University of California, Los Angeles
Department of Statistics

Statistics C173/C273 **Instructor: Nicolas Christou**

Data with trend

- As we discussed earlier, if the *intrinsic stationarity* assumption holds, which implies

  $$E(Z(s+h) - Z(s)) = 0$$

  and

  $$Var(Z(s+h) - Z(s)) = 2\gamma(h)$$

  we can write

  $$\text{Var}(Z(s+h) - Z(s)) = E(Z(s+h) - Z(s))^2$$

  and therefore we can use the method of moments estimator for the variogram (also called the classical estimator):

  $$2\hat{\gamma}(h) = \frac{1}{N(h)} \sum_{N(h)} (Z(s_i) - Z(s_j))^2,$$

  where the sum is over $N(h)$ such that $s_i - s_j = h$.

- But what if there is a tend in our data? For example, the values may increase from north to south, or northeast to southwest, etc. We will have to take this into account when computing the variogram. Why?

  It can be shown that the formula for computing the sample variogram is also equal to:

  $$2\hat{\gamma}(h) = \frac{1}{N(h)} \sum_{N(h)} (Z(s_i) - Z(s_j))^2 - \hat{\mu}^2_{\text{diff}}$$

  By assuming a constant mean ($\mu_{diff} = 0$) it is like adding a positive quantity to the variogram. Adding a square term will result to a parabola, and therefore a parabolic variogram is an indication of a presence of a trend in our data.

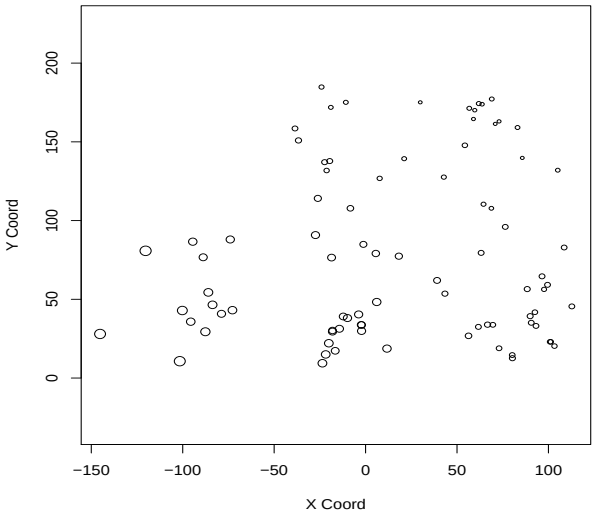
- **Example 1: The Wolfcamp aquifer data: See Cressie (1993, pp. 212–214).** The U.S. Department of Energy proposed (in the 1980s) a nuclear waste site to be in Deaf Smith County in Texas (bordering New Mexico). The contamination of the aquifer was a concern, and therefore the piezometric-head data were obtained at 85 locations by drilling a narrow pipe through the aquifer. The measures are in feet above sea level. See figure below for the location:

  The Texas panhandle:

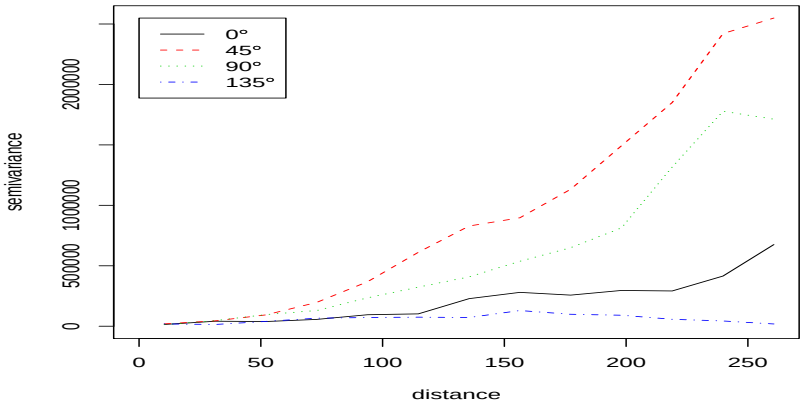
  Please access the data from:

```
> a <- read.table("http://www.stat.ucla.edu/~nchristo/statistics_c173_c273/
  wolfcamp.txt", header=T)
> b <- as.geodata(a)
> points(b)
```

  The data points:

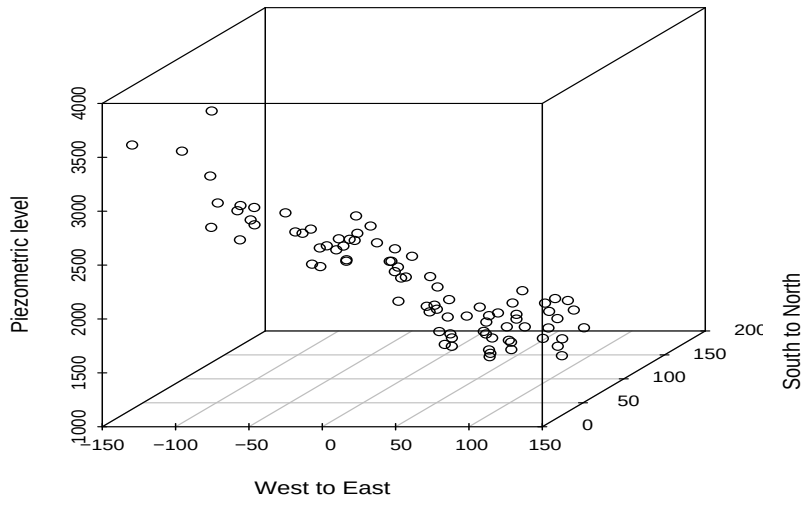


```
> var1 <- variog4(b)
> plot(var1)
```



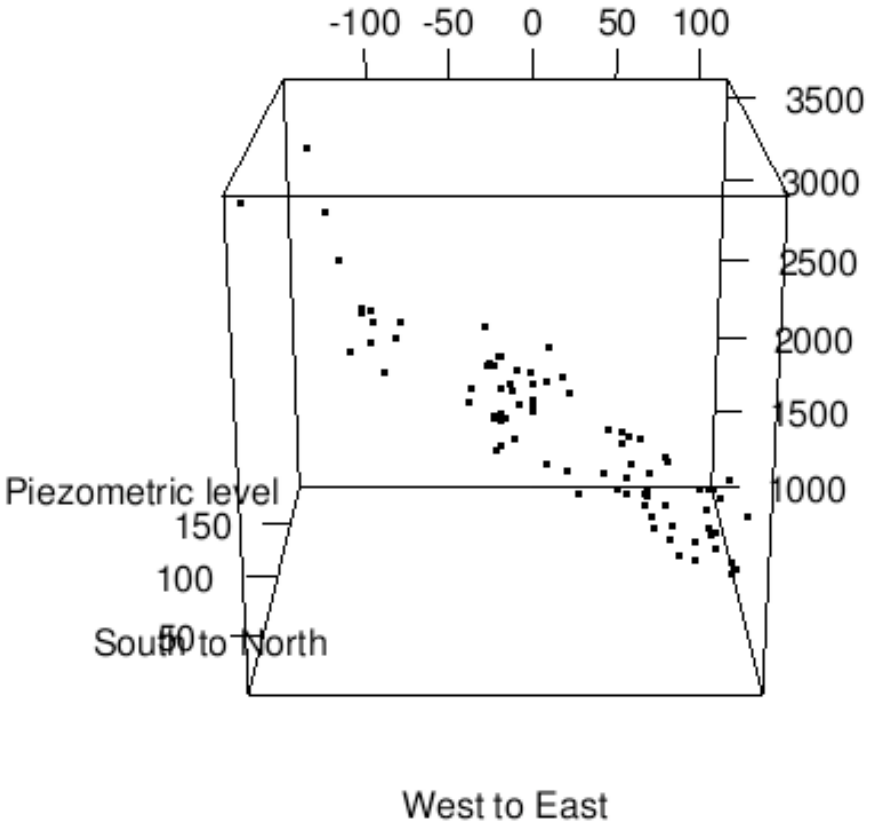
The parabolic shape of the sample variograms indicates that clearly there is a trend in our data. See the 3D figure below. In order to construct this 3D graph we must load the `library(scatterplot3d)`.

```
> library(scatterplot3d)
> scatterplot3d(a$x,a$y,a$level, xlab="West to East",
  ylab="South to North", zlab="Piezometric level")
```



Similarly we can use the `rgl` library for more interactive plots:

```
> library(rgl)
> plot3d(a$x,a$y,a$level, xlab="West to East",
  ylab="South to North", zlab="Piezometric level", size=3)
```



We should try to "detrend" the data by fitting a plane through them. We can fit a linear surface to the data by regressing the data against the x and y coordinates. Or, we can fit a quadratic or cubic surface. Once the surface is estimated we can subtract the observed data from the predicted data and get the residuals. It is hoped that the residuals do not show any trend. All these, can be done within the function `variog`, but we can verify the results by running the regression of the data on the x, y coordinates, obtaining the residuals, and using a new data frame with `x, y, res`.

Using regression in R:

```
> q <- lm(a$level ~ a$x+a$y)
> c <- as.data.frame(cbind(a$x, a$y, q$res))
> names(c) <- c("x", "y", "res")
> d <- as.geodata(c)
```

Now we can compute the variogram on the de-trended residuals (using the geodata object d).

```
> var1 <- variog(d)
```

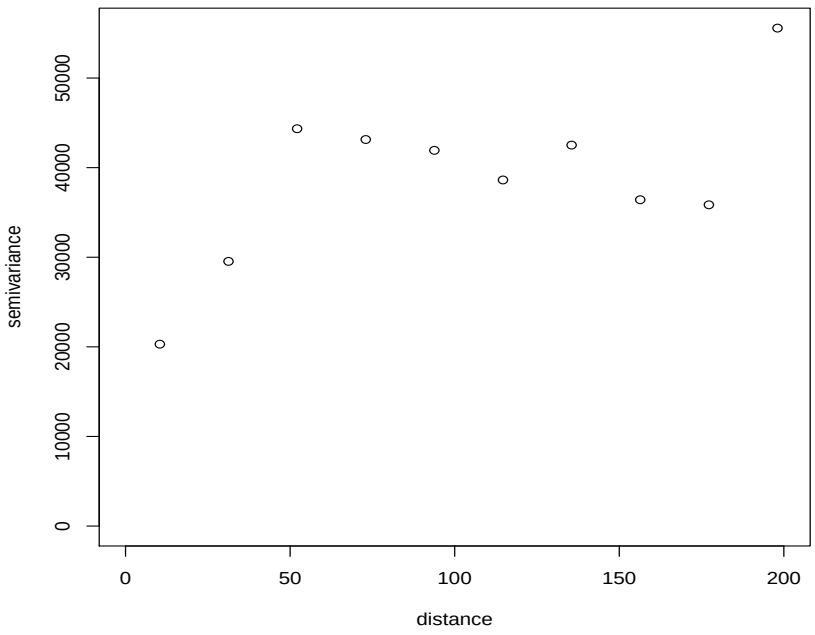
Or, we can simply use the function variog on the initial geodata object b as follows:

```
> vario_detrend <- variog(b, trend="1st")
```

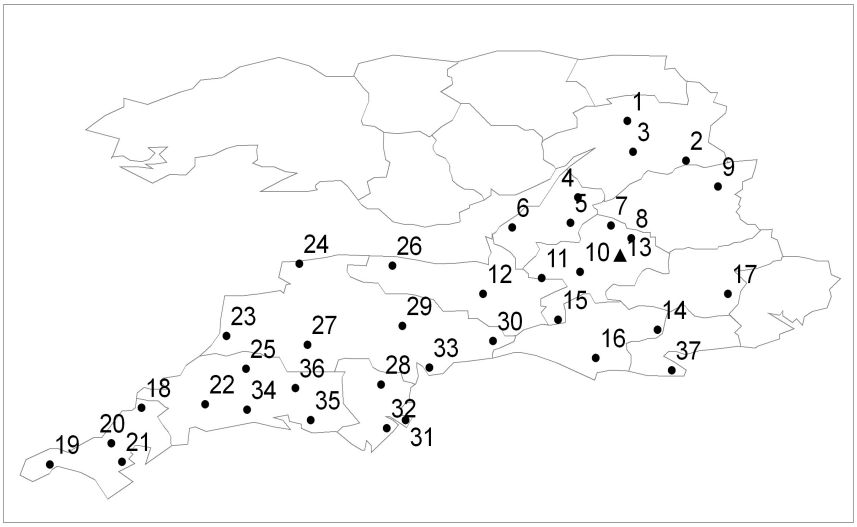
The argument 1st is used for fitting a linear surface. Similarly you can use 2nd for second order polynomial on the coordinates. We can also use the argument trend to fit a surface based on external variables. For example trend=∼x1+x2+x3, where x_1, x_2, x_3 are any external variables.

Here is the new variogram plot after we have detrended our data:

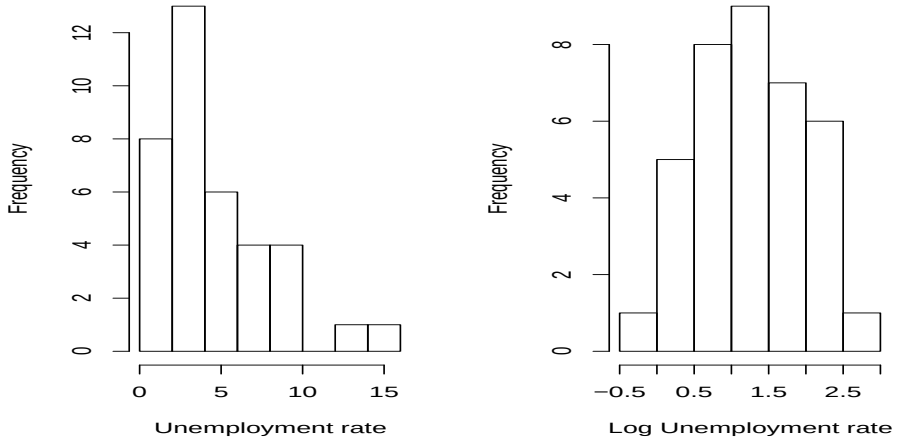
```
> plot(vario_detrend)
```



- **Example 2: Southwest of England 1967 unemployment data.**
  For this data set the percentage of the total workforce unemployed in January, 1967, (see Cliff and Ord (1973)) in the 37 employment areas in the southwest of England is used (see figure below).



We first transform the unemployment rate using logarithms as shown below:



Please access the data from:

```
> a <- read.table("http://www.stat.ucla.edu/~nchristo/statistics_c173_c273/
  unemp_data.txt", header=T)
```

Convert the data frame into a geodata object and compute the sample variogram on the 4 major directions. Use different values for `uvec` and comment on the shape of the sample variograms. Fit a surface to the data, de-trend the residuals, and compute again the variograms.

Robust estimation of the variogram

Cressie and Hawkins (1980) proposed the following estimator for the variogram which is robust to outliers compare to the classical estimator:

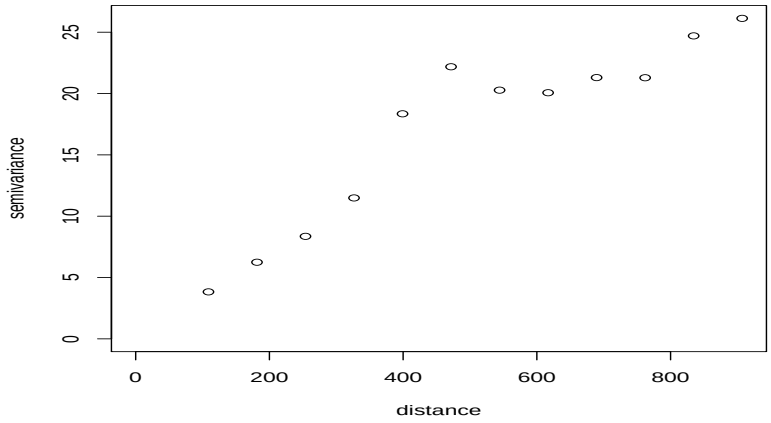
$$2\bar{\gamma}(h) = \frac{\left\{\frac{1}{N(h)}\sum_{N(h)}|Z(s_i) - Z(s_j)|^{\frac{1}{2}}\right\}^4}{\frac{0.457+0.494}{N(h)}}$$

where the sum is over $N(h)$ such that $s_i - s_j = h$.

To use this estimator in `R` we use the argument `estimator.type` as follows:

```
> variogram1 <- variog(b, estimator.type="modulus")
```

The above will compute the robust omnidirectional variogram for the data `data_var.txt`. Here it is:



The classical omnidirectional variogram for the same data is shown below: