Homework 4 - Group 12

Francesco Natali, Leonardo Agate, Lorenzo Bartocci

2024-10-25

Load the dataset and the file containing the estimation

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

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

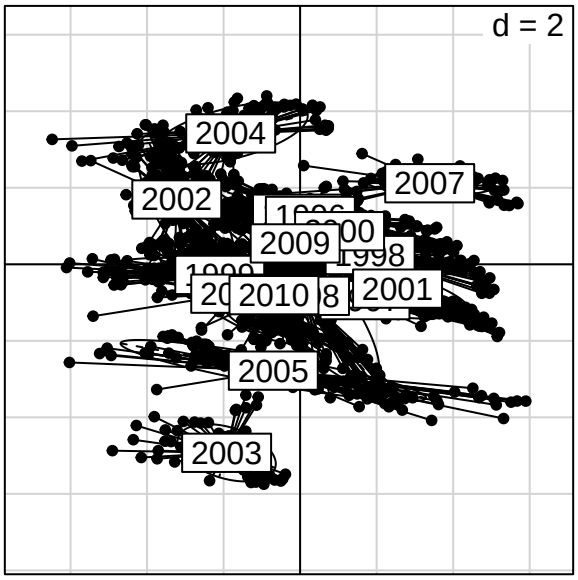
```
shrimp_data_2002 <- shrimpsdata[shrimpsdata$ANNO == 2002, ]
shrimp_data_2008 <- shrimpsdata[shrimpsdata$ANNO == 2008, ]
```

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)
```



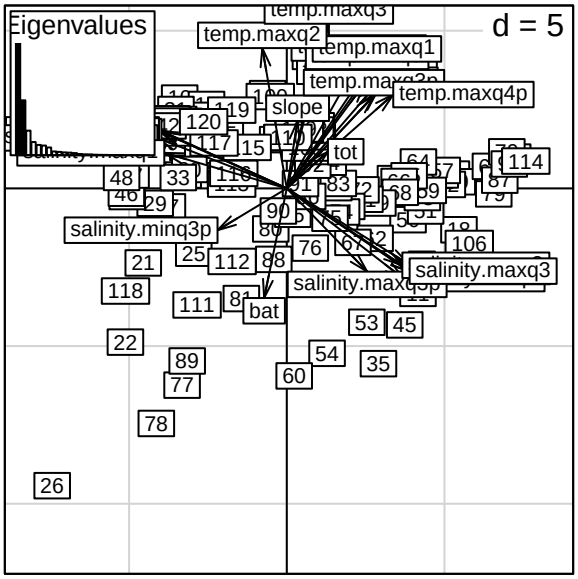
2008

We now start our analysis for the year 2008. 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_08 <- dudi.pca(df = shrimp_data_2008[, -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_08, xax = 1, yax = 2, clab.row = 0.7, clab.col = 0.7) #for better reading
```



Covariates like salinity.maxq3, temp.maxq2, and temp.maxq3 have strong loadings along the directions close to "tot" on the first two components. We can now visualize the loadings of the covariates with the principal components

```
print(pca_08$c1)
```

```
##                       CS1         CS2
## X                0.25480046 -0.15842934
## Y               -0.26164575  0.11859392
## salinity.minq3p -0.14459168 -0.08727478
## salinity.minq4p -0.26689416  0.12215922
## salinity.minq1  -0.26693351  0.12000550
## salinity.minq2   0.24359189 -0.18830384
## salinity.minq3   0.25562391 -0.17345555
## salinity.maxq3p  0.16776905 -0.17371043
## salinity.maxq4p -0.27013243  0.10534018
## salinity.maxq1  -0.25576446  0.07412357
## salinity.maxq2   0.24118128 -0.16049099
## salinity.maxq3   0.25647049 -0.17272536
## bat             -0.04729551 -0.22883416
## dist             0.05926173  0.24884029
```

```
## slope            0.02253986  0.14072970
## temp.minq3p      0.16167172  0.21791757
## temp.minq4p      0.15019961  0.26637631
## temp.minq1       0.18532884  0.21454993
## temp.minq2       0.18715206  0.19920271
## temp.minq3       0.16772099  0.24903398
## temp.maxq3p      0.18197359  0.19399802
## temp.maxq4p      0.22222741  0.19595660
## temp.maxq1       0.18820698  0.26448716
## temp.maxq2      -0.05112144  0.29286055
## temp.maxq3       0.08777050  0.34366654
## tot              0.08584905  0.07657020
```

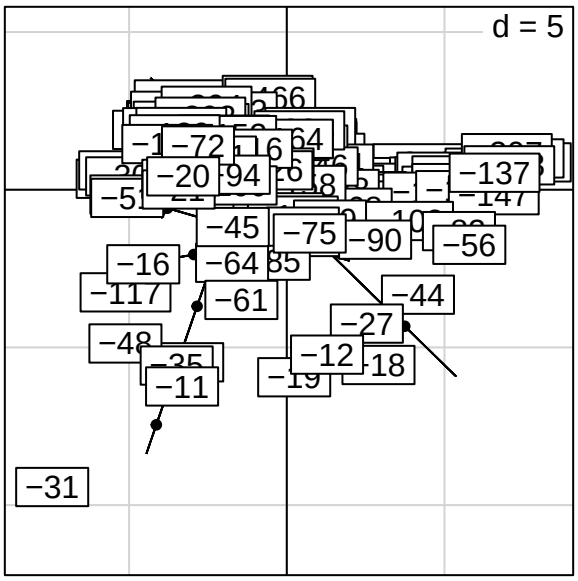
The second component (CS2) shaw that temp.maxq3 and temp.maxq2 have high positive loadings on CS2, suggesting that these temperature-related covariates are significant in explaining variance in the data. But also salinity.minq4p and salinity.maxq4p have high loadings on both CS1 and CS2, indicating their potential influence on biomass. We now plot the covariates with the highest loadings related to the first and second component respectively

```
shrimp_data_2008$salinity.maxq3 <- factor(shrimp_data_2002$salinity.maxq3)
s.class(pca_08$li, fac = shrimp_data_2002$salinity.maxq3)

shrimp_data_2008$temp.maxq3 <- factor(shrimp_data_2008$temp.maxq3)
s.class(pca_08$li, fac = shrimp_data_2008$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_2008$bat <- factor(shrimp_data_2008$bat)
s.class(pca_08$li, fac = shrimp_data_2008$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

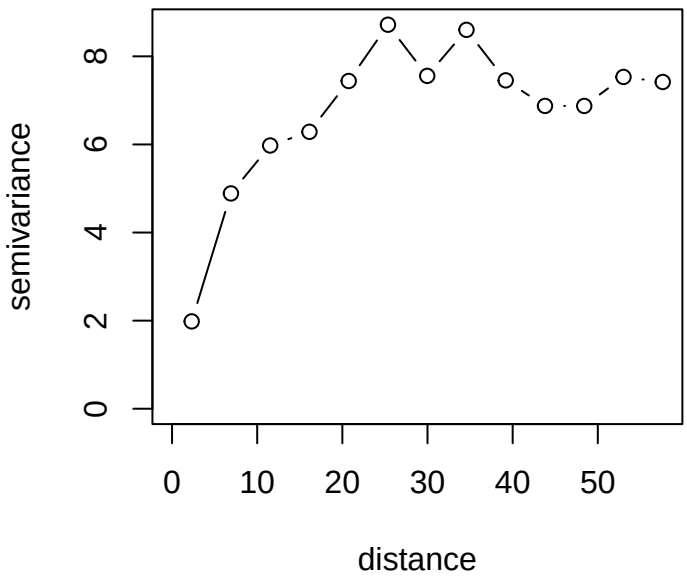
```
shrimp_data_2008$log_tot <- log(shrimp_data_2008$tot + 1)
```

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

```
shrimp_geodata_2008_log <- as.geodata(shrimp_data_2008,
                                      coords.col = c("X", "Y"),
                                      data.col = "log_tot",
                                      covar.col = c("bat","salinity.maxq3","temp.maxq3","slope"))
```

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_2008 <- variog(shrimp_geodata_2008_log, trend = "1st", max.dist = 60)
plot(variogram_2008,type="b")
```



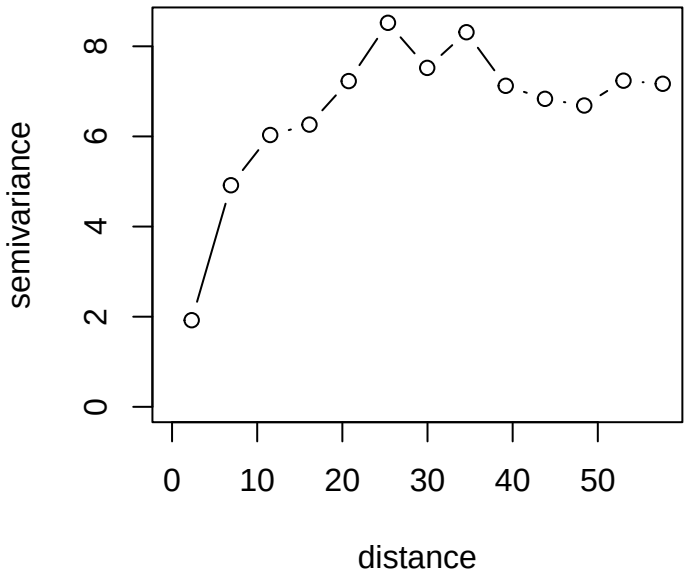
and we include our chosen covariates in the trend

```
trend_matrix_2008 <- cbind(shrimp_data_2008$bat,
                           shrimp_data_2008$salinity.minq3,
                           shrimp_data_2008$temp.maxq3,
                           shrimp_data_2008$slope)

variogram_2008 <- variog(shrimp_geodata_2008_log, trend = "1st",
                             trend.d = trend_matrix_2008, max.dist = 60)
```

We now plot the variogram with second order trend

```
variogram_2_2008 <- variog(shrimp_geodata_2008_log, trend = "2nd", max.dist = 60)
plot(variogram_2_2008,type="b")
```



and we include our chosen covariates in the trend

```
trend_matrix_2008 <- cbind(shrimp_data_2008$bat,
                           shrimp_data_2008$salinity.minq3,
                           shrimp_data_2008$temp.maxq3,
                           shrimp_data_2008$slope)

variogram_tot2_2008 <- variog(shrimp_geodata_2008_log, trend = "2nd",
                             trend.d = trend_matrix_2008, max.dist = 60)
```

If we now compare the two variograms, we can notice how their pverall 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_2008)
```

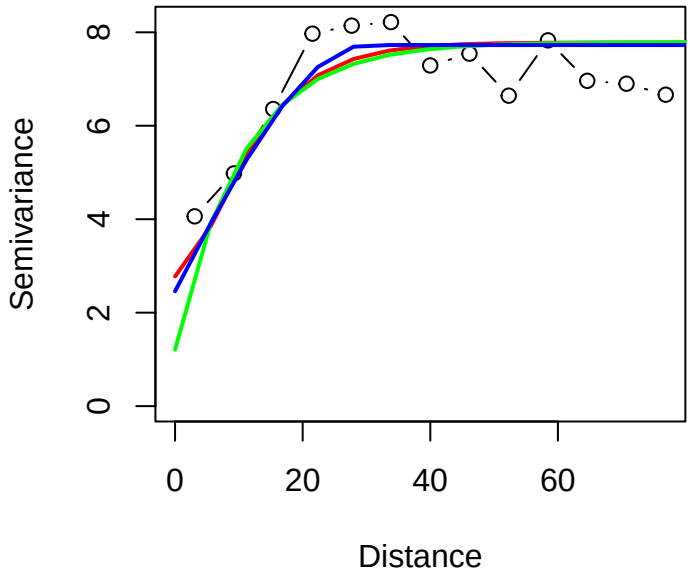
The three models that give a better fit overall are the matern (with parameter k = 1.5), the exponential (i.e. the particular case of a matern with k = 0.5) and the spherical model

```
fit_matern_lik_2008 <- likfit(shrimp_geodata_2008_log,
                              cov.model = "matern",
                              ini.cov.pars = c(4.9, 4.4),
                              nugget = 3.78, kappa = 1.5)
fit_exponential_lik_2008 <- likfit(shrimp_geodata_2008_log,
                                   cov.model = "exponential",
                                   ini.cov.pars = c(6.9, 6.3),
                                   nugget = 3.78)
fit_spherical_lik_2008 <- likfit(shrimp_geodata_2008_log,
                                 cov.model = "spherical",
                                 ini.cov.pars = c(6.4, 8.7),
                                 nugget = 3.78)
```

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

```
plot(variogram_2008, type = "b", xlab = 'Distance', ylab = 'Semivariance')

lines(fit_matern_lik_2008, col = "red", lwd = 2)
lines(fit_exponential_lik_2008, col = "green", lwd = 2)
lines(fit_spherical_lik_2008, col = "blue", lwd = 2)
```



At first sight, all three models don't look too different from each other and seem to be a good approximation of the variogram; hence, in order to make our final choice, we now compute the RMSE of each running first a cross validation

```
vv.mat.2008<-xvalid(shrimp_geodata_2008_log,model=fit_matern_lik_2008)
vv.exp.2008<-xvalid(shrimp_geodata_2008_log,model=fit_exponential_lik_2008)
vv.sph.2008<-xvalid(shrimp_geodata_2008_log,model=fit_spherical_lik_2008)
```

Calculate the Mean Squared Error (MSE) for each model

```
MSE_mat_2008 <- mean(vv.mat.2008$std.error^2)
MSE_exp_2008 <- mean(vv.exp.2008$std.error^2)
MSE_sph_2008 <- mean(vv.sph.2008$std.error^2)
```

Calculate the Root Mean Squared Error (RMSE)

```
RMSE_mat_2008 <- sqrt(MSE_mat_2008)
RMSE_exp_2008 <- sqrt(MSE_exp_2008)
RMSE_sph_2008 <- sqrt(MSE_sph_2008)
```

```
RMSE_mat_2008
```

```
## [1] 1.007362
```

```
RMSE_exp_2008
```

```
## [1] 1.005225
```

```
RMSE_sph_2008
```

```
## [1] 1.008564
```

As we can see, all MSE values are quite good. The exponential 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.2008 <- trend.spatial(~ bat + salinity.minq3 + temp.maxq3 + slope,
                              geodata = shrimp_geodata_2008_log)
trend.l.2008 <- trend.spatial(~ grid_2008$bat + grid_2008$salinity.minq3 +
                                grid_2008$temp.maxq3 + grid_2008$slope)
```

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 exponential model

```
krige.2008 <- krige.control(
  cov.model = "exponential",
  cov.pars = c(6.9, 6.3),
  nugget = 3.78,
  trend.d = trend.d.2008,
  trend.l = trend.l.2008
)
```

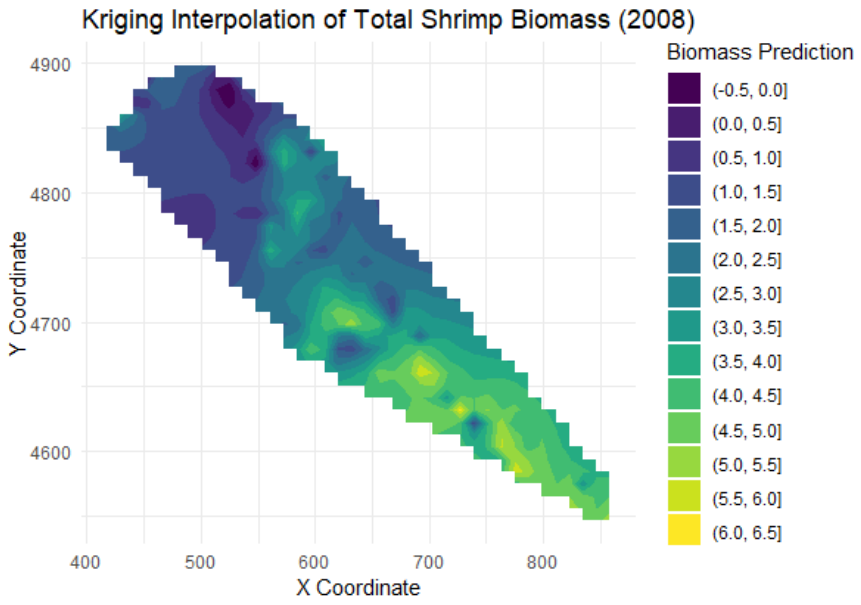
Run krige.conv with the updated locations

```
krig_2008 <- krige.conv(shrimp_geodata_2008_log,
                        locations = as.matrix(grid_2008[, c("X", "Y")]), krige = krige.2008)
```

Finally, we can visualize the kriging result for the exponential

```
cc_exp_2008 <- data.frame(X = grid_2008$X, Y = grid_2008$Y, Z = krig_2008$predict)

ggplot(cc_exp_2008, aes(x = X, y = Y, z = Z)) +
  geom_contour_filled() +
  labs(
    title = "Kriging Interpolation of Total Shrimp Biomass (2008)",
    x = "X Coordinate",
    y = "Y Coordinate",
    fill = "Biomass Prediction") +
  theme_minimal()
```



In 2008, the spatial distribution of biomass shows a broader spread of high biomass values. There is an apparent shift or expansion of areas with values between 5.0 and 6.5, especially towards the southern region. The gradient from high to low biomass is more diffused compared to 2002, suggesting changes in shrimp distribution patterns. This could indicate a response to changing environmental conditions. Shrimp may have moved to areas where environmental conditions remained within their optimal range. There is also a high concentration of biomass near the Ligurian coast with respect to the 2002 interpolation. The southern coast region benefits from relatively stable oceanographic conditions respect to the northen coast, including lower exposure to strong currents or deep-water upwellings. This stability provides a more suitable environment for shrimp, enabling them to concentrate in these areas without being dispersed by unfavorable water movements. 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_08 <- data.frame(
  X = grid_2008$X * 1000,
  Y = grid_2008$Y * 1000,
  Z = krig_2008$predict
)
```

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

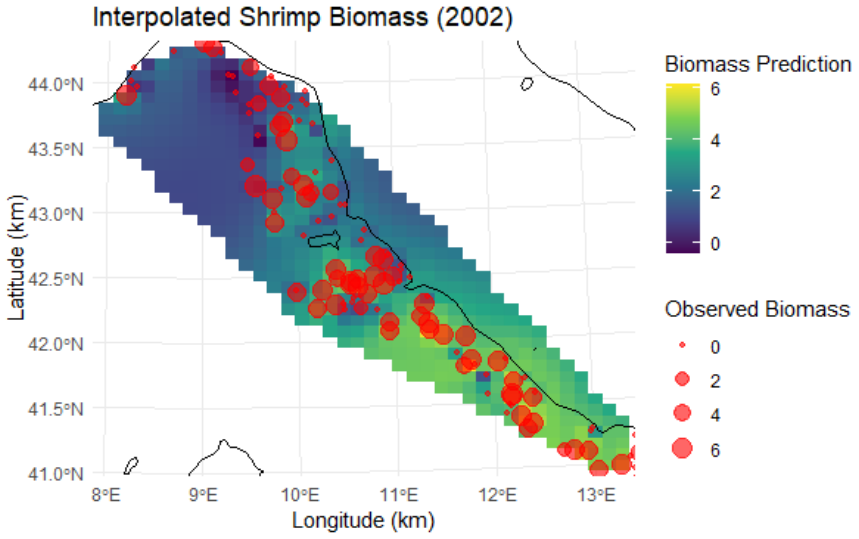
```
x_min <- min(krig_result_df_08$X) - 10000
x_max <- max(krig_result_df_08$X) + 10000
y_min <- min(krig_result_df_08$Y) - 10000
y_max <- max(krig_result_df_08$Y) + 10000
```

```
ggplot() + # Plot the kriging predictions as filled contours
  geom_raster( data =krig_result_df_08, 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_df08 <- data.frame(
  X = shrimp_geodata_2008_log$coords[, 1] * 1000,
  Y = shrimp_geodata_2008_log$coords[, 2] * 1000,
  Biomass = shrimp_geodata_2002_log$data
)

# Plot with Italy map, kriging results, and observed points
ggplot() +
  # Kriging predictions as a raster layer
  geom_raster(data = krig_result_df_08, 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_df08, 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()
```

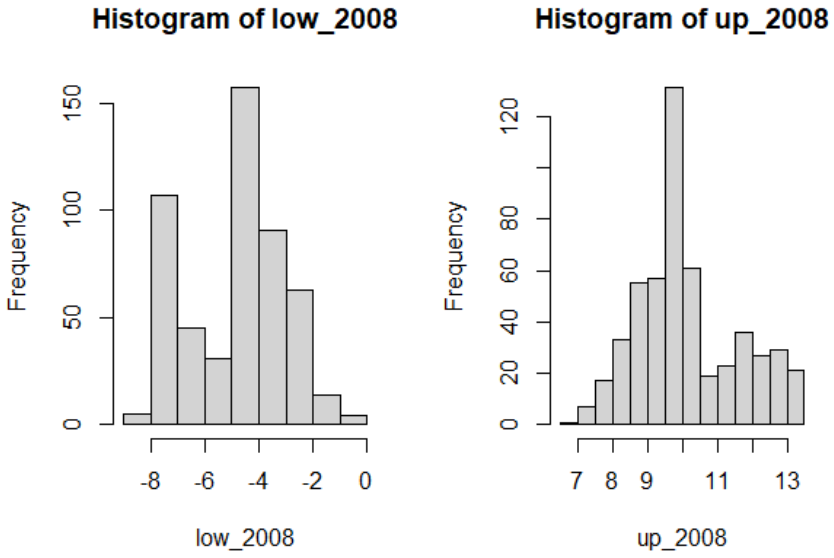


The figure provides a comprehensive visualization of the spatial distribution of shrimp biomass for the year 2008, derived through kriging interpolation. There seems to be a good alignment between the observed biomass values (red points) and the predicted values (background gradient). The kriging framework has successfully interpolated shrimp biomass across the region, capturing spatial trends influenced by environmental factors. The visualization suggests that the model is reasonably predictive, but further validation (e.g., cross-validation) is necessary to confirm its accuracy. ## Confidence Interval Given the results obtained in the previous steps, we can also build confidence intervals at the 95% level

```
low_2008 <- krig_2008$predict- 1.96*sqrt(krig_2008$krige.var)
up_2008 <- krig_2008$predict + 1.96*sqrt(krig_2008$krige.var)
```

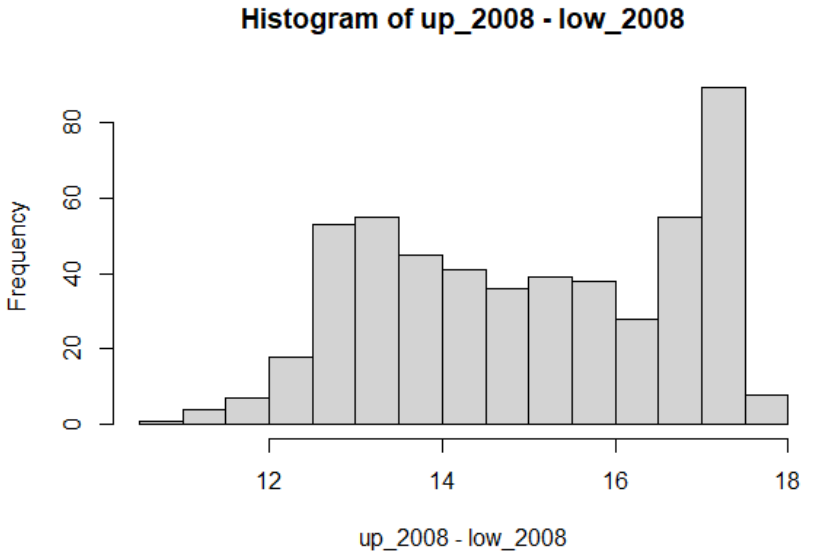
We can at first build one map for the lower bound and one map for the other, and then we can compare both

```
hist(low_2008)
hist(up_2008)
```



The distribution of the lower bound (low) is skewed towards the lower end, with most values concentred between -8 and-4. 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 more symmetric and centered around values between 9 and 11, with fewer extreme values compared to the lower bound. 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_2008-low_2008)
```



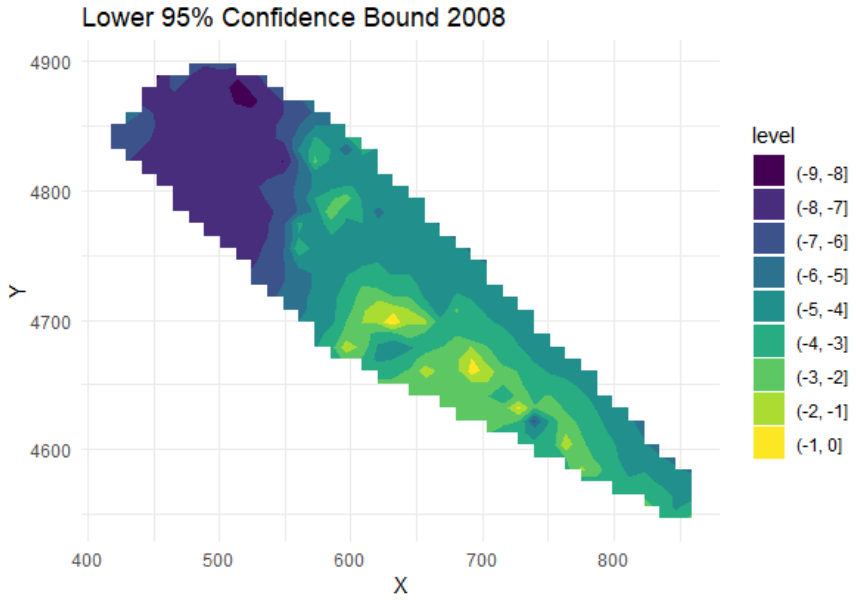
The histogram of the difference (up - low) between the upper and lower bounds is unimodal and primarily concentrated between 12 and 18, indicating a consistent width of the confidence intervals across most of the grid. The difference histogram suggests that uncertainty in biomass predictions is fairly uniform across the

study area, which can be a positive sign if the model is well-calibrated. Now is necessary to create data frames for plotting

```
cc_lower_2008 <- data.frame(X = grid_2008$X, Y = grid_2008$Y, Z = low_2008)
cc_upper_2008 <- data.frame(X = grid_2008$X, Y = grid_2008$Y, Z = up_2008)
```

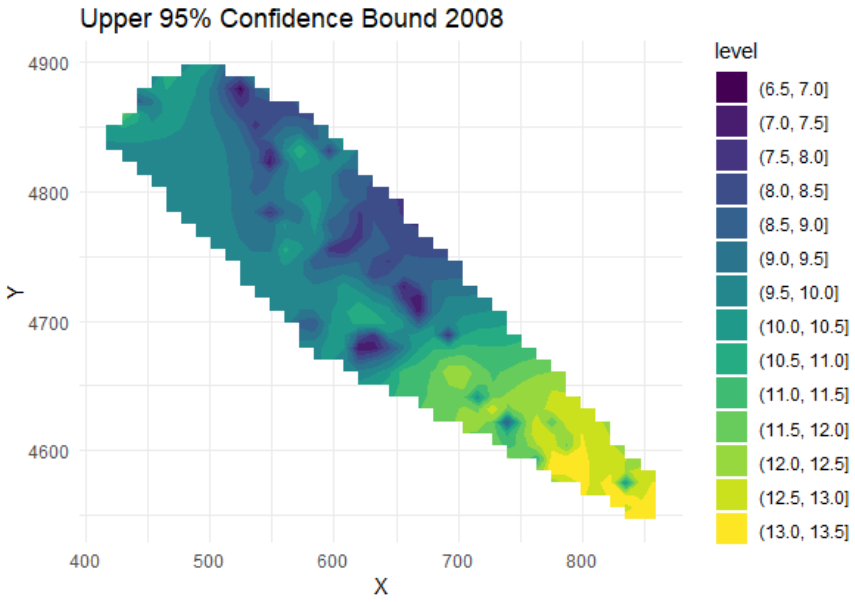
Plot lower bound

```
ggplot(cc_lower_2008, aes(x = X, y = Y, z = Z)) +
  geom_contour_filled() +
  labs(title = "Lower 95% Confidence Bound 2008") +
  theme_minimal()
```



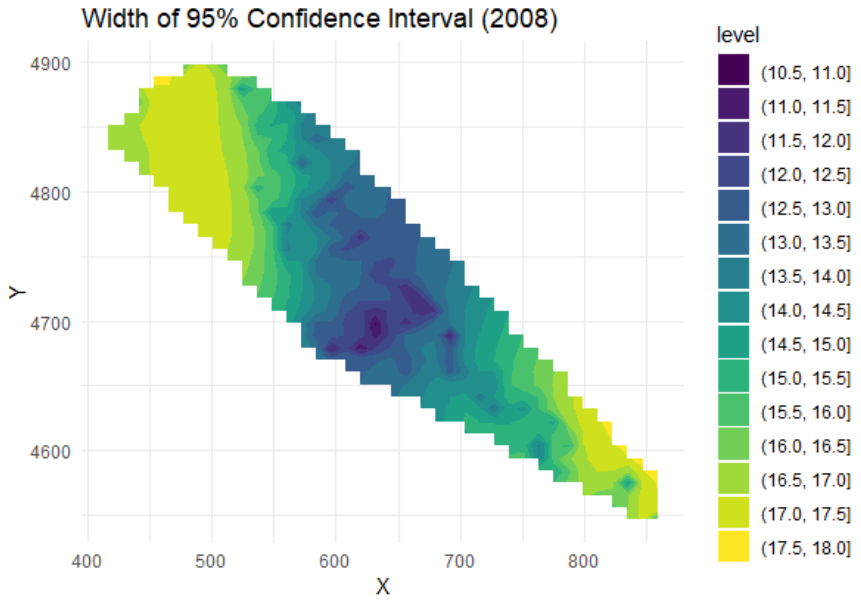
This map depicts the spatial variation in the lower 95% confidence bound for total biomass along the Gaeta-Genova coast in 2008. The lighter regions (green to yellow), with values ranging from approximately -1 to 0 (log scale), represent areas with higher biomass even under conservative estimates. These zones are likely associated with more productive or nutrient-rich habitats, such as shallow coastal zones or estuarine areas, where conditions like moderate salinity and sunlight penetration promote biomass growth. In contrast, darker regions (purple to blue), with values from approximately -9 to -4, indicate areas with lower conservative biomass estimates. These regions may correspond to deeper waters or zones with less favorable salinity and nutrient conditions, which can limit the accumulation of biomass. Depth is a key factor influencing biomass distribution, as shallow coastal areas typically experience greater sunlight penetration, enabling photosynthesis and supporting a richer food web for marine life. The lighter zones on this map likely indicate biologically productive areas where biomass remains relatively high even when accounting for variability. These areas may benefit from moderate salinity levels influenced by freshwater inflows, creating environments that are conducive to sustaining higher biomass levels. Overall, the map highlights productive regions (in green to yellow) along the Gaeta-Genova coastline, likely reflecting the interplay of favorable bathymetry, salinity, and nutrient availability.

```
ggplot(cc_upper_2008, aes(x = X, y = Y, z = Z)) +
  geom_contour_filled() +
  labs(title = "Upper 95% Confidence Bound 2008") +
  theme_minimal()
```



This map showcases the upper 95% confidence bound for shrimp density (log scale) along the Gaeta-Genova coast in 2008, offering an optimistic projection of potential biomass distribution. The regions shaded in green to yellow, with values ranging from 11 to 13.5 on the log scale, represent areas with the highest potential biomass. These zones are likely associated with optimal environmental conditions, such as favorable bathymetry (moderate depths), adequate salinity levels, and nutrient-rich waters, which together promote higher biomass productivity. Conversely, darker regions (shades of purple to blue), with values between 6.5 and 8.0, indicate areas where the potential biomass is lower, even under favorable conditions. These areas may be linked to less suitable environmental factors, such as greater depths, reduced sunlight penetration, or lower nutrient availability, which limit biomass accumulation. Bathymetry plays a central role in biomass distribution, as shallower areas are generally more productive due to greater sunlight availability that supports photosynthesis, fostering robust marine food webs. Additionally, moderate salinity, often influenced by freshwater inflows in estuarine or near-coastal areas, enhances nutrient availability, creating conditions conducive to higher biomass. This map reflects these influences, highlighting areas of potential biological richness (green to yellow zones) and offering insights into how depth and salinity gradients shape biomass distribution along the coastline.

```
interval_width_2008 <- up_2008- low_2008
cc_interval <- data.frame(X = grid_2008$X, Y = grid_2008$Y, Z = interval_width_2008)
ggplot(cc_interval, aes(x = X, y = Y, z = Z)) +
  geom_contour_filled() +
  labs(title = "Width of 95% Confidence Interval (2008)") +
  theme_minimal()
```



This map illustrates the width of the 95% confidence interval for biomass estimates along the coastal stretch from Gaeta to Liguria in the northwestern Mediterranean in 2008. The width of the confidence interval serves as an indicator of the uncertainty associated with biomass predictions: High Uncertainty (Yellow, 17.5–18.0): Predominantly observed along the outer edges, these regions exhibit the widest confidence intervals, reflecting significant variability in biomass estimates. Such uncertainty may arise from dynamic environmental factors, including fluctuating currents, variable nutrient availability, or human activities like fishing that impact local biomass levels. Moderate Uncertainty (Green, 13.0–16.5): Found in intermediate regions, these areas represent moderate confidence in biomass estimates. The variability may result from transitional oceanographic conditions, proximity to mixed ecological zones, or the influence of seasonal factors.Low Uncertainty (Purple to Blue, 10.5–12.5): Concentrated centrally along the coast, these regions have the narrowest confidence intervals, indicating stable and predictable conditions that enhance the precision of biomass estimates. These zones are likely characterized by consistent environmental factors, such as optimal depth, stable nutrient distribution, and salinity levels, which create favorable and predictable habitats for marine life.