Rodrigo Arriaza
Arvin Rastegar
Johanna Weiss

Project

Question 1

A first numerical summary of the data reveals that the elevation of the islet ranges from a minimum of -9.5 meters, so below sea level, to a maximum of 39.6 meters with a mean of 15.97 meters and a median of 16.35 meters. These characteristics are calculated from a total of 148 observations. The histogram containing all the elevations can be seen in 1. The data was checked for outliers, however, the extreme points in the data are coherent with the extrema to be expected in the elevations of a landscape and are, therefore, not removed from the data set.

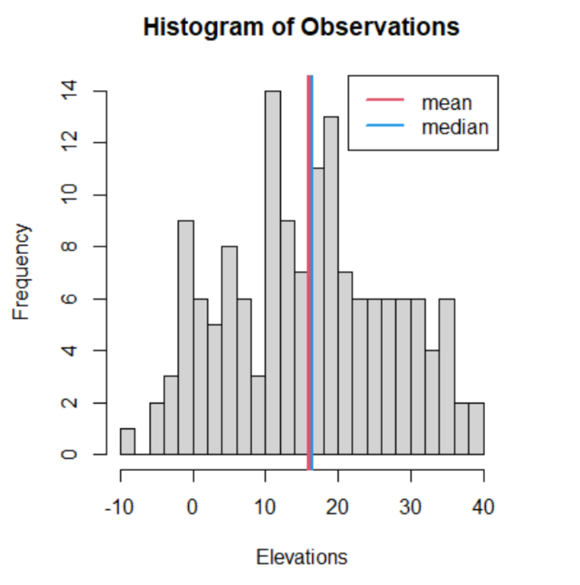


Figure 1: Histogram of the elevation data

The distribution of the observations is approximately symmetric, slightly skewed to the left as the median is a little greater than the mean. The interquartile range of the elevations is from 7.67 meters to 24.18 meters, which means that 50% of all the observations lay in that interval. The observations plotted with respect to their locations can be seen in 2. The plot suggests that the wider part of the islet is marked by two mountains, whose elevation lowers towards the narrow tip of the islet on the opposite side. In the center of the islet is a smaller elevation, possibly a hill.

Figure 3 shows the marginal relation of the longitude and latitude with the elevation data. This is a binned representation of the data, in which the distribution of each bin of coordinates is shown in a boxplot. Looking at the boxplot with respect to the y-coordinates, it can be seen that data is fairly spread out which can be seen by larger interquartile ranges and whiskers. Towards the center of the island, the variance seems to decrease a little. The boxplots also show the overall shape of the islet from the x-coordinate perspective, namely

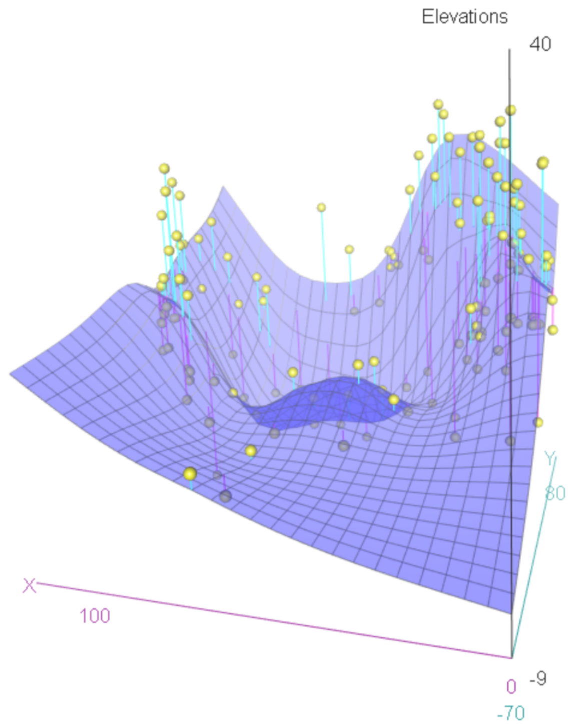


Figure 2: 3D plot of the elevation data

the two mountains at the edge of the islet and a smaller elevation in between. From the y-coordinate perspective, it can be seen that the island has high elevations in the northern end and decreases strongly in height towards the southern tip of the islet. The tip of the islet is very narrow and contains very few observations and therefore practically no variance in the boxplots, whereas the wider end of the islet has larger boxplots with more spread out interquartile range and whiskers.

Question 2

To explore the small scale variability of the elevation data, we investigate if there is a trend. As a first model, a linear regression is fitted to the data. However, as one can assume from the scatter plot in the previous section, the data does not exactly follow a linear relationship. In fact, only about 8% of the variation can be explained by a linear model. Using a polynomial regression of order 3 seems to be a significantly better representation of the data. After removing the trend, the residual data is explored with regard to the small variability. Figure 4 (left) shows the cloud variogram of the data representing the dissimilarities between value pairs against their distance. From the plot, it can be seen that most of the data is located in the bottom left corner, which possibly indicates that closely located data points are more similar than those that are farther part. It needs to be said, however, that due to the shape

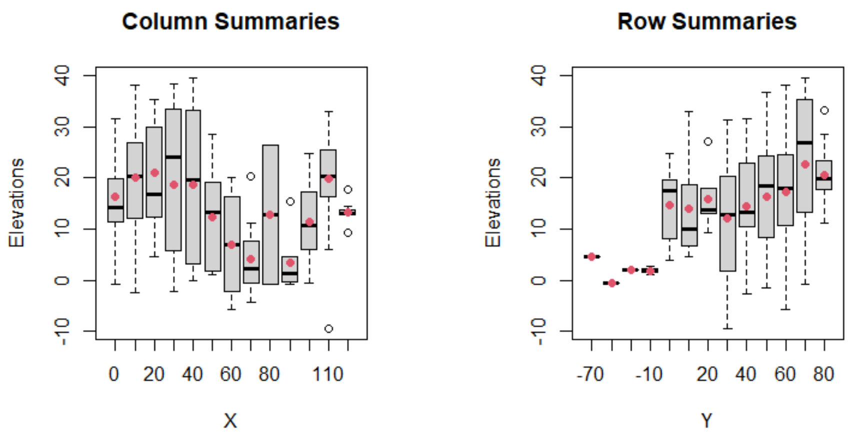


Figure 3: Boxplots of the elevation data

of the islet, there are more observations laying closely together.

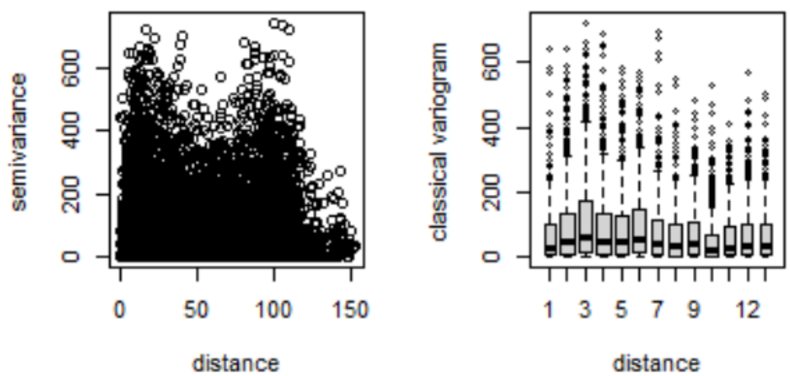


Figure 4: Boxplots of the elevation data

Figure 4 (right) shows the cloud variogram binned into classes according to their distance and the average dissimilarity resulting from each of the bins. It can be seen that with all boxplots, the boxes extend nearly to the lower extreme, which shows that the data is fairly consistent below the median. The median is varying slightly for the different boxplots, but overall stays on one level. The longer whiskers on the upper side suggest that there is a larger variance among greater values of dissimilarity. Since this holds for all the boxplots, it cannot be concluded that a smaller distance of the data pairs indicates higher dissimilarity of the elevation.

Finally, similarly as seen in 3, if we add trend lines to the elevation points we can see that there is a clear trend in 5, specially in the y-axis, so the further north we go the higher is the

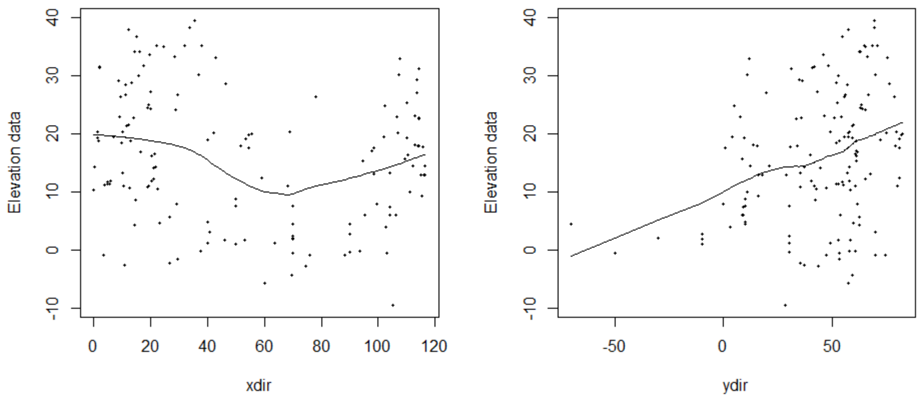


Figure 5: Trend in the elevation by x-coordinates(right) and y-coordinates(left)

elevation of the islet; whereas in the x-axis the further away we move from the middle, the higher the elevation. Thus we can confirm that the process is not constant.

Question 3

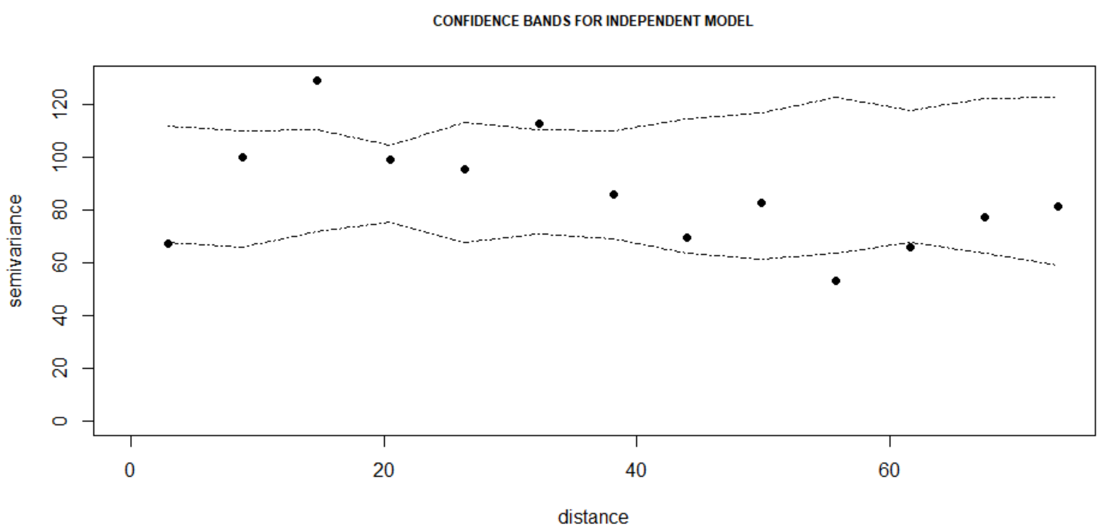


Figure 6: Semivariance vs distance plot to test spatial correlation

To explore the spatial independence we run a Monte Carlo test from random permutations of the residual data. From them we calculate variogram envelopes to test the hypothesis of constant variogram which would imply no spatial correlation.

As we can see in 6, not all the points are inside the confidence interval and therefore we can conclude that the elevation is not entirely independent from the coordinates, i.e. there is a spatial correlation. This seems to be mainly present for smaller distances, and approximately independent for distances between 20 and 50 meters. This goes along with the hypothesis that closely located data points are more similar than distant ones. For larger distances the data seems to show a negative correlation which might, however, be due to the low number of observations at the tip of the islets which affects the behavior of the correlations with respect to larger distances.

Question 4

We proposed the following four variograms with these parameters and the corresponding sum of squares:

Model	Nugget	Sill	Range	R^2
Exponential	0.00	95.60	2.25	315.72
Gaussian	0.00	95.5	2.66	312.66
Spherical	25.06	70.47	6.91	312.65
Matern	0.00	95.59	1.50	313.47

Table 1: Summary of four theoretical variograms

As we can see from Table 1 the best fitting variogram is the spherical with a nugget of 25.06 a sill of 70.47 and a range of 6.91.

From all of the variograms we see that the value of the range is very low, which again confirms that less distant points have a stronger spatial correlation than more distant ones. Also we see a decrease in the semivariance after the distance 40, which may be explained by the existence of a valley in the middle.

Question 5

As discovered in Question 2, the large scale variation of the geostatistical process is not constant, as it can be approximated by a polynomial regression model. This means, for the prediction of the elevation, the universal Kriging can be used. Using the two best performing variograms from Question 4, the elevations of the whole area of study can be predicted using the Kriging. The results can be seen in Figure 8.

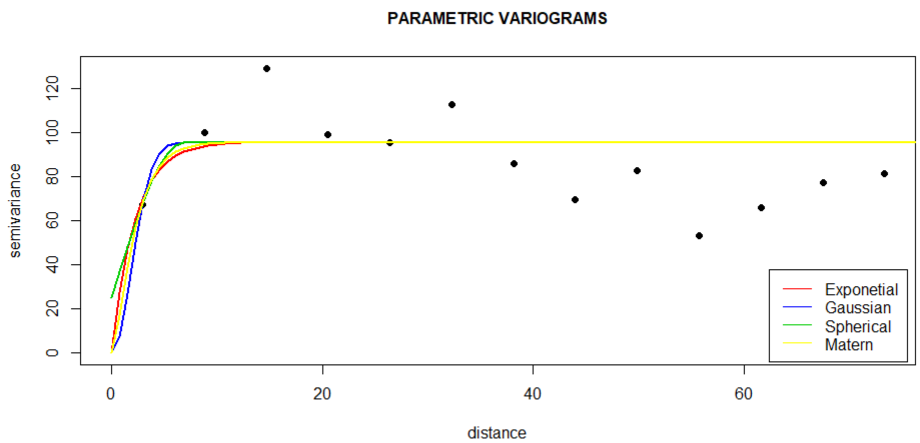
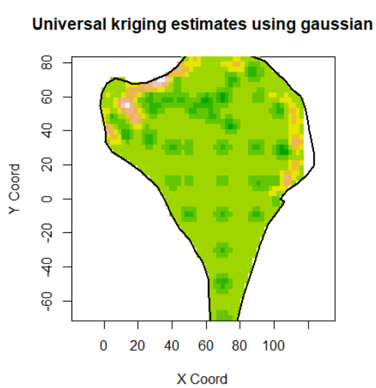


Figure 7: Parametric variograms

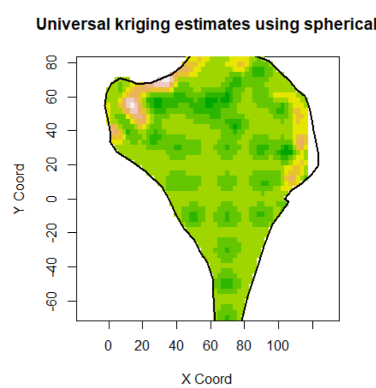
Model	VC1	VC2	VC3
Gaussian	-0.0335	1.0895	9.1253
Spherical	-0.0295	1.0784	9.0687

Table 2: Numerical comparison of two kriging

In 2 we can see that both models provide a good fit, since $VC1 \approx 0$, $VC2 \approx 1$ and VC3 has a value of 9 which can be considered to be low. Additionally, the spherical model is a slightly better fit than the gaussian one, which can also be seen graphically in 8.



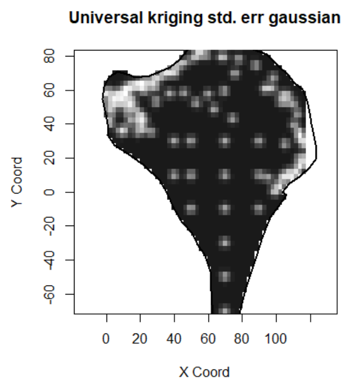
(a) Predictions using gaussian model



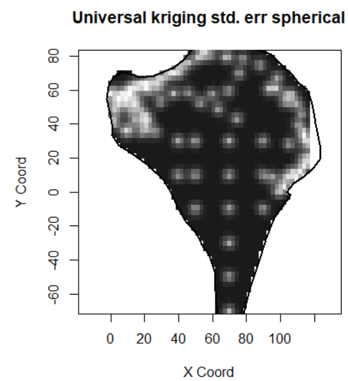
(b) Predictions using spherical model

Figure 8: Comparing the best two variograms predictions

Standard Errors



(a) Standard errors of gaussian model



(b) Standard errors of spherical model

Figure 9: Comparing standard errors of best two models

From 9 we see that the standard errors of both models are very similar and are higher towards the edges of the islet. This can be explained as we see from 8 that at the top edges of the islet the elevation changes more abruptly than in the middle of it. This is the cause why the variance of the predictions are higher.