Geostatistical Data

Recall the model $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$, where

- $\mathbf{s} = (x, y)$ denotes the coordinates of the sample site. Here (x, y) may be Euclidean coordinates or latitude and longitude.
- $Z(\mathbf{s})$ denotes the variable of interest at the location $\mathbf{s}$. Note that this is written as a function of the location $\mathbf{s}$.
- D denotes the region of interest, which contains an (uncountably) infinite number of sites.
- Observations can only be taken on a finite collection of sample sites $\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_n$.
- Geostatistical data are continuous spatial data; i.e., between any two sites in D, we can find another site in D.

Model & Assumptions

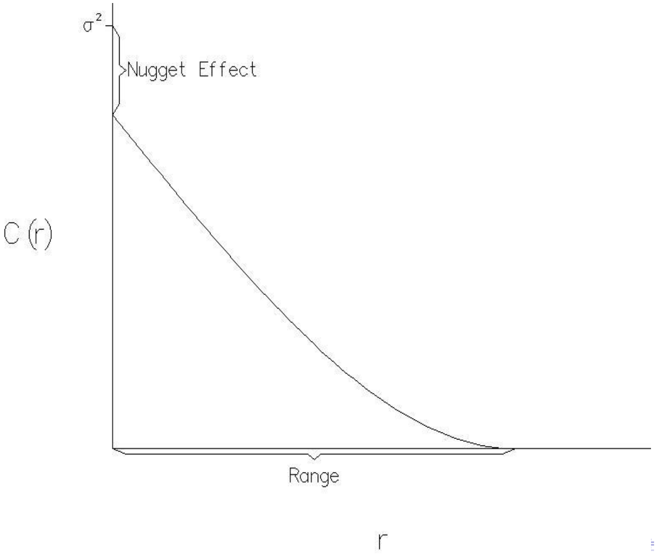
Consider the simple model $Z(\mathbf{s}) = \mu + \varepsilon(\mathbf{s})$

- μ is the population mean.
- $\varepsilon(\mathbf{s})$ is a zero-mean random error the spatial location $\mathbf{s}$.
  - $E\left\{\varepsilon\left(\mathbf{s}\right)\right\} = 0;\ \mathbf{s} \in D$.
  - $var\left\{\varepsilon\left(\mathbf{s}\right)\right\} = \sigma^2;\ \mathbf{s} \in D$.
  - $C(\mathbf{s} - \mathbf{u}) = cov\left\{\varepsilon\left(\mathbf{s}\right), \varepsilon\left(\mathbf{u}\right)\right\};\ \mathbf{s}, \mathbf{u} \in D$ only depends on the difference in the locations (distance and direction) of the pair of sites $\mathbf{s}, \mathbf{u} \in D$.
- $Z\left(\mathbf{s}\right)$ has the same variance and covariance. $\left\{Z\left(\mathbf{s}\right) : \mathbf{s} \in D\right\}$ is said to be *second-order stationary*.
- $\left\{Z\left(\mathbf{s}\right) : \mathbf{s} \in D\right\}$ is *isotropic* if $C(\mathbf{s} - \mathbf{u}) = C(||\mathbf{s} - \mathbf{u}||)$

Features of the Covariance Function

- **Range:** r_0 is the range if $C(r) = 0$ for all $r \geq r_0$.
- Pairs of sites further than r_0 apart are uncorrected.
- **Nugget Effect:** $\sigma^2 - \lim_{r \to 0} C(r)$, which may be attributed to:
  - Microscale variation: Variation at spatial scales shorter than that separating the sample sites;
  - Measurement error: Variation due to errors in measuring the variable.

A Typical Plot of Covariance Function



Effect of Spatial Dependence on Estimation

Consider a stationary time series $Z_1, \cdots, Z_n$ with mean μ and covariance function $C(h) = cov(Z_i, Z_{i+h}) = \sigma^2\rho(h)$.

- $\bar{Z}$ is an unbiased estimator for μ.
- If data were independent, $var(\bar{Z}) = \sigma^2/n$.
- If the data are not independent, then

$$var(\bar{Z}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} cov(Z_i, Z_j) = \frac{\sigma^2}{n}\left\{1 + \frac{2}{n}\sum_{h=1}^{n-1}(n-h)\rho(h)\right\}$$

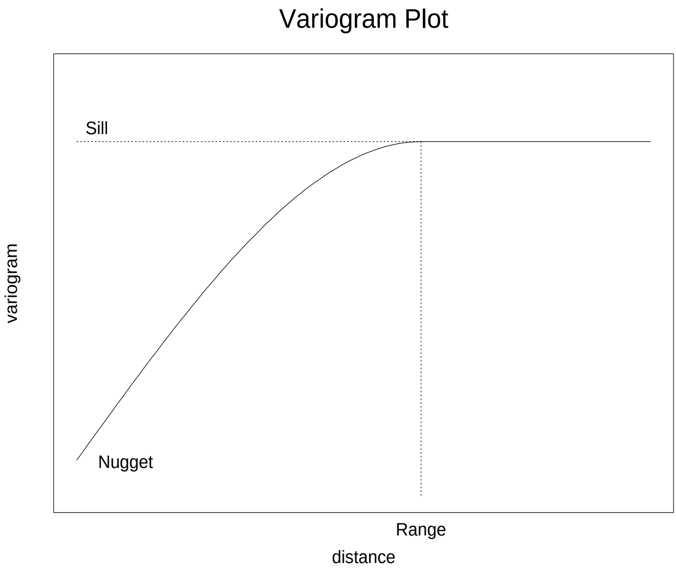
- As $n \to \infty$, $n \times var(\bar{Z}) \to \sigma^2 \sum_{h=-\infty}^{\infty} \rho(h) \gg \sigma^2$.
- Correlation is bad for estimation and inference, but as we will see, it is (not so surprisingly) good for prediction!

Variogram

- The *variogram* is defined to be

$$2\gamma(\mathbf{s}-\mathbf{u}) = var\{Z(\mathbf{s}) - Z(\mathbf{u})\}.$$

- **Range:** Range of spatial correlation.
- **Nugget Effect:** The nugget effect is due to microscale variation (variation between locations closer together than the sample sites) and/or measurement error.
- **Sill:** The sill is equal to $2\sigma^2$, and so measures the variability in the data.


Variogram Plot
Sill
Nugget
variogram
Range
distance

Intrinsic Stationarity

- A random field is *intrinsic* if

$$E\left\{Z(\mathbf{s}) - Z(\mathbf{u})\right\} = 0;$$

$$var\left\{Z(\mathbf{s}) - Z(\mathbf{u})\right\} = 2\gamma(\mathbf{s} - \mathbf{u}) \;\; \mathbf{s}, \mathbf{u} \in D.$$

- Intrinsic stationarity allows for the possibility that $\sigma^2 = \infty$. In this case, the covariance function is undefined.
- All second-order stationary random fields are intrinsic, but intrinsic random fields need not be second-order stationary.
- If Z is second-order stationary, then $2\gamma(\mathbf{h}) = 2\sigma^2 - 2C(\mathbf{h})$.
- If Z is isotropic, then $2\gamma(\|\mathbf{s} - \mathbf{u}\|) = var\left\{Z(\mathbf{s}) - Z(\mathbf{u})\right\}$.

Conditions for Valid Variogram

- Consider linear combinations of the form

$$Y = \sum_{i=1}^{m} a_i Z(\mathbf{u}_i),$$

where $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_m$ is any finite collection of sites, and $a_1, a_2, \cdots, a_n$ is any finite collection of constants.

- The variance of Y is given by

$$var(Y) = var\left(\sum_{i=1}^{m} a_i Z(\mathbf{u}_i)\right) = \sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j C(\mathbf{u}_i - \mathbf{u}_j)$$

- For a valid model, we require $var(Y) \geq 0$.

- **Definition:** A function $C(\cdot)$ is *positive definite* if

$$\sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j C(\mathbf{u}_i - \mathbf{u}_j) \geq 0.$$

- **Bochner's Theorem:** The function $C(\cdot)$ is positive definite if

$$C(\mathbf{h}) = \sigma^2 \int \cos(\mathbf{h}'\omega)\, f(\omega)\, d\omega$$

  where $f(\cdot)$ is a probability density function (called the *spectral density function*).
- **Definition:** A function $C(\cdot)$ is a *valid* covariance function if and only if it is positive definite, i.e. we can find a density function $f(\cdot)$ such that $C(\cdot)$ satisfies the above expression.

- **Definition:** A function $2\gamma(\cdot)$ is a *valid* variogram if an only if it is conditionally negative definite; that is,

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

given that $\sum_{i=1}^{m} a_i = 0$.
- If $C(\cdot)$ is a valid covariance function, then $2\gamma(\mathbf{h}) = 2\sigma^2 - 2C(\mathbf{h})$ is a valid variogram.
- Bochner's Theorem has been used to verify the validity of many covariance and variogram models.

Variogram Models

- Power variogram. $2\gamma(r) = c_0 + ar^{\alpha}$.
  - c_0 is the nugget effect.
  - The power variogram has no sill, and so the variance of the process is infinite.
  - Often due to existence of trend.
- Spherical variogram

$$2\gamma(r) = \begin{cases} c_0 + c_s \left\{ \frac{3}{2}\left(\frac{r}{a_s}\right) - \frac{1}{2}\left(\frac{r}{a_s}\right)^3 \right\} & ; \quad 0 < r \leq a_s \\ c_0 + c_s & ; \quad r \geq a_s \end{cases}$$

  - c_0 is the nugget effect.
  - a_s is the range of spatial correlation
  - $c_0 + c_s$ is the sill.

- Matérn Class of Variograms

$$2\gamma(r) = c_0 + c_1 \left(1 - \frac{(r/2\alpha)^\nu}{2\Gamma(\nu)} K_\nu(r/\alpha)\right)$$

  - $K_\nu(\cdot)$ is a modified Bessel function of the second kind.
  - The nugget effect is c_0, and the sill is $c_0 + c_1$.
  - ν controls the smoothness of the random field.

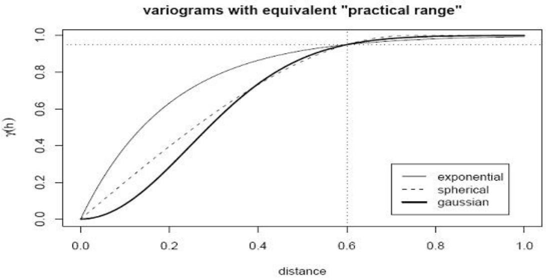
- Exponential variogram ($\nu = .5$)

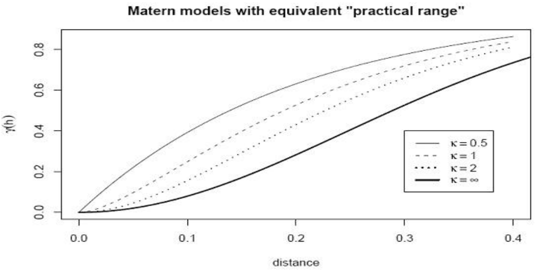
$$2\gamma(r) = c_0 + c_1(1 - \exp\{-r/\alpha\})$$

- Gaussian variogram ($\nu = \infty$)

$$2\gamma(r) = c_0 + c_1(1 - \exp\{-(r/\alpha)^2\})$$

- The (practical) range are 3α and $\sqrt{3}\alpha$, respectively.


variograms with equivalent "practical range"
γ(h)
0.0
0.2
0.4
0.6
0.8
1.0
distance
exponential
spherical
gaussian


Matern models with equivalent "practical range"
γ(h)
0.0
0.2
0.4
0.6
0.8
0.1
0.3
distance
κ = 0.5
κ = 1
κ = 2
κ = ∞

MOM Variogram Estimator

Recall: Definition of the variogram of a random field

$$2\gamma(\mathbf{s}-\mathbf{u}) = var\left\{Z(\mathbf{s}) - Z(\mathbf{u})\right\}; \;\; \mathbf{s}, \mathbf{u} \in D.$$

Note that if $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$ is an intrinsic stationary random field, then

$$2\gamma(\mathbf{s}-\mathbf{u}) = E\left\{|Z(\mathbf{s}) - Z(\mathbf{u})|^2\right\}$$

This suggests the following estimator for the variogram:

$$2\widehat{\gamma}(\mathbf{h}) = \frac{1}{N_{\mathbf{h}}} \sum |Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^2,$$

where

- the sum is over all pairs of sites **h** apart;
- $N_{\mathbf{h}}$ is the number of such pairs of sites.

Equivalently, we may write

$$2\widehat{\gamma}(\mathbf{h}) = \frac{\sum_{i<j} \left| Z\left(\mathbf{s}_i\right) - Z\left(\mathbf{s}_j\right)\right|^2 I\left(\mathbf{s}_i - \mathbf{s}_j = \mathbf{h}\right)}{\sum_{i<j} I\left(\mathbf{s}_i - \mathbf{s}_j = \mathbf{h}\right)},$$

where

$$I\left(\mathbf{s}_i - \mathbf{s}_j = \mathbf{h}\right) = \begin{cases} 1; & \text{if } \mathbf{s}_i - \mathbf{s}_j = \mathbf{h} \\ 0; & \text{if } \mathbf{s}_i - \mathbf{s}_j \neq \mathbf{h} \end{cases}$$

For isotropic random fields, estimate

$$\begin{aligned} 2\widehat{\gamma}\left(r\right) &= \frac{1}{N_r} \sum \left| Z\left(\mathbf{s}_i\right) - Z\left(\mathbf{s}_j\right)\right|^2 \\ &= \frac{\sum_{i<j} \left| Z\left(\mathbf{s}_i\right) - Z\left(\mathbf{s}_j\right)\right|^2 I\left(\left\|\mathbf{s}_i - \mathbf{s}_j\right\| = r\right)}{\sum_{i<j} I\left(\left\|\mathbf{s}_i - \mathbf{s}_j\right\| = r\right)} \end{aligned}$$

where the sum is over all pairs of sites distance r apart and N_r is the number of such pairs of sites.

Properties of MOM Variogram Estimator

Assume that $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$ is intrinsically stationary.

1. Unbiasedness: $E\{2\widehat{\gamma}(\mathbf{h})\} = 2\gamma(\mathbf{h})$.
2. Consistency: if $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$ is ergodic, then $2\widehat{\gamma}(\mathbf{h}) \rightarrow 2\gamma(\mathbf{h})$ almost surely as $n \rightarrow \infty$ and $D \uparrow \mathrm{R}^d$.
3. Asymptotic normality: $2\widehat{\gamma}(\mathbf{h})$ converges to a normal distribution as $n \rightarrow \infty$ and $D \uparrow \mathrm{R}^d$.
   - For Gaussian processes, Cressie (1985, *Mathematical Geology* **17,** 563) gives the variance-covariance matrix of $\{2\widehat{\gamma}(\mathbf{h})\}$.
   - For non-Gaussian processes, resampling methods can be used (Guan, Sherman and Calvin, 2004) to estimate asymptotic the variance/covariance.

Some Notes

- If data are on an irregular lattice, then the variogram can be estimated by a smoothing method, in which case the estimator will be asymptotically unbiased.
- Covariance can be estimated analogously.
- Robust variogram estimators are available for Gaussian data.
- The form of asymptotics is called increasing-domain asymptotics, i.e., the study region becomes increasingly large, in contrast to the infill asymptotics, i.e., increasingly dense samples are taken from a fixed study region.