Project 2: Water environment in Skive fjord II

Read the accompanying document: Project2Formalities.pdf

Skive fjord is part of the Limfjord system, which separates the northern part of Jutland. The fjord is monitored intensively through the national maritime monitoring program and therefore much data is available for analyzing initiatives for improving the water environment, as well as investigating different biological relations in the fjord.

The data for this project consists of monthly observations of 4 variables. Table .1 provides an overview of the variables and what they represent. The data is available in the file `dataSkivefjord.csv`
In project 1 we worked with the emissions of nitrate to Skive fjord. In order to improve the water environment legislation was implemented and we investigated if this has lead to significant decreases in nitrate emissions into Skive fjord.

One very apparent indicator of the state of the water environment is phytoplankton, since the water becomes green and disgusting, if the concentration of phytoplankton is high. We will in this project focus on phytoplankton as indicator of the water environment. The amount of phytoplankton is determined through the amount of chlorophyll

Variable	**Representation**	**Unit**
`year`	Year of the observation	
`month`	Month of the observation	
`totalP`	Total Phosphor concentration in Skive fjord	g/m^3
`chlorophyll`	Chlorophyll concentration in Skive fjord	g/m^3

Table .1: Variables included in the data set.

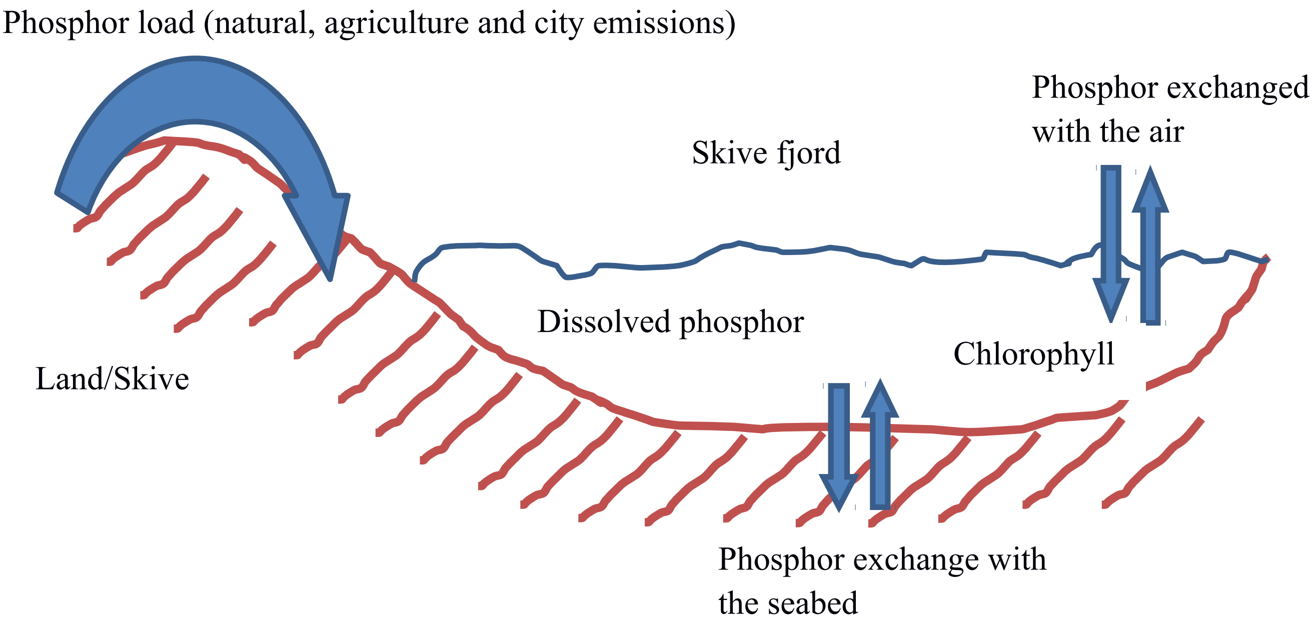
Phosphor load (natural, agriculture and city emissions)
Phosphor exchanged with the air
Skive fjord
Dissolved phosphor
Chlorophyll
Land/Skive
Phosphor exchange with the seabed

Figure .1: Simplified illustration of the phosphor circuit in Skive fjord.

(green pigment in plankton). The amounts are measured as a concentration: mass per volume.

Phytoplankton need nutrients, such as nitrate and phosphor, together with light in order to grow. Furthermore the temperature has an impact on their growth rate. Figure .1 holds a simplified illustration of the phosphor circuit in the fjord.

The variable `totalP` is thus a measure of the total amount of phosphor in Skive fjord (again as a concentration). In practice:

`totalP` = *Phosphor part of phytoplankton measured by chlorophyll + dissolved non-organic phosphor*

Analysis

Find the R code in the accompanying file: skivefjord2.R

Read the data

```
## Read the data
D <- read.table("dataSkivefjord.csv", sep=",", header=TRUE)

## Use for the description of the data
summary(D)
str(D)
```

a) Give a very short summary of the data: include for example the number of observations, which variables and which period.

Carry out a logarithmic transformation by adding a new variable `logChlorophyll` to `D`, which is the logarithm to the chlorophyll observation. In the following we will carry out a simple linear regression, where `logChlorophyll` is the dependent variable (model output) and the `totalP` is the explanatory variable (model input). Hence we want to develop a model, which describe the amount of phytoplankton as a function of the total amount of phosphor.

b) Make a scatter plot of the `totalP` and `logChlorophyll` data points, which indicates their relation:

```
## Make a logarithmic transformation of the chlorophyll concentration
D$logChlorophyll <- log(D$chlorophyll)
## Make (yourself) a scatter plot of logChlorophyll vs. totalP
```

c) Define a linear regression model having `logChlorophyll` as the dependent variable (Y_i) and `totalP` as explanatory variable (x_i) and state the assumptions of the model (write in mathematical notation).

d) Estimate the coefficients in the model, usually β_0 and β_1, and interpret the estimated values: write up the applied equations (also with the values inserted in the formulas). What do the estimated values tell about the relation between the phosphor concentration (`totalP`) and the logarithm of chlorophyll concentration (`logChlorophyll`)?

e) Verify that you calculated correctly with the following calculations:

```r
## A simple linear regression
fit <- lm(logChlorophyll ~ totalP, data=D)
## A summary of the result
summary(fit)
```

Summarize with a short text the result by including:

- The estimated standard deviations of the coefficient parameter estimates (usually denoted with $\hat{\beta}_0$ and $\hat{\beta}_1$).
- The estimate of the standard deviation of the errors (ϵ_i).
- The degrees of freedom used for the estimate of the standard deviation of the errors.
- The explained variance.

f) Carry out a model validation by analyzing the residuals to verify if the model assumptions are met:

```r
## Make a QQ-plot
qqnorm(fit$residuals)
qqline(fit$residuals)

## Plot the residuals vs. the fitted values, i.e. the
## expected (predicted by the model) values of logChlorophyll
plot(fit$fitted.values, fit$residuals)
```

Assess if the assumptions of the model are met: Is the model suitable or should be improved? Justify your assessments from the QQ-plot and the scatter plot of residuals vs. fitted values. (See Section 5.6 in the eNotes).

g) Calculate a 95% confidence band for the coefficients in the model (usually denoted with β_0 og β_1). Write up the formulas, insert values and check with R:

```r
## t quantile for calculation of confidence bands
qt(0.975, fit$df.residual)
## Function for calculation of parameter confidence bands
confint(fit, level=0.95)
```

h) Test a hypothesis to conclude if the explanatory variable contribute significantly to the description of the dependent variable. State the hypothesis, determine the value of the test statistic and p-value. Write up the formulas, insert values and check with R:

```r
## The critical value of the test statistic
qt(0.975, fit$df.residual)
```

i) The phosphor concentration was in December 2006 measured to: $\texttt{totalP}_{\text{december 2006}} = 0.0703\ \text{g}/\text{m}^3$.

Calculate a prediction and a 95% prediction interval for the chlorophyll concentration i December 2006. Write up the formulas, insert values and check with R:

```r
## Prediction of chlorophyll concentration in December 2006
exp(predict(fit, newdata=data.frame(totalP=0.0703)))
## 95% prediction interval
PI <- predict(fit, newdata=data.frame(totalP=0.0703),
              interval="prediction", level=0.95)
## Lower bound
exp(PI[,"lwr"])
## Upper bound
exp(PI[,"upr"])
```