

# The pairwise relative semivariogram



Edzer Pebesma

Aug 29, 2011

## 1 Introduction

The general relative variogram (Deutsch and Journel, 1997) is defined as

$$\gamma(h) = \frac{1}{2N_h} \sum_{i=1}^{N_h} \left( \frac{2(Z(s_i) - Z(s_i + h))}{Z(s_i) + Z(s_i + h)} \right)^2.$$

It is claimed to reveal spatial structure (correlation) better when data are skewed and/or clustered. The `cluster.dat` data set used in this vignette, from the GSLIB distribution<sup>1</sup>, seems to confirm this.

From version 1.02 on, R package `gstat` provides computation of the *pairwise relative semivariogram*. The following code provides an example and verification of the computation using direct R code and using the GSLIB program `gamv`.

The following code imports the `cluster.dat` data from GSLIB, which has been converted to have a single-line header containing column names, packaged with the R `gstat` package, and converts it into a `SpatialPointsDataFrame` object:

```
> library(gstat)
> cluster = read.table(system.file("external/cluster.txt", package="gstat"),
+                      header = TRUE)
> summary(cluster)
```

X	Y	Primary	Secondary
Min. : 0.50	Min. : 0.50	Min. : 0.060	Min. : 0.1800
1st Qu.: 9.50	1st Qu.:14.25	1st Qu.: 0.700	1st Qu.: 0.7875
Median :25.50	Median :27.00	Median : 2.195	Median : 2.3750
Mean :23.32	Mean :25.61	Mean : 4.350	Mean : 4.1402
3rd Qu.:35.50	3rd Qu.:36.50	3rd Qu.: 5.327	3rd Qu.: 5.5800
Max. :48.50	Max. :48.50	Max. :58.320	Max. :22.4600
Declustering_Weight			
Min. :0.252			
1st Qu.:0.445			

---

<sup>1</sup>F77 source code for Linux, downloaded Aug 28, 2011 from <http://www.gslib.com/>

```

Median :1.012
Mean   :1.000
3rd Qu.:1.416
Max.   :2.023

> library(sp)
> coordinates(cluster) = ~X+Y

```

The following commands specify a sequence of lag boundaries that correspond to the GSLIB conventions, and compute a regular variogram using these boundaries:

```

> bnd = c(0,2.5,7.5,12.5,17.5,22.5,27.5,32.5,37.5,42.5,47.5,52.5)
> variogram(Primary~1, cluster, boundaries = bnd)

```

	np	dist	gamma	dir.hor	dir.ver	id
1	149	1.527974	58.07709	0	0	var1
2	624	5.472649	54.09188	0	0	var1
3	989	10.150607	48.85144	0	0	var1
4	1249	15.112173	40.08909	0	0	var1
5	1148	20.033244	42.45081	0	0	var1
6	1367	25.020160	48.60365	0	0	var1
7	1311	29.996102	46.88879	0	0	var1
8	1085	34.907219	44.36890	0	0	var1
9	904	39.876469	47.34666	0	0	var1
10	611	44.716540	38.72725	0	0	var1
11	219	49.387310	30.67908	0	0	var1

To compute the relative pairwise variogram, the logical argument PR (*pairwise relative*) needs to be set to TRUE:

```

> variogram(Primary~1, cluster, boundaries=bnd, PR = TRUE)

      np      dist      gamma dir.hor dir.ver   id
1  149  1.527974  0.3608431     0      0 var1
2  624  5.472649  0.6307083     0      0 var1
3  989 10.150607  0.8376443     0      0 var1
4 1249 15.112173  0.7769083     0      0 var1
5 1148 20.033244  0.8774599     0      0 var1
6 1367 25.020160  0.8961016     0      0 var1
7 1311 29.996102  0.9002297     0      0 var1
8 1085 34.907219  0.9604305     0      0 var1
9  904 39.876469  0.9055426     0      0 var1
10 611 44.716540  0.7554474     0      0 var1
11 219 49.387310  0.8226759     0      0 var1

```

Figure 1 shows the two variograms, as plots, side by side

## 2 Verification with plain R code

The following R code reproduces the relative pairwise semivariogram values for the first three lags, i.e. 0-2.5, 2.5-7.5 and 7.5-12.5.

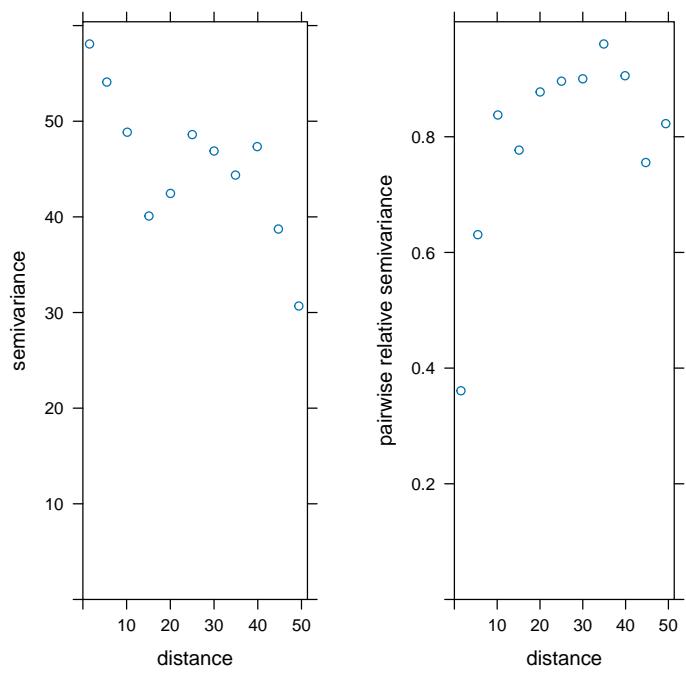


Figure 1: Regular variogram (left) and pairwise relative variogram (right) for the GSLIB data set `cluster.dat`.

```

> z = cluster$Primary
> d = spDists(cluster)
> zd = outer(z, z, "-")
> zs = outer(z, z, "+")
> pr = (2 * zd / zs )^2
> prv = as.vector(pr)
> dv = as.vector(d)
> mean(prv[dv > 0 & dv < 2.5])/2

[1] 0.3608431

> mean(prv[dv > 2.5 & dv < 7.5])/2

[1] 0.6307083

> mean(prv[dv > 7.5 & dv < 12.5])/2

[1] 0.8376443

```

### 3 Verification with GSLIB

In a verification with the GSLIB (Deutsch and Journel, 1997) code of `gamv`, the following file was used:

```

Parameters for GAMV
*****
START OF PARAMETERS:
./data/cluster.dat
file with data
1 2 0
columns for X, Y, Z coordinates
1 3
number of variables,column numbers
-1.0e21 1.0e21
trimming limits
gamv.out
file for variogram output
10
number of lags
5.0
lag separation distance
2.5
lag tolerance
1
number of directions
0.0 90.0 50.0 0.0 90.0 50.0
azm,atol,bandh,dip,dtol,bandv
0
standardize sills? (0=no, 1=yes)

```

```

2
number of variograms
1 1 1
tail var., head var., variogram type
1 1 6
tail var., head var., variogram type

```

Running this program with these parameters gave the following output:

Semivariogram		tail:Primary	head:Primary	direction	1
1	.000	.00000	280	4.35043	4.35043
2	1.528	58.07709	298	8.62309	8.62309
3	5.473	54.09188	1248	5.41315	5.41315
4	10.151	48.85144	1978	4.42758	4.42758
5	15.112	40.08909	2498	4.25680	4.25680
6	20.033	42.45081	2296	3.74311	3.74311
7	25.020	48.60365	2734	4.09575	4.09575
8	29.996	46.88879	2622	4.15950	4.15950
9	34.907	44.36890	2170	3.77190	3.77190
10	39.876	47.34666	1808	4.54173	4.54173
11	44.717	38.72725	1222	5.15251	5.15251
12	49.387	30.67908	438	4.56539	4.56539
Pairwise Relative		tail:Primary	head:Primary	direction	1
1	.000	.00000	280	4.35043	4.35043
2	1.528	.36084	298	8.62309	8.62309
3	5.473	.63071	1248	5.41315	5.41315
4	10.151	.83764	1978	4.42758	4.42758
5	15.112	.77691	2498	4.25680	4.25680
6	20.033	.87746	2296	3.74311	3.74311
7	25.020	.89610	2734	4.09575	4.09575
8	29.996	.90023	2622	4.15950	4.15950
9	34.907	.96043	2170	3.77190	3.77190
10	39.876	.90554	1808	4.54173	4.54173
11	44.717	.75545	1222	5.15251	5.15251
12	49.387	.82268	438	4.56539	4.56539

As can be seen, the values in the third column (semivariogram for the first section, pairwise relative semivariogram for the second) correspond to the output generated by `variogram` of package `gstat`. Two differences with respect to the `gstat` output are:

- for the first lag with distance zero, GSLIB reports that the semivariance value is zero based on 280 point pairs;
- the number of point pairs in GSLIB is double the number reported by `gstat`.

The ground for these differences seems that the GSLIB `gamv` uses a single routine for computing variograms as well as cross variograms and cross covariances. For cross variograms or covariograms, considering two variables  $Z_a$  and  $Z_b$  each having  $N$  observations, the  $N^2$  point pairs  $Z_a(s_i), Z_b(s_i + h)$  and  $Z_a(s_i + h), Z_b(s_i)$  need to be evaluated, and all contribute information.

For direct (non-cross) variograms or covariograms,  $Z_a = Z_b$  and the  $N^2$  pairs considered contain the  $N$  trivial pairs  $(Z(s_i) - Z(s_i))^2 = 0$ , which contribute no information, as well as all duplicate pairs, i.e. in addition to  $(Z(s_i) - Z(s_i+h))^2$ , the identical pair  $(Z(s_i + h) - Z(s_i))^2$  is also considered. This leads to correct variogram value estimates, but incorrect unique point pair numbers. (Data set `cluster` contains  $N = 140$  observations.)

In contrast, `gstat` considers (and reports) only the number of unique pairs for each lag.

## References

- Deutsch, C.V., A.G. Journel, 1997. GSLIB: Geostatistical Software Library and User's Guide, second edition. Oxford University Press.