Case study on normalisation and trend analyses for OSPAR RID data

Claudia von Brömssen

Applied Statistics and Mathematics
Dept. of Energy and Technology
Swedish University of Agricultural Sciences (SLU)

Case study on normalisation and trend analyses for OSPAR RID data

Claudia von Brömssen

claudia.von.bromssen@slu.se

Contents

1. Introduction

The objective of this report is to present statistical methods for normalisation, trend analysis and breakpoint analysis and illustrate these methods on some case studies. It includes general recommendations on methodology that can be used for OSPAR RID data and a suggestion how to conduct such analysis in the open source software R.

2. The RID database

2.1 Reported loads and river flow

For OSPAR reports the load of substances is reported as annual loads:

$$Load = Q_r \frac{\sum_{i=1}^{n} C_i Q_i}{\sum_{i=1}^{n} Q_i}$$

where C_i is the measured concentration in sample i, Q_i is the corresponding flow for sample i and Q_r is the mean flow rate for the period in question, i.e. annual flow. n is the number of samples taken during the period (OSPAR Commission, 2008).

Annual loads are usually computed from monthly or quarterly data. River flow is monitored continuously or daily in most countries, but local differences do occur.

3. Flow normalisation

Flow normalisation for time series of loads is relevant when the interest is not on the observed annual load to a receiving water body including short-term variation caused by meteorological or hydrological conditions but instead on long-term changes over time. Statistical methodology for normalisation stretches from simple ratios, over linear regression models to flexible non-linear models.

3.1 Choice of normalisation period

When data is normalised the goal is to establish reasonable loads for a situation when meteorological or hydrological conditions were at a constant level over time. This is obtained by removing the short-term variations created by annual flow conditions, while containing the mean of the loads at a level that would be observed for a mean flow $\bar{q}$. The mean flow can be the average flow over the entire observed time period or for a chosen reference period. For subsequent trend test the choice of mean flow level is not important since it is constant over time and therefore does not influence the temporal structure.

If the goal of the analysis is to estimate normalised annual loads at reasonable levels in order to compare inputs to receiving waters from different countries or to add up inputs from different sources it is important that the same reference periods are used. It could be problematic, for example, to use period A for one input series and period B (which is some years shorter) for another, if the excluded years were especially dry or wet. The reference period should be chosen to be as long as possible to give a relevant estimate of the conditions during an 'average' year. Also, using

normalisation, it is assumed that flow levels do not change systematically over time, i.e. they do not exhibit a trend on their own.

3.2 Methods

Normalisation can be conducted using statistical regression models or by simple ratios between loads and flow. Three approaches are described and compared below.

3.2.1. Flow normalisation by ratios

Loads can be normalised by taking the ratio between the load and the flow rate (Uhlig and Kuhbier, 2001):

$$RL_i = \frac{L_i}{q_i}$$

This gives a measure on the concentration scale. To normalise this value to a normal flow year the ratio is multiplied by an appropriate long term average or reference value for flow $\bar{q}$:

$$NRL_i = \frac{L_i}{q_i} \cdot \bar{q} \tag{A}$$

3.2.2. Flow normalisation by linear regression

A more common way to do flow-normalisation is by linear regression, assuming a linear relationship between the response variable, here nutrient load, and the flow variable. The model used is then

$$L_i = \alpha + \beta_2 q_i \tag{B}$$

for years i=1, n. Parameter α is the intercept and β describes the relationship between discharge and load. An increase of one unit in discharge leads to a modelled β units increase in load.

For the model above it is assumed that the mean load is approximately constant over time, which in longer time series usually is not the case. To account for this the model can be extended to:

$$L_i = \alpha + f(t_i) + \beta_2 q_i \tag{C}$$

where $f(t_i)$ represent the mean levels changing over time and β_2 describes again the relationship to loads. The structure of $f(t_i)$ can be either linear, and is than replaced by $\beta_1 t_i$ in the formula, or non-linear but smooth. See section 4 for details.

Normalised loads are computed by removing the interannual variation of flow rate.

$$NL_i = L_i - \beta q_i + \beta \bar{q}$$

where $\bar{q}$ is a flow rate mean. This means that for normalised loads the effect of the annual flow rates is removed and the effect of a 'normal' year is added to retain correct levels of mean loads (Stålnacke and Grimvall, 2001).

3.2.3. Flow normalisation by non-linear regression

The relationship between flow rate and loads could also be assumed to be non-linear. In that case the function describing this relationship could be modelled as a smooth or any other kind of non-linear function

$$L_i = \alpha + f(t_i) + g(q_i)$$

Normalised loads can be produced from such models in a similar way as above but using the mean of the function $g(q_i)$ to restore correct levels for the loads. Again, the trend function can be either linear or smooth.

3.2.4. Assumptions in models

Normalisation models establish a function to describe the relationship between flow and loads and use the estimated relationship to remove effects of flow from the series. Since significance testing is not an important part of the normalisation step, we do not need data that are normally distributed or have equal variance. However, parameter estimation gets more stable if the distribution is at least symmetric and if variances are not too different. Since RID data are annual sums of loads their distribution is usually rather symmetric. It is also advisable to conduct normalisation only if there is a clear relationship to flow and the data is free from influential data points, see case study 3.3.2 for examples.

A further requirement in statistical models is that data used is independent. This is clearly not fulfilled in time series data, since measurements are made on at the same station or area at several times points. Annual data, however, exhibit often quite small correlations in time and therefore it is not proposed that the estimation of temporal correlation is included in these models.

3.3 Case study

3.3.1. Case study: The river X.

Figure 1 (top) shows that the temporal pattern for flow rate and total nitrogen loads look very similar. In Figure 1 (bottom) total nitrogen loads are plotted against the flow rate, indicating a strong dependence of load on flow. In fact, the correlation between these two variables is 0.88. The plot of nitrogen loads does, therefore, rather reflect patterns in discharge and not in anthropogenic forcing.
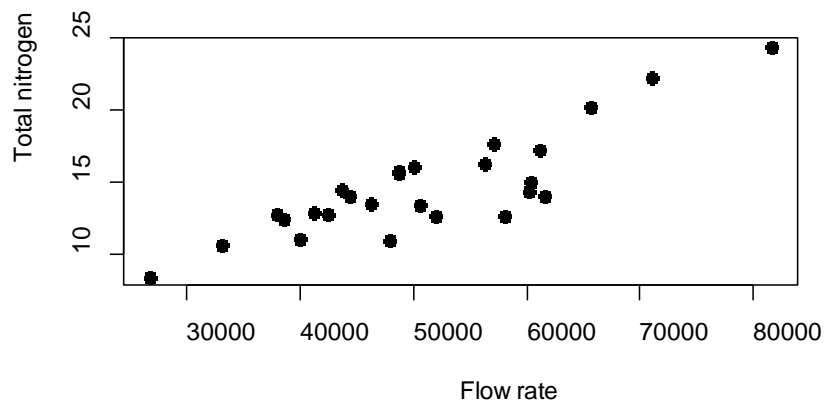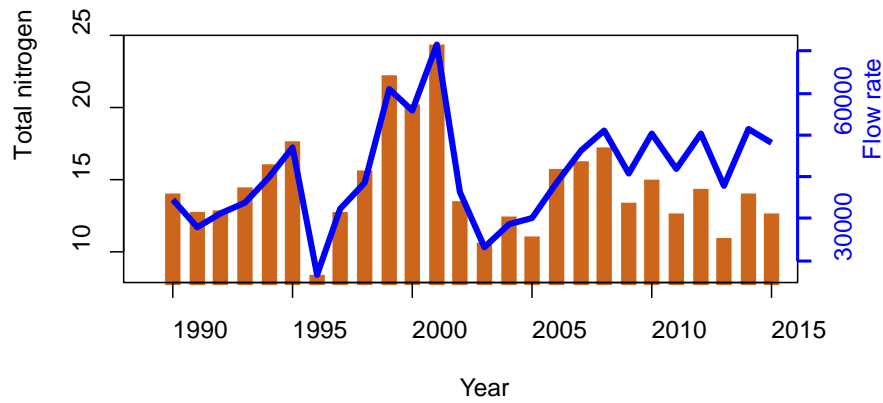
Figure 1: Top: Total nitrogen and flow rate against time. Bottom: Total nitrogen against the flow rate.

Normalisation is conducted by the three approaches discussed above. Figure 2 shows the observed total nitrogen loads and the normalised loads. The normalised loads from the 3 different models are very similar.
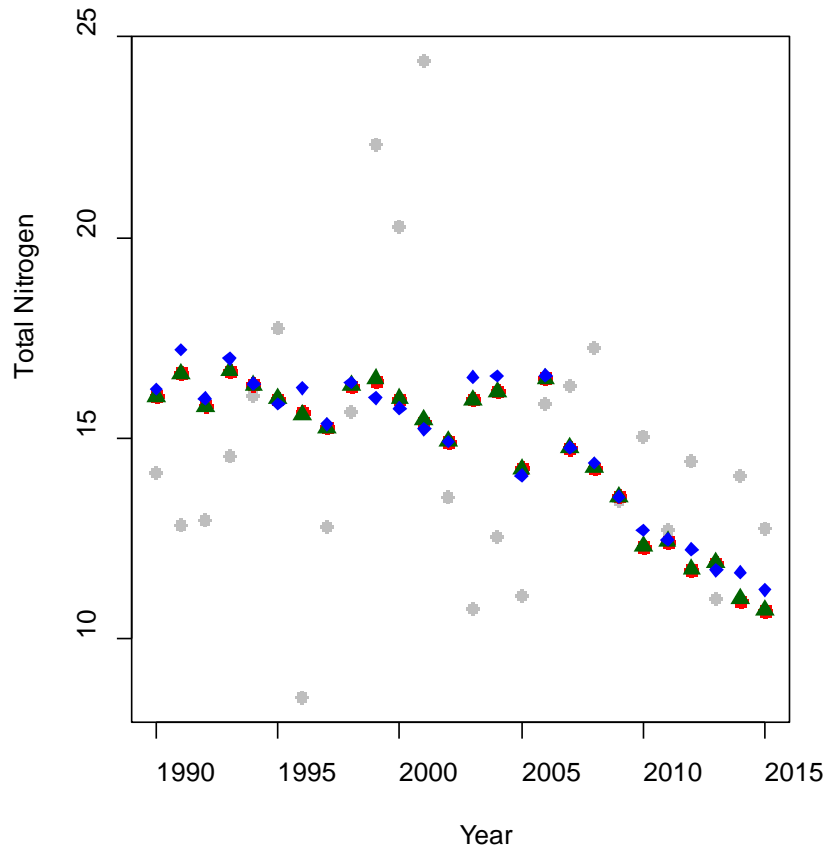
Figure 2: The original observed total nitrogen loads (grey circles) and normalised loads produced by the three methods: ratios (A, blue), linear regression with linear trend (B, red) and linear regression with a non-linear trend (C, green).

Since the dependence of total nitrogen loads on flow rate is strong, the choice of normalisation method is less important. With both methods B and C trend lines are estimated simultaneously and these are plotted in Figure 3. In this case it can be seen that the linear trend overestimates normalised values in the beginning and the end of the series, and underestimates values in the middle. The smooth trend line represents normalised loads better. If there is any knowledge about changes in emissions around year 2006 the series could also be split there and the two parts could be analysed separately. In that case linear trends for both parts are to be preferred , since the separate series would be too short to fit non-linear trends (Larsen and Svendsen, 2013).
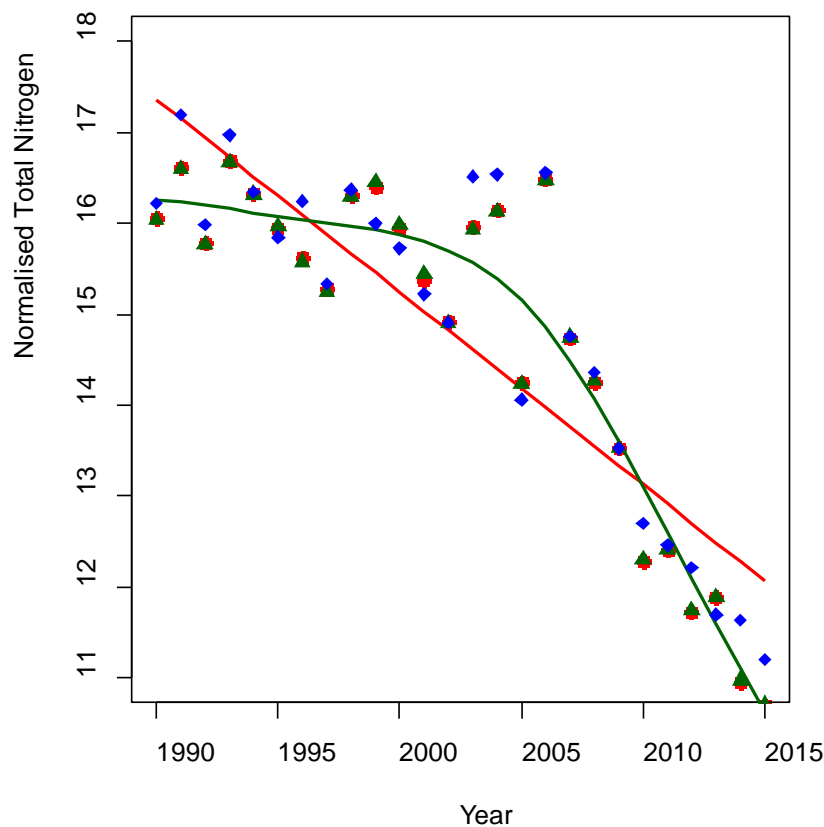
Figure 3: Normalised total nitrogen loads and estimated trend curves: linear (red) and smooth (green) trend. The normalised values as above.

3.3.2. Case study: The sea area Y.

Total phosphorus input to Y is presented in Figure 4. In this case there are three data points that need special attention. The observations 1990 and 1991 lie on the same level (1.1) and are considerably higher than other values with comparable flow rates. The observation in 1999 has a very low level of phosphorus at the same time as the flow rate is quite large. To be able to discuss effects of these deviating observations we distinguish between the terms outlier and leverage point:

- outlier: an outlier is a value that lies at a considerably higher or lower value compared to other (surrounding) values
- leverage point: is an outlier that in addition lies at a – for computations or models – crucial location

For this data we can say that the observations for 1990 and 1991 are leverage points in the trend analysis, since they constitute high values at the start of the series and thereby influence trend estimates considerably. The observation in 1999 is not equally influential for the trend estimation and can be considered an outlier.

Regarding the normalisation step the observation in 1999 is a leverage point since it influences the slope of the normalisation function (Figure 4, bottom and Figure 5), while the observations in 1990 and 1991 are outliers. The analysis below is done in two parts: Part 1 uses all observations

6

and Part 2 presents the analysis when the three observations in 1990, 1991 and 1999 are removed.

Part 1: all observations included
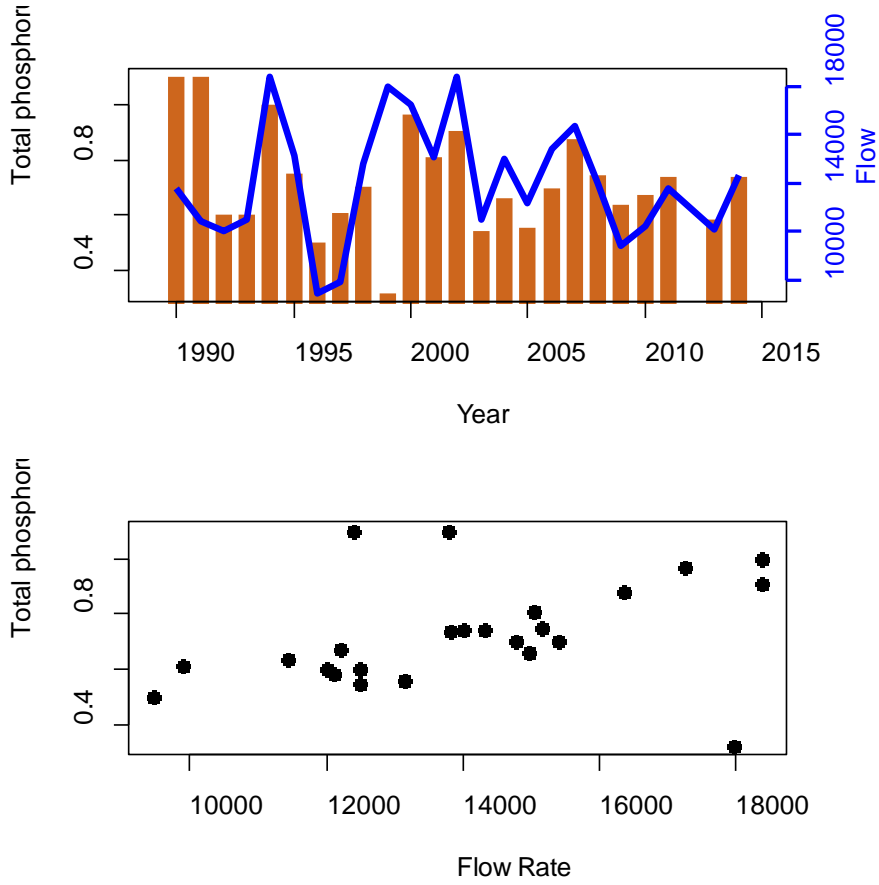
Figure 4: Top: Total phosphorus and flow rate against time. Bottom: Total phosphorus against flow rate.
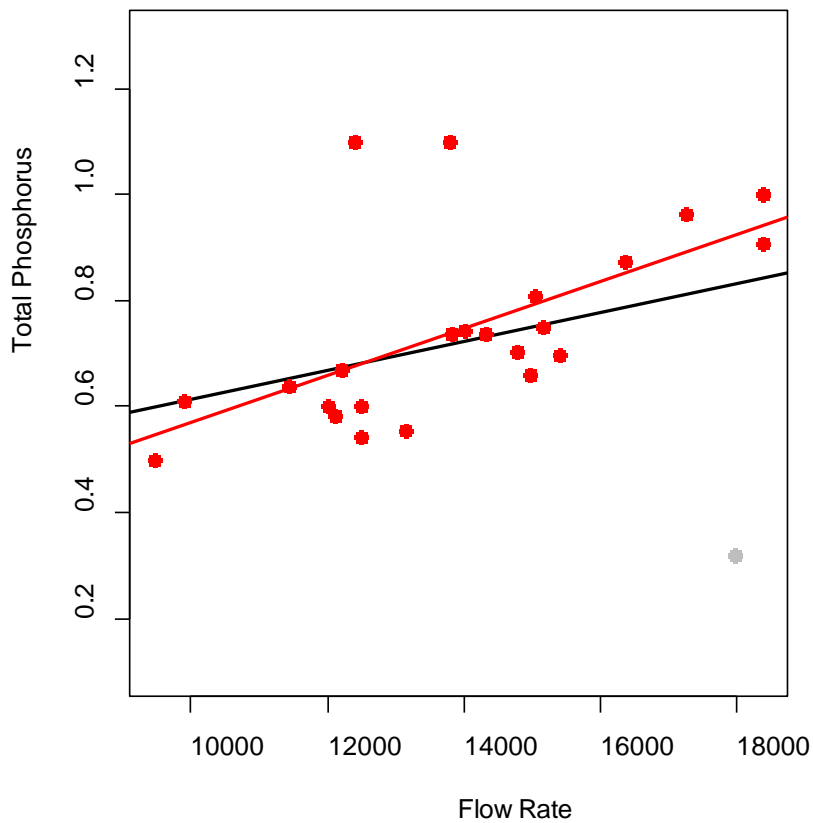
Figure 5: The relationship of total phosphorus to flow rate. Black line: the regression fit for all data, red line: the regression fit if the leverage point in 1999 is removed (grey point).

In cases like this it should be determined if the leverage point is correct and why it does not fit into the common picture of the flow-load relationship. The exclusion of the point should be considered, since it worsens the estimated relationship for the remaining points and, by that, could introduce artefacts in the normalised data.

Figure 6: The original observed total phosphorus loads (grey circles) and normalised loads produced by the three methods: ratios (A, blue), linear regression with linear trend (B, red) and linear regression with a non-linear trend (C, green).

In Figure 6 we also see that there is a larger spread for the normalised values as there has been for the data in river X, since the normalisation model is influenced in different ways by how the trend function is estimated due to these leverage points. Especially, it should be noted that the normalised values of the outlying observations for 1991 and 1999 actually are even more deviating than the original observations.

Part 2: influential observations removed

To illustrate the effect of the 3 influential observations further we repeat the study with data for 1992-2014 and a missing observation 1999. Data is presented in Figure 7. For this data the flow-load relationship is much clearer and the different normalisation methods lead to very similar results (Figure 8).
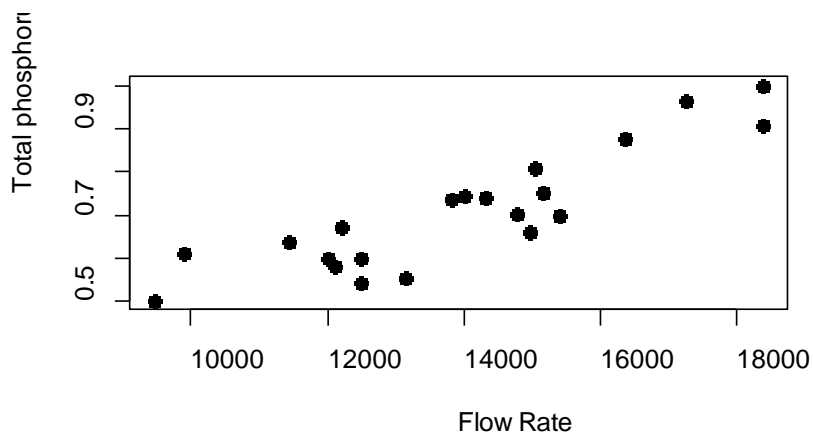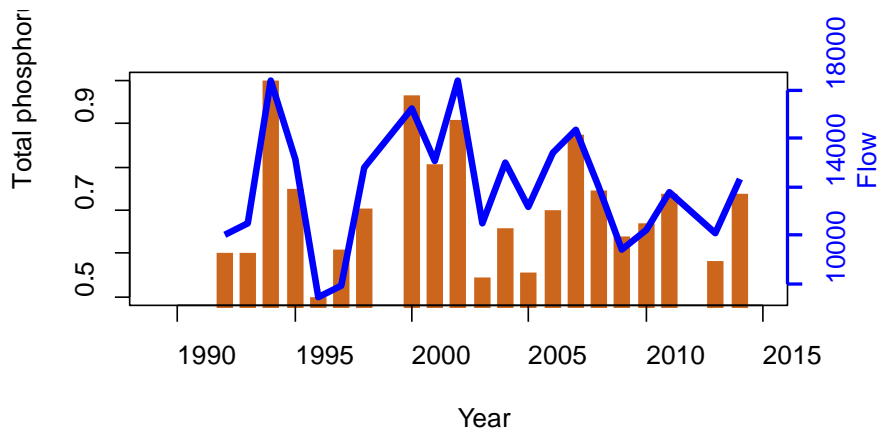
Figure 7: Top: Total phosphorus and flow rate against time. Bottom: Total phosphorus against flow rate. 3 observations are removed (1990, 1991 and 1999).
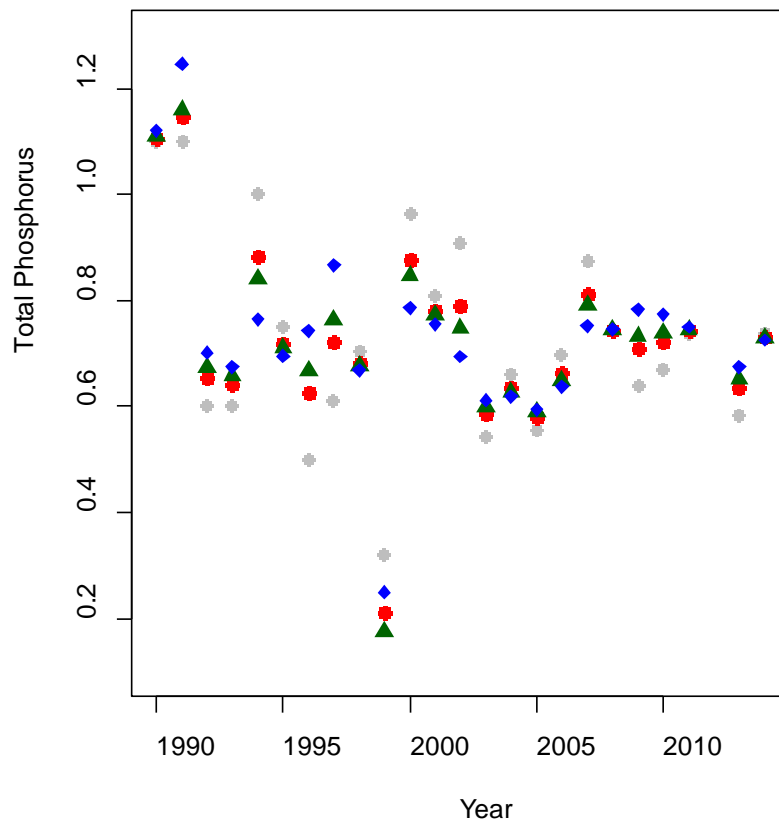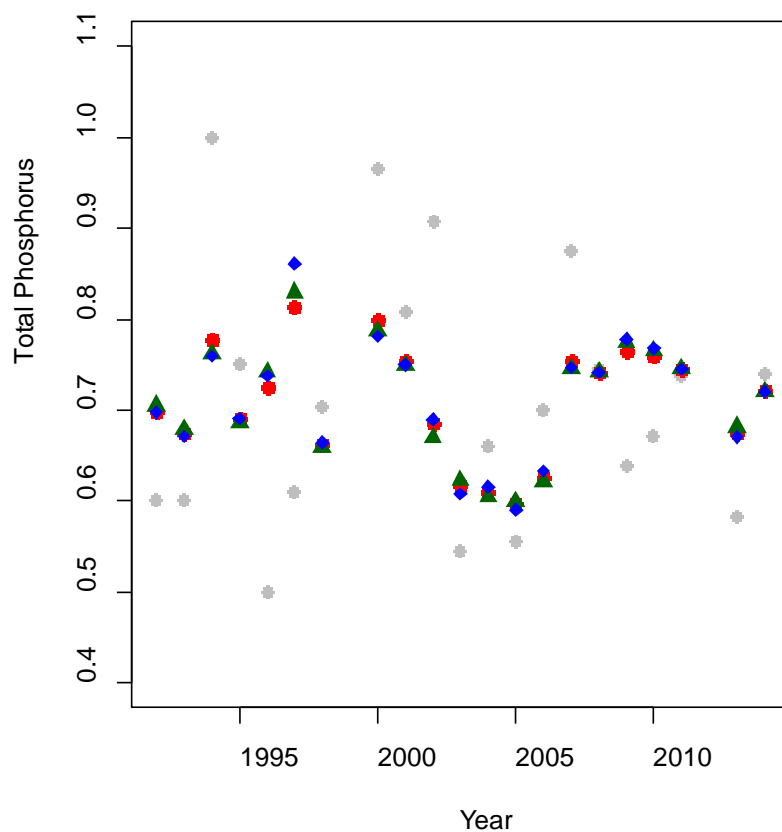
Figure 8: The original observed total phosphorus loads (grey circles) and normalised loads produced by the three methods: ratios (A, blue), linear regression with linear trend (B, red) and regression with a non-linear trend (C, green).

Nonlinear relation between load and flow

For both series also a non-linear relationship between load and flow was tested. For this a thin plate spline fit in a generalised additive model was used. Thin plate spline fits have the advantage that they reduce to linear fits if that is the best function to describe the data. For both case studies (for the sea area Y with and without leverage points) the spline fit resulted in a linear fit.

3.4 Recommendation for RID data

The RID database contains annual load data both for inputs to sea areas and for individual rivers. Loads are in their nature, dependent on the predominant runoff conditions during the year and can therefore vary strongly between years. If the goal is to compare annual inputs into receiving waters over time, short-term variations due to such hydrological conditions are often not interesting and impair the possibility to detect trends due to the increased variation in the series. Normalisation is therefore recommended if possible, i.e. if the information on flow rate conditions for individual years is available and the relationship between load and flow is reasonably strong.

The method chosen for normalisation is generally of minor importance when the dependencies between flow and loads are strong. Typically some kind of regression model is chosen, but simple

division by annual flow can be made, especially if time series are short. When regression models are used the relationship between flow and load can be assumed linear or non-linear. The uniquely most common assumption, which is also recommended for RID data, is the one of a linear relationship. A linear relationship is intuitively meaningful due to the way the loads are computed and held for the two case studies presented. Non-linear relationships were tested, but did not contribute to better model.

Linear relationships between flow and load are quite robust to individual outliers. Leverage points are observations that are outliers in a very influencing position within the data. For data from sea area Y we had an observation with a high annual flow, but a rather low annual load. Since this observation lies far from the linear relationship fitted for the remaining data it influences the outcome substantially and the correctness for such observations should be checked. The normalisation step should always be accompanied by a plot of the flow-load relationship.

4. Trend analysis

Trend analysis can be conducted on total loads or on total normalised loads. Since interannual variation is removed by flow normalisation it is generally easier to see and statistically detect anthropogenic trends in normalised loads.

4.1 Methods

4.1.1 Nonparametric trend tests: Mann-Kendall tests and Sen's slope

Non-parametric methods relax the distributional assumptions in tests. In the case of Mann-Kendall tests (Mann, 1945; Hirsch and Slack, 1984) underlying data does not need to be normally distributed. Instead computations are made on the ranks of observations, which is the same as using signs for differences between pairs of observations:

$$MK = \sum_{j<i} sgn(y_i - y_j)$$

where sgn(x) is the sign function

$$sgn(x) = \begin{cases} 1, & if\ x > 0 \\ 0, & if\ x = 0 \\ -1, & if\ x < 0 \end{cases}.$$

The resulting statistics MK is used to test if the observed trend is statistically significant or not. Mann-Kendall tests can detect linear or monotone trends.

In addition Sen's slope (or Theil-Sen slope, Theil, 1959, Sen, 1968, Gilbert, 1987)) can be computed to quantify the average yearly increase or decrease in the series. Sen's slope is the median annual change over the entire time period. It does not say anything about the shape of the trend.

Partial Mann-Kendall tests

If normalisation by a model is not possible or not chosen, an adjustment for flow can be made within the Mann-Kendall test using the so-called partial Mann-Kendall test (Libiseller and Grimvall, 2002).

For this the Mann-Kendall statistics MK for the load is adjusted for the Mann-Kendall statistics of flow, leaving a test statistics for the part of the trend that cannot be explained by changes in flow.

Assumptions in the analysis:
Data does not need to follow any specific distribution, but observations do still need to be independent of each other. Otherwise, if observations are positively correlated, p-values will be too low leading to an increased risk for falsely positive tests. The p-value of the test is based on a normal distribution which holds for the MK statistics if there are at least 10 years of annual data.

4.1.2 Linear trends: Linear regression
Linear regression can be used if the temporal development in the series is approximately linear. The model fitted is

$$L_i = \alpha + \beta_1 \cdot t_i$$

The slope estimate β_1 gives the magnitude of the linear trend and can be tested to see if the change in time is significant or not.

Percentage change
The estimated amount of increase or decrease as a percentage of the initial levels can be computed from the estimated parameter $\hat{\beta}_1$ or using predicted values from the model:

$$PC_{lin\,reg} = 100 \cdot \frac{(start - end)}{start}$$

where *start* indicates the reference value of the series, often the modelled load for the first year in the series and *end* the modelled load for the last year.

Assumptions in the analysis:
Traditional linear regression assumes normally distributed errors, i.e. residuals in the model should be approximately normal. Furthermore observations need to be independent and variation around the regression line needs to be approximately constant. If these assumptions are violated the p-values for the trend tests cannot be relied on.

4.1.3 Non-linear trend: Estimation of smooth curves
If trends cannot be assumed to be linear some version of smoother can be used. Typical choices are loess or spline smoothing (Hastie and Tibshirani, 1986; Wood, 2006; Cleveland and Devlin, 1988). Generally the model can be written as

$$L_i = f(t_i)$$

where $f()$ is any type of smooth function, meaning that the function is governed by observed levels at the same time as the modelled mean is not allowed to change too quickly between years.

Percentage change
Smoothers are non-parametric methods, in the meaning that they do not estimate parameters such as intercept or slope, and therefore significance testing of trends is more difficult. As before modelled start and end values can be used to compute the percentage change over the entire time period:

$$PC_{smooth} = 100 \cdot \frac{(start - end)}{start}$$

Assumptions in the analysis

Model outputs are not used for statistical testing in this case and therefore there is no requirement for data to be independent or normally distributed. However, estimates are usually more stable as data distributions are more symmetric and independent.

4.1.4 Percentage change

As described above percentage change in the time series can be computed from modelled values. In the same manner percentage change can be computed directly from the normalised (or non-normalised) values, i.e

$$PC_{norm} = 100 \cdot \frac{(start - end)}{start}$$

where *start* and *end* stand for the normalised values for the first and last year.

4.2 Case study

4.2.1 Case study: The river X.

Mann-Kendall test and Sen's slope

Different normalisation procedures for the river X resulted in very similar normalised time series for total nitrogen. Therefore it is not essential which method is used to normalise the series. Here we use normalised data by the linear regression approach (B). As seen from the p-value in Output 1 the trend is highly significant (p-value: 0.0000125). The Sen's slope is estimated to be -0.1975 indicating a median decrease of 0.1975 units per year. Note that no information is given about if the decrease is constant over time or the result of a sudden drop. Therefore it is necessary to combine Mann-Kendall tests with plots over normalised values.

Output 1: The results for the non-parametric Mann-Kendall trend test and Sen's slope for total nitrogen in river X using the rkt package in R. P-value in yellow, Sen's slope in green.

Standard model
Tau = -0.6266667
Score = -188
var(Score) = 1833.333
2-sided p-value = 1.257464e-05
Theil-Sen's (MK) or seasonal/regional Kendall (SKT/RKT) slope= -0.1974725

Trend analysis by linear regression

If a linear trend can be assumed, trend testing can be done by linear regression. In that case it is best to combine the normalisation and trend testing step in the same model, which has been shown to work better than stepwise approaches.

In Output 2 we can see that the trend in the data is highly significant (p-value: 0.00000098) and the mean annual decrease is 0.1982, i.e. very similar results to the nonparametric test and slope.

Output 2: Results of a linear regression including the normalisation using the lm function in R. p-value in yellow and slope estimate in green.

Call:
lm(formula = TotN ~ year + flowRate, data = riverX)

Residuals:
     Min   1Q Median   3Q   Max
-1.55262 -0.68161 -0.02781 0.63740 2.40213

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.966e+02 5.893e+01   6.73 9.17e-07 ***
year   -1.982e-01 2.958e-02   -6.70 9.81e-07 ***
flowRate   2.937e-04 1.760e-05   16.68 5.69e-14 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.011 on 22 degrees of freedom
Multiple R-squared: 0.9272,   Adjusted R-squared: 0.9206
F-statistic: 140.2 on 2 and 22 DF, p-value: 3.024e-13

Since we from the normalisation study already know that the linear trend assumption does not hold for river X (see also Figure 9), the mean annual decrease should not be interpreted as an ongoing decrease of this magnitude, but should again be judged in connection with the plot of normalised data, which indicates rather constant levels before 2005 and a steeper drop thereafter.
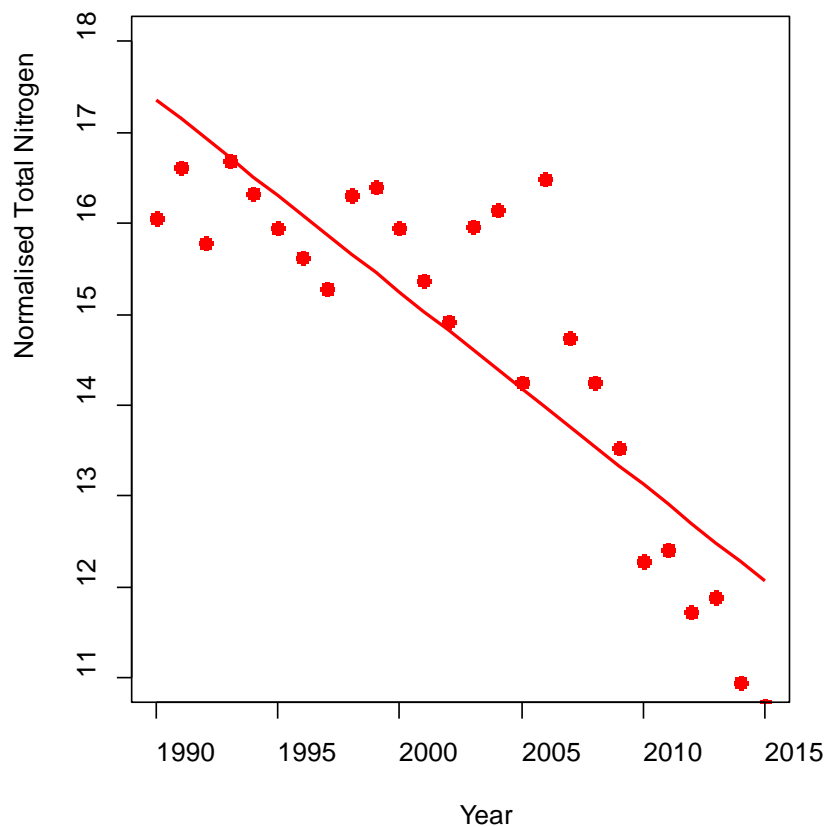
Figure 9: Normalised total nitrogen loads and linear trend line for river X.

The percentage of decrease over the entire time period can be computed from the normalised values for 1990 and 2014:

$$PC_{norm} = 100 \cdot \frac{10.86223 - 15.97286}{15.97286} = -31.99573$$

or from the modelled values of the linear trend model for the same years:

$$PC_{lin\_reg} = 100 \cdot \frac{12.41485 - 17.17067}{17.17067} = -27.69733$$

Here, however, it must be noted that the estimation for the modelled values is influenced by the misspecification of the model. The trend is not linear and the estimation of percentage change for the entire time period is therefore less reliable.

Trend analysis by non-linear smooth regression

Also in the smooth trend model both trend estimation and normalisation can be done at the same time. Fitting a smooth function to describe trends in time allows more flexible structures over time (Figure 10). This comes with the drawback that there is no parameter in the model that can be used to test if the trend is significant. Normalised values from this model can be trend tested using Mann-Kendall tests to compensate for this.

Output 3: Output from the smooth model fitted by the gam function in the mgcv package in R.

Family: gaussian
Link function: identity

Formula:
TotN ~ s(year) + flowRate

Parametric coefficients:
         Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.416e-02 5.896e-01   0.058   0.954
flowRate   2.913e-04 1.137e-05 25.614   <2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
        edf Ref.df   F p-value
s(year) 3.177 3.928 38.63 2.27e-14 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) = 0.97   Deviance explained = 97.5%
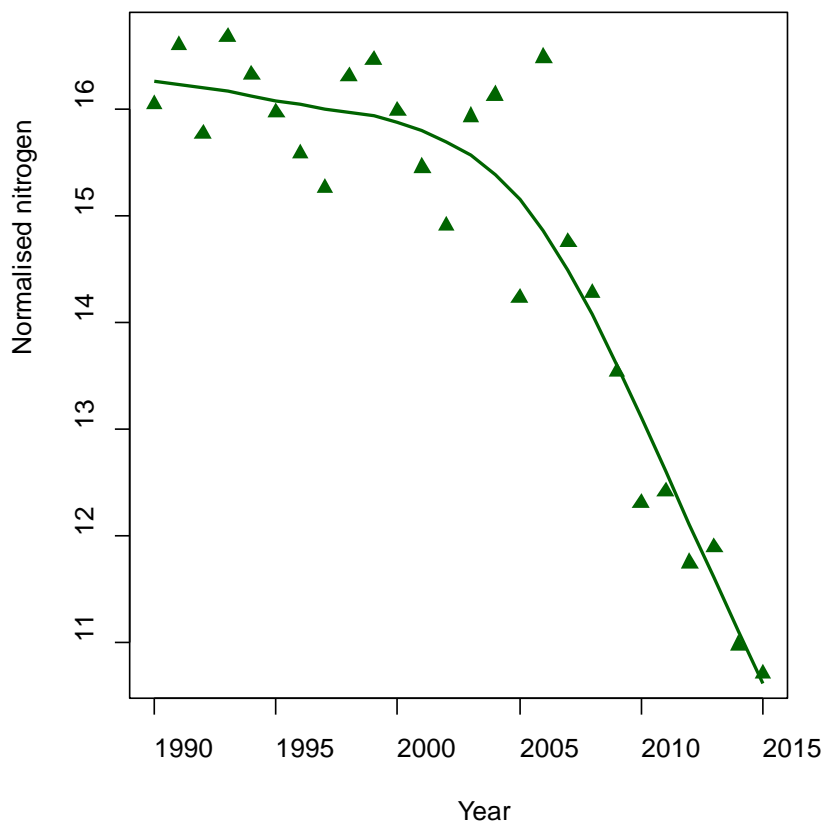GCV = 0.49143 Scale est. = 0.38966   n = 25

Figure 10: Normalised values and trend curve from the smooth model for river X..

Percentage changes for the entire time period can again be computed using normalised or modelled values. Since normalised values produced from the different models are very similar also the

17

estimated percentage change for these values is very similar to the ones produced by the linear model:

$$PC_{norm} = 100 \cdot \frac{10.888 - 15.958}{15.958} = -31.769$$

The smooth model is not bound by any a priori assumptions of the trend and therefore the modelled values are more reliable than from the linear regression model. The percentage change for the modelled values is computed as:

$$PC_{smooth} = 100 \cdot \frac{10.978 - 16.181}{16.181} = -32.152$$

4.2.2 Case study: the sea area Y.

As seen in the normalisation step, results for Y are influenced by 3 observations, which should be considered and quality checked. If the decision is to remove these three observations the normalisation results are no longer dependent on model choice. Also the trend analysis is influenced by the outlying observations, especially by the two initial data points in 1990 and 1991. Below we see the results when all observations are included (Part 1) and how results change without these observations (Part 2). The corresponding linear and nonlinear trend curves are given in Figure 11 (all observations) and Figure 12 (without influential observations).
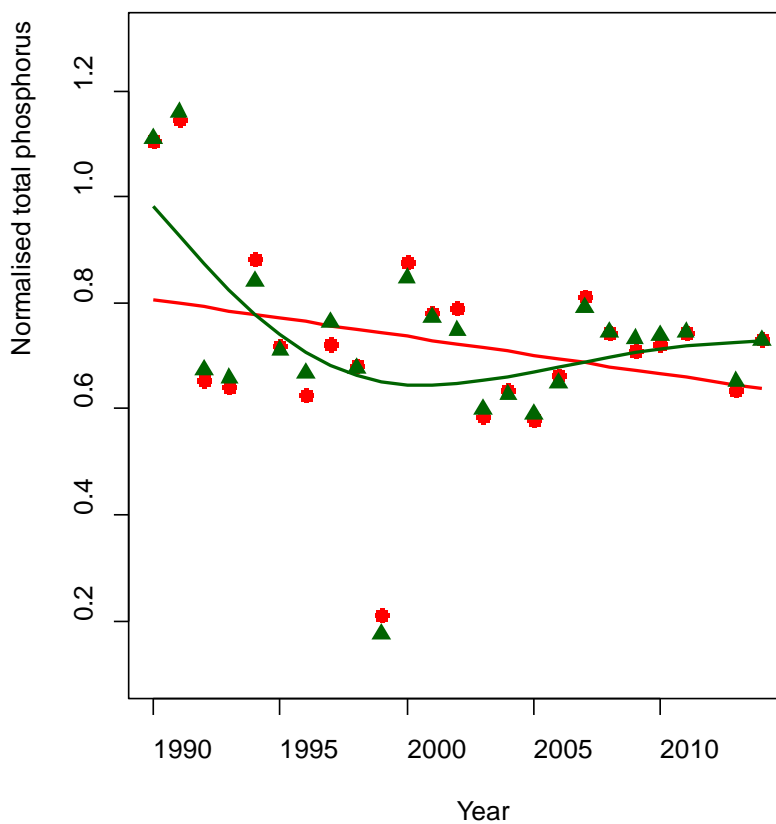
Figure 11: Normalised total nitrogen loads and estimated trend curves: linear (red) and smooth (green). The normalised values as presented in the normalisation section.

Figure 11 shows that the smooth trend curve is more easily influenced by the outlying observations, especially the leverage points in the beginning of the series, but also the linear regression line is affected (many points in the beginning of the series lie below the regression line, and most points in the end of the series lie above). In Figure 12 the estimated trend curves are more stable and do not show any trend. The smooth curve adapts more readily to changes in mean values that persist for some years.
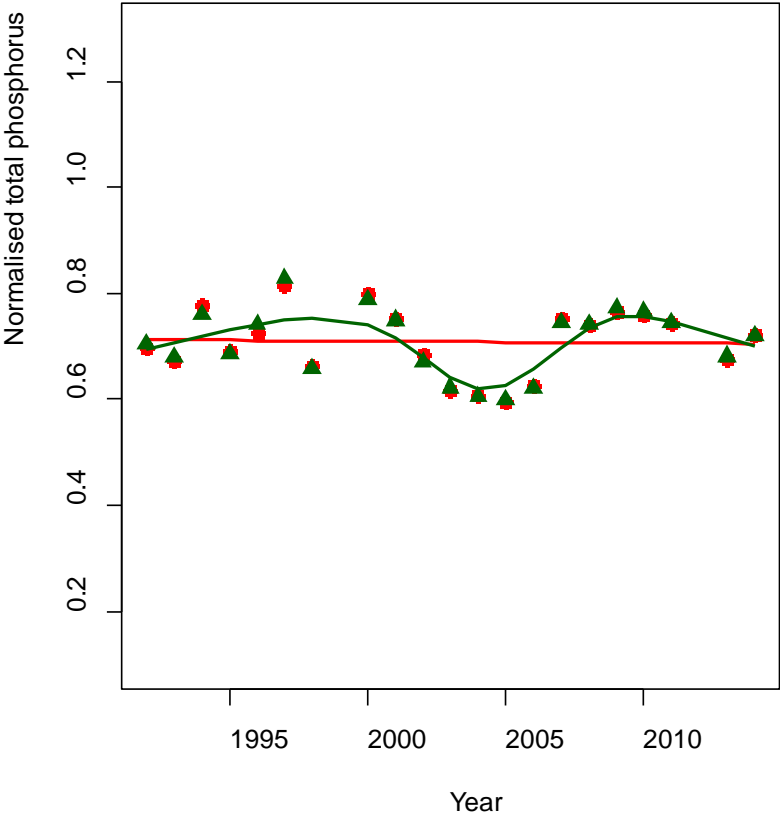
Figure 12: Normalised total nitrogen loads and estimated trend curves: linear (red) and smooth (green). The normalised values as presented in the normalisation section.

Mann-Kendall test and Sen's slope

Part 1: all observations
In Table 1 we see that the trend test result is slightly dependent on the choice of normalisation method. However all lead to the conclusion that there is no significant trend in the data. In this case the non-parametric nature of the Mann-Kendall test is an advantage, since it is computed on ranks of observations instead of original observations and this lowers the influence of the strongly deviating observations in the beginning of the series. Since the remaining part does not show any clear further downward trend the results are non-significant.

Table 1: Results of the Mann-Kendall test for the three types of normalised data

	Ratio normalised	linear trend normalised	smooth trend normalised
MK trend test: p-value	0.31	0.41	0.26
Sen's slope	-0.0026	-0.0039	-0.0028

Part 2: influential observations removed

When the three deviating observations are removed the results for the trend tests do not vary much for different normalisation methods and results show clearly that there is no significant trend in the data. Results are given in Table 2.

Table 2: Results of the Mann-Kendall test for the three types of normalised data

	Ratio normalised	linear trend normalised	smooth trend normalised
MK trend test: p-value	0.88	0.83	0.79
Sen's slope	-0.0004	-0.0009	-0.0007

Trend analysis by linear regression

Part 1: all observations

If trend tests are done by linear regression we see that neither the linear trend nor the relationship between flow and load is significant (Output 3). Only about 10% of all variation in phosphorus load can be explained by this model. The estimated trend curve is given in Figure 11.

Output 3: Results of a linear regression including the normalisation using the lm function in R. p-value in yellow and slope estimate in green, R^2 value in blue.

Call:
lm(formula = mean1 ~ year + flow1, data = seaareaY)

Residuals:
    Min   1Q Median   3Q   Max
-0.53138 -0.08531 0.01247 0.08643 0.34564

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.434e+01 1.038e+01   1.382   0.1816
year   -6.992e-03 5.183e-03 -1.349   0.1917
flow1   2.726e-05 1.520e-05   1.794   0.0872 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1792 on 21 degrees of freedom
Multiple R-squared: 0.1933,   Adjusted R-squared: 0.1165
F-statistic: 2.516 on 2 and 21 DF, p-value: 0.1048

The computed percentage changes are quite high since they are strongly influenced by the high value in 1990. Using modelled values of the trend curve instead of normalised values lowers this effect somewhat. The percentage changes need however to be interpreted together with the plots over the

data and the non-significant trend tests made on the data. Hence, the conclusion must be that there is a high uncertainty about the magnitude of decrease in sea area Y.
For normalised values:

$$PC_{norm} = 100 \cdot \frac{0.732 - 1.107}{1.107} = -33.89$$

For modelled values from the linear regression model:

$$PC_{lin\_reg} = 100 \cdot \frac{0.6385166 - 0.8063325}{0.8063325} = -20.81224$$

Part 2: influential observations removed

As already noticed before the relationship between flow and load becomes clearer when the three deviating observations are removed. This leads to a stronger estimated relationship between flow and load in the model at the same time as the trend estimate gets smaller. For this data 77% of the variation can be explained by the model (Output 4). The trend curve was shown in Figure 12.

Output 4: Results of a linear regression including the normalisation using the lm function in R. p-value in yellow and slope estimate in green, R^2 value in blue.

Call:
lm(formula = mean1 ~ year + flow1, data = seaareaY)

Residuals:
    Min   1Q Median   3Q   Max
-0.11164 -0.03852 0.01459 0.04689 0.10346

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.985e-01 4.418e+00   0.158   0.876
year   -3.450e-04 2.204e-03 -0.157   0.877
flow1   5.025e-05 6.066e-06   8.285 1.48e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06662 on 18 degrees of freedom
Multiple R-squared: 0.7925,   Adjusted R-squared: 0.7695
F-statistic: 34.38 on 2 and 18 DF, p-value: 7.123e-07

Since the initial high values are removed this has a substantial effect on the computations of percentage change, both for the normalised values (3.34%) and for the modelled values (-1.07%). The percentage change is now computed with 1992 as start year and 2014 as end year.
For normalised values:

$$PC_{norm} = 100 \cdot \frac{0.721 - 0.698}{0.698} = 3.34$$

For modelled values from the linear regression model:

$$PC_{lin\_reg} = 100 \cdot \frac{0.705 - 0.713}{0.713} = -1.07$$

21

Part 1: all observations

Similar results as for the linear regression are obtained by the non-linear smooth regression. The trend curve was shown in Figure 11. Since the smooth curve, however, adapts more easily to single or small groups of observations the modelled value for 1990 is higher than for the linear regression and therefore the percentage change is also estimated to be higher, but still less than for normalised values.

For normalised values:

$$PC_{norm} = 100 \cdot \frac{0.729 - 1.109}{1.109} = -34.255$$

For modelled values from the smooth regression model:

$$PC_{smooth} = 100 \cdot \frac{0.727 - 0.984}{0.984} = -26.074$$

Part 2: 3 observations removed

Again removing the 3 observations leads to no changes in the levels for total phosphorus loads. The trend curve was shown in Figure 12. The estimates are 0.79% and 2.1% , and far from statistically significant (as shown with Mann-Kendall) or practically interesting.

For normalised values:

$$PC_{norm} = 100 \cdot \frac{0.72 - 0.705}{0.705} = 2.1$$

For modelled values from the smooth regression model:

$$PC_{smooth} = 100 \cdot \frac{0.701 - 0.696}{0.696} = 0.79$$

4.3 Recommendation for RID data

When time series get longer it is usually a disadvantage to assume strict functional forms for the shape of the trend. A linear trend, for example, suggests that the decrease or increase is constant over the entire time period, which is often not the case, as seen from the case study on river X. The RID database contains 25 years of data and therefore the trend models should be flexible and fitted by a smooth curve. Smoothers have been used for trend analysis for a long time and there are several software packages designed for environmental data, e.g. the RTrend package or the MULTITREND program for Excel (Grimvall et al., 2009). Here, similar models within the general additive models framework (GAM, mgcv package in R (Wood, 2006)) were used to fit smooth trends.

Using smooth trends to visualise and analyse trends comes with the drawback that significance testing is not readily available. Changes between the start and the end of the series can be presented as percentage changes. If these values are predicted from the model significance tests could be based on the model uncertainty, but this is not discussed here. Instead, the results from the smooth

trend fitting and the computed percentage change should be presented together with a non-parametric trend test on normalised data. For this the Mann-Kendall test is recommended.

Computing percentage change from the beginning to the end of a series can be done for normalised data or for modelled values. Generally modelled values are more stable, since they are less influenced by other confounding factors that were not included in the model, e.g. meteorological or hydrological influences other than flow. If other factors are neglectable the results for the two approached are expected to be very similar.

5. Breakpoint detection

The analysis of break points in a series can have two different starting points:

a) using the knowledge that the series is interrupted due to any planned or unplanned event within the sampling period. Examples of such events could be the addition or removal of substantial point sources or the change of methodology or change of accredited laboratory used to make the analyses.

b) not using any a priori knowledge in order to find so far unknown reasons for a change in mean level or other data characteristics. Example of such reasons could be shifts in the ecological systems or unknown problems with measurement equipment.

The presence of breaks in series is an important issue in trend analysis, since abrupt changes in levels can be mistaken as trends (e.g von Brömssen et al., 2017). Available methods are described and discussed below.

5.1 Methods

5.1.1 Methods that have a priori information about the location of a break point
If there are known reasons that can result in a level shift in time series this knowledge can be used to improve the trend analysis.

A general approach is to include the time point of change in the model used to describe the trend. This is possible if e.g. a linear or smooth regression model is used. It needs, however, be noticed that it usually is problematic to estimate a level shift at the same time as a trend estimate, since these two estimates influence each other and this procedure can lead to an underestimation of the magnitude of the level shift and an overestimation of the trend magnitude or vice versa. Clearly, this is still better than to ignore shifts in level.

Another possibility is to analyse different time periods separately from each other (Larsen and Svendsen, 2013). For RID data this is probably the better approach, since it demands less modelling effort and can be conducted as long as there are reasonably long series before and after a break point.

For both approaches above the time point of changes needs to be known. At this stage no information is present in the RID database about events that could lead to breaks in single or multiple series.

CUSUM methods

CUSUM methods (McGilchrist and Woodyer, 1975) are sequential methods computing the cumulative sum of the response variable over time. By this a plot is created from which it can be seen if the level of a series deviates from earlier levels. CUSUM plots are mainly used in industry for quality assurance of production processes, assuming that a variable lies on a constant level until a problem arises. The deviation from the constant level is observed in the CUSUM plot after a number of time steps. The plots are also used in environmental assessment, but are less useful there due to the presence of trends, high inter-annual variation and possible outlying observation due to other reasons than process changes. CUSUM plots can however be used in dynamic adaptive monitoring for specified systems, i.e. by defining a threshold that should not be passed for a particular lake or river. Then the goal is to quickly detect and evaluate sudden changes in mean levels (Mac Nally and Hart, 1997).

Pettitt's test

Pettitt's test (Pettitt, 1979) is used to determine the location of a break point in a series. The test assumes constant levels before and after this breakpoint and the computation of the test is based on ranked observations rather than observed values, which makes it more robust against outliers in the series. Pettitt's test is used retrospectively, i.e. it is usually not used to detect deviations in real time, but in a follow-up analysis. A simulation study (von Brömssen et al., 2017) showed that Pettitt's test often indicate significant break points at location where no break was simulated and generally works badly when the series additionally includes a trend.

5.2 Case study: Cadmium loads in sea areas Z1 and Z2

Cadmium loads to the sea area Z1 drop dramatically in 2003, which is easily seen in the plotted data (Figure 13, top). A CUSUM method can correctly locate this break point (Figure 13, bottom). Similarly a single high observation in cadmium loads in sea area Z2 is identified as break point by CUSUM (Figure 14). In both cases the plot of the original data gives more information than the CUSUM results. For the data for Z1 quality needs to be assured, since the break comes after a two-year period with no observations and at the same time as flow observations are started to be reported. For Z2 we are looking at an outlying observation, since no data is available for the time period before 1990. Even for this quality needs to be assured.

For the plots below the efp function in the package strucchange in R is used. This function allows missing values in data series.
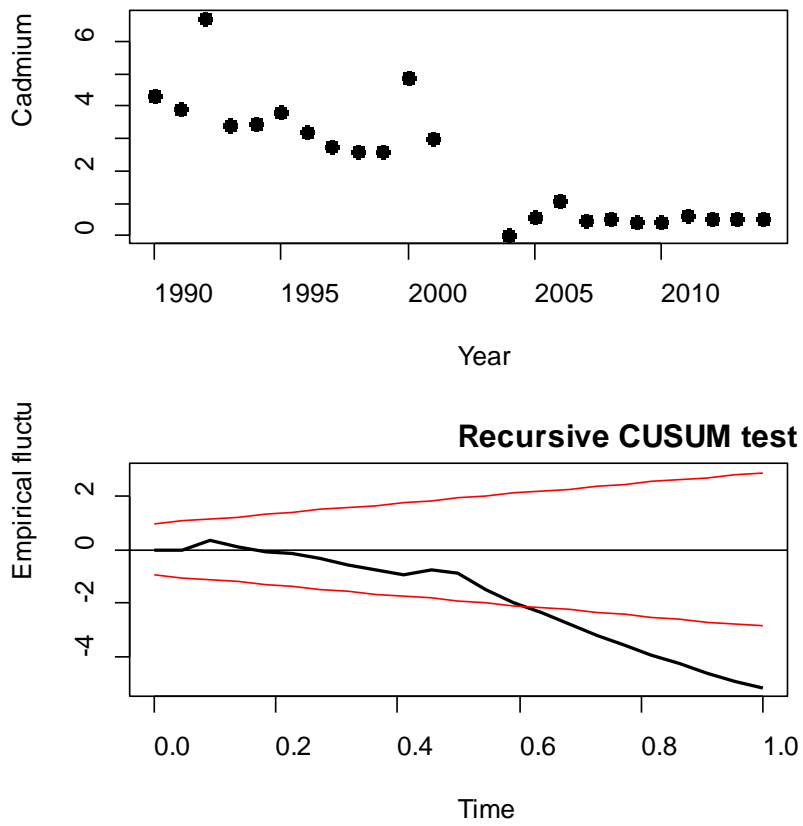
Figure 13: Cadmium loads in sea area Z1. A dramatic reduction in 2004 after a period of 2 years without data (Top). The break is detected and correctly located by a CUSUM method, when the black line crossed the red line (Bottom).
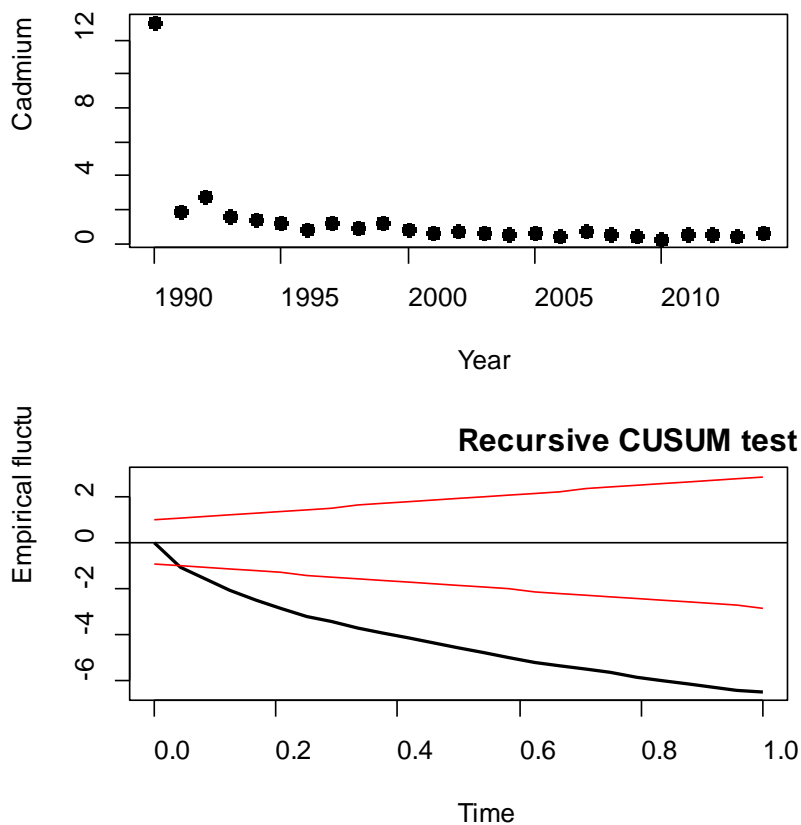
Figure 14: Cadmium loads in sea area Z2. A single high value is observed in 1990 (top) and identified as break point by CUSUM (bottom).

Pettitt's test could not be used for the Z1cadmium loads, since this series has missing values.

For the Z2 cadmium loads Pettitt's test suggests a highly significant break after 11 years (Output 5), which is hard to interpret from the plot of the data (Figure 15).

Output5: The Pettitt's test computed by the trend package in R. P-value in yellow, indicated time point of change in green.

Pettitt's test for single change-point detection

data: seaareaZ2$mean1
K = 154, p-value = 0.0003148
alternative hypothesis: true change point is present in the series
sample estimates:
probable change point at tau
          11

Figure 15: The time series of cadmium loads to sea area Z2 with the break point at year 2000 suggested by the Pettitt test.

5.3 Recommendation for RID data

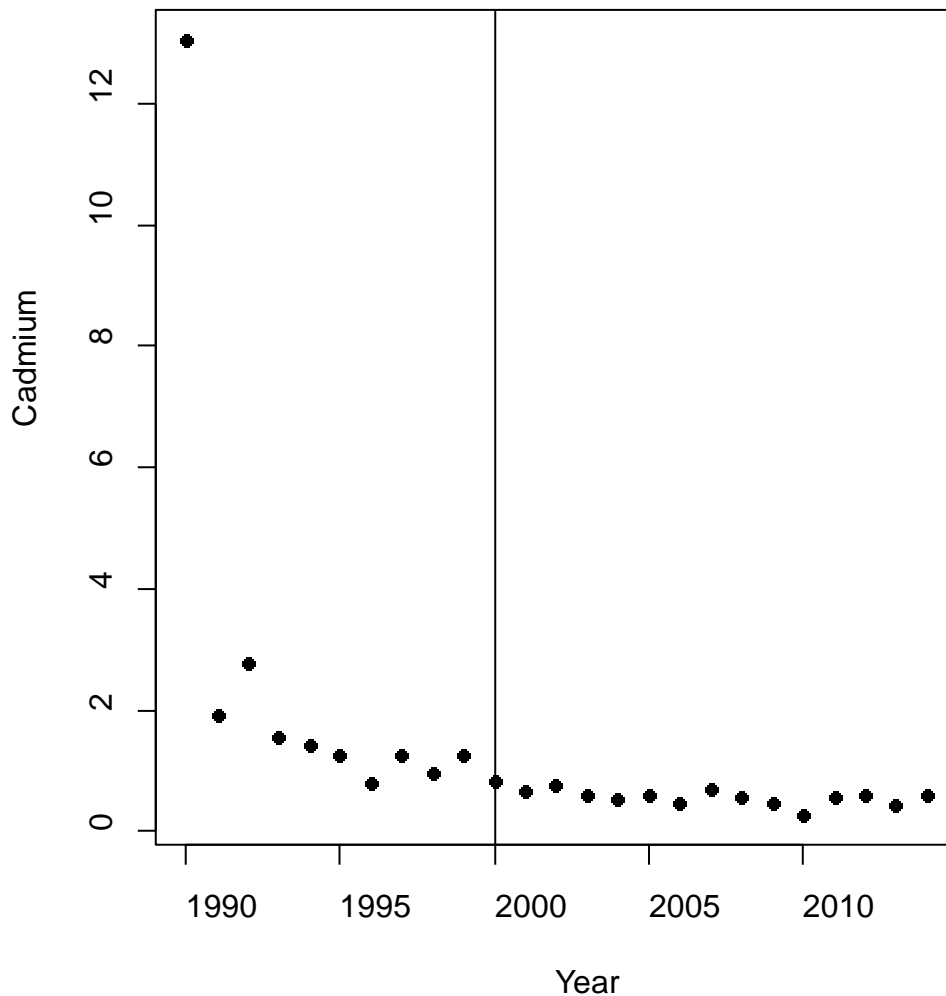OSPAR RID data consist of about 25 years of annual loads for a number of rivers and sea areas. Generally it cannot be expected that system changes without a priori knowledge could be identified on data with this aggregation levels unless the effect is extremely strong and can be clearly observed visually. Furthermore none of the available methods is really appropriate for RID data: CUSUM works best adaptively while Pettitt's test works poorly even on long series, especially if trends are present (von Brömssen et al., 2017). In shorter series, like RID data, tests will, most likely not be able to identify any break points unless there are outlying observations and even then they seem to suggest breaks at unreasonable time points.

Further it must be noticed that many of the break point detection methods in R do not support the presence of missing values in the data. This is true, for example, for the local.trend function in the pastecs package (a CUSUM method, also used in the TTAinterfaceTrendAnalysis), the cusum function in the qcc package and the Pettitt's test function pettitt in the trend package. The efp function in the package strucchange can, however, be used to compute the CUSUM method if missing values are present.

If the goal of a breakpoint analysis is to find a time point where a trend is starting or levelling out the best approach in time series of RID type is to rely on visual inspection, for example, enhanced by a smooth trend curve fit and combined with knowledge about the sources and processes behind the data.

6. Other considerations

The number of observations used to compute the annual mean is available in the RID database. If the number of observations varies within a series this information could be used to define the uncertainty around the computed mean values. This could be incorporated in a regression analysis framework. In most cases, however, the number of available observations is expected to be the same within the series and therefore this is not studied further in this report.

If there are several inputs to the same water body it could be interesting to sum the inputs together to quantify the total riverine input to this water body. This can be done for the reported loads or for normalised loads. When using normalised loads it is, however, necessary to define a common reference time period to compute the mean flow on, so that normalised inputs are on a comparable level. Even if common reference time series are defined it must be kept in mind that countries bordering to the same water body can have large geographical distances and therefore can have different meteorological conditions and therefore reference time series, again, need to be chosen to be as long as possible to receive an average flow that is representative.

7. Conclusions

RID annual load data is reported by participating countries since 1990. The loads are computed from observed (monthly or seasonal) concentration and weighted by observed flow rate. Therefore annual loads are often highly dependent on the predominant flow processes. If the goal of an analysis is to compare input to sea areas over time the interannual variation due to changed flow is not interesting and should be removed. This process is called normalization. Normalisation is often based on linear regression model assuming a linear relation between loads and flow and such models have shown to work well for RID data. The exact form of the normalisation model is often not important as long normalization is conducted. It is, however, important to be aware of the influence of individual deviating observations on the model.

Normalised data is further analysed to compare input from rivers or to sea areas over time. For this trend analysis smooth trend functions should be assumed, since completely linear trends are seldom observed in time series as long as RID data. Many functions for fitting smooth trends have also the property that they simplify to linear trends if that is the best fitting function.

When smooth functions are fitted as trend one disadvantage is that it is not obvious how to test if the trend is significant. The question if there is a significant up- or downward trend can be answered by a follow-up non-parametric trend test, e.g. a Mann-Kendall test. Together with the result from the Mann-Kendall test an estimate for the Theil-Sen's slope is often received and can be used to quantify

the median change per year. Another possibility is to estimate the percentage change in levels for the last year compared to levels during the first year. Percentage change estimation is clearly very dependent on which levels are observed during the first year and therefore it needs to be certified that these levels are correct. A more robust version is to use modelled values for the first and last year using the trend analysis model. Even though also modelled values are influenced by outlying observations in the beginning and the end of the series these values are affected by a lesser degree.

Breakpoint analysis is sometimes suggested in connection to trend analysis. Typical breakpoint detection method search for potential break over the entire series, but have a higher probabilities to find such breaks near the middle of the series. In short series with a lot of variation, such as RID data, breakpoint detection methods can lead to either no significant results, since series are too short, or to significant breaks due to individual outlying observations. Furthermore many of the available methods only work on series without missing data. For RID data the application of breakpoint detection methods is not recommended. Instead, if there are reason to believe that the series can be split in two parts due to added or removed point sources, changes in analysis methods or other drastic changes in the series, normalisation and trend analysis should be done separately for the two parts.

8. References:

Cleveland, W., Devlin, S., 1988. Locally Weighted Regression - an Approach to Regression-Analysis by Local Fitting. J. Am. Stat. Assoc. 83, 596–610. doi:10.2307/2289282

Gilbert, R.O., 1987. Sen's Nonparametric Estimator of Slope, in: Statistical Methods for Environmental Pollution Monitoring. John Wiley & Sons, pp. 217–219.

Grimvall, A., Libiseller, C., Wahlin, K., 2009. MULTITREND. Linköpings universitet.

Hastie, T., Tibshirani, R., 1986. Generalized Additive Models. Stat. Sci. 1, 297–310. doi:10.1214/ss/1177013604

Hirsch, R.M., Slack, J.R., 1984. A Nonparametric Trend Test for Seasonal Data With Serial Dependence. Water Resour. Res. 20, 727–732. doi:10.1029/WR020i006p00727

Larsen, S., Svendsen, L., 2013. Statistical aspects in relation to Baltic Sea Pollution Load Compilation. Task 1 under HELCOM PLC-6 (No. 33). Aarhus University, DCE – Danish Centre for Environment and Energy.

Libiseller, C., Grimvall, A., 2002. Performance of partial Mann-Kendall tests for trend detection in the presence of covariates. Environmetrics 13, 71–84. doi:10.1002/env.507

Mac Nally, R., Hart, B.T., 1997. Use of CUSUM Methods for Water-Quality Monitoring in Storages. Environ. Sci. Technol. 31, 2114–2119. doi:10.1021/es9609516

Mann, H.B., 1945. Nonparametric Tests Against Trend. Econometrica 13, 245–259. doi:10.2307/1907187

McGilchrist, C.A., Woodyer, K.D., 1975. Note on a Distribution-free CUSUM Technique. Technometrics 17, 321–325.

OSPAR Comission, 2008. Comprehensive Study on Riverine Inputs and Direct Discharges (RID): Presentation and Assessment of the OSPAR Contracting Parties' RID 2006 DAta (No. 376/2008).

Pettitt, A.N., 1979. A Non-Parametric Approach to the Change-Point Problem. J. R. Stat. Soc. Ser. C Appl. Stat. 28, 126–135.

Sen, P.K., 1968. Estimates of the regression coefficient based on Kandall's tau. J. Am. Stat. Assoc. 63, 1379–1389. doi:10.2307/2285891

Stålnacke, P., Grimvall, A., 2001. Semiparametric approaches to flow normalization and source apportionment of substance transport in rivers. Environmetrics 12, 233–250. doi:10.1002/env.459

Theil, H., 1959. A rank-invariant method of linear and polynomial regression analysis. I, II, III. Nederl Akad Wetensch 53, 386–392.

Uhlig, S., Kuhbier, P., 2001. Trend methods for the assessment of effetiveness of reduction measures in the water system (UBA-FB No. 29822244), Environmental research of the federa ministry of the envionment, nature conservation and nuclear safety. Federal Environmental Agency, Berlin.

von Brömssen, C., Fölster, J., Futter, M., McEwan, K., 2017. Statistical models for evaluating suspected artefacts in long term environmental data. manuscript.

Wood, S., 2006. Generalized Additive Models An Introduction with R. CRC Press, Hoboken.

Appendix A: R code for the extraction and analysis of RID data

Extraction of data from the RID database

Data from the RID database are saved as text files, containing loads and flow data respectively. Files available are called:

for loads: RIDdatabase_utdragnovember2016.txt

for flow: Flow rate november2016.txt).

```
RID <- read.table("Z:\my
documents\Projekt\HaV\faktablad\Analysis\RIDdatabase_utdragnovember2016.txt
",sep="\t", na.strings="", header=TRUE)

RIDflow <- read.table("Z:\\my
documents\\Projekt\\HaV\\faktablad\\Analysis\\Flow rate
november2016.txt",sep="\t", na.strings="", header=TRUE)
```

In the database values for the mean, as well as lower and upper levels are stored as text variables and need therefore be converted to numerical values before analysis

```
RID$mean1<-as.numeric(as.character(RID$mean))
RID$lower1<-as.numeric(as.character(RID$lower))
RID$upper1<-as.numeric(as.character(RID$upper))
RIDflow$flow1<-as.numeric(as.character(RIDflow$flowRate))
```

Data stored in the RID database is ordered hierarchically from Country to 'main area' to individual rivers. Both sea areas and rivers from different countries can have the same areaID. To extract relevant data it is therefore necessary to select the country and within the country the correct areaID and 'Input table', where available input tables are

Table 5a: Sewage effluents to the maritime area
Table 5b: Industrial effluents to the maritime area
Table 5c: Total direct discharges to the maritime area
Table 6a: Main riverine inputs to the maritime area
Table 6b: Inputs of tributary rivers to the maritime area
Table 6c: Total riverine inputs to the maritime area (For France: Table 6c represents unmonitored areas)

The examples here have no connection to the analysis presented in the main text except that the same script was used for those analysis.

Example 1: Extracting total phosphorus loads (determinantID=12) for sea area 183 in the UK

```
RID_183_TP<-subset(RID, determinandID==12 & country== "UK" & areaID
== 183 & Input_Table =='6c')
```

Extracting flow data for the same area:

```
RIDflow_183<-subset(RIDflow, country =="UK" & areaID == 183)
```

31

In case mean values are missing in the database, but lower and upper bounds are given the mean is computed as the mean of lower and upper

```
RID_183_TP$mean1[is.na(RID_183_TP$mean1)]<-
(RID_183_TP$lower1[is.na(RID_183_TP$mean1)]+RID_183_TP$upper1[is.na(
RID_183_TP$mean1)])/2
```

Load and flow data is combined

```
RID_183<-merge(RID_183_TP, RIDflow_183, by.x="year", by.y="year",
sort=TRUE, all=TRUE)
```

Only relevant variables are kept in the dataset: year, mean load, flow, lower and upper bound and the dataset is renamed to RIDn (the code after renaming the data set can be used for any sea area or river).

```
RID_183 <- subset(RID_183, select = c(year, mean1, flow1, lower1,
upper1))
```

Example 2: Extracting total nitrogen loads (determinantID=11) for river 98 in Sweden):

Data for a river are extracted by determining the areaID for the river and selecting Input Table 6a. The remaining part corresponds to the selection of data for a sea area.

```
RID_98_TN<-subset(RID, determinandID==11 & country == "Sweden" &
areaID == 98 & Input_Table =='6a')
```

```
RIDflow_98<-subset(RIDflow, country == "Sweden" & areaID == 98)
```

```
RID_98_TN$mean1[is.na(RID_98_TN$mean1)]<-
(RID_98_TN$lower1[is.na(RID_98_TN$mean1)]+RID_98_TN$upper1[is.na(RID
_98_TN$mean1)])/2
```

```
RID_98<-merge(RID_98_TN, RIDflow_98, by.x="year", by.y="year",
sort=TRUE, all=TRUE)
```

```
RID_98 <- subset(RID_98, select = c(year, mean1, flow1, lower1,
upper1))
```

Plotting data

The following code can be used for any river of sea area. Therefor the dataset is renamed to the general RIDn. The first part produces plots as presented in Figure 1 and 4.

```
RIDn<-RID_183
```

or

```
RIDn<-RID_98
```

Plots over total loads and flow and their relationship are made:

```
par(lend=1)

par(mfrow=c(2,1))

par(mar=c(5,4,2,4))

plot(RIDn$year, RIDn$mean1, xlab="Year", type="h", lwd=10,
ylab="Total phosphorus", xlim=c(1989, 2015), col="chocolate3")

par(new=T)

plot(RIDn$year, RIDn$flow1, axes=F, ylab="", type="l", lwd=4,
xlim=c(1989,2015),xlab="", col="blue")

axis(side = 4, line=0, col="blue", lwd=2, col.axis="blue")

mtext(side = 4, line=2, 'Flow', col="blue" )

plot(RIDn$flow1, RIDn$mean1, xlab="Flow Rate", ylab="Total
phosphorus", cex=1.2, pch=16 )
```

Removal of deviating observations or subsetting a series

If some observations should be removed at this stage of the analysis it can easily be done making a subset of the data, e.g. by removing the year 1990-1992.

```
RIDn<-subset(RIDn, year!=1990 & year!=1991 & year!=1992)
```

Flow normalisation and trend analysis

For the normalisation model the mgcv package is used and for computing Mann-Kendall tests the rkt package is used. Both packages need to be installed in R (Packages -> Install packages) and invoked:

```
library(mgcv)

library(rkt)
```

The flow reference value is needed to bring normalised values on the correct level. Here the mean of the available flow data for this series is used.

```
flowmean<-mean(RIDn$flow1, na.rm=T)
```

The normalisation model is given in mod_G with a smooth trend and a linear relationship to flow. Normalised values are computed from the model:

```
par(mfrow=c(1,1))

mod_G<-gam(mean1~s(year)+flow1, data=RIDn)

RIDn$normalised_G<-RIDn$mean1-
mod_G$coefficients["flow1"]*RIDn$flow1+mod_G$coefficients["flow1"]*f
lowmean
```

33

A plot over original and normalised values including a trend line is produced by the following code:

```
plot(RIDn$year, RIDn$mean1, xlab="Year", ylab="Observed and
normalised total phosphorus", xlim=c(1989, 2015), cex=1.2, pch=16)

points(RIDn$year, RIDn$normalised_G, xlab="Year", cex=1.5, pch=18,
col="red")

lines(RIDn$year,predict(mod_G,
newdata=data.frame(year=RIDn$year[RIDn$flow1!="NA"],
flow1=flowmean)), col="red")
```

The Mann-Kendall test is computed on the normalised values by:

```
rkt(RIDn$year, RIDn$normalised_G)
```

Computation of percentage change

Percentage change is computed using normalised (pcnorm) and modelled (pcmod) values. For example, if a series is observed between 1993 and 2014:

```
startyear=1993

endyear=2014

start<-RIDn$normalised_G[RIDn$year==startyear]

end<-RIDn$normalised_G[RIDn$year==endyear]

pcnorm=100*(end-start)/start
```

```
start<-predict(mod_G, newdata=data.frame(year=startyear,
flow1=flowmean))

end<-predict(mod_G, newdata=data.frame(year=endyear,
flow1=flowmean))

pcmod=100*(end-start)/start
```

Write:

```
pcnorm
```

```
pcmod
```

to print the results to the console.