Introduction to Spatial Statistics

Bin Li

IIT Lecture Series

M.A. Oliver and R. Webster (2014)

Catena 113 (2014) 56–69

Contents lists available at ScienceDirect

Catena

journal homepage: www.elsevier.com/locate/catena

A tutorial guide to geostatistics: Computing and modelling variograms and kriging

M.A. Oliver [a], R. Webster [b,*]

[a] Soil Research Centre, Department of Geography and Environmental Science, University of Reading, Reading RG6 6DW, Great Britain, United Kingdom
[b] Rothamsted Research, Harpenden AL5 2JQ, Great Britain, United Kingdom

ARTICLE INFO

Article history:
Received 4 June 2013
Received in revised form 18 September 2013
Accepted 22 September 2013

Keywords:
Geostatistics
Sampling
Variogram
Model fitting
Trend
Kriging

ABSTRACT

Many environmental scientists are analysing spatial data by geostatistical methods and interpolating from sparse sample data by kriging to make maps. They recognize its merits in providing unbiased estimates with minimum variance. Several statistical packages now have the facilities they require, as do some geographic information systems. In the latter kriging is an option for interpolation that can be done at the press of a few buttons. Unfortunately, the ease conferred by this allows one to krige without understanding and to produce unreliable and even misleading results. Crucial for sound kriging is a plausible function for the spatial covariances or, more widely, of the variogram. The variogram must be estimated reliably and then modelled with valid mathematical functions. This requires an understanding of the assumptions in the underlying theory of random processes on which geostatistics is based. Here we guide readers through computing the sample variogram and modelling it by weighted least-squares fitting. We explain how to choose the most suitable functions by a combination of graphics and statistical diagnostics. Ordinary kriging follows straightforwardly from the model, but small changes in the model function and its parameters can affect the kriging error variances. When kriging is automated these effects remain unknown. We explain the choices to be made when kriging, i.e. whether the support is at points or over blocks, and whether the predictions are global or within moving windows.

Random process

- Features of the environment, such as soil, are the product of many physical, chemical and biological processes.
- These processes are physically determined, but their interactions are so complex that the variation appears to be random.
- Two common approaches:
  - Treat them as random processes and fit statistical models such as *kriging*.
  - Use geostatistical simulation to approximate these complex processes.
- Here we only focus on the statistical modelling approach, particularly *variogram* and *kriging*.

Stationarity

- Allows us to assume that there is the same degree of variation from place to place.
- A simple model to represent the random process:

$$Z(s) = \mu + \epsilon(s)$$

  - $Z(s)$ is a random variable of interest at place s.
  - μ is the mean of the process.
  - $\epsilon(s)$ is a random error with mean 0 and a **semivariance** $\gamma(h)$.
- Stationary model with *trend* (also called *external drift*): $Z(s) = f(s) + \epsilon(s)$, where $f(s)$ is the trend component.

Semivariance

- *Semivariance* (also called *variogram*) is defined as:

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

- $Z(s)$ and $Z(s+h)$ are values of Z at places s and $s+h$.
- Lag h: separation between samples in both direction and distance.
- Assumption: the semivariance **only** depends on h.
- Isotropic variogram: $\gamma(|h|) = \frac{1}{2} Var[Z(s) - Z(s+h)]$, where $|h|$ is the distance of h.

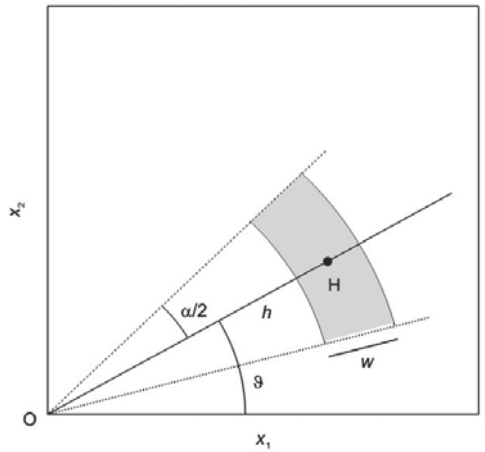
Variogram

- Two types of variogram
  - Theoretical variogram models: e.g. spherical, power, exponential models. Continuous and usually smooth.
  - Sample variogram: estimate of $\gamma(h)$ from the sample data. A finite set of discrete lags (i.e. h).
- Sample variogram can be computed by the method of moments attributed to Matheron (1965):

$$\hat{\gamma}(h) = \frac{1}{2m(h)} \sum_{j=1}^{m(h)} [Z(s_j) - Z(s_j + h)]^2 ,$$

where $m(h)$ is the number of paired samples are lag h.

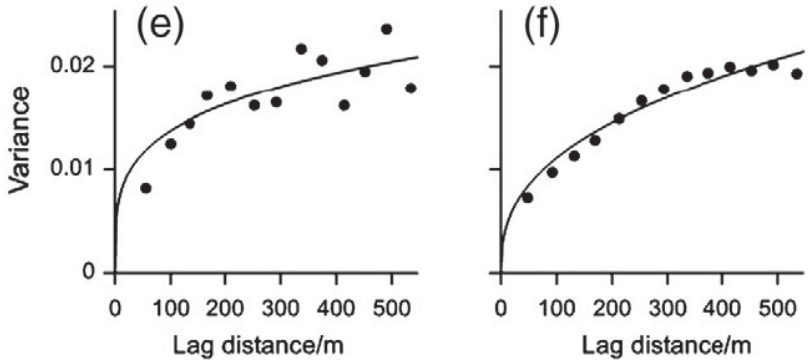
Lag interval and bin width



The geometry in 2D for discretizing the lag into bins by distance and direction. Shaded area is one bin.

Figure from Oliver and Webster (2014).

Variogram (cont.)



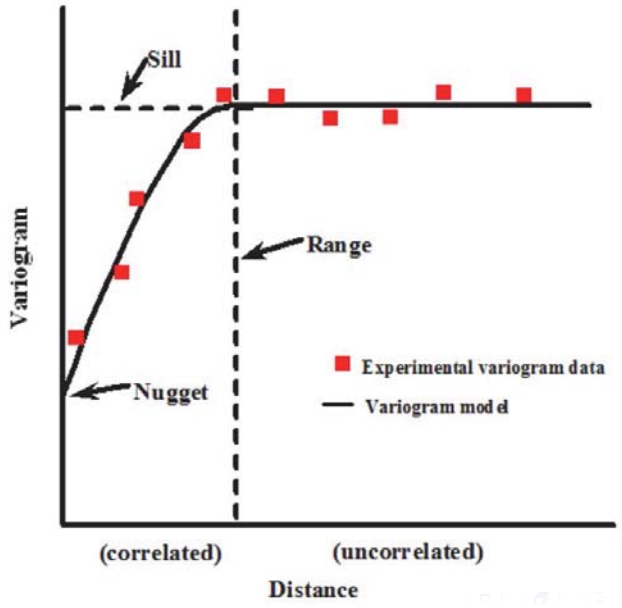
Left(right): computed from 87 (434) samples. Curves: power variogram model.

Figure from Oliver and Webster (2014).

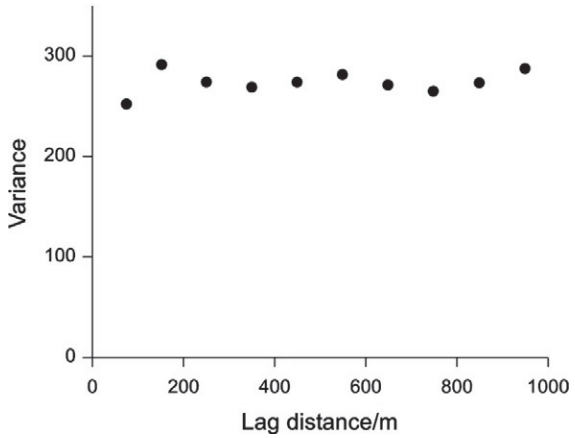
Sill, nugget and range

- Variogram shows the spatial correlation from the collected data or specified model.
- **Range**: the distance where the variogram first flattens out.
  - Samples separated within the range are spatially correlated.
  - Outside range: not spatially correlated.
- **Nugget**: semivariance at distance "zero".
  - Environmental variability at the scale of sampling (geological micro-structure).
  - Error inherent in the measurements (sampling design and sampling unit size).
- **Sill**: the value variogram attains at the range (the value on y-axis).
  - *Partial sill*: Sill minus nugget effect

Illustration of a variogram



Pure nugget Figure from Oliver and Webster (2014).



- No spatial correlation exists;
- sampling interval is larger than the correlation range.

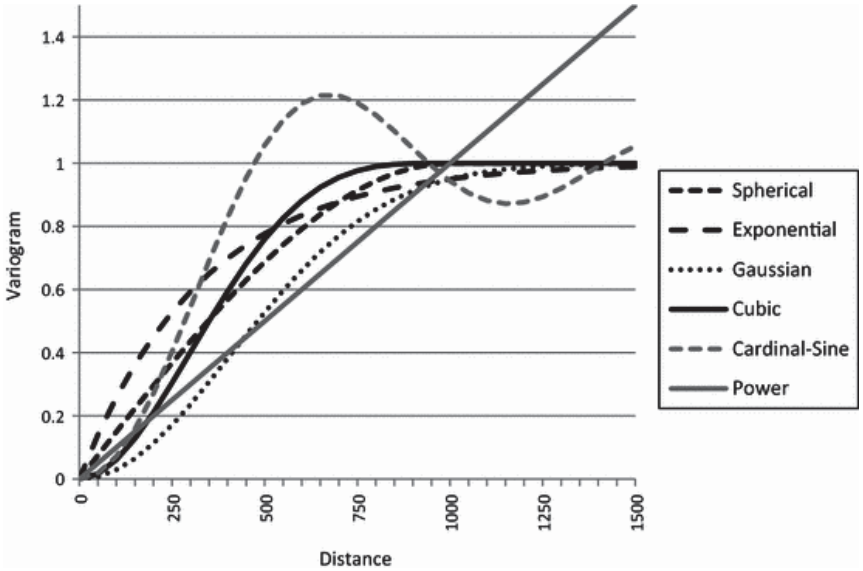
A variogram model

One of the most popular functions is the *isotropic spherical-plus-nugget model*.

$$\begin{aligned}\gamma(h) &= c_0 + c\left\{\frac{3h}{2r} - \frac{1}{2}\left(\frac{h}{r}\right)^3\right\} \text{ for } 0 < h \leq r \\ &= c_0 + c \text{ for } h > r \\ &= 0 \text{ for } h = 0,\end{aligned}$$

- in which $h = |h|$ is the lag distance;
- c_0 is the nugget;
- $c_0 + c$ is the sill;
- c is the partial sill;
- r is the range.

Variogram models



How to model variogram?

- Plot the sample variogram.
- Choose several models that appear to have the right shape and fit each model based on the sampled data.
- Plot the fitted models together with the sample variogram and assess whether the fit looks reasonable.
- If more than one models seem to fit well, then
  - we can choose the one with the smallest RSS;
  - or run CV and choose the one that produce a mean squared error closest to mean kriging variance.
- Spherical, exponential, Gaussian and power models are popular variogram models. In case of having outlying samples, use robust variogram estimators, such as Cressie-Hawkins, Dowd, and Genton estimators.

Two robust variogram estimators

- Cressie-Hawkins estimator (Cressie and Hawkins, 1980).

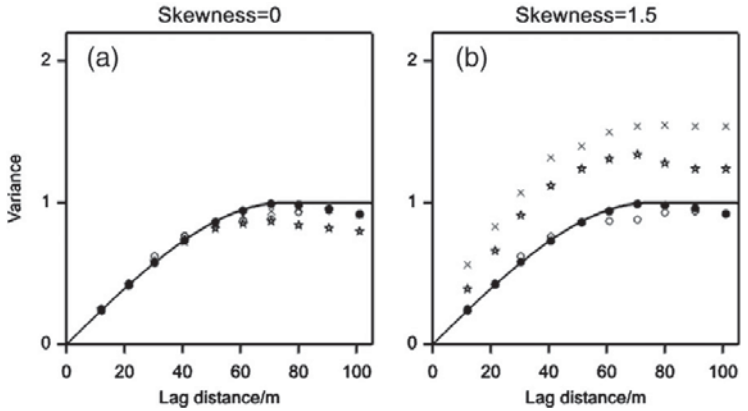
$$2\hat{\gamma}_{CH}(h) = \frac{\left\{\frac{1}{m(h)}\sum_{j=1}^{m(h)} |z(x_j)| - z(x_j+h)|^{1/2}\right\}^4}{0.457 + \frac{0.494}{m(h)} + \frac{0.045}{m^2(h)}}$$

- Dowd's estimator (Dowd, 1984). 2

$$\hat{\gamma}_D(h) = 2.198\{\text{median}|y_j(h)|\}^2,$$

where $y_j(h) = z(x_j) - z(x_j+h)$, $j = 1, 2, \ldots, m(h)$.

Robust variogram estimators



x: Matheron's method of moments; dots: Cressie and Hawkins; o: Dowd; *: Genton

Figure from Oliver and Webster (2014).

Anisotropy

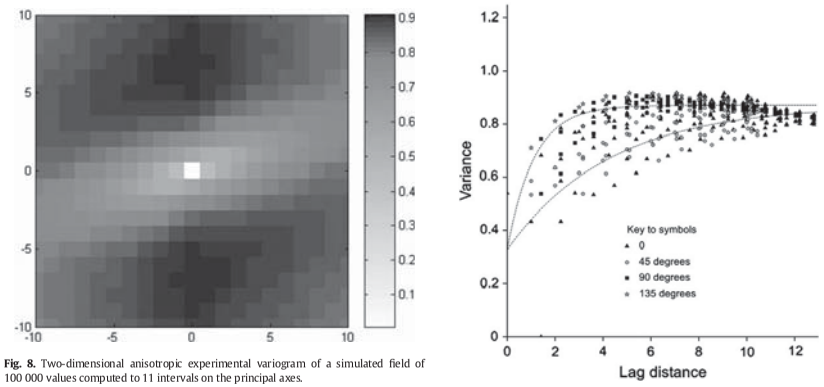


Fig. 8. Two-dimensional anisotropic experimental variogram of a simulated field of 100 000 values computed to 11 intervals on the principal axes.

Left: 2D anisotropic sample variogram of a simulated field.
Right: 90^o and 135^o are on the top; 0^o and 45^o are at the bottom.

Figure from Oliver and Webster (2014).

Kriging

- The most popular methods for spatial prediction.
- Solves a set of linear equations, which contains variogram, to provide the best linear unbiased predictions (BLUP). Best in the sense of minimum variance.
- Returns the observed values at sampled locations.
- Interpolates the values at unsampled locations using the sampled data and the experimental or modeled variogram.
- Provides the standard errors of the interpolated values.
- Serves well in most situations with its assumptions easily satisfied.
- Robust w.r.t. moderate departures from those assumptions and a less than optimal choice of the variogram model.