Assignment 1 Spatial Epidemiology

Elevation of a small islet in Delta d'Ebre

Ferrara Lorenzo, Lucchini Marco

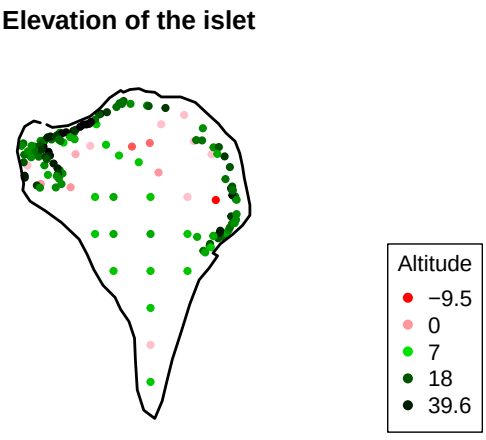
25-10-2022

Description

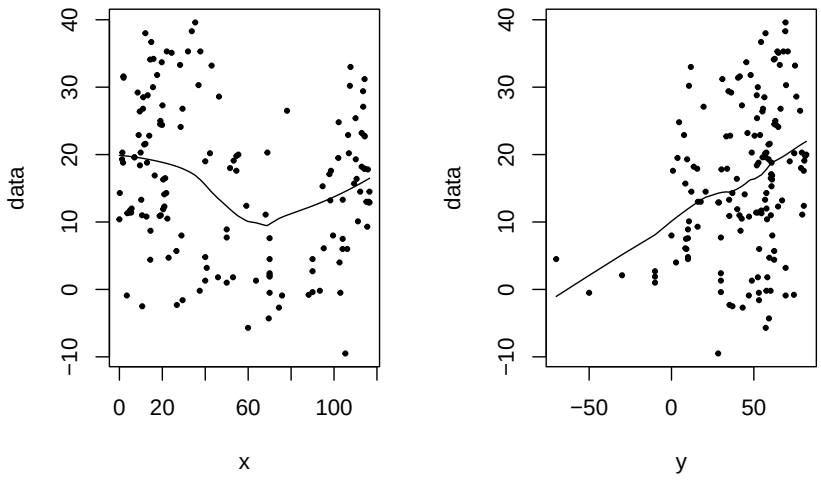
The data were collected during a study of the settlement pattern of common terns on a small islet in the Delta d'Ebre (Hernandez and Ruiz, range3), particularly in the mouths of the Ebre river. The islet was inspected at two-day intervals throughout the range0 breeding season. The data include the location of each nest, its elevation above sea level, and elevations at a number of additional points (points without nest) on the islet. In the file called elevationsIslet.txt, contains the information of the coordinates and elevation above sea, and in file, called poly84.txt contains the coordinates of the borders of the islet. The aim is to predict the elevation above sea level along the small islet using a kriging interpolation.

1) Explore the requirement of stationary mean of the process. In case this requirement is not met, detrend the data to ensure that the process is stationary in mean. Discuss the results and show the plot of the results

Firstly, we plot the locations to see the spatial distribution of the data



We also look at the distribution in relation to the x-coordinates (E-W) and y-coordinates(N-S).



The plots show a concentration of high values in the extreme east and west of the islet and we can observe a greater density in the north.

The process doesn't seem stationary, indeed there is a clear quadratic trend along the x direction. In addition we try using the y direction as regressor, to find a possible linear or quadratic trend.

```
##
## Call:
## lm(formula = data ~ x + y + I(x^2) + I(y^2), data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.639  -6.417  -0.964   6.024  22.744
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.4502678  2.8429177   6.138 7.78e-09 ***
## x           -0.4938096  0.1103379  -4.475 1.55e-05 ***
## y            0.0349038  0.0570014   0.612    0.541
## I(x^2)       0.0040706  0.0009267   4.393 2.17e-05 ***
## I(y^2)       0.0019508  0.0008841   2.206    0.029 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.19 on 143 degrees of freedom
## Multiple R-squared:  0.2026, Adjusted R-squared:  0.1803
## F-statistic: 9.082 on 4 and 143 DF,  p-value: 1.446e-06
```

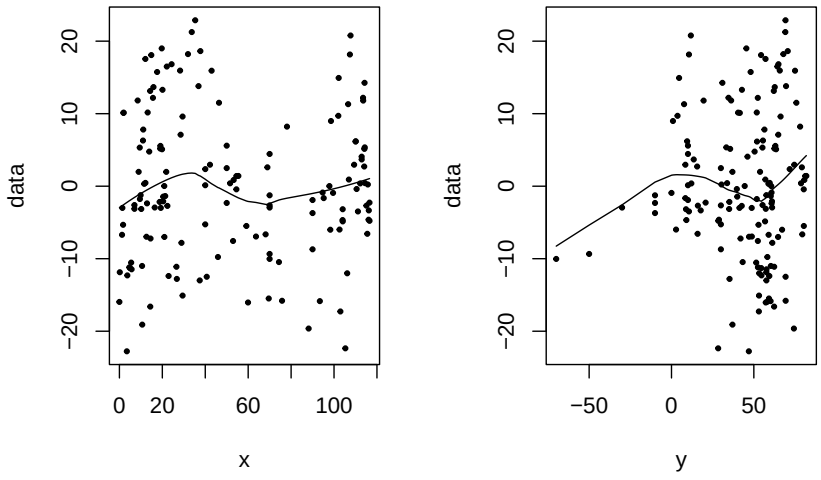
The linear term in y doesn't seem significant, so we remove it.

```
##
```

```
## Call:
## lm(formula = data ~ x + I(x^2) + I(y^2), data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.768  -6.656  -1.081   6.166  22.879
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.3842839  2.3938539   7.680 2.23e-12 ***
## x           -0.5192823  0.1019735  -5.092 1.09e-06 ***
## I(x^2)       0.0042621  0.0008705   4.896 2.59e-06 ***
## I(y^2)       0.0023697  0.0005587   4.241 3.96e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.17 on 144 degrees of freedom
## Multiple R-squared:  0.2005, Adjusted R-squared:  0.1838
## F-statistic: 12.04 on 3 and 144 DF,  p-value: 4.456e-07
```

Now we have obtained a model in which all regressors seem to be significant. We save the residuals of our linear model and look at the de-trended data:

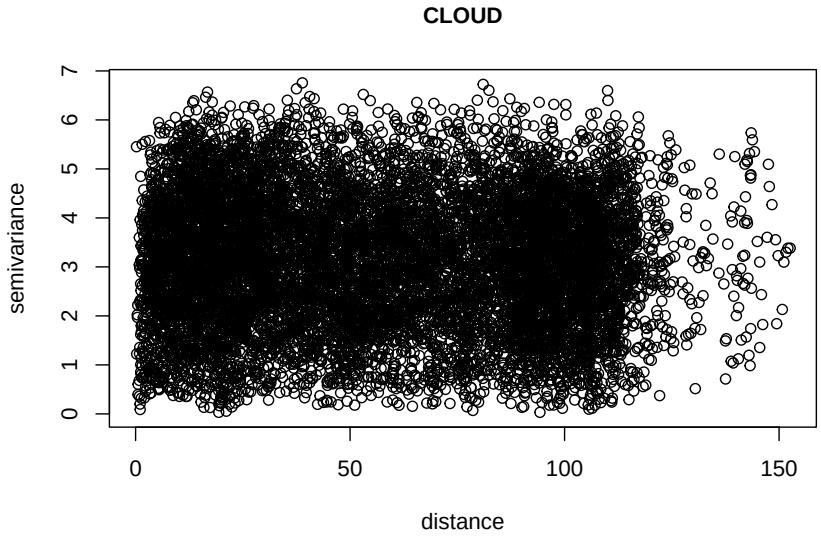
What we now obtain is a more homogeneous distribution of the data around the value 0.



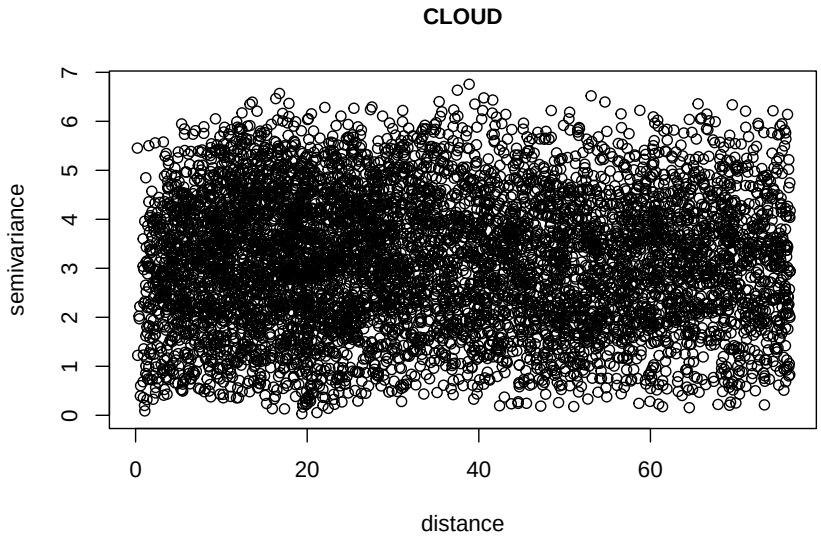
We don't notice any particular trend so this new dataset seems stationary!

2) Explore the spatial dependence of the elevation variable using the variogram cloud and bins and the empirical variogram. Discuss the results and plot them

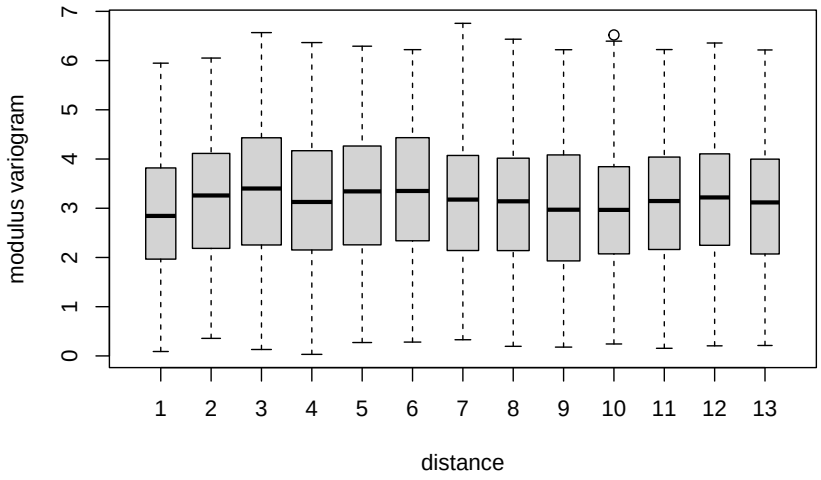
We try to analyse the spatial dependence through a Variogram Cloud, using the Robust Estimator (since the Classical one might consider outliers some values which are not necessarily outliers)



The variability at small distances appears a bit greater than that for larger distances. Therefore we reduce the density of the plot by reducing the maximum distance over which the variances are calculated.



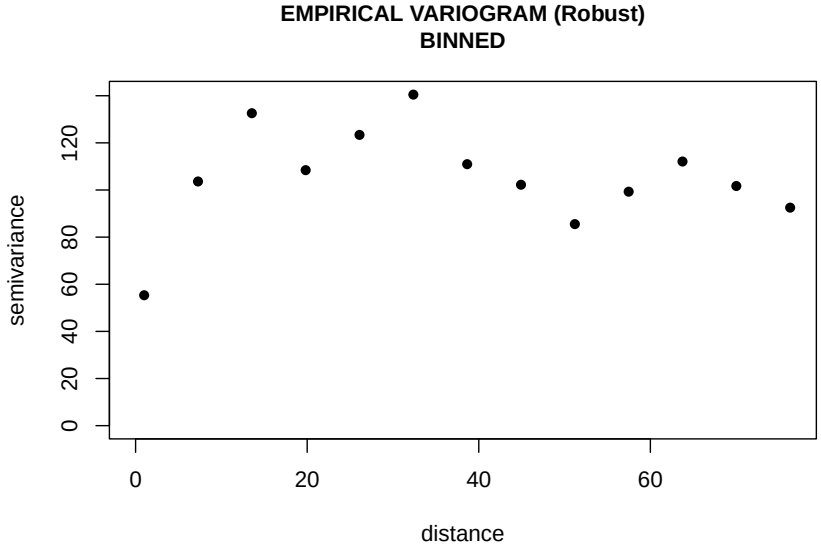
We also have a look at the boxplots:



All the bins have at least $pairs.min = 30$ observations each, indeed they have:

```
##  [1] 422 650 731 755 660 514 460 473 497 447 457 415 377
```

Therefore the Empirical Variogram is:

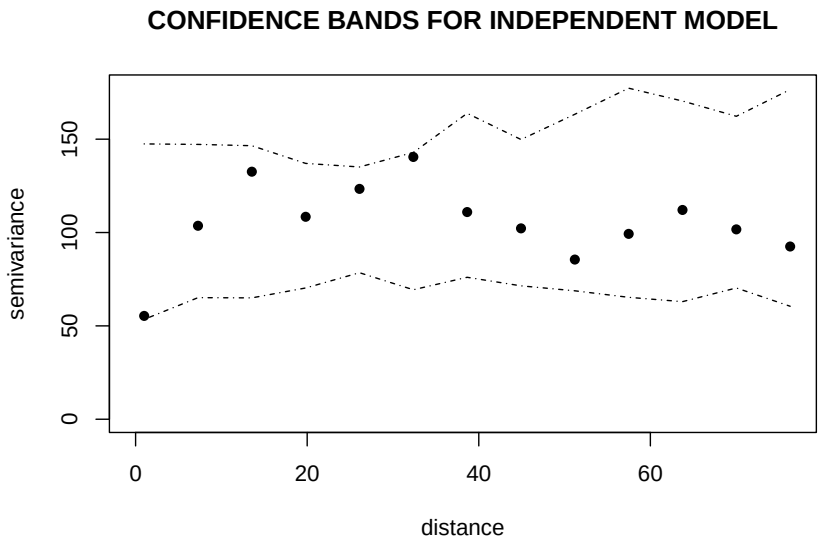


From what we can observe in the variogram of the residuals, the values increase until a certain distance, and then they are constant. That's ok since it's the behaviour that we expect from a stationary variogram.

3) Check the hypothesis of the spatial independence

To check the hypothesis of the spatial independence we use a Monte Carlo approach

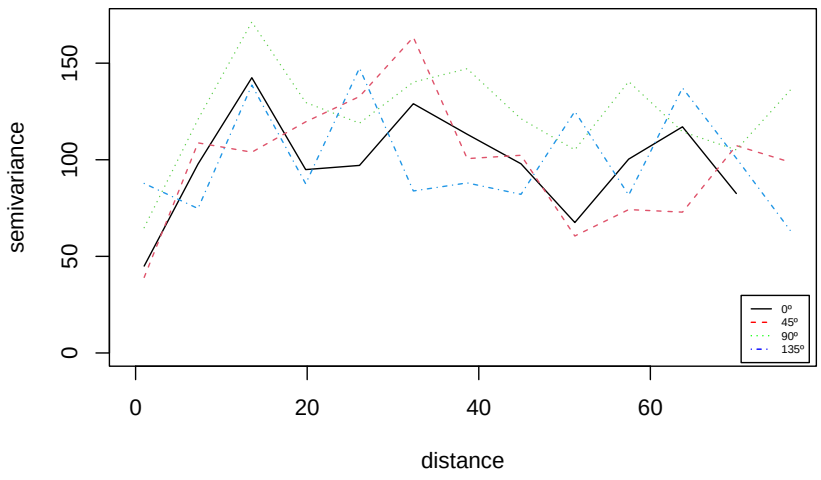
```
## variog.env: generating 2000 simulations by permutating data values
## variog.env: computing the empirical variogram for the 2000 simulations
## variog.env: computing the envelops
```



All the values of the empirical variogram are inside the envelope, therefore the process has no spatial dependence.

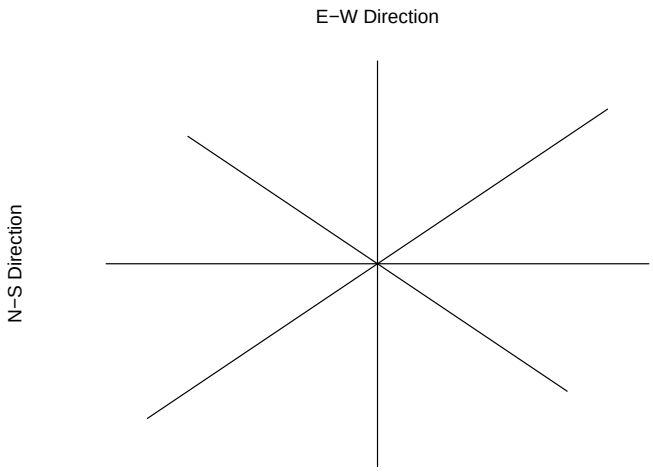
4) Check the isotropy property of the process. Comment the results, it's not necessary to overcome the anisotropy.

To check the anisotropy we need to compute the directional variogram in the 4 main directions: 0º, 45º,90º and 135º.



The directional variograms don't seem to be perfectly overlapping: they might have a different sill, in particular the variogram seems to have a higher value along the 90° direction, so we have Geometrical Anistropy.

Let's also analyse the range observing the Rose Diagram:



We notice that the Rose diagram is not perfectly circular: it is slightly elliptical, with major range in the 45° and 90° direction, so we also have Zonal Anistropy.

In conclusion this is an example of combined anisotropy, so we can't make the assumption of isotropic process, which is necessary for the correct use of the kriging techniques, since the theoretical variograms used for krigring are based on isotropic models.

5) Propose four theoretical variograms and estimate the parameters via restricted maximum likelihood or weighed least square. Select the two variograms which best fit the data. Explain the parameters of the chosen variogram (sill, nugget, range and kappa).

We'll use a Restricted Maximum Likelihood approach

Exponential

```
## likfit: estimated model parameters:
##     beta    tausq  sigmasq      phi
## "-2.631" "34.598" "71.657" " 6.083"
## Practical Range with cor=0.05 for asymptotic range: 18.2225
##
## likfit: maximised log-likelihood = -531.4
```

Gaussian

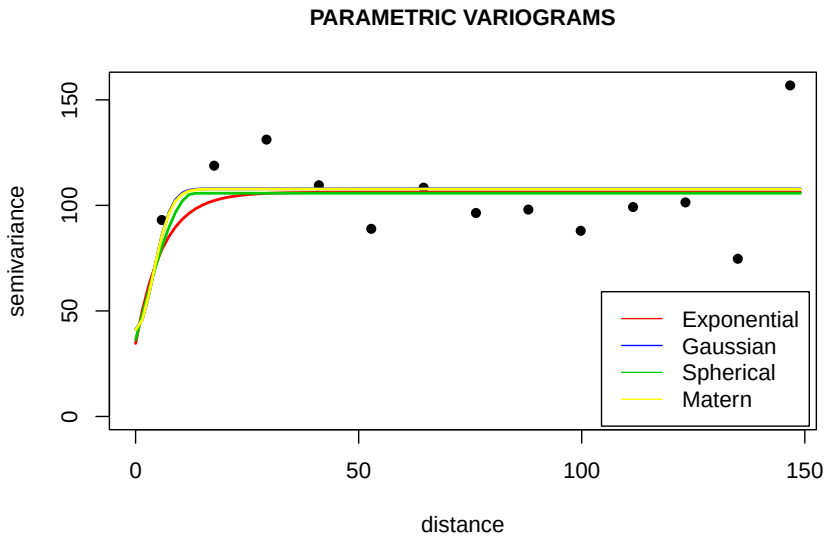
```
## likfit: estimated model parameters:
##     beta    tausq  sigmasq      phi
## "-2.786" "41.157" "66.712" " 5.601"
## Practical Range with cor=0.05 for asymptotic range: 9.693811
##
## likfit: maximised log-likelihood = -529.8
```

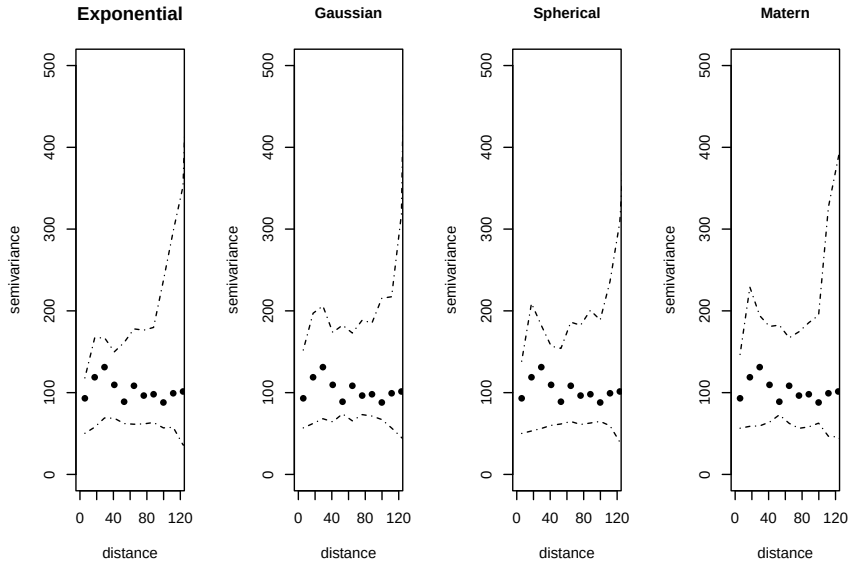
Spherical

```
## likfit: estimated model parameters:
##     beta    tausq  sigmasq      phi
## "-2.758" "36.363" "69.395" "12.992"
## Practical Range with cor=0.05 for asymptotic range: 12.99251
##
## likfit: maximised log-likelihood = -530.3
```

Matern

```
## likfit: estimated model parameters:
##      beta     tausq   sigmasq       phi     kappa
## "-2.7802" "40.9994" "66.7522" " 0.5757" "24.1330"
## Practical Range with cor=0.05 for asymptotic range: 9.889255
##
## likfit: maximised log-likelihood = -529.8
```





All the simulated variograms contain the empirical variogram, so we don't have evidence to exclude any model.

Let's compare the different models:

```
##          model loglikelihood
## 1°      matern     -529.7873
## 2°    gaussian     -529.7890
## 3°   spherical     -530.2743
## 4° exponential     -531.3895
```

The two fitted models with the highest loglikelihood are the Matern model and the Gaussian model. So we'll use these two to perform the kriging prediction step.

Let's have a look again at the result of the two fits:

```
## likfit: estimated model parameters:
##     beta    tausq  sigmasq      phi
## "-2.786" "41.157" "66.712" " 5.601"
## Practical Range with cor=0.05 for asymptotic range: 9.693811
##
## likfit: maximised log-likelihood = -529.8

## likfit: estimated model parameters:
##      beta     tausq   sigmasq       phi     kappa
## "-2.7802" "40.9994" "66.7522" " 0.5757" "24.1330"
## Practical Range with cor=0.05 for asymptotic range: 9.889255
##
## likfit: maximised log-likelihood = -529.8
```

Both models have a nugget effect $\tau^2 \approx 41$ and a partial sill $\sigma^2 \approx 66.7$.

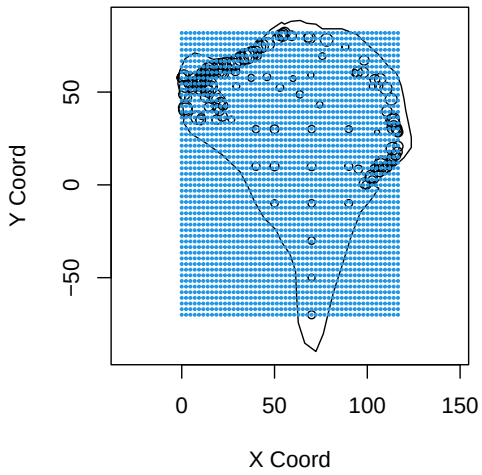
And they have almost the same Effective Range $\approx 9.7 - 9.9$, but the Gaussian model has a higher range than the Matern one: $\phi_G = 5.601$ vs $\phi_M = 0.5757$ thanks to the presence of a quite high value of the smoothness parameter: $k \approx 24.1$.

6) Predict the elevations along all the area of study using the two variogram selected in point 4. Discuss the type of kriging chosen:

a. Compare both kriging predictions using cross-validation, and propose the best model.

b. Show the predictions and their standard errors.

First generate a grid where we'll perform our kriging predictions:



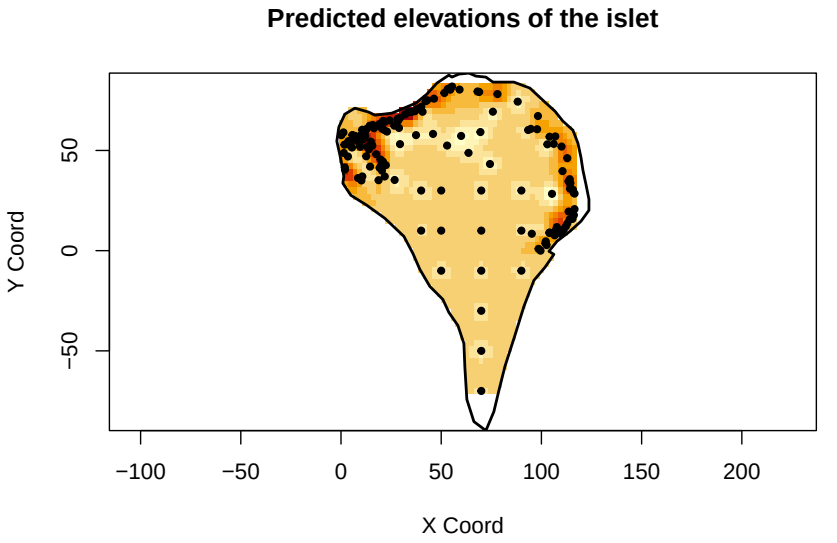
The residuals used in the variogram estimation were obtained after a quadratic model, so we'll perform a Universal Kriging on the original data using a 2^{nd} order trend.

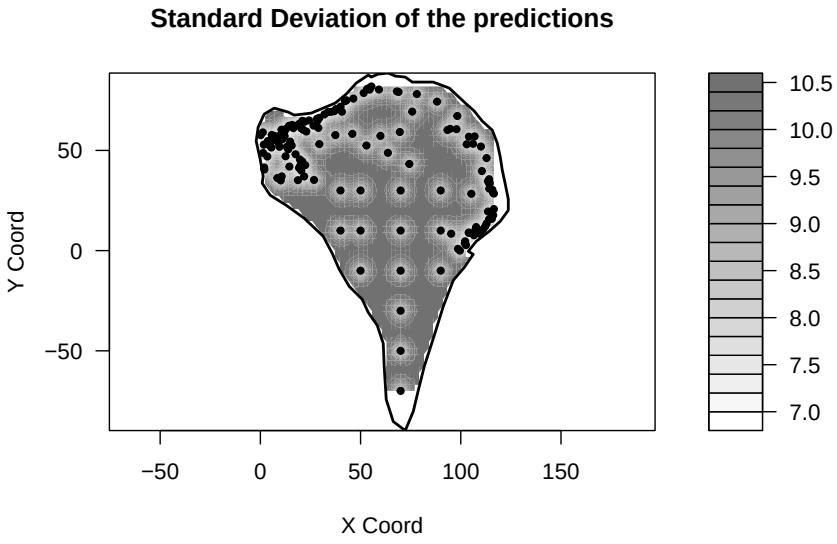
So we now compare the performance of the two kriging methods using a cross-validation approach: the 3 goodness indexes we obtain from the procedure are:

```
##   index   Gaussian     Matern
## 1   VC1 0.03902553 0.03877583
## 2   VC2 1.03294158 1.03249363
## 3   VC3 8.68269629 8.67909161
```

Therefore we decide to choose the Matern model as it has a lower VC3, a VC1 closer to 0 and a VC2 closer to 1.

Let's proceed to visualize the predictions of the Matern Kriging and their standard errors:





From the prediction plot we deduce that there's a mountain range on the north coast of the islet and a plain in the center and the south.

We also observe that, coherently with the theory, the variances of the prediction are lower near the observations.