Machine learning approach for predicting zooplankton abundance in the Baltic Sea

Dušan Sovilj

Aalto University School of Science

March 23, 2011

Hello, my name is...

- Dušan Sovilj
- ...and I have a problem with zooplankton
- Aalto University School of Science, Department of Information and Computer Science
- Research focus on machine learning
  - time series prediction
  - variable selection
- `http://users.ics.tkk.fi/dusans/`

My focus under AMBER

- Zooplankton prediction (acartia, temora, pseudocalanus,...)
- ... taken as time series problem
- Using machine learning methods and models for the job (neural networks the most famous)
- Idea in a nutshell:

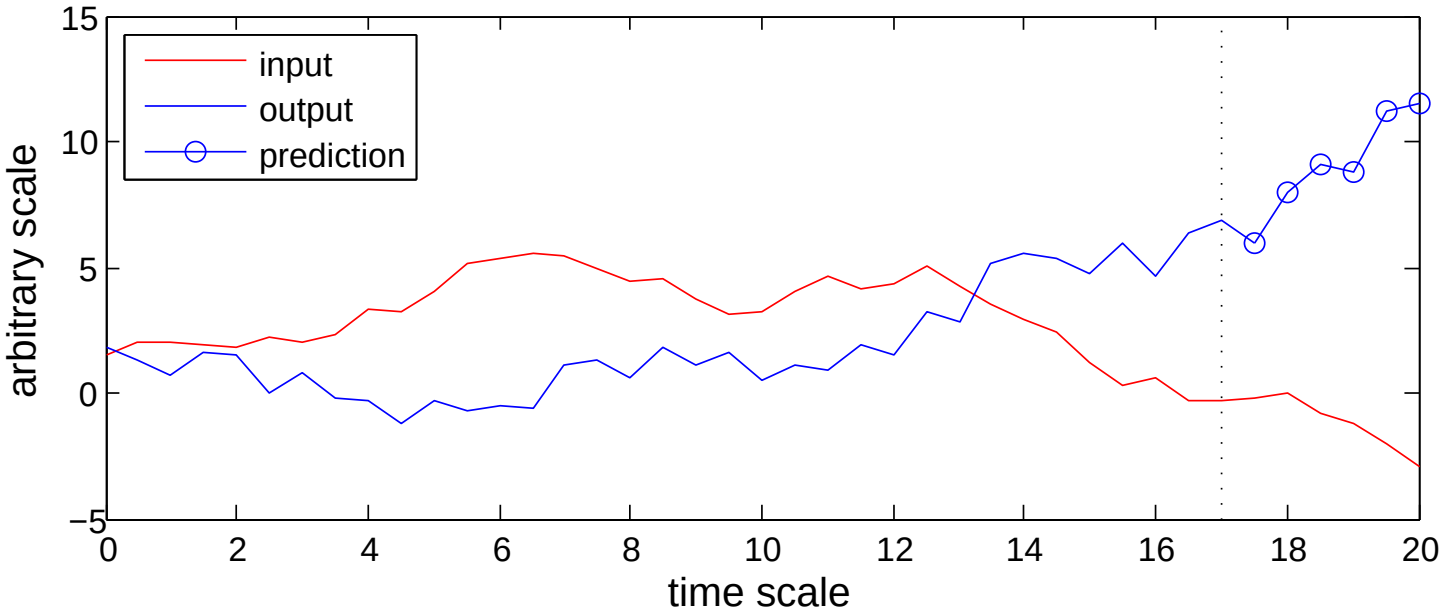
  Given some measurements / samples / datapoints in (input,output) format, predict the value of the output for some *new* input samples
- In the zooplankton prediction, the input is the climate index / indices (AO / NAO / BSI), while the output is the zooplankton abundance
- Focus on spring values of the species
  the "easier" of the tasks

My focus under AMBER

- Zooplankton prediction (acartia, temora, pseudocalanus,...)
- ... taken as time series problem
- Using machine learning methods and models for the job (neural networks the most famous)
- Idea in a nutshell:

  Given some measurements / samples / datapoints in (input,output) format, predict the value of the output for some *new* input samples
- In the zooplankton prediction, the input is the climate index / indices (AO / NAO / BSI), while the output is the zooplankton abundance
- Focus on spring values of the species
  the "easier" of the tasks

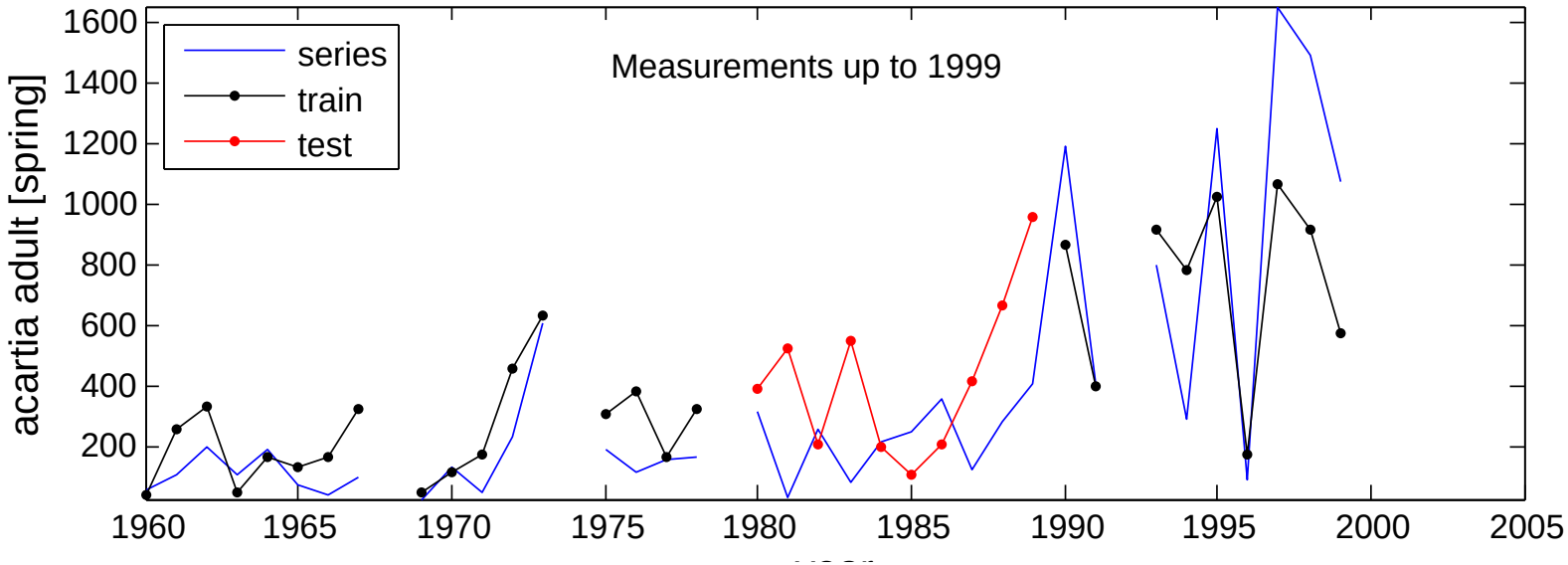
Visually

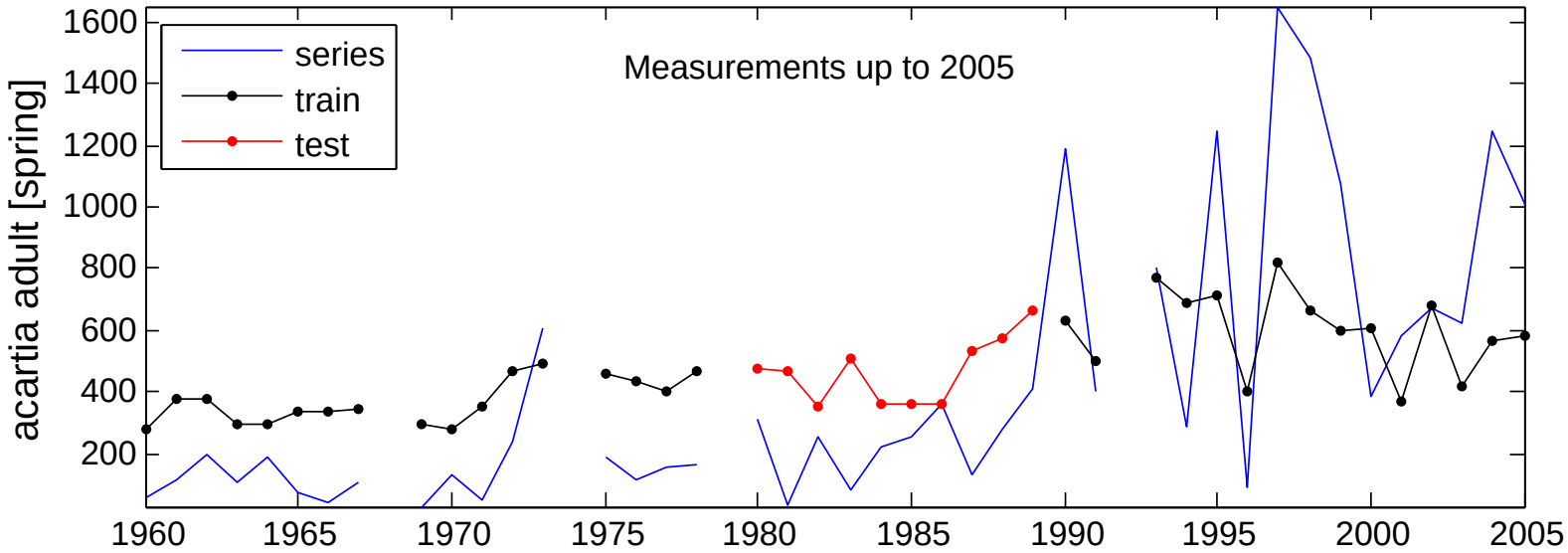

input
output
prediction
arbitrary scale
time scale
15
10
5
0
−5
0
2
4
6
8
10
12
14
16
18
20

Assumptions & results discussion

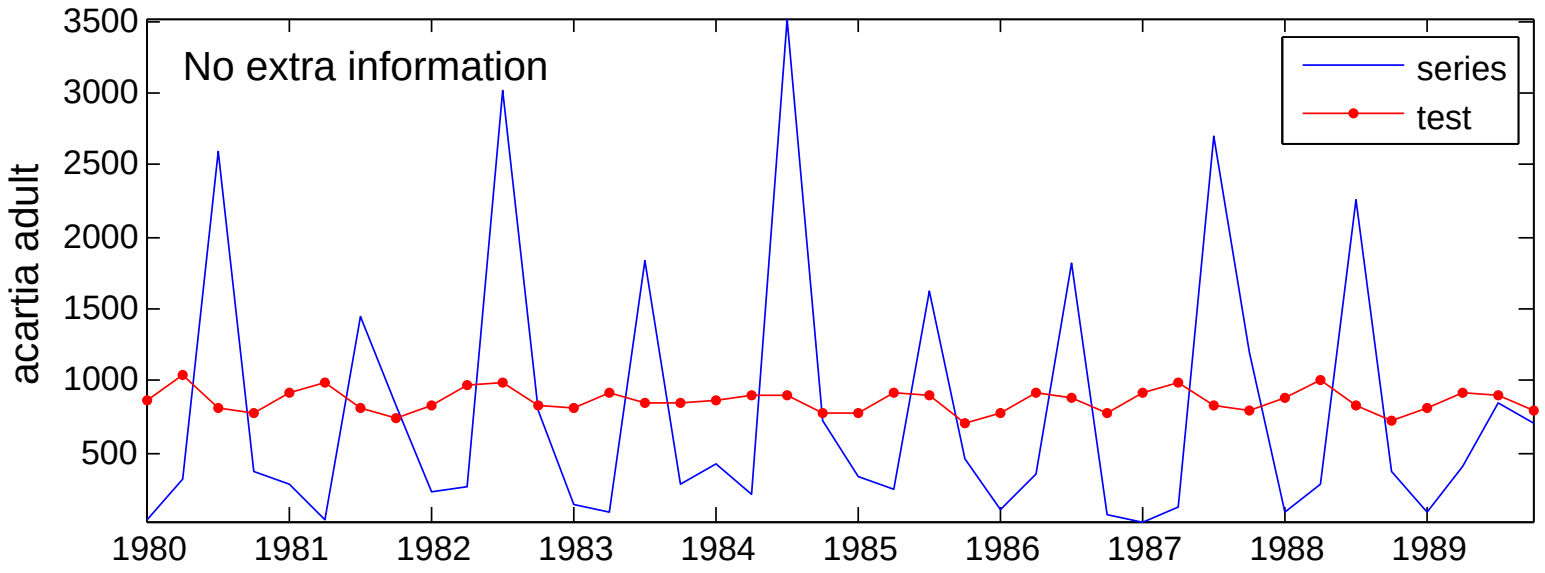
- Looking from machine learning perspective $\longrightarrow$ accuracy
- Taking into account the goal of the problem (CI $\longrightarrow$ ZP):
  - modeling restriction
  - one form of assumption
- Preprocessing the time series (yes / no?)
  - other assumptions taken as “truth”

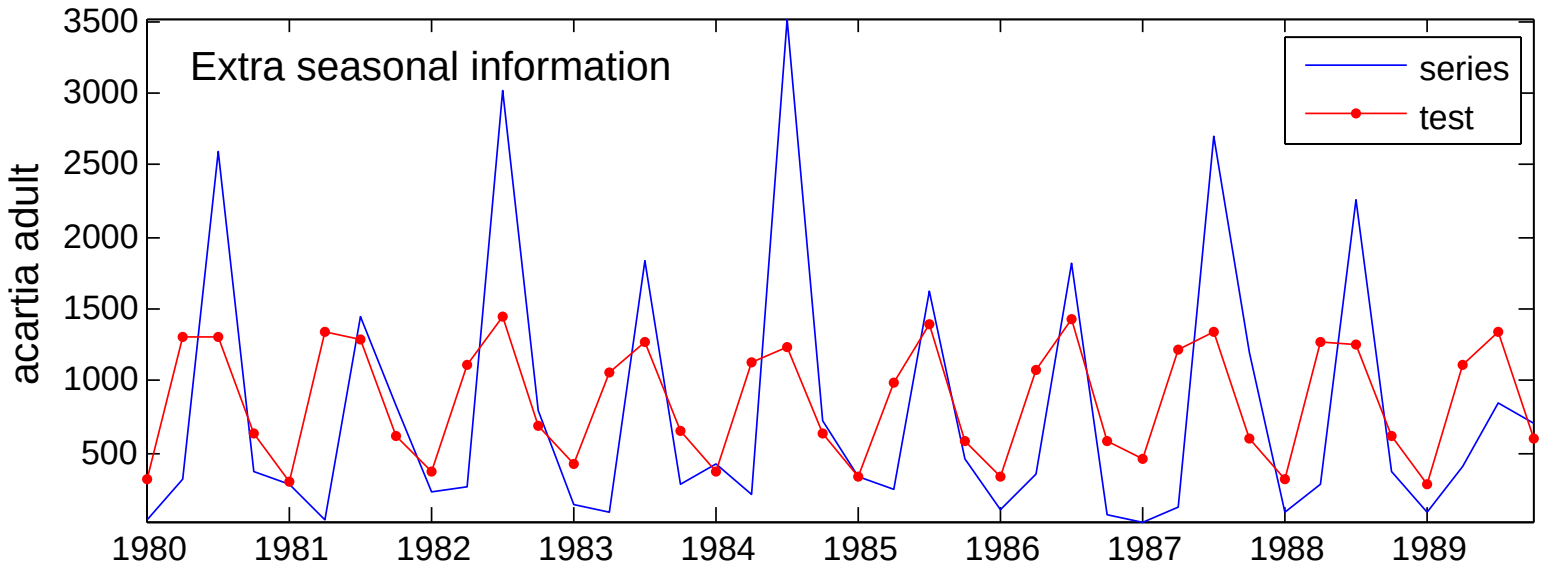
Results – time period influence





Results – additional assumptions about problem





Goal in different fields

- What is the goal of prediction?
  - Machine learning – accuracy (minimize loss function)
  - Oceanography/Biology – accuracy + interpretability (plausibility constraints)
- What is interpretability?
- A priori knowledge is too important to be neglected
- Constraints in mathematical terms can be incorporated into a loss function
- Constraints can be directly specified into the model, or inferred from the data (in this case we have interpretability)
- Constraints == model structure (i.e. assumptions about the problem)

Goal in different fields

- What is the goal of prediction?
  - Machine learning – accuracy (minimize loss function)
  - Oceanography/Biology – accuracy + interpretability (plausibility constraints)
- What is interpretability?
- A priori knowledge is too important to be neglected
- Constraints in mathematical terms can be incorporated into a loss function
- Constraints can be directly specified into the model, or inferred from the data (in this case we have interpretability)
- Constraints == model structure (i.e. assumptions about the problem)

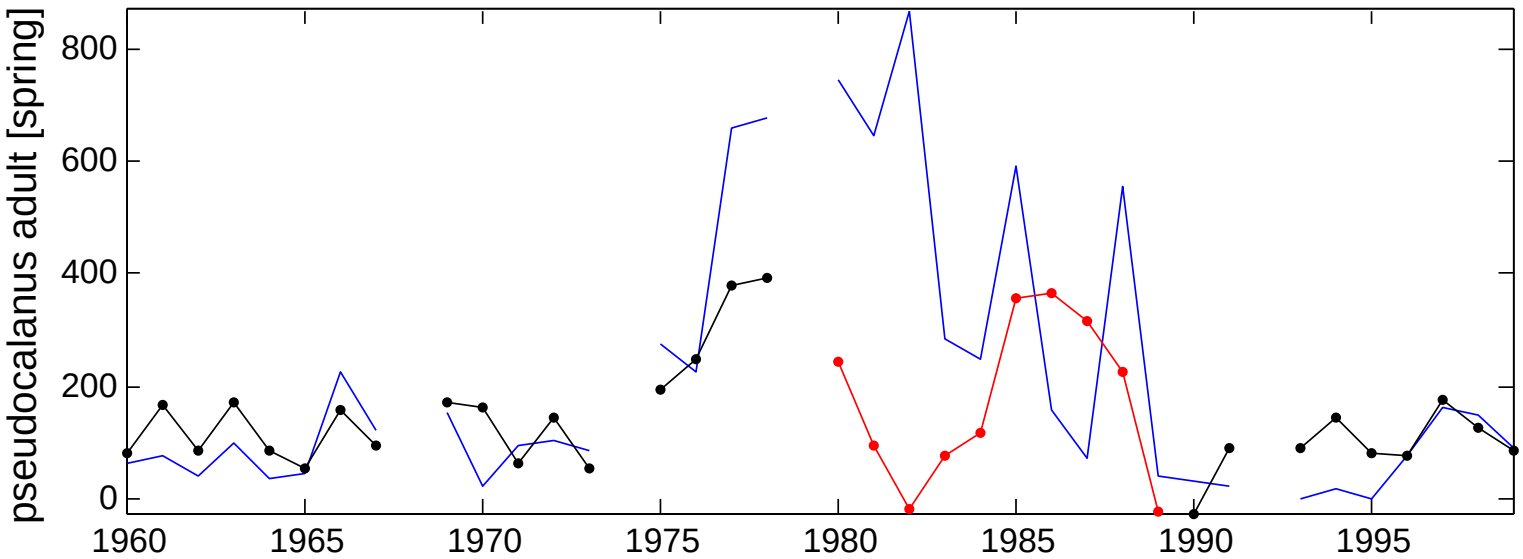
Goal in different fields

- What is the goal of prediction?
  - Machine learning – accuracy (minimize loss function)
  - Oceanography/Biology – accuracy + interpretability (plausibility constraints)
- What is interpretability?
- A priori knowledge is too important to be neglected
- Constraints in mathematical terms can be incorporated into a loss function
- Constraints can be directly specified into the model, or inferred from the data (in this case we have interpretability)
- Constraints == model structure (i.e. assumptions about the problem)

Next focus

- Include as much information as possible
  $\longrightarrow$ domain knowledge
- Runoff, temperature, salinity, ...
- Probabilistic graphical models (Bayesian networks) – assumptions specified directly into the model (relationships)

Other things



- Some species cannot be predicted at all
  $\longrightarrow$ is this completely independent from others?
- Samples, samples, samples (I want more)
- Imagine fixed number of points in increasingly higher dimensions / factors