# Stats 544 - Homework 2

Franklyn Dunbar

Feb 14 2020

## Problem 1 - Fe/Mg

By hand, I calculated the cross-correlation values for h=(0,1) of both $C_{UV}$ and $C_{VU}$ respectively. We can see from the formula for cross-correlation that we will have non-symmetric results from the cross-co variance (the denominator, in this case). My results were $P_{UV} = 0.27$, and $P_{VU} = 0.47$.

The two cross correlations are different because the formula uses a different subset of data. $P_{UV}$ skips the U values in the last row, while $P_{VU}$ skips the vales V values in the last row. The subset we use to calculate the cross-correlation metric depends on the ordering. As the data set becomes larger, the cross correlations approach symmetry.

## LANDSAT

### EDA

For my EDA, I did a quick plot of the bands and took a co-variance matrix to quantify numerical relationships between the bands and get an idea of the variance in an individual bands measurements. This seemed like a reasonable way to do this as individual variances could measure how much information is captured (more variance could imply stronger features) and the co-variances could provide a numerical measure of agreement between the bands. The interpolated plots show agreement between bands 1, 2, and 3 while bands 4,5,and 6 show significantly more disagreement in their plots. The band which captured the largest surface variances was band 5 (513.94), and the largest co-variance was between bands 4 and 5 (333.19). The smallest surface variance was captured by band 2 (19.8), and the smallest co variance was between bands 6 and 2 (11.786). I feel that it would be reasonable to say that we could use these co-variance's to make a guess at the level of agreement between differing bands, and the interpolated plots compliment the covariance matrix in terms of variance in an individual plots scale and to some degree the level of agreement between each pair of plots.

## Spatial Auto-correlation

Included below are my computed correlogram, covariogram, and semi-covariogram for the respective spatial relationship of bands 2 and 4 with a lag of 1 along the respective x and y axes. Given that this is raster data, it seemed like a reasonable choice to use a distance and angular tolerance of 0. Band 2 shows similar agreement along both axes, while Band 4 shows stronger agreement along the x axis in comparison with the y axis.

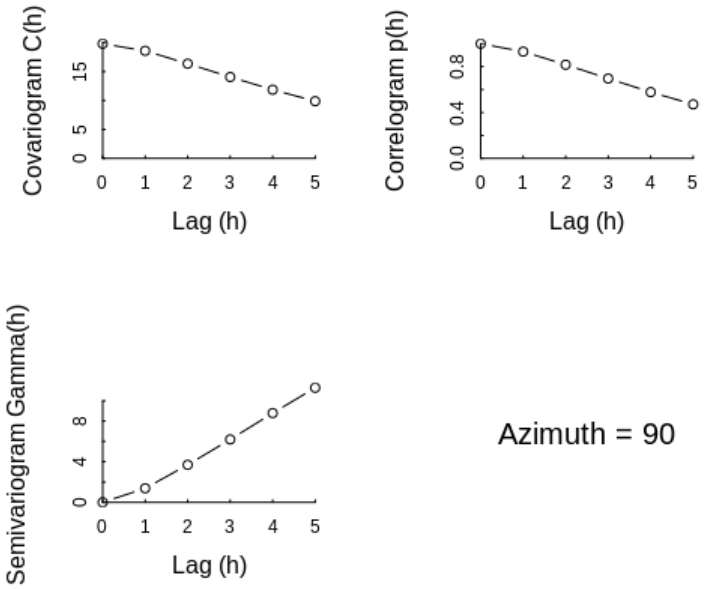



Figure 1: SA for a lag of 1 along x axis (Band 2)

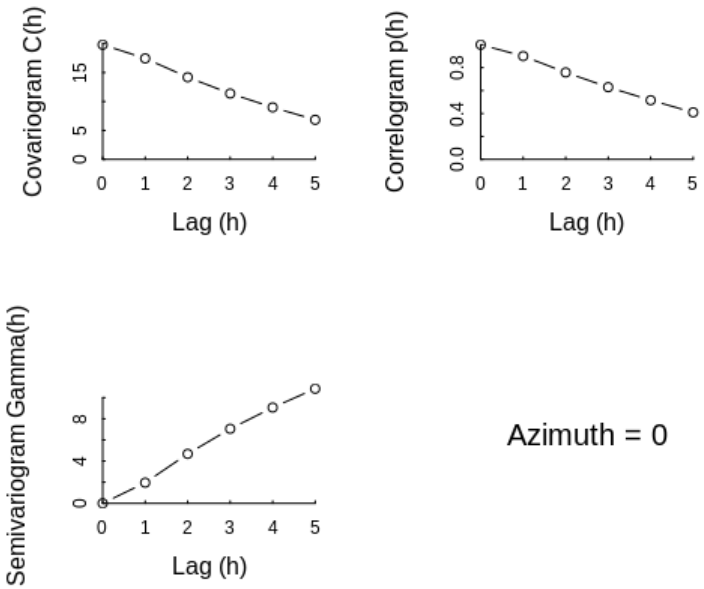



Figure 2: SA for a lag of 1 along y axis (Band 2)

Figure 3: Band 2

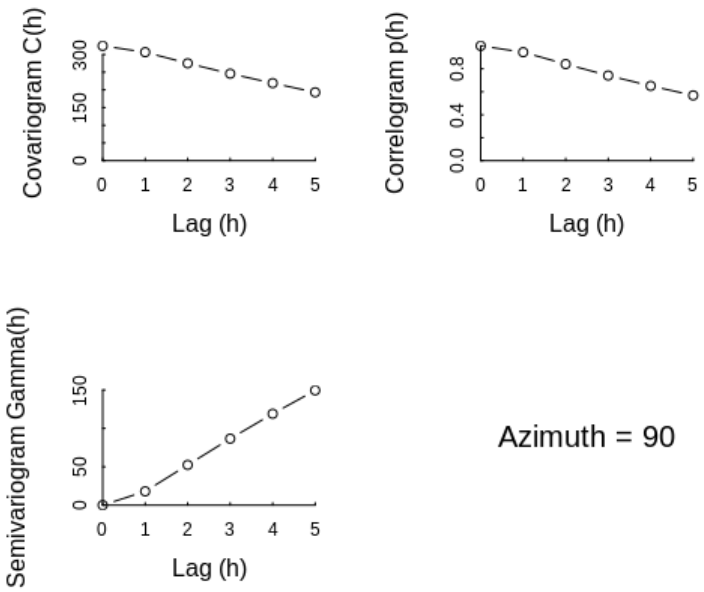



Figure 4: SA for a lag of 1 along x axis (Band 4)

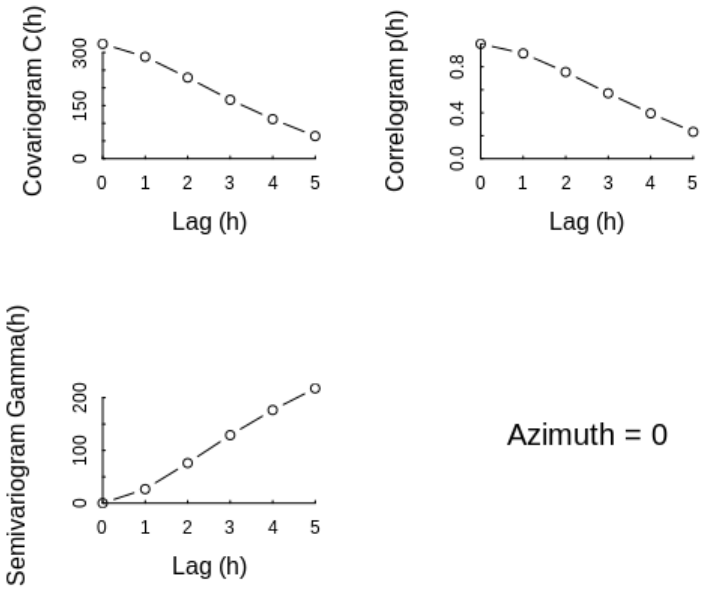



Figure 5: SA for a lag of 1 along y axis (Band 4)

Figure 6: Band 4

### Spatial Cross Auto-correlation

The cross correllogram, covariogram, and semi-variogram were calculated for the pairing of bands 2 and 4 along their respective axes. The cross metrics show reasonable agreement at low lags along both axes, with both showing a stronger negative slope after the first lag. Though there is some agreement with my EDA, I would say each metric exhibits a stronger negative slope than anticipated after looking at the co-variances. As an after-the-fact assessment, it would seem more ideal to have lower agreement between bands as this would allow for more unique features between images.

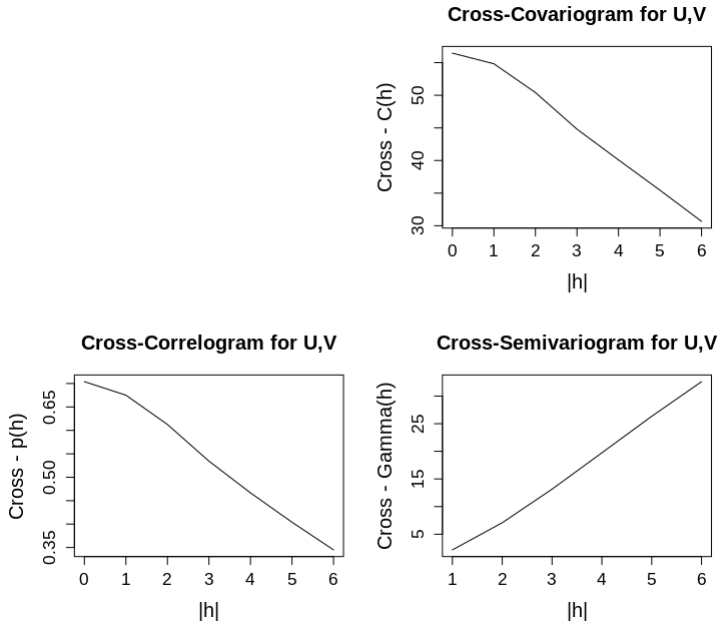



Figure 7: SCA for a lag of 1 along x axis (Band 2 v 4)

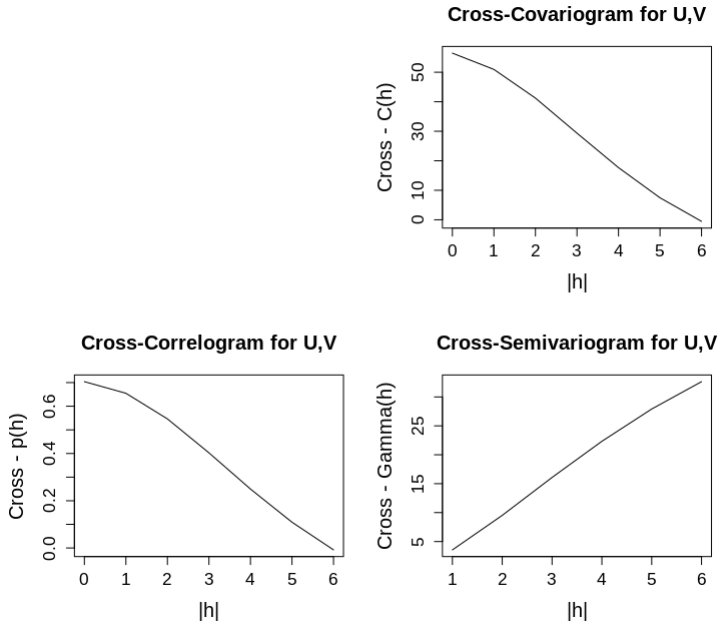


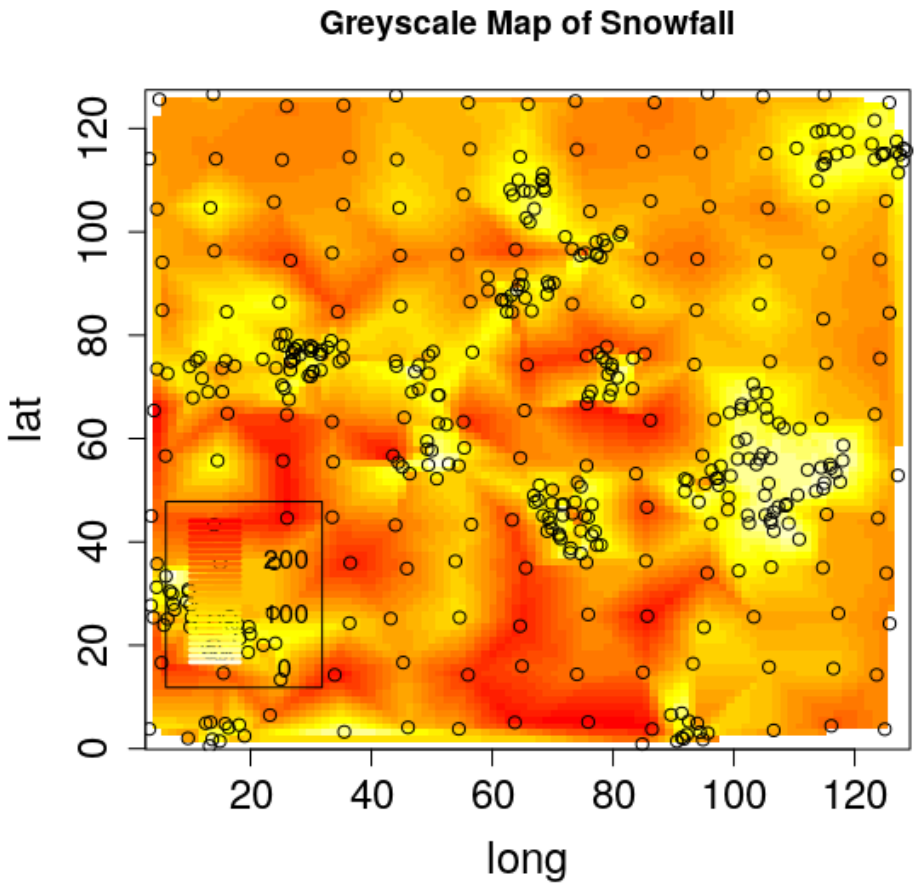

Figure 8: SCA for a lag of 1 along y axis (Band 2 v 4)

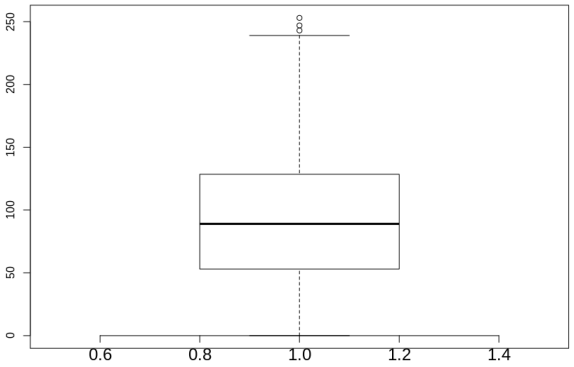
Figure 9: Band 4

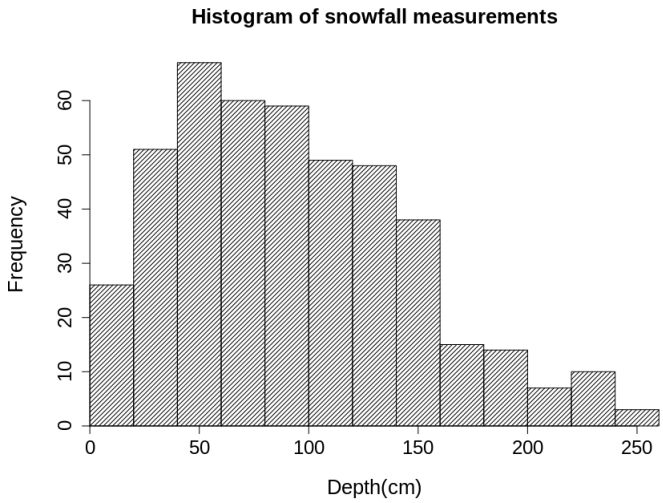
## Visual Greenness

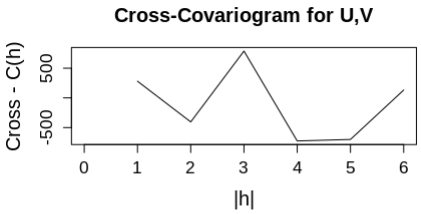
### EDA

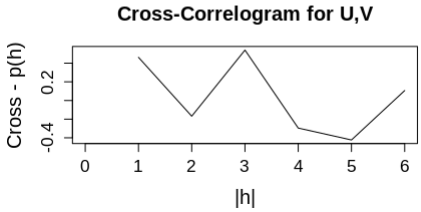
For my EDA I created a heat map, box plot, and histogram of the data. From this somewhat noisy heat map, it would be hard to argue that there are any clear spatial trends in the interpolated values, but one could say the values are clustered. From the box plot and histogram it would appear that the data is normal but skewed to the right (mean is larger than the inferred median), which could be due to the outliers displayed in the box plot. Taking calculating the spatial auto correlation along both axes proved to be somewhat tedious; only by tinkering with the distance and angular tolerances did any moderately interpretable results appear.

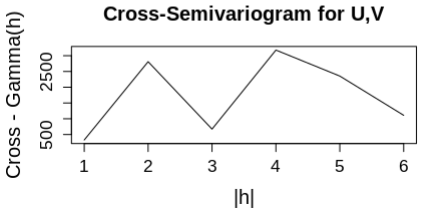


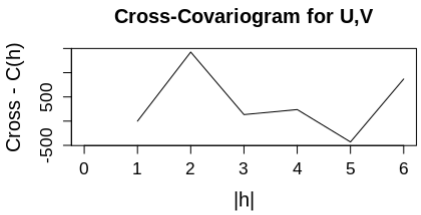



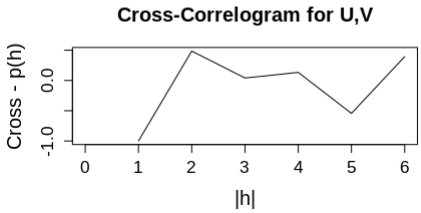

Histogram of snowfall measurements
Frequency
0
10
20
30
40
50
60
0
50
100
150
200
250
Depth(cm)

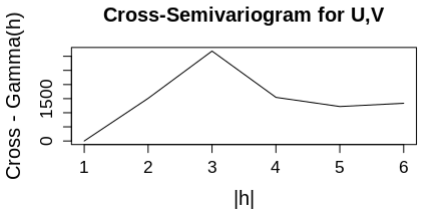

Cross-Covariogram for U,V
Cross - C(h)
-500
500
0
1
2
3
4
5
6
|h|


Cross-Correlogram for U,V
Cross - p(h)
-0.4
0.2
0
1
2
3
4
5
6
|h|


Cross-Semivariogram for U,V
Cross - Gamma(h)
500
2500
1
2
3
4
5
6
|h|


Cross-Covariogram for U,V
Cross - C(h)
-500
500
0
1
2
3
4
5
6
|h|


Cross-Correlogram for U,V
Cross - p(h)
-1.0
0.0
0
1
2
3
4
5
6
|h|

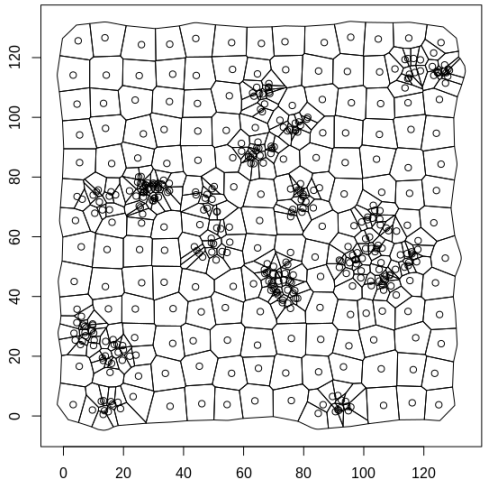

Cross-Semivariogram for U,V
Cross - Gamma(h)
0
1500
1
2
3
4
5
6
|h|

### Sample Mean

The sample mean of the snow depth is 93.73 cm. Given that the study was intentionally focused towards areas with low "visual greenness", it is clearly biased data. It could be argued that a strong negative correlation between VG and snow depth would imply that our sample mean is an underestimate.

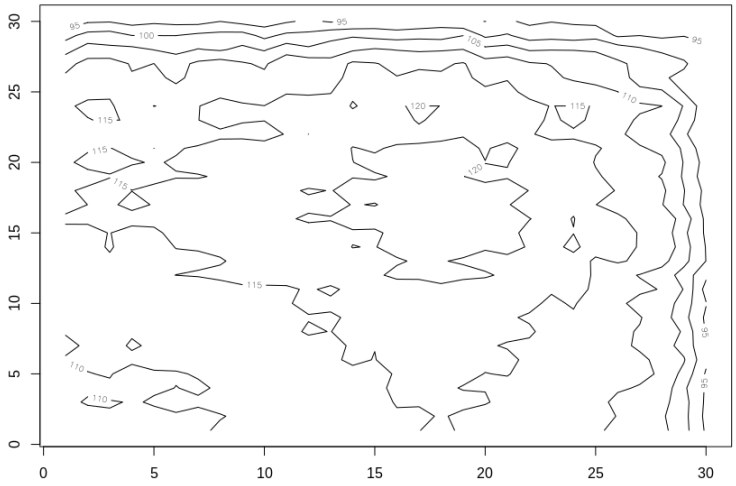
### Polygonal Declustering

With a peel of 7, the edge effects appear to be mitigated well. The global mean estimate was 125.7. We can see that in less clustered areas each point has a larger weight (in this case, cell area), and in more clustered areas cells have proportionally smaller cell areas.




### Cell Declustering

Given the contour plot produced as recommended in the homework, I would say that the most reasonable global estimate lies between 120 and 125. This is the largest contour allotted by the largest grid size (estimates become substantially smaller as grid sizes increase), because as the grid size increases there is convergence towards the sample mean.




### Polygonal vs Cell Declustering

Do the use of these so-called declustering methods seem to improve global estimates of snow depth for the region? yes. By accounting for the bias induced by the oversampling of areas with a particular characteristic, we can see that these reasonable approaches to estimating a global mean are substantially different than our initial sample mean.

## Gravity Measurements

| | n | $\bar{y}$ | s | CV | min | Q1 | M | Q3 | max | $\hat{\rho}_{\hat{y},y}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| True | 241 | -157 | 48 | 28.0 | -244 | -199 | -159 | -117 | -66 | |
| Triangulation | 241 | -156 | 44 | 28.2 | -237 | -196 | -155 | -119 | -74 | 0.92 |
| True | 255 | -157 | 48 | 30.6 | -244 | -199 | -159 | -117 | -66 | |
| Polygonal | 255 | -158 | 48 | 30.4 | -244 | -199 | -167 | -116 | -66 | 0.97 |
| LSM (r=$2{\cdot}10^4$) | 255 | -157 | 47 | 30.0 | -244 | -199 | -160 | -117 | -60 | 0.98 |
| LSM (r=$4{\cdot}10^4$) | 255 | -157 | 45 | 28.7 | -228 | -200 | -158 | -118 | -74 | 0.98 |
| LSM (r=$6{\cdot}10^4$) | 255 | -157 | 44 | 28.0 | -225 | -199 | -156 | -118 | -75 | 0.97 |
| LSM (r=$8{\cdot}10^4$) | 255 | -157 | 42 | 26.8 | -224 | -200 | -156 | -117 | -80 | 0.96 |
| LSM (r=$10{\cdot}10^4$) | 255 | -158 | 40 | 25.3 | -221 | -199 | -155 | -120 | -87 | 0.95 |
| IDS (r=$2{\cdot}10^4$) | 255 | -157 | 47 | 30.0 | -244 | -199 | -161 | -117 | -69 | 0.98 |
| IDS (r=$4{\cdot}10^4$) | 255 | -157 | 46 | 29.2 | -232 | -199 | -159 | -118 | -72 | 0.98 |
| IDS (r=$6{\cdot}10^4$) | 255 | -157 | 46 | 29.2 | -228 | -200 | -157 | -118 | -73 | 0.98 |
| IDS (r=$8{\cdot}10^4$) | 255 | -157 | 45 | 28.7 | -227 | -199 | -157 | -119 | -74 | 0.98 |
| IDS (r=$10{\cdot}10^4$) | 255 | -157 | 44 | 28.0 | -226 | -199 | -157 | -118 | -77 | 0.98 |

Table 1: Summary statistics for the five estimation methods. I worked with Tommy on this one.

| | n | $\bar{y}$ | s | min | Q1 | M | Q3 | max | MAE | MSE |
|---|---|---|---|---|---|---|---|---|---|---|
| Triangulation | 241 | -1 | 19 | -88 | -4 | 3 | 9 | 44 | 11 | 343 |
| Polygonal | 255 | -1 | 11 | -33 | -8 | -1 | 6 | 33 | 9 | 122 |
| LSM (r=$2{\cdot}10^4$) | 255 | 0 | 9 | -33 | -4 | 0 | 5 | 33 | 6 | 73 |
| LSM (r=$4{\cdot}10^4$) | 255 | 0 | 11 | -38 | -8 | -1 | 6 | 37 | 8 | 115 |
| LSM (r=$6{\cdot}10^4$) | 255 | 0 | 12 | -31 | -8 | -1 | 7 | 44 | 10 | 150 |
| LSM (r=$8{\cdot}10^4$) | 255 | 0 | 14 | -29 | -10 | -2 | 8 | 44 | 11 | 201 |
| LSM (r=$10{\cdot}10^4$) | 255 | -1 | 16 | -34 | -10 | -1 | 9 | 45 | 13 | 257 |
| IDS (r=$2{\cdot}10^4$) | 255 | 0 | 9 | -25 | -5 | -1 | 5 | 31 | 6.5 | 74 |
| IDS (r=$4{\cdot}10^4$) | 255 | 0 | 9 | -30 | -6 | -1 | 5 | 33 | 6.9 | 83.4 |
| IDS (r=$6{\cdot}10^4$) | 255 | 0 | 10 | -29 | -6 | -1 | 5 | 37 | 7.3 | 92.3 |
| IDS (r=$8{\cdot}10^4$) | 255 | 0 | 10 | -29 | -7 | -1 | 5 | 38 | 7.7 | 102.5 |
| IDS (r=$10{\cdot}10^4$) | 255 | 0 | 11 | -28 | -7 | -1 | 5 | 38 | 8.1 | 112.5 |

Table 2: Summary statistics of the residuals given by the five estimation methods. I worked with Tommy on this one.

| | Polygonal | ULSM | IDS | Triangulation |
|---|---|---|---|---|
| mean-res | -1 | 0 | 0 | 0 |
| SD-res | 11 | 11 | 9 | 19 |
| min-res | -33 | -38 | 30 | -88 |
| Q1-res | -8 | -8 | -6 | -4 |
| M-res | -1 | -1 | -1 | 3 |
| Q3-res | 6 | 6 | 6 | 9 |
| Max -res | 33 | 37 | 33 | 44 |
| MSE | 122 | 155 | 83.4 | 343 |
| MAE | 9 | 8 | 6.9 | 11 |

### Comparing Global Estimates

It would appear from the cross validation results that the best general purpose method for this particular data set is Inverse distance weighting. This method has the lowest absolute error across all radii for both mean and standard deviation estimates.

The Inverse Distance Method uses all sample points to weight each sample, which are inversely proportional to their respective distance from the prediction site, as compared to the Local Sample Mean which simply takes the mean of all samples within a given region. As the radii increases, LSM's assumption that all samples within the radial distance from the prediction point directly affect with global mean estimate

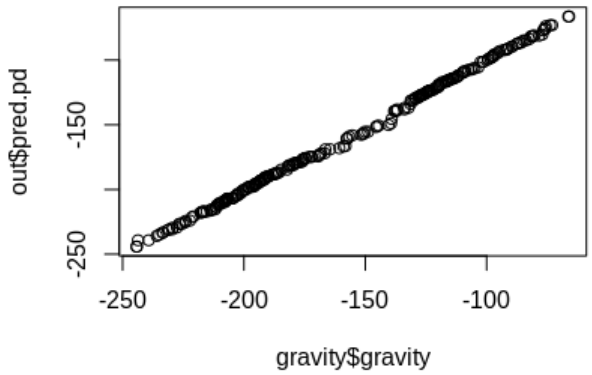
fails because in fact they may not necessarily do so.

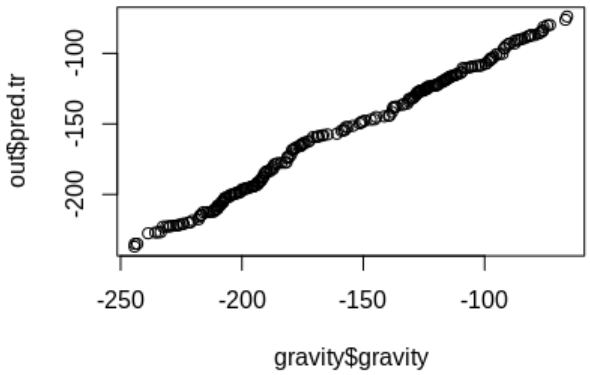
The following table was created for the meta-stats for each respective method at powers of 0.2, 0.5, 1 2, 5, and 10 at a fixed radial distance of 20,000. From this table we can see that a power of 0.5 seems to perform the best.

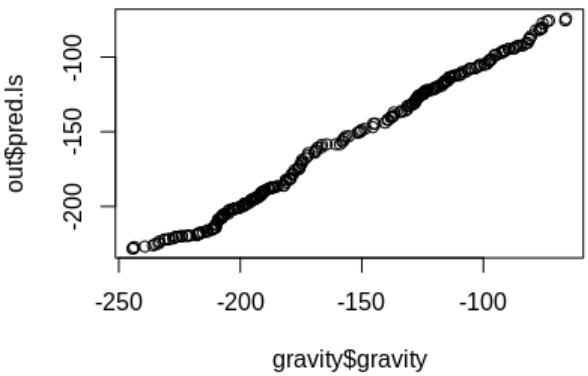
| | 0.2 | 0.5 | 1 | 2 | 5.0 | 10.0 |
|---|---|---|---|---|---|---|
| Mean-res | 0 | 0 | 0 | 0 | 0 | -1 |
| SD-res | 9 | 8 | 10 | 9 | 9 | 10 |
| MSE | 72.3 | 72 | 101 | 73.9 | 85.6 | 101 |
| MAE | 6.2 | 6.2 | 7.8 | 6.5 | 7.1 | 7.8 |
| Corr | 0.984 | 0.984 | 0.978 | 0.984 | 0.981 | 0.978 |

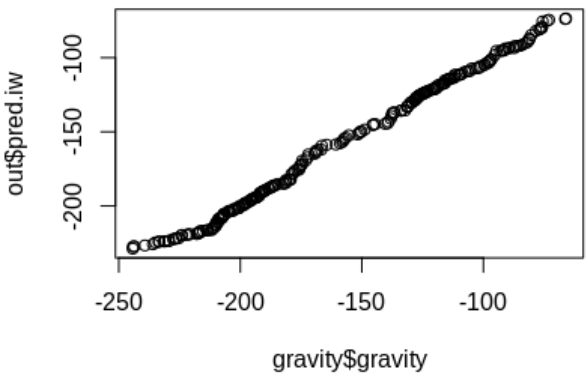
## QQ Plots

We can see in the plots below that our predictions are in good agreement with the true gravity values. There is some discrepancy at the higher values for triangulation, local sample mean and inverse distance weighting. Polygonal de-clustering is clearly the most accurate estimation method.



### Scatter Plots

The lowess spline fit relative to the 45 degree is indicative of global and/or conditional bias. The plots below show that Polygonal declustering does not have any particularly obvious conditional bias, whereas triangulation shows conditional bias for the higher values and the inverse distance methods have some bias at the median values.