Alternatives to the Classical Variogram Estimator

In attempting to fit an omnidirectional (or directional) variogram to the Walker Lake U concentration data, it was apparent that due to the hole effect, some preferential clustering due to the sampling scheme, and quite possibly some underlying relationship between the mean and variance (the proportional effect), the classical variogram estimator may not be appropriate. This handout outlines two alternatives to the classical variogram estimator, discussing their forms and when they should be considered. These two forms are called:

1. Relative Variograms (page 165, Chapter 5 - Bailey & Gatrell)
2. Robust Variogram Estimator

Relative Variograms: The major problem leading to the use of relative variograms is the proportional effect. This occurs whenever the local means are fairly stable, but drift across the region of interest with some relationship to the variance. There are three types of relative variograms, each of which is used for a slightly different purpose, as outlined below.

1. **Local Relative Variograms**: If the response values are preferentially clustered, or in some way represent clusters of separate populations, it may be more appropriate to compute variograms over each clustered region individually, and then combine them in some weighted fashion based on the local cluster means and the number of pairs used in each variogram calculation.

   Suppose there are m separate regions or populations, from which the variogram and mean have been calculated, and are denoted: $\widehat{\gamma}_i(\boldsymbol{h})$ and $\overline{y}_i$ respectively, $i = 1, \ldots, m$. The local relative variogram is given by the following weighted average:

$$\widetilde{\gamma}_{LR}(\boldsymbol{h}) = \frac{\sum_{i=1}^{m} n_i(\boldsymbol{h}) \frac{\widehat{\gamma}_i(\boldsymbol{h})}{\overline{y}_i^2}}{\sum_{i=1}^{n} n_i(\boldsymbol{h})}, \quad \text{where:}$$

   $n_i(\boldsymbol{h})$ = the number of sample pairs used in the i^{th} region.

   - This estimator (of what?) is just the average of the regional variograms, weighted by both the local means and the number of pairs used in the variogram calculations.
   - A major problem with this estimator is that it assumes that there are a sufficient number of pairs of sites in each region to compute a variogram. This is often not the case.
   - The estimator, as written, assumes that there is a linear relationship between the local mean and local standard deviation (since it weights $\widehat{\gamma}_i(\boldsymbol{h})$ by the square of the mean). This is the common type of proportional effect for concentration or log data. If some other relationship is present (such as proportionality between the local mean and *variance*), the estimator given should be adjusted to reflect the relationship present.

2. **General Relative Variograms**: To circumvent the possibility of having too few pairs to compute a local relative variogram, the general relative variogram adjusts the variogram based on the local mean of *all points used in calculating that variogram.* This is more of a global than local adjustment for the proportional effect. The resulting estimator is given by:

$$\widetilde{\gamma}_{GR}(\boldsymbol{h}) = \frac{\widehat{\gamma}(\boldsymbol{h})}{\overline{y}(\boldsymbol{h})^2}, \quad \text{where:}$$

   $\overline{y}(\boldsymbol{h})$ is the mean of all values used in the calculation of $\widehat{\gamma}(\boldsymbol{h})$, given by:

$$\overline{y}(\boldsymbol{h}) = \frac{1}{2n(\boldsymbol{h})} \sum_{(i,j)|\boldsymbol{h_{ij}} \approx \boldsymbol{h}} (v_i + v_j) = \frac{m_{+\boldsymbol{h}} + m_{-\boldsymbol{h}}}{2}.$$

   - Effectively, all this adjustment does is scale the values of the variogram for the orientation $\boldsymbol{h}$ by some constant. I find it hard to believe that this would correct most erratic behavior in the variogram.

3. **Pairwise Relative Variogram**: As with the previous two relative variograms, the pairwise relative variogram adjusts the variogram by the squared mean to reduce any proportional effects present. The adjustment in this case is done on a pairwise basis (as opposed to regionally or globally as with the other two types). The resulting estimator is given by:

$$\widetilde{\gamma}_{PR}(\boldsymbol{h}) = \frac{1}{2n(\boldsymbol{h})} \sum_{(i,j)|\boldsymbol{h_{ij}} \approx \boldsymbol{h}} \left[\frac{(v_i - v_j)^2}{\left(\frac{v_i + v_j}{2}\right)^2} \right].$$

   - The adjustment made to the classical variogram estimator occurs in the denominator of the summand above. Dividing by the squared average of the pairs of points helps to reduce the effect of the larger contributions to the variogram, which should in turn reduce any erratic behavior.
   - It should be noted that if both v_i and v_j are zero, the estimator is undefined. To avoid this, Isaaks & Srivastava suggest choosing a lower bound for the denominator, so that small data values do not become overinflated.

Some Final Comments on Relative Variograms

- In viewing the three relative variograms computed for the Walker Lake U concentration data, two seem to provide some improvement, whereas one does not. The local relative variogram still reveals a hole effect, whereas both the general relative and pairwise relative variograms show the increase and leveling off we expect in a variogram.

- There is really no theoretical basis for any of the three relative variograms introduced. In other words, it is not clear exactly what these relative variograms are estimating. This has limited and even prevented their use in practice.

- As the impetus for using a relative variogram is the presence of a proportional effect in the data, it might be better to first attempt a transformation of the data to correct the proportional effect. Once this is accomplished, one should be able to compute omnidirectional and directional variograms without worrying about the problems encountered due to a proportional effect.

- Only the pairwise relative variogram appears to be supported in **R** and can be invoked using the **PR=T** option within the **variogram** function.

Robust Variogram Estimator: Throughout statistics, we encounter estimators of population parameters, termed "classical", which represent the optimal estimators under special conditions, such as normality of the response variable. For example, it is well known that the sample mean and variance are optimal estimators of the population mean and variance whenever the corresponding data are normally distributed.

With variogram estimation, if the response variable is normally distributed, then the classical variogram estimator, given by:

$$\widehat{\gamma}(\boldsymbol{h}) = \frac{1}{2n(\boldsymbol{h})} \sum_{(i,j)|\boldsymbol{h_{ij}} \approx \boldsymbol{h}} (v_i - v_j)^2,$$

is optimal. As with the sample variance, if the data are contaminated either by outliers or severe skewness, then the classical variogram estimator may be greatly affected. In other words, it is not robust to departures from normality.

Cressie & Hawkins (1980, *Mathematical Geology* 12:115-125) proposed a robust estimator of the population variogram, $\gamma(\boldsymbol{h})$, which is based on the square root of the absolute value of the differences, $|V_i - V_j|^{1/2}$, rather than the squares of the differences.

- One other reason for considering estimators other than the classical variogram estimator is that under normality:

$$(V_i - V_j)^2/2\gamma(\boldsymbol{h}) \;\sim\; \chi^2(1),$$

  a distribution which is highly skewed. It is easier to estimate the expectation of a symmetric distribution, and taking the fourth root of $(V_i - V_j)^2$ achieves the necessary symmetry.

- The expectation of $|V_i - V_j|^{1/2}$ is then estimated as either the sample mean:

$$\frac{1}{n(\boldsymbol{h})} \sum_{(i,j)|\boldsymbol{h_{ij}} \approx \boldsymbol{h}} |v_i - v_j|^{1/2},$$

  or the sample median of $|V_i - V_j|^{1/2}$. This sample mean (or median) is next raised to the fourth power to put it back in the original scale of the variogram:

$$\left\{ \frac{1}{n(\boldsymbol{h})} \sum_{(i,j)|\boldsymbol{h_{ij}} \approx \boldsymbol{h}} |v_i - v_j|^{1/2} \right\}^4.$$

- When nonlinear transformations such as this are done, the expectation of the back-transformed statistic is not exactly the same as the back-transformed mean. Using a Taylor series expansion (often referred to as the Delta Method), the expectation of this back-transformed statistic is:

$$E\left\{\frac{1}{n(\boldsymbol{h})}\sum_{(i,j)|\boldsymbol{h_{ij}} \approx \boldsymbol{h}} |V_i - V_j|^{1/2}\right\}^4 \approx 2\left(.457 + \frac{.494}{n(\boldsymbol{h})}\right)\gamma(\boldsymbol{h}).$$

Consequently, Cressie & Hawkins' final robust variogram estimator is:

$$\overline{\gamma}(\boldsymbol{h}) = \frac{1}{2\left(.457 + \frac{.494}{n(\boldsymbol{h})}\right)}\left\{\frac{1}{n(\boldsymbol{h})}\sum_{(i,j)|\boldsymbol{h_{ij}} \approx \boldsymbol{h}} |v_i - v_j|^{1/2}\right\}^4,$$

which is *approximately* unbiased for $\gamma(\boldsymbol{h})$. There is a similar version based on medians.

Some Notes on the Robust Variogram Estimator

1. You can choose the robust variogram estimator in **R** by entering the option: **cressie=T** within the variogram function. For example:

```
coordinates(walk470) ~ x+y
dat.var <- variogram(v ~ x+y, data=walk470, width=5, cutoff=120, cressie=T)
```

2. The robust variogram estimator is better for estimating the shape of the variogram, especially in the presence of outliers, but has been noted to underestimate the nugget effect and the overall sill.

3. In a paper by Genton (1998, Mathematical Geology), an alternative robust variogram estimator was proposed which appears to estimate the shape of the variogram better than Cressie & Hawkins' $\overline{\gamma}(\boldsymbol{h})$ in the presence of outliers.

4. An article by Basu et. al. (1997, Journal of Agricultural, Biological, and Environmental Statistics 2:490-512) showed the dramatic effects even a single outlier can have on both the classical variogram estimator $\widehat{\gamma}(\boldsymbol{h})$ and the robust variogram estimator $\overline{\gamma}(\boldsymbol{h})$. They conclude that it is better to remove any known outliers before computing the semivariogram instead of depending on the "robustness" of the robust variogram estimator.

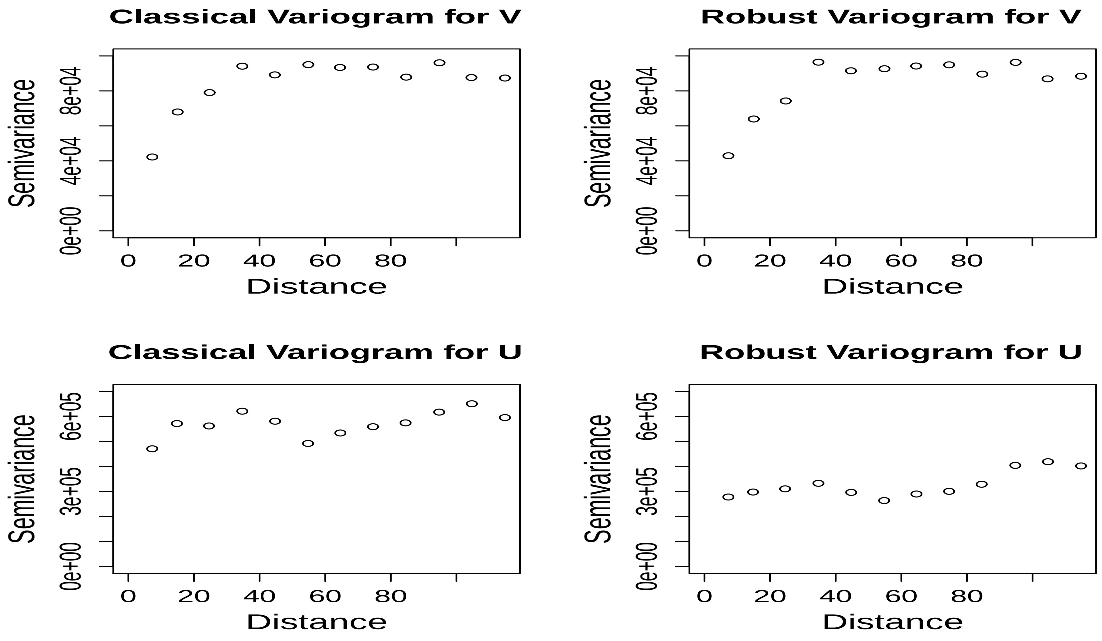
The use of the Cressie-Hawkins robust estimator is illustrated for the Walker Lake data using the following code, which creates the four semivariograms given on the next page.

```
library(gstat)               # Loads gstat library for "variogram" function
walk470 <- read.table("Data/walk470.txt",header=T)
library(sp)                  # Loads sp library for "coordinates" function
coordinates(walk470) = ~x+y  # Assigns coordinates for use in "variogram"
```

```
vario <- as.list(4)          # Initiate a list of length 4.
ylims.V <- c(0,100000)       # y-axis limits for V-semivariances
vario[[1]] <- variogram(v~   # Variogram for the V-data
  x+y,data=walk470,width=10,cutoff=120)
vario[[2]] <- variogram(v~   # Robust variogram (cressie=T) for the V-data
  x+y,data=walk470,width=10,cressie=T,cutoff=120)

ylims.U <- c(0,700000)       # y-axis limits for U-semivariances
u <- walk470$u               # U-data
vario[[3]] <- variogram(u~x+y,data=walk470[!is.na(u),],width=10,cutoff=120)
vario[[4]] <- variogram(u~x+y,data=walk470[!is.na(u),],width=10,cressie=T,
                        cutoff=120)

vtype <- rep(c("Classical",  # Labels for variogram titles on plots,
  "Robust"),2)               #   repeated twice
varname <- c("V","V","U","U")# Variable names for the 4 plots
par(mfrow=c(2,2))            # Sets up a 2x2 graphics window
for (i in 1:4){              # Loops through the 4 variograms fit
  v = vario[[i]]             # Assigns v to variogram output i
  plot(gamma~dist,v,xlim=c(0,max(v$dist)), # Plots the ith semivariogram
    xlab="Distance",ylab="Semivariance",   #   with axis labels and titles
    ylim=get(paste("ylims.",varname[i],    #   cutomized using the variable
    sep="")),main=paste(vtype[i],          #   names in "vtype" & "varname"
    " Variogram for ",varname[i],sep=""),
    cex.main=1.5,cex.axis=1.3,cex.lab=1.6,mgp=c(2.7,1,0))
}                            # End of for loop
```


Classical Variogram for V
Robust Variogram for V
Classical Variogram for U
Robust Variogram for U
Semivariance
Distance
0e+00
4e+04
8e+04
0e+00
3e+05
6e+05