DATA.ML.420

Lab Exercise 1

Pen and paper:

1.

Median polish

Start

1	6	3	0
5	9	2	0
6	4	7	0
0	0	0	0

Subtract column medians

-4	0	0	0
0	3	-1	0
1	-2	4	0
5	6	3	0

Subtract row medians

-4	0	0	0
0	3	-1	0
0	-3	3	1
0	1	-2	5

2.

I get the following values for sample covariance function:

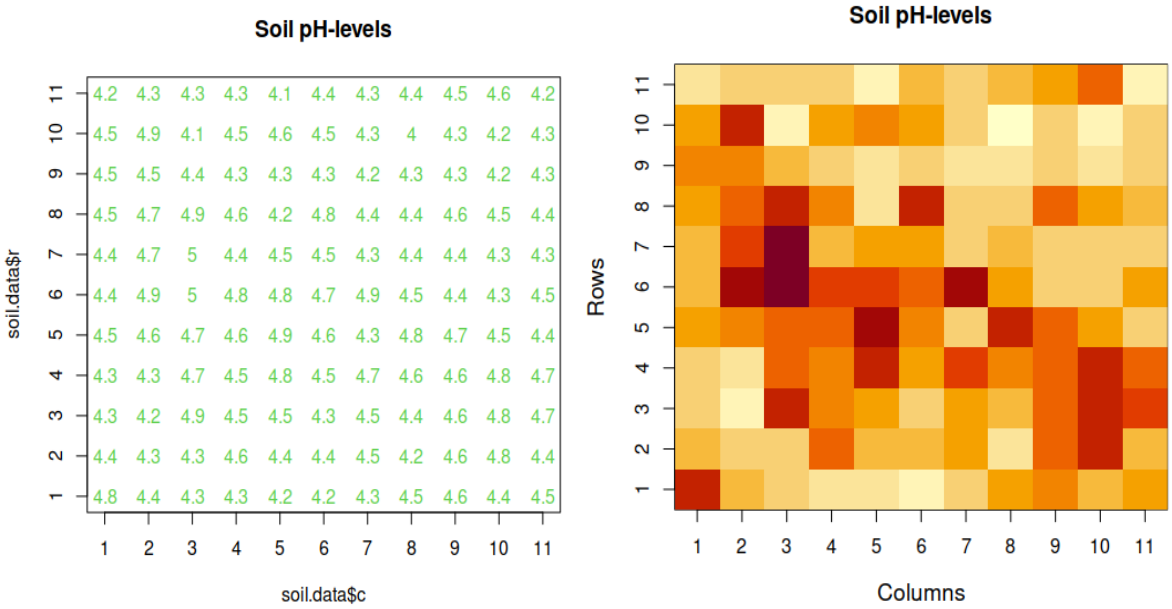
$h = (0,0) : \hat{C}(h) = 3.5$
$h = (1,0) : \hat{C}(h) = 2/3$
$h = (0,1) : \hat{C}(h) = 5/6$
$h = (-1,-1) : \hat{C}(h) = -4/3$
$h = (1,-1) : \hat{C}(h) = 5/3$
$h = (2,0) : \hat{C}(h) = -11/4$
$h = (-2,-1) : \hat{C}(h) = -2$
$h = (2,-1) : \hat{C}(h) = -3/2$
$h = (3,0) : \hat{C}(h) = -3$
$h = (-3,-1) : \hat{C}(h) = 0$
$h = (3,-1) : \hat{C}(h) = 0$

For example, the case $h = (3,0) : \hat{\gamma}(h) = 25/4 = 6.25$, but $\hat{C}(0) - \hat{C}(h) = 3.5 + 3 = 6.5$.

R-exercises (outputs)

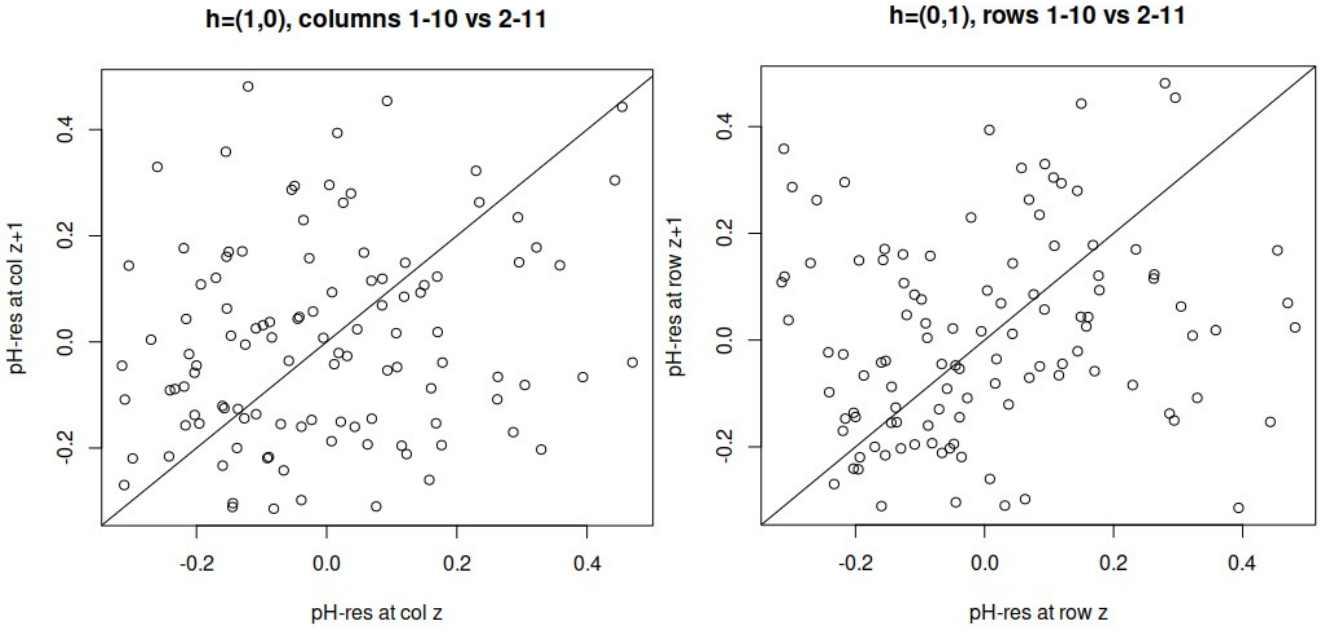
1.

i) Exploratory analysis



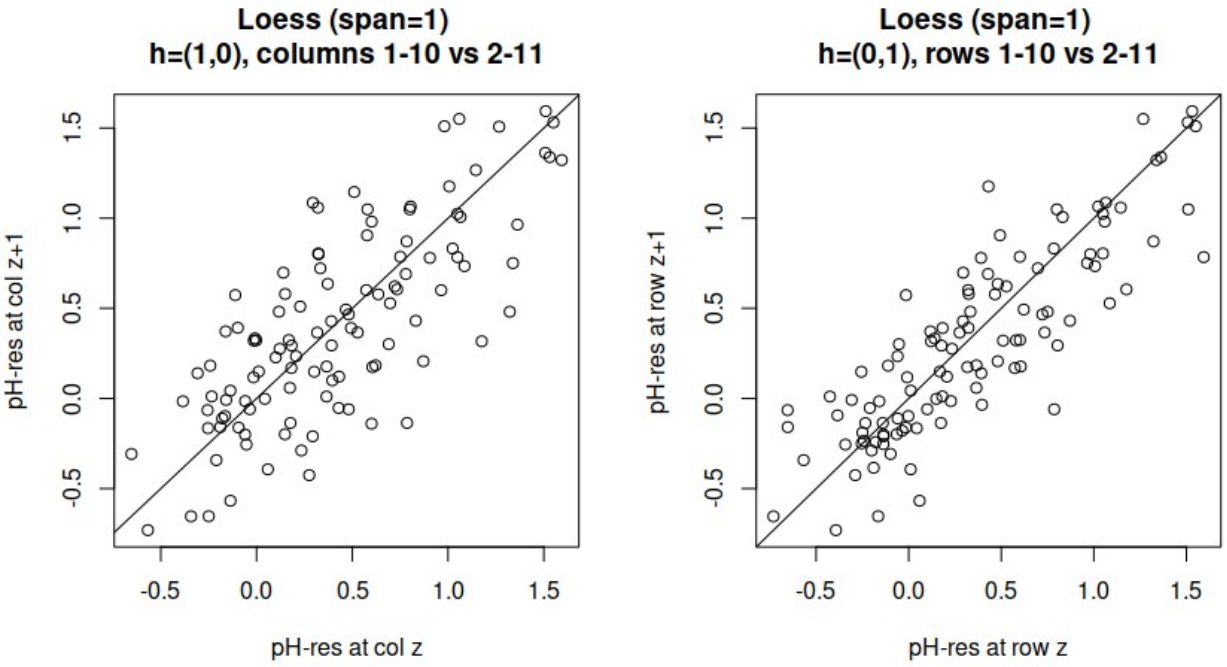
There are a little bit larger pH-levels observed at the lower-triangle (near the diagonal) of the grid.

ii)



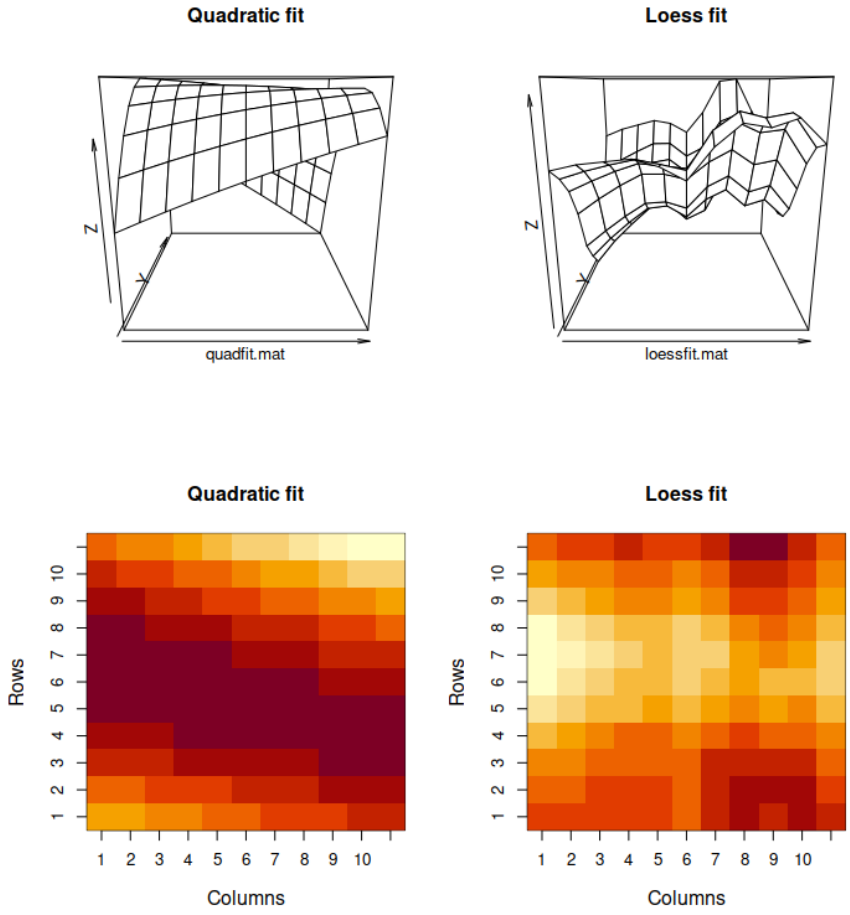
All residuals seem to be reasonably close to 0 in both cases and the distribution of points looks quite random. Therefore we can deduce that the quadratic regression model fits the data reasonably well.

iii)



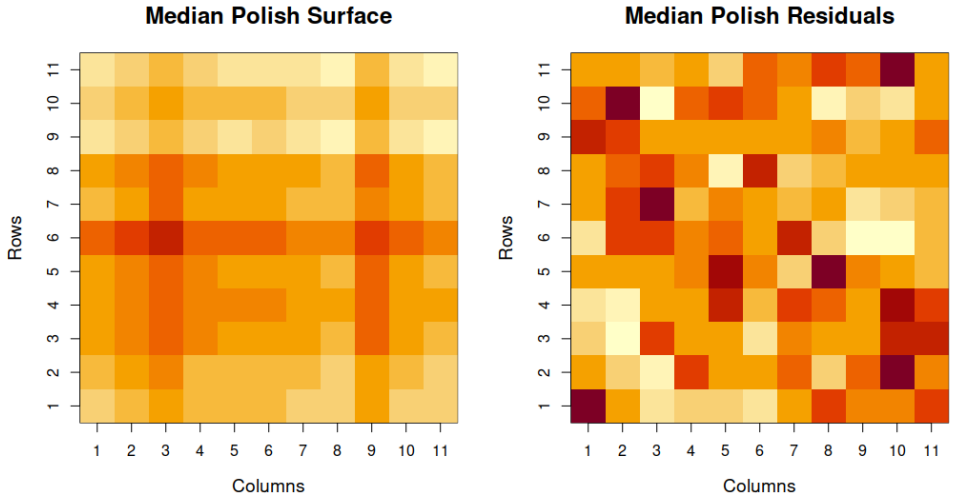
I selected the smoothness (span) to be 1 since this value tends to produce residuals closer to zero. However, residuals still seem to be quite further away from 0 compared to quadratic model. There also seems to be some kind of positive trend going on among residuals that is not preferable (?). Therefore we can deduce that the quadratic regression model fits the data better than the loess model.

iv)



Based on the grayscale plot, it seems that the polynomial regression model tends to show the large pH-level trend at diagonal region.

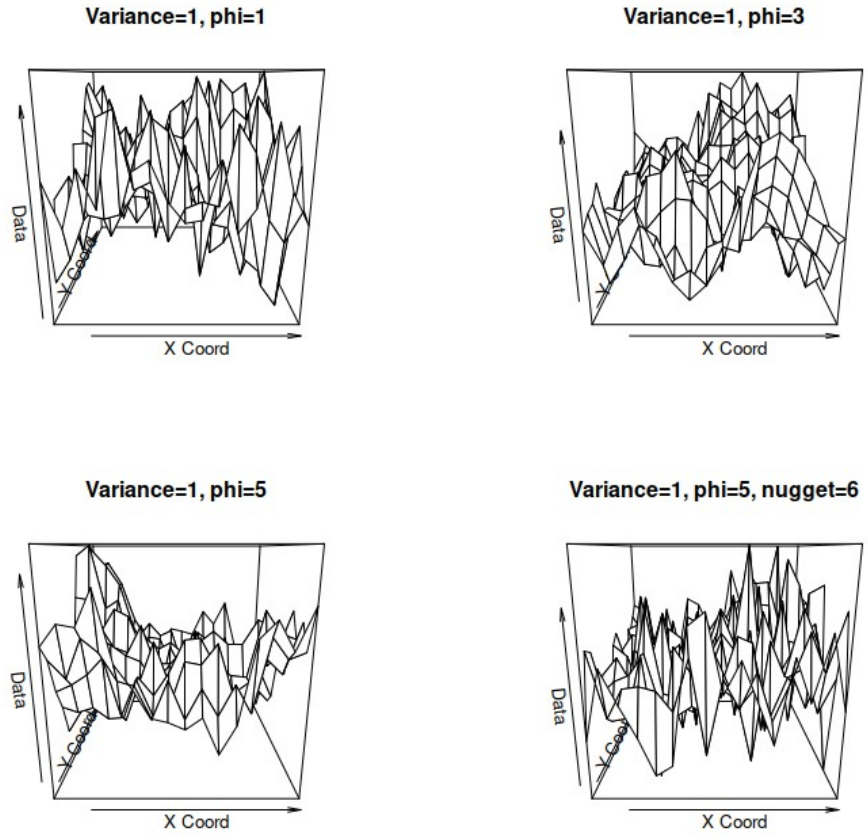
v)



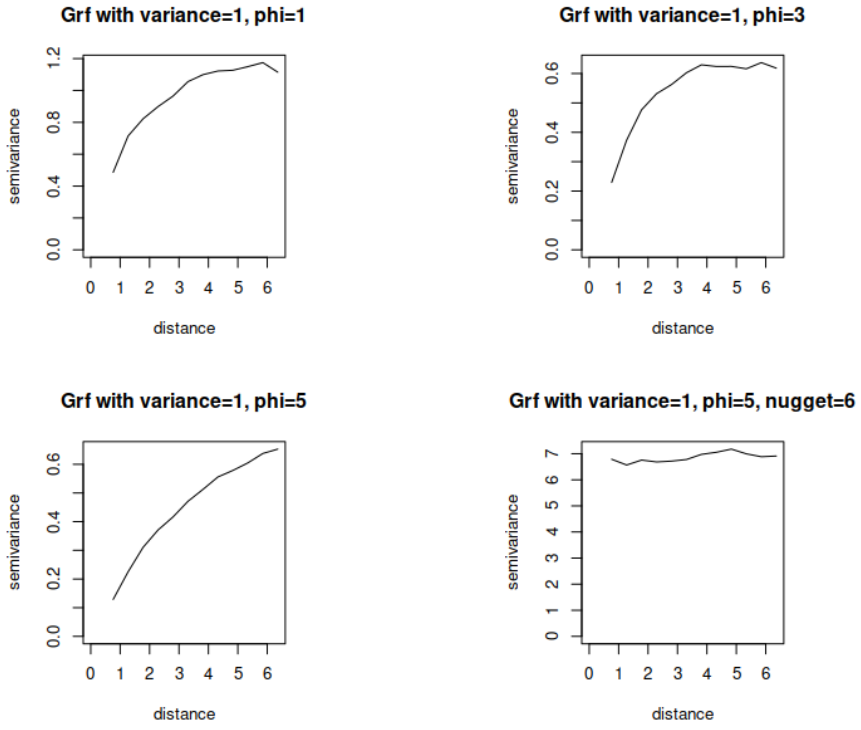
The median polish method seems to produce a more point-like correspondences for regions where high pH-levels were measured, unlike the methods used in part iv).

2.

i)

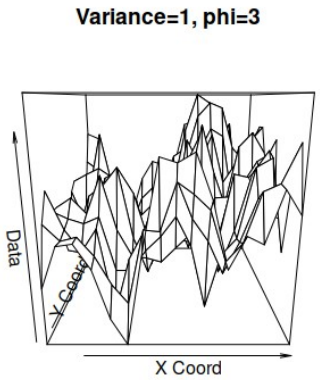

Variance=1, phi=1
Variance=1, phi=3
Variance=1, phi=5
Variance=1, phi=5, nugget=6
Data
X Coord

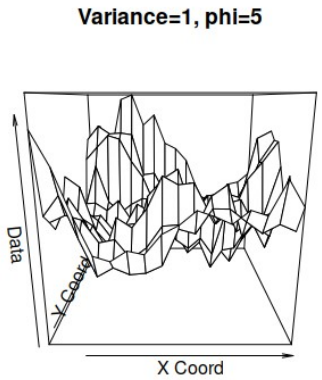
ii)

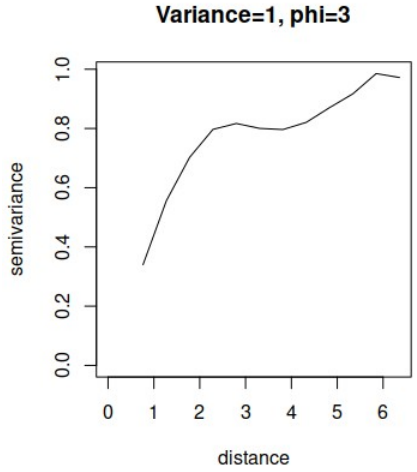

Grf with variance=1, phi=1
Grf with variance=1, phi=3
Grf with variance=1, phi=5
Grf with variance=1, phi=5, nugget=6
semivariance
distance

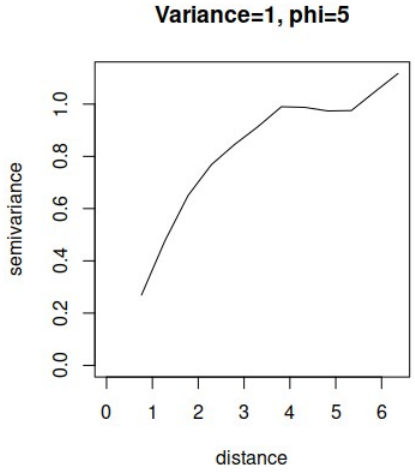
It seems that grf with variance=1 and phi=5 has the largest effective range while the same random field with nugget=6 has the smallest effective range. On the latter case, the samples seem to be mostly independent. Among the grf(1,1) and grf(1,3) the latter one should have theoretically larger effective range, but it is hard to notice based on empirical data.

iii)









Theoretically the semivariogram with spherical covariance structure has a range equal to phi parameter. This phi location in distance axis corresponds pretty well to the saddle point in the semivariance curve.