Modelling fish and macroinvertebrates off Peru

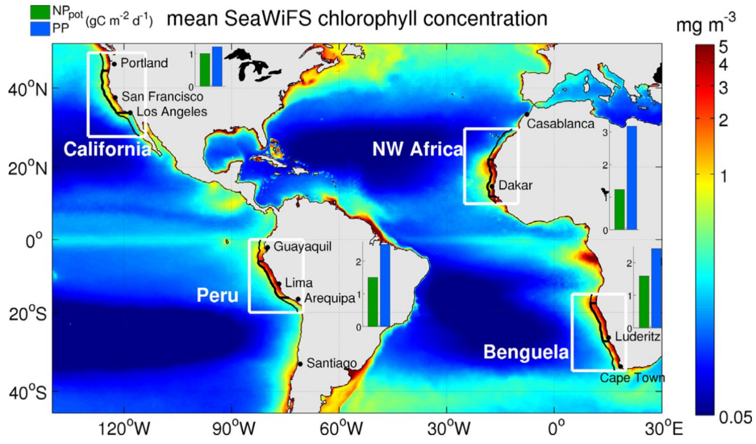
Benjamin Müller and Mariana Hill

Outline

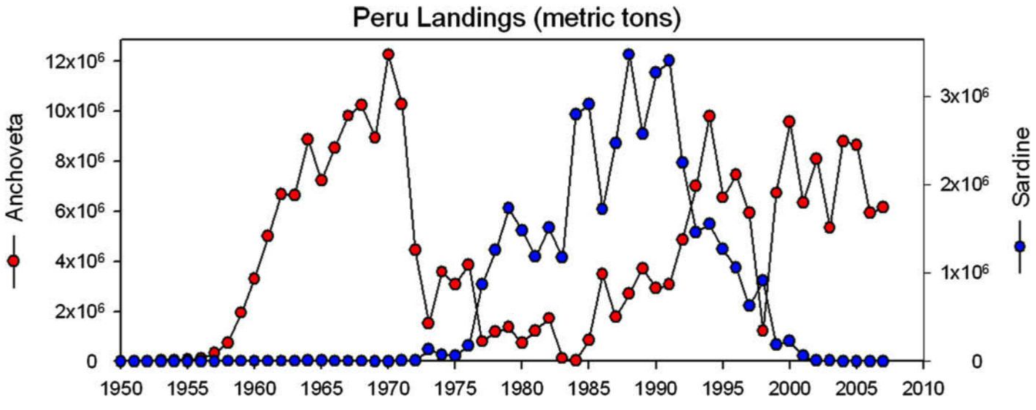
- Introduction: Peruvian fisheries
- Introduction: Habitat modelling
- First dataset (abundance)
- heatmaps
- seasonal movements
- spatial panel model
- INLA (first steps)
- Random forest (absolute abundance)
- Data distribution
- Random forest (presence absence)
- Outlook
- Second dataset (landings)
- Time-series prediction with RNN
- Outlook

Introduction: Peruvian fisheries

- Most productive Eastern Boundary Upwelling System
- Small pelagic fish
- Constrained by a shallow oxygen minimum zone
- Largest single species fishery (anchovy)
- 10% of global catches
- Interannual fluctuations and fisheries collapses
- Environmental drivers still poorly understood



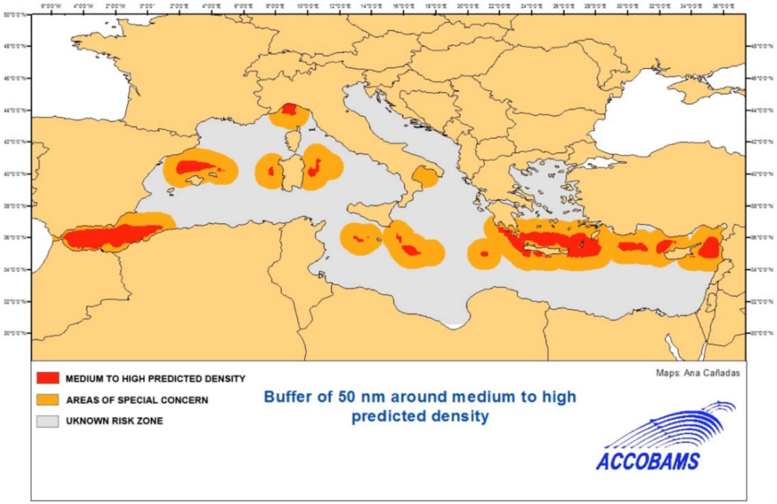
Messié & Chávez (2015)



Chávez et al. (2008)

Introduction: Habitat modelling

- Potential distribution of marine organisms
- Common environmental variables: distance from the coast, water depth, primary production, etc.
- Statistical methods (GAM and linear models)
- Predictions about marine animals where more complex models are not available
- Usually used as forcing for dynamic models
- How does species distribution change over time?



Habitat suitability for Cuvier’s Beaked Whale (UNEP 2015)

Modeling and Prediction of Habitat Suitability for *Ferula gummosa* Medicinal Plant in a Mountainous Area

Majid Mohammady, Hamid Reza Pourghasemi, Saleh Yousefi, Emran Dastres, Mohsen Edalat, Soheila Pouyan & Saeedeh Eskandari

Natural Resources Research **30**, 4861–4884 (2021) | Cite this article

“…the RF model was the best one for assessing *Ferula gummosa* habitat suitability.”

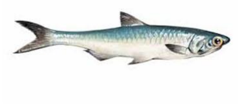
Cruises off Peru

- Peruvian EEZ
- Area of anchovy occurrence
- Acoustic indices are used to predict total biomass over the whole region
- 1985 to 2012

- Species:
  - Anchovy, sardine, jack mackerel, chub mackerel, hake, demersal fish, zooplankton, lightfish, lantern fish, Humboldt squid, catfish, white anchovy
- Environmental variables:
  - Distance from the coast, distance from the shelf, chlorophyll, surface oxygen, water depth, depth of the oxycline, sea surface temperature, sea surface salinity
- Challenges:
  - Spatial and temporal dataset
  - Sampling locations change every year

Which animals again?

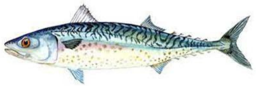
Anchovy



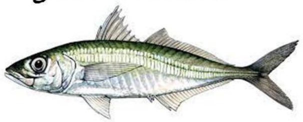
Sardine

Vinciguerria

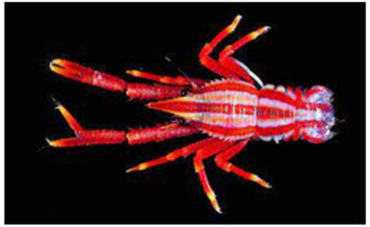
Chub Mackerel

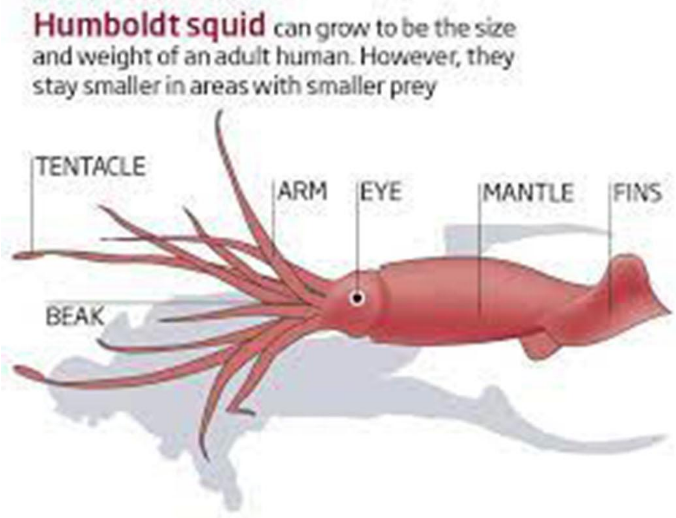


Jack Mackerel



Red Squat Lobster

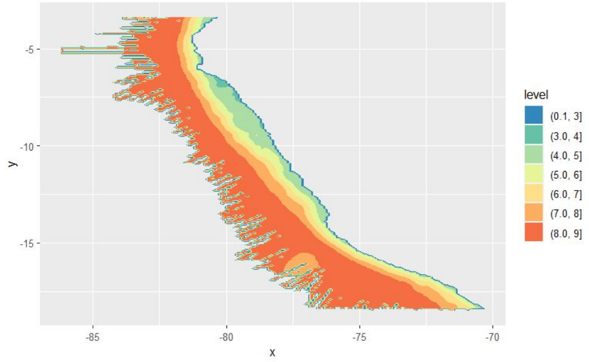




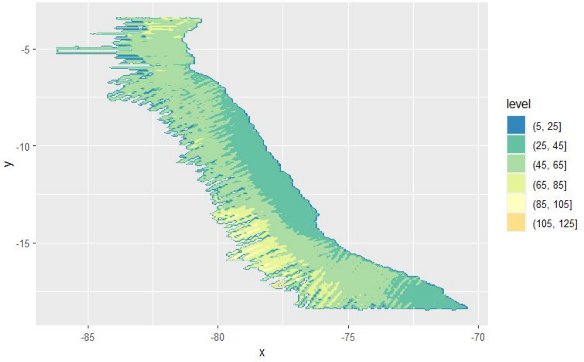
smaller squids

Heatmaps - explaining variables

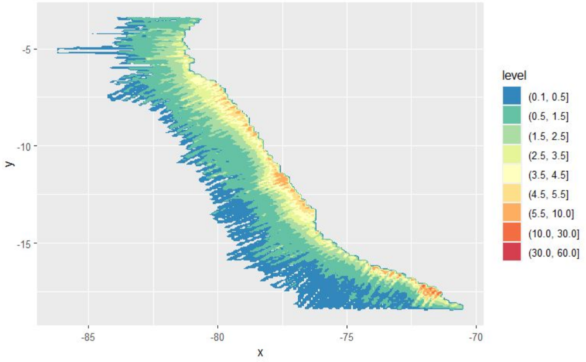
depth



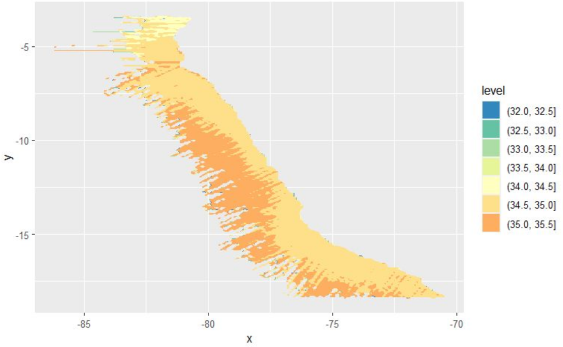
oxygen



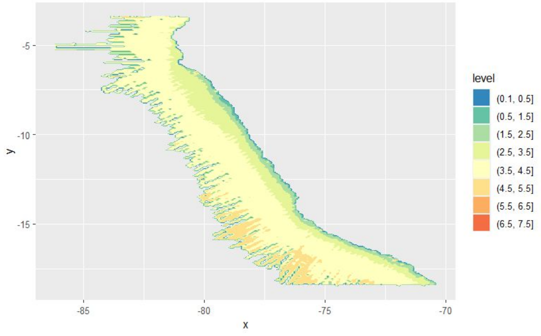
chlorophyll



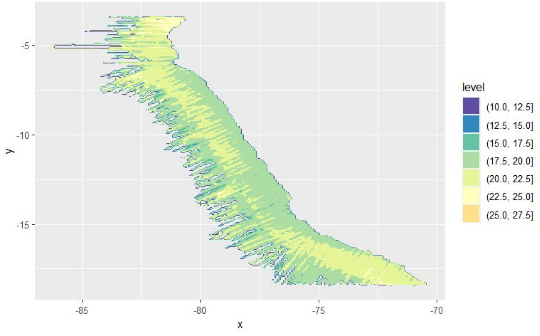
sea surface salinity



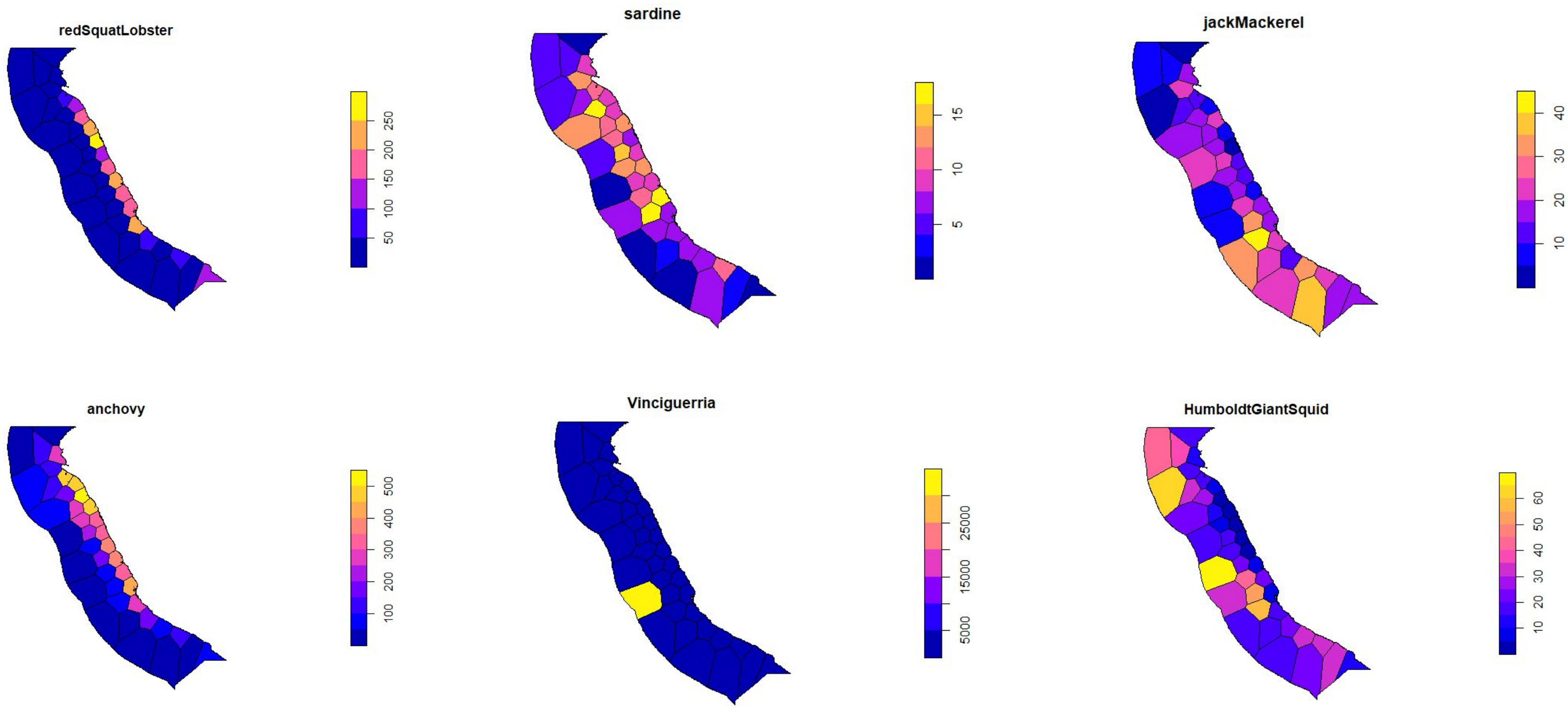
depth oxycline



sea surface temperature



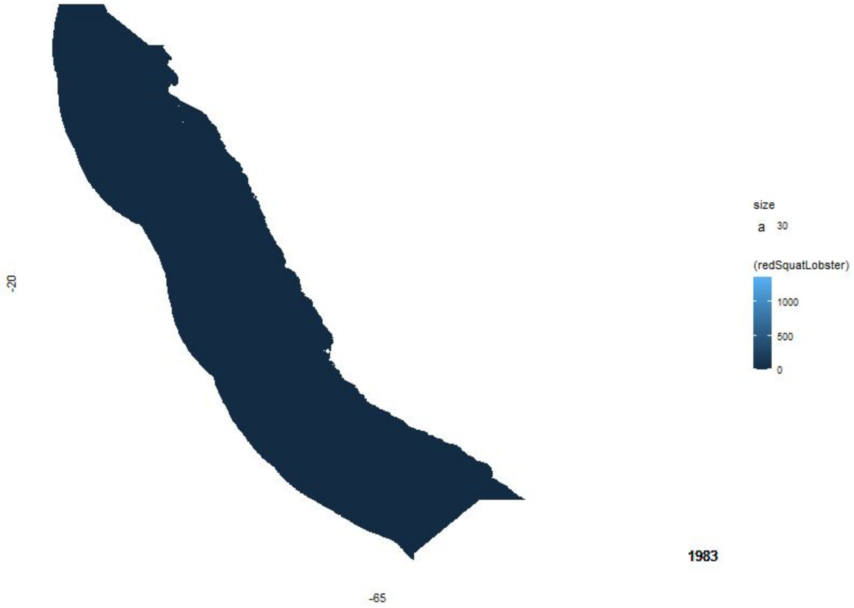
Chloropeth maps - marine animals



Chloropeth Maps - animation over years

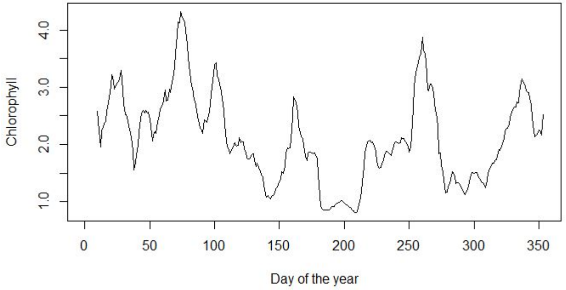
anchovy over years

red squat lobster over years

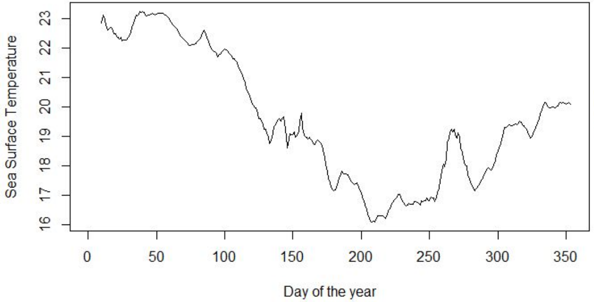


Seasonal Movements - Explaining Variables

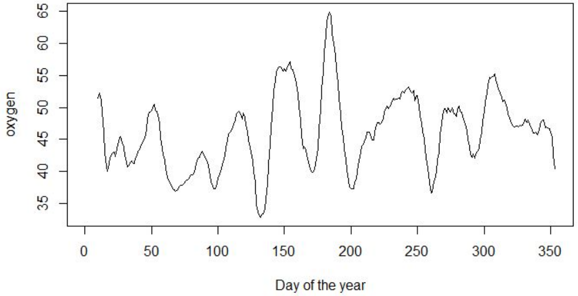
Chlorophyll



Sea Surface Temperature

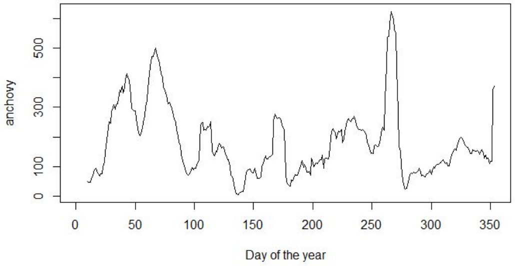


Oxygen

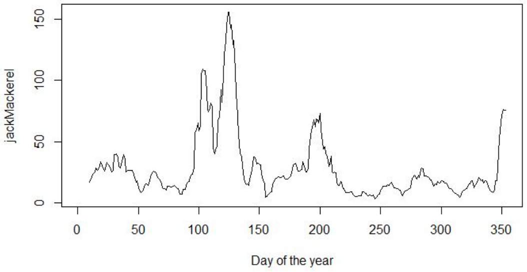


Seasonal Movements - Marine Animals

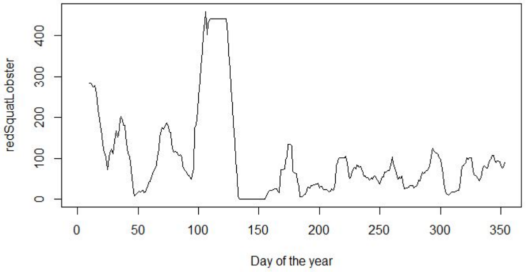
anchovy



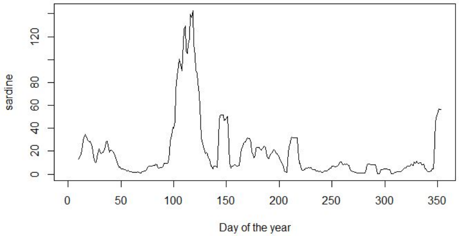
jack mackerel



red squat lobster

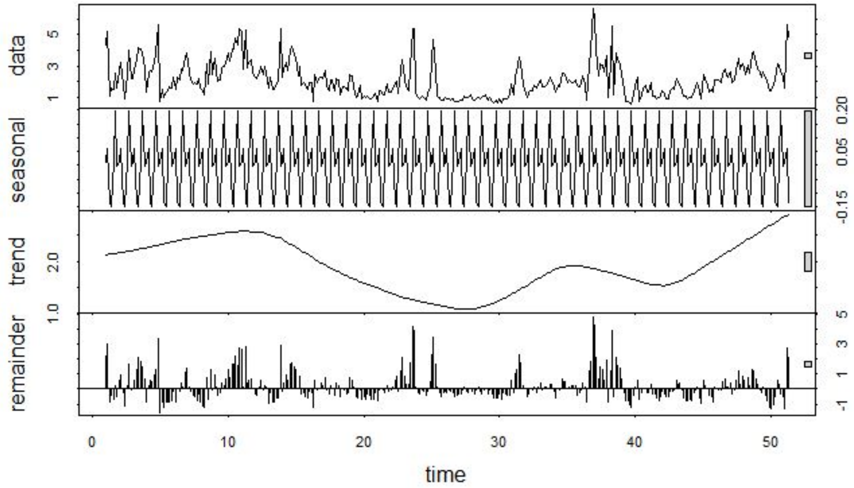


sardine



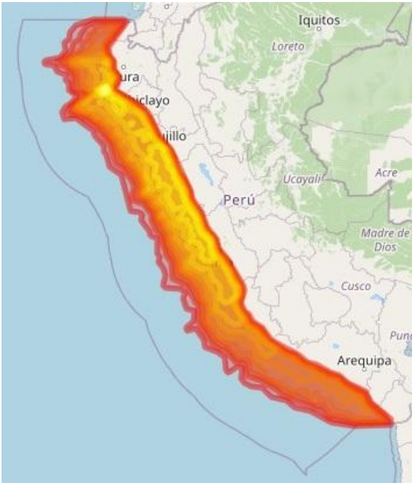
Seasonal Movements - Chlorophyll Decomposition

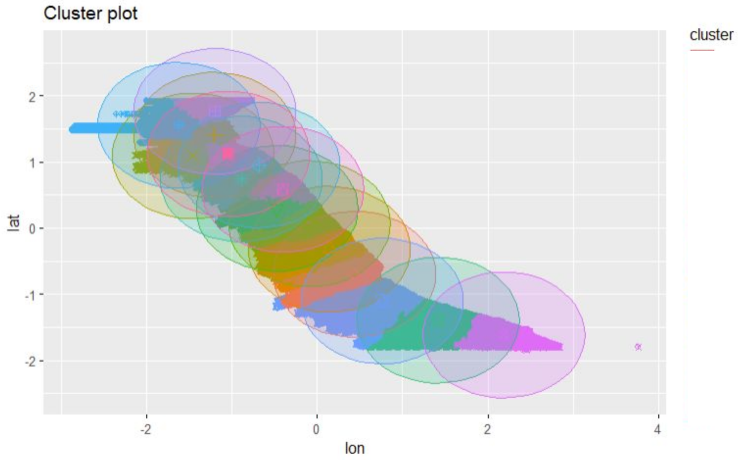
- winter is in the middle of the year in Peru
- less chlorophyll in winter



aggregation per region

create a tesselation with k means clustering and voronoi tesselation





only longitude, latitude variables used for clustering

clustering shown with k =15, voronoi with k = 40, optimal k?

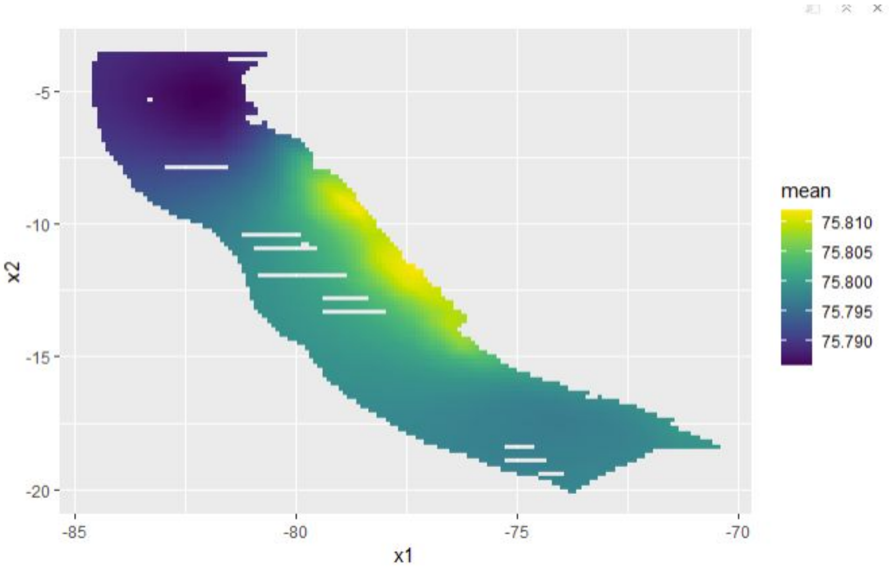
Spatial Panel Regression Results

- using spatial regression alone increase R^2 from 0.1 to 0.8
- panel data aggregated into 80 regions and 30 years
- amount of red squat lobsters in dependence of

```
Coefficients:
                        Estimate Std. Error t-value  Pr(>|t|)    
(Intercept)           31.6835095  6.7246085  4.7116 2.458e-06 ***
distToCoast           -0.0275152  0.0852918 -0.3226  0.746998    
depth                 -0.0066095  0.0033272 -1.9865  0.046976 *  
oxygen                -0.5102885  0.3124521 -1.6332  0.102432    
chlorophyll           -0.8943744  1.0880353 -0.8220  0.411072    
seaSurfaceTemperature -4.1159285  1.5729290 -2.6167  0.008878 ** 
seaSurfaceSalinity     4.0199545  0.9942104  4.0434 5.269e-05 ***
depthOxycline         -0.3158990  0.1794258 -1.7606  0.078304 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

INLA - first steps - results

result is an interpolation map (kriging)

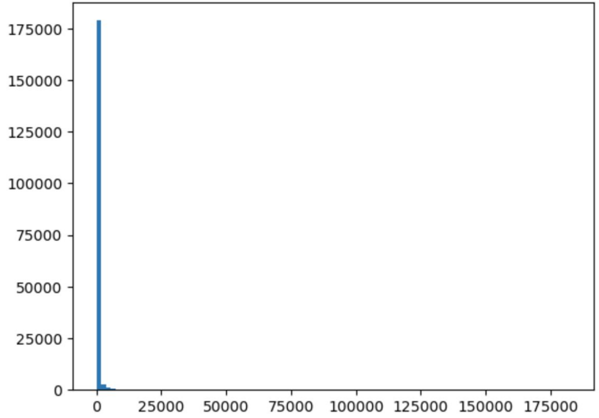


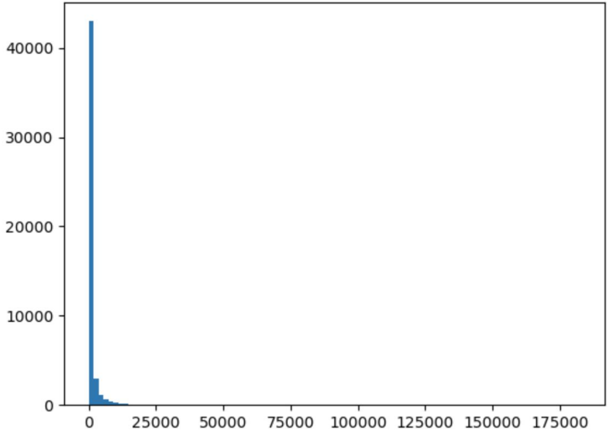
Data distribution

zeros

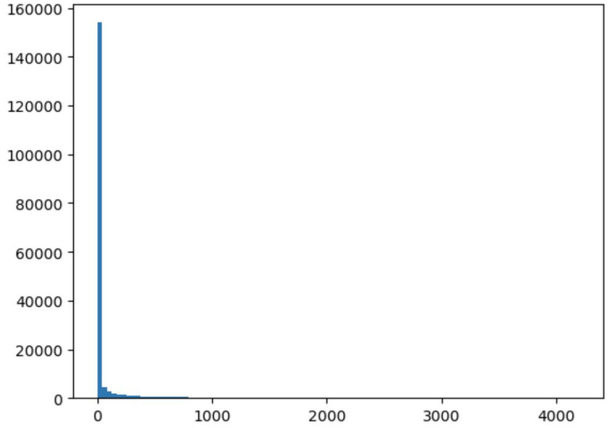
no zeros

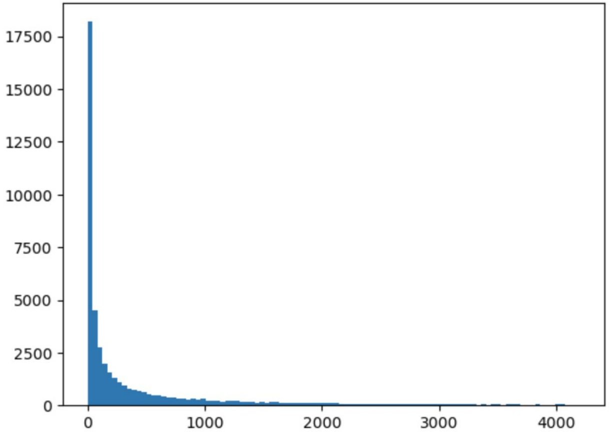
outliers





no outliers





Random forest - abundance

Without outliers, anchovy predicted abundance:

100 estimators

R2 train: 0.848

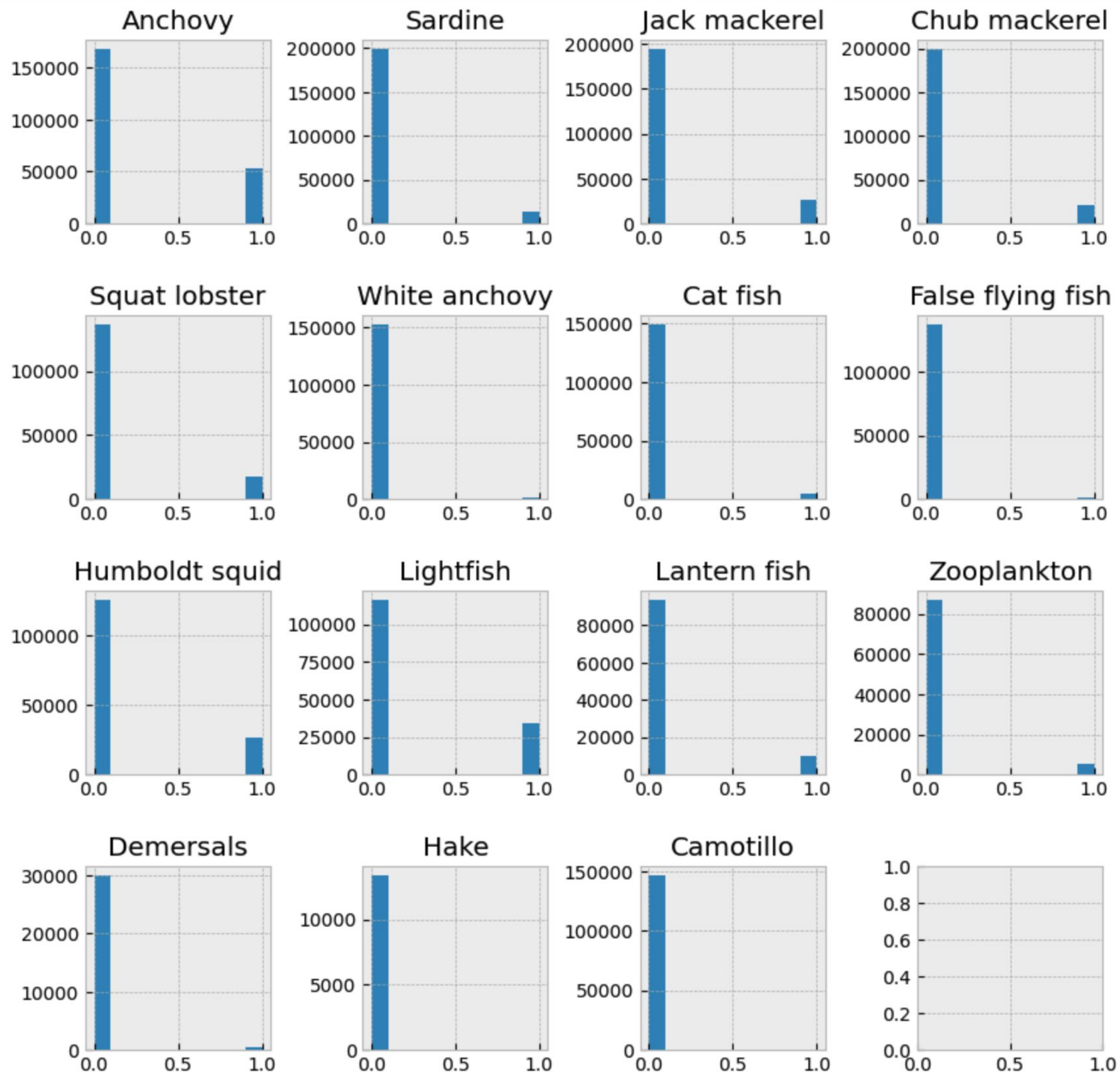
R2 test: 0.37

1000 estimators

R2 train: 0.85

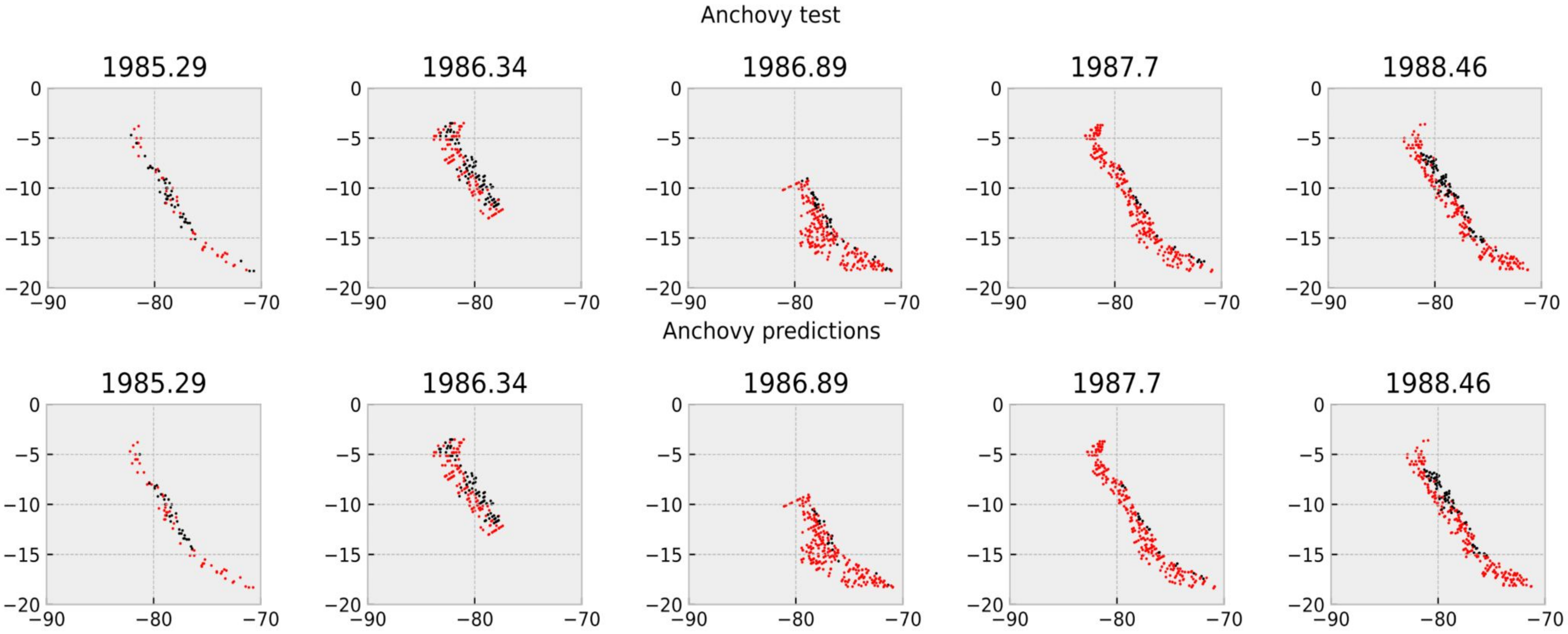
R2 test: 0.37 :(

Data distribution: presence - absence



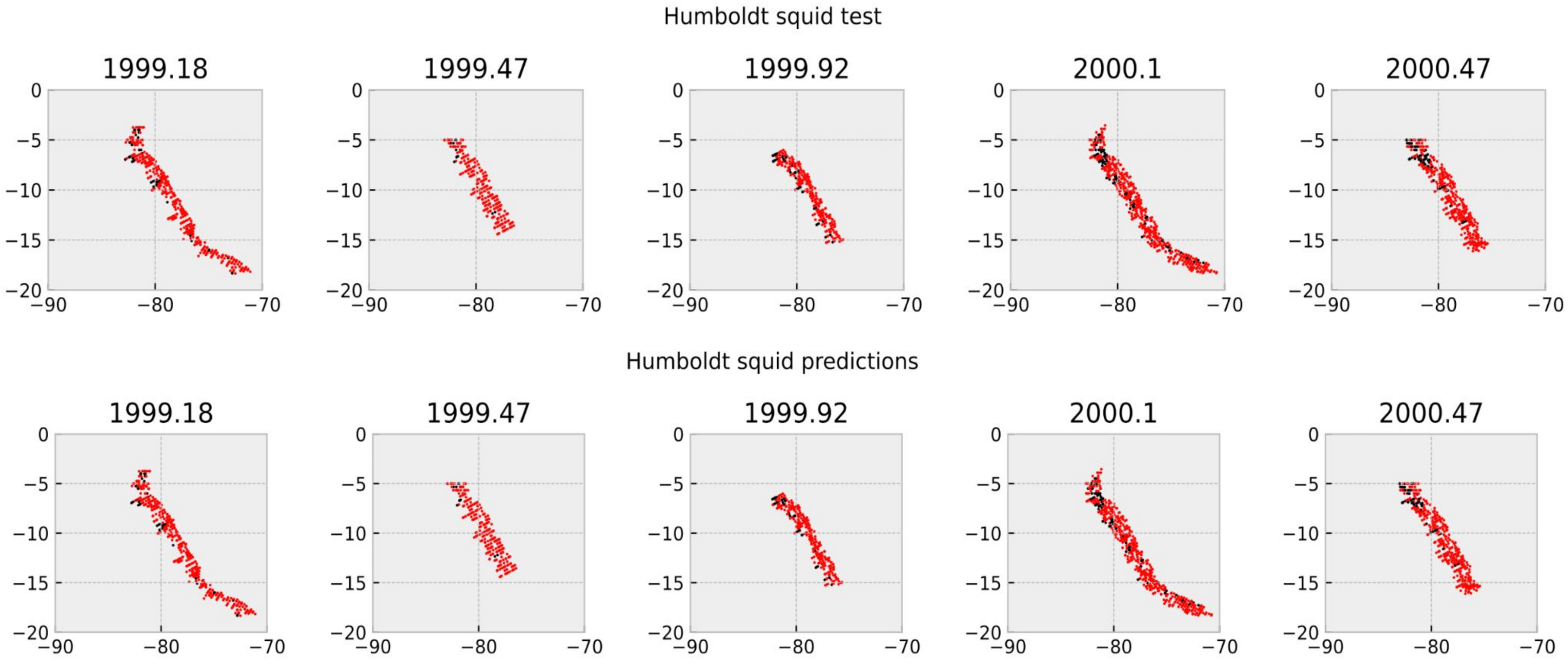
Random forest - presence absence

n estimators: 100
test accuracy: 0.96
red: absence (0)
black: presence (1)



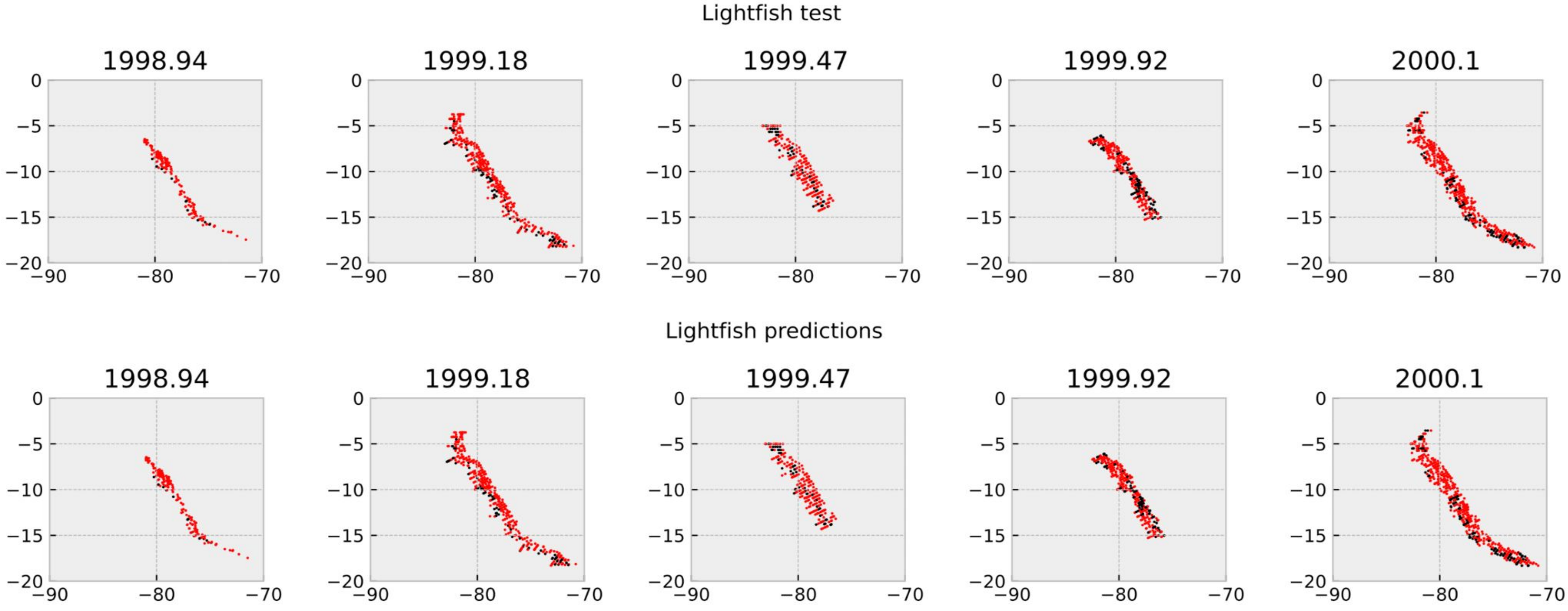
Random forest - presence absence

n estimators: 100
test accuracy: 0.97
red: absence (0)
black: presence (1)



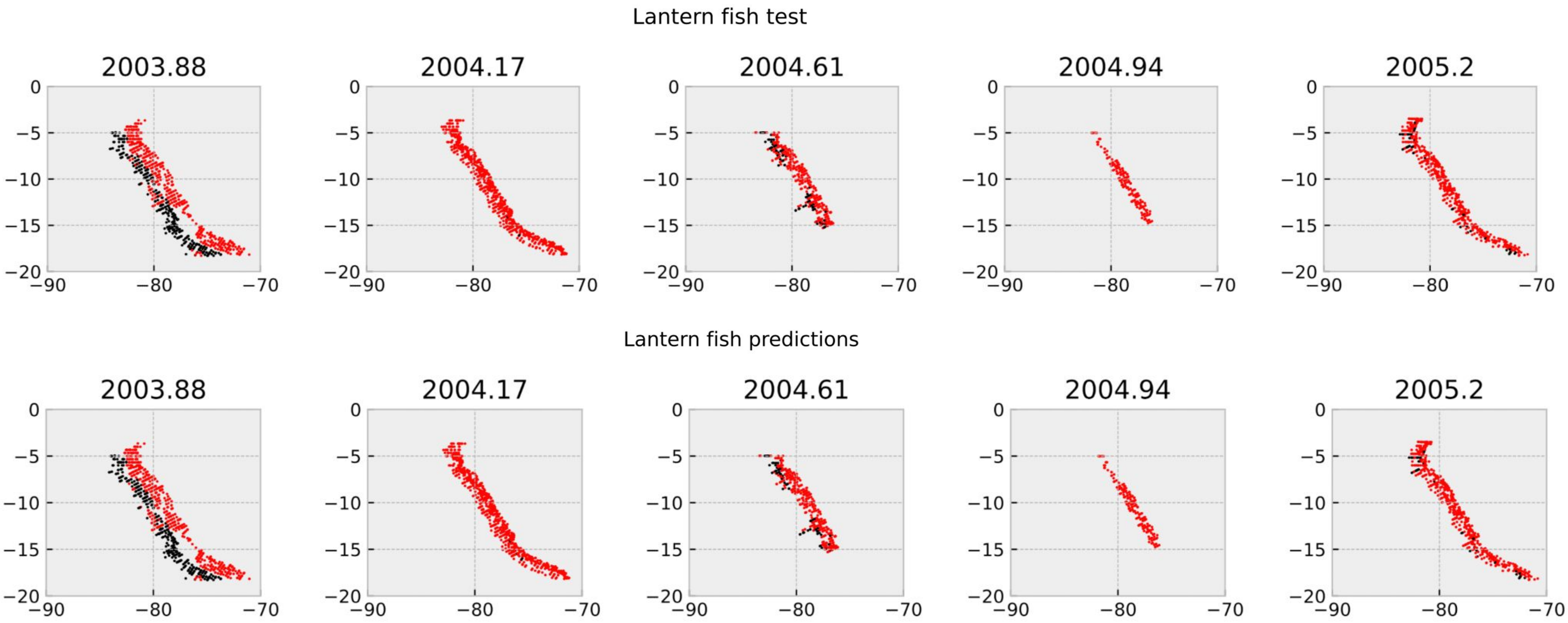
Random forest - presence absence

n estimators: 100
test accuracy: 0.97
red: absence (0)
black: presence (1)



Random forest - presence absence

n estimators: 100
test accuracy: 0.99
red: absence (0)
black: presence (1)



Random forest - presence absence

n estimators: 100
test accuracy: 0.98
red: absence (0)
black: presence (1)

Can we actually use a RF to predict the timeseries?

- Train: 1985 to 1997
- Test: 1998 to 2008
- 100 and 1000 estimators
- Accuracy: 0.75

- Train: 1985 to 2003
- Test: 2004 to 2008
- 100 and 1000 estimators
- Accuracy: 0.75
- Precision: 0.84
- Recall: 0.58

Conclusion and outlook

- Abundance prediction does not work - not real link between observed abundance and environmental drivers
- Data is highly imbalanced
- Presence absence works - this is the standard method
- But test and train data are too similar
- The RF can’t predict the timeseries unless the data is splitted randomly

Next steps:

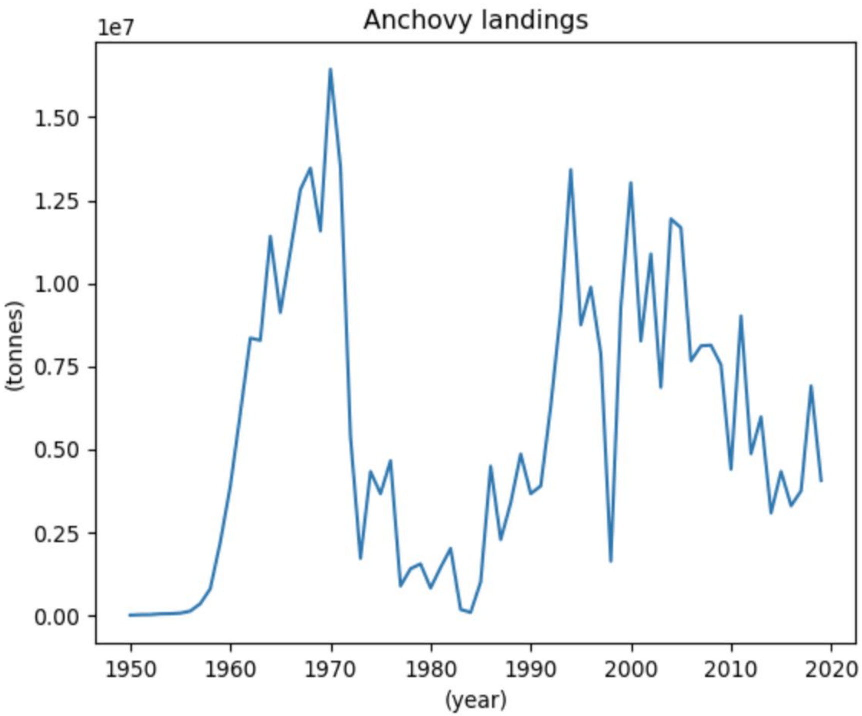
- Get probability of occurrence instead of just presence-absence
- Apply the model to predict potential fish distribution from scenarios
- Refine the model - deal with very imbalanced data?
- Try to predict abundance?

Fish landings timeseries

Yearly timeseries of landings from different species in the world downloaded from:

Sea Around Us

https://www.seaaroundus.org/data/#/eez/604?chart=catch-chart&dimension=taxon&measure=tonnage&limit=10

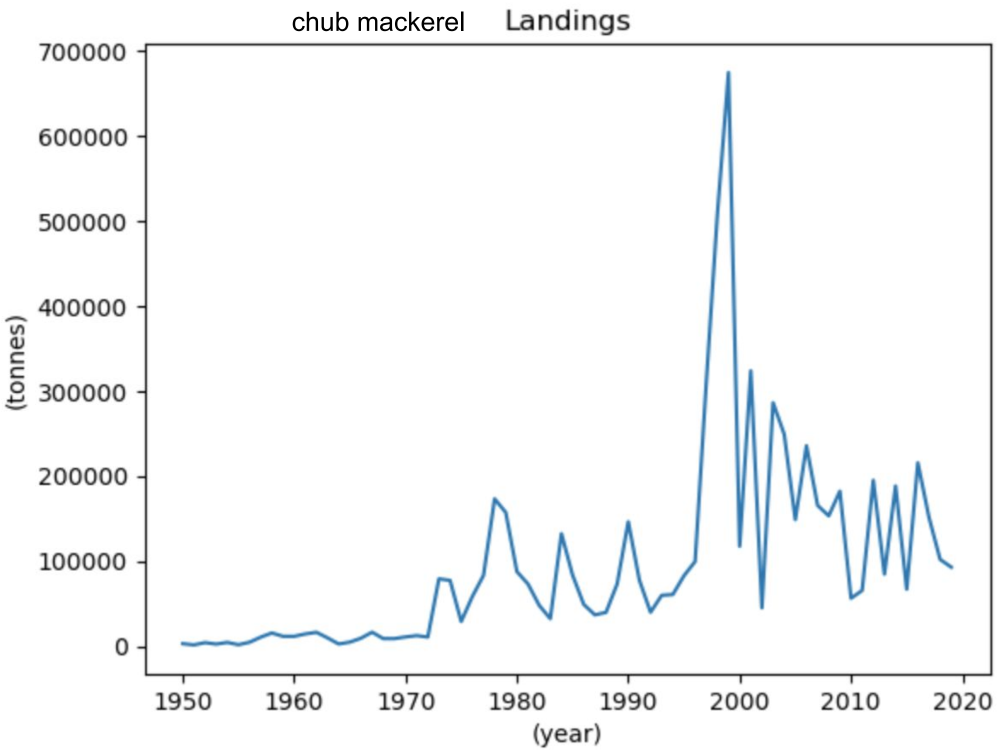


Fish landings timeseries

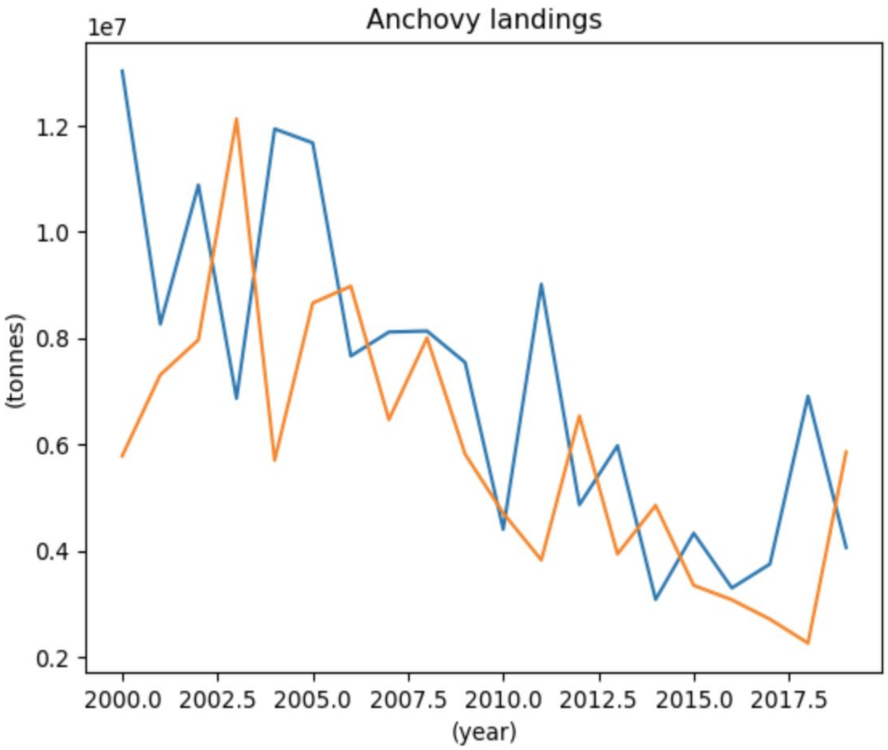
Yearly timeseries of landings from different species in the world downloaded from:

Sea Around Us

https://www.seaaroundus.org/data/#/eez/604?chart=catch-chart&dimension=taxon&measure=tonnage&limit=10



Predicting anchovy landings with a RNN

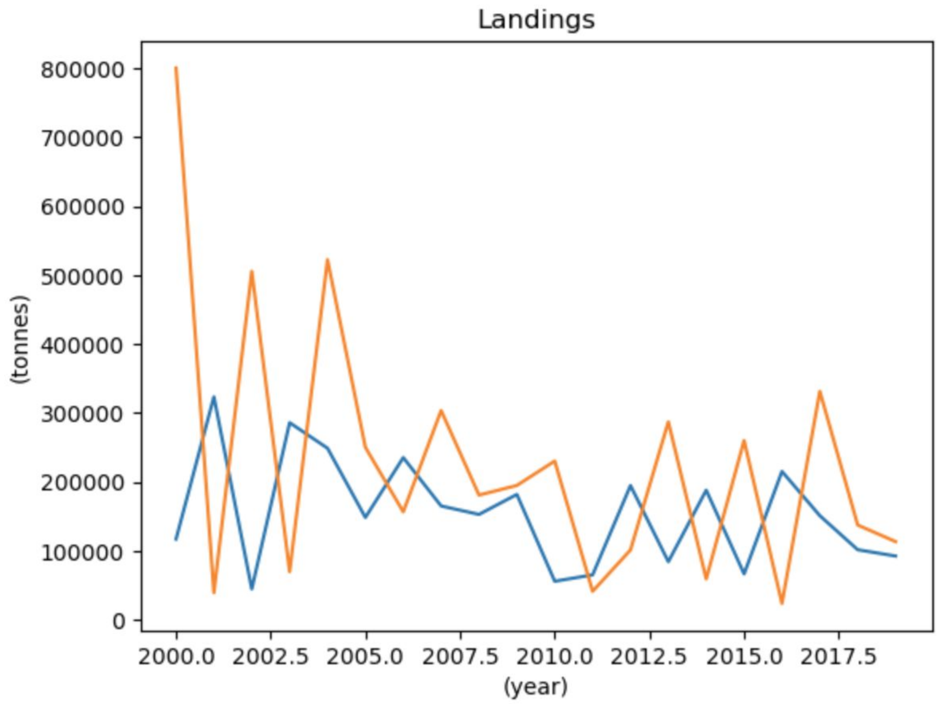


window size: 5
train: 1950 - 1999
test: 2000 - 2019

1000 epochs
learning rate: 4e-4

y_test
prediction

Predicting chub mackerel landings with a RNN



window size: 5
train: 1950 - 1999
test: 2000 - 2019

1000 epochs
learning rate: 4e-4

y_test
prediction

Conclusion and outlook

- Mean and trend are simulated but there is a mismatch between the predicted and test series

Next steps:

- Try other models
- Apply to more species
- Use monthly data

Any questions?