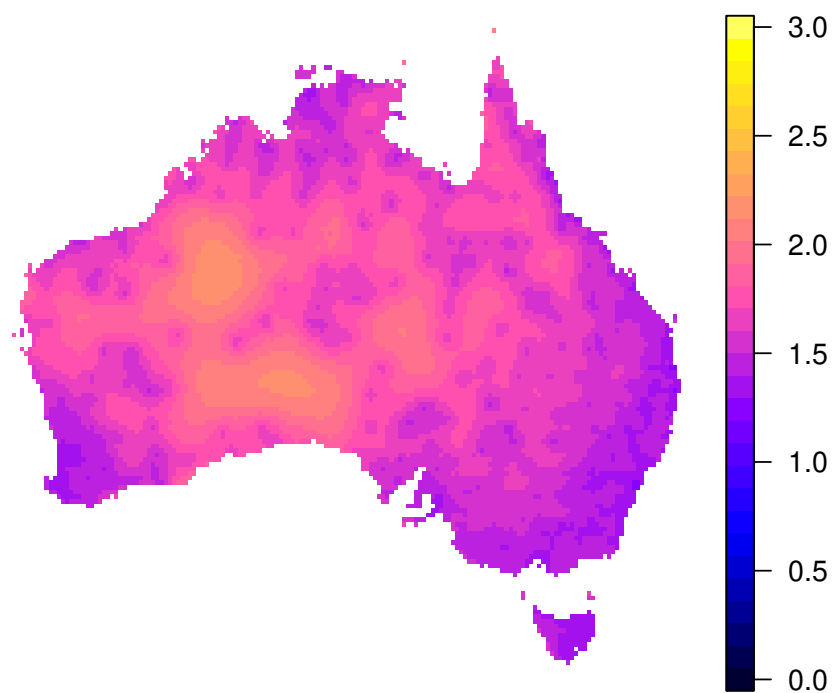
Figure 2: Map of the kriging standard error (18) for the kriging predictor shown in Figure 1.

The constant-mean assumption (14) can be generalized to $E(Y(\mathbf{s})) \equiv \mathbf{x}(\mathbf{s})'\boldsymbol{\beta}$, for $\mathbf{s} \in D^G$, which is a linear regression where the regression coefficients $\boldsymbol{\beta}$ are unknown and the covariate vector $\mathbf{x}(\mathbf{s})$ includes the entry 1. Under this assumption on $\mathrm{E}(Y(\mathbf{s}))$, ordinary kriging is generalized to *universal kriging*, also notated as $\hat{Y}(\mathbf{s}_0)$. Figure 1 shows the universal-kriging predictor of Australian temperature in the month of January 2009, mapped over the whole continent D^G, where the spatial locations $D^{G*} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ of weather stations that supplied the data $\mathbf{Z}$ are superimposed. Formulas for $\hat{Y}(\mathbf{s}_0)$ can be found in, for example, Chilès and Delfiner (2012, Section 3.4).

The optimized MSPE (17) is called the *kriging variance*, and its square root is called the *kriging standard error*:

$$\sigma_k(\mathbf{s}_0) \equiv \left(\mathrm{E} \left(\hat{Y}(\mathbf{s}_0) - Y(\mathbf{s}_0) \right)^2 \right)^{1/2}, \text{ for any } \mathbf{s}_0 \in D^G. \tag{18}$$
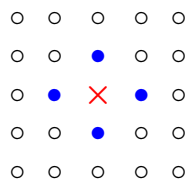
Figure 2 shows a map over D^G of the kriging standard error associated with the kriging predictor mapped in Figure 1. It can be shown that a smaller $\sigma_k(\mathbf{s}_0)$ corresponds to a higher density of weather stations near $\mathbf{s}_0$. While ordinary and universal kriging produce an optimal linear predictor, there is an even better predictor, the *best optimal predictor* (*BOP*), which is the best of all the best predictors obtained under extra constraints (e.g., linearity). From Bayes' Rule (10), the predictor that minimizes the MSPE (16) without any constraints is $Y^*(\mathbf{s}_0) \equiv \mathrm{E}(Y(\mathbf{s}_0) \mid \mathbf{Z})$, which is the mean of the predictive distribution. Notice that the BOP, $Y^*(\mathbf{s}_0)$, is unbiased, namely $\mathrm{E}(Y^*(\mathbf{s}_0)) = \mathrm{E}(Y(\mathbf{s}_0))$, without having to constrain it to be so.

2.2 Lattice Processes

In this section, we assume that the spatial locations D are given by D^L, a known countable subset of $\mathbb{R}^d$. This usually represents a collection of grid nodes, pixels, or small areas and the spatial locations associated with them; we write the countable set of all such locations as $D^L \equiv \{\mathbf{s}_1, \mathbf{s}_2, \ldots\}$. Each $\mathbf{s}_i$ has a set of neighbors, $\mathcal{N}(\mathbf{s}_i) \subset D^L \setminus \mathbf{s}_i$, associated with it, and whose locations are spatially proximate (and note that a location is not considered to be its own neighbor). Spatial-statistical dependence between locations in lattice processes is defined in terms of these neighborhood relations.

Typically, the neighbors are represented by a spatial-dependence matrix $\mathbf{W}$ with entries $w_{i,j}$ nonzero if $\mathbf{s}_j \in \mathcal{N}(\mathbf{s}_i)$, and hence the diagonal entries of $\mathbf{W}$ are all zero. The non-diagonal entries of $\mathbf{W}$ might be, for example, inversely proportional to the distance, $\|\mathbf{s}_i - \mathbf{s}_j\|$, or they might involve some other way of moderating dependence based on spatial proximity. For example, they might be assigned the value 1 if a neighborhood relation exists and 0 otherwise. In this case, $\mathbf{W}$ is called an adjacency matrix, and it is symmetric if $\mathbf{s}_j \in \mathcal{N}(\mathbf{s}_i)$ whenever $\mathbf{s}_i \in \mathcal{N}(\mathbf{s}_j)$ and *vice versa*.

Consider a lattice process in $\mathbb{R}^2$ defined on the finite grid $D^L = \{(x,y) : x, y = 1, \ldots, 5\}$. The first-order neighbors of the grid node (x,y) in the interior of the lattice are the four adjacent nodes, $\mathcal{N}(x,y) = \{(x-1,y), (x,y-1), (x+1,y), (x,y+1)\}$, shown as:

$$
\begin{array}{ccccc}
\circ & \circ & \circ & \circ & \circ \\
\circ & \circ & \bullet & \circ & \circ \\
\circ & \bullet & \times & \bullet & \circ \\
\circ & \circ & \bullet & \circ & \circ \\
\circ & \circ & \circ & \circ & \circ \\
\end{array}
$$

where grid node $\mathbf{s}_i$ is represented by $\times$, and its first-order neighbors are represented by $\bullet$. Nodes $\times$ situated on the boundary of the grid will have less than four neighbours.

The most common type of lattice process is the *Markov random field* (MRF), which has a conditional-probability property in the spatial domain $\mathbb{R}^d$ that is a generalization of the temporal Markov property found in section, "Spatio-Temporal Processes." A lattice process $\{Y(\mathbf{s}) : \mathbf{s} \in D^L\}$ is a MRF if, for all $\mathbf{s}_i \in D^L$, its conditional probabilities satisfy

$$\left[Y(\mathbf{s}_i) \mid \mathbf{Y}(D^L \setminus \mathbf{s}_i)\right] = \left[Y(\mathbf{s}_i) \mid \mathbf{Y}(\mathscr{N}(\mathbf{s}_i))\right], \tag{19}$$

where $\mathbf{Y}(A) \equiv \{Y(\mathbf{s}_j) : \mathbf{s}_j \in A\}$. The MRF is defined in terms of these conditional probabilities (19), which represent statistical dependencies between neighbouring nodes that are captured differently from those given by the variogram or the covariance function. Specifically,

$$\left[Y(\mathbf{s}_i) \mid \mathbf{Y}(D^L \setminus \mathbf{s}_i)\right] = \frac{\exp\{-f(Y(\mathbf{s}_i), \mathbf{Y}(\mathscr{N}(\mathbf{s}_i)))\}}{C}, \tag{20}$$

where C is a normalizing constant that ensures the right-hand side of (20) integrates (or sums) to 1. Equation (20) is also known as a Gibbs random field in statistical mechanics since, under regularity conditions, the Hammersley-Clifford Theorem relates the joint probability distribution to the Gibbs measure (Besag, 1974). The function $f(Y(\mathbf{s}_i), \mathbf{Y}(\mathscr{N}(\mathbf{s}_i)))$ is referred to as the potential energy, since it quantifies the strength of interactions between neighbors. A wide variety of MRF models can be defined by choosing different forms of the potential-energy function (Winkler, 2003, Section 3.2). Note that care needs to be taken to ensure that specification of the model through all the conditional probability distributions, $\{[Y(\mathbf{s}_i) \mid \mathbf{Y}(\mathscr{N}(\mathbf{s}_i))] : \mathbf{s}_i \in D^L\}$, results in a valid joint probability distribution, $\left[\{Y(\mathbf{s}_i) : \mathbf{s}_i \in D^L\}\right]$ (Kaiser and Cressie, 2000).

Revisiting the previous simple example of a first-order neighborhood structure on a regular lattice in $\mathbb{R}^2$, notice that grid nodes situated diagonally across from each other are conditionally independent. Hence, D^L can be partitioned into two sub-lattices D_1^L and D_2^L, such that the values at the nodes in D_1^L are independent given the values at the nodes in D_2^L and vice versa (Besag, 1974; Winkler, 2003, Section 8.1):

$$
\begin{matrix}
\bullet & \circ & \bullet & \circ & \bullet \\
\circ & \bullet & \circ & \bullet & \circ \\
\bullet & \circ & \bullet & \circ & \bullet \\
\circ & \bullet & \circ & \bullet & \circ \\
\bullet & \circ & \bullet & \circ & \bullet
\end{matrix}
$$

This forms a checkerboard pattern where $\{Y(\mathbf{s}) : \mathbf{s} \in D_1^L\}$ at nodes D_1^L represented by $\bullet$ are mutually independent, given the values $\{Y(\mathbf{u}) : \mathbf{u} \in D_2^L\}$ at nodes D_2^L represented by $\circ$.

Besag (1974) introduced the conditional autoregressive (CAR) model, which is a Gaussian MRF that is defined in terms of its conditional means and variances. We refer the reader to LeSage and Pace (2009) for discussion of a different lattice-process model, known as the simultaneous autoregressive (SAR) model, and a comparison of it with the CAR model. We define the CAR model as follows: For $\mathbf{s}_i \in D^L$, $Y(\mathbf{s}_i)$ is conditionally Gaussian defined by its first and second moments,

$$\mathrm{E}(Y(\mathbf{s}_i) \mid \mathbf{Y}(\mathscr{N}(\mathbf{s}_i))) = \sum_{\mathbf{s}_j \in \mathscr{N}(\mathbf{s}_i)} c_{i,j} Y(\mathbf{s}_j) \tag{21}$$

$$\mathrm{var}(Y(\mathbf{s}_i) \mid \mathbf{Y}(\mathscr{N}(\mathbf{s}_i))) = \tau_i^2, \tag{22}$$

where $c_{i,j}$ are spatial autoregressive coefficients such that the diagonal elements $c_{1,1} = \cdots = c_{n,n} = 0$, and $\{\tau_i^2\}$ are the scale parameters for the locations $\{\mathbf{s}_i\}$, respectively. Under an important regularity condition (see below), this specification results in a joint probability distribution that is multivariate Gaussian. That is,

$$\mathbf{Y} \sim Gau(\mathbf{0}, (\mathbf{I} - \mathbf{C})^{-1}\mathbf{M}), \tag{23}$$

where Gau (μ, Σ) denotes a Gaussian distribution with mean vector μ and covariance matrix Σ; the matrix $\mathbf{M} \equiv \text{diag}(\tau_1^2, \ldots, \tau_n^2)$ is diagonal; and the regularity condition referred to above is that the coefficients $\mathbf{C} \equiv \{c_{i,j}\}$ in (21) have to result in $\mathbf{M}^{-1}(\mathbf{I} - \mathbf{C})$ being a symmetric and positive-definite matrix. With a first-order neighborhood structure, such as shown in the simple example above in $\mathbb{R}^2$, the precision matrix is block-diagonal, which makes it possible to sample efficiently from this Gaussian MRF using sparse-matrix methods (Rue and Held, 2005, Section 2.4).

The data vector for lattice processes is $\mathbf{Z} \equiv (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))'$, where $D^{L*} \equiv \{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \subset D^L$. As for the previous subsection, the data model is $[\mathbf{Z} \mid \{Y(\mathbf{s}) : \mathbf{s} \in D^L\}]$ which, to emphasize dependance on parameters θ, we reactivate earlier notation and write it as $[\mathbf{Z} \mid Y, \theta]$. Now, if we write the lattice-process model as $[\{Y(\mathbf{s}) : \mathbf{s} \in D^L\} \mid \theta] \equiv [Y \mid \theta]$, then estimation of θ follows by maximizing the likelihood, $\mathscr{L}(\theta) \equiv \int [\mathbf{Z} \mid Y, \theta] [Y \mid \theta] \, dY$.

Regarding spatial prediction, $Y^*(\mathbf{s}_0) \equiv E(Y(\mathbf{s}_0) \mid \mathbf{Z}, \theta)$ is the best optimal predictor of $Y(\mathbf{s}_0)$, for $\mathbf{s}_0 \in D^L$ and known θ (e.g., Besag et al., 1991). Note that $\mathbf{s}_0$ may not belong to D^{L*}, and hence $Y^*(\mathbf{s}_0)$ is a predictor of $Y(\mathbf{s}_0)$ even when there is no datum observed at the node $\mathbf{s}_0$. Inference on unobserved parts of the process Y is just as important for lattice processes as it is for geostatistical processes.

2.3 Spatial Point Processes and Random Sets

A spatial point process is a countable collection of *random locations* $D \equiv D^P \subset \mathscr{D}$. Closely related to this random set of points is the counting process that we shall call $\{N(A) : A \subset \mathscr{D}\}$, where recall that $\mathscr{D}$ indexes all possible locations of interest, and now we assume it is bounded. For example, if A is a given subset of $\mathscr{D}$, and two of the random points $\{\mathbf{s}_i\}$ are contained in A, then $N(A) = 2$. Since $D^P = \{\mathbf{s}_i\}$ is random and A is fixed, $N(A)$ is a random variable defined on the non-negative integers.

Clearly, the joint distributions $[N(A_1), \ldots, N(A_m)]$, for any subsets $\{A_j : j = 1, \ldots, m\}$ contained in $\mathscr{D}$ (possibly overlapping) and for any $m = 0, 1, 2, \ldots$, are well defined. Spatial dependence can be seen through the spatial proximity between the $\{A_j\}$. Consider just two fixed subsets, A_1 and A_2 (i.e., $m = 2$) and, to avoid ambiguity caused by potentially sharing points, let $A_1 \cap A_2$ be empty. Then no spatial dependence is exhibited if, for any disjoint A_1 and A_2, there is statistical independence; that is,

$$[N(A_1), N(A_2)] = [N(A_1)] [N(A_2)]. \tag{24}$$

The basic point process known as the *Poisson point process* has the independence property (24), and its associated counting process satisfies

$$[N(A)] = \exp\{-\lambda(A)\} \frac{\lambda(A)^{N(A)}}{N(A)!}; \quad A \subset \mathscr{D}, \tag{25}$$

where $\lambda(A) \equiv \int_A \lambda(\mathbf{s}) d\mathbf{s}$. In (25), $\lambda(\cdot)$ is a given intensity function defined according to:

$$\lambda(\mathbf{s}) \equiv \lim_{|\delta\mathbf{s}| \to 0} \frac{\text{E}(Y(\delta\mathbf{s}))}{|\delta\mathbf{s}|}, \tag{26}$$

where $\delta\mathbf{s}$ is a small set centered at $\mathbf{s} \in \mathscr{D}$, and whose volume is $|\delta\mathbf{s}|$.

In (25), the special case of $\lambda(s) \equiv \lambda$, for all $\mathbf{s} \in \mathscr{D}$, results in a *homogeneous Poisson point process*, and a simulation of it is shown in Figure 3. The simulation was obtained using an equivalent probabilistic representation for which the count random variable $N(\mathscr{D})$, for $\mathscr{D} =$
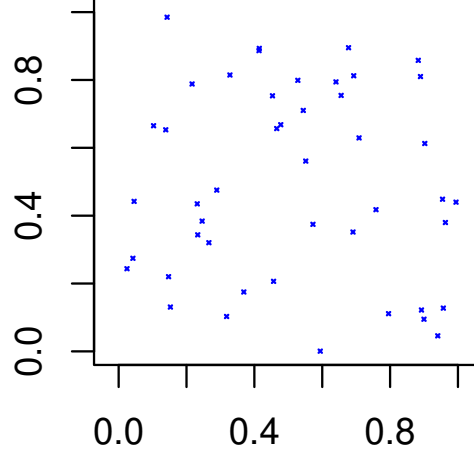
Figure 3: A realization on the unit square $\mathscr{D} = [0,1] \times [0,1]$, of the homogenous Poisson point process (25) with parameter $\lambda = 50$; for this realization, $N(\mathscr{D}) = 46$.

$[0,1] \times [0,1]$, was simulated according to (25). Then, conditional on $N(\mathscr{D})$, $\{\mathbf{s}_1, \ldots, \mathbf{s}_{N(\mathscr{D})}\}$ was simulated independently and identically according to the uniform distribution,

$$[\mathbf{u}] = \begin{cases} \frac{1}{\lambda(\mathscr{D})} \ ; & \mathbf{u} \in \mathscr{D} \\ 0 \ ; & \text{elsewhere.} \end{cases} \tag{27}$$

This representation explains why the homogenous Poisson point process is commonly referred to as a *Completely Spatially Random* (CSR) process, and why it is used as a baseline for testing for the absence of spatial dependance in a point process. That is, before a spatial model is fitted to a point pattern, a test of the null hypothesis that the point pattern originates from a CSR process, is often carried out. Rejection of CSR then justifies the fitting of spatially dependent point processes to the data (e.g., Ripley, 1981; Diggle, 2013).

Much of the early research in point processes was devoted to establishing test statistics that were sensitive to various types of departures from the CSR process (e.g., Cressie, 1993, Section 8.2). This was followed by researchers' defining and then estimating spatial-dependence measures such as the second-order-intensity function and the K-function (e.g., Ripley, 1981, Chapter 8), where inference was often in terms of method-of-moments estimation of these functions. More efficient likelihood-based inference came later; Baddeley et al. (2015) give a comprehensive review of these methodologies for point processes. From a modeling perspective, particular attention has been paid to the log Gaussian Cox point processes; here, $\lambda(\cdot)$ in (25) is random, such that $\{\log(\lambda(\mathbf{s})) : \mathbf{s} \in \mathscr{D}\}$ is a Gaussian process (e.g., Møller and Waagepetersen, 2003). This model leads naturally to hierarchical Bayesian inference for $\lambda(\cdot)$ and its parameters (e.g., Gelfand and Schliep, 2018).

If an attribute process, $\{Y(\mathbf{s}_i) : \mathbf{s}_i \in D^P\}$, is included with the spatial point process D^P, one obtains a so-called *marked point process* (e.g., Cressie, 1993, Section 8.7). For example, the

study of a natural forest where both the locations $\{\mathbf{s}_i\}$ and the sizes of the trees $\{Y(\mathbf{s}_i)\}$ are modeled together probabilistically, results in a marked point process where the "mark" process is a spatial process $\{Y(\mathbf{s}_i) : \mathbf{s}_i \in D^P\}$ of tree size. Now, Bayes' Rule given by (10), where both Y and D $(= D^P)$ are random, should be used to make inference on Y and D^P through the predictive distribution $[Y, D^P \mid \mathbf{Z}]$. Here, $\mathbf{Z}$ consists of the number of trees, the trees' locations, and their size measurements, as denoted in (5). After marginalization, we can obtain $[D^P \mid \mathbf{Z}]$, the predictive distribution of the spatial point process D^P.

A spatial point process is a special case of a random set, which is a random quantity in Euclidean space that was defined rigorously by Matheron (1975). Some geological processes are more naturally modeled as set-valued phenomena (e.g., the facies of a mineralization), however inference for random-set processes has lagged behind those for spatial point processes. It is difficult to define a likelihood based on set-valued data, which has held back statistically efficient inferences; nevertheless, basic method-of-moment estimators are often available. The most well known random set that allows statistical inference from set-valued data is the Boolean Model (e.g., Cressie and Wikle, 2011, Section 4.4).

2.4 Multivariate Spatial Processes

The previous subsections have presented single spatial statistical processes but, as models become more realistic representations of a complex world, there is a need to express interactions between multiple processes. This is most directly seen by modeling vector-valued "Geostatistical Processes," $\{\mathbf{Y}(\mathbf{s}) : \mathbf{s} \in D^G\}$, and vector-valued "Lattice Processes," $\{\mathbf{Y}(\mathbf{s}_i) : \mathbf{s}_i \in D^L\}$, where the k-dimensional vector $\mathbf{Y}(\mathbf{s}) \equiv (Y_1(\mathbf{s}), \ldots, Y_k(\mathbf{s}))'$ represents the multiple processes at the generic location $\mathbf{s} \in \mathscr{D}$. Vector-valued spatial point processes, discussed in Section 2.3, can be represented as a set of k point processes, $\{\{\mathbf{s}_{1,i}\}, \ldots, \{\mathbf{s}_{k,i}\}\}$, and these are presented in Baddeley et al. (2015, Chapter 14). If we adopt a hierarchical-statistical-modeling approach, it is possible to construct multivariate spatial processes whose component univariate processes could come from any of the three types of spatial processes presented in the previous three subsections. This is because, at a deeper level of the hierarchy, a core *multivariate geostatistical process* can control the spatial dependance for processes of any type, which allows the possibility of hybrid multivariate spatial statistical processes.

In what follows, we describe briefly two approaches to constructing multivariate geostatistical processes, one based on a joint approach and the other based on a conditional approach. We consider the case $k = 2$, namely the bivariate spatial process $\{(Y_1(\mathbf{s}), Y_2(\mathbf{s}))' : \mathbf{s} \in D^G\}$, for illustration. The joint approach involves directly constructing a valid spatial statistical model from $\boldsymbol{\mu}(\mathbf{s}) \equiv (\mu_1(\mathbf{s}), \mu_2(\mathbf{s}))' \equiv (\mathrm{E}(Y_1(\mathbf{s})), \mathrm{E}(Y_2(\mathbf{s})))'$, for $\mathbf{s} \in D^G$, and from

$$\mathrm{cov}(Y_l(\mathbf{s}), Y_m(\mathbf{u})) \equiv \mathrm{C}_{lm}(\mathbf{s}, \mathbf{u}); \quad l, m = 1, 2, \tag{28}$$

for $\mathbf{s}, \mathbf{u} \in D^G$. The bivariate-process mean $\boldsymbol{\mu}(\cdot)$ is typically modeled as a vector linear regression; hence it is straightforward to model the bivariate mean once the appropriate regressors have been chosen.

Analogous to the univariate case, the set of covariance and cross-covariance functions, $\{\mathrm{C}_{11}(\cdot, \cdot), \mathrm{C}_{22}(\cdot, \cdot), \mathrm{C}_{12}(\cdot, \cdot), \mathrm{C}_{21}(\cdot, \cdot)\}$, have to satisfy positive-definiteness conditions for the bivariate geostatistical model to be valid, and it is important to note that, in general, $\mathrm{C}_{12}(\mathbf{s}, \mathbf{u}) \neq \mathrm{C}_{21}(\mathbf{s}, \mathbf{u})$. There are classes of valid models that exhibit symmetric cross-dependance, namely $\mathrm{C}_{12}(\mathbf{s}, \mathbf{u}) = \mathrm{C}_{21}(\mathbf{s}, \mathbf{u})$, such as the linear model of co-regionalization (Gelfand et al., 2004). These are not reasonable models for ore-reserve estimation when there has been preferential mineralization in the ore body.

The joint approach can be contrasted with a conditional approach (Cressie and Zammit-Mangion, 2016), where each of the k processes is a node of a directed acyclic graph that guides the conditional dependance of any process, given the remaining processes. Again consider the bivariate case (i.e., $k = 2$), where there are only two nodes such that $Y_1(\cdot)$ is at node 1, $Y_2(\cdot)$ is at node 2, and a directed edge is declared from node 1 to node 2. Then the appropriate way to model the joint distribution is through

$$[Y_1(\cdot), Y_2(\cdot)] = [Y_2(\cdot) \mid Y_1(\cdot)][Y_1(\cdot)], \tag{29}$$

where $[Y_2(\cdot) \mid Y_1(\cdot)]$ is shorthand for $[Y_2(\cdot) \mid \{Y_1(\mathbf{s}) : \mathbf{s} \in D^G\}]$.

The geostatistical model for $[Y_1(\cdot)]$ is simply a univariate model based on a mean function $\mu_1(\cdot)$ and a valid covariance function $C_{11}(\cdot, \cdot)$, which was discussed in "Geostatistical Processes." Now assume that $Y_2(\cdot)$ depends on $Y_1(\cdot)$ as follows: For $\mathbf{s}, \mathbf{u} \in D^G$,

$$\mathrm{E}[Y_2(\mathbf{s}) \mid Y_1(\cdot)] \equiv \mu_2(\mathbf{s}) + \int_{D^G} b(\mathbf{s}, \mathbf{v})\, (Y_1(\mathbf{v}) - \mu_1(\mathbf{v}))\, \mathrm{d}\mathbf{v}, \tag{30}$$

$$\mathrm{cov}(Y_2(\mathbf{s}), Y_2(\mathbf{u}) \mid Y_1(\cdot)) \equiv C_{2|1}(\mathbf{s}, \mathbf{u}), \tag{31}$$

where $C_{2|1}(\cdot, \cdot)$ is a valid univariate covariance function and $b(\cdot, \cdot)$ is an integrable interaction function. The conditional-moment assumptions given by (30) and (31) follow if one assumes that $(Y_1(\cdot), Y_2(\cdot))'$ is a bivariate Gaussian process.

Cressie and Zammit-Mangion (2016) show that, from (30) and (31),

$$C_{12}(\mathbf{s}, \mathbf{u}) = \int_{D^G} C_{11}(\mathbf{s}, \mathbf{v}) b(\mathbf{u}, \mathbf{v})\, \mathrm{d}\mathbf{v} \tag{32}$$

$$C_{21}(\mathbf{s}, \mathbf{u}) = \int_{D^G} C_{11}(\mathbf{v}, \mathbf{u}) b(\mathbf{v}, \mathbf{s})\, \mathrm{d}\mathbf{v} \tag{33}$$

$$C_{22}(\mathbf{s}, \mathbf{u}) = C_{2|1}(\mathbf{s}, \mathbf{u}) + \int_{D^G} \int_{D^G} b(\mathbf{s}, \mathbf{v}) C_{11}(\mathbf{v}, \mathbf{w}) b(\mathbf{u}, \mathbf{w})\, \mathrm{d}\mathbf{v}\, \mathrm{d}\mathbf{w}, \tag{34}$$

for $\mathbf{s}, \mathbf{u} \in D^G$. Along with $\mu_1(\cdot)$, $\mu_2(\cdot)$, and $C_{11}(\cdot, \cdot)$, these functions (32)–(34) define a valid bivariate geostatistical process $[Y_1(\cdot), Y_2(\cdot)]$. A notable property of the conditional approach is that asymmetric cross-dependance (i.e., $C_{12}(\mathbf{s}, \mathbf{u}) \neq C_{21}(\mathbf{s}, \mathbf{u})$) occurs if $b(\mathbf{s}, \mathbf{u}) \neq b(\mathbf{u}, \mathbf{s})$.

In summary, the conditional approach allows multivariate modeling to be carried out validly by simply specifying $\boldsymbol{\mu}(\cdot) = (\mu_1(\cdot), \mu_2(\cdot))'$ and two valid univariate covariance functions, $C_1(\cdot, \cdot)$ and $C_{2|1}(\cdot, \cdot)$. The strengths of the conditional approach are that only univariate covariance functions need to be specified (for which there is a very large body of research; e.g., Cressie and Wikle, 2011, Chapter 4), and that only integrability of $b(\cdot, \cdot)$, the interaction function, needs to be assumed (Cressie and Zammit-Mangion, 2016).

3   Spatial Discretization

Although geostatistical processes are defined on a continuous spatial domain D^G, this can limit the practical feasibility of statistical inferences due to computational and mathematical considerations. For example, kriging from an n-dimensional vector of data involves the inversion of an $n \times n$ covariance matrix, which requires order n^3 floating-point operations and order n^2 storage in available memory. These costs can be prohibitive for large spatial datasets; hence, spatial discretization to achieve scalable computation for spatial models is an active area of research.
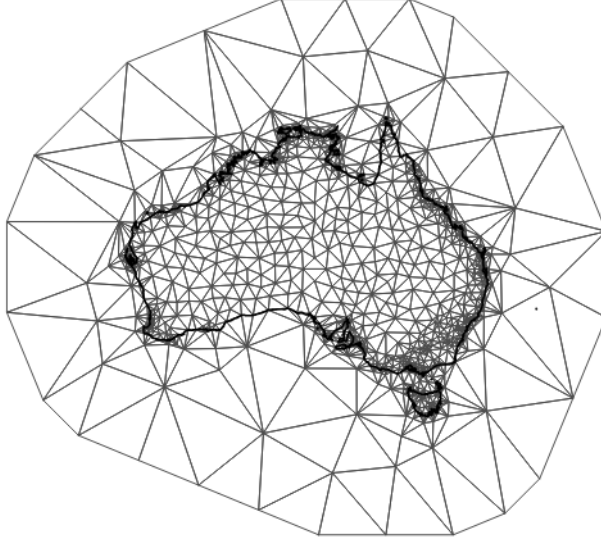
Figure 4: Discretization of the spatial domain $\mathscr{D}$, a convex region around Australia, into a triangular lattice.

In practical applications, spatial statistical inference is required up to a finite spatial resolution. Many approaches take advantage of this by dividing the spatial domain $\mathscr{D}$ into a lattice of discrete points in $\mathscr{D}$, as shown in Figure 4. As a consequence of this discretization, a geostatistical process can be approximated by a lattice process, such as a Gaussian MRF (e.g., Rue and Held, 2005, Section 5.1), however sometimes this can result in undesirable discretization errors and artifacts. More sophisticated approaches have been developed to obtain highly accurate approximations of a geostatistical (i.e., continuously indexed) spatial process evaluated over an irregular lattice, as we now discuss.

Let the original domain $\mathscr{D}$ be bounded and suppose it is tesselated into the areas $\{A_j \subset \mathscr{D} : j = 1, \ldots, m\}$ that are small, non-overlapping *basic areal units* (BAUs; Nguyen et al., 2012), so that $\mathscr{D} = \cup_{j=1}^m A_j$, and $A_j \cap A_k$ is an empty set for any $j \neq k \in \{1, \ldots, m\}$; Figure 4 gives an example of triangular BAUs. Spatial basis functions $\{\phi_\ell(\cdot) : \ell = 1, \ldots, r\}$, can then be defined on the BAUs. For example, Lindgren et al. (2011) used triangular basis functions where $r > m$, while fixed rank kriging (FRK; Cressie and Johannesson, 2008; Zammit-Mangion and Cressie, 2021) can employ a variety of different basis functions for $r < m$, including multi-resolution wavelets and bisquares.

Vecchia approximations (e.g., Datta et al., 2016; Katzfuss et al., 2020) are also defined using a lattice of discrete points $D^L \subset D^G \subset \mathscr{D}$, that include the coordinates of the observed data $D^{G*} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ and the prediction locations $\{\mathbf{s}_{n+1}, \ldots, \mathbf{s}_{n+p}\}$. Let $[\mathbf{X}] \equiv [\mathbf{Z}, Y]$, where data $\mathbf{Z}$ and spatial process Y are associated with the lattice $D^L \equiv \{\mathbf{s}_1, \ldots, \mathbf{s}_n, \mathbf{s}_{n+1}, \ldots, \mathbf{s}_{n+p}\}$. It is a property of joint and conditional distributions that this can be factorized into a product:

$$[\mathbf{X}] = \prod_{i=1}^{n+p} [X(\mathbf{s}_i) \mid X(\mathbf{s}_1), \ldots, X(\mathbf{s}_{i-1})]. \tag{35}$$

In the previous section, the set of spatial coordinates D^L had no fixed ordering. However, a Vecchia approximation requires that an artificial ordering is imposed on $\{\mathbf{s}_1, \ldots, \mathbf{s}_{n+p}\}$. Let the ordering be denoted by $\{\mathbf{s}_{(1)}, \ldots, \mathbf{s}_{(n+p)}\}$, and define the set of neighbors $\mathcal{N}(\mathbf{s}_{(i)}) \subset \{\mathbf{s}_{(1)}, \ldots, \mathbf{s}_{(i-1)}\}$, similarly to "Lattice Processes," except that these neighborhood relations are not reciprocal: If

13

for $j < i$, $\mathbf{s}_{(j)}$ belongs to $\mathcal{N}(\mathbf{s}_{(i)})$, then $\mathbf{s}_{(i)}$ cannot belong to $\mathcal{N}(\mathbf{s}_{(j)})$. As part of the Vecchia approximation, a fixed upper bound $q \ll n$ on the number of neighbours is chosen. That is, $|\mathcal{N}(\mathbf{s}_{(i)})| \leq q$, so that the lattice formed by $\{\mathcal{N}(\mathbf{s}_i) : i = 1, ..., n + p\}$ is a directed acyclic graph, which results in a partial order in $\mathcal{D}$ (Cressie and Davidson, 1998).

The joint distribution $[\mathbf{X}]$ given by (35) is then approximated by:

$$\prod_{i=1}^{n+p} [X(\mathbf{s}_{(i)}) \mid \mathbf{X}(\mathcal{N}(\mathbf{s}_{(i)}))] \equiv [\widetilde{\mathbf{X}}], \tag{36}$$

which is a Partially-Ordered Markov Model (POMM; Cressie and Davidson, 1998). This Vecchia approximation, $[\widetilde{\mathbf{X}}]$, is a distribution coming from a valid spatial process on the original, uncountable, unordered index set D^G (Datta et al., 2016), which means that it can be used as a geostatistical process model with considerable computational advantages. For example, it can be used as a random log-intensity function, $\log(\lambda(\mathbf{s}))$, in a hierarchical point-process model, or it can be combined with other processes to define models described in "Multivariate Spatial Processes." However, in all of these contexts it should be remembered that the resulting predictive process, $[\widetilde{\mathbf{X}} \mid \mathbf{Z}]$, is an approximation to the true predictive process, $[\mathbf{X} \mid \mathbf{Z}]$.

4 Spatio-Temporal Processes

The section titled "Multivariate Spatial Processes" introduced processes that were written in vector form as,

$$\mathbf{Y}(s) \equiv (Y_1(\mathbf{s}), ..., Y_k(\mathbf{s}))'; \ \mathbf{s} \in D^G. \tag{37}$$

In that section, we distinguished between the joint approach and the conditional approach to multivariate-spatial-statistical modeling and, under the conditional approach, we used a directed acyclic graph to give a blueprint for the multivariate spatial dependance.

Now, consider a spatio-temporal process,

$$\{Y(\mathbf{s};t) : \mathbf{s} \in D^G; \ t \in \mathcal{T}\}, \tag{38}$$

where $\mathcal{T}$ is a temporal index set. Clearly, if $\mathcal{T} = \{1, 2, ...\}$ then (38) becomes a spatial process of time series, $\{Y(\mathbf{s}; 1), Y(\mathbf{s}; 2), \cdots : \mathbf{s} \in D^G\}$. If $\mathcal{T} = \{1, 2, ..., k\}$, and we define $Y_j(\mathbf{s}) \equiv Y(\mathbf{s}; j)$, for $j = 1, ..., k$, then the resulting spatio-temporal process can be represented as a multivariate spatial process given by (37). Not surprisingly, the same dichotomy of approach to modeling statistical dependance (i.e., joint versus conditional) occurs for spatio-temporal processes as it does for multivariate spatial processes.

Describing all possible covariances between Y at any spatio-temporal "location" $(\mathbf{s};t)$ and any other one $(\mathbf{u};v)$, amounts to treating "time" as simply another dimension to be added to the d-dimensional Euclidean space, $\mathbb{R}^d$. Taking this approach, spatio-temporal statistical dependence can be expressed in $(d + 1)$-dimensional space through the covariance function,

$$C(\mathbf{s};t,\mathbf{u};v) \equiv \text{cov}(Y(\mathbf{s};t), Y(\mathbf{u};v)); \qquad \mathbf{s}, \mathbf{u} \in D^G, \ t, v \in \mathcal{T}. \tag{39}$$

Of course, the time dimension has different units than the spatial dimensions, and its interpretation is different since the future is unobserved. Hence, the joint modeling of space and time based on (39) must be done with care to account for the special nature of the time dimension in this *descriptive approach* to spatio-temporal modeling.

From current and past spatio-temporal data $\mathbf{Z}$, predicting past values of Y is called *smoothing*, predicting unobserved values of the current Y is called *filtering*, and predicting future

values of Y is called *forecasting*. The Kalman filter (Kalman, 1960) was developed to provide fast predictions of the current state using a methodology that recognises the ordering of the time dimension. Today's filtered values become "old" the next day when a new set of current data are received. Using a dynamical approach that we shall now describe, the Kalman filter updates yesterday's optimal filtered value with today's data very rapidly, to obtain a current optimal filtered value.

The best way to describe the dynamical approach is to discretize the spatial domain. The previous section, "Spatial Discretization," describes a number of ways this can be done; here we shall consider the discretization that is most natural for storing the attribute and location information in computer memory, namely a fine-resolution lattice D^L of pixels or voxels (short for "volume elements"). Replace $\{Y(\mathbf{s};t) : \mathbf{s} \in D^G, t = 1, 2, \dots\}$ with $\{Y(\mathbf{s};t) : \mathbf{s} \in D^L, t = 1, 2, \dots\}$, where $D^L \equiv \{\mathbf{s}_1, \dots, \mathbf{s}_m\}$ are the centroids of elements of small area (or small volume) that make up D^G. Often the areas of these elements are specified to be equal, having been defined by a regular grid. As we explain below, this allows a *dynamical approach* to constructing a statistical model for the spatio-temporal process Y on the discretized space-time cube, $\{\mathbf{s}_1, \dots, \mathbf{s}_m\} \times \{1, 2, \dots\}$.

Define $\mathbf{Y}_t \equiv (Y(\mathbf{s};t) : \mathbf{s} \in D^L)'$, which is an m-dimensional vector. Because of the temporal ordering, we can write the joint distribution of $\{Y(\mathbf{s};t) : \mathbf{s} \in D^L, t = 1, \dots, k\}$ from $t = 1$ up to the present time $t = k$, as

$$[\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_k] = [\mathbf{Y}_1][\mathbf{Y}_2 \mid \mathbf{Y}_1] \dots [\mathbf{Y}_k \mid \mathbf{Y}_{k-1}, \dots, \mathbf{Y}_2, \mathbf{Y}_1], \tag{40}$$

which has the same form as (35). Note that this conditional modeling of space and time is a natural approach, since time is completely ordered. The next step is to make a Markov assumption, and hence (40) can be written as

$$[\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_k] = [\mathbf{Y}_1] \prod_{j=2}^{k} [\mathbf{Y}_j \mid \mathbf{Y}_{j-1}]. \tag{41}$$

This is the same Markov property that we previously discussed in "Lattice Processes," except it is now applied to the completely ordered one-dimensional domain, $\mathscr{T} = \{1, 2, \dots\}$, and $\mathscr{N}(j) = j - 1$. The Markov assumption makes our approach dynamical: It says that the present, conditional on the past, in fact only depends on the "most recent past." That is, since $\mathscr{N}(j) = j - 1$, the factor $[\mathbf{Y}_j \mid \mathbf{Y}_{j-1}, \dots, \mathbf{Y}_2, \mathbf{Y}_1] = [\mathbf{Y}_j \mid \mathbf{Y}_{j-1}]$, which results in the model (41).

For further information on the types of models used in the descriptive approach given by (39) and the types of models used in the dynamical approach given by (41), see Cressie and Wikle (2011, Chapters 6–8) and Wikle et al. (2019, Chapters 4 and 5). The statistical analysis of observations from these processes is known as *spatio-temporal statistics*. Inference (estimation and prediction) from spatio-temporal data using R software can be found in Wikle et al. (2019).

5   Conclusion

Spatial-statistical methods distinguish themselves from spatial-analysis methods found in the geographical and environmental sciences, by providing well calibrated quantification of the uncertainty involved with estimation or prediction. Uncertainty in the scientific phenomenon of interest is represented by a spatial *process model*, $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$, defined on possibly random D in $\mathbb{R}^d$, while measurement uncertainty in the observations $\mathbf{Z}$ is represented in a *data model*. In "Introduction," we saw how these two models are combined using Bayes' Rule (10), or the simpler version (11), to calculate the overall uncertainty needed for statistical inference.

With some exceptions (e.g., Cressie and Kornak, 2003), spatial-statistical models (1) rarely consider the case of measurement error in the locations in $\mathscr{D}$. Here we focus on a spatial-statistical model for the location error: Write the observed locations as $D^* \equiv \{\mathbf{u}_i : i = 1,...,n\}$; in this case, a part of the data model is $[D^* \mid D]$, and a part of the process model is $[D]$. Finally then, the data consist of both locations and attributes and are $\mathbf{Z} \equiv \{(\mathbf{u}_i, Z(\mathbf{u}_i)) : i = 1,\ldots,n\}$, the spatial process model is $[Y,D]$, and the data model is $[\mathbf{Z} \mid Y,D]$. Then Bayes' Rule given by (10) is used to infer the unknown Y and D from the predictive distribution $[Y,D \mid \mathbf{Z}]$.

There are three main types of spatial process models: geostatistical processes where uncertainty is in the process Y, which is indexed continuously in $D = D^G$; lattice processes where uncertainty is also in Y, but now Y is indexed over a countable number of spatial locations $D = D^L$; and point processes where uncertainty is in the spatial locations $D = D^P$. Multiple spatial processes can interact with each other to form a multivariate spatial process. Importantly, processes can vary over time as well as spatially, forming a spatio-temporal process.

As the size of spatial datasets have been increasing dramatically, more and more attention has been devoted to scalable computation for spatial-statistical models. Of particular interest are methods that use "Spatial Discretization" to approximate a continuous spatial domain, D^G. There are other recent advances in spatial statistics that we feel are important to mention, but their discussion here is necessarily brief.

Physical barriers can sometimes interrupt the statistical association between locations in close spatial proximity. Barrier models (Bakka et al., 2019) have been developed to account for these kinds of discontinuities in the spatial correlation function. Other methods for modeling nonstationarity, anisotropy, and heteroskedasticity in spatial process models are an active area of research.

It can often be difficult to select appropriate prior distributions for the parameters of a stationary spatial process, for example its correlation-length scale. Penalised complexity (PC) priors (Simpson et al., 2017) are a way to encourage parsimony by favoring parameter values that result in the simplest model consistent with the data. The likelihood function of a point process or of a non-Gaussian lattice model can be both analytically and computationally intractable. Surrogate models, emulators, and quasi-likelihoods have been developed to approximate these intractable likelihoods (Moores et al., 2020).

Copulas are an alternative method for modeling spatial dependence in multivariate data, particularly when the data are non-Gaussian (Krupskii and Genton, 2019). One area where non-Gaussianity can arise is in modeling the spatial association between extreme events, such as for temperature or precipitation (Tawn et al., 2018; Bacro et al., 2020).

As a final comment, we reflect on how the field of geostatistics has evolved, beginning with applications of spatial stochastic processes to mining: In the 1970s, Georges Matheron and his Centre of Mathematical Morphology in Fontainebleau were part of the Paris School of Mines, a celebrated French tertiary-education and research institution. To see what the geostatistical methodology of the time was like, the interested reader could consult Journel and Huijbregts (1978), for example. Over the following decade, geostatistics became notationally and methodologically integrated into statistical science and the growing field of spatial statistics (e.g., Ripley, 1981; Cressie, 1993). It took one or two more decades before geostatistics became integrated into the hierarchical-statistical-modeling approach to spatial statistics (e.g., Cressie and Wikle, 2011, Chapter 4). The presentation given in our review takes this latter viewpoint and explains well known geostatistical quantities such as the variogram and kriging in terms of this advanced, modern view of geostatistics. We also include a discussion of uncertainty in the spatial index set as part of our review, which offers new insights into spatial-statistical modeling. Probabilistic difficulties with geostatistics, of making inference on a possibly non-

countable number of spatial random variables from a finite number of observations, can be finessed by discretizing the process. In a modern computing environment, this is key to doing spatial-statistical inference (including kriging).

References

Bacro JN, Gaetan C, Opitz T, Toulemonde G (2020) Hierarchical space-time modeling of asymptotically independent exceedances with an application to precipitation data. Journal of the American Statistical Association 115(530):555–569, DOI 10.1080/01621459.2019. 1617152

Baddeley A, Rubak E, Turner R (2015) Spatial Point Patterns: Methodology and Applications with R. Chapman & Hall/CRC Press, Boca Raton, FL

Bakka H, Vanhatalo J, Illian JB, Simpson D, Rue H (2019) Non-stationary Gaussian models with physical barriers. Spatial Statistics 29:268–288, DOI 10.1016/j.spasta.2019.01.002

Besag J (1974) Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 36(2):192–236

Besag J, York J, Mollié A (1991) Bayesian image restoration, with two applications in spatial statistics. Annals of the Institute of Statistical Mathematics 43:1–20

Chilès JP, Delfiner P (2012) Geostatistics: Modeling Spatial Uncertainty, 2nd edn. John Wiley & Sons, Hoboken, NJ

Cressie N (1990) The origins of kriging. Mathematical Geology 22(3):239–252, DOI 10.1007/ BF00889887

Cressie N (1993) Statistics for Spatial Data, revised edn. John Wiley & Sons, Hoboken, NJ

Cressie N, Davidson JL (1998) Image analysis with partially ordered Markov models. Computational Statistics & Data Analysis 29(1):1–26

Cressie N, Johannesson G (2008) Fixed rank kriging for very large spatial data sets. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70(1):209–226, DOI 10.1111/j.1467-9868.2007.00633.x

Cressie N, Kornak J (2003) Spatial statistics in the presence of location error with an application to remote sensing of the environment. Statistical Science 18(4):436–456, DOI 10.1214/ss/ 1081443228

Cressie N, Wikle CK (2011) Statistics for Spatio-Temporal Data. John Wiley & Sons, Hoboken, NJ

Cressie N, Zammit-Mangion A (2016) Multivariate spatial covariance models: A conditional approach. Biometrika 103(4):915–935, DOI 10.1093/biomet/asw045

Datta A, Banerjee S, Finley AO, Gelfand AE (2016) Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. Journal of the American Statistical Association 111(514):800–812, DOI 10.1080/01621459.2015.1044091

Diggle PJ (2013) Statistical Analysis of Spatial and Spatio-Temporal Point Patterns, 3rd edn. Chapman & Hall/CRC Press, Boca Raton, FL

Gelfand AE, Schliep EM (2018) Bayesian inference and computing for spatial point patterns. In: NSF-CBMS Regional Conference Series in Probability and Statistics, Institute of Mathematical Statistics and the American Statistical Association, Alexandria, VA, vol 10, pp i–125

Gelfand AE, Schmidt AM, Banerjee S, Sirmans C (2004) Nonstationary multivariate process modeling through spatially varying coregionalization. Test 13(2):263–312

Journel AG, Huijbregts CJ (1978) Mining Geostatistics. Academic Press, London, UK

Kaiser MS, Cressie N (2000) The construction of multivariate distributions from Markov random fields. Journal of Multivariate Analysis 73(2):199–220

Kalman RE (1960) A new approach to linear filtering and prediction problems. Transactions of the ASME – Journal of Basic Engineering 82:35–45

Katzfuss M, Guinness J, Gong W, Zilber D (2020) Vecchia approximations of Gaussian-process predictions. Journal of Agricultural, Biological and Environmental Statistics 25(3):383–414, DOI https://doi.org/10.1007/s13253-020-00401-7

Krupskii P, Genton MG (2019) A copula model for non-Gaussian multivariate spatial data. Journal of Multivariate Analysis 169:264–277

LeSage J, Pace RK (2009) Introduction to Spatial Econometrics. Chapman & Hall/CRC Press, Boca Raton, FL

Lindgren F, Rue H, Lindström J (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73(4):423–498, DOI 10.1111/j.1467-9868.2011.00777.x

Matheron G (1963) Principles of geostatistics. Economic Geology 58(8):1246–1266

Matheron G (1975) Random Sets and Integral Geometry. John Wiley & Sons, Hoboken, NJ

Møller J, Waagepetersen RP (2003) Statistical Inference and Simulation for Spatial Point Processes. Chapman & Hall/CRC Press, Boca Raton, FL

Moores MT, Pettitt AN, Mengersen KL (2020) Bayesian computation with intractable likelihoods. In: Case Studies in Applied Bayesian Data Science, Springer-Verlag, Berlin, pp 137–151

Nguyen H, Cressie N, Braverman A (2012) Spatial statistical data fusion for remote sensing applications. Journal of the American Statistical Association 107(499):1004–1018, DOI 10.1080/01621459.2012.694717

Ripley BD (1981) Spatial Statistics. John Wiley & Sons, Hoboken, NJ

Rue H, Held L (2005) Gaussian Markov Random Fields: Theory and Applications. Chapman & Hall/CRC Press, Boca Raton, FL

Simpson D, Rue H, Riebler A, Martins TG, Sørbye SH (2017) Penalising model component complexity: A principled, practical approach to constructing priors. Statistical Science 32(1):1–28

Tawn J, Shooter R, Towe R, Lamb R (2018) Modelling spatial extreme events with environmental applications. Spatial Statistics 28:39–58

Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. Economic Geography 46(suppl):234–240

Upton G, Fingleton B (1985) Spatial Data Analysis by Example, Volume 1: Point Pattern and Quantitative Data. John Wiley & Sons, Hoboken, NJ

Wikle CK, Zammit-Mangion A, Cressie N (2019) Spatio-Temporal Statistics with R. Chapman & Hall/CRC Press, Boca Raton, FL

Winkler G (2003) Image Analysis, Random Fields and Markov Chain Monte Carlo Methods: A Mathematical Introduction, 2nd edn. Springer-Verlag, Berlin

Zammit-Mangion A, Cressie N (2021) FRK: An R package for spatial and spatio-temporal prediction with large datasets. Journal of Statistical Software, in press