Semivariogram Modeling

A GRASS tutorial on `s.sv` and `m.svfit`

James Darrell McCauley*

15 Nov 1994

Abstract

The following is a tutorial for `s.sv` and `m.svfit`, GRASS sites programs for semivariogram calculation and modeling of one variable (scattered data) in $\mathbb{R}^2$. The first few sections (§1–5) of this tutorial defines the methods used by these programs and last section (§6) shows how they are used with a data set from Cressie [2].

1 Introduction

This tutorial assumes that the reader has taken at least an introductory course in statistics and has some background in semivariogram modeling. It also assumes that the reader has reviewed the manual pages for `s.sv` and `m.svfit`. An excellent practical introduction to geostatistics has been written by Isaaks and Srivastava [4]. Cressie [2] provides a larger, more complete, and more mathematically based reference book. Sample data sets from each of these works are available from the author so that a user may learn from both the references and this software simultaneously.

2 Definition of Semivariance

If we consider a stochastic process Z, as a function of spatial coordinate $\boldsymbol{s}$, then the *variogram* $2\gamma(\cdot)$ is defined as

$$2\gamma\left(\boldsymbol{s}_1 - \boldsymbol{s}_2\right) \equiv \operatorname{var}\left[Z(\boldsymbol{s}_1) - Z(\boldsymbol{s}_2)\right]. \tag{1}$$

When the process $Z(\cdot)$ is instrinsically stationary, the variogram may be defined as [2]:

$$2\gamma\left(\boldsymbol{s}_1 - \boldsymbol{s}_2\right) \equiv \mathrm{E}\left[Z(\boldsymbol{s}_1) - Z(\boldsymbol{s}_2)\right]. \tag{2}$$

*USDA Graduate Fellow, Department of Agricultural Engineering, Purdue University (`mccauley@ecn.purdue.edu`)

If we define the *lag* $\boldsymbol{h}$ as the distance and angle between $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, then the *semivariogram* is a plot of γ as a function of $\boldsymbol{h}$.

3 Estimators of Semivariance

Under the constant mean assumption, Matheron [6] used the following (classical) estimator:

$$2\hat{\gamma}(\boldsymbol{h}) \equiv \frac{1}{|N(\boldsymbol{h})|} \sum_{N(\boldsymbol{h})} \left[Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j)\right]^2. \tag{3}$$

where

$$N(\boldsymbol{h}) \equiv \{Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j) : \boldsymbol{s}_i - \boldsymbol{s}_j = \boldsymbol{h}; i, j = 1, \ldots, n\}$$

and $|N(\boldsymbol{h})|$ is the number of distinct pairs lagged by the vector $\boldsymbol{h}$.

However, this is not the only estimator used by geostatisticians. For example, Cressie and Hawkins [1] defined the following *robust* estimator:

$$2\bar{\gamma}(\boldsymbol{h}) \equiv \frac{\left[\frac{1}{|N(\boldsymbol{h})|} \sum_{N(\boldsymbol{h})} |Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j)|^{\frac{1}{2}}\right]^4}{0.457 + \frac{0.494}{|N(\boldsymbol{h})|}} \tag{4}$$

(termed the *Mean Fourth Root* estimator). Also, Cressie [2] gives another *robust* estimator:

$$2\tilde{\gamma}(\boldsymbol{h}) \equiv \left[\text{med}\{|Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j)|^{\frac{1}{2}} \;:\; (\boldsymbol{s}_i, \boldsymbol{s}_j) \in N(\boldsymbol{h})\}\right]^4 / B(\boldsymbol{h}) \tag{5}$$

(termed the *Median Fourth Root* estimator) where

$$B(\boldsymbol{h}) = 0.457 + \frac{0.494}{|N(\boldsymbol{h})|}.$$

Deutsch and Journel [3] give the *Semimadogram* (mean absolute difference):

$$2\check{\gamma}(\boldsymbol{h}) \equiv \frac{1}{|N(\boldsymbol{h})|} \sum_{N(\boldsymbol{h})} |Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j)|, \tag{6}$$

the *Semirodogram* (root of difference):

$$2\check{\gamma}(\boldsymbol{h}) \equiv \frac{1}{|N(\boldsymbol{h})|} \sum_{N(\boldsymbol{h})} \sqrt{|Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j)|}, \tag{7}$$

the *General Relative Semivariogram*:

$$2\acute{\gamma}(\boldsymbol{h}) \equiv \frac{\hat{\gamma}}{\left(\frac{m_{-}\boldsymbol{h} + m_{+}\boldsymbol{h}}{2}\right)^2}, \tag{8}$$

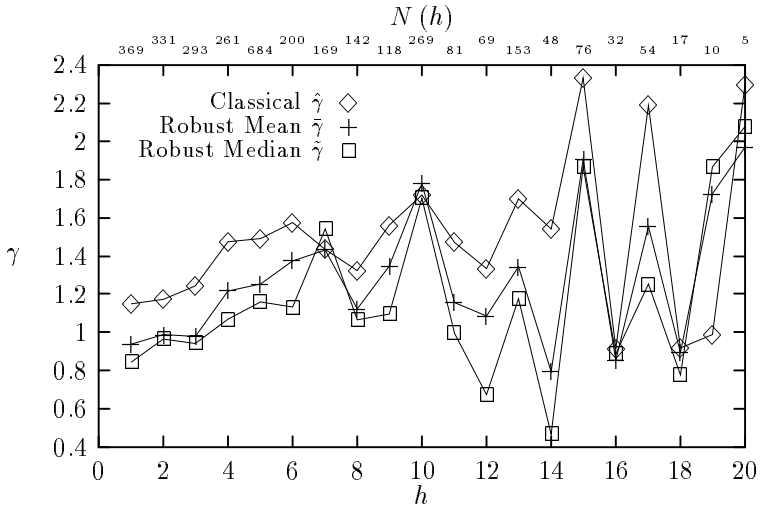

Figure 1: Comparison of semivariance estimators[1](after [2]).

where

$$m_{-\boldsymbol{h}} = \frac{1}{|N(\boldsymbol{h})|} \sum_{N(\boldsymbol{h})} Z(\boldsymbol{s}_i)$$

and

$$m_{+\boldsymbol{h}} = \frac{1}{|N(\boldsymbol{h})|} \sum_{N(\boldsymbol{h})} Z(\boldsymbol{s}_j),$$

and the *Pairwise Relative Semivariogram*

$$2\dot{\gamma}(\boldsymbol{h}) \equiv \frac{1}{|N(\boldsymbol{h})|} \sum_{N(\boldsymbol{h})} \frac{Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j)}{\left(\frac{Z(\boldsymbol{s}_i)+Z(\boldsymbol{s}_j)}{2}\right)^2} \tag{9}$$

Currently, only eqn. 3 (see fig. 1) is used in `m.svfit`. The others (eqn. 5 and 9) may be added in future versions.

4 Lag Settings

Important to note here is that, for stability reasons, the lag distance is defined as "any integral multiple of the sampling interval [7]." In other words, given

[1]For validation, the results on page 82 of Cressie's book [2] were calculated using this software. A lag vector of $(1 \pm 0, 90 \pm 1°)$ was used. Results were *exactly* the same for the number of significant digits given.

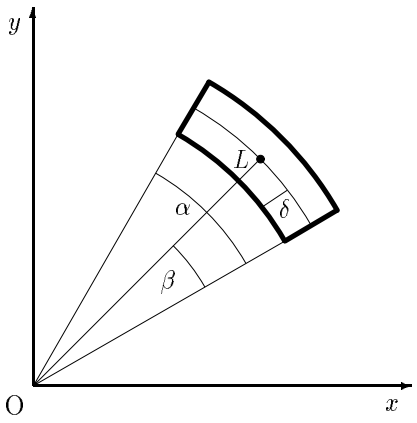

Figure 2: Groupings of lags by distance and direction (after [7]). The distance OL is an integral multiple of the sampling interval in the α direction. Journel and Huijbregts [5] recommend that the number of distinct paris of $(\boldsymbol{s}_i, \boldsymbol{s}_j)$ in the "tolerance" region be at least 30.

a vector $\boldsymbol{h}$ (the minimum lag vector), semivariance is computed at $\boldsymbol{h}$, $c\boldsymbol{h}$, $2c\boldsymbol{h}$, $3c\boldsymbol{h}$, $\ldots nc\boldsymbol{h}$.

For irregularly–spaced data, this will not work since most points will never be separated exactly by any multiple of $\boldsymbol{h}$. Usually, for two–dimensional data, lag directions may be grouped, as in figure 2. If one sample point is at the origin and another sample points is $L \pm \delta$ units away in the $\alpha \pm \beta$ direction, then these two may be grouped together.

Because of the calculation of the angle for $\boldsymbol{h}$, `s.sv` may not work properly for lat-long data.

The sample variogram computed by `s.sv` may be directional or omnidirectional, depending upon if an argument is given for the `angle` option. For omnidirectional semivariograms, only $L \pm \delta$ is used as the pairing criteria (i.e., only $||\boldsymbol{h}||$ is considered).

5 Model Fitting

Several models are used to represent semivariograms. Usually, *isotropy* is assumed so that the vector $\boldsymbol{h}$ becomes a scalar. Cressie [2] gives six basic models.

Linear model (valid in $\mathbb{R}^d$, $d \geq 1$):

$$\gamma(\boldsymbol{h};\boldsymbol{\theta}) = \begin{cases} 0, & \boldsymbol{h} = \mathbf{0} \\ c_o + c_l||\boldsymbol{h}||, & \boldsymbol{h} \neq \mathbf{0} \end{cases} \tag{10}$$

$\boldsymbol{\theta} = [c_o, c_l]'$, where $c_o \geq 0$ and $b_l \geq 0$.

Spherical Model (valid in $\mathbb{R}^1$, $\mathbb{R}^2$, and $\mathbb{R}^3$):

$$\gamma(\boldsymbol{h};\boldsymbol{\theta}) = \begin{cases} 0, & \boldsymbol{h} = \mathbf{0} \\ c_o + c_s\left\{\frac{3}{2}\left(||\boldsymbol{h}||/a_s\right) - \frac{1}{2}\left(||\boldsymbol{h}||/a_s\right)^3\right\}, & 0 \leq ||\boldsymbol{h}|| \leq a_s \\ c_o + c_s & ||\boldsymbol{h}|| \geq a_s \end{cases} \tag{11}$$

$\boldsymbol{\theta} = [c_o, c_s, a_s]'$, where $c_o \geq 0$, $c_s \geq 0$, and $a_s \geq 0$.

Exponential model (valid in $\mathbb{R}^d$, $d \geq 1$):

$$\gamma(\boldsymbol{h};\boldsymbol{\theta}) = \begin{cases} 0, & \boldsymbol{h} = \mathbf{0} \\ c_o + c_e\left\{1 - \exp\left(-||\boldsymbol{h}||/a_e\right)\right\}, & \boldsymbol{h} \neq \mathbf{0} \end{cases} \tag{12}$$

$\boldsymbol{\theta} = [c_o, c_e, a_e]'$, where $c_o \geq 0$, $c_e \geq 0$, and $a_e \geq 0$.

Rational quadratic model (valid in $\mathbb{R}^d$, $d \geq 1$):

$$\gamma(\boldsymbol{h};\boldsymbol{\theta}) = \begin{cases} 0, & \boldsymbol{h} = \mathbf{0} \\ c_o + c_r||\boldsymbol{h}||\left(1 + ||\boldsymbol{h}||^2/a_r\right), & \boldsymbol{h} \neq \mathbf{0} \end{cases} \tag{13}$$

$\boldsymbol{\theta} = [c_o, c_r, a_r]'$, where $c_o \geq 0$, $c_r \geq 0$, and $a_r \geq 0$.

*Hole effect (*or *wave) model* (valid in $\mathbb{R}^1$, $\mathbb{R}^2$, and $\mathbb{R}^3$):

$$\gamma(\boldsymbol{h};\boldsymbol{\theta}) = \begin{cases} 0, & \boldsymbol{h} = \mathbf{0} \\ c_o + c_w\left\{1 - a_w \sin\left(||\boldsymbol{h}||/a_w\right)/||\boldsymbol{h}||\right\}, & \boldsymbol{h} \neq \mathbf{0} \end{cases} \tag{14}$$

$\boldsymbol{\theta} = [c_o, c_w, a_r]'$, where $c_o \geq 0$, $c_w \geq 0$, and $a_r \geq 0$.
Power model (valid in $\mathbb{R}^d$, $d \geq 1$):

$$\gamma(\boldsymbol{h};\boldsymbol{\theta}) = \begin{cases} 0, & \boldsymbol{h} = \mathbf{0} \\ c_o + c_p||\boldsymbol{h}||^\lambda, & \boldsymbol{h} \neq \mathbf{0} \end{cases} \tag{15}$$

$\boldsymbol{\theta} = [c_o, c_p, \lambda]'$, where $c_o \geq 0$, $c_p \geq 0$, and $0 \leq \lambda < 2$.

Weighted least-squares data fitting may be used, where $|N(\boldsymbol{h})|$ is the weight

applied to each observation. If a least-squares problem is defined as

$$\min_x \sum_{j=1}^{m} \left[(b - Ax)_j \right]^2$$

then a weighted least-squares problem is defined as

$$\min_x \sum_{j=1}^{m} \left[w_j \, (b - Ax)_j \right]^2$$

In this application,

$$\min_c \sum_{j=1}^{m} \left[|N(\boldsymbol{h})|_j \, (\gamma - Ac)_j \right]^2 .$$

Recall $\boldsymbol{\theta} = \left[[c]' \sqcup a \right]'$ where $\sqcup$ denotes matrix augmentation with the parameter vector $[c]$ and the range parameter a.

For `m.svfit` to use weighted least squares, the `-w` flag should be used.

6 Example Application: Coal Ash Data

As with most software, usage may best be described by a short example. For `s.sv` and `m.svfit`, we will begin with the coal ash measurements (slightly modified) used by Cressie [2]. A GRASS database location was created for this data and they were imported using `s.in.ascii`.

The remainder of this example assumes that you are already running the GRASS shell (though the shell prompt is indicated by the string "`%`").

```
% s.out.ascii -d coalash | head -3
   1 16 11.170000
   1 15 9.920000
   1 14 10.210000
```

After starting a graphics monitor (using `d.mon`), we may display site locations using `d.sites` or by following the following dialogue with `g.gnuplot`:

```
% g.gnuplot
   gnuplot:  set nokey
   gnuplot:  set title "Locations of Coalash Samples"
   gnuplot:  set xlabel "Easting"
   gnuplot:  set ylabel "Northing"
   gnuplot:  plot '< s.out.ascii coalash'
```

The plot of locations produced using `g.gnuplot` is shown in figure 3. The data appear gridded, though not every point on the 16×23 grid has a value.

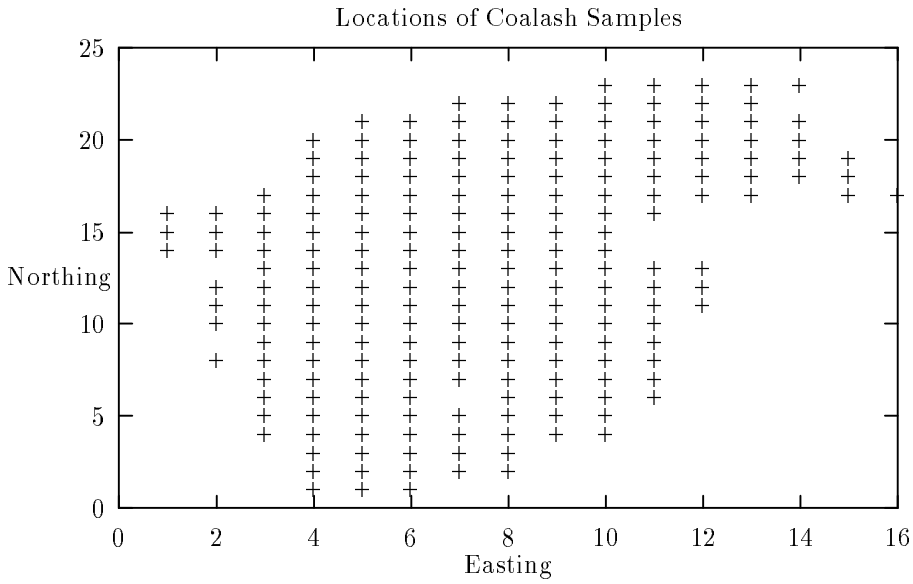

Figure 3: Location of coalash samples.

6.1 Know Your Data

The first and most important step in geostatistical analysis is becoming familiar with the data. This includes much more than just knowing the filename and path! It can sometimes be very useful to know answers to: Who collected your data? When was it collected (hour, day, month, season, year)? How was it collected? What instrument was used? How accurate is the instrument for location? for dependent variables? Answers to these questions and more can help with interpretation of statistics calculated by `s.univar` and `s.sv`. For the coalash data, `s.univar` computes the following:

number of points	208
mean	9.77856
standard deviation	1.27643
coefficient of variation	13.0534
skewness	1.17259
kurtosis	5.8772
mean of squares	97.2416
mean of absolute values	9.77856
minimum	7
first quartile	8.96
median	9.785
third quartile	10.575
maximum	17.61

Also useful here would be a histogram or perhaps a probability plot (see `s.probplt`). Checking for normality, `s.normal` computes a chi-square statistic of $x^2 = 396$ with $\nu = 29$ degrees of freedom. Since $\chi^2_{29,0.05} = 42.6$, we may conclude, at an $\alpha = 0.05$ level, that the data are not normal.

6.2 Sample Semivariogram for Coalash

The following command computes the sample semivariogram with a nominal lag distance of 1 and saves a `g.gnuplot` data and instruction file to `ash.dat` and `ash.gp`, respectively, in the current working directory:

```
% s.sv -p coalash.dat lag=1 gr=ash
```

The sample semivariogram is plotted in the GRASS graphics window. It has been reproduced in figure 4. Notice that after about $||\boldsymbol{h}|| = 16$, $\gamma(\cdot)$ drops off. This may be because $N(\boldsymbol{h})$ also drops off since fewer pairs are separated by larger lags. We may wish to trim the sample semivariogram at this point.

6.3 Semivariogram Model for Coalash

For this example, we will fit a spherical model (eqn. 11) to the trimmed sample semivariogram using `m.svfit`. For this step, after deciding upon a particular model, we select a range by trail-and-error until the model fits satisfactory.

The following command plots the fitted spherical model with a range of 10:

```
% s.sv -q coalash.dat lag=1 | head -16 | m.svfit -wp m=2 r=10
16 samples found, 10 below range
Weighted least squares reg ...       100%
  model = Spherical
  range = 10
   sill = 1.63511
```

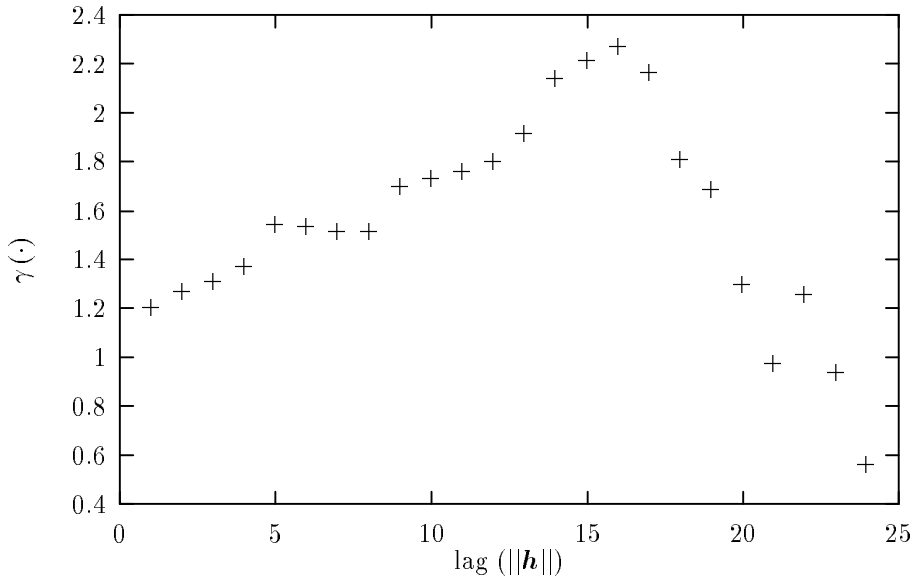

Figure 4: Sample semivariogram of coalash samples.

```
nugget = 1.09554
    c1 = 0.539571
```

The graphical output has been reproduced in figure 5. This command may be repeated several times for different values for the range.

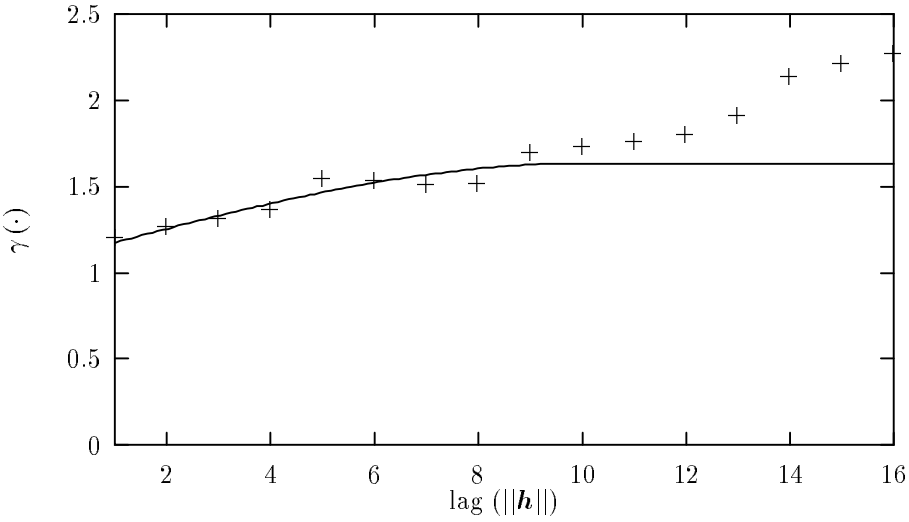

Figure 5: Spherical semivariogram of coalash samples.

References

[1] Noel Cressie and Douglas M. Hawkins. Robust estimation of the variogram: I. *Mathematical Geology*, 12(2):115–125, 1980.

[2] Noel A. C. Cressie. *Statistics for Spatial Data*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, NY, 1991.

[3] Clayton V. Deutsch and André G. Journel. *GSLIB: Geostatistical Software Library and User's Guide*. Oxford University Press, New York, NY, 1992.

[4] Edward H. Isaaks and R. Mohan Srivastava. *An Introduction to Applied Geostatistics*. Oxford University Press, Oxford, 1989.

[5] A. G. Journel and C. Huijbregts. *Mining Geostatistics*. Academic Press, New York, New York, 1978.

[6] G. Matheron. Principles of geostatistics. *Econ. Geol.*, 58:1246–1266, 1963.

[7] R. Webster. Quantitative spatial analysis of soil in the field. *Advances in Soil Science*, 3:1–70, 1985.