Chapter Two Notes

- Recall the general spatial model or process:

$$\left\{ Z(\mathbf{s}) : \mathbf{s} \in D \subset \mathcal{R}^d \right\}.$$

  The spatial process is a collection of random variables indexed by spatial coordinates, i.e. it is a stochastic process.

- Generally the random variables are correlated with one another, at least over short distances, and the correlation structure is specified by a probabilistic mechanism. The set of random variables and the specification of the correlation structure is called a *random function*: a collection (possibly uncountable) of spatially indexed random variables with a dependence structure specified by a probability model.

- The idea of a random function is different from what you are probably used to in standard data analysis classes. Here is an example from Schabenberger and Gotwwa in a geostatistical setting. You walk into a room and observed a pile of sand on top of a table. The surface of the pile is not level but has peaks and valleys. A set of locations $\mathbf{s}_1, \cdots, \mathbf{s}_n$ is chosen (not necessarily at random) and the depth $Z(\mathbf{s}_i), i = 1, \cdots, n$ is measured. We consider the spatial observations $Z(\mathbf{s}_i), i = 1, \cdots, n$ not as a sample of size n from some probability distribution but as a partially observed realization of a single surface generated by the random function. In other words, we observe only part of a sample of size 1.

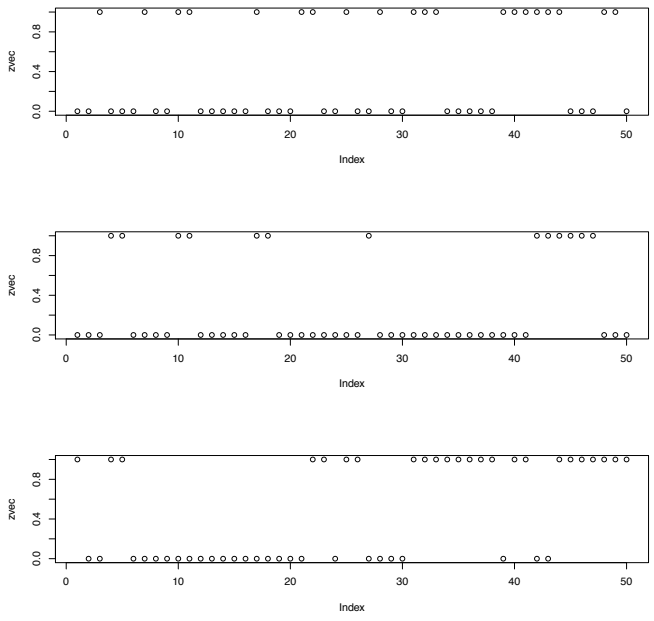
  What does it mean to talk about the mean depth at location $\mathbf{s}_i$, i.e $E[Z(\mathbf{s}_i)] = \mu(\mathbf{s}_i)$? What probability distribution are we taking expectation with respect to? Conceptually we imagine walking into the room over and over again and observing different piles of sand produced by the probabilistic process. We go to location $\mathbf{s}_i$ on each occasion and measure the depth. The long run average of these depths is $\mu(\mathbf{s}_i)$. In the words of Schabenberger and Gotway: "the expectation tells us that we need to repeat the process of pouring sand over and over again and consider the expected value with respect to the probability distribution of all *surfaces* so generated" (emphasis added).

  Lattice Process Example: This example is from a text by Isaaks and Srivastava (An Introduction to Applied Geostatistics). We suppose a set of equally spaced points in one dimension. Define

$$\begin{aligned} Z(0) &= \begin{cases} 0 & \textit{with probability } 1/2 \\ 1 & \textit{with probability } 1/2 \end{cases} \\ Z(s) &= \begin{cases} Z(s-1) & \textit{with probability } 3/4 \\ 1 - Z(s-1) & \textit{with probability } 1/4 \end{cases} \end{aligned}$$

  Each random variable ($Z(s)$) is binary (0 or 1). We have an equal probability of starting with a 0 or 1. At each following location we stay at the same value with

probability 0.75 *and switch with probability* 0.25. *The plot below shows 3 realizations of this random function.*



The $Z(s) \sim$ *Bernoulli*$(1/2)$ *for* $s = 0, 1, 2, \cdots$. *We start at* 0 *or* 1 *with equal probability and at location* s *over the long run we have as much chance of seeing a* 0 *as we do a* 1. *However, the random variables are not independent.*

The pairs $(Z(s), Z(s+1))$ *have the same joint probability distribution. The set of possible outcomes is*

$$\{(0,0),(0,1),(1,0),(1,1)\}$$

with associated probabilities of

$$\{3/8, 1/8, 1/8, 3/8\}$$

clearly indicating the tendency for $Z(s+1)$ *to remain at the value observed for* $Z(s)$.

The random function defined above generates joint distributions for collections of Z. *In particular, it generates joint distributions for the pairs* $(Z(s), Z(s+h))$ *for* $h = 1, 2, \cdots$.

Of course, in practice we do not observe the random function. At best, we will observe a single realization of the function and at worst we will observe only a part of a single realization. We must make strong assumptions to proceed with statistical analysis.

- Stationarity Assumptions

– Strict (or strong) Stationarity: A spatial process

$$\left\{Z(\mathbf{s}) : \mathbf{s} \in D \subset \mathcal{R}^d\right\}$$

satisfies the assumption of strict stationarity if

$$P\left(Z\left(\mathbf{s}_1\right) < z_1, Z\left(\mathbf{s}_2\right) < z_2, \cdots, Z\left(\mathbf{s}_k\right) < z_k\right) =$$

$$P\left(Z\left(\mathbf{s}_1 + \mathbf{h}\right) < z_1, Z\left(\mathbf{s}_2 + \mathbf{h}\right) < z_2, \cdots, Z\left(\mathbf{s}_k + \mathbf{h}\right) < z_k\right)$$

for all k and $\mathbf{h}$.

– Second-order (weak) stationarity: A spatial process

$$\left\{Z(\mathbf{s}) : \mathbf{s} \in D \subset \mathcal{R}^d\right\}$$

is said to be second-order stationary if

$$E\left[Z\left(\mathbf{s}\right)\right] = \mu, \text{ for all } \mathbf{s} \in D$$

and

$$\text{Cov}\left[Z\left(\mathbf{s}\right), Z\left(\mathbf{s} + \mathbf{h}\right)\right] = C\left(\mathbf{h}\right).$$

In words, the mean of the random field is a constant and the covariance function $C\left(\mathbf{h}\right)$ is a function only of the separation (or lag) vector $\mathbf{h}$. One consequence of second-order stationarity is that

$$\text{Var}\left[Z\left(\mathbf{s}\right)\right] = \text{Cov}\left[Z\left(\mathbf{s}\right), Z\left(\mathbf{s}\right)\right] = C\left(\mathbf{0}\right)$$

exists and is constant over the random field.

– $C\left(\mathbf{s}_i - \mathbf{s}_j\right)$ is a covariance function if and only if C is positive definite:

$$\sum_{i=1}^{k}\sum_{j=1}^{k} a_i a_j C\left(\mathbf{s}_i - \mathbf{s}_j\right) \geq 0$$

for any finite collection of spatial locations $\mathbf{s}_i, i = 1, \cdots, k$ and real numbers $a_i, i = 1, \cdots, k$.

– Covariance functions of second-order stationary spatial processes satisfy the following properties:

  * $C\left(\mathbf{0}\right) \geq 0$
  * $C\left(\mathbf{h}\right) = C\left(-\mathbf{h}\right)$
  * $C\left(\mathbf{0}\right) \geq \left|C\left(\mathbf{h}\right)\right|$
  * $C\left(\mathbf{h}\right) = \text{Cov}\left[Z\left(\mathbf{s}\right), Z\left(\mathbf{s} + \mathbf{h}\right)\right] = \text{Cov}\left[Z\left(\mathbf{0}\right), Z\left(\mathbf{h}\right)\right]$

* If $C_j(\mathbf{h}), j = 1, \cdots, k$ are covariance functions then

$$\sum_{j=1}^{k} b_j C_j(\mathbf{h})$$

is a covariance function if all $b_j \geq 0$.

* If $C_j(\mathbf{h}), j = 1, \cdots, k$ are covariance functions then

$$\Pi_{j=1}^{k} C_j(\mathbf{h})$$

is a covariance function.

* If $C(\mathbf{h})$ is a covariance function in $\mathcal{R}^d$ it is a covariance function in $\mathcal{R}^p$ for $p < d$.

– Intrinsic Stationarity: A spatial process

$$\left\{Z(\mathbf{s}) : \mathbf{s} \in D \subset \mathcal{R}^d\right\}$$

satisfies the assumption of intrinsic stationarity if

$$E[Z(\mathbf{s})] = \mu$$

and

$$\frac{1}{2}\mathrm{Var}[Z(\mathbf{s}) - Z(\mathbf{s}+\mathbf{h})] = \gamma(\mathbf{h}).$$

Intrinsic stationarity implies a constant mean spatial process and a semivariogram that depends only on $\mathbf{h}$. $\gamma(\mathbf{s}_i - \mathbf{s}_j)$ is a semivariogram of an intrinsically stationary process if and only if it is conditionally negative definite:

$$\sum_{i=1}^{k}\sum_{j=1}^{k} a_i a_j \gamma(\mathbf{s}_i - \mathbf{s}_j) \leq 0$$

for any finite collection of spatial locations $\mathbf{s}_i, i = 1, \cdots, k$ and real numbers $a_i, i = 1, \cdots, k$ such that $\sum_{i=1}^{k} a_i = 0$.

– Some properties of semivariograms of second order stationary processes can be deduced from the properties of covariance functions listed above, e.g. γ is a symmetric function and $\gamma(\mathbf{0}) = 0$.

- Strict stationarity implies second-order stationarity and intrinsic stationarity. Second-order stationarity implies intrinsic stationarity. However, intrinsic stationarity does not imply second-order stationarity or strict stationarity and second-order stationarity does not imply strict stationarity except for one special case (see below).

- Isotropy: If the covariance function and semivariogram depend only on the Euclidean distance between 2 points $\mathbf{s}$ and $\mathbf{s}+\mathbf{h}$ ($\|\mathbf{h}\|$) then the functions (and the associated spatial process) are said to be *isotropic*. Otherwise the process is *anisotropic*, i.e. the functions depend on direction as well as the distance separating 2 locations. We write $C(\|\mathbf{h}\|)$ and $\gamma(\|\mathbf{h}\|)$ to denote isotropic covariance functions and semivariograms, respectively. Technically, $C(\mathbf{h})$ and $C(\|\mathbf{h}\|)$ are 2 different functions for processes in dimensions $d \geq 2$ but we will not worry about those details.

- Gaussian Random Fields: A random field

$$\left\{Z(\mathbf{s}) : \mathbf{s} \in D \subset \mathcal{R}^d\right\}$$

  is a Gaussian random field (GRF) if $Z(\mathbf{s}_i), i = 1, \cdots, k$ have a multivariate Gaussian distribution with means $\mu(\mathbf{s}_i)$ and variances $\sigma^2(\mathbf{s}_i)$, respectively for all locations $\mathbf{s} \in D$. This is the one case in which second-order stationarity implies strict stationarity because the means and variances completely determine the multivariate Gaussian distribution.

Spatial Continuity and Differentiability

- This material is a summary of a section in Chapter 2 in Schabenberger and Gotway and is germane to continuous spatial processes, i.e. the Z's are continuous random variables in a continuous spatial domain D.

- The continuity and differentiability properties of a spatial process are determined by its correlation structure. In particular, the behavior of C and γ near the origin are informative about the continuity properties of Z. This is an important consideration because statistical modelling will ultimately require specification of a correlation (covariance, semivariogram) model and choice of such a model is also an assumption about the continuity properties of the spatial process.

- The discussion in Section 2.3 is a bit confusing. We will not worry too much about the details but we will try to clarify a few things. A second order stationary process $Z(\mathbf{s})$ is said to be *mean square continuous* if

$$\lim_{\mathbf{h}\to\mathbf{0}} E\left[(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s}))^2\right] = 0.$$

  Because

$$E\left[(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s}))^2\right] = 2C(\mathbf{0}) - 2C(\mathbf{h}) = 2\gamma(\mathbf{h})$$

  we can conclude that a second-order stationary process is mean square continuous if and only if C (and γ) are continuous at the origin.

  Discontinuous processes will behave erratically. The discontinuity of C and γ at the origin is manifest by the nugget effect. As the text points out this is bothersome to

many who work in this area. For example Yaglom (1987) writes: "Such discontinuous random processes must naturally behave extremely irregularly and can hardly serve as a suitable model for any reasonable applied problem." He then spends the rest of a 2 volume treatise on random functions ignoring such processes.

- The text briefly discusses mean square differentiability but the discussion is confusing and incorrect in a few places. A one dimensional process is said to be *mean square differentiable* if

$$Z^{'}(s) = \lim_{h \to 0} \frac{Z(s+h) - Z(s)}{h}$$

exists and $Z^{'}$ is the *mean square derivative*. Yaglom (1987), Stein (1999), and others provide proofs of the following results:

  - The mean square derivative will exist if and only if $C^{"}(0)$ (and $\gamma^{"}(0)$) is finite. It can be shown that $-C^{"}(h)$ is the covariance function of the mean square derivative $Z^{'}(h)$.
  - Existence of the mean square derivative requires

$$\lim_{h \to \infty} \frac{2\gamma(h)}{h^2} = 0.$$

These results can be extended to higher dimensions, although it is not all that straightforward. They also hold for intrinsically stationary processes that are not second order stationary (although the results must be stated in terms of the variogram 2γ). The multiple dimension version of the second point above is that, for isotropic semivariograms

$$\lim_{\|\mathbf{h}\| \to \infty} \frac{2\gamma(\mathbf{h})}{\|\mathbf{h}\|^2} = 0.$$

The example given in the text is of the power semivariogram model:

$$\gamma(\mathbf{h};\ \boldsymbol{\theta}) = \begin{cases} 0 & \mathbf{h} = \mathbf{0} \\ \theta_1 + \theta_2 \|\mathbf{h}\|^{\theta_3} & \mathbf{h} \neq \mathbf{0} \end{cases}$$

This is a semivariogram in all dimensions $d \geq 1$ for $0 < \theta_3 < 2$. One thing to note about this semivariogram is that it is not the semivariogram of a second-order stationary process because the sill does not exist, i.e. the process does not have a variance.

Random Fields in the Spatial Domain

- Schabenberger and Gotway discuss 2 representations of spatial processes considered in the spatial domain: the model representation and the convolution representation. We will only consider the first. The convolution representation is not addressed much in the rest of this text and is addressed not at all in most other texts that deal with the statistical analysis of spatial data.

- We have seen that fairly strong assumptions are required for statistical analysis of data generated by a spatial process: second order and/or intrinsic stationarity. One of the most restrictive of the assumptions is that

$$E\left[Z\left(\mathbf{s}\right)\right]=\mu.$$

This assumption in particular seems unreasonable, i.e. it seems likely that the mean will depend on $\mathbf{s}$,

$$E\left[Z\left(\mathbf{s}\right)\right]=\mu\left(\mathbf{s}\right).$$

- A standard approach is to consider $\mu\left(\mathbf{s}\right)$ to be a large scale *trend* surface. The trend is modelled and removed and the remaining error terms (or residuals) are assumed to be the (partial) realization of a stationary spatial process with mean 0. Specifically, we assume

$$\mathbf{Z}\left(\mathbf{s}\right)=\mathbf{f}\left(\boldsymbol{X},\mathbf{s},\boldsymbol{\beta}\right)+e\left(\mathbf{s}\right)$$

where

$$\mathbf{Z}\left(\mathbf{s}\right)=\left[Z\left(\mathbf{s}_1\right),Z\left(\mathbf{s}_2\right),\cdots,Z\left(\mathbf{s}_n\right)\right]',$$

$$\mathbf{f}\left(\boldsymbol{X},\mathbf{s},\boldsymbol{\beta}\right)=\left[\begin{array}{c} f\left(\mathbf{x}_1,\mathbf{s}_1,\boldsymbol{\beta}\right) \\ f\left(\mathbf{x}_2,\mathbf{s}_2,\boldsymbol{\beta}\right) \\ \vdots \\ f\left(\mathbf{x}_n,\mathbf{s}_n,\boldsymbol{\beta}\right) \end{array}\right],$$

$\boldsymbol{\beta}$ is a vector of parameters, and $\mathbf{e}\left(\mathbf{s}\right)$ is a random vector with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}\left(\boldsymbol{\theta}\right)$. $\boldsymbol{\theta}$ indicates that the covariance model will depend on unknown parameters that must be estimated. This is a general expression allowing for possibly nonlinear relationships between the response variable and its associated covariates. It includes linear models and generalized linear models (binary regression models, count regression models, etc.). The mean structure of the model

$$E\left(\mathbf{Z}(\mathbf{s})\right)=\boldsymbol{\mu}(\mathbf{s})$$

represents trend. Variance/covariance and stationarity properties are made for the error terms $\boldsymbol{\epsilon}(\mathbf{s})$.

- Geostatistical Data: We consider a model

$$\begin{aligned} Z\left(\mathbf{s}\right) &= \mu\left(\mathbf{s}\right)+e\left(\mathbf{s}\right) && (1)\\ &= \mu\left(\mathbf{s}\right)+W\left(\mathbf{s}\right)+\eta\left(\mathbf{s}\right)+\epsilon\left(\mathbf{s}\right) && (2)\end{aligned}$$

where

  - $\mu(\mathbf{s})\equiv E\left[Z(\mathbf{s})\right]$ represents large scale variation (trend).

- $W(\mathbf{s})$ is a 0-mean, mean-square continuous, intrinsically stationary process. The range of the semivariogram γ_W may not exist (i.e. W is not necessarily second order stationary). If the range exists it is larger than $\min\{\|\mathbf{s}_i - \mathbf{s}_j\|\} : 1 \leq i < j \leq n$. W represents smooth small scale variation.
- $\eta(\mathbf{s})$ is a 0-mean, intrinsically stationary process, uncorrelated with W whose variogram range exists and is smaller than $\min\{\|\mathbf{s}_i - \mathbf{s}_j\|\} : 1 \leq i < j \leq n$. This represents micro-scale variation.
- $\epsilon(\mathbf{s})$ is a 0-mean white noise process, uncorrelated with W and η which represents measurement error.

Without replicate observations we cannot estimate σ^2_ϵ and it follows that σ^2_η and σ^2_ϵ are not estimable. The model that is generally considered is then

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + W(\mathbf{s}) + \epsilon^*(\mathbf{s})$$

where ϵ^* consists of micro-scale and measurement error variation.

Two types of models are defined.

- Signal Model: $Z(\mathbf{s}) = S(\mathbf{s}) + \epsilon(\mathbf{s})$ where the *signal* S is a noiseless version of Z. Engineers refer to this type of model as a state process.
- Mean Model: $Z(\mathbf{s}) = \mu(\mathbf{s}) + e(\mathbf{s})$ which is the type of model you are more familiar with. Generally, we cannot say much about W and η aside from recognizing that they are there. Some of the local scale variation represented by those terms may, in fact, be at least partially accounted for by a flexible mean function. A natural question that arises then is when to account for the *correlation structure* in the mean function, when to account for it in the error structure, and when to possibly do a bit of both. This is the source of Cressie's now semi-classical comment: "one modeler's mean function is another modeler's covariance structure."

- Schabenberger and Gotway provide a bit of advice on this last question using Cliff and Ord's characterization of reaction and interaction models. Impacts attributable to external factors (e..g. elevation) are naturally accounted for in the mean structure. Impacts attributable to interactions with neighbors should be accounted for in the covariance structure. But as always, rules such as this are made to be broken.

- Lattice Data: This is the realm of spatial auto-regression models. Define Z by

$$Z(\mathbf{s}_i) = \sum_{j=1}^{n} b_{ij}\left(Z(\mathbf{s}_j) - \mu(\mathbf{s}_i)\right) + \epsilon(\mathbf{s}_i)$$

where b_{ij} is generally taken to be a contiguity measure, in which case the matrix $\mathbf{B}$ will either the same as or proportional to the contiguity matrix $\mathbf{W}$ we dealt with

in Chapter 1. As an extension to this model, which incorporates covariates, we suppose

$$(\mathbf{I} - \mathbf{B})(\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\boldsymbol{\beta}) = \boldsymbol{\epsilon}(\mathbf{s}).$$

If $\text{Var}[\boldsymbol{\epsilon}(\mathbf{s})] = \sigma^2\mathbf{I}$ then $\text{Var}[\mathbf{Z}(\mathbf{s})] = \sigma^2(\mathbf{I} - \mathbf{B})^{-1}\left(\mathbf{I} - \mathbf{B}'\right)^{-1}$. Even though the error terms are assumed to be uncorrelated the Z's are not, specification of $\mathbf{B}$ *induces* a correlation structure among the Z's. There is another type of autoregression model know as as a Conditional Autoregression Model (CAR) as opposed to the Simultaneous Autoregression Model (SAR) just outlined.

- We will have more to say about these various models later in the course.