James K. Pringle
140.663 Geostatistics
Dr. Curriero
Assignment 2
April 4, 2014

Assignment 2

Questions 2-4

2 Large Scale vs. Small Scale Spatial Variation

The data file `Q2 Data.csv` has 6 variables with 200 records. Variables y_1 and y_2 are two simulated spatial data sets (outcome variables) with coordinates *coordx* and *coordy*. The variables x_1 and x_2 are potential explanatory variables or covariates for each outcome respectively.

(a) Read the data file `Q2 Data.csv` into your current `R` session using the `read.csv` command.

Done.

(b) Perform some exploratory spatial data analysis on the variables y_1 and y_2 with focus on large scale spatial variation and interpret your findings. Recall in the spatial literature, the terms large scale and small scale spatial variation are often taken to mean first order and second order variation.

We do not have covariates other than the x and y coordinates. Therefore the plot of Data vs. x-coordinate and y-coordinate vs Data should be sufficient to describe the large scale variation. In the first set of plots (Figure 1), it does not appear that there is any trend between either x- or y-coordinate with the data. The scatterplots appear to be a blob of uncorrelated points. This would indicate that there is not any large scale spatial variation.

There is a difference in the second set of plots (Figure 2) in that the x-coordinate seems to have a linear relationship with the data. That means that there is large scale variation with the covariates available.

(c) Continue from (b) above and explore small scale spatial variation. Estimate and plot semivariograms for y_1 and y_2. Interpret results from both (b) and (c).

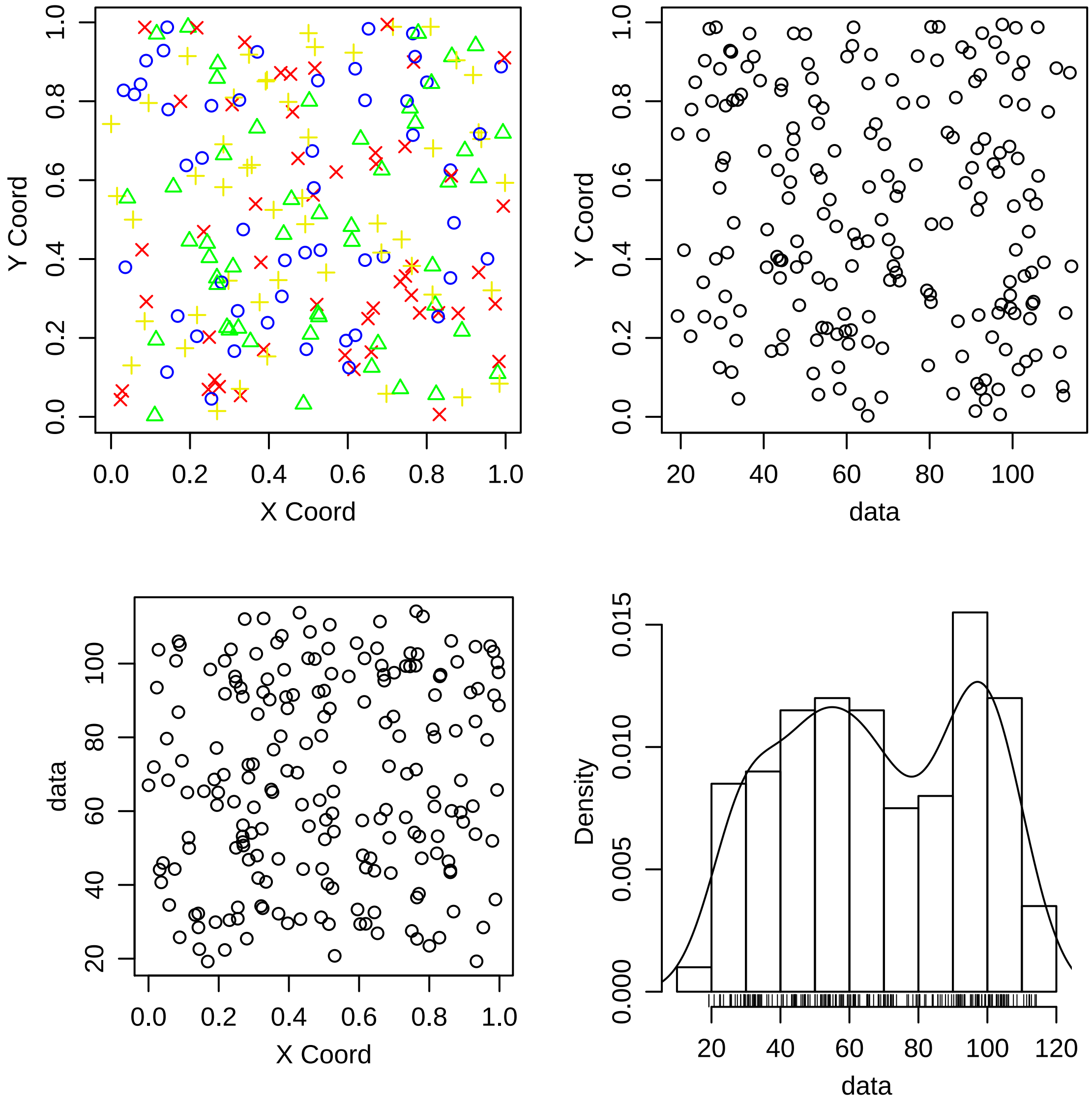


Figure 1: Plot of simulated Y_1 data

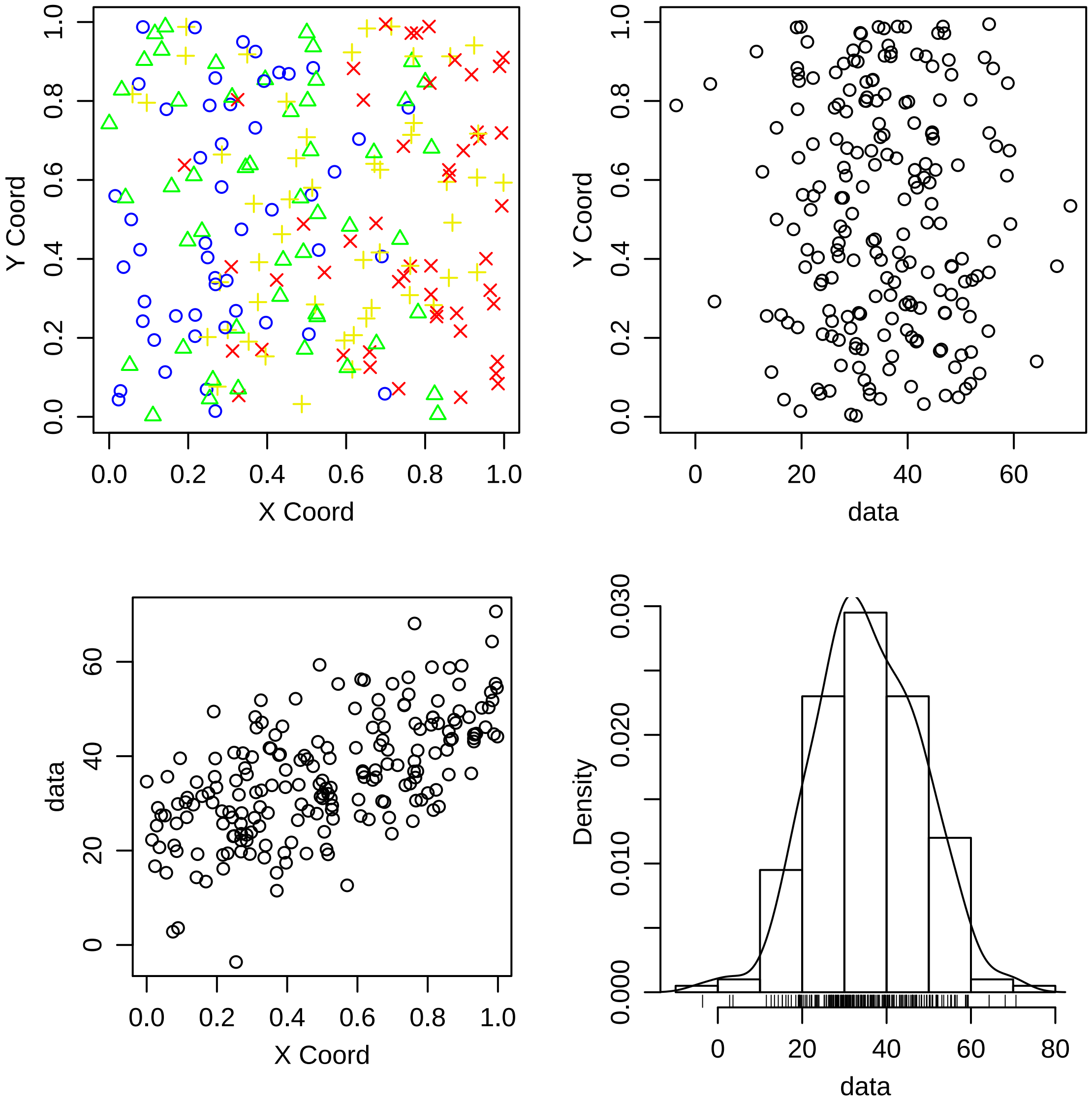
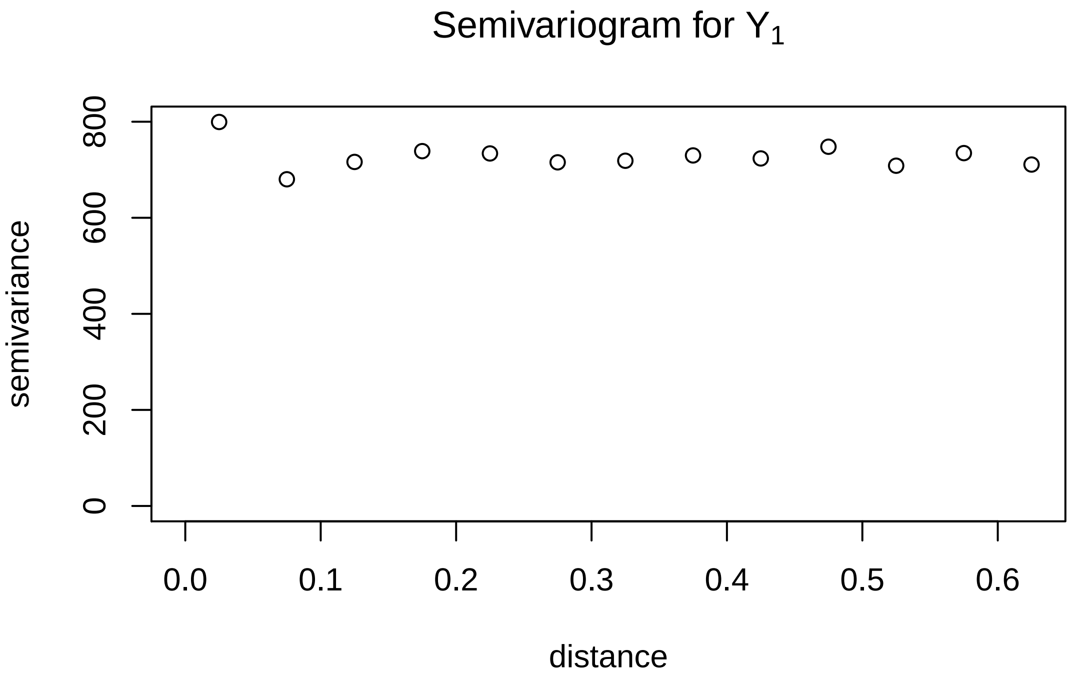
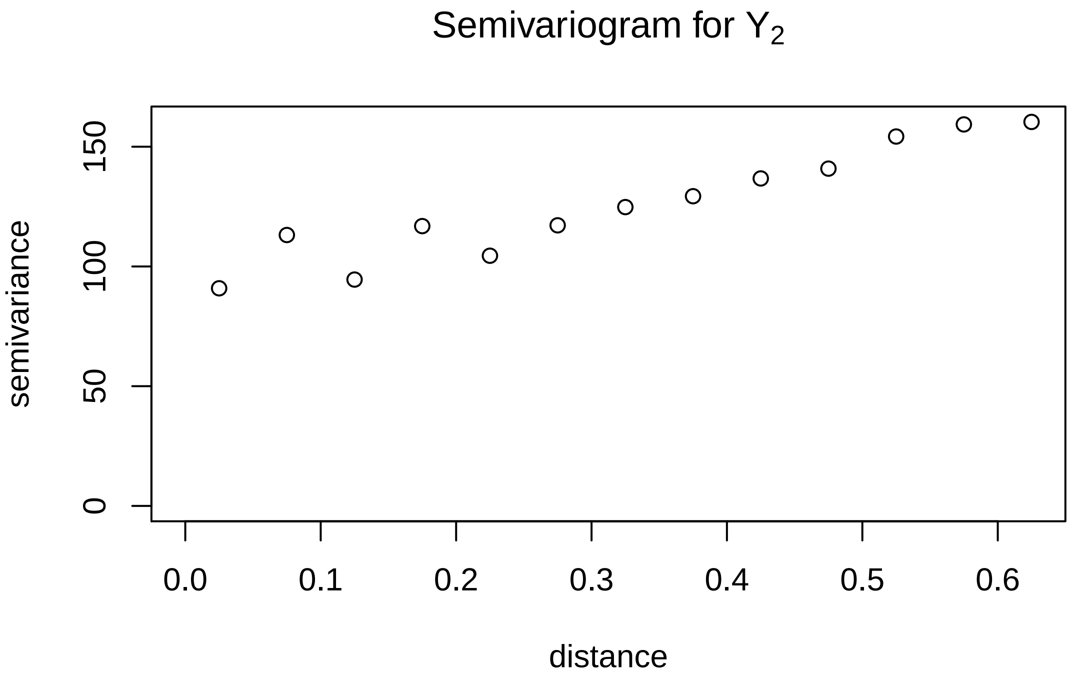


Figure 2: Plot of simulated Y_2 data

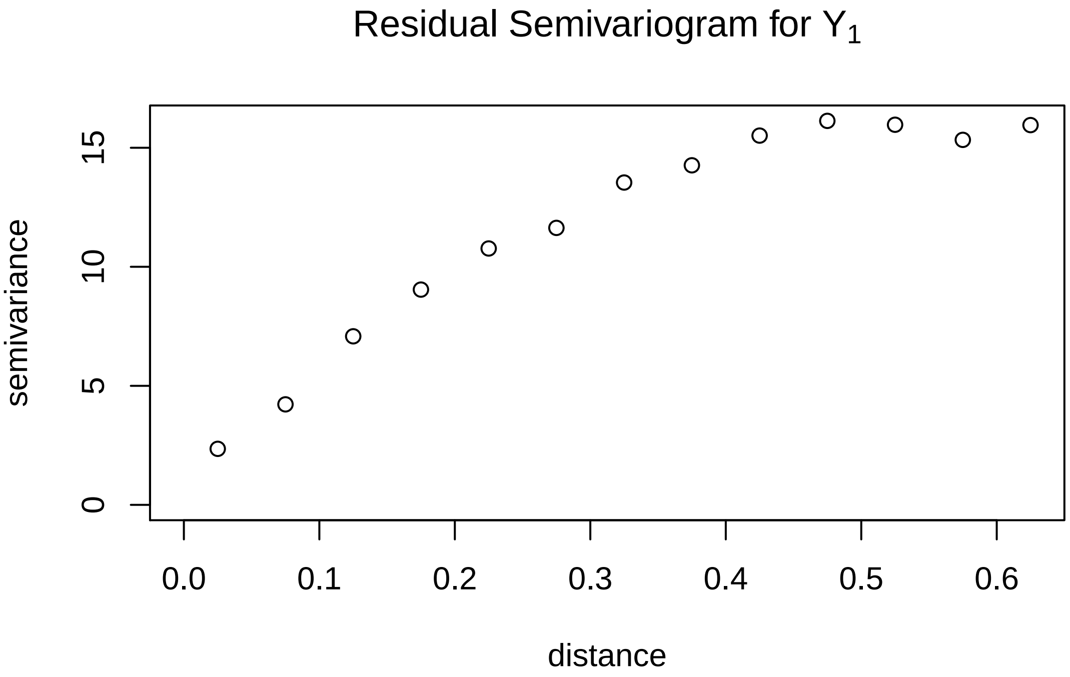


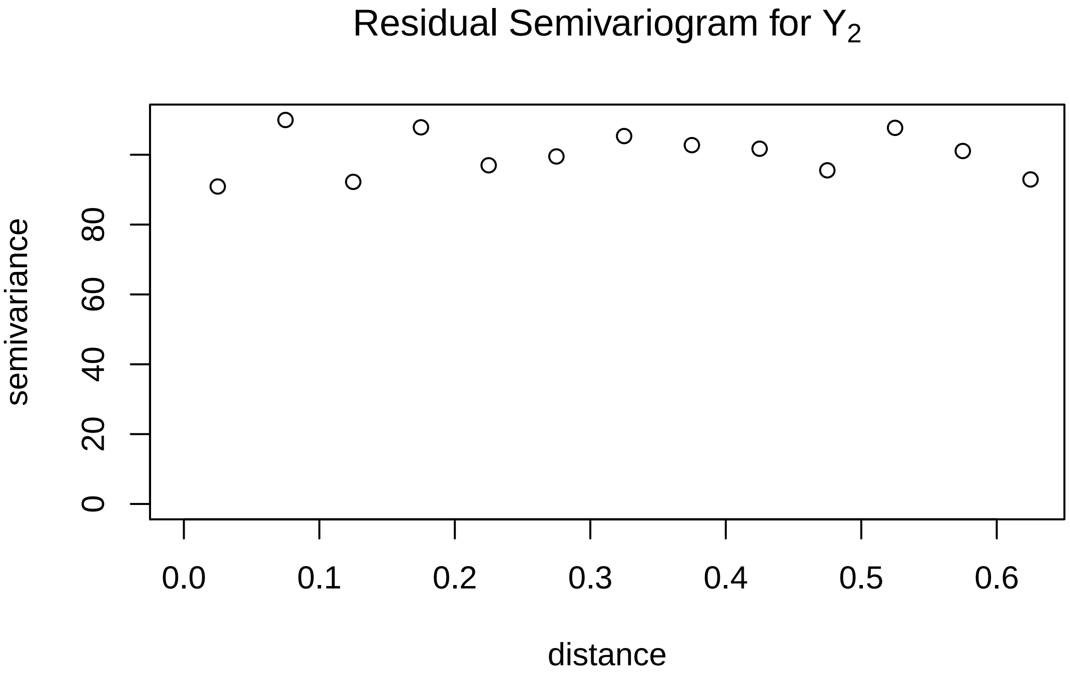


The semivariogram for Y_1 is a flat line. This implies that Y_1 is not spatially dependent, or that any spatial dependence is dwarfed by (is much smaller in comparison to) the variance of the data. That is consistent with the plots in (b), since the Y_1 values seem pretty uniformly scattered throughout the domain.

The semivariagram for Y_2 shows an increasing trend. This implies spatial dependence, that points closer together have more similar values than points farther apart. This reinforces the notion that there is large scale variation associated with the x-coordinate as discussed in (b).

(d) Perform simple linear regressions for y_1 using x_1 as a covariate and for y_2 using x_2 as the covariate. Estimate and plot the residual semivariograms for each. Interpret.





For Y_1, the residuals show spatial dependence. In imprecise terms, after accounting for the variation of X_1, the left over variation (small scale variation) is spatially dependent. Since the variagram of Y_1 is a flat line, it may be possible to conclude that variation due to X_1 is much larger than the spatial variance.

The variogram of the residuals of Y_2 is flat. For Y_2, after accounting for the variation due to X_2 (the x-coordinate), there is no evidence of spatial dependence in the residuals. This means there is no small scale spatial variation in the data, that the data has only large scale spatial variation.

(e) Below are the statistical models used to simulate these data,

$$\begin{aligned} Y_1(s) &= 20 + 3X_1(s) + \epsilon_1(s), \qquad & \epsilon_1(s) \sim N(0, \sigma_1^2) \quad & corr(\epsilon_1(s_i), \epsilon_1(s_j)) \neq 0 \\ Y_2(s) &= 20 + 3X_2(s) + \epsilon_2(s), \qquad & \epsilon_2(s) \sim N(0, \sigma_2^2) \quad & corr(\epsilon_2(s_i), \epsilon_2(s_j)) = 0, \end{aligned}$$

where coordinates $s = (coordx, coordy)$, $X_1(s)$ are uniform random numbers generated independent of Y_1, $X_2(s)$ are actually the $coordx$ values scaled up by a factor of 10, $\epsilon_1(s)$ is a spatially dependent Normal random variable, and $\epsilon_2(s)$ is a Normal random variable spatially independent. Thus $Y_1(s)$ is generated from adding a spatially unstructured variable $X_1(s)$ to spatially structured errors $\epsilon_1(s)$. In contrast, $Y_2(s)$ is generated from adding a spatially structured variable $X_2(s)$ to spatially unstructured errors $\epsilon_2(s)$. With this knowledge, comment on the behavior of the respective semivariograms of the outcome variables compared to their respective residual semivariograms. That is compare the spatial dependence structure of Y_1 with the spatial dependence of the residuals from the regression of Y_1 on X_1, and the spatial dependence of Y_2 with the spatial dependence of the residuals from the regression of Y_2 on X_2.

Given this knowledge of how the outcome variables Y_1 and Y_2 are simulated, it is easier to interpret the variograms. The variogram of Y_1 appears to indicate no spatial variation. That is because the magnitude of the large scale component is much larger than the magnitude of the spatially structured errors. Therefore, after regression onto X_1, the residuals reveal the spatially structured nature of the errors. That is why the variogram of the residuals Y_1 are spatially dependent.

The variogram of Y_2 shows spatial dependence because it is built into the Y_2 by means of X_2 (the x-coordinate). After regression onto X_2, the residuals reveal the spatially independent nature of the errors. This is seen by the flat trend in the variogram of the residuals.

3 Understanding the Covariance Matrix

Consider the following spatial design of locations.

Location	(x,y)
1	(1,1)
2	(1,2)
3	(1,3)
4	(2,1)
5	(2,2)
6	(2,3)
7	(3,1)
8	(3,2)
9	(3,3)

Lets assume we know the spatial dependance for this design is characterized by an exponential semivariogram with nugget of zero, sill of 25, and range of 2,

$$\gamma(h) = 25(1 - \exp(-h/2)).$$

a. Use the known semivariogram model and the relation $\gamma(\|\mathbf{h}\|) = C(\mathbf{0}) - C(\|\mathbf{h}\|)$ to estimate the covariogram $C(\|\mathbf{h}\|)$ and correlogram $\rho(\|\mathbf{h}\|)$. Also provide estimates for the variogram and the number of data points falling within each distance class $N(\|\mathbf{h}\|)$. The table on the following page is provided to fill in and to further clarify what is being asked. The sample locations from this 3×3 regular grid were used to dictate the distance classes.

It is shown later that $C(0) = 25$. Thus

$$C(h) = 25 \exp(-h/2)$$

and

$$\rho(h) = C(h)/C(0) = \exp(-h/2)$$

The table is as follows:

Distance		Variogram	Semivariogram	Covariogram	Correlogram
$\|\mathbf{h}\|$	$N(\|\mathbf{h}\|)$	$2\gamma(\|\mathbf{h}\|)$	$\gamma(\|\mathbf{h}\|)$	$C(\|\mathbf{h}\|)$	$\rho(\|\mathbf{h}\|)$
1	12	19.673	9.837	15.163	0.607
1.414	8	25.344	12.672	12.328	0.493
2	6	31.606	15.803	9.197	0.368
2.236	8	33.653	16.827	8.173	0.327
2.828	2	37.842	18.921	6.079	0.243

b. Use results from the completed table as well as the spatial design of these 9 locations to fill in values for the distance matrix and the covariance matrix Σ. Empty 9×9 matrices (with 81 entries) are provided for convenience. View these matrices as each having 9 rows and 9 columns for spatial locations 1 to 9 and cross reference accordingly. For example, row 3 column 5 corresponds to the location 3 location 5 pair.

$$\text{Distance} = \begin{bmatrix} 0 & 1 & 2 & 1 & 1.414 & 2.236 & 2 & 2.236 & 2.828 \\ 1 & 0 & 1 & 1.414 & 1 & 1.414 & 2.236 & 2 & 2.236 \\ 2 & 1 & 0 & 2.236 & 1.414 & 1 & 2.828 & 2.236 & 2 \\ 1 & 1.414 & 2.236 & 0 & 1 & 2 & 1 & 1.414 & 2.236 \\ 1.414 & 1 & 1.414 & 1 & 0 & 1 & 1.414 & 1 & 1.414 \\ 2.236 & 1.414 & 1 & 2 & 1 & 0 & 2.236 & 1.414 & 1 \\ 2 & 2.236 & 2.828 & 1 & 1.414 & 2.236 & 0 & 1 & 2 \\ 2.236 & 2 & 2.236 & 1.414 & 1 & 1.414 & 1 & 0 & 1 \\ 2.828 & 2.236 & 2 & 2.236 & 1.414 & 1 & 2 & 1 & 0 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 25 & 15.163 & 9.197 & 15.163 & 12.328 & 8.173 & 9.197 & 8.173 & 6.079 \\ 15.163 & 25 & 15.163 & 12.328 & 15.163 & 12.328 & 8.173 & 9.197 & 8.173 \\ 9.197 & 15.163 & 25 & 8.173 & 12.328 & 15.163 & 6.079 & 8.173 & 9.197 \\ 15.163 & 12.328 & 8.173 & 25 & 15.163 & 9.197 & 15.163 & 12.328 & 8.173 \\ 12.328 & 15.163 & 12.328 & 15.163 & 25 & 15.163 & 12.328 & 15.163 & 12.328 \\ 8.173 & 12.328 & 15.163 & 9.197 & 15.163 & 25 & 8.173 & 12.328 & 15.163 \\ 9.197 & 8.173 & 6.079 & 15.163 & 12.328 & 8.173 & 25 & 15.163 & 9.197 \\ 8.173 & 9.197 & 8.173 & 12.328 & 15.163 & 12.328 & 15.163 & 25 & 15.163 \\ 6.079 & 8.173 & 9.197 & 8.173 & 12.328 & 15.163 & 9.197 & 15.163 & 25 \end{bmatrix}$$

c. Answer the following

(i) In this example what would be the covariance between data that are separated by a distance of 2.236?

The covariance of data separated by 2.236 is 8.173, as can be seen by the table in "a."

(ii) What would be the covariance between data at location 1 and location 7?

The covariance between the data at location 1 and location 7 is equal to Σ_{17} and Σ_{71}, which is 9.197.

(iii) What is the variance represented by this semivariogram?

The variance represented by this semivariogram is the variance of the data, var$(Z(s))$. According to class lecture (slides 85, 104), var$(Z(s)) = C(0) = \tau^2 + \sigma^2$. So the variance of the data is the total sill, i.e. 25.

(iv) Based on the known semivariogram, at what distance would points become approximately uncorrelated?

For exponential semivariograms, the 3ϕ is the distance where points become approximately uncorrelated, where ϕ is the range (see slide 100). Here $3\phi = 3 \cdot 2 = 6$

d. Now suppose we assumed for this design that the data were independent with variance 25. Fill in the entries of the covariance matrix Σ. A second empty 9×9 matrix is provided for convenience.

If the data are assumed independent, then cov$(Z(s_i), Z(s_j)) = 0$ where $s_i \neq s_j$. Hence,

$$\Sigma = \begin{bmatrix} 25 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 25 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 25 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 25 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 25 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 25 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 25 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 25 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 25 \end{bmatrix}$$

4 The Wolfcamp Aquifer Data

This data is already in R. Typing the commands `data(wolfcamp)` and `help(wolfcamp)` will load the data and show a brief description. The article *Geostatistics* by N. Cressie 1989 *American Statistician*, 43, 197-202 posted on the course website will provide some background. R commands that produce all output required is provided separately. Use these as a guide in generating other results or graphs you wish.

a. Review the article *Geostatistics* by Cressie. This is not necessarily an easy read but focus on the Introduction, Section on the case study, Concluding Remarks, and the general methodology he proposed to krige this data. See the answer to (b) below to help with understanding the read.

   Done.

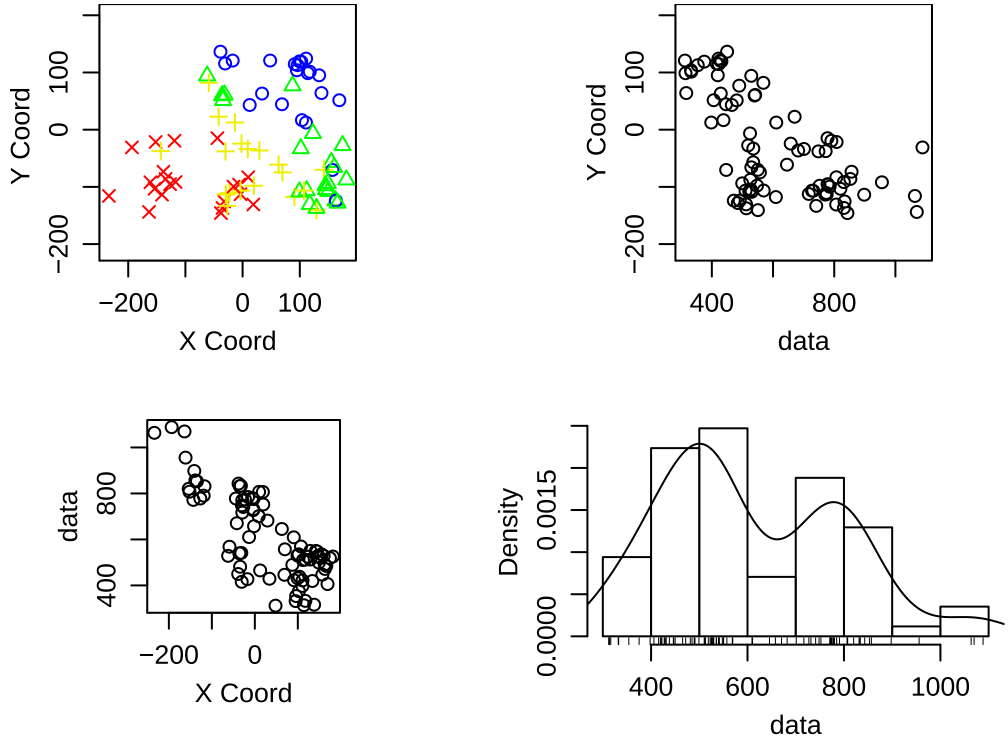
b. Describe the modeling approach used by Cressie to krige the Wolfcamp Aquifer data. Be specific to discuss large scale and small scale variation and any assumptions made about each. I would suggest writing out a model with descriptions (like those shown in class) that depicts Cressies approach. Answer below.

   It appears that Cressie is only estimating the data $Y(s)$ (piezometric-head data) through the variogram (no covariates). This is equivalent to

$$Y(s) = b_0 + \epsilon(s)$$

   However, the error terms have a more complicated structure than what we have discussed in class. Here, the process is modeled to show anisotropy. The variogram of the data is a mixture of "variograms" in two different directions (with tolerance in the direction).

c. Perform some exploratory spatial data analysis, both with a focus on large scale and small scale variation.

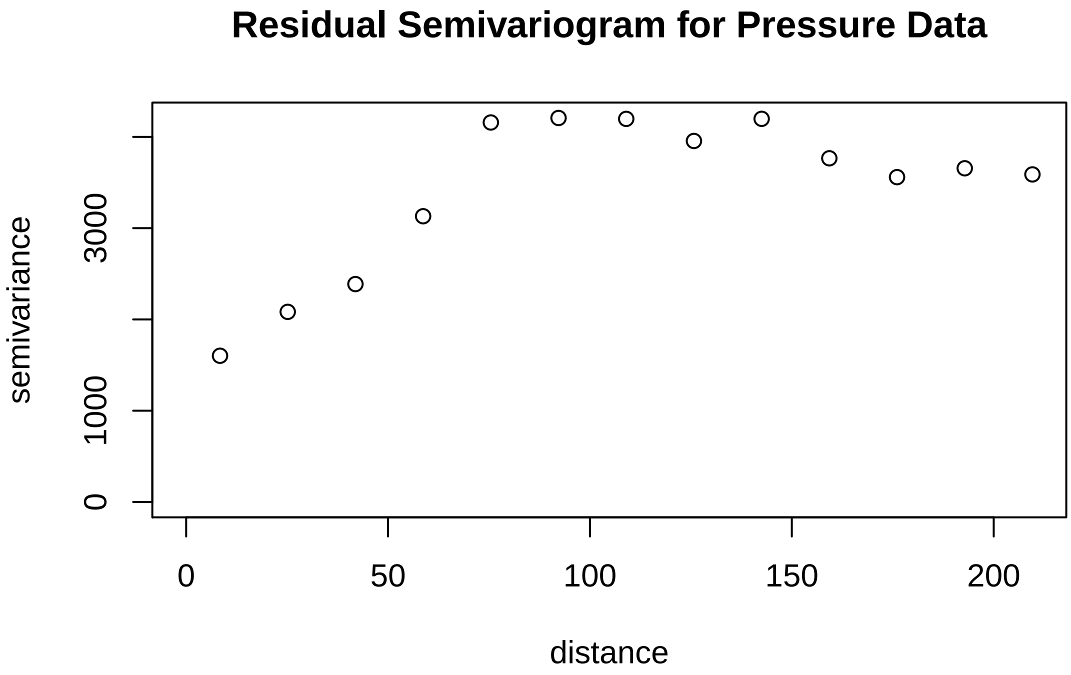


According to this plot, there appears to be a trend surface based on the x- and the y-coordinate. Performing a linear regression on these two covariates (with no interaction) gives

```
##
## Call:
## lm(formula = pressure ~ xcoord + ycoord, data = wolf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -111.99  -50.30   -9.33   48.51  197.99
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 607.7707     7.5222    80.8   <2e-16 ***
## xcoord       -1.2784     0.0655   -19.5   <2e-16 ***
## ycoord       -1.1387     0.0774   -14.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.3 on 82 degrees of freedom
## Multiple R-squared:  0.891,Adjusted R-squared:  0.888
## F-statistic:  335 on 2 and 82 DF,  p-value: <2e-16
```

The effect of both coordinates is a highly significant, negative linear relationship. According to the R^2 value, these two covariates explain 62% of the variation of the data. This shows large scale spatial variation.

However, a variogram of the residuals shows there is still small scale spatial variation.



d. Suggest an alternative kriging approach than that taken in the article. Again be specific to write out a model with assumptions, etc. that describes your approach. Discuss why you selected this alternative.

In my approach, I would try to add covariates for a "universal kriging" model. I do not know much about groundwater pressure, but I would think that the elevation of the ground water would be important (and that it would vary spatially). I chose this covariate because the paper states the contaminants would flow "downhill" toward Amarillo, Texas. That got me interested in elevation. So my model would be

$$Y(s) = b_0 + b_1 X(s) + \epsilon(s)$$

where $X(s)$ is the elevation of the groundwater at location s. I would still check the "directional variogram" like Cressie did, and model it as he did if two directions have different variograms.

e. Is the data available to generate a kriged map of pressure based on your kriging model suggested in (d)? Explain.

It appears that the elevation data needed is not stored in R. I do not know what methods exist for finding this elevation, but according to `https://water.usgs.gov/edu/gwhowtofind.html`, "a target area must be thoroughly tested and studied to identify hydrologic and geologic features important to the planning and management of the resource." This means that it would be difficult (in time and money) to collect the data I would like to have.