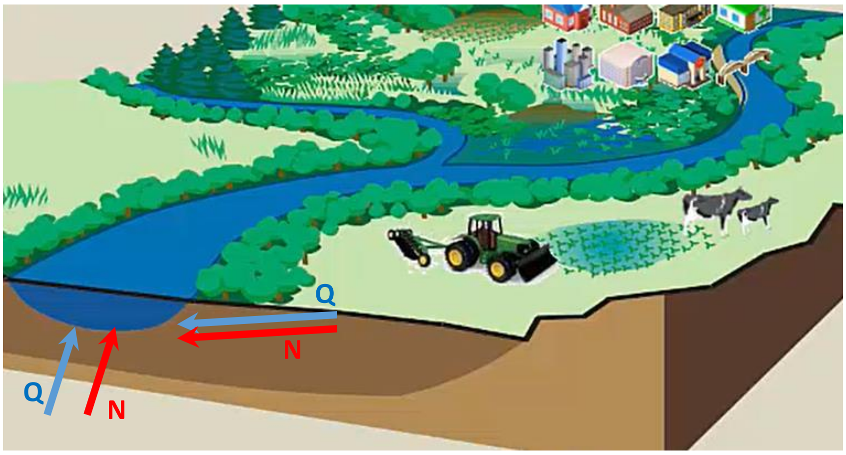
Exercise for Water and Environment module 3, Applied Statistics, spring 2023

Problem formulation

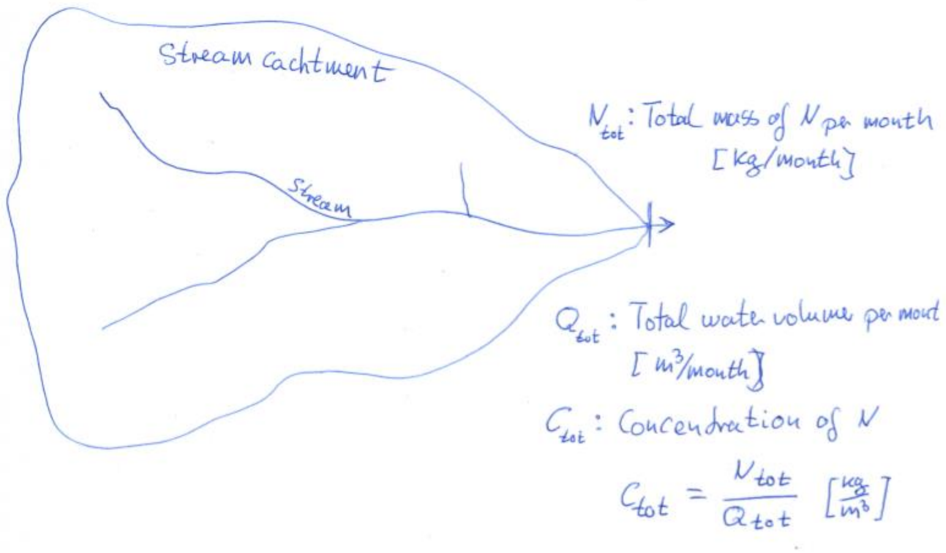
Stream water transport of nitrogen (N) has been measure for many streams in Denmark during several years in order to monitor the N leaching from agriculture areas. As the water in the stream has different origin the N contribution will come from different types of water. In general, a fraction of the water is coming from deeper ground water while another fraction is a discharge from more surface near reservoirs. As illustrated below:



It would improve interpretation of transport mechanisms if the N concentration could be estimated separately for flow from these two water reservoirs. During the period of measuring there have been some reduction in N surface application in the farming system, so a hypothesis could be that the resulting decrease in water concentration of N is more rapidly reflected as a decrease in concentration level for the water that comes from surface near water sources compared to water coming from deeper reservoirs.

Model Theory

The model is made for a cross section of the stream where a measuring station has be recording volume of water and mass of N as monthly values as illustrated in this drawing:



A mass balance is made for the stream water, where the source of N is divided up into respectively the source coming with the surface near water (*surf*) (drains, root zones and upper groundwater) and the source coming from base flow (*base*) (deeper ground water). The mass balance for the N at the cross section is:

$$C_{tot}|_{i,j} \cdot Q_{tot}|_{i,j} = C_{base}|_{i,j} \cdot Q_{base}|_{i,j} + C_{surf}\big|_{i,j} \cdot Q_{surf}\big|_{i,j} + P_{i,j} \tag{1}$$

Where i is index for year (i=1 is 1990, i=2 is 1991, etc.) and j is index for month during year (j=1: January, j=2: February, etc.). $C_{tot}|_{i,j}$ is the flow weighted average bulk concentration as it is measured in the water; $Q_{tot}|_{i,j}$ is the total volume of water passing the cross section per time (per month); $Q_{base}|_{i,j}$ is the volume of water per time that has entered the stream from deep ground water (base flow); $C_{base}|_{i,j}$ is the base flow concentration; $Q_{surf}\big|_{i,j}$ is the volume of water per time coming from the surface near flow (drains, root zones, upper groundwater); $C_{surf}\big|_{i,j}$ is the surface near water concentration; $P_{i,j}$ is contribution from points sources (wastewater treatment facilities, etc.). Eq. 1 can be rewritten as:

$$C_{tot}|_{i,j} - \frac{P_{i,j}}{Q_{tot}|_{i,j}} = C_{base}|_{i,j} \cdot \frac{Q_{base}|_{i,j}}{Q_{tot}|_{i,j}} + C_{surf}\big|_{i,j} \cdot \frac{Q_{surf}\big|_{i,j}}{Q_{tot}|_{i,j}} \tag{2}$$

The water volume is divide in the two sources:

$$Q_{tot}|_{i,j} = Q_{base}|_{i} + Q_{surf}\big|_{i,j} \tag{3}$$

The eq. 3 is rearranged:

$$Q_{surf}\big|_{i,j} = Q_{tot}|_{i,j} - Q_{base}|_{i} \tag{4}$$

Eqs. 2 and 4 can now be combined and rearranged as:

$$C_{tot}|_{i,j} - \frac{P_{i,j}}{Q_{tot}|_{i,j}} = C_{base}|_{i} \cdot \left(1 - X|_{i,j}\right) + C_{surf}\big|_{i,j} \cdot X|_{i,j}, \text{ where } X|_{i,j} = \frac{Q_{tot}|_{i,j} - Q_{base}|_{i}}{Q_{tot}|_{i,j}} \tag{5}$$

It is assumed that the base flow concentration ($C_{base}|_i$) and base flow ($Q_{base}|_i$) is constant during the year but may change from year to year (thus only having index i and not index j).

The Eq. 5 can be further rearranged to:

$$C_{tot}|_{i,j} - \frac{P_{i,j}}{Q_{tot}|_{i,j}} = C_{base}|_i + \left(C_{surf}\big|_{i,j} - C_{base}|_i\right) \cdot X|_{i,j} \tag{6}$$

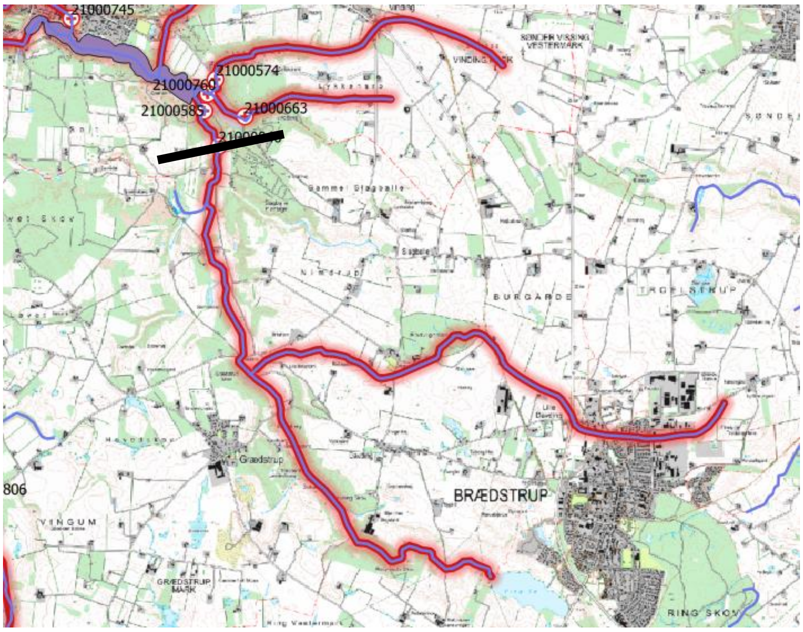
If the point sources $(P_{i,j})$ are known then the variable X can be considered as a continues independent variable in the linear equation having the form:

$$Y|_{i,j} = \alpha|_i + \beta|_{i,j} \cdot X|_{i,j} \tag{7}$$

Comparing Eqs. 6 and 7 it is seen that $\alpha|_i$ is an estimated for the base flow concentration that is assumed constant during an year, but may change from year to year. The coefficient $\beta|_{i,j}$ is seen to estimate the difference between surface concentration and base flow concentration.

Data set

The data set is for the main inlet to Lake Bryrup (Denmark), see the file NModelData.xlsx



The point sources are neglected. The base is also assumed constant during the year and is estimated using the lowest flow for the year ($Q_{YearMin}|_i$). Hint: You may do some of the data manipulaton in Excel in case you find tjhis more easy.

1. Make a column in the data set for X, assuming that the base flow is equal $Q_{YearMin}|_i$ and a column for $Ctot$.
2. Explain how the following model in R is reflecting Eq. 7:

```
Ctot ~ X * Year + X:Month,
```

3. Make the coefficient table
4. How will you asses the validity of this model?
5. Make the full time series (as a curve) where you estimate *Ctot* for each month and where you also plot the measured values of *Ctot* (as dots). Evaluate how well the model can simulate the data. Hint:
   ```
   Prediction<-predict(LinearModel.1)
   NModelData$Prediction<-Prediction
   ```
6. Estimate *Cbase* as a function of year. Is there a significant ($p<0.05$) relation between *Cbase* and year? Is the time relation linear during all the years? Is the slope significant ($p<0.05$) different from 0? What is the value of the slope.
7. Estimate *Csurf* for January as a function of years. Is this the relation between year and *Csurf* linear during the whole period? Is the slope of a linear relation significant ($p<0.05$) different from zero. Compare the value of the slope with the slope for *Cbase* (Question 6).
8. Estimate *Csurf* as function of the month for year 1990. How will you explain the seasonal dynamic?
9. During the period, some regulation of agricultural praxis has been implemented to reduce the leaching of N. Is this reflected in the estimated values of the slope for respectively *Cbase* and *Csurf*? Please suggest an explanation for the different value of the slope for respectively *Cbase* and *Csurf*.