SAPIENZA
UNIVERSITÀ DI ROMA

SPATIAL STATISTICS AND STATISTICAL TOOLS
FOR ENVIRONMENTAL DATA
A.Y. 2024/2025

A STATISTICAL FRAMEWORK FOR UNDERSTANDING PARAPENAEUS LONGIROSTRIS: BAYESIAN AND MLE APPLICATIONS

Group 12

LORENZO BARTOCCI   FRANCESCO NATALI
LEONARDO AGATE

TABLE OF CONTENTS

SAPIENZA
UNIVERSITÀ DI ROMA

01 DATASET INTRODUCTION

MEDITS SURVEY

The MEDITS (Mediterranean International Trawl Survey) program conducts standardized trawl surveys in the Mediterranean Sea. It aims to monitor demersal and benthic fish species, crustaceans, and cephalopods.

The dataset includes 29 variables:

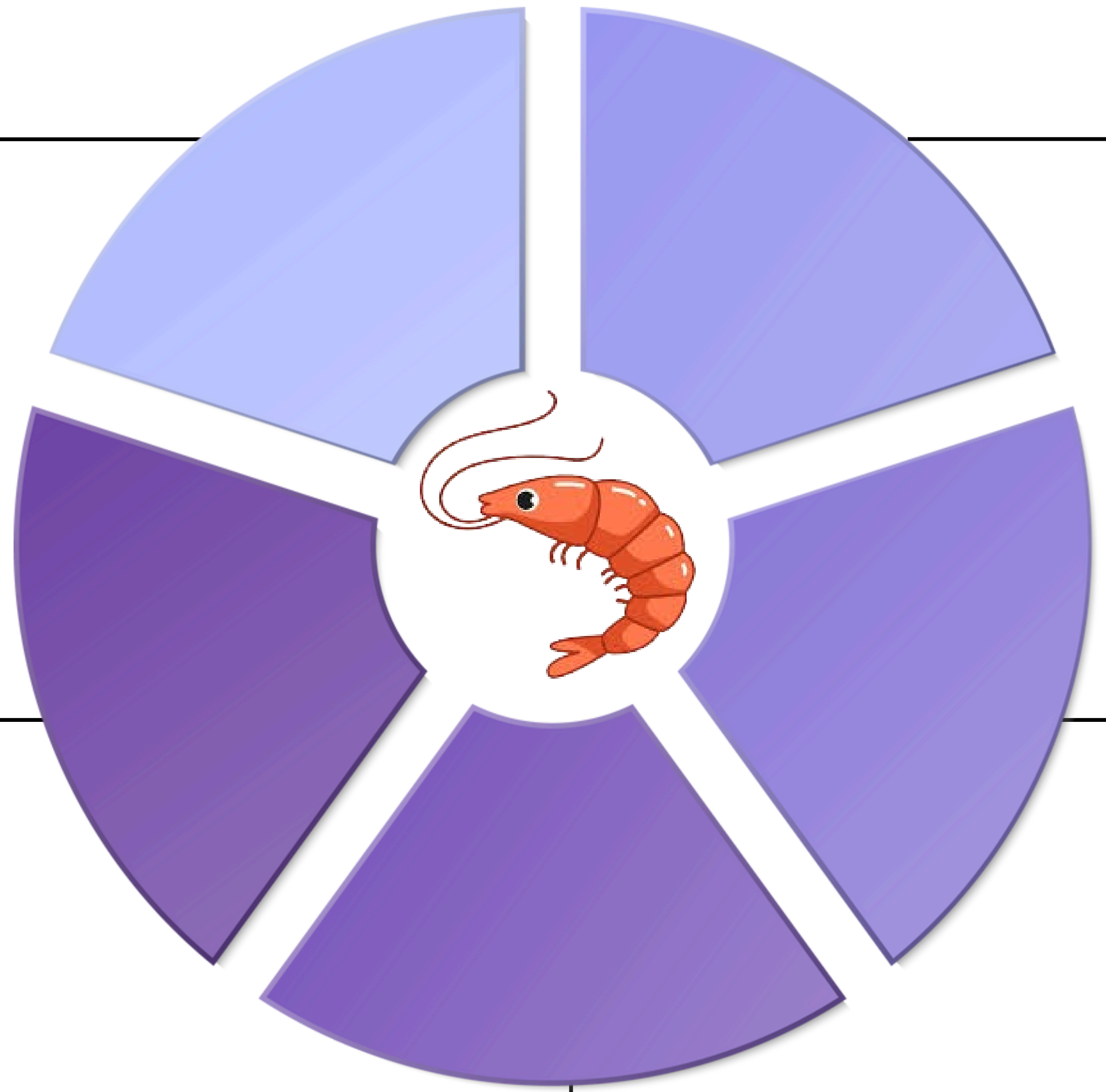
- Spatial data like **bathymetry**, distance, and slope
- Biological variables such as Spawners (adults) and Recruits (juveniles)
- Environmental factors like salinity and temperature

The main focus will be on shrimps' biomass (tot) for the years 2002 and 2008

MEDITS Project: **https://www.sibm.it/SITO%20MEDITS/Medits%20programme.htm**

SAPIENZA UNIVERSITÀ DI ROMA

About Parapenaeus longirostris
(commonly known as the deep-water rose shrimp)



Reaches 13–15 cm, with females larger than males

Lives at 50–700 m, mainly 200–400 m depth

Thrives at 12–16°C and salinity of 35–39 PSU

Prefers soft, muddy, or sandy-mud seabeds

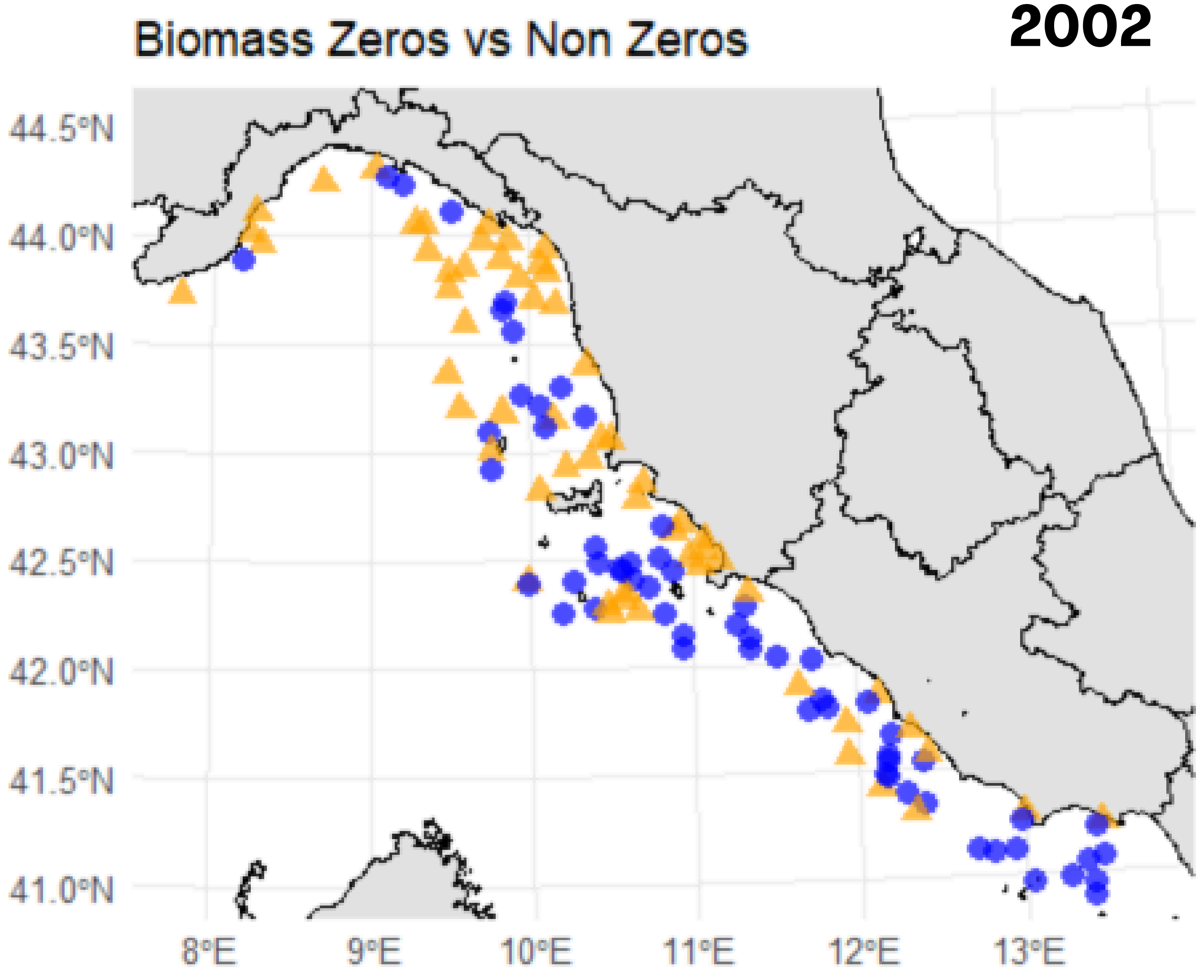
Indicator of climate change, habitat degradation, and overfishing

SAPIENZA
UNIVERSITÀ DI ROMA

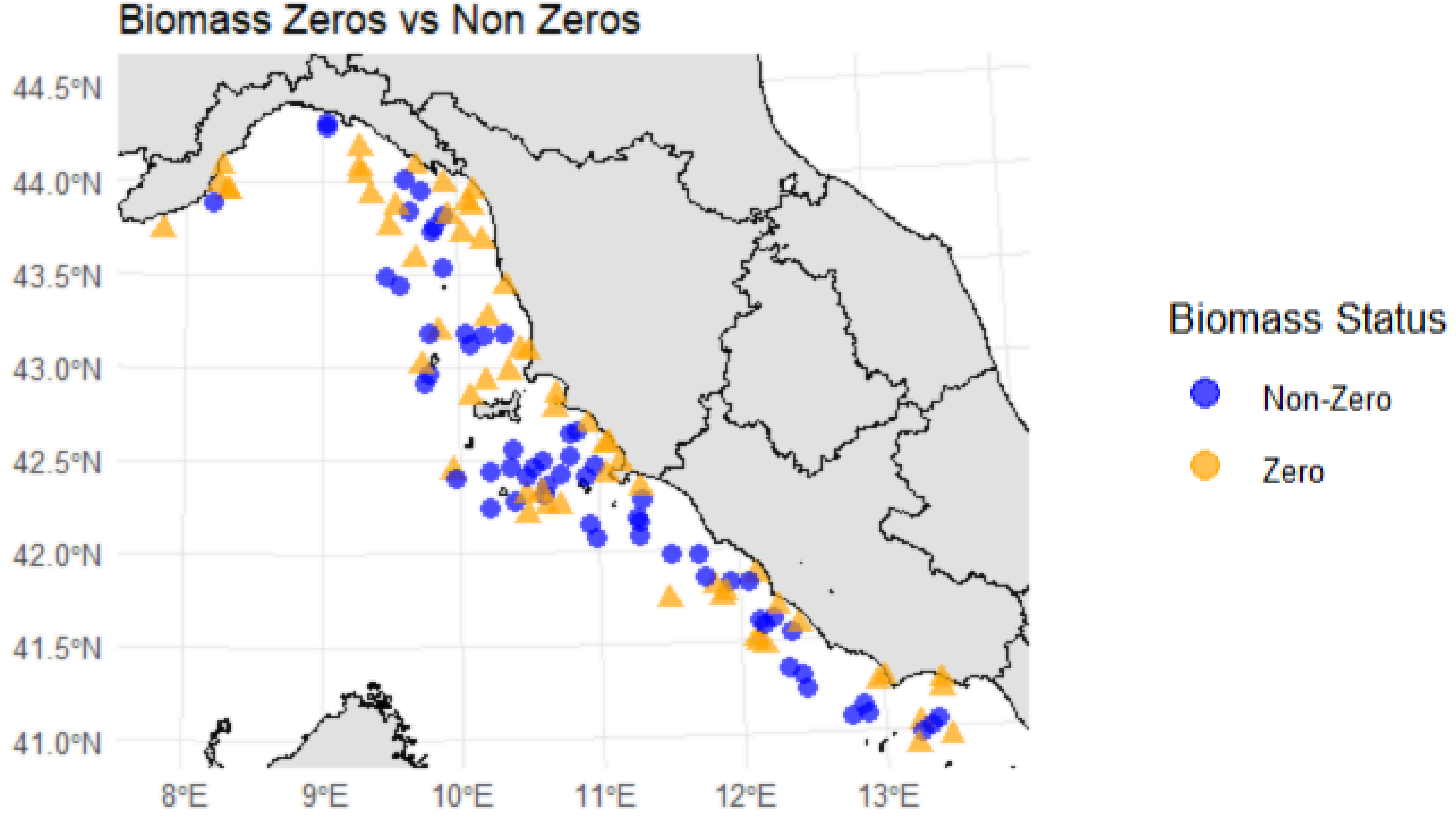
02 EXPLORATORY DATA ANALYSIS

Geo-locaton of the Total Biomass

2002

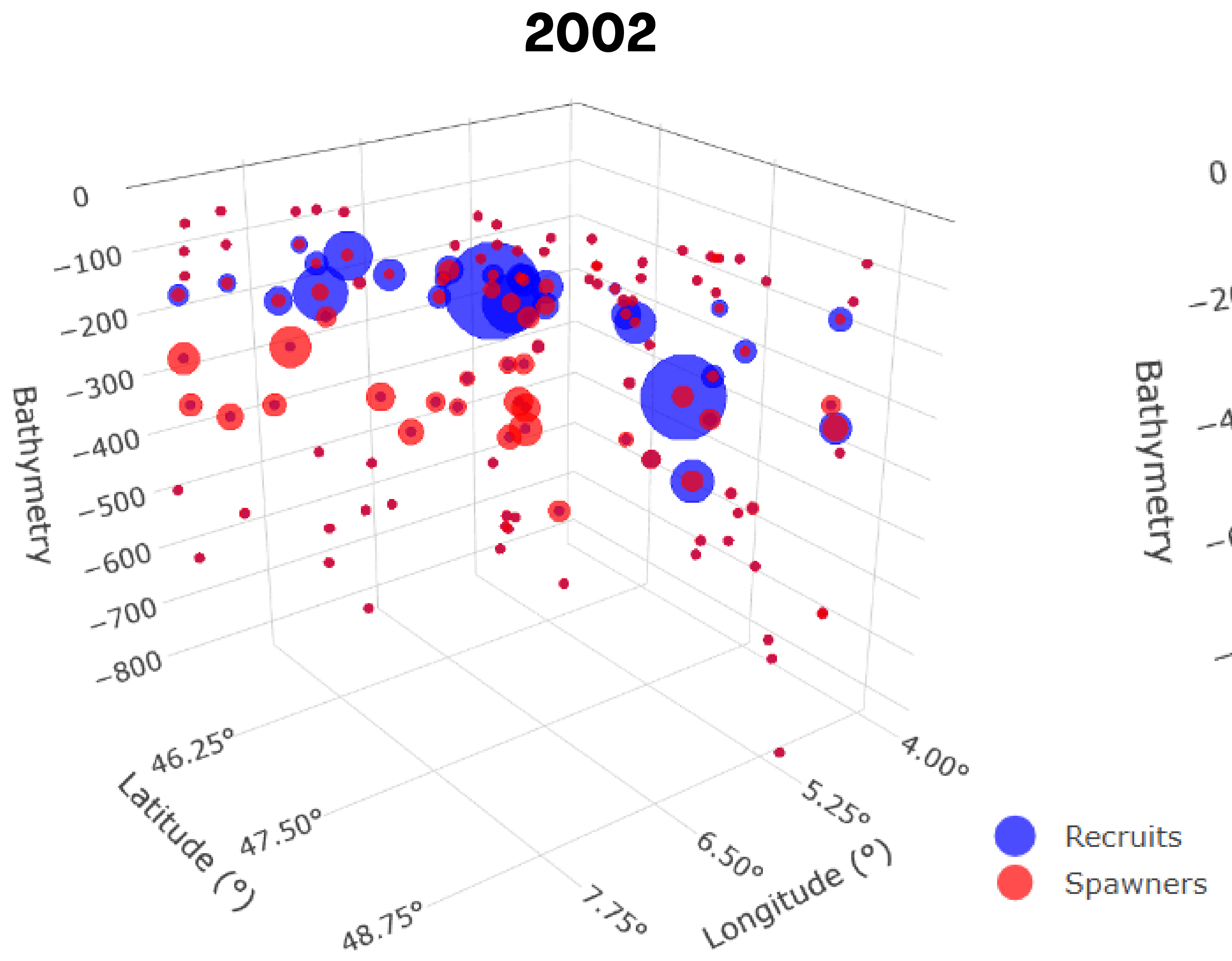


2008



SAPIENZA
UNIVERSITÀ DI ROMA

Spatial Distribution of Spawners and Recruits of Shrimps in 2002 and 2008



SAPIENZA
UNIVERSITÀ DI ROMA

Spatial Distribution of Spawners and Recruits of Shrimps in 2002 and 2008

Tot: Recruits + spawners

- **Recruits**: Density of juvenile shrimps, measured in kg/km^2
- **Spawners**: Density of reproductive, mature shrimps, measured in kg/km^2

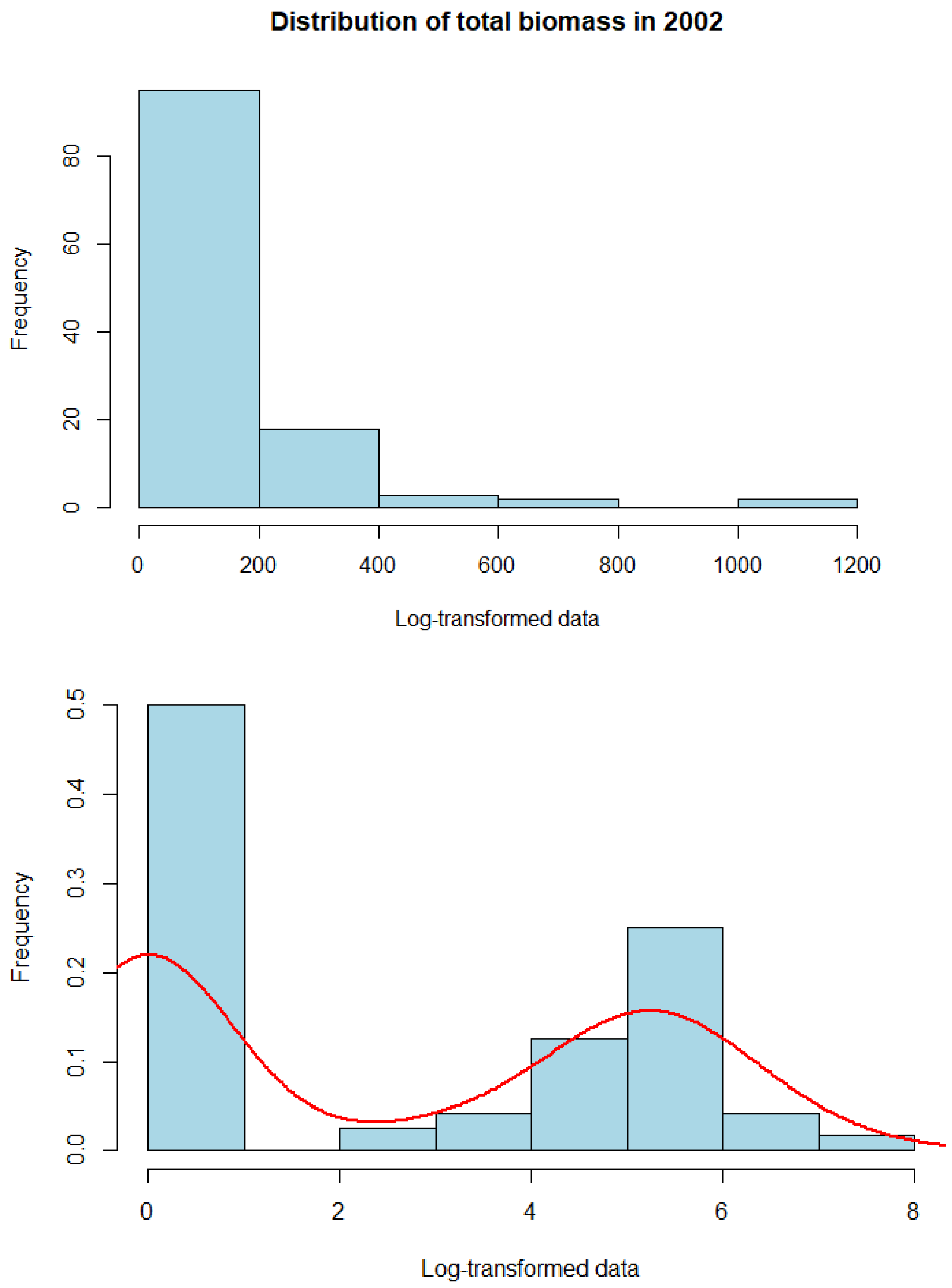
Medium-sized recruits' clusters in 2002, large clusters of spawners in 2008

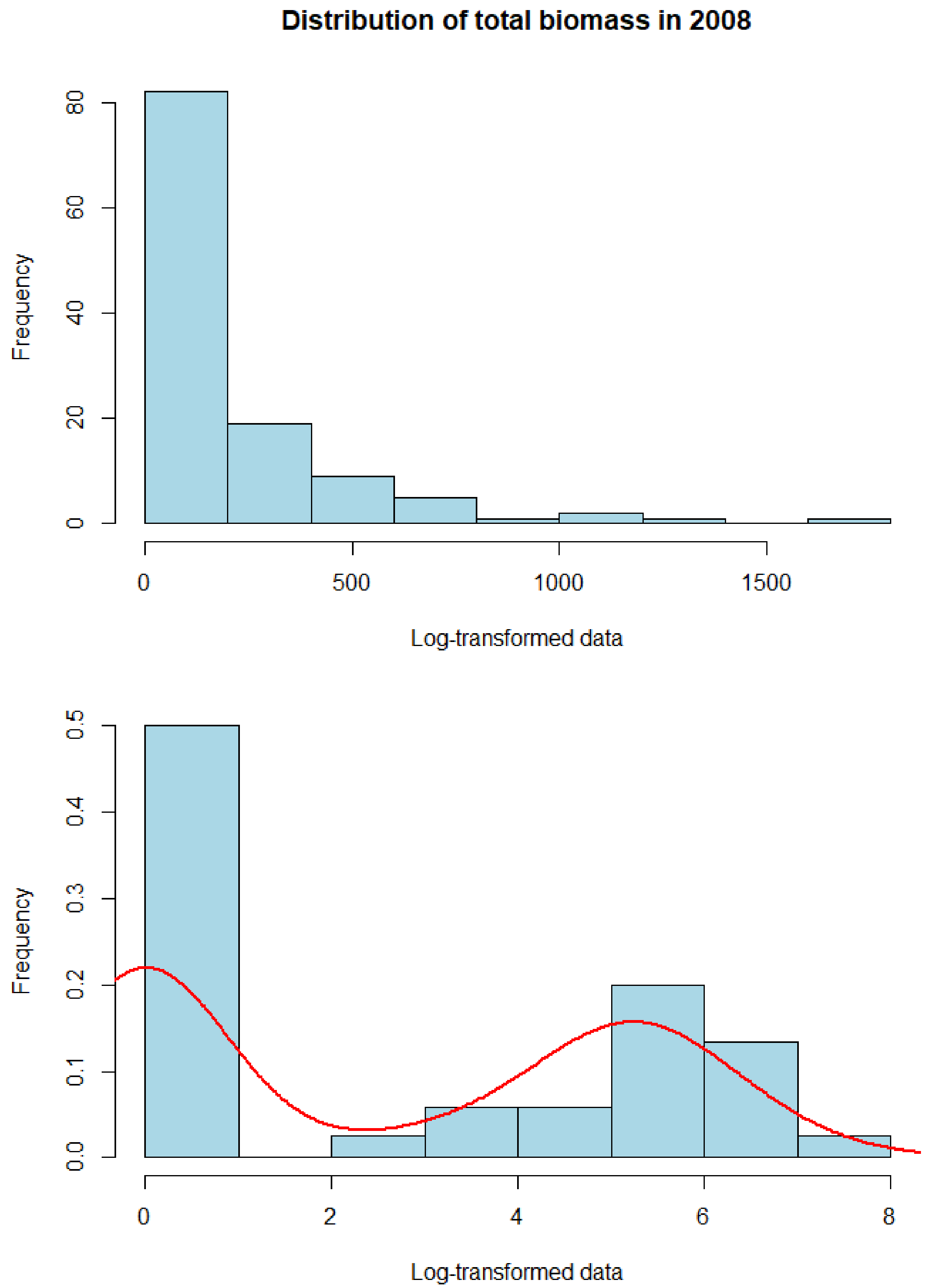
- Plausible explanations: **Overfishing** (**2002**), predator pressure, life cycle dynamics
- Shrimps' biomass concentrated at **200-500 m depth** in both years

We cannot use them as covariates

Sapienza Università di Roma

Original data set scale *vs* *log-transformation*





Original data set scale vs log-transformation

Original scale (tot)

- Overall increase in biomass in 2008 (outliers on right-tail)
- No Gaussian assumptions for this dataset (lots of zero's)
- Very asymmetric on the original scale ---> Tweedie distribution
- Assume data is distributed as a Lognormal:

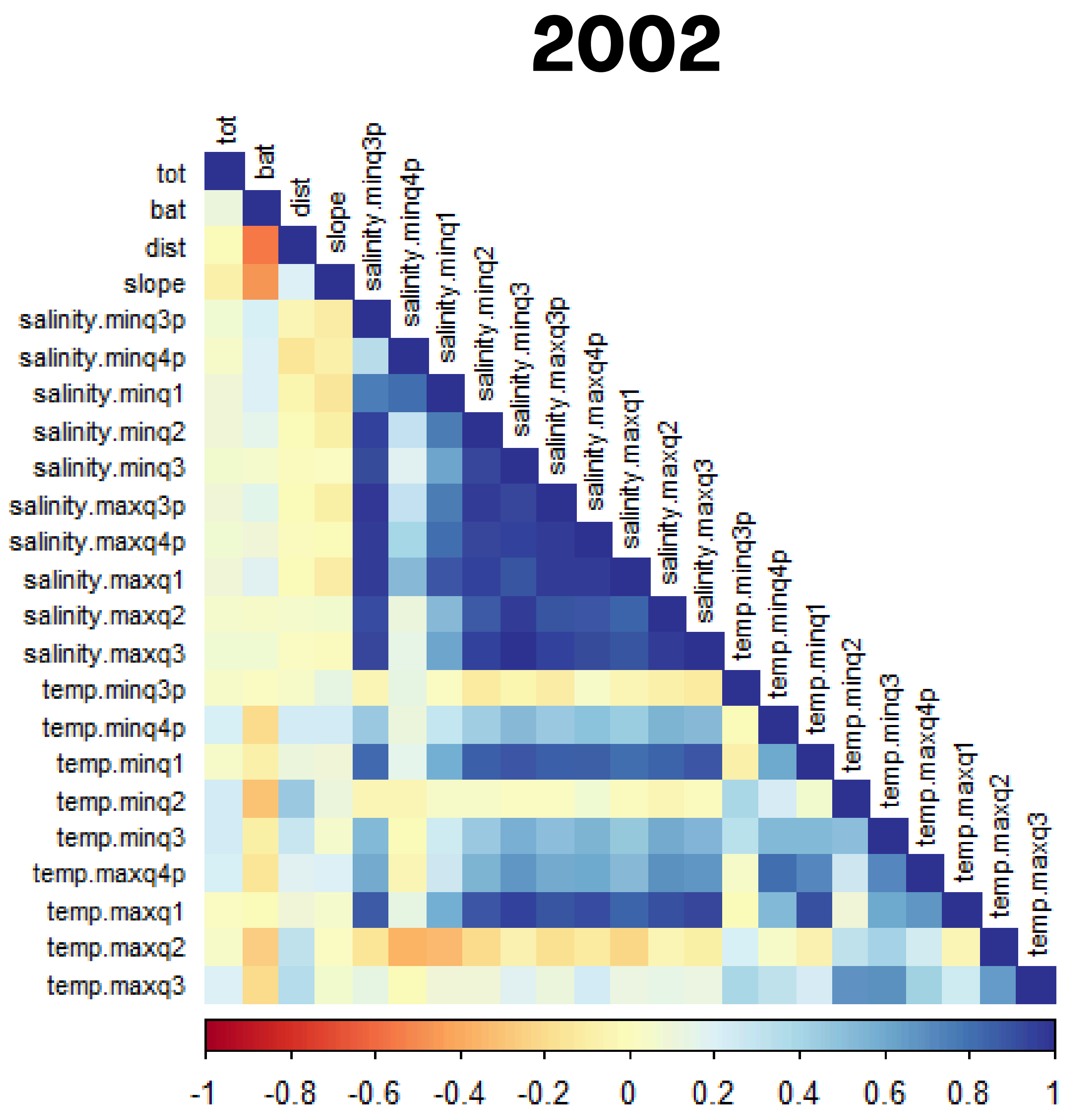
  Log transformation of the data ---> log(x+1)

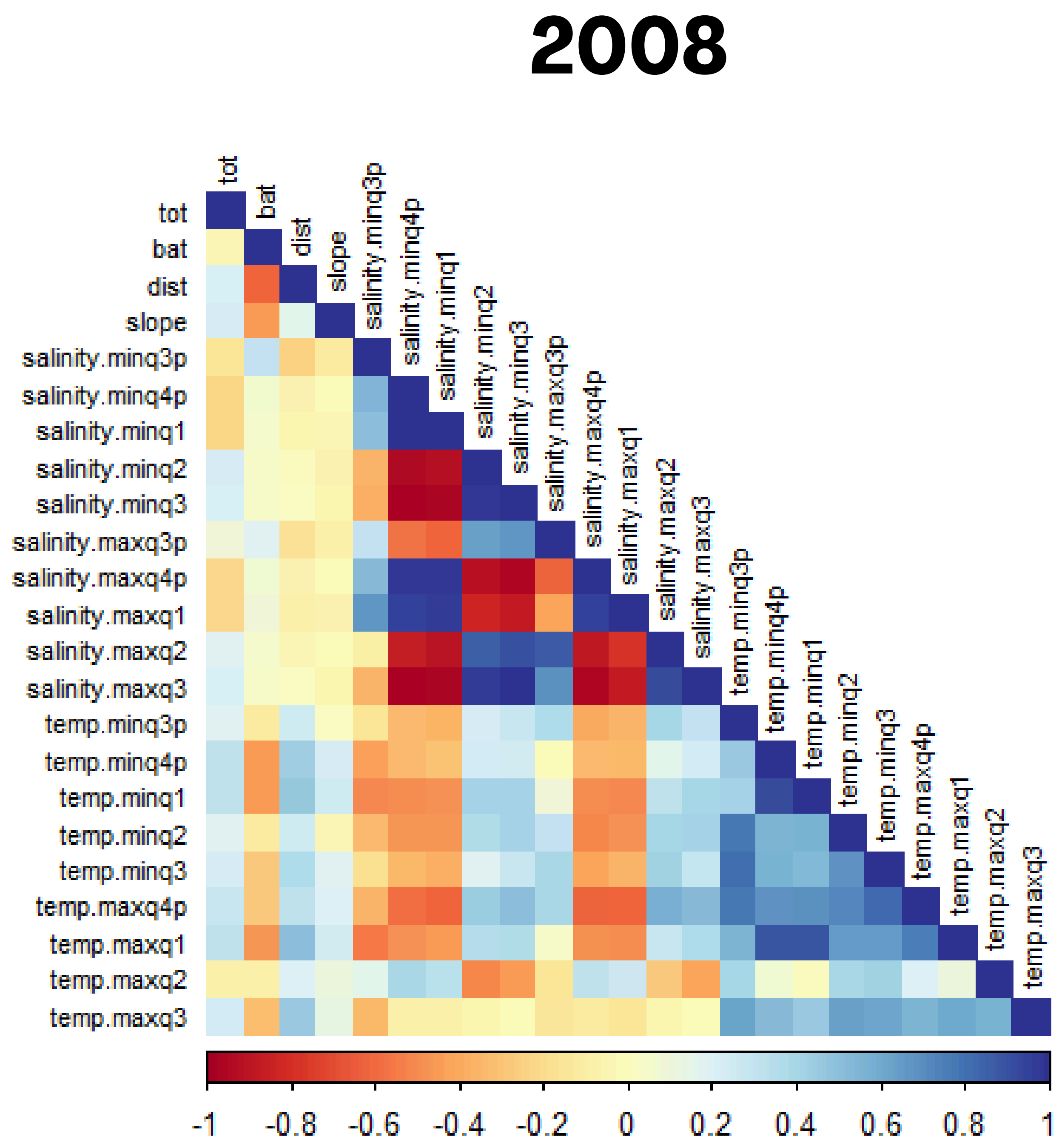
log(tot+1)

- Preservation of zero's and of symmetric distribution for non-zero values
- Lognormal: link between Mean and Variance.

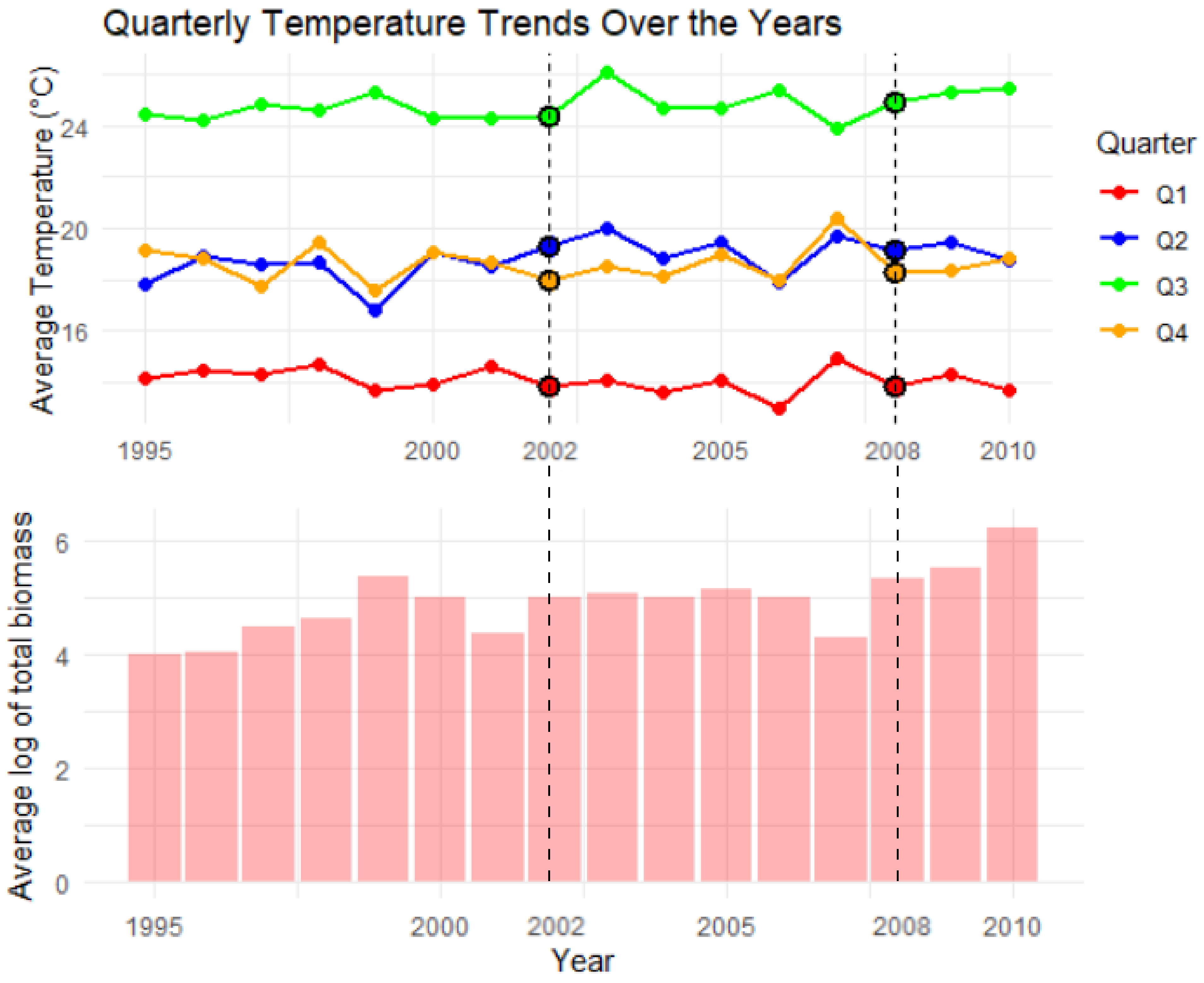
SAPIENZA UNIVERSITÀ DI ROMA

Correlation Matrices



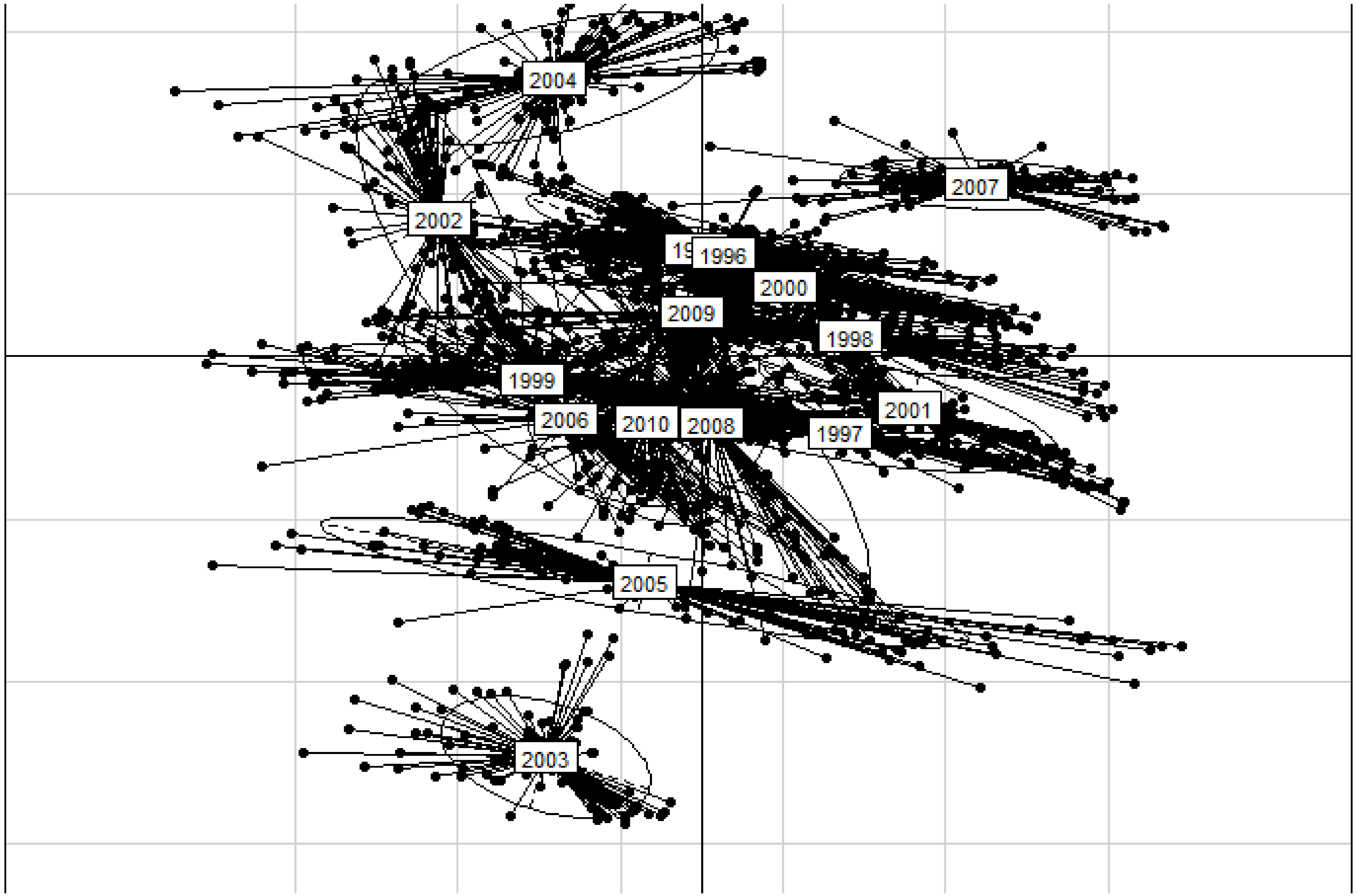


Temperature Trends and Biomass Dynamics (1995-2010)



SAPIENZA UNIVERSITÀ DI ROMA

02 PRINCIPAL COMPONENT ANALYSIS



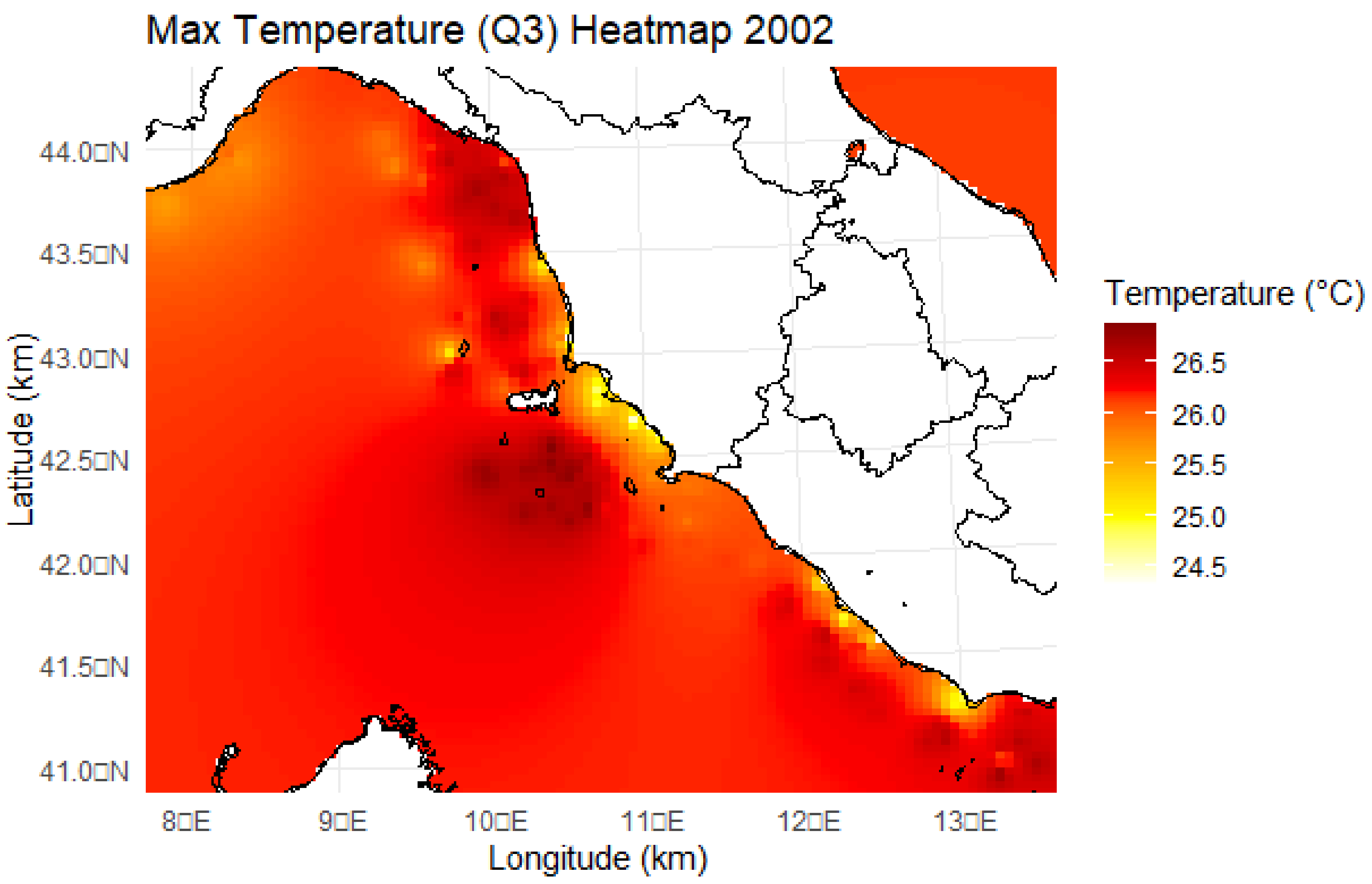
Choice of the covariates: loadings of the PCA

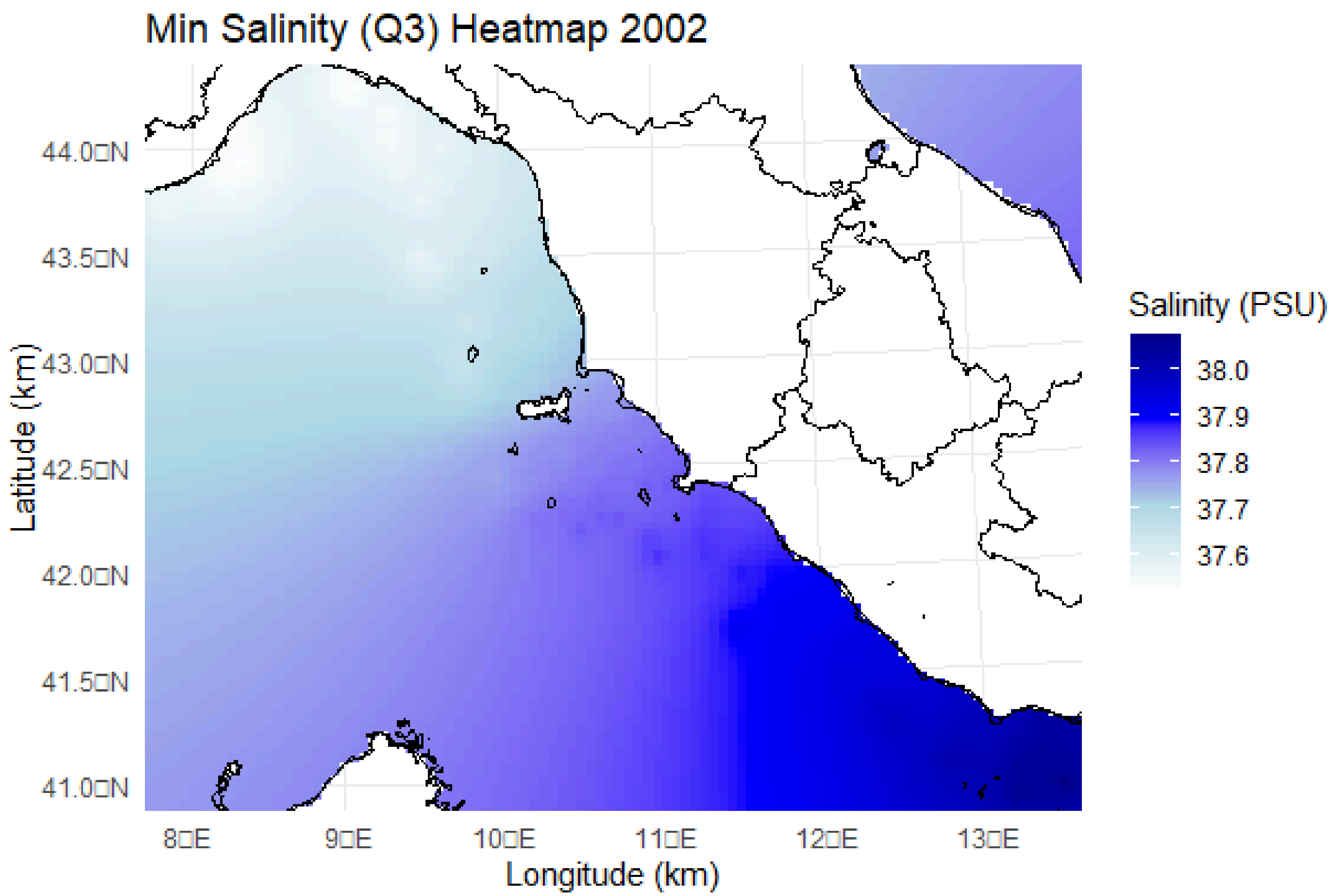
- **First 2 components capture a large proportion of the total variance** (**e.g., 68–70%**)

(2002)	CS1	CS2
salinity.minq3p	0.2679	0.0775
salinity.minq3	**0.2713**	0.0166
salinity.maxq4p	0.2713	0.0298
bat	0.0184	**0.2538**
dist	0.0153	**-0.2876**
temp.minq2	0.0245	-0.2946
temp.maxq3p	-0.0122	-0.2789
temp.maxq3	0.0684	**-0.4262**

(2008)	CS1	CS2
salinity.minq4p	-0.2669	0.1221
salinity.maxq3	**0.2565**	-0.1727
dist	0.0592	0.2488
bat	-0.0473	**-0.2288**
slope	0.0255	**0.2419**
temp.minq3	0.1677	0.1790
temp.maxq1	0.1882	**0.3436**
temp.maxq3	0.0877	0.2645

Heatmap 2002: Temperature and Salinity covariates

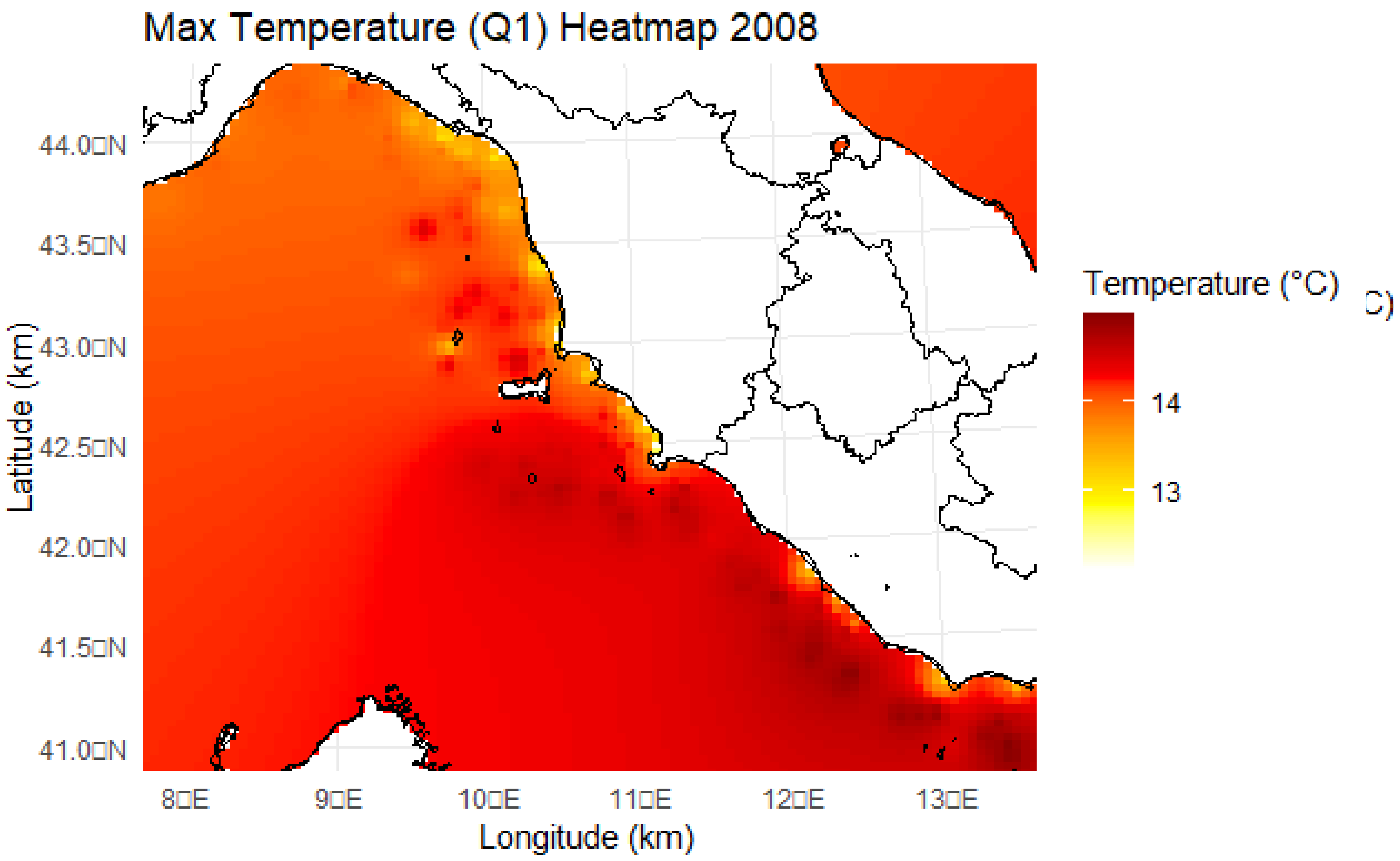


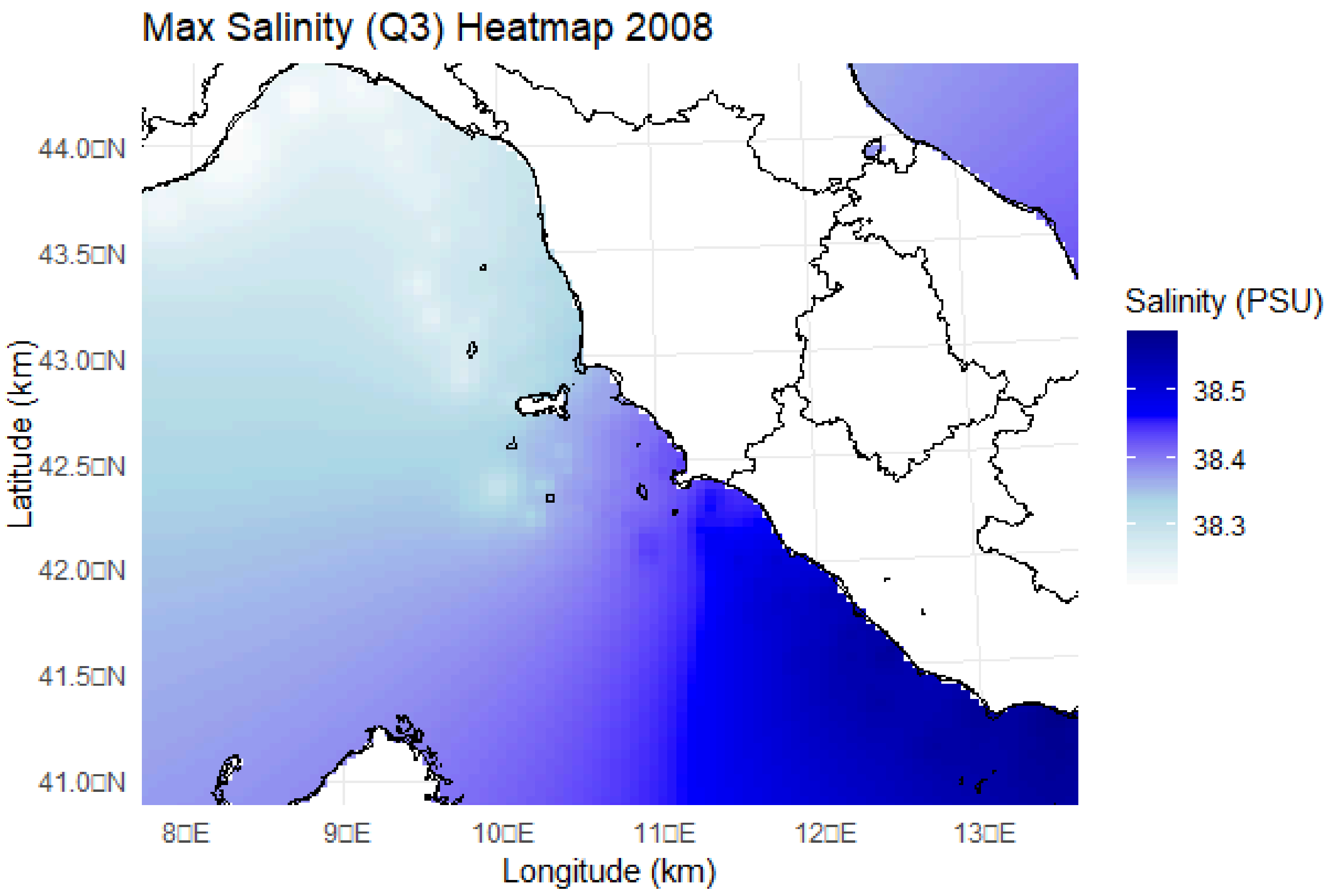


- **Temperature: Peak 26.5°C offshore, particularly in central areas**
- **Salinity: Lower ranged (37.6–38.0 PSU) near the coast.**

SAPIENZA
UNIVERSITÀ DI ROMA

Heatmap 2008: Temperature and Salinity covariates





- **Temperature: 13–14°C in winter, typical of winter conditions**
- **Salinity: Peaks at 38.5 PSU in central areas**

SAPIENZA
UNIVERSITÀ DI ROMA

Spatial Mixed Effect Model

$$Z(s) = X(s)\beta + W(s) + \varepsilon(s)$$

The spatial process $Z(s)$ is described by the following components:

- Fixed effect: $X(s)\beta$, representing the large-scale variation
- Measurement error: $\varepsilon(s) \sim N_n(0, \tau^2)$ where τ^2 is the variogram nugget effect
- Spatial random effect $W(s)$: $W(s) \sim N_n(0, \sigma^2 \cdot H_{11}(\phi))$ where σ^2 is the sill, which captures spatial dependence between locations
- Goal: To work on this Mixed Effect model and estimate it according to two different perspectives (M.L. - Maximum Likelihood and Bayesian approach)

Marginal Model

Starting from the spatial mixed effect model:

$$\boldsymbol{Z}(\boldsymbol{s}) = \boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta} + \boldsymbol{W}(\boldsymbol{s}) + \boldsymbol{\epsilon}(\boldsymbol{s}),$$

the marginal model is obtained by integrating out the spatial random effect W(s)

The covariance structure of the marginal model is given by:

$$\Sigma_{11} = \sigma^2 H_{11}(\phi) + \tau^2 I$$

where

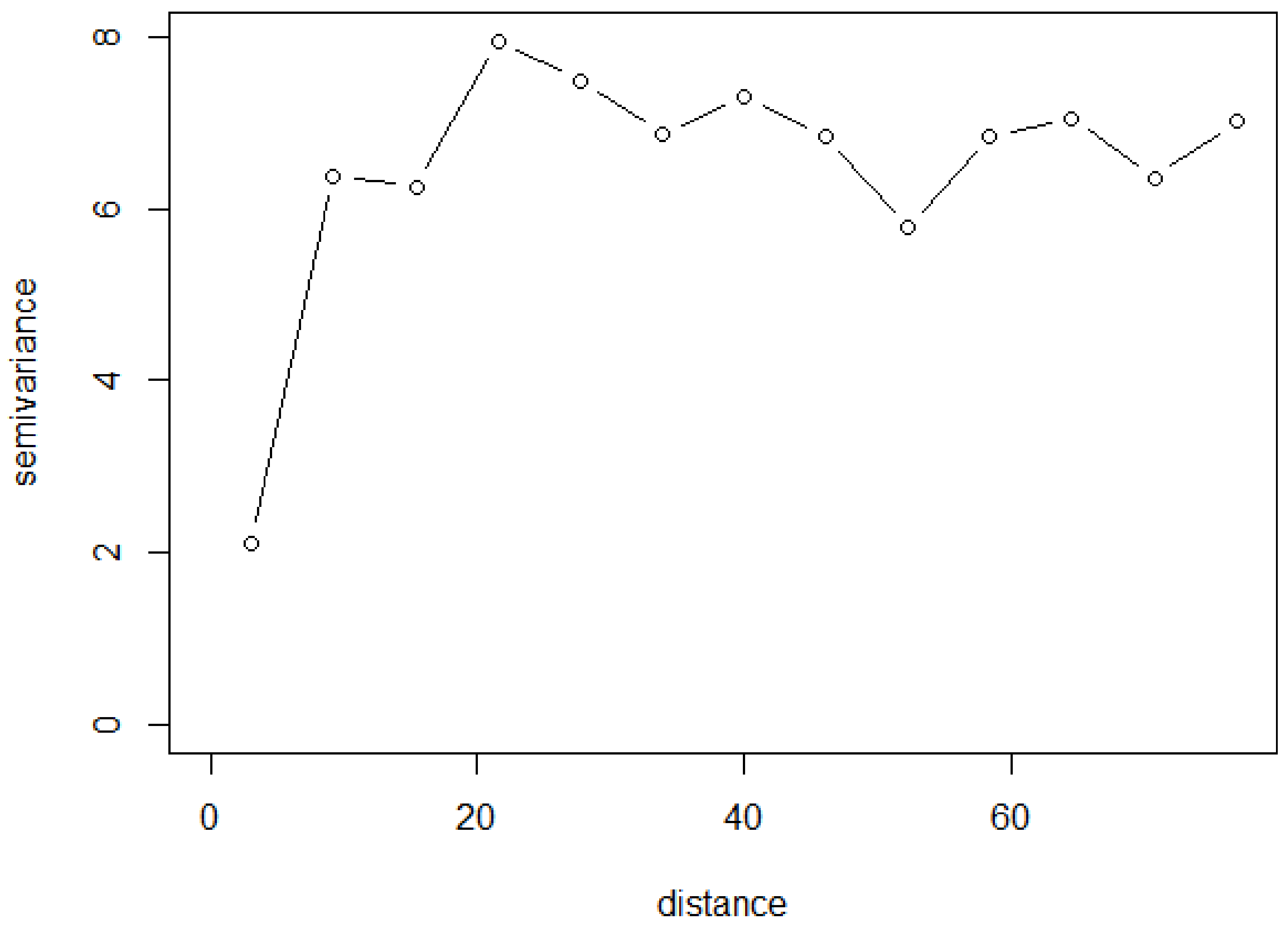
- $\sigma^2 H_{11}(\phi)$ captures spatial dependence,
- $\tau^2 I$ accounts for measurement error.

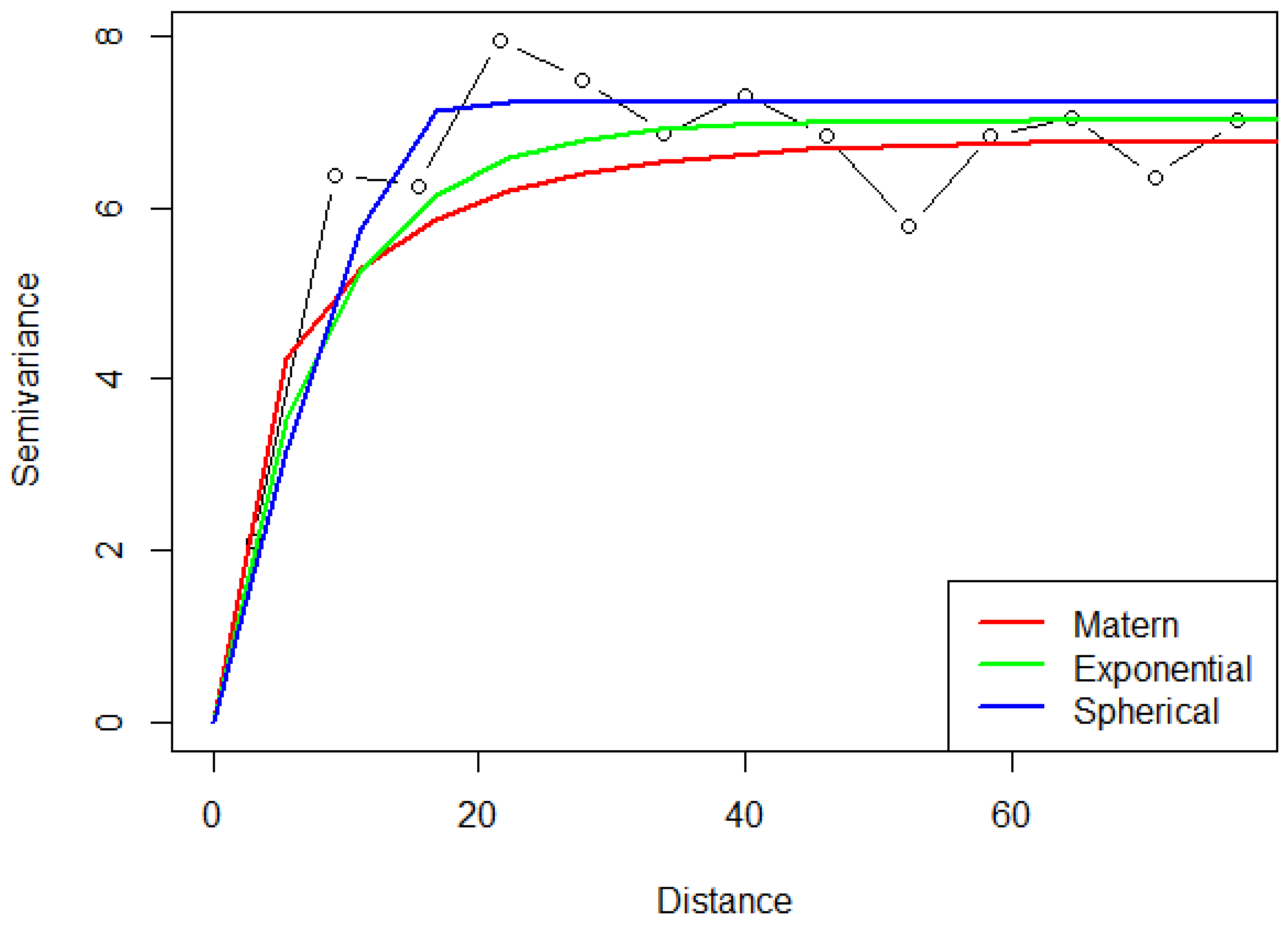
Under Gaussian assumptions, the marginal distribution of Z(s) is: $Z(s) \sim N(X(s)\beta, \Sigma_{11})$

KEY POINT: The marginal model provides a simplified covariance structure by incorporating spatial variability and measurement error into a single term.

SAPIENZA
UNIVERSITÀ DI ROMA

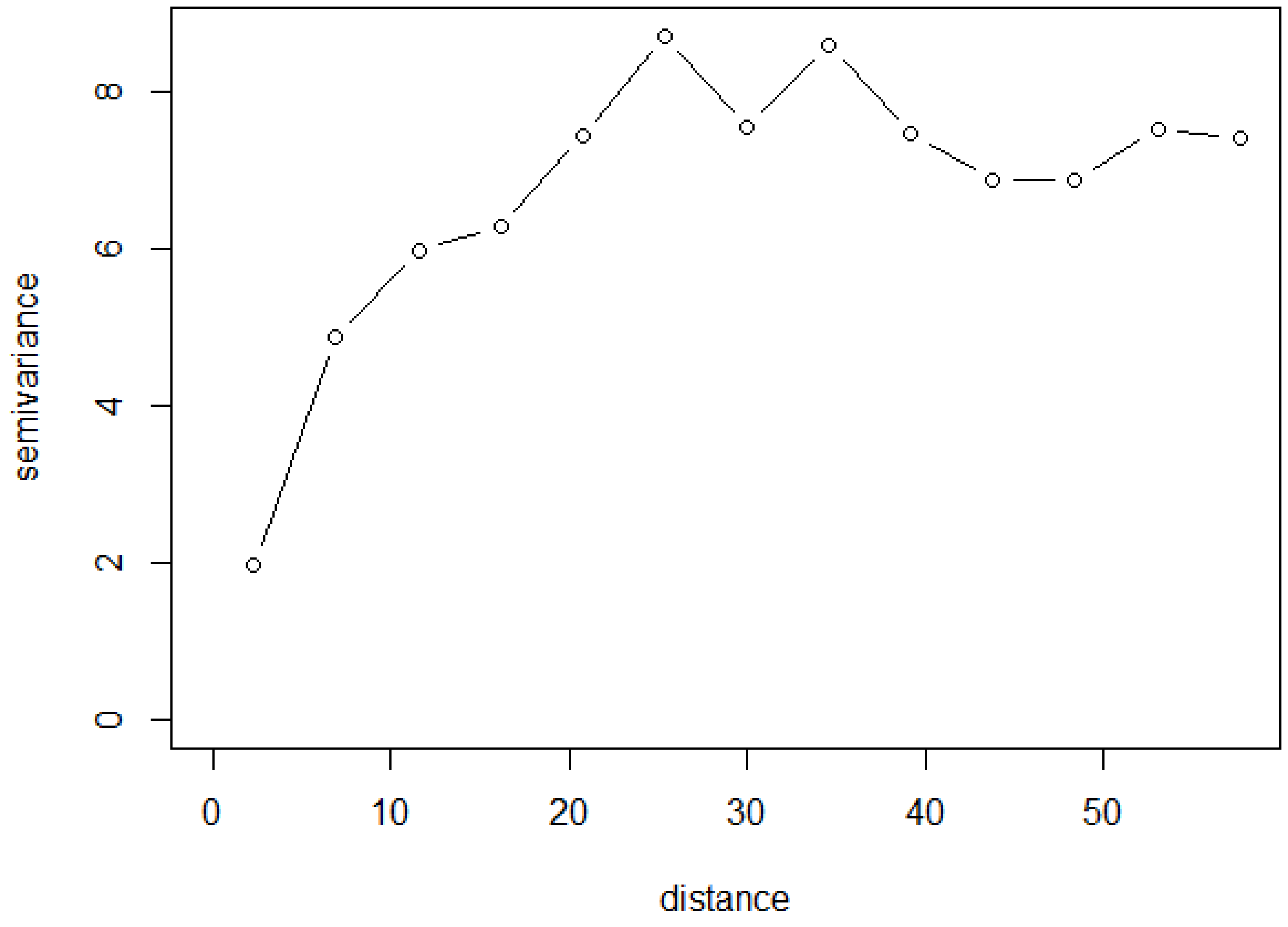
Choice of Variogram model - 2002

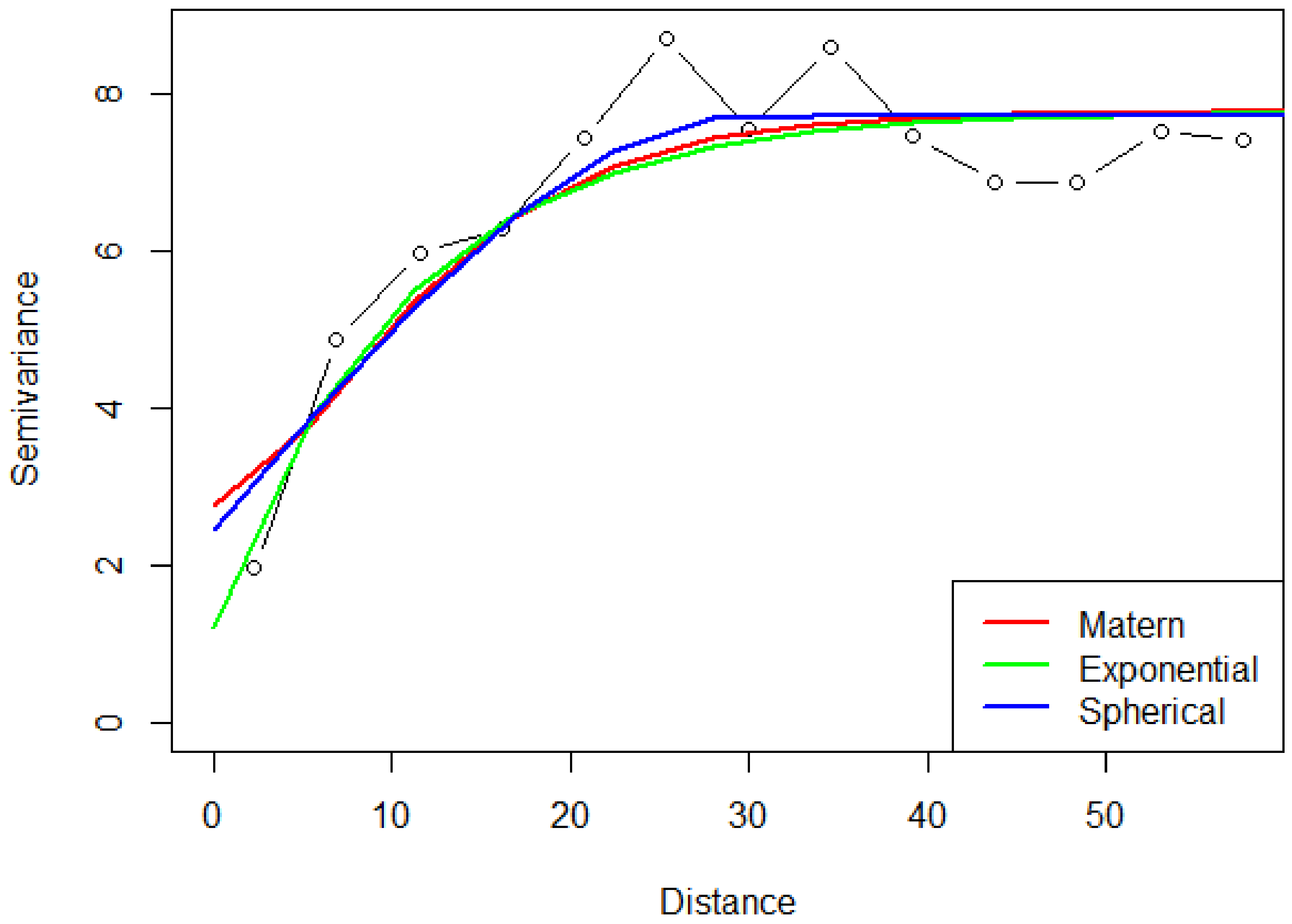




	Matern (k=0.2)	Exponential	Spherical
RMSE	2.3229	2.3778	2.3965
CV	1.025	1.047	1.057

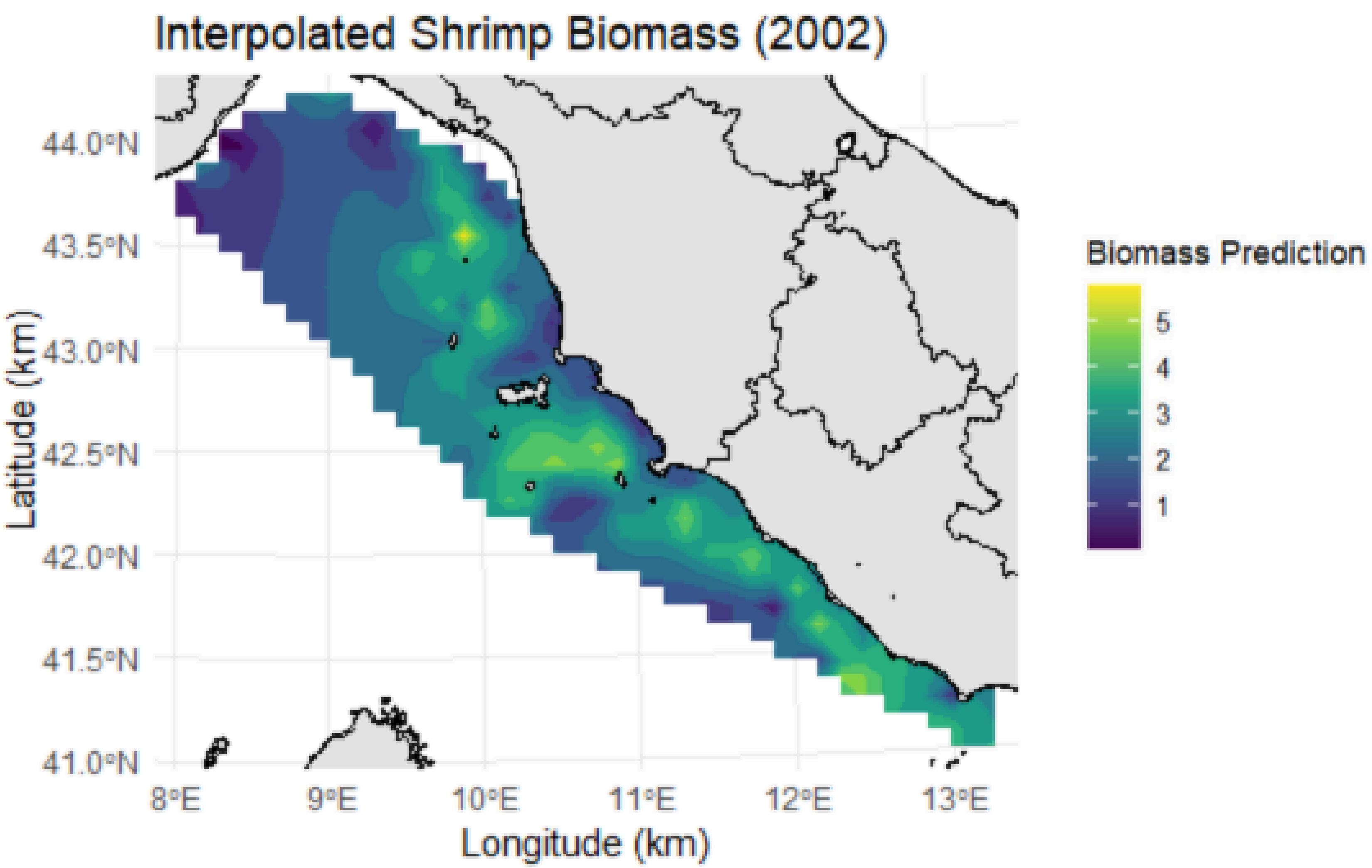
Choice of Variogram model - 2008

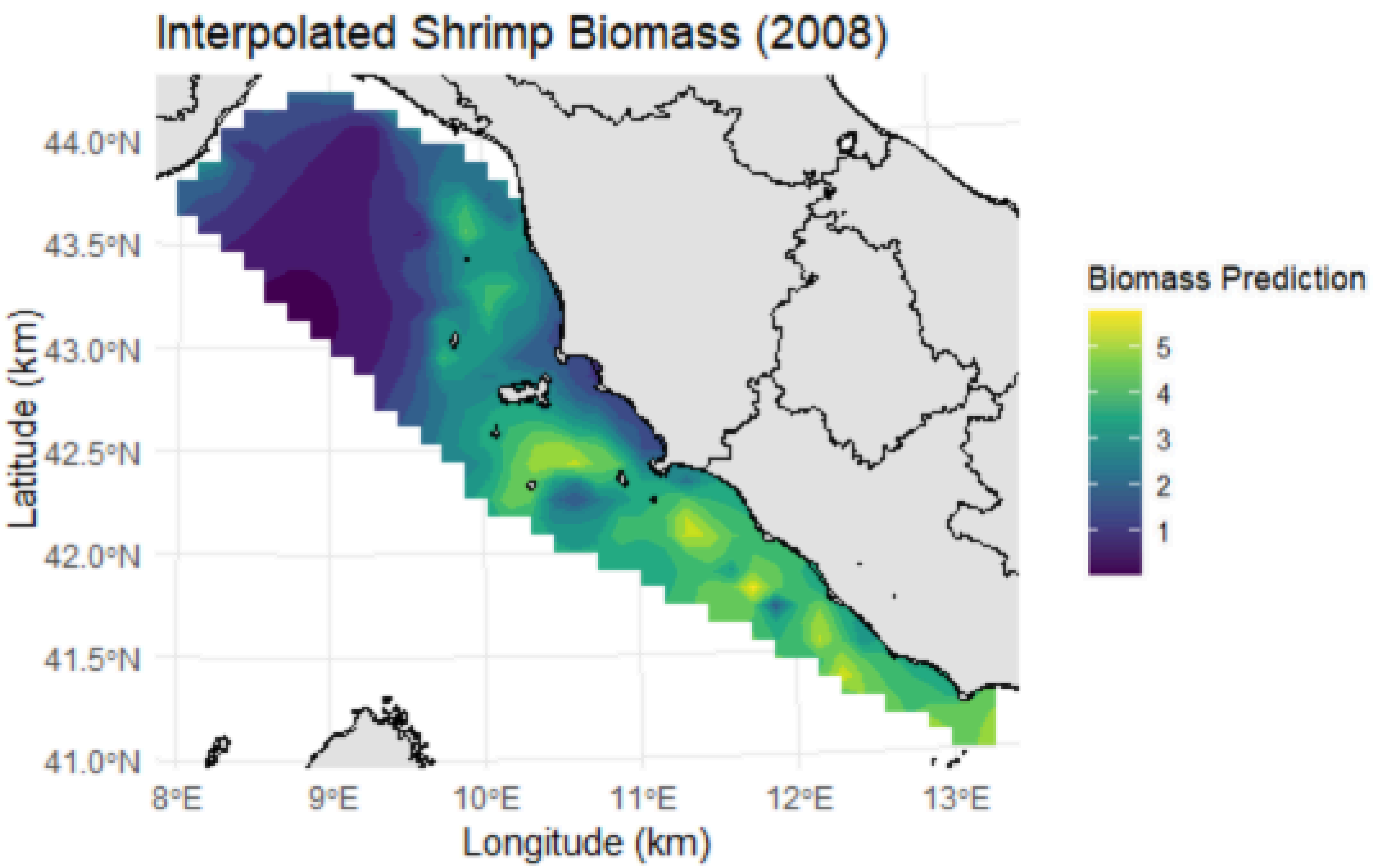




	Matern (k=0.3)	Exponential	Spherical
RMSE	2.656	2.576	2.795
CV	0.766	1.014	1.057

03 INTERPOLATION RESULTS



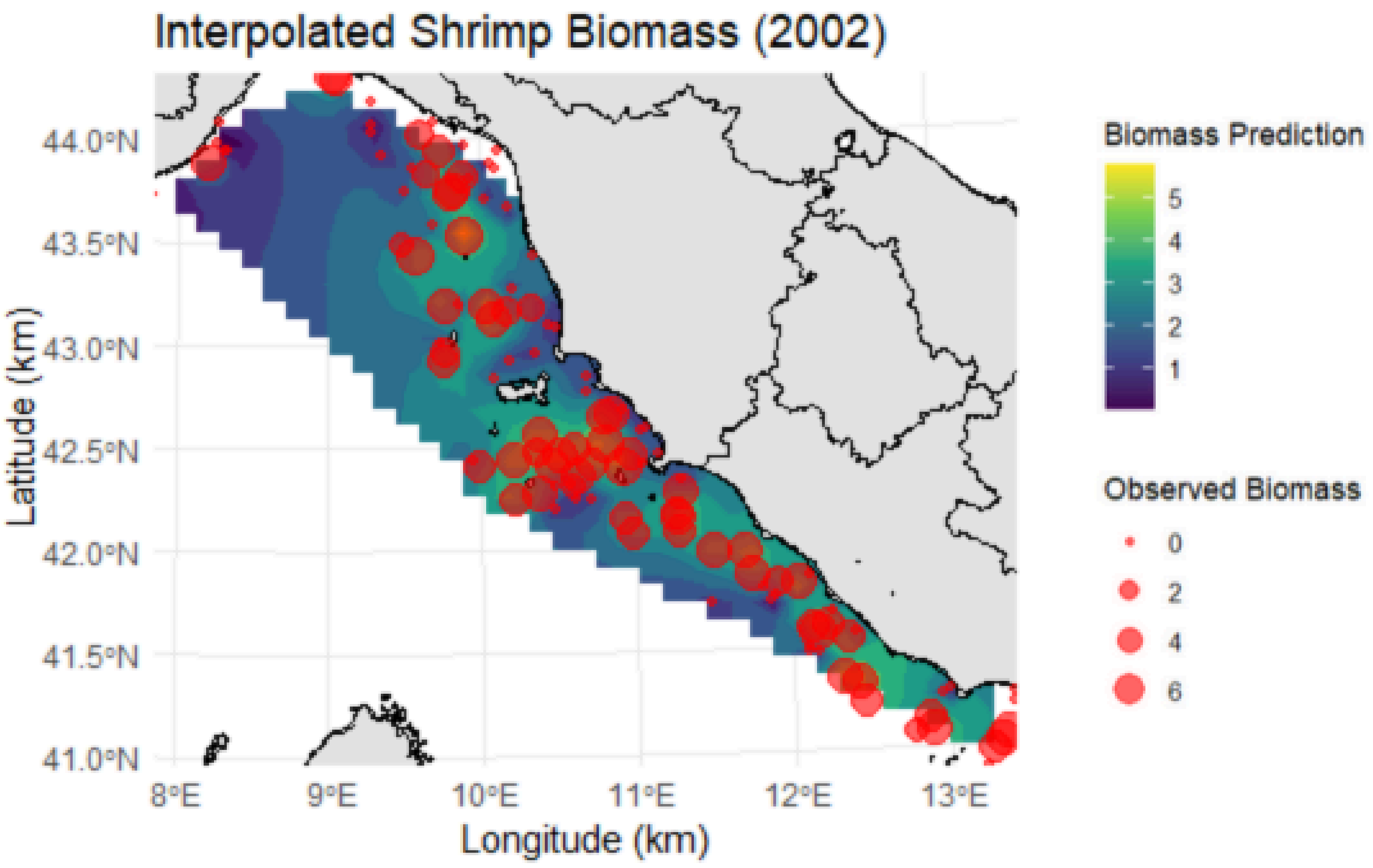


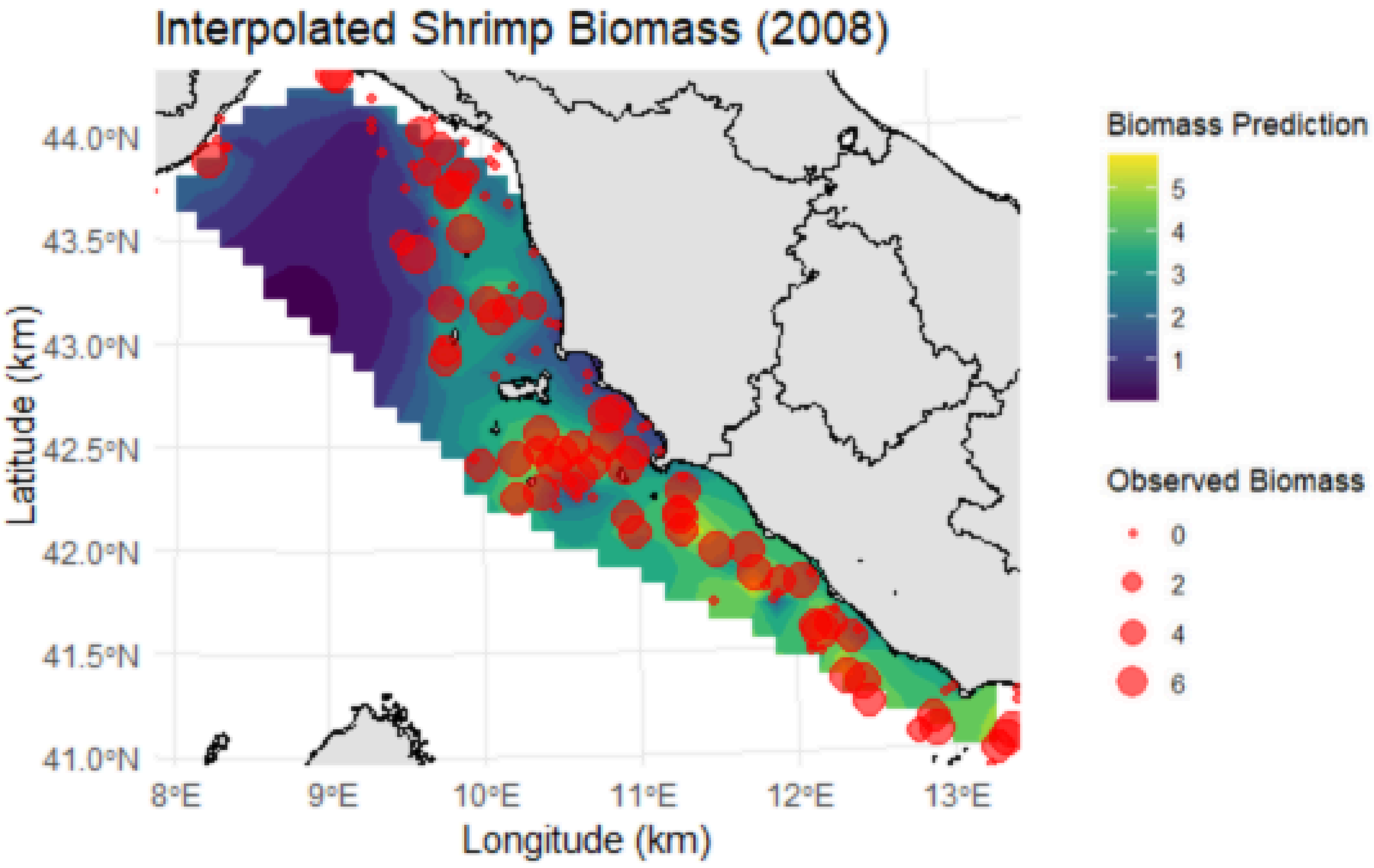
- **Higher concentration of biomass near Tuscany and Lazio**
- **Lower biomass in Liguria: steep bathymetric gradient**
- **Tuscany: more stable and favorable temperature and salinity profiles**

- **Broader spread of high biomass values overall**
- **Shift of higher concentration towards southern regions (Lazio), especially southeast, with lower values stay in nothern areas**

SAPIENZA
UNIVERSITÀ DI ROMA

Interpolated Shrimp Biomass compared with the Observed one

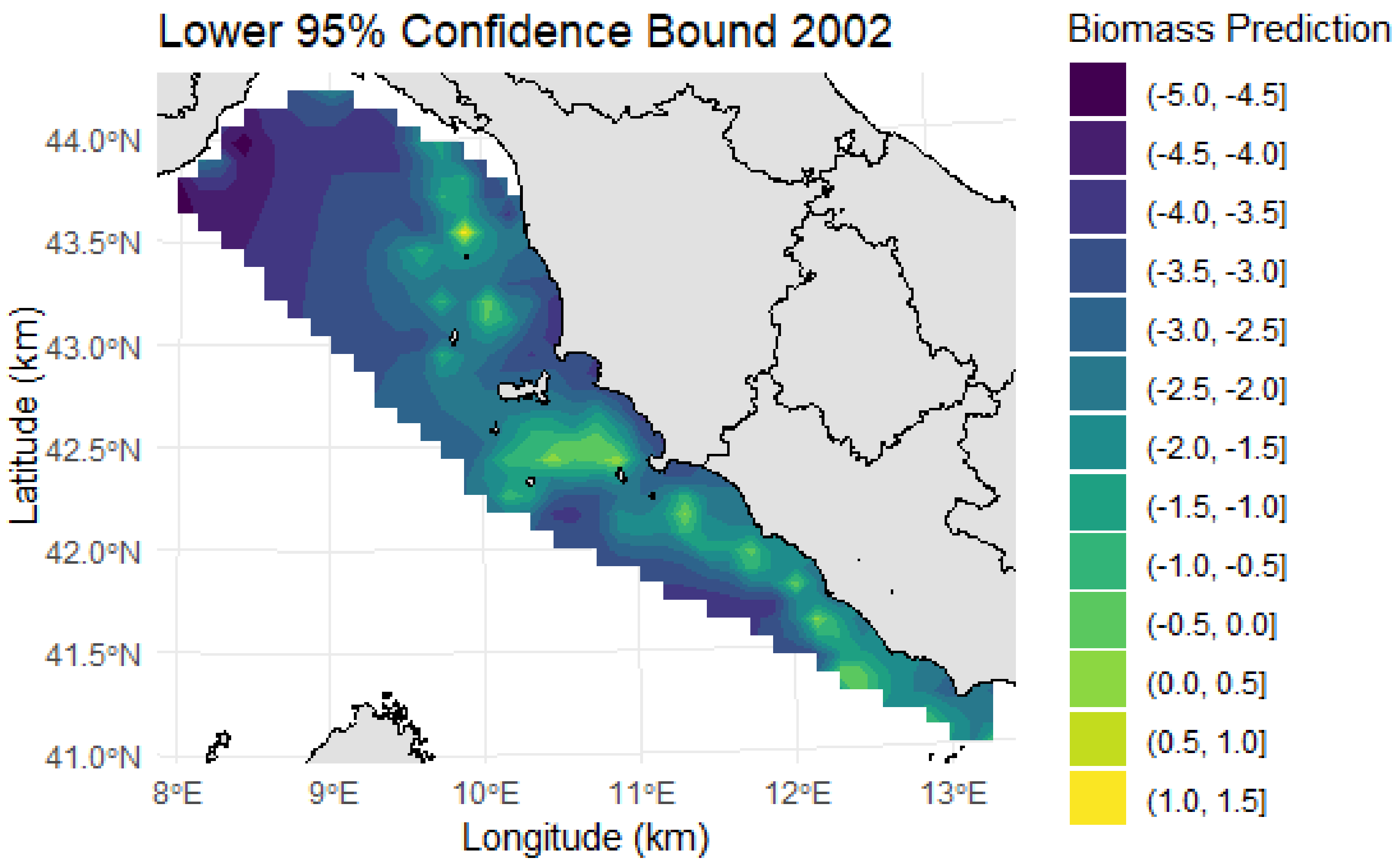


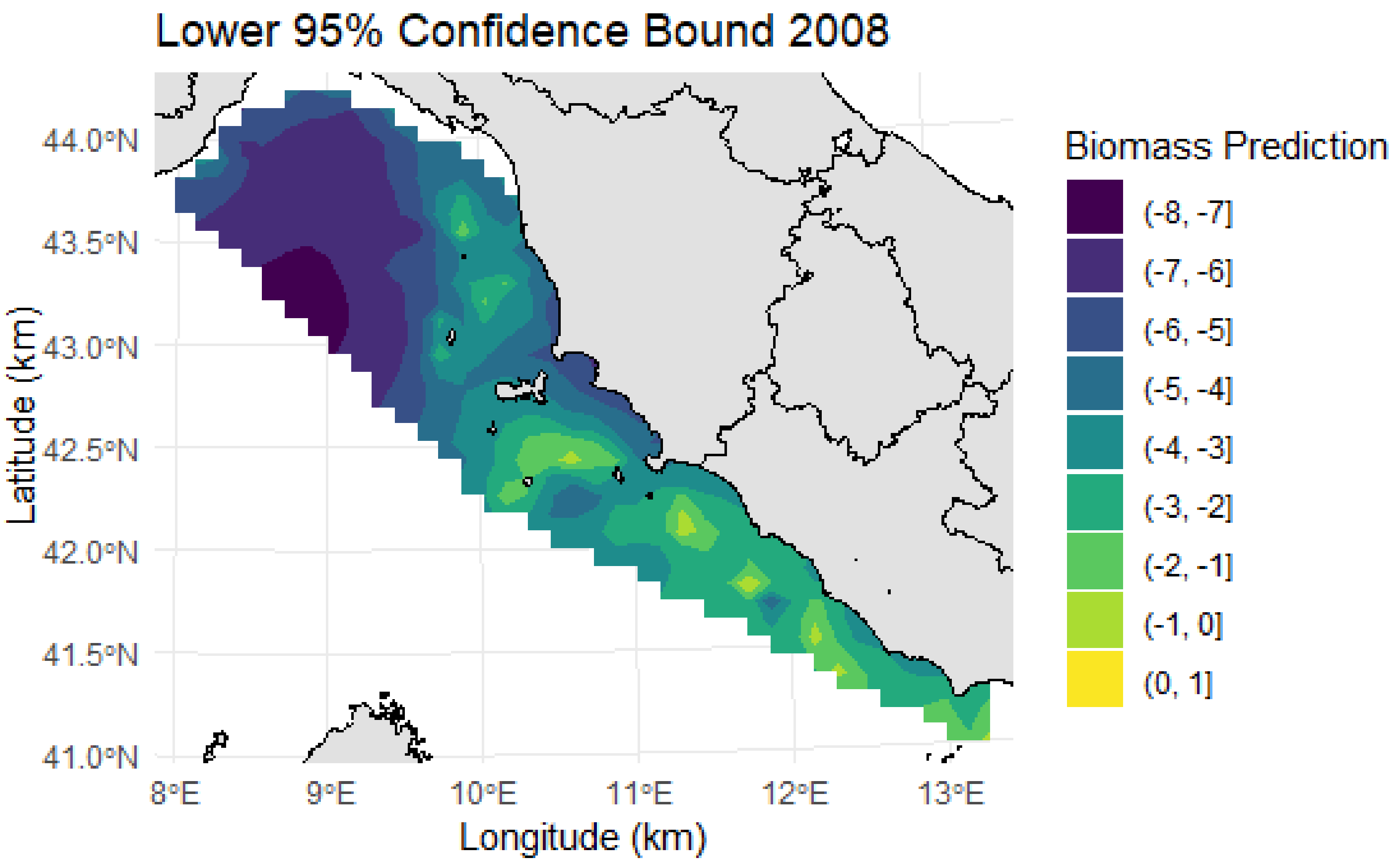


- **Distribution of red points ➡ equivalence between higher observed and predicted biomass values**
- **Spatial variability in biomass has been reasonably captured by kriging**
- **Few small discrepancies are still present**

Sapienza Università di Roma

Confidence Intervals - Lower Bounds

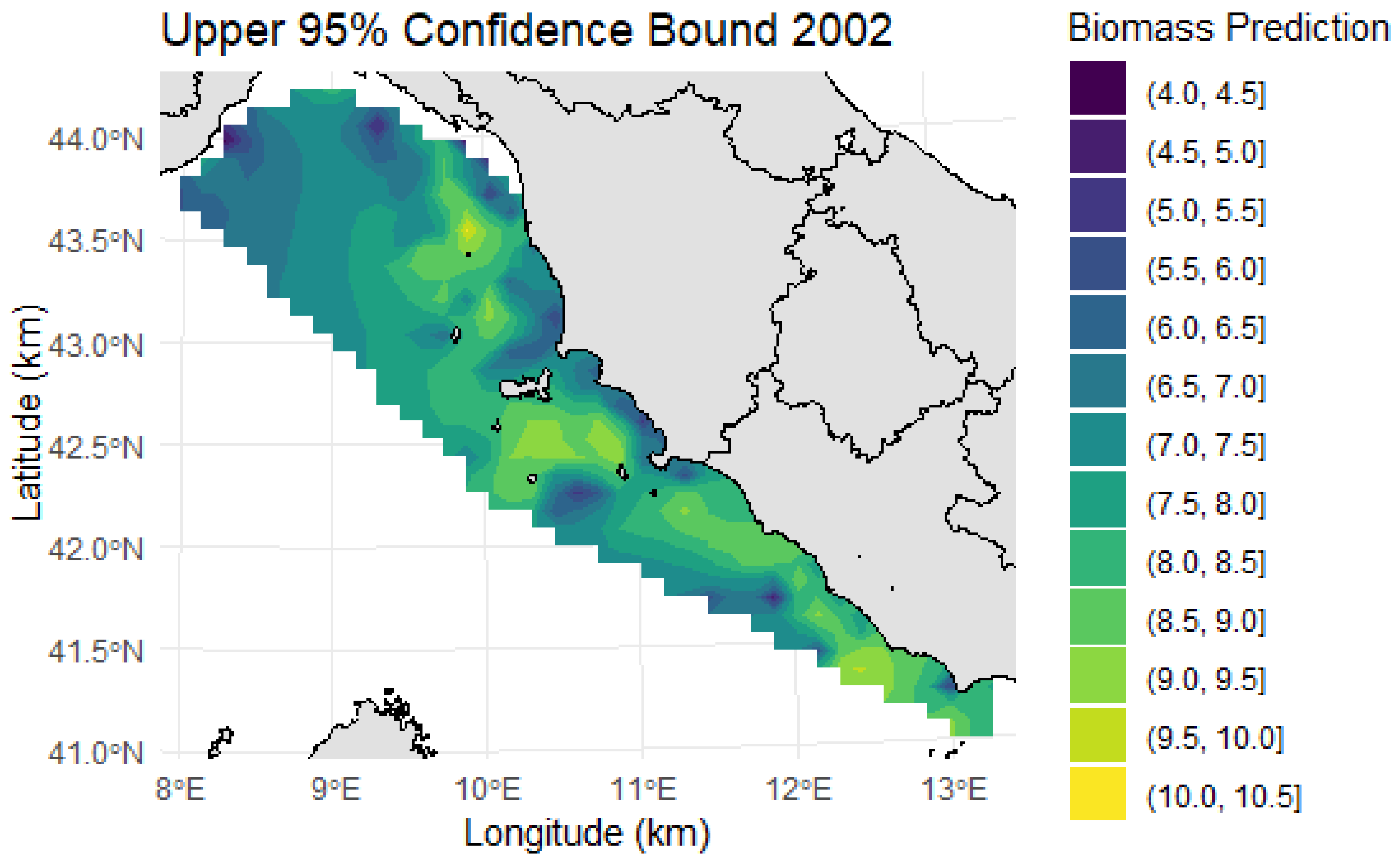


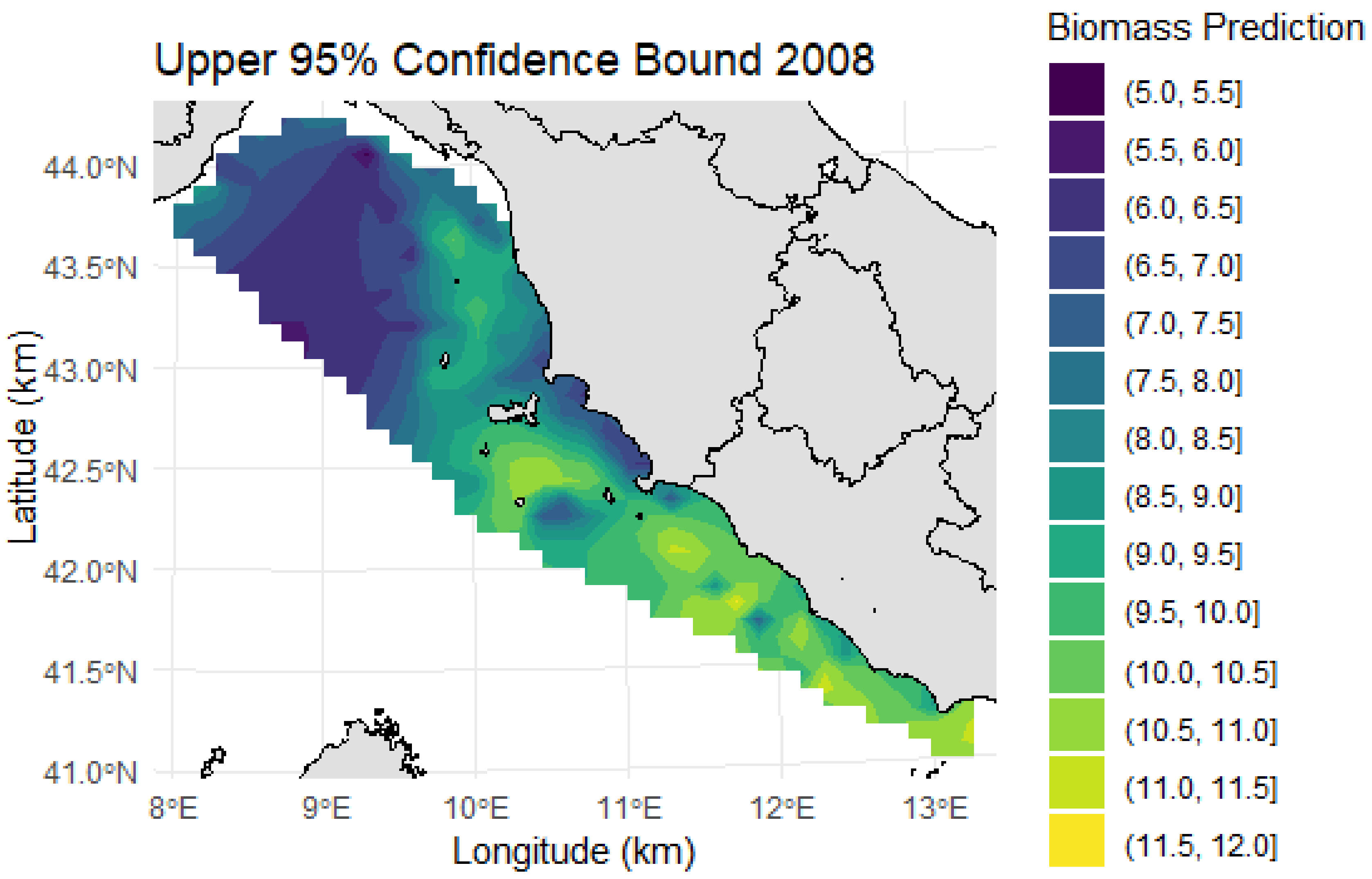


- **Lowest biomass values are more extreme in 2008**
- **Darker zones: less favorable conditions**
- **Lighter colors: areas where biomass remains high when accounting for variability**

Sapienza Università di Roma

Confidence Intervals - Upper Bounds





- **2008: higher upper bounds (more outliers)**
- **Gradual increase in the estimated vaues through the years**
- **The difference between lower and upper limits is small: estimates are likely precise with low uncertainty**

04 BAYESIAN KRIGING

Why Bayesian Kriging?

- **Inclusion of prior knowledge in our modelling:** $\pi(\beta,\omega)=\pi(\beta)\pi(\tau^2)\pi(\sigma^2,\phi)$

Provides predictions and associated **uncertainty quantification** at unmeasured locations

- **The posterior distribution, obtained by integrating over the parameters:**

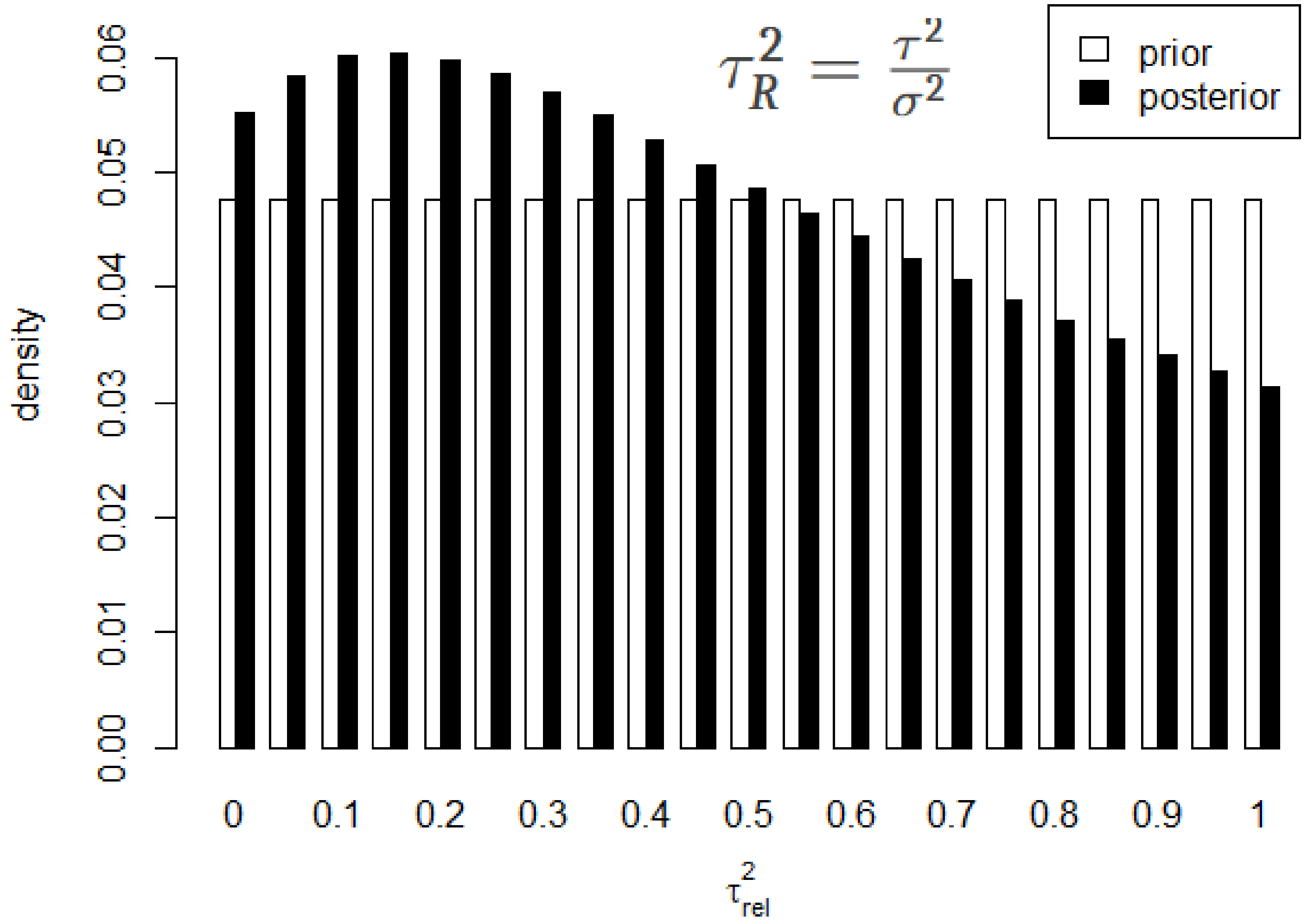
$$\pi(\beta,\omega \mid z)=\frac{\int p(Z \mid W,\beta,\tau^2)p(W \mid \sigma^2,\phi)\pi(\beta)\pi(\tau^2)\pi(\sigma^2,\phi)\,dW}{\int\cdots\int p(Z \mid W,\beta,\tau^2)p(W \mid \sigma^2,\phi)\pi(\beta)\pi(\tau^2)\pi(\sigma^2,\phi)\,dW\,d\beta d\tau^2 d\sigma^2 d\phi}$$

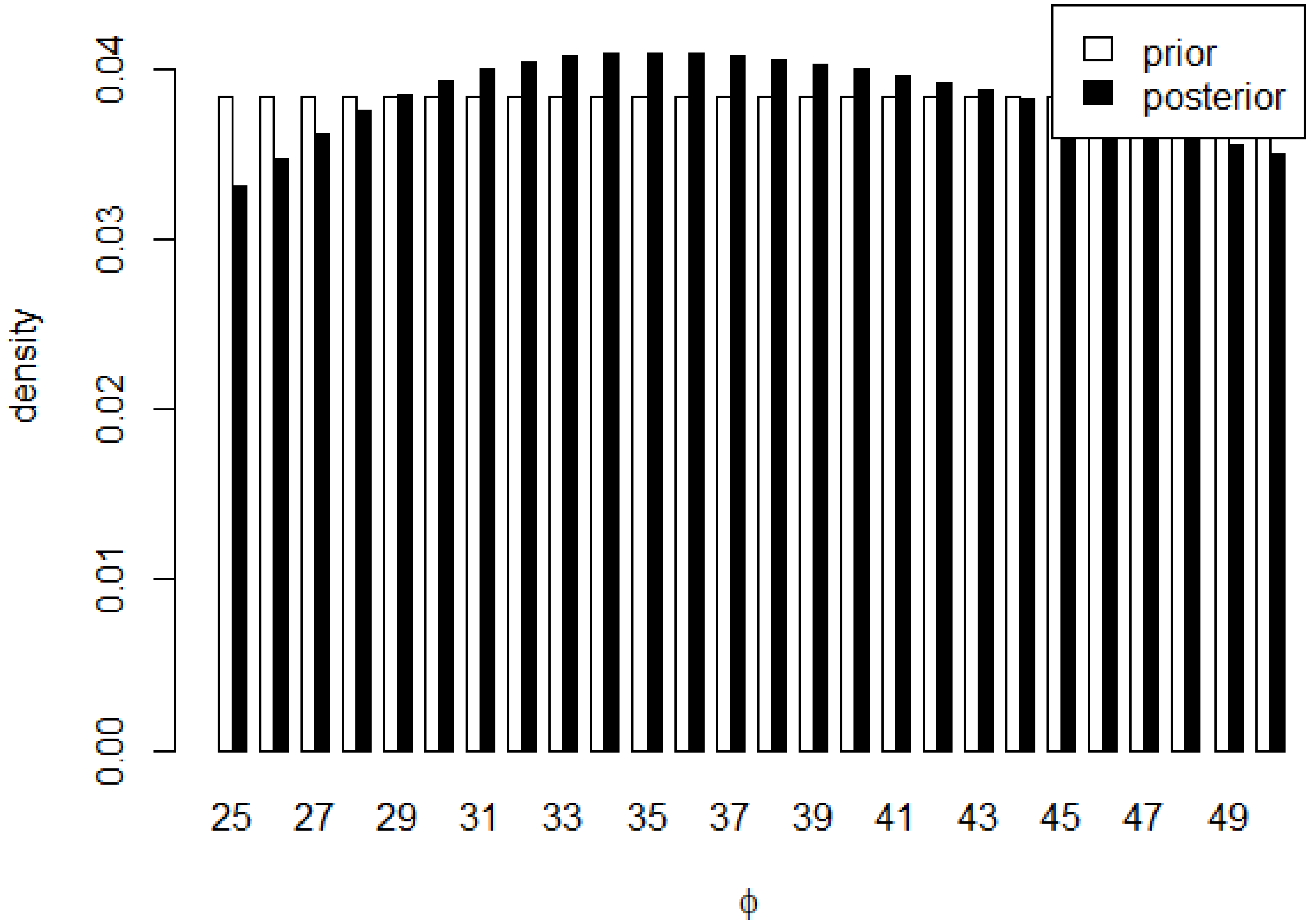
The Bayesian Kriging equation integrates the spatial model: $Z(s)=\mu(s)+\epsilon(s)$

Where:

- $\mu(s)=X(s)\beta$ **represent the deterministic trend**
- $\epsilon(s)\sim GP(0,C(h))$ **captures spatially random effects with covariance** $C(h)=\sigma^2\rho(h)$

Prior and Posterior - 2002

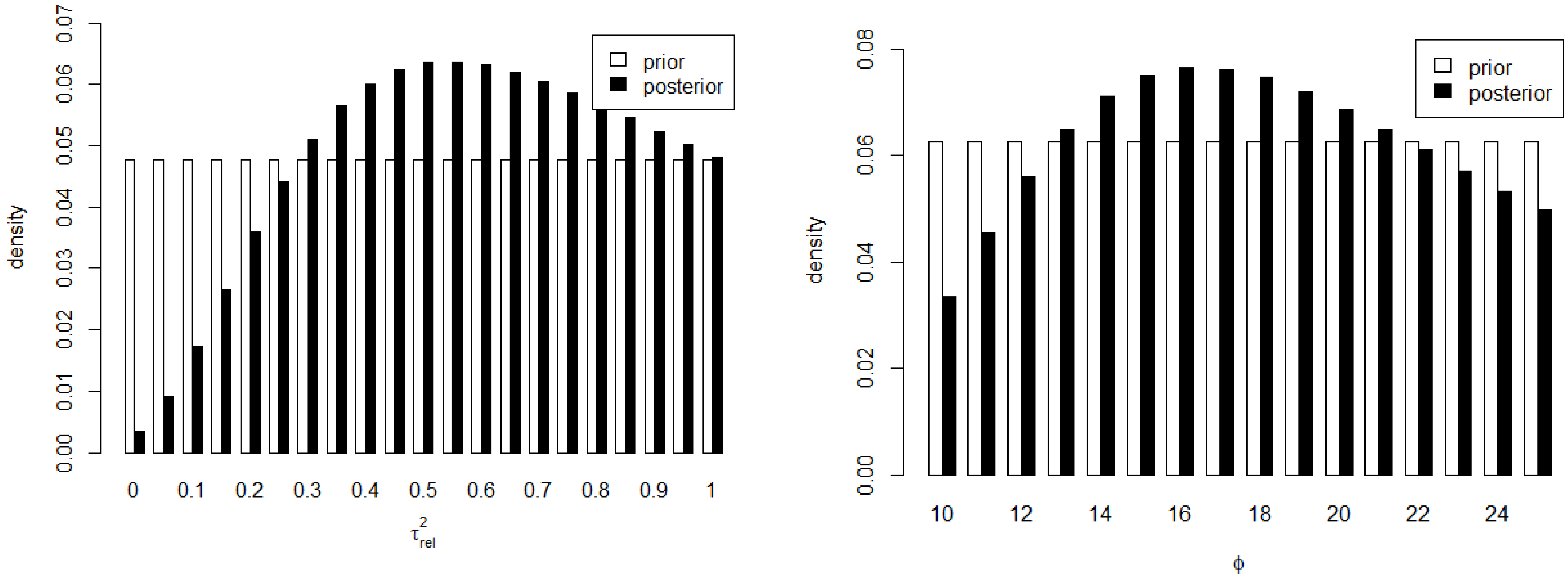




- **Beta Prior: Normal prior with mean zero and large variance (non-informative prior belief)**
- **Sigma-squared prior: Reciprocal prior (standard non-informative choice)**
- **Phi Prior: Uniform prior over discretized range**

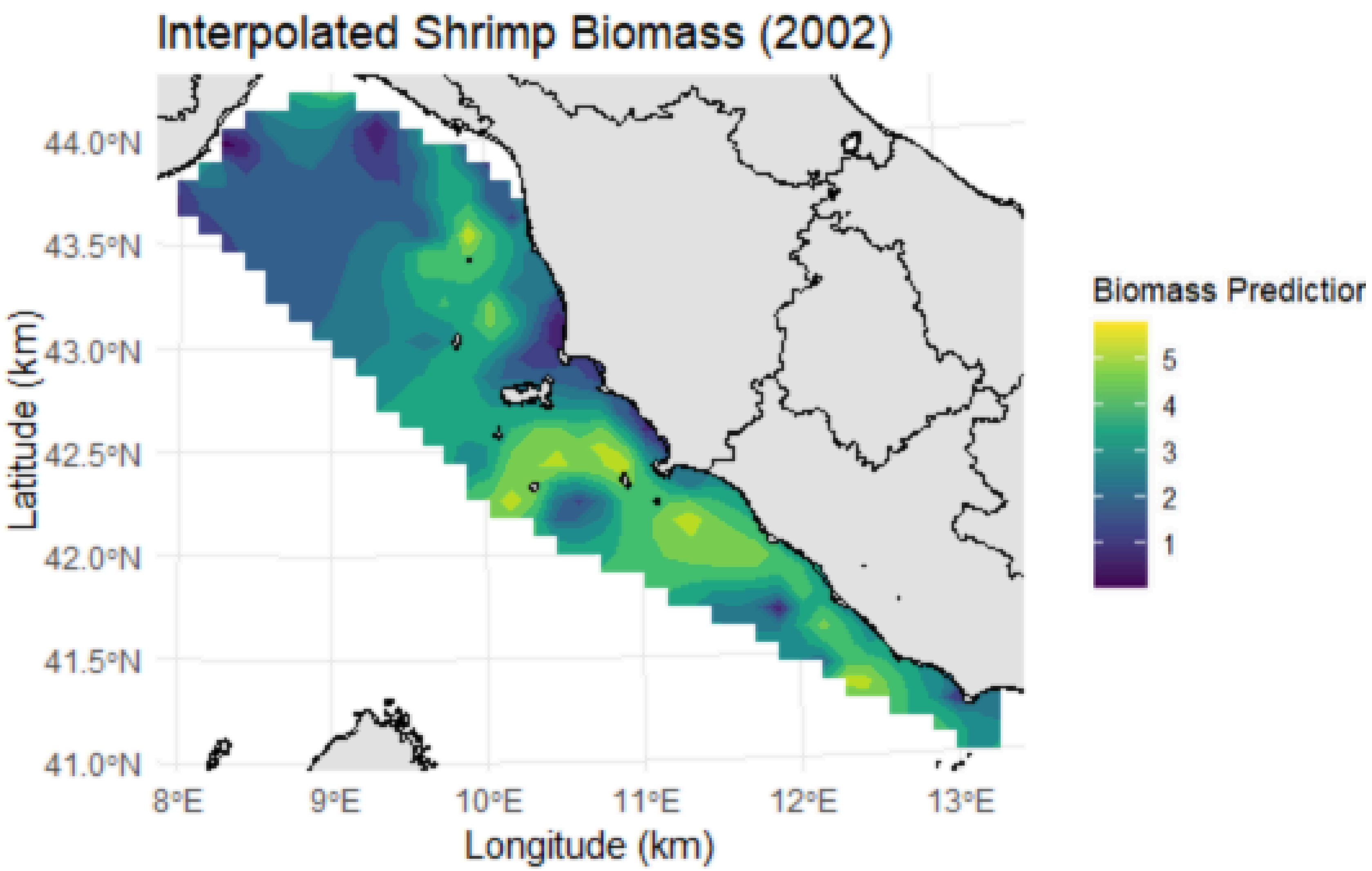
Posterior should have well-defined peaks, indicating that Bayesian inference has converged to a stable stimate

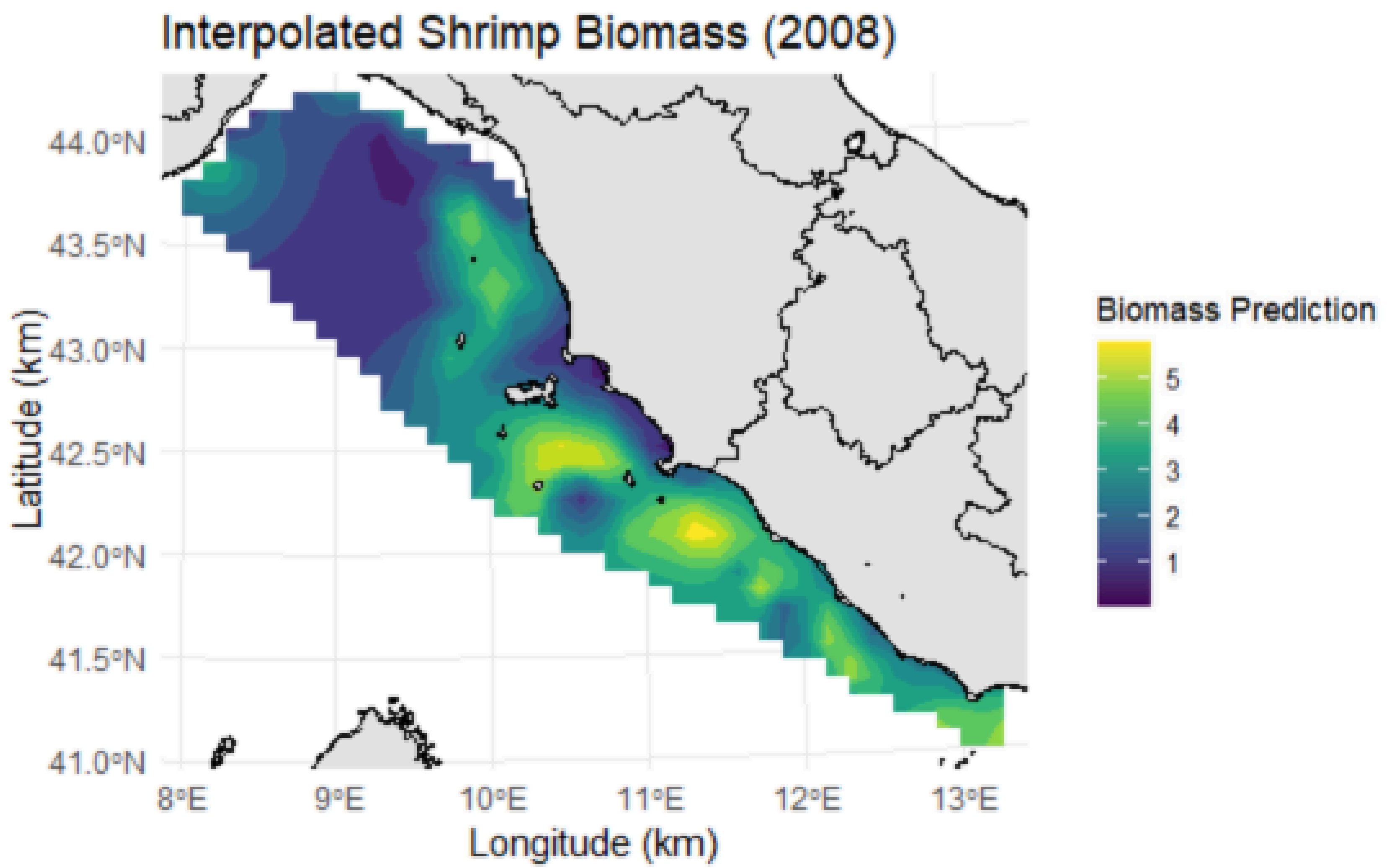
Prior and Posterior - 2008



- **Same prior choice as 2002**
- **Again, with the chosen covariates the posterior has clear peaks**

04 INTERPOLATION RESULTS



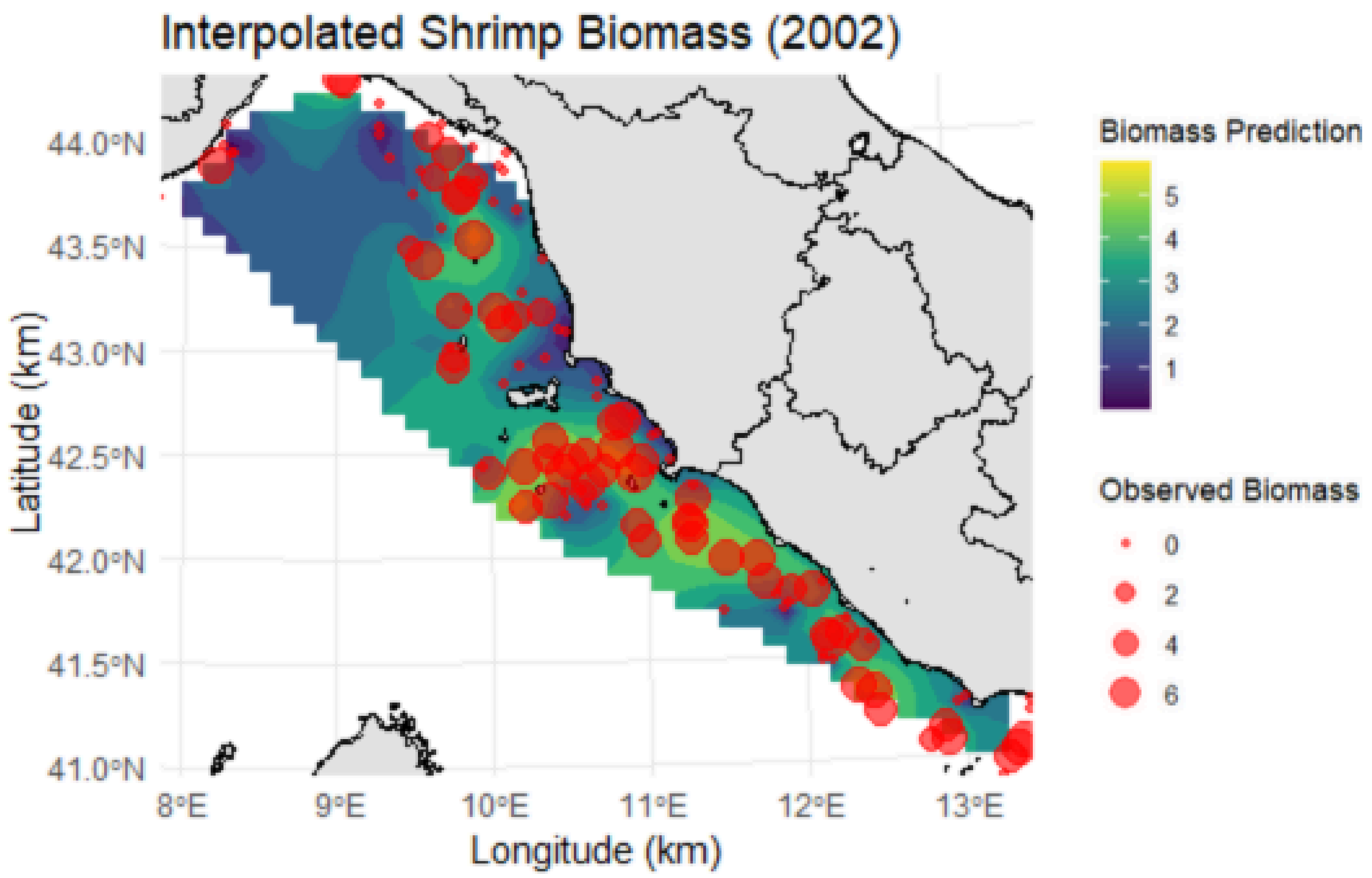


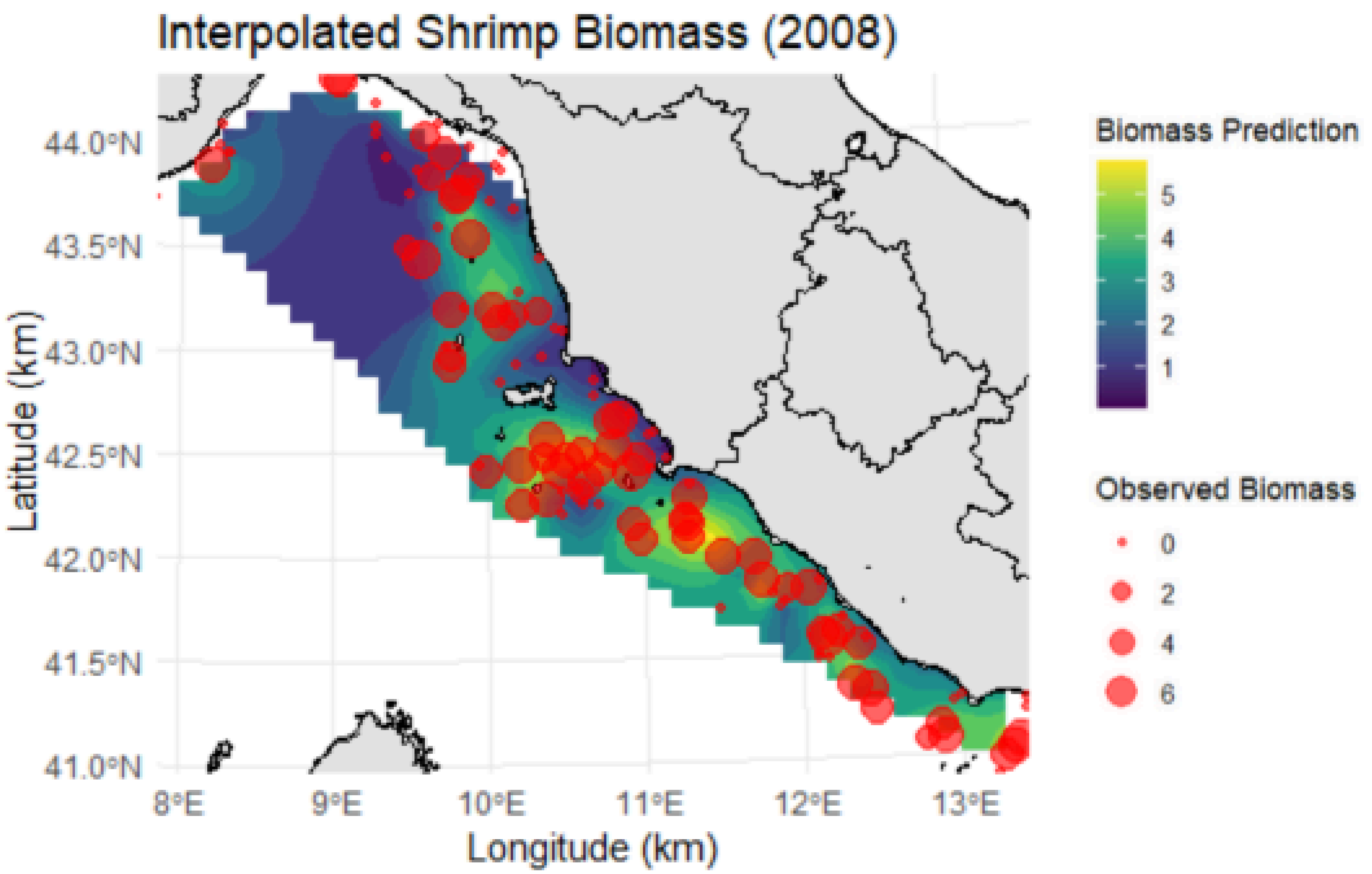
- **Distinct spatial patterns**
- **Medium and highest biomass concentrations along Tuscany and Lazio (shrimp hotspots)**
- **Lower biomass values near Genoa**

- **Hotspots in central and southeastern areas (larger clusters than 2002 in Island of Elba)**
- **Lower biomass in Liguria: steep bathymetric gradient**
- **Slight shift towards southeast**

SAPIENZA
UNIVERSITÀ DI ROMA

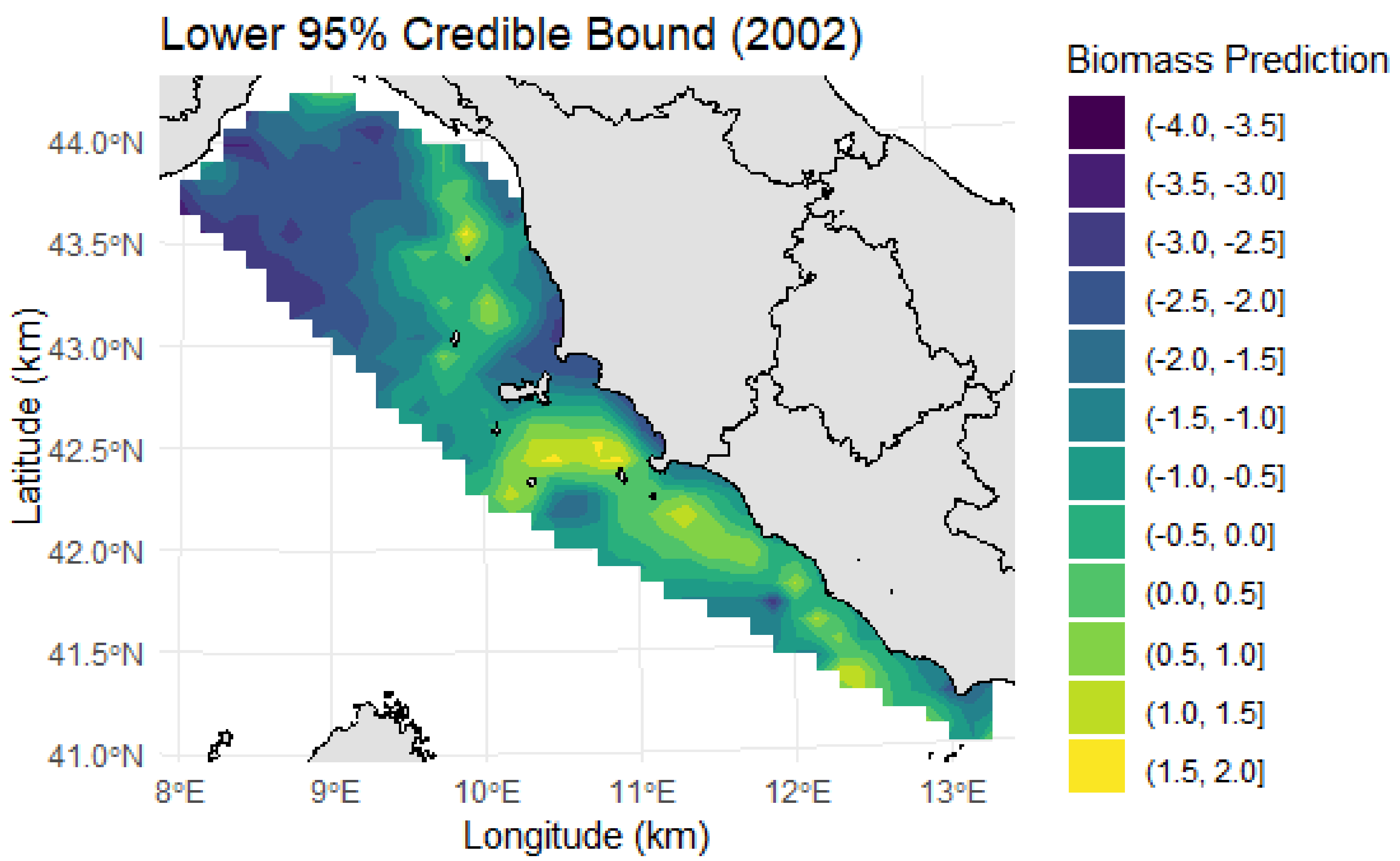
Interpolated Shrimp Biomass compared with the Observed one

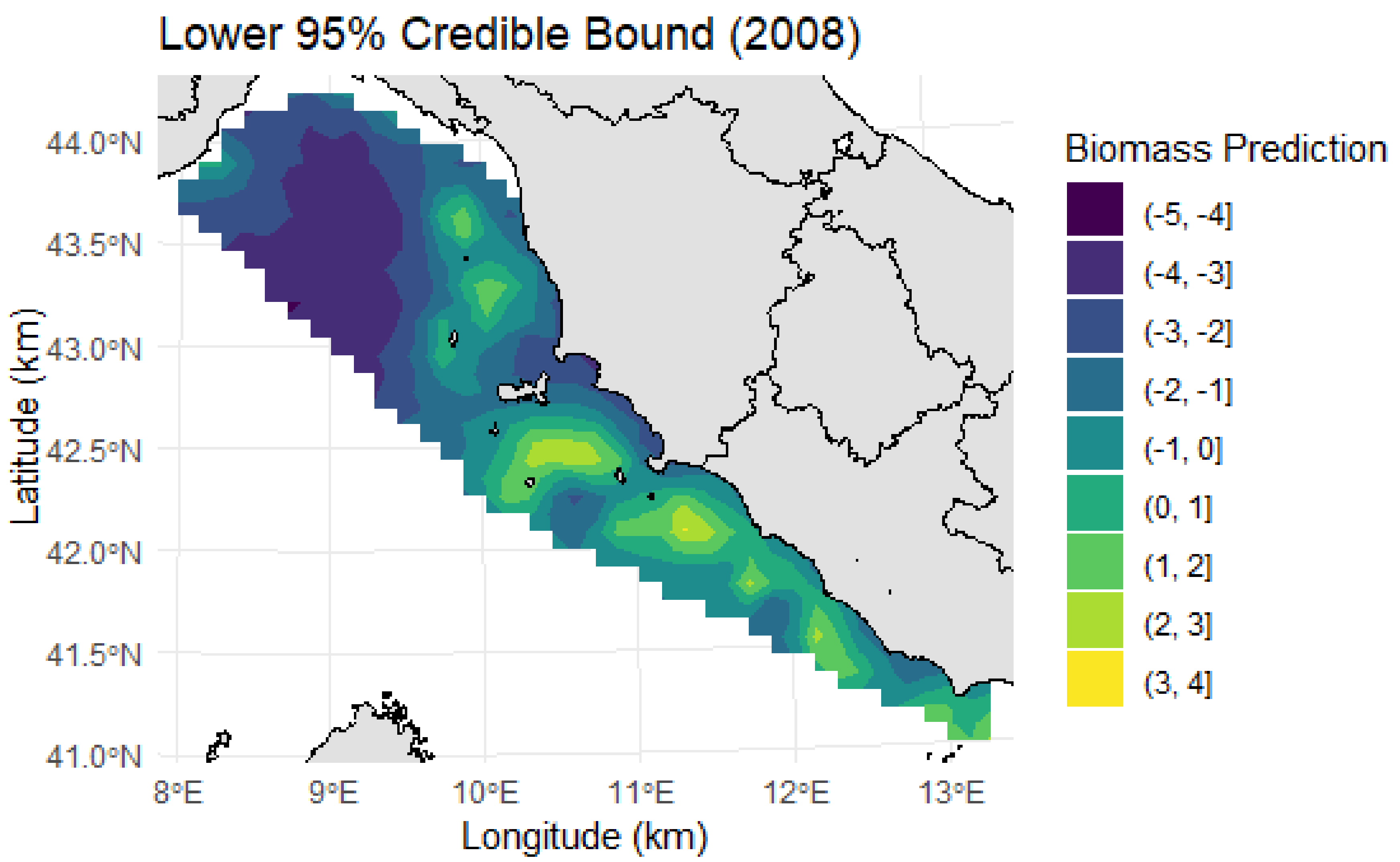




- **Distribution of red points → equivalence between higher observed and predicted biomass values**
- **Spatial variability in biomass has been reasonably captured by kriging**
- **Central and souther sections align well with the model's predictions, but few small discrepancies are still present**

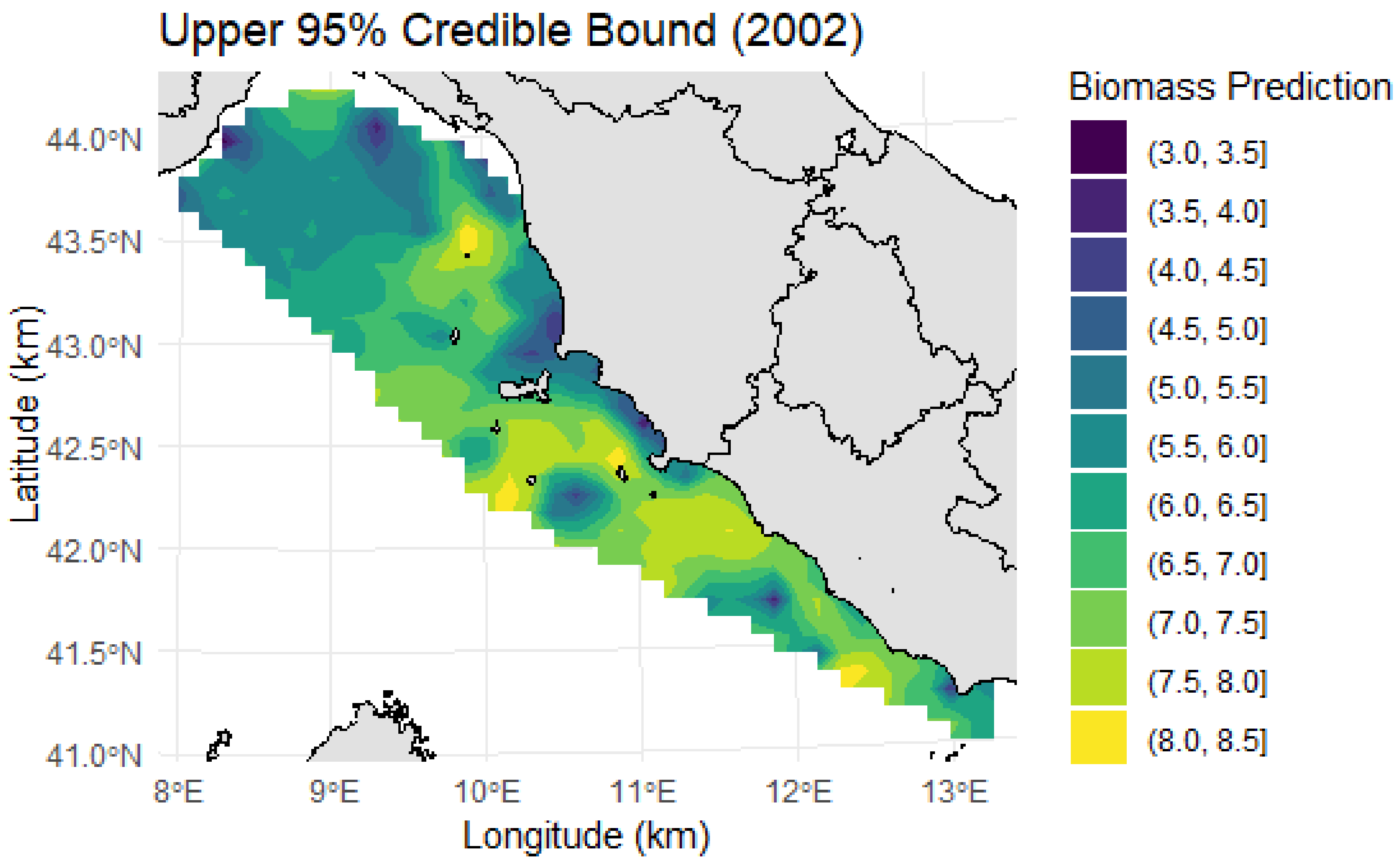
Credible Intervals - Lower Bounds

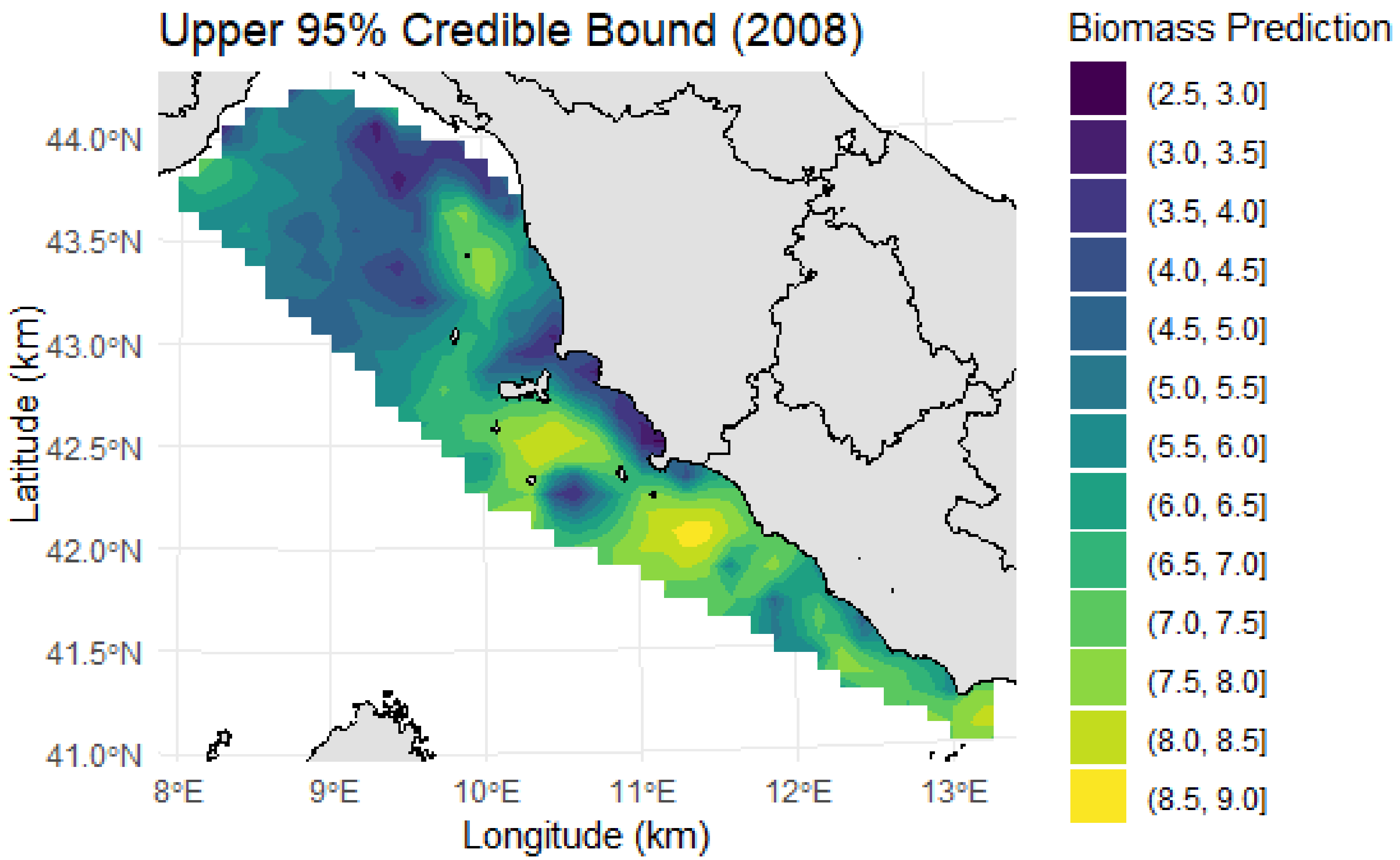




- **Lowest biomass values are more extreme in 2002**
- **Darker zones: less favorable conditions**
- **Lighter colors: areas where biomass remains high when accounting for variability**

Credible Intervals - Upper Bounds





- **Credible intervals are quite narrow, indicating low uncertainty in the model's estimates**
- **Gradual increase in the estimated vaues through the years**
- **Spatial consistency between the maps for the two years**

Sapienza Università di Roma

05 COMPARISON: MLE *VS.* BAYESIAN KRIGING

2002

- **Similar results** for Bayesian and MLE Kriging interpolation
- Biomass hotspots in **central** areas (Island of Elba and northern Lazio)
- Bayesian presents less spots of low biomass near Liguria

2008

- Results between the two approaches are quite **different**
- In ML Kriging highest concentrations **more spread** towards coasts of Lazio
- Bayesian approach: more concentrated in **central areas**
- Similar results with respect to northern regions

Comparison between approaches: RMSE and Interval Scores

MEAN RMSE	Frequentist	Bayesian
2002	2.2617	2.0759
2008	2.3812	2.1481

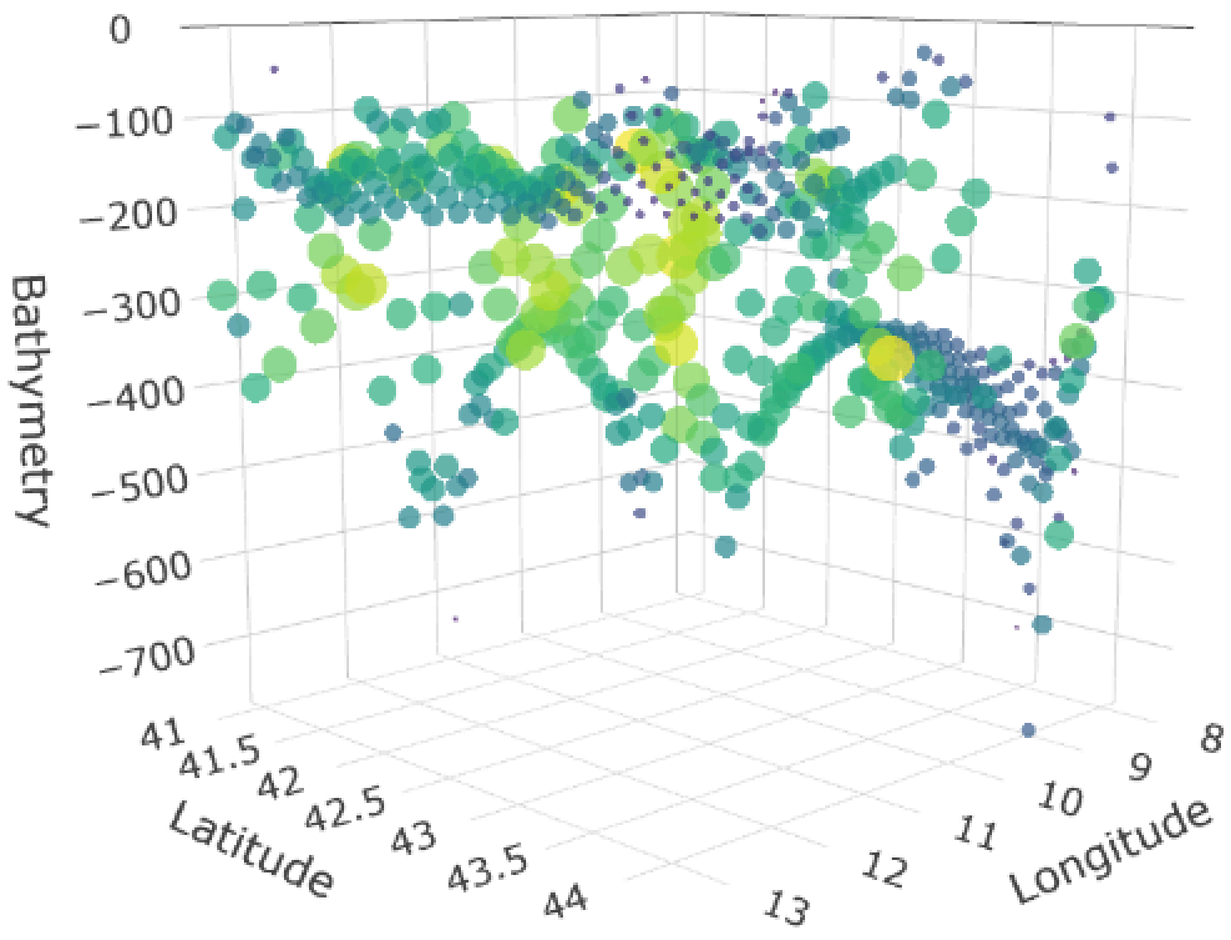
- **Better values for the Bayesian approach (slight differences)**
- **This may be due to Bayesian methods' ability to incorporate prior information, handle uncertainty, and procude more robust estimates**

MEAN INTERVAL SCORES	Frequentist	Bayesian
2002	9.731	8.384
2008	9.708	9.95

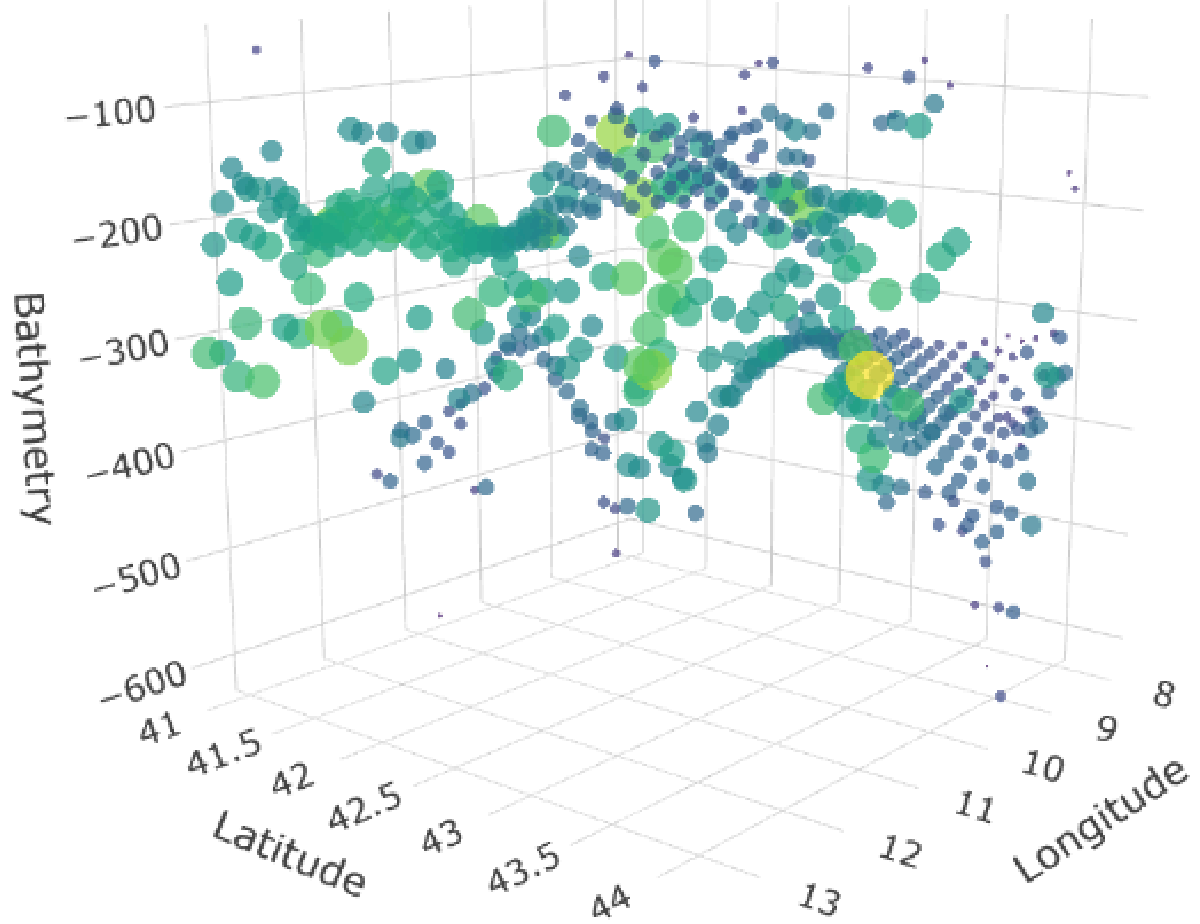
- **Bayesian model produces narrower and more accurate intervals in both years**
- **In 2008, the difference is not substantial**

SAPIENZA UNIVERSITÀ DI ROMA

3D Interactive Comparison of Kriging Interplation



BAYESIAN KRIGING



MLE KRIGING

Scan the QR code to interact!

SAPIENZA
UNIVERSITÀ DI ROMA

06 CONCLUSIONS & TAKE-HOME MESSAGE

Maximum Likelihood framework

- **Pros**: It accounts for **spatial relationships** through the **Covariance**, delivers direct estimates of uncertainty based on it
- **Cons**: **Computational complexity** due to estimation of model parameters

Bayesian approach

- **Advantages:** Incorporates **prior knowledge** and provides a reasonable description of **uncertainty** through the posterior
- **Drawbacks:** Higher **computational complexity** compared to other models, without prior knowledge defining a **good prior** coud be tricky

The Bayesian model is the most suitable for the shrimp dataset (2002-2008) based on RMSE and interval scores, but its advantage over the ML approach is not significant.

SAPIENZA
UNIVERSITÀ DI ROMA

THANK YOU FOR YOUR ATTENTION !
ANY QUESTIONS?

You can ask them only if you recognize the reference!