# Homework 4 - Group 12

Francesco Natali, Leonardo Agate, Lorenzo Bartocci

2024-10-25

Load the dataset and the file containing the estimation grids

```
load("shrimpsfull.RData")
load("AllGrids.RData")
```

We now filter data for our years of interest, i.e. 2002 and 2008

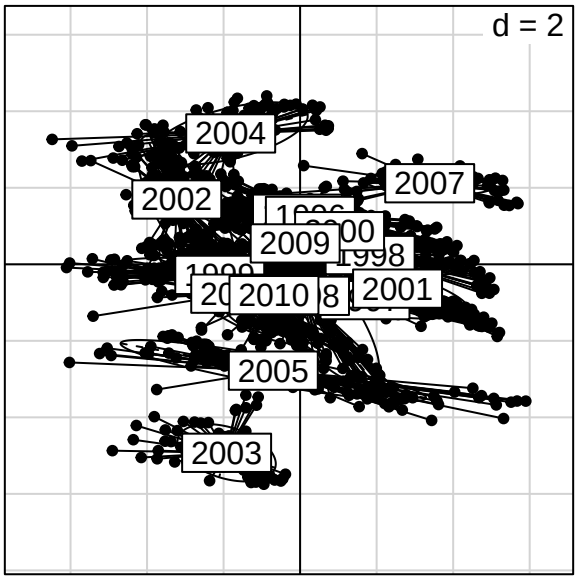
```
shrimp_data_2002 <- shrimpsdata[shrimpsdata$ANNO == 2002, ]
shrimp_data_2008 <- shrimpsdata[shrimpsdata$ANNO == 2008, ]
```

## Run a multivariate analysis

Let's first see what the year 2002 and 2008 were like in general

```
pca_result1 <- dudi.pca(df = shrimpsdata[, -c(3:5)], scannf = FALSE, nf = 3)
scatter(pca_result1)
```

```
shrimpsdata$ANNO <- factor(shrimpsdata$ANNO)
s.class(pca_result1$li, fac = shrimpsdata$ANNO)
```



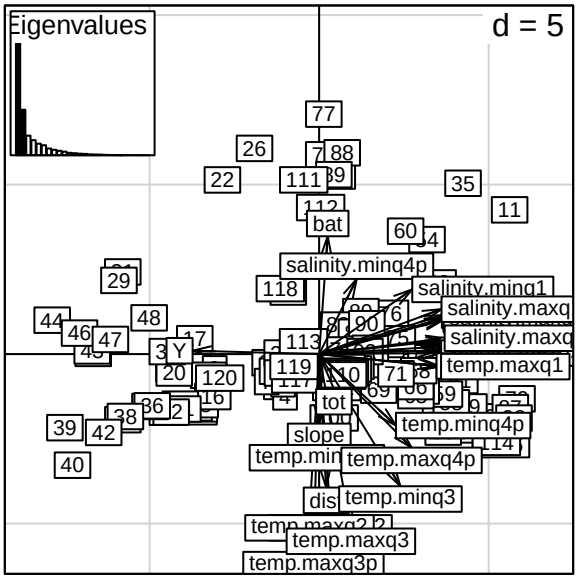
## 2002

We now start our analysis for the year 2002. At first, we explore data with a multivariate analysis, and perform PCA with at least two components (since we can see from the plot that the first two eigenvalue explain enough variance)

```
pca_02 <- dudi.pca(df = shrimp_data_2002[, -c(3:5)], scannf = FALSE, nf = 2)
```

Plot the PCA with the first two components, with the scatter() function plotting the first two axes

```
scatter(pca_02, xax = 1, yax = 2, clab.row = 0.7, clab.col = 0.7) #for better reading
```




We can now visualize the loadings of the covariates with the principal components

```
print(pca_02$c1)
```

```
##                         CS1          CS2
## X                0.2599973604  0.008107193
## Y               -0.2726727032  0.007247737
## salinity.minq3p  0.2679361665  0.077573121
## salinity.minq4p  0.0810293203  0.161934630
## salinity.minq1   0.2006441258  0.139417641
## salinity.minq2   0.2645247852  0.083932015
## salinity.minq3   0.2713874074  0.016641922
## salinity.maxq3p  0.2684184067  0.074014001
## salinity.maxq4p  0.2713338777  0.029874554
## salinity.maxq1   0.2637213654  0.096783936
## salinity.maxq2   0.2629046335  0.002242238
## salinity.maxq3   0.2704925441  0.034171859
## bat              0.0184509541  0.253880452
## dist             0.0153259204 -0.287605355
## slope            0.0001904954 -0.149774928
## temp.minq3p     -0.0048714908 -0.195809671
```

```
## temp.minq4p      0.1655394326 -0.153964807
## temp.minq1       0.2530604209 -0.025630456
## temp.minq2       0.0245327161 -0.346888513
## temp.minq3       0.1751849037 -0.284763959
## temp.maxq3p     -0.0122666819 -0.426162384
## temp.maxq4p      0.1991661493 -0.201524714
## temp.maxq1       0.2627312776 -0.026991684
## temp.maxq2      -0.0208165790 -0.346937020
## temp.maxq3       0.0684273869 -0.375863381
## tot              0.0318399052 -0.075153497
```

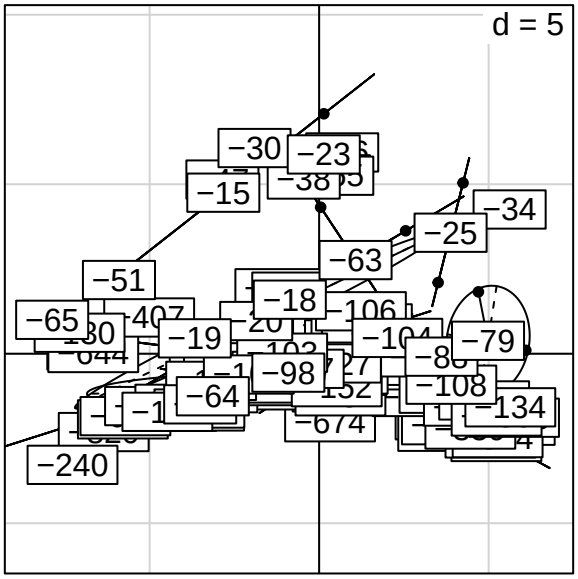
The first component (CS1) is strongly associated with salinity levels and geographical coordinates (specifically salinity.minq3p, salinity.maxq3, and salinity.minq3). This suggests a spatial gradient in salinity conditions. The second component (CS2) is linked to temperature extremes (high loadings on temp.maxq3p, temp.maxq3, and temp.minq2) and physical habitat features like bathymetry and distance from the coast. This indicates a temperature gradient and other physical factors influencing shrimp biomass. We now plot the covariates with the highest loadings related to the first and second component respectively

```
shrimp_data_2002$salinity.minq3 <- factor(shrimp_data_2002$salinity.minq3)
s.class(pca_02$li, fac = shrimp_data_2002$salinity.minq3)

shrimp_data_2002$temp.maxq3 <- factor(shrimp_data_2002$temp.maxq3)
s.class(pca_02$li, fac = shrimp_data_2002$temp.maxq3)
```

The s.class plot likely shows the grouping of observations based on biomass categories or classes. Each class represents a range of biomass values, and points in the plot are grouped according to similarities in bathymetry, salinity, and temperature. If the points form distinct clusters, this suggests that certain combinations of bathymetry, salinity, and temperature values correspond to specific biomass levels. These clusters support the idea that biomass has a structured spatial pattern influenced by these environmental covariates.

```
shrimp_data_2002$bat <- factor(shrimp_data_2002$bat)
s.class(pca_02$li, fac = shrimp_data_2002$bat)
```



The s.class plot for bathymetry supports the choice of including bathymetry in the variogram model for biomass estimation. It provides evidence of spatial depth clusters that might correspond to biomass clustering. ## Variogram choice Now, before obtaining and choosing a model for our empirical variogram, it is better to work on a log scale rather than on the original scale of the data set, because log transforming our data means to assume that it is distributed as a lognormal distribution, where mean and variance are not independent anymore: We then log-transform total biomass

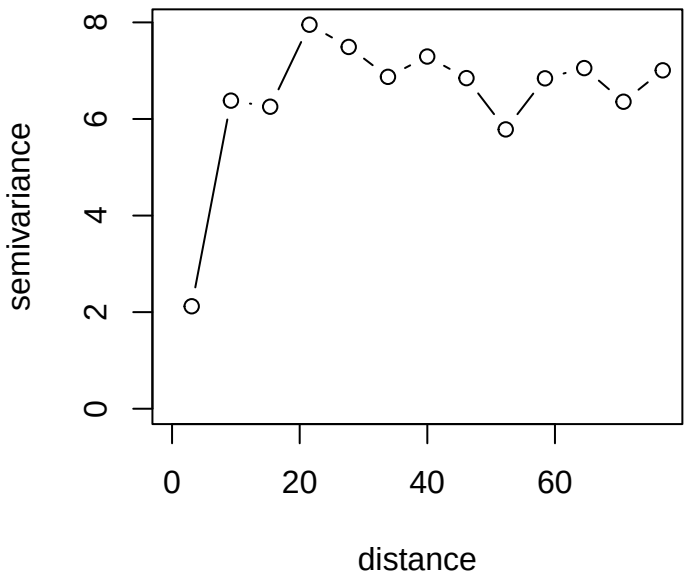
```
shrimp_data_2002$log_tot <- log(shrimp_data_2002$tot + 1)
```

After showing the values of the loadings in the PCA, we create geodata object with only depth (bat), salinity.minq3, temp.maxq3 and dist as covariates

```
shrimp_geodata_2002_log <- as.geodata(shrimp_data_2002,
                                      coords.col = c("X", "Y"),
                                      data.col = "log_tot",
                                      covar.col = c("bat","salinity.minq3","temp.maxq3","dist"))
```

Now, based on this log transformation, we can plot our empirical variogram. Before doing that, we have also to choose our trend in order to have a stationary and isotropic variogram, which is the basis for the kriging; first we plot the variogram with first order trend

```
variogram_2002 <- variog(shrimp_geodata_2002_log, trend = "1st", max.dist = 80)
plot(variogram_2002,type="b")
```



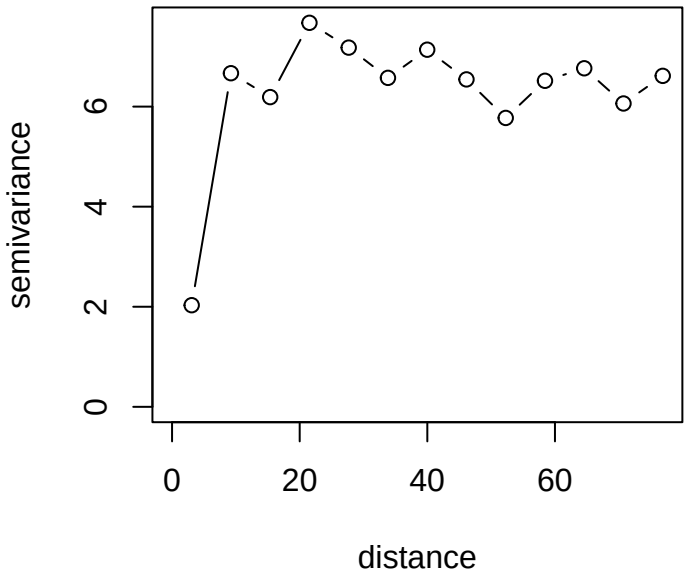

and we include our chosen covariates in the trend

```
trend_matrix_2002 <- cbind(shrimp_data_2002$bat,
                           shrimp_data_2002$salinity.minq3,
                           shrimp_data_2002$temp.maxq3,
                           shrimp_data_2002$dist)

variogram_tot_2002 <- variog(shrimp_geodata_2002_log, trend = "1st",
                             trend.d = trend_matrix_2002, max.dist = 80)
```

We now plot the variogram with second order trend

```
variogram_2_2002 <- variog(shrimp_geodata_2002_log, trend = "2nd", max.dist = 80)
plot(variogram_2_2002,type="b")
```




and we include our chosen covariates in the trend

```
trend_matrix_2002 <- cbind(shrimp_data_2002$bat,
                           shrimp_data_2002$salinity.minq3,
                           shrimp_data_2002$temp.maxq3,
                           shrimp_data_2002$dist)

variogram_2_2002 <- variog(shrimp_geodata_2002_log, trend = "2nd",
                           trend.d = trend_matrix_2002, max.dist = 80)
```

If we now compare the two variograms, we can notice how their overall behaviour is pretty similar and that they have equal nugget effect; however, the second order trend has a slightly lower sill (which is still reached at distance 25 ca.) which might be indicating an overparametrization of the model. Therefore, we leave the first order trend as our preferred choice. In order to see which model fits the variogram best we display the function

```
eyefit(variogram_2002)
```

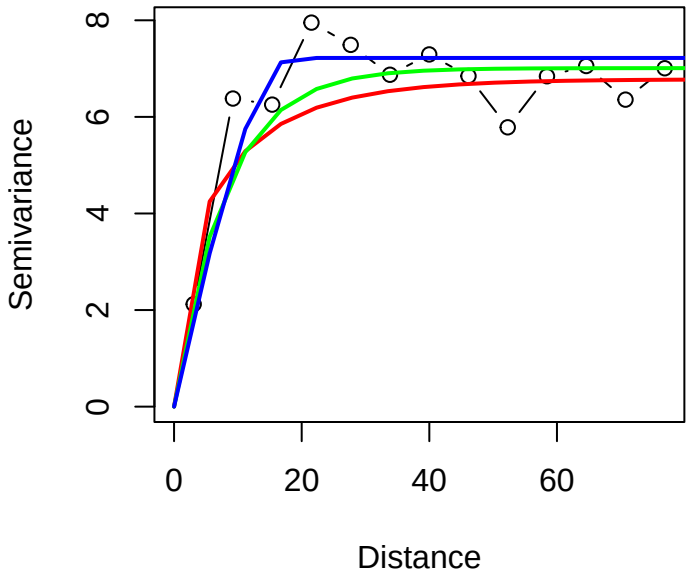
The three models that give a better fit overall are the matern (with parameter k = 0.2), the exponential (i.e. the particular case of a matern with k = 0.5) and the spherical model

```
fit_matern_lik_2002 <- likfit(shrimp_geodata_2002_log,
                              cov.model = "matern",
                              ini.cov.pars = c(5.16, 6.78),
                              nugget = 2, kappa = 0.2)
fit_exponential_lik_2002 <- likfit(shrimp_geodata_2002_log,
                             cov.model = "exponential",
                             ini.cov.pars = c(6.9, 4.6),
                             nugget = 2)
fit_spherical_lik_2002 <- likfit(shrimp_geodata_2002_log,
                                 cov.model = "spherical",
                                 ini.cov.pars = c(6.8, 12.3),
                                 nugget = 2)
```

we now plot our empirical variogram with all the three models for a better graphical comparison

```
plot(variogram_2002, type = "b", xlab = 'Distance', ylab = 'Semivariance')

lines(fit_matern_lik_2002, col = "red")
lines(fit_exponential_lik_2002, col = "green")
lines(fit_spherical_lik_2002, col = "blue")
```




At first sight, the matern looks like the better fit overall; we now compute the RMSE of each running first a cross validation

```
vv.mat.2002<-xvalid(shrimp_geodata_2002_log,model=fit_matern_lik_2002)
vv.exp.2002<-xvalid(shrimp_geodata_2002_log,model=fit_exponential_lik_2002)
vv.sph.2002<-xvalid(shrimp_geodata_2002_log,model=fit_spherical_lik_2002)
```

Calculate the Mean Squared Error (MSE) for each model

```
MSE_mat_2002 <- mean(vv.mat.2002$std.error^2)
MSE_exp_2002 <- mean(vv.exp.2002$std.error^2)
MSE_sph_2002 <- mean(vv.sph.2002$std.error^2)
```

Calculate the Root Mean Squared Error (RMSE)

```
RMSE_mat_2002 <- sqrt(MSE_mat_2002)
RMSE_exp_2002 <- sqrt(MSE_exp_2002)
RMSE_sph_2002 <- sqrt(MSE_sph_2002)
```

```
RMSE_mat_2002
```

```
## [1] 0.976253
```

```
RMSE_exp_2002
```

```
## [1] 0.9946067
```

```
RMSE_sph_2002
```

```
## [1] 1.018708
```

as expected, the matern model is the one with the lowest RMSE and will therefore be our chosen model (with first order trend). ## Kriging interpolation We can now run kriging interpolation for the estimation of total biomass. We include again the chosen covariates in the trend

```
trend.d.2002 <- trend.spatial(~ bat + salinity.minq3 + temp.maxq3 + dist,
                              geodata = shrimp_geodata_2002_log)
trend.l.2002 <- trend.spatial(~ grid_2002$bat + grid_2002$salinity.minq3 +
                                grid_2002$temp.maxq3 + grid_2002$dist)
```

Now we check how to implement things in the kriging function. We first control for the kriging: We have to build a krige.control list telling to the krig.conv all the elements that are required. Start with the matern model

```
krige.2002 <- krige.control(
  cov.model = "matern",
  cov.pars = c(5.16, 6.78),
  nugget = 2,
  trend.d = trend.d.2002,
  trend.l = trend.l.2002
)
```

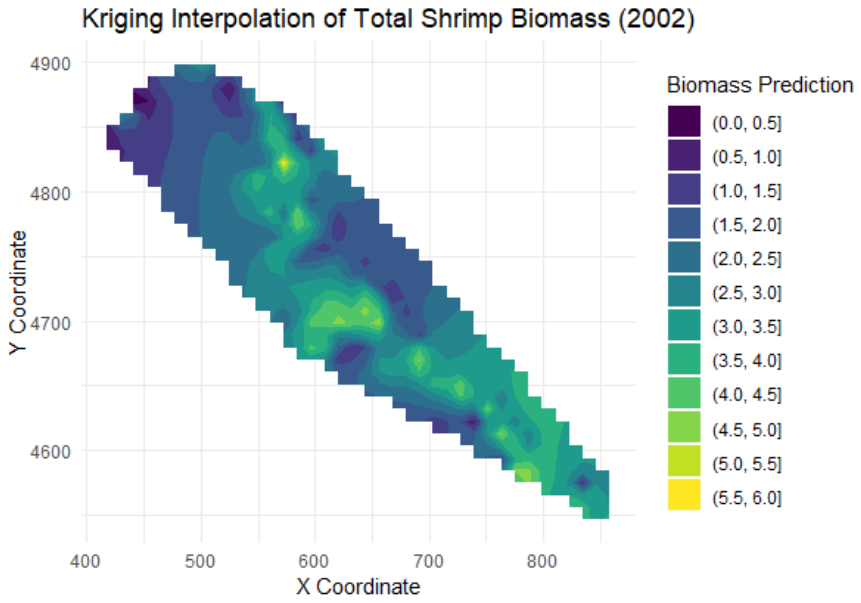
Run krige.conv with the updated locations

```
krig_2002 <- krige.conv(shrimp_geodata_2002_log,
                        locations = as.matrix(grid_2002[, c("X", "Y")]), krige = krige.2002)
```

Finally, we can visualize the kriging result for the matern

```
cc_mat_2002 <- data.frame(X = grid_2002$X, Y = grid_2002$Y, Z = krig_2002$predict)

ggplot(cc_mat_2002, aes(x = X, y = Y, z = Z)) +
  geom_contour_filled() +
  labs(
    title = "Kriging Interpolation of Total Shrimp Biomass (2002)",
    x = "X Coordinate",
    y = "Y Coordinate",
    fill = "Biomass Prediction") +
  theme_minimal()
```




In 2002, the total shrimp biomass appears to be more concentrated in the central parts of the study area (Lazio e Toscana coast). Brigest areas represent regions with the highest levels of biomass, while darkest areas indicate regions with lower levels. High biomass values (ranging from 4.0 to 6.0) are observed in distinct zones, possibly suggesting favorable environmental conditions that supported dense shrimp populations. The coastal regions, especially in the middle and upper parts of the study area, have a gradient where biomass decreases as you move offshore. It is possible to see that the areas of low biomass intensity are in particular along the Liguria coast, this could be due to a combination of unsuitable features, like the fact that the Ligurian coast is characterized by a steep continental shelf, leading to a rapid increase in depth offshore. Shrimp species typically prefer shallower and more gradually sloping habitats where nutrient-rich waters are more abundant. The steep bathymetric gradient in Liguria may reduce the extent of suitable habitats for shrimp. The opposite situation can be noted in the coast of Tuscany region tends to have more stable and favorable temperature and salinity profiles compared to northern areas like Liguria. These conditions are essential for shrimp breeding and larval development. Stable environmental conditions promote high recruitment and the maintenance of large shrimp populations

Now, in order to obtain a better and more precise geographical representation, we plot the map of Italy to see how the shrimp biomass is distributed along the areas of the Tirrenean Sea.

```
italy <- ne_countries(country = "Italy", scale = "medium", returnclass = "sf")
italy <- st_transform(italy, crs = 32632)
```

Convert kriging results to a grid

```
krig_result_df_02 <- data.frame(
  X = grid_2002$X * 1000,
  Y = grid_2002$Y * 1000,
  Z = krig_2002$predict
)
```

Define the bounds for the area of interest based on your prediction data

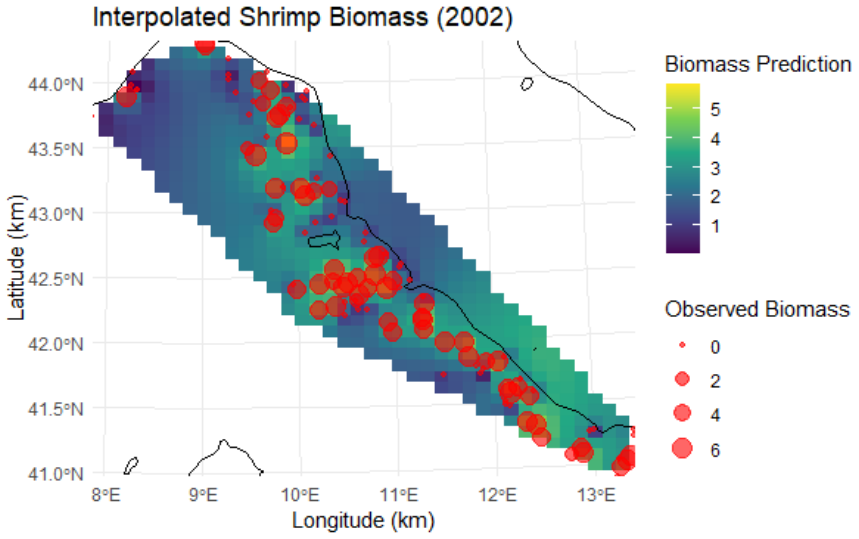
```
x_min <- min(krig_result_df_02$X) - 10000
x_max <- max(krig_result_df_02$X) + 10000
y_min <- min(krig_result_df_02$Y) - 10000
y_max <- max(krig_result_df_02$Y) + 10000
```

```
ggplot() +
  geom_raster( data =krig_result_df_02, aes(x = X, y = Y, fill = Z) ) +
  scale_fill_viridis_c( option ="viridis", name = "Biomass Prediction" ) +
  geom_sf( data = italy, fill = NA, color = "black",lwd = 0.7 ) +
  coord_sf( xlim = c(x_min, x_max), ylim = c(y_min, y_max), expand =FALSE ) +
  labs( title = "Interpolated Shrimp Biomass (2002)", x = "Longitude (km)", y ="Latitude (km)" ) +
  theme_minimal()
```

Add the observed biomass

```
observed_points_df02 <- data.frame(
  X = shrimp_geodata_2002_log$coords[, 1] * 1000,
  Y = shrimp_geodata_2002_log$coords[, 2] * 1000,
  Biomass = shrimp_geodata_2002_log$data
)

# Plot with Italy map, kriging results, and observed points
ggplot() +
  geom_raster(data = krig_result_df_02, aes(x = X, y = Y, fill = Z)) +
  scale_fill_viridis_c(option = "viridis", name = "Biomass Prediction") +
  geom_sf(data = italy, fill = NA, color = "black", lwd = 0.7) +
  geom_point(data = observed_points_df02, aes(x = X, y = Y, size = Biomass),
             color = "red", alpha = 0.6) +
  scale_size_continuous(name = "Observed Biomass", range = c(1, 5)) +
  coord_sf(xlim = c(x_min, x_max), ylim = c(y_min, y_max), expand = FALSE) +
  labs(title = "Interpolated Shrimp Biomass (2002)", x = "Longitude (km)", y = "Latitude (km)") +
  theme_minimal()
```




The figure provides a comprehensive visualization of the spatial distribution of shrimp biomass for the year 2002, derived through kriging interpolation. The distribution of the red points indicates that higher observed biomass values generally coincide with areas of high predicted biomass. This overlap suggests that the kriging model has reasonably captured the underlying spatial variability in shrimp biomass. However, there are some discrepancies where observed biomass values do not align perfectly with the predictions. These discrepancies may be due to limitations in the available data or model parameters

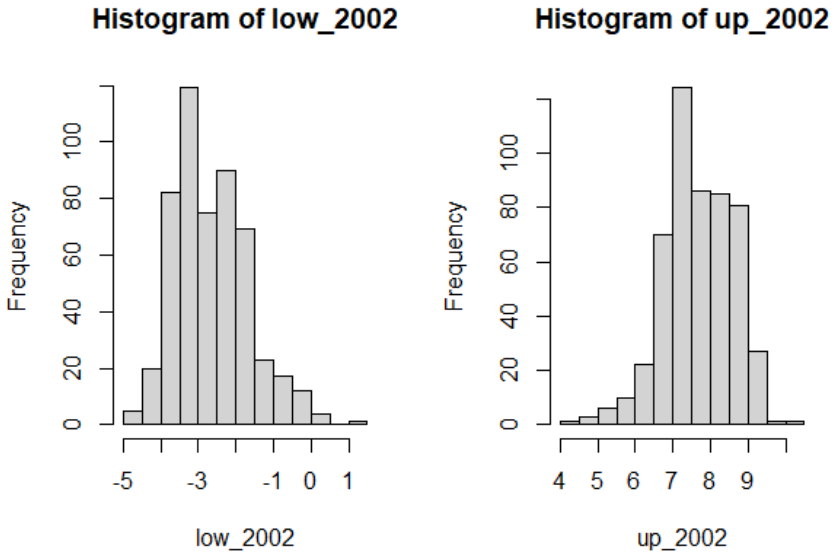
## Confidence Interval

Given the results obtained in the previous steps, we can also build confidence intervals at the 95% level

```
low_2002 <- krig_2002$predict- 1.96*sqrt(krig_2002$krige.var)
up_2002 <- krig_2002$predict + 1.96*sqrt(krig_2002$krige.var)
```

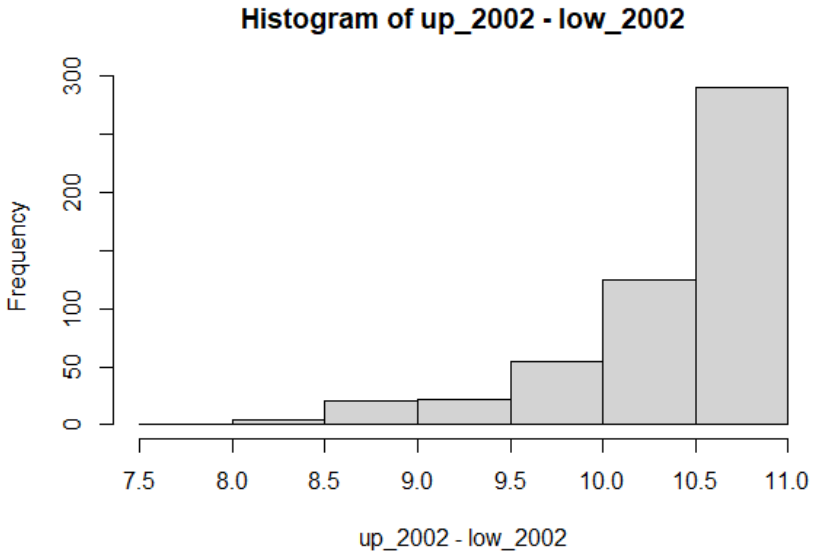
We can at first build one map for the lower bound and one map for the other, and then we can compare both

```
hist(low_2002)
hist(up_2002)
```




The distribution of the lower bound (low) is skewed towards the lower end, with most values ranging from around -5 to 0. This indicates that the lower bound of the biomass predictions is generally on the lower side, reflecting uncertainty and variability in areas with potentially low biomass. Then the upper confidence interval (up) values are distributed in a more compact range, from approximately 4 to 9, with a peak around 7 to 8. This shows that the upper limits of biomass predictions are more consistent and tend to reflect a higher level of potential shrimp biomass.

```
hist(up_2002-low_2002)
```

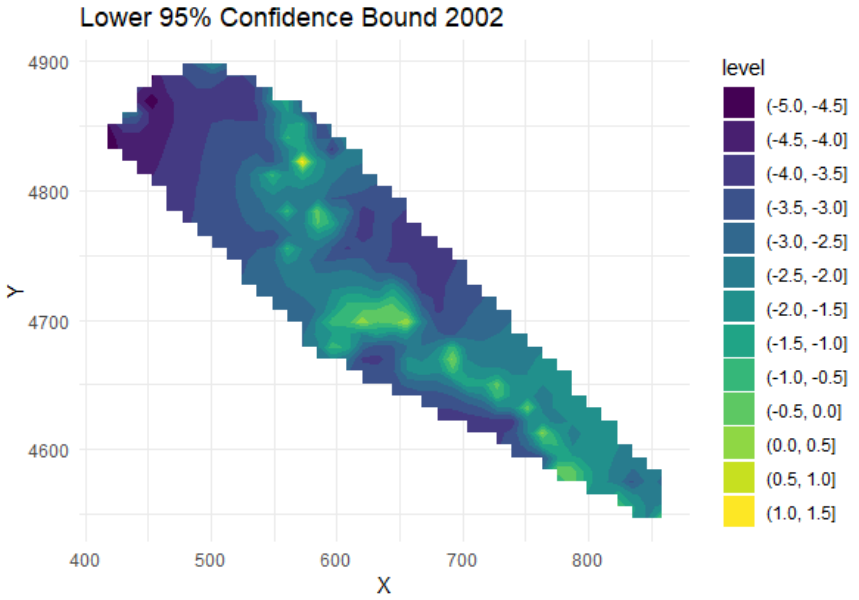



The histogram of the difference (up - low) between the upper and lower bounds are quite large, with a clear right-skewed distribution peaking between 10 and 11. This suggests that the width of the confidence intervals is substantial. Now is necessary to create data frames for plotting

```
cc_lower_2002 <- data.frame(X = grid_2002$X, Y = grid_2002$Y, Z = low_2002)
cc_upper_2002 <- data.frame(X = grid_2002$X, Y = grid_2002$Y, Z = up_2002)
```

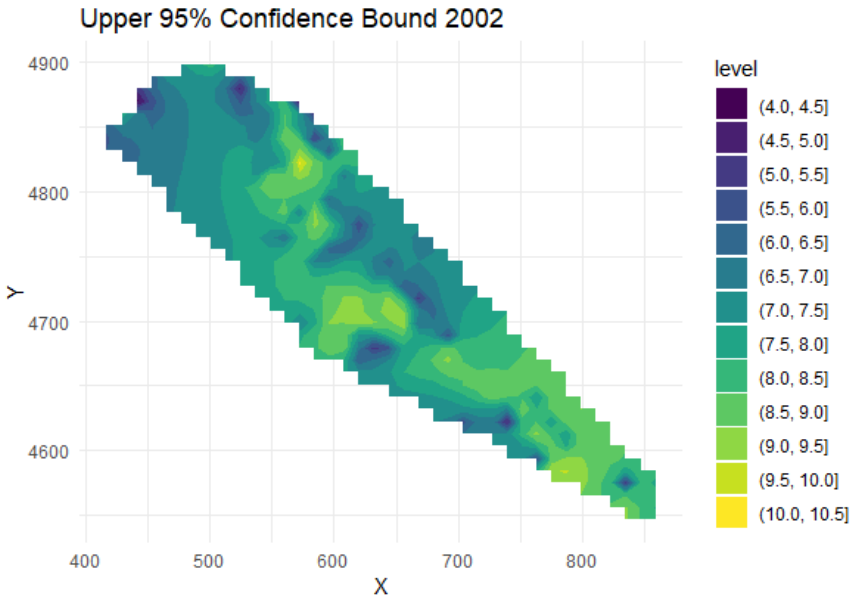
Plot lower bound

```
ggplot(cc_lower_2002, aes(x = X, y = Y, z = Z)) +
  geom_contour_filled() +
  labs(title = "Lower 95% Confidence Bound 2002") +
  theme_minimal()
```



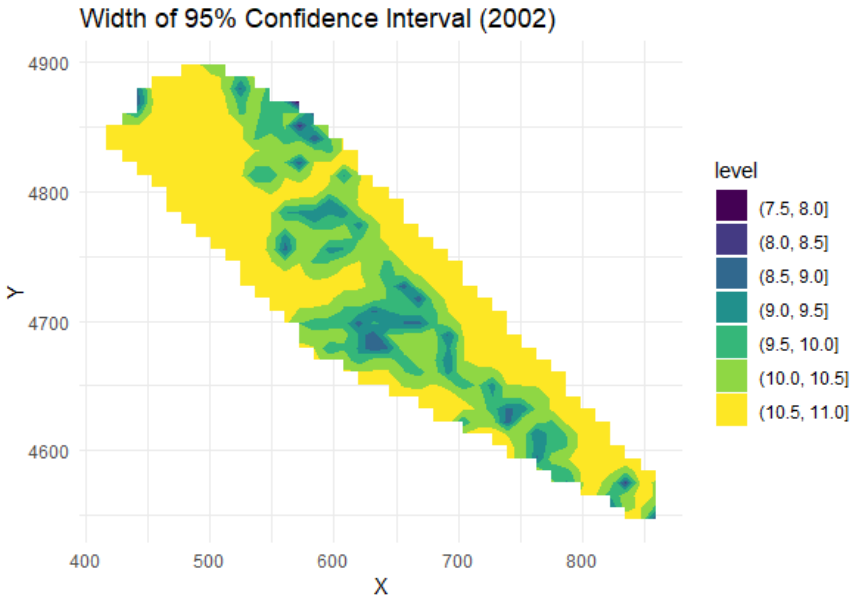

This map illustrates the spatial variation in the lower 95% confidence bound for total biomass in 2002. The color gradient transitions from dark purple (representing lower values, between -5 and -1) to bright yellow (indicating higher values, between 1 and 1.5). Regions shaded in green to yellow mark areas with relatively higher conservative biomass estimates, suggesting these zones may serve as biomass hotspots. These favorable areas likely correspond to optimal environmental conditions such as suitable bathymetry and salinity levels. In contrast, darker purple to blue regions denote locations with consistently lower biomass estimates, implying less favorable conditions, potentially due to factors like deeper water, suboptimal salinity, or nutrient scarcity. Overall, this spatial representation along the Gaeta-Genova coast highlights the variability in biomass distribution, emphasizing regions with robust conservative estimates and identifying areas where environmental conditions might limit biomass growth

```
ggplot(cc_upper_2002, aes(x = X, y = Y, z = Z)) +
  geom_contour_filled() +
  labs(title = "Upper 95% Confidence Bound 2002") +
  theme_minimal()
```




The Upper 95% Confidence Bound map presents an optimistic projection of potential shrimp density (log scale) across spatial locations. The lighter shades, ranging from yellow to light green, denote regions with higher upper bounds (7–10 log scale), indicating areas with the potential for significant total biomass under ideal environmental conditions. These zones are likely associated with optimal habitats, including shallower coastal shelves or areas near estuaries where depth and salinity are favorable. In contrast, darker shades, from purple to blue, represent areas with lower upper bounds (4–6 log scale), suggesting limited potential for high biomass even under favorable conditions. Such areas may correspond to deeper waters, less suitable salinity gradients, or generally suboptimal habitats for shrimp. This map offers valuable insights into spatial variations in biomass potential along the Gaeta-Genova coastline, emphasizing areas with higher growth opportunities and identifying regions with limited biomass expectations

```
interval_width_2002 <- up_2002- low_2002
cc_interval <- data.frame(X = grid_2002$X, Y = grid_2002$Y, Z = interval_width_2002)
ggplot(cc_interval, aes(x = X, y = Y, z = Z)) +
  geom_contour_filled() +
  labs(title = "Width of 95% Confidence Interval (2002)") +
  theme_minimal()
```




This map displays the spatial variation in biomass uncertainty for shrimp populations along the coastal stretch from Gaeta to Genoa in the northwestern Mediterranean during 2002. The width of the 95% confidence interval reflects the variability or uncertainty associated with the kriging biomass estimates. Regions shaded in yellow, where the confidence interval widths range from 10 to 11, represent areas of highest uncertainty. This considerable variability suggests that shrimp biomass estimates in these zones are less reliable, possibly due to fluctuating or poorly understood environmental conditions. Hypothetically, these regions may experience highly dynamic influences, such as variable currents, inconsistent nutrient supply, or irregular salinity patterns. Conversely, areas depicted in shades of green to blue, with confidence interval widths between 6 and 9, exhibit moderate uncertainty. This level of variability implies that these regions may have relatively stable environmental factors, such as consistent depth profiles, predictable salinity gradients, or moderate nutrient availability, contributing to more dependable biomass predictions. The purple areas, characterized by the narrowest confidence interval widths (8-8.5), denote zones of high certainty in biomass estimates. These regions are likely governed by stable and predictable environmental conditions, such as well-defined bathymetric features, steady nutrient influxes, or established ecological factors that have been extensively documented. The overall spatial pattern emphasizes areas where biomass estimates are more reliable versus those with greater uncertainty, underscoring the influence of environmental complexity and the importance of further understanding key ecological drivers affecting shrimp distributions.