STATISTICAL ASPECTS IN RELATION TO BALTIC SEA POLLUTION LOAD COMPILATION

Task under HELCOM PLC-8 project

Technical Report from DCE – Danish Centre for Environment and Energy No. 224 2021

AARHUS UNIVERSITY
DCE – DANISH CENTRE FOR ENVIRONMENT AND ENERGY

STATISTICAL ASPECTS IN RELATION TO BALTIC SEA POLLUTION LOAD COMPILATION

Task under HELCOM PLC-8 project

Technical Report from DCE – Danish Centre for Environment and Energy No. 224 2021

Søren E. Larsen[1]
Lars M. Svendsen[2]

[1]Aarhus University, Department of Ecoscience
[2]DCE - Danish Centre for Environment and Energy

AARHUS UNIVERSITY
DCE – DANISH CENTRE FOR ENVIRONMENT AND ENERGY

Data sheet

Series title and no.:	Technical Report from DCE – Danish Centre for Environment and Energy No. 224
Title:	Statistical aspects in relation to Baltic Sea Pollution Load Compilation
Subtitle:	Task under HELCOM PLC-8 project
Authors:	Søren E. Larsen[1] & Lars M. Svendsen[2]
Institutions:	Aarhus University, [1]Department of Ecoscience & [2]DCE - Danish Centre for Environment and Energy
Publisher:	Aarhus University, DCE – Danish Centre for Environment and Energy ©
URL:	http://dce.au.dk/en
Year of publication:	December 2021
Editing completed:	November 2021
Referees:	Henrik Tornbjerg, Department of Ecoscience, Aarhus University
Quality assurance, DCE:	Jesper R. Fredshavn, DCE - Danish Centre for Environment and Energy, Aarhus University
Financial support:	Helsinki Commission (HELCOM) Baltic Marine Environment Protection Commission has kindly provided funding
Please cite as:	Larsen, S.E, & Svendsen, L.M. 2021. Statistical aspects in relation to Baltic Sea Pollution Load Compilation. Task under HELCOM PLC-8 project. Aarhus University, DCE – Danish Centre for Environment and Energy, 60 pp. Technical Report No. 224 http://dce2.au.dk/pub/TR224.pdf
	Reproduction permitted provided the source is explicitly acknowledged

Abstract:	HELCOM periodic pollution load compilation (PLC) assessments report status and development in total annual runoff and total annual waterborne and airborne nutrient inputs to the Baltic Sea. This report deals with statistical methods for assessing and evaluating time series of annual runoff and nutrient inputs. Methods included are hydrological normalization of nutrient time series, trend analysis, change point analysis and a method for testing fulfilment of HELCOM Baltic Sea Action Plan (BSAP) nutrient reduction targets (MAI and NIC). Further is described how to fill in data gaps and to estimate the total uncertainty in nutrient inputs. These statistical methods are also included in the revised PLC guidelines.

Keywords:	Statistical methods, nutrients, runoff, hydrological normalization, trend estimations, uncertainty of nutrient inputs, data gaps, outliers, trend analysis, estimation of changes, evaluation off fulfilling BSAP reduction targets, HELCOM
Layout:	Graphic Group, AU Silkeborg
Front page photo:	Pixabay
ISBN:	978-87-7156-642-0
ISSN (electronic):	2244-999X
Number of pages:	60
Internet version:	The report is available in electronic format (pdf) at http://dce2.au.dk/pub/TR224.pdf

Contents

Summary

The HELCOM project for the Eight Baltic Sea Pollution Load Compilation (PLC-8) includes the tasks of preparing a comprehensive assessment of the water- and airborne inputs of nutrients and their sources to the Baltic Sea during the monitoring period of 1995-2022. The project includes a follow up on the implementation of the HELCOM nutrient reduction scheme and assessment of the effectiveness of measures to reach the BSAP targets.

The tasks in the PLC-8 project call for standardized and appropriate methodologies, especially statistical methods. Statistical methods, in the PLC-8 project, are used for testing and estimating the trends in nutrient input time series including test for change points, and to give an estimate of the uncertainty in nutrient inputs datasets. Statistical methods are used to evaluate the progress of fulfilling the HELCOM BSAP nutrient reduction targets (Maximum Allowable Inputs (MAI)) and national Nutrient Input Ceilings (NIC) taking into account the uncertainty and variation of the nutrient inputs over time.

This is the fourth version of the collection of statistical aspects and methods first compiled as a part of the former HELCOM PLC-6 project. The report presents the statistical methods in a theoretical mathematical setup supplied with numerous examples of their use on nutrient inputs to the Baltic Sea. The methods included: testing for data outliers and filling in data gaps, hydrological normalization of time series of nutrient inputs, estimating uncertainty of nutrient inputs, trend analysis and modelling of the normalized inputs over time, testing for change points, and finally testing for fulfillment of the reduction targets (MAI and NIC).

Compared to the former version of the report, this new version includes a new method for normalization based on statistical methods from time series analysis, revision of the estimation of uncertainty, and updated descriptions of all the other statistical methods. Further, data examples are updated. Finally, the report now includes an annex with suggestions and examples of how to implement the statistical methods in the software R and SAS.

Sammenfatning

HELCOM projektet vedrørende ”Eight Baltic Sea Pollution Load Compilation (PLC-8)” indeholder en række opgaver i forbindelse med af udarbejde en omfattende vurdering af input af næringsstoffer via vandløb og luft og deres kilder til Østersøen i moniteringsperioden 1995-2022. Projektet indeholder en opfølgning af implementeringen af HELCOM’s næringsstof reduktionsmål og en evaluering af effektiviteten af foranstaltningerne til at nå BSAP målene.

Opgaverne i forbindelse med PLC-8 projektet kræver standardiserede og relevante metoder, især herunder statistiske metoder. De statistiske metoder i PLC-8 projektet anvendes til at teste og estimere udviklingstendenser i tidsserier med næringsstof input og de indeholder endvidere til test for changepoints og til at estimere usikkerheden i næringsstof datasættene. De statistiske metoder skal også anvendes til at evaluere fremskridt mod opfyldelse af HELCOM’s BSAP reduktionsmål for næringsstoffer (henholdsvis de maksimale tilladte udledning per Østersø havområde = Maximum Allowable Inputs (MAI) og de nationale udledningslofter til havområderne = Nutrient Input Ceilings (NIC)), hvor man inddrager usikkerheder og variationen over tid af næringsstof inputtet.

Dette er den fjerde version af samlingen af statistiske metoder og aspekter, som blev udgivet første gang i forbindelse med det tidligere HELCOM projekt PLC-6. Rapporten præsenterer de statistiske metoder i et teoretisk matematisk set up suppleret med adskillelige eksempler på anvendelsen af disse metoder på næringsstof input til Østersøen. Metoderne inkluderer: testning af outliers og udfyldning af data huller, hydrologisk normalisering af tidsrækker af næringsstof input, estimering af usikkerhed på input, test for udviklingstendenser og modellering af de normaliserede værdier over tid, testning af changepoints og endelig testning af opfyldelse af reduktionsmålene (MAI og BIC).

Sammenlignet med den seneste version af rapporten så indeholder denne nye version en nyudviklet metode til normalisering som er baseret på statistiske metoder fra teorien om analyse at tidsrækker, en revision af hvordan usikkerhed estimeres og opdateret beskrivelser af alle de andre statistiske metoder. Endvidere er alle dataeksemplerne opdateret og slutteligt indeholder rapporten nu et anneks som indeholder forslag og eksempler på hvordan de statistiske metoder kan implementeres i de to statistiske software: R og SAS.

1. Introduction

One of the key pressures related to the eutrophication and quality of the water of the Baltic Sea is waterborne (and airborne) nutrient inputs. In the Baltic Sea Action Plan from 2007 (BSAP (2007)), eutrophication targets were set, and based on these preliminary maximum allowable inputs, country-allocated nutrient reduction targets were developed and adopted. In HELCOM Copenhagen Ministerial Declaration from 3. October 2013, Contracting Parties decided on revised nitrogen and phosphorus input reduction targets.

In order to implement the following commitments:

- In Article 3 and Article 16 of the Convention on the Protection of the Marine Environment of the Baltic Sea Area, 1992 (Helsinki Convention)
- Baltic Sea Action Plan (BSAP), HELCOM Ministerial Meeting, Copenhagen, Denmark, 3. October 2013
- The HELCOM Monitoring and Assessment Strategy, HELCOM Ministerial Meeting, Copenhagen, 3. October, Denmark.

HELCOM 41-2020 approved the HELCOM project for the Eight Baltic Sea Pollution Load Compilation (PLC-8) (Outcome HOD 41-2020 item 5.12).

The overall task of the PLC-8 project is to prepare a comprehensive assessment of the water- and airborne inputs of nutrients and selected hazardous substances and their sources to the Baltic Sea during the period 1995-2022 including follow up implementation of the HELCOM nutrient reduction scheme and assessment of the effectiveness of measures to reach the BSAP targets. The PLC-8 project is organized in working packages with the following tasks:

1. Data reporting and establishing datasets
    a) Monitoring and reporting of national annual / periodical data.
    b) Annual updating PLC-Water database and data on atmospheric inputs (PLC-Air).
    c) Establishing the periodic assessment data sets.
    d) Update of background information including information on measures.
2. Assessments based on annual data
    a) Annual BSEF (Baltic Sea Environmental Factsheet) on actual waterborne nutrient inputs.
    b) Update of HELCOM indicator on inputs of nutrients to the Baltic Sea sub-basins.
    c) Assessment of the progress towards national nutrient input ceilings (NIC).
    d) Assessment of nutrient inputs of big rivers.
    e) Assessment of inputs of selected hazardous substances.
3. Assessments based on periodic data
    a) Assessment of sources of nutrients.
    b) Assessment of the effectiveness of measures.
4. Methodological support
    a) Updating guidelines and a statistical methodology report.
    b) Intercalibration on heavy metals and nutrients analysis.

All these tasks and objectives call for a standardized and appropriate methodology, including statistical methods. Statistical methods are used e.g. when assessing trend in nutrient input time series including test for change points, to identify the extent of trends, and estimate uncertainty in nutrient inputs datasets. Further statistical methods are needed for the evaluation of progress fulfilling HELCOM BSAP nutrient reduction targets (Maximum Allowable Inputs (MAI) and national Nutrient Input Ceilings (NIC) taking into account the uncertainty on nutrient inputs. The statistical methods are supporting tools in the PLC assessments to allow the most qualified decisions to be made regarding possible trends and acceptance of fulfilling reduction requirements.

The report “Statistical aspects in relation to Baltic Sea Pollution Load Compilation” (Larsen & Svendsen, 2013) was developed as a part of the former HELCOM PLC-6 project. During the first follow-ups of progresses towards fulfilling BSAP nutrient reduction scheme and the PLC-6 assessment, the statistical methods has been further developed, and an updated report was developed under the PLC-7 project (Larsen & Svendsen, 2019). A further development have taken place under the PLC-7 assessments, with some revised statistical methodology and new methods included in the present revised version of the 2019 report.

This report describes and includes a theoretical treatment of the statistical methods to be applied in the PLC assessments and BSAP nutrient reduction scheme. Focus points are assessment of waterborne input – and for the follow up of BSAP nutrient reduction scheme evaluating Contracting Parties fulfilment of Maximum Allowable Inputs (MAI) and Nutrient Input Ceilings (NIC) (e.g. (HELCOM, 2020) and (Svendsen et al., 2020)). The described methods include flow normalization of nutrient inputs, filling in data gaps, testing for outliers, trends and change points, calculating changes in inputs in a time series, estimation of dataset uncertainty, and, finally, how to test whether reduction targets are fulfilled taking into account the uncertainty on inputs. Examples of quantified uncertainty on riverine inputs are also included. A brief summary of the statistical methods is included in the PLC guidelines (HELCOM (in prep.)). The statistical methods are also applicable for hazardous substances inputs.

The statistical procedure for analyzing trends in the normalized nutrient input values plays an important role in the pollution input compilation assessments. The preparation of the data for trend analysis should include an assessment of the data quality, and this report includes proposals on how to fill in gaps / missing data in input time series and how to test for outliers in the data (chapter 2).

Furthermore, a study of the variability in the data sets behind the time series is important for assessing the size of the different components of variance. If some components can be reduced, the trend analysis will be more precise, and chapter 3 includes and discusses methods to estimate variance components and total uncertainty.

A final step in the preparation of the data is hydrological normalization of the yearly inputs in order to remove some of the effects of climate in the trends and to smooth out the input time series. This is described in chapter 4.

A number of different trend analysis methods, both non-parametric and parametric, exist. In HELCOM nutrient input assessments, the non-parametric method based on Kendall's tau has been used as a first step in detecting and testing for trends in the first MAI and NIC (CART country allocated reductions requirement) assessment. This method is known as the Mann-Kendall's trend test. Trend methods are described in chapter 5. In the more recent MAI and NIC assessments the Mann-Kendall trend method is only used as a preliminary tool analyzing possible trends in the TN and TP input time series and for analyzing possible trends in runoff time series. ***Otherwise trend analysis, as estimating trend line (slope, intersect etc.), and fulfillment of MAI and NIC is based on linear regression and parametric testing.***

In chapter 6 the method used, e.g. in (HELCOM, 2020) and (Svendsen et al., 2020), assessing and testing fulfilment of BSAP nutrient reduction targets (fulfilment and MAI and NIC) is presented. The chapter also includes a definition of a traffic light system for inputs to determine for which Baltic Sea main basins Contracting Parties (or catchments) MAI or NIC are:

1) fulfilled

2) not possible to judge if they are fulfilled due to statistical uncertainty

3) not fulfilling maximum allowable inputs (MAI) or nutrient inputs ceilings country per basin (NIC).

In chapter 7, we illustrate the proposed methods by a step-by-step analysis of real input data from the PLC water database to exemplify the practical use of some of the proposed methodologies.

In a concluding chapter, chapter 8, we discuss the different methods presented for normalizing, trend testing and estimating variance components, filling gaps, and testing the fulfilment of reduction targets. We provide recommendations on which methods to use for the different statistical tasks involved in preparing pollution load compilations.

The report includes Annex 1 with an in-depth mathematical treatment of the Mann-Kendall trend test, Annex 2 with selected percentiles of the t-distribution for different combinations of degrees of freedom, and Annex 3 with examples of SAS and R programs applied for some of the statistical analysis.

Mathematical symbols are defined and described in the relevant sections of the report.

The report includes several changes compared with the former version (Larsen & Svendsen, 2019):

- Chapter 1: Minor updates of text including the main aims of PLC 8 project
- Chapter 2: Example on applying Dixons outliers test have been included
- Chapter 3: Major revision including extended paragraphs on how to calculate bias and precision and total uncertainty and including examples. Extended examples on using Harmels formula on Danish data. Included several tables with examples based on PLC data on

bias and precision on TN or TP annual input data for different catchment sizes for catchment dominated by bedrock and soil, respectively, that countries can use.

- Chapter 4: It is clarified how the present used normalization methodology are used in the MAI and NIC assessments. A revised normalization method has been added, which will be tested before decision of its use in future MAI and NIC assessments.
- Chapter 5: All figures updated based on PLC data 1995-2018 and extra examples included. Revised and rather extended description of change point (break point) methodology and testing for trend, including describing trend detection when the time series are segmented due to change points, to describe how the latest assessment of MAI and NIC evaluations have been performed.
- Chapter 6: Figures and tables updated using 1995-2018 data. Text slightly revised/extended.
- Chapter 7: Extended with new examples on how to apply the revised normalization methodology. Includes new detailed examples on using the statistical methodologies.
- Chapter 8: Discussion and recommendation slightly updated and extended to reflect changes in the report.
- Chapter 9: Some references added
- Annex 3: New annex including several examples of R procedures and SAS programs (script) for the different statistical test mentioned in the report allowing PLC IG to make some of these analysis in their national data.
- Reference list updated.

The authors want to express their gratitude for the funding provided by Helsinki Commission and Aarhus University. Further, we want to thank participants of the PLC-8 implementation group, HELCOM PRESSURE and HELCOM RedCore DG for comments and inputs, and HELCOM Secretariat for help and support.

2. Data gaps and outliers

The reliability and power of all statistical methods and statistical analyses, for example normalization of time series of nutrient inputs and trend analysis of the resulting time series are greatly enhanced when conducting an initial inspection and analysis of the data quality. In general, data quality is ensured by checking the data for gaps, i.e. missing values, and for suspect values, i.e. outliers. When investigating suspicious values, the data are checked for analytical errors or errors in the data storing process, for consistency with previously reported data and with data from other comparable sources, and for errors when transferring data between databases.

A first task in the establishment of a data quality routine is the precise identification of gaps in the dataset (which variables are missing and what is the length of the missing period or periods?), followed by a determination of the type of gap or gaps (not measured, measured but not reported, etc.). Data gaps in time series on nutrient input may occur for a number of different reasons:

- Measurements are missing from the catchment or a sub-catchment for certain periods of time.
- Measurements of nutrient concentrations are missing.
- Runoff has not been measured.
- Nutrient input and runoff data are both missing for a certain period or periods.
- Measurements could not be made due to external conditions (e.g. ice cover) or flooding.
- Data have not been reported for unknown reasons.
- Concentrations and/or runoff values were evaluated as suspicious and have therefore been omitted from the calculation of inputs by the data provider and alternative inputs have not been estimated.

Several different general methods are available for filling in data gaps. Depending on type, any of the following methods can be applied to fill in the gaps:

- The mean value of a statistical distribution. The distribution is determined either by including all relevant data on the given catchment or from a shorter time series, for instance when estimating missing data from point sources in the beginning or end of a time series.

- The mean of adjacent values. If x_a and x_c are perceived as two time series values with x_b missing, then:

$$x_b = \frac{x_a + x_c}{2} \tag{2.1}$$

- Linear interpolation. If x_a and x_b are perceived as two adjacent values to n missing values, then the k^{th} missing value (from x_a) can be estimated as:

$$x_k = x_a + k \cdot \frac{x_b - x_a}{n+1} \tag{2.2}$$

- If runoff (q) is known and a good relationship can be established between nutrient input and runoff, this can be used to estimate missing values.

- A *q-q* relationship can be used to estimate missing runoff values; a good *q-q* relationship can often be established from a nearby river.

- A load-load relationship with another river for which high correlation can be verified.

- Model estimations of unmonitored catchment inputs, if possible – otherwise, inputs can be estimated from data from neighboring catchments with similar conditions.

- Assignment of a real value in the interval between zero and the limit of detection (LOD)/limit of quantification (LOQ) to observations below a limit of detection/limit of quantification. The PLC guidelines (see chapter 1) describe how to handle concentrations under LOD/LOQ when calculating loads.

Most methods for trend analysis, like the non-parametric Mann-Kendall's trend method and linear regression (see chapter 5), can handle missing values, preferably in the middle and not at the end of the time series (e.g. either the first 2 or the last 2 years). The trend test will only be negligibly affected with few missing values. The statistical power of the trend tests decreases if the time series includes gaps, as it is more difficult to prove a real significant trend at reduced statistical power. If many missing values have been estimated and the inserted values are identical for many years, a trend test should not be performed, as variation will be much smaller than when the data are based on real observations.

Above, various methods for filling in gaps have been described. Usually, the circumstance will decide which method to choose, but the following rank is suggested:

1. A model approach – i.e. a regression type model – to estimate nutrient load or flow.
2. Linear interpolation.
3. Values from a look-up table or values provided by experts.
4. No filling in of gaps. The time series is used as it is and assessments are made afterwards.

Outliers are data values that are extreme compared to other reported values for the same locality (country, basin, catchment, etc.) and can only be determined and flagged by conducting a formal outlier test using for instance:

- Dixon's 4 sigma (σ) test: Outliers are the values outside the interval consisting of the mean ±4 times the standard deviation.
- A box and whisker diagram.
- Experience-based definition of maximum and minimum values that is not likely to be exceeded or fallen below.
- Water quality standards (interval values or limits), if available.

It is important to note that outliers are not necessarily faulty data, but data requiring extra careful evaluation prior to be used in statistical analyses.

Suspect or dubious values are values that do not fulfill the requirement of being determined as a formal outlier by the outlier tests but differ significantly from the remaining values in the time series, or values that are unreliable; for instance, a load value for the reported runoff or data from a neighboring catchment. Suspect or dubious values may occur if measurements in a sub-catchment have been made for only a limited period, if changes in laboratory or laboratory standards have occurred, or if changes have been made in other measurement methods, resulting in an abrupt change in data values. Or, if the same value occurs for a number of successive years. In addition, calculation mistakes may occur due to use of wrong units, faulty water samples, laboratory mistakes, etc. Suspect or dubious values should be corrected and treated as a formal outlier unless they can be proven correct.

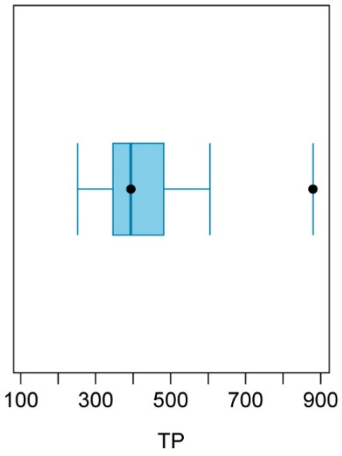


Figure 2.1. Box and whisker plot of the series of TP loads (tons) in the Pregolya River during 1995-2018.

If a dubious value is determined, deemed to be wrong and omitted from assessments, and if it is not possible for the Contracting Party to correct the value, it should be removed from the PLC database by the Contracting Party. If a reported data value is determined to be an outlier and deemed to be omitted from assessments, the outlier can be replaced in the assessment using a method from the list on data gaps. Usually, filling in data gaps or replacing suspect data cannot substitute measured data; thus, if possible, preferably measured or consistent model data should be found and used. It should be stressed that filled-in data gaps must be clearly marked in the PLC database.

In order to illustrate the use of the Dixon outlier test, we consider the TP loads from the Pregolya River (Russia). I figure 2.1 we show a box and whisker plot of the TP loads during the period 1995-2018. Figure shows one outlier at a value of 880 for TP. The Dixon test in R shows that the value in 2017 (880 tonnes TP) is determined as an outlier in the series of yearly TP load values (with a significance value of P=0.007). The series of TN loads during the same period in the Pregolya River have no outliers.

3. Uncertainty of inputs (yearly input from a specific country or area)

Time series of nutrient inputs demonstrate a certain amount of year-to-year variation due to the contributions from a large number of different components. One such component is a possible trend in inputs over time, and time series are therefore, by standard, detrended before analysis of the variance components from other sources, since trend-induced variations are not of basic interest in estimating the variance components. Detrending means that a significant trend in the time series data is removed before calculating the uncertainty. This can be done in different ways; one is to calculate the residuals between the observed values and the predicted values using a model for the trend, e.g. a linear trend model.

Variation appears within the yearly values – and it is thus assumed that the yearly inputs are sampled from the same population of inputs with a given mean value and a given variation (after detrending). This variation is, in fact, an estimate of a part of the total uncertainty of a given yearly input, i.e. the standard error of the mean. The other part is a possible bias in the calculation of the yearly loads. This part is in principle not known from the values of the time series.

Total uncertainty (bias plus variation) is an extremely complex sum (based on certain assumptions) of a number of different uncertainty components:

- Uncertainty due to field sampling (uncertainty from field sampling / measurements of concentrations of nutrients, metals and other substances, uncertainty from measurements of water velocity and stage, etc.).
- Laboratory uncertainty (from the applied analysis method in the laboratory or from changing laboratories over time).
- Uncertainty deriving from the sampling set-up (how often, where and when, sampling location, time) and the methods for calculating runoff (either stage-discharge relationship or other methods) and load (based on combined concentrations and runoff).
- Variation introduced by year-to-year differences in climate (amount, type, and distribution of rainfall and changes in accumulated pools (snow / ice, soil and groundwater)).
- Uncertainty from estimation of unmeasured inputs (bias from omitting unmeasured inputs and uncertainty of the methods applied for estimating unmonitored inputs).
- Uncertainty of inputs from direct point sources, including sampling, analytical errors, etc.
- And probably, several other components contributing to uncertainty.

Awareness exists in most countries of analysis (laboratory) uncertainty, at least regarding nutrients. This is relatively well documented but may not be one of the components contributing the most to total uncertainty. Most other components are complex, and some of them are very difficult to estimate in practice due to unavailability of empirical data. Uncertainty can be diminished by optimizing, for instance, time and location of sampling and implementation of a monitoring program taking into account variations in concentrations and runoff. An optimized monitoring program may introduce more strategic monitoring and more precise and modern techniques as well as an optimized methodology for estimating inputs from unmonitored areas, strategic measuring being most important factor to decrease uncertainty.

Knowing the size of the different uncertainty components is not necessary when testing for trends and for compliance with set targets. Variance component analysis is used in statistics when the researcher seeks to optimize the sampling design in a hierarchical sampling regime and / or to test for effects (treatment, emission reducing measures or other factors) using the correct sums of squares.

In the PLC assessments, it would be useful to compare the total uncertainty of detrended nutrient input time series among countries, among sub-basins, etc., to determine if time series have the same level of uncertainty or if some countries, sub-basins, etc., have significantly lower or higher uncertainties. Investigation of the size of the different variance components would be highly useful for determining the reasons for the differences. The main result of such an exercise would be an overall improved data quality with more complete and consistent data sets from all Contracting Parties.

All Contracting Parties are requested to submit estimates of uncertainties for yearly inputs of TN and TP as well as for yearly runoff values, but it is not requested to report on the individual uncertainty components listed above.

For this purpose, we need a standardized methodology for estimating the uncertainties in the national datasets from measured areas.

The calculation of the total uncertainty is done by using the statistical principle "Propagation of errors". This principle can be explained as:

Let X be the sum of n stochastically independent measured inputs X_i

$$X = \sum_{i=1}^{n} X_i. \tag{3.1}$$

All the X_i variables are considered stochastic. The variance of the sum X can be calculated as:

$$\sigma_X^2 = Var(X) = \sum_{i=1}^{n} \sigma_{X_i}^2. \tag{3.2}$$

The standard deviation is then calculated as:

$$\sigma_X = \sqrt{\sum_{i=1}^{n} \sigma_{X_i}^2}. \tag{3.3}$$

And the relative standard deviation (denoted the precision) is calculated as

$$100 \cdot \frac{\sigma_X}{X} = \frac{100}{\sum_{i=1}^{n} X_i} \sqrt{\sum_{i=1}^{n} \sigma_{X_i}^2}. \tag{3.4}$$

The calculation of the total inputs from the monitored areas constitute of measurements from n stations in streams, as defined in (3.1). The relative bias and relative precision of the sum of X_i can then be calculated as

$$bias\ (\%) = \frac{100}{\sum_{i=1}^{n} X_i} \sum_{i=1}^{n} bias_i \cdot X_i, \tag{3.5}$$

$$precision\ (\%) = \frac{100}{\sum_{i=1}^{n} X_i} \sqrt{\sum_{i=1}^{n} (precision_i \cdot X_i)^2}. \qquad (3.6)$$

Here $bias_i$ and $precision_i$ are the individual biases and precisions (given in decimal notation) for each river indexed by i. Bias is the consistent under- or overestimation of the true value of the mean for example. Precision is a measure of the size of the closeness of the individual measurement values. The total uncertainty can then be calculated as

$$uncertainty\ (\%) = \frac{100}{\sum_{i=1}^{n} X_i} \sqrt{\sum_{i=1}^{n} (bias_i \cdot X_i)^2 + (precision_i \cdot X_i)^2}. \qquad (3.7)$$

The total uncertainty is the measure of the closeness of the measurements to the true value (bias plus precision). Theoretically, the total uncertainty is the Root Mean Squared Error ($RMSE$) as defined by the equation

$$RMSE(\widehat{X}) = \sqrt{(\sigma^2 + \beta^2)}, \qquad (3.8)$$

Where

$\sigma^2 = E_X\left(\hat{X} - E_X(\hat{X})\right)^2$ = the variance (precision) and

$\beta = E_X(\hat{X}) - X$ = the bias.

The symbol E_X means the theoretical mean value with respect to the stochastic variable X, and $\hat{X}$ is the estimate of X.

This implies that the uncertainty (%) in (3.7) is the $RMSE$ (%).

Below is a small example of the use of formulas (3.5), (3.6) and (3.7) for calculating the bias, precision and uncertainty of the sum of X_1, X_2 and X_3 :

	Load	Bias	Precision
X_1	10	-5%	5%
X_2	15	-5%	8%
X_3	20	-5%	10%

Bias (%) = $100/45 \cdot (-0.05 \cdot 10 - 0.05 \cdot 15 - 0.05 \cdot 20) = -5\%$

Precision (%) = $100/45 \cdot \sqrt{((0.05 \cdot 10)^2 + (0.08 \cdot 15)^2 + (0.1 \cdot 20)^2))} = 100/45 \cdot \sqrt{(0.25 + 1.44 + 4)} = 5.3\%$

Uncertainty (%) = $100/45 \cdot \sqrt{(0.25 + 0.5625 + 1 + 0.25 + 1.44 + 4)} = 6.1\%$

One such methodology for estimating the uncertainty of data from monitored rivers has been described in a paper by Harmel et al. (2009). The method is called DUET-H/WQ, which is based on the $RMSE$ propagation method just explained. It is a fair approximation to the true value, which is, as mentioned, often very complicated to derive. In DUET-H/WQ, the total uncertainty of individual measured inputs is estimated by the formula:

$$EP = \sqrt{E_Q^2 + E_C^2 + E_{PS}^2 + E_A^2 + E_{DPM}^2}, \qquad (3.9)$$

where according to Harmel et al. (2009):

E_Q=Total uncertainty of the discharge measurement (%)

E_C= Total uncertainty of sample collection (%)

E_{PS}=Total uncertainty of sample preservation / storage (%)

E_A=Total uncertainty of laboratory analysis (%)

E_{DPM}=Total uncertainty of data processing and data management (%), i.e. input calculation or model uncertainty (see Silgram and Schoumans (ed., 2004)).

Then, the total uncertainty for aggregated data can be estimated by the formula:

$$EP_{total} = \frac{100}{\sum_{i=1}^{n} x_i} \sqrt{\sum_{i=1}^{n} \left(x_i \cdot \frac{EP_i}{100}\right)^2} \tag{3.10}$$

where EP_{total} is given as %.

EP_{total} = Total uncertainty for the sum $x = \sum_{i=1}^{n} x_i$

x_i = Yearly input from a catchment or a country.

The Contracting Parties will need to gather information on the different uncertainties, either from empirical data or from national or international papers and reports based on the same kind of data, i.e. riverine measurements based on more or less similar methods.

Furthermore, uncertainties regarding input estimates from unmonitored areas need to be described in order to estimate the total uncertainty for the whole catchment area. Uncertainty on direct inputs can be estimated using the same formula as above.

The uncertainties for many of the components listed above are not quantified or estimated, but the uncertainty on individual water flow quantifications are well known and should in most cases be lower than ± 5% (Herschy (2009) and WMO (2008)). The precision on daily water flow depends on the number of discharge observations, and is estimated for open gauging stations in streams channels in Denmark to be about 8% (given as standard deviation) with 10 annual discharge observations (measurements of discharge), about 6% with 12 measurements to less than 1% with more than 40 annual measurements (Kronvang et al. 2014). For modelled water flow, the uncertainty might be higher. For chemical analysis the requirement in Denmark is that the total (expanded) uncertainty for total nitrogen and total phosphorus is less than 15% (or 0.1 mg N l^{-1} and 0.01 mg P l^{-1} at low concentration values in freshwater, respectively 5 mg N l^{-1} and 1 mg P l^{-1} at low concentration values in waste water).

As mentioned in the beginning of this chapter, the precision may also be estimated from the variance of a time series of inputs without trends or a detrended time series. It is the standard error of the mean input throughout the period. The two estimates can then be compared. However, the last method described do not include a systematical measurement bias, e.g. in runoff or in nutrient inputs. Rather, it estimate the variation around an average value.

In a situation where the given time series of inputs show a significant positive serial correlation, the standard error is underestimated and the precision is accordingly underestimated. In this report, we assume that the serial correlation in a yearly time series of nutrient inputs is low; the basic calculation of the standard error is therefore used as a close approximation to the true value of the standard error.

The method by Harmel et al. (2009) is illustrated by two examples: 1) total uncertainty for a river with high measurement precision and 2) total uncertainty for a river with low measurement precision – see table 3.1. High measurement precision stands for a low value of formula (3.6) and vice versa.

Table 3.1. Illustration of the method by Harmel et al. (2009) with 2 examples of variance components in formula (3.1). Example 1 with low total uncertainty (river with high measurement precision) and example 2 with high uncertainty (river with low measurement precision)

Variance components	Example 1	Example 2
E_Q	5%	50%
E_C	5%	100%
E_{PS}	5%	30%
E_A	5%	25%
E_{DPM}	5%	50%

In Example 1 (table 3.1) *EP* is 11% and in Example 2 *EP* is 129% when using formula 3.9 Total uncertainty of assuming a constant monthly input of 2500 tons (x_i) is 3% for Example 1 and 36% for Example 2. Total uncertainties were calculated using formula 3.10

Another method of calculating the total uncertainty is illustrated using Danish data for total nitrogen (TN) and total phosphorus (TP) inputs to the marine areas around Denmark. The total input to the Danish marine environment is a sum of two components. One component is from the monitored catchment area and the other is from the unmonitored area. The inputs from the unmeasured area is estimated by using a model. A Monte Carlo study (Kronvang & Bruhn, 1996) based on daily samples has shown that for Danish streams categorized by their catchment area, the following values for bias and precision are valid for TN load calculated using the linear interpolation method:

0-50 km^2: Bias: -1% to -3%; Precision: 1-3%
50-200 km^2: Bias: -0.7% to -3%; Precision: 1-3%
>200 km^2: Bias: -1% to -4%; Precision: 2-5%

These number are valid for the yearly load from one stream station and include the uncertainty of laboratory analysis, yearly variation of concentrations and stream discharge and uncertainty from the method for calculating yearly load (by linear interpolation). The uncertainty from the measurement of the concentration in the stream (placement of the sampling site horizontal and vertical in the stream) is not included and therefore 2% is added to the precision in the 3 categories.

Using the formulae (3.5-3.7), it can be calculated that, for the monitored area (210 stations) the total bias is -1% to –3%, the total precision is 0.7% to 1.2% and the total uncertainty is 0.7% to 1.3%. For an average stream station the bias is -1% to -3%, the precision is 3% to 5% and the uncertainty is 3.2% to 5.8%.

The TN input from the unmonitored area is based on model estimates for 1286 very small catchments covering the rest of the Danish area (39%). The year load from each small catchment is calculated using the formula

$$L = Ndiffuse_{model} + R_{lake} + R_{stream} + N_{waste} - R_{total}, \qquad (3.11)$$

$Ndiffuse_{model}$ = the estimated nitrogen inputs from the model

R_{lake}= Estimated nitrogen retention in lakes

R_{stream} = Estimates nitrogen retention in streams

N_{waste} = Nitrogen inputs from wastewater

R_{total} = Total nitrogen retention.

In table 3.2 are shown bias and precision for the components in formula (3.11 based on both numerical calculations, the study by Kronvang & Bruhn (1996) and estimates.

Table 3.2. Bias and precision for nitrogen inputs in formula (3.11) based on both numerical calculations, estimates and Kronvang & Bruhn (1996).

Components	Bias (%)	Precision (%)
Model	15 to 20	12 to 15
Retention lake	-5 to 5	40
Retention stream	-5 to 10	40
Retention total	-5	40
Point source: Industry	-1 to -3	1 to 10
Point source: Wastewater	-1 to -3	1 to 10
Point source: Fishfarms	-1 to -3	1 to 20
Point source: Rainwater	-5	40

Using the formulae (3.5) to (3.7) and the bias and precision indicated in table 3.2 the total bias for the unmonitored area is calculated to 20% to 28%, the total precision is 0.8% to 2.0% and the total uncertainty is 1.2% to 2.2%. For an average small unmonitored catchment, the bias is 27%, precision 15% to 20% and the uncertainty 31% to 34%.

For the total Danish catchment area, combing the calculated bias, precision and uncertainty for both the monitored and unmonitored areas and using special versions of formulae (3.7) to (3.9), we get a total bias of 7.4% to 12.8%, a total precision of 0.5% to 1.1% and a total uncertainty of 7.4% to 12.8% on TN inputs.

With respect to total phosphorus (TP), calculations show that for the measured area the bias is -6 to -3%, the precision is 1 – 2% and the uncertainty is then 1 – 2.5%. For the unmeasured area the bias is between -5 and 30%, the precision is 1 – 3% and the uncertainty is 1 – 4%. These calculations are based on the following values of bias and precision from Kronvang & Bruhn (1996) for TP load (using linear interpolation method):

0-50 km2:	Bias: -16% to -27%;	Precision: 18-37%
50-200 km2:	Bias: -2% to -5%;	Precision: 9-13%
>200 km2:	Bias: -2% to -4%;	Precision: 3-8%

It is suggested to use this Danish approach estimating uncertainties for all monitored rivers discharging to the Baltic Sea in the HELCOM contracting countries.

Two methods for calculating the yearly load of nutrients are applied in these countries.

- Linear interpolation method: Denmark, Estonia, Germany and Sweden.
- Monthly mean method: Finland, Latvia, Lithuania, Poland, and Russia.

In tables 3.3-3.6 values for bias and precision are given for the two load calculation methods for different catchment sizes, number of yearly samples, catchment dominated by bedrock or dominated by soils for annual load of TN and TP, respectively.

Table 3.3. Bias and precision for yearly load based on the linear interpolation calculation method. TN (black) – TP (red). Catchment dominated by soils.

Catchment size (km^2)	Number of samples	Bias (%)	Precision (%)
0 – 50	<18	-1 → -3	2 → 5
		-17 → -25	22 → 40
0 – 50	>18	0 → -2	1 → 3
		-15 → -20	18 → 37
50 – 200	<18	-1 → -3	2 → 5
		-4 → -8	12 → 18
50 – 200	>18	0 → -2	1 → 3
		-2 → -5	9 → 13
200 – 1000	<18	-1 → -3	4 → 7
		-3 → -7	6 → 12
200 – 1000	>18	0 → -2	2 → 5
		-1 → -3	3 → 8
>1000	<18	-1 → -3	4 → 7
		-3 → -7	6 → 12
>1000	>18	0 → -2	2 → 5
		-1 → -3	3 → 8

Table 3.4. Bias and precision for yearly load based on the monthly mean calculation method. TN (black) – TP (red). Catchment dominated by soils.

Catchment size (km^2)	Number of samples	Bias (%)	Precision (%)
0 – 50	<18	-2 → -4	2 → 5
		-20 → -30	25 → 45
0 – 50	>18	-1 → -3	1 → 3
		-16 → -27	20 → 40
50 – 200	<18	-2 → -4	2 → 5
		-4 → -10	15 → 20
50 – 200	>18	-1 → -3	1 → 3
		-2 → -5	10 → 15
200 – 1000	<18	-2 → -4	4 → 7
		-4 → -8	8 → 14
200 – 1000	>18	-1 → -3	2 → 5
		-2 → -4	5 → 10
>1000	<18	-2 → -4	4 → 7
		-4 → -8	8 → 14
>1000	>18	-1 → -3	2 → 5
		-2 → -4	5 → 10

Table 3.5. Bias and precision for yearly load calculation based on the linear interpolation calculation method. TN (black) – TP (red). Catchment dominated by bedrock.

Catchment size (km^2)	Number of samples	Bias (%)	Precision (%)
0 – 50	<18	-2 → -4	3 → 6
		-20 → -30	25 → 45
0 – 50	>18	-1 → -3	2 → 4
		-17 → -24	18 → 37
50 – 200	<18	-2 → -4	3 → 6
		-5 → -10	14 → 20
50 – 200	>18	-1 → -3	2 → 4
		-3 → -6	10 → 15
200 – 1000	<18	-2 → -4	6 → 9
		-4 → -9	8 → 14
200 – 1000	>18	-1 → -3	3 → 6
		-2 → -4	4 → 10
>1000	<18	-2 → -4	6 → 9
		-4 → -9	8 → 14
>1000	>18	-1 → -3	3 → 6
		-2 → -4	4 → 10

Table 3.6. Bias and precision for yearly load based on the monthly mean calculation method. TN (black) – TP (red). Catchment dominated by bedrock.

Catchment size (km^2)	Number of samples	Bias (%)	Precision (%)
0 – 50	<18	-3 → -5	3 → 6
		-23 → -35	30 → 50
0 – 50	>18	-2 → -4	2 → 4
		-18 → -30	23 → 45
50 – 200	<18	-3 → -5	3 → 6
		-5 → -12	18 → 23
50 – 200	>18	-2 → -4	2 → 4
		-3 → -6	12 → 18
200 – 1000	<18	-3 → -5	6 → 9
		-6 → -10	10 → 16
200 – 1000	>18	-2 → -4	3 → 6
		-3 → -5	7 → 12
>1000	<18	-3 → -5	6 → 9
		-6 → -10	10 → 16
>1000	>18	-2 → -4	3 → 6
		-3 → -5	7→ 12

4. Hydrological normalization of nutrient inputs

The annual riverine inputs of nutrients show large variations between the reported years. Variation in runoff is a major reason behind this and is mainly caused by weather effects on hydrological factors such as precipitation, including accumulation and melting of snow / ice, and evapotranspiration, but also by temperature, etc. To remove the main part of the variation introduced by hydrological factors, the annual nutrient inputs are flow-normalized. Normalization should not be applied on inputs from point sources discharging directly to the sea, and it is needed to subtract major point sources from the inputs before normalization is performed on riverine loads. After normalization loads from the deducted point source are added.

Normalization of riverine loads is a statistical method whose result is a new time series of nutrient inputs where the major part of the hydrology-introduced variation has been removed. The normalized time series has a reduced between-year variation and the trend analysis is thus much more precise. Significant trends in the normalized series can probably mainly be attributed to an effect of human activities.

Different methods for normalizing inputs are described in Silgram & Schoumans (ed., 2004), chapter 4. In this report, we focus on methods based on empirical data. The empirical hydrological normalization method is based on the regression of annual loads and annual runoff; thus, the method normalizes the loads to an average runoff (averaged over the whole time series period). In this way, the variation attributable to the annual amount of runoff is removed, whereas the effect of differences in the distribution of runoff over the year is not removed.

In Silgram & Schoumans (ed., 2004), the normalization is based on un-transformed loads and runoffs. In our experience, the regression explains slightly more of the variation if both annual input and annual runoff values are transformed by the natural logarithmic function before normalizing.

The hydrological normalization should be regarded as a prerequisite for analysing trends. The trend analysis is a two-step process including: 1) the normalization and 2) the actual trend analysis.

According to Silgram & Schoumans (ed., 2004), the empirical hydrological normalization method should be based on the linear relationship between annual runoff (Q) and the annual load (L) of a nutrient:

$$L_i = \alpha + \beta \cdot Q_i + \varepsilon_i, \tag{4.1}$$

α and β = Parameters associated with linear regression

ε_i = Model error in the linear regression.

Then, the normalized load L_N is calculated as:

$$L_{Ni} = L_i - (Q_i - \bar{Q}) \cdot \hat{\beta}, \tag{4.2}$$

$\bar{Q}$ = Average runoff for the whole time series period

Q_i = Runoff in year i

^ = Indicates that it is an estimated parameter.

Normally, the relationship is modelled after log-log transformation, reducing the influence of large loads and runoff values giving, as mentioned, a slightly more precise fit with residuals that are more likely to be Gaussian distributed, which is a statistical prerequisite for the regression method. Thus, normalization should be based on a log-log regression between load and runoff:

$$\log L_i = \alpha + \beta \cdot \log Q_i + \varepsilon_i. \tag{4.3}$$

To avoid large negative values when log transforming very small load or runoff values, it is suggested to multiply load and runoff with 1000 before log transforming.

Formula 4.3 gives the following formula for normalized loads:

$$L_{Ni} = \exp\left(\log L_i - \left(\log Q_i - \overline{\log Q}\right) \cdot \hat{\beta}\right) \cdot \exp(0.5 \cdot MSE)., \tag{4.4}$$

In the above formula (4.3) and (4.4), “log” is the natural logarithmic function, “exp” is the exponential function, and MSE stands for Mean Squared Error and is derived by the regression analysis (Snedecor & Cochran, 1989). MSE is calculated in all standard statistical software programs and is in general defined as:

$$MSE = \frac{1}{n-p} \sum_{i=1}^{n} (x_i - \hat{x_i})^2, \tag{4.5}$$

n = Number of observations in the time series

p = Number of parameters in the model, in standard regression p=2

$x_{i=}$ Observed value

$\hat{x_i}$ = Modeled value from linear regression.

In this report x_i would be $\log L_i$ and $\hat{x_i}$ would be $\widehat{\log L_i}$, and log the natural logarithm function.

The factor “exp(0.5·MSE)” in the formulae is a bias correction factor and is derived as described by Ferguson (1986). The factor is needed in order to back-transform to a mean value and not to a geometric mean, whose calculation does not require this factor. If $\exp(0.5 \cdot MSE) > 1.25$, this indicate that the fit in formula (4.3) is not very good and it is probably better to use formula (4.1) and (4.2). The main reason for using the natural logarithmic function for transformation is stabilization of the variance among residuals. Without the transformation, residuals are often distributed with a heavy tail to the right.

The models presented in (4.1) and (4.3) will result in a pattern in the model residuals if the time series is non-stationary (with a trend). In time series with trends, the relationship between L and Q changes over time. Furthermore, the residuals will also be serial correlated to a strong degree. This can be seen in the following example (figure 4.1). It illustrates estimated linear relationship between diffuse TN load and discharge for the sum of all monitored Danish rivers in the period 1990-2018. It can be seen from the figure that the relationship changes from the beginning of the period to the end. In this case, the model residuals resulting in applying the model in (4.1), will start with large positive residuals decreasing more or less linearly to large negative residuals as time elapses. Indicating a poor fit of the model to the time series data.

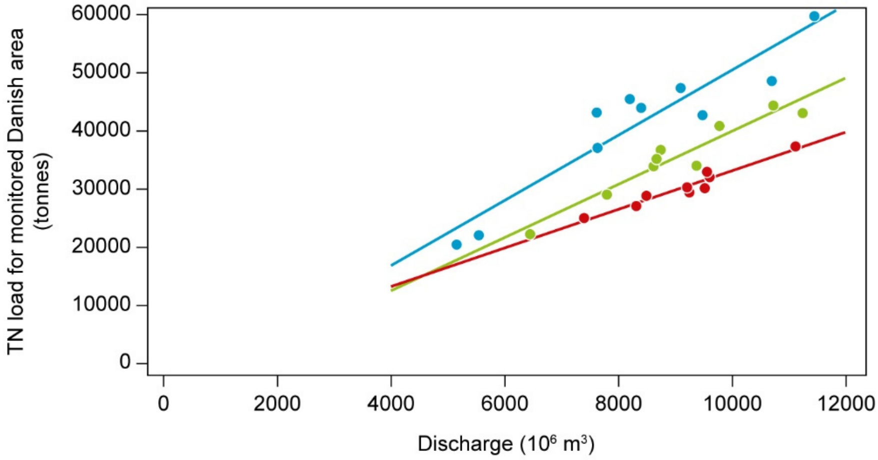


Figure 4.1. Estimated linear relationship between diffuse TN load and discharge for the sum of all monitored Danish rivers. Data is from the period 1990-2018. The total period is divided into 3 sub-periods: 1990-1999 (blue), 2000-2008 (green) and 2009-2018 (red).

In order to deal with these model problems, we introduce models and methods from the statistical theory of serial correlated time series (Box et al., 2015). We apply the method of differencing the time series, which is a transformation of the time series used to stabilize the mean of the time series. So for all years $(i, i = 2, \cdots, n)$, we calculate $L_i - L_{i-1}$ and $Q_i - Q_{i-1}$ and fit the model

$$L_i - L_{i-1} = \alpha + \beta \cdot (Q_i - Q_{i-1}) + \varepsilon_i. \tag{4.6}$$

This model, generally, results in a much more appropriate distribution of model residuals over time. The model in (4.6) can be written as $(i, i = 2, \cdots, n)$

$$L_i = \alpha + \beta \cdot (Q_i - Q_{i-1}) + L_{i-1} + \varepsilon_i \tag{4.7}$$

$$\Updownarrow$$

$$L_i = \beta \cdot Q_i + \alpha \cdot i + \frac{1}{(1-B)} \varepsilon_i \tag{4.8}$$

Where

$$\frac{1}{(1-B)} \varepsilon_i = \varepsilon_i + \varepsilon_{i-1} + \cdots + \varepsilon_1.$$

The B is the so-called back-shift operator used in time series mathematics (Box *et al.*, 2015) and it is defined as

$$(1-B)L_i = L_i - L_{i-1}$$

Differencing the data results in what is called a random walk, which in the simplest form is $x_i = x_{i-1} + \varepsilon_i$ with ε_i as a white noise, i.e. independent identical distributed as $N(0,\sigma^2)$. So the most simple random walk is a stochastic process depending of the value just before plus a random shock, a Gaussian distributed value with zero mean (=0) and a given variance (σ^2). A random walk can also be viewed as a simple autoregressive process of order 1. Differencing is equivalent to a stochastic differential equation in continuous time.

The model in (4.6-4.8) can also be used with logarithmic transformed loads and flow. Whether to transform or not can be determined by looking at the residuals of the model with untransformed data.

The normalized loads, using the model in (4.6-4.8) are calculated for untransformed data as $(i, i = 1, \cdots, n)$

$$L_{Ni} = \hat{\beta} \cdot \bar{Q} + \hat{\alpha} \cdot i + r_i = L_i - (Q_i - \bar{Q}) \cdot \hat{\beta} \tag{4.9}$$

where the r_i $(i, i = 1, \cdots, n)$ is defined as

$$r_i = L_i - \left(\hat{\alpha} \cdot i + \hat{\beta} \cdot Q_i\right). \tag{4.10}$$

It turns out that

$$r_1 = L_1 - \left(\hat{\alpha} + \hat{\beta} \cdot Q_1\right)$$

and for $(i, i = 2, \cdots, n)$

$$r_i = r_1 + \sum_{j=2}^{i} \left(L_j - \left(\hat{\alpha} + \hat{\beta} \cdot \left(Q_j - Q_{j-1}\right) + L_{j-1}\right)\right).$$

The right side in (4.9) is equal to (4.2) but remember that $\hat{\beta}$ is from another model.

And for transformed data

$$L_{Ni} = \exp\left(\hat{\beta} \cdot \log\bar{Q} + \hat{\alpha} \cdot i + r_i\right) \cdot \exp(0.5 \cdot MSE) \tag{4.11}$$

Where r_i is defined as in (4.10) but now with logarithmic transformed data.

The parameter α can almost always be tested equal to zero and is in fact associated with a trend in the size of point source inputs, so in time series without point sources this parameter can be disregarded in the model (4.6).

In the case of using the transformed data for normalization, and in the case of a non-linear normalization model, the normalized values will not in average be equal to the un-normalized values, i.e.

$\frac{1}{n}\sum_{i=1}^{n} L_i \neq \frac{1}{n}\sum_{i=1}^{n} L_{Ni}$.

So, the normalized values have to be corrected for bias. This can be done in two different ways:

$$L_{NBi} = L_{Ni} + (\bar{L} - \overline{L_N}) \tag{4.12}$$

or

$$L_{NBi} = L_{Ni} \frac{\bar{L_\iota}}{\bar{L_{\iota N}}} \tag{4.13}$$

The first way (formula 4.12) is preferred.

In the PLC-5 assessment, the method described in (4.3) and (4.4) was used (Svendsen et al., 2013).

Hydrological normalization should be carried out river / catchment-wise, i.e. nutrient loads should be normalized for each river / catchment separately. If the normalization is performed country-wise or sub-basin-wise, the result will not be exactly the same as the catchment-wise normalized nutrient loads summed to country or sub-basin level. There will be a minor difference in the results.

To illustrate the method defined in (4.6-4.11), we have used TN load data from the River Aalbek (Germany) and TP load data from the River Vistula (Poland). Figure 4.2 shows scatter plots and the linear relation between differenced loads and differenced flows. For the River Aalbek the relation is quite good, for River Vistula, there is also a clear relationship but the scatter around the fitted line is larger. Figure 4.3 shows the normalized time series together with the unnormalized loads. Note the large reduction in between-year variation in the normalized time series. As can be seen in figure 4.3(a) the difference between the new method described above (4.6-4.11) and the until now used method (Gustafson, 2019) is very small. This is due to a constant relationship between TN load and runoff over time. There is a larger difference between the two methods in figure 4.3(b). In some periods, the new introduced normalizing method gives larger values, in other periods the opposite can be noticed. This is due to changes in the relationship between runoff and TP load.

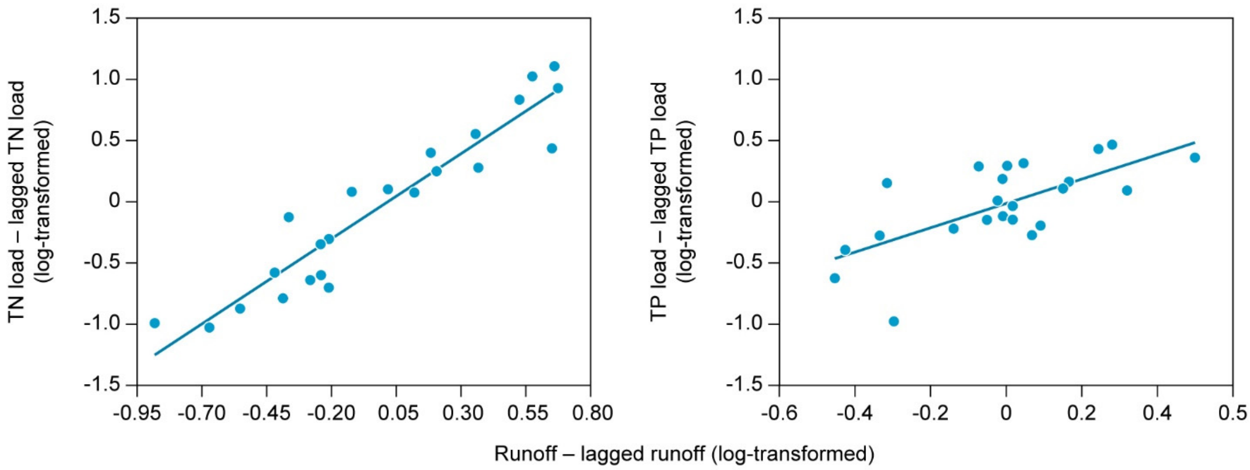


Figure 4.2. Scatter plots of annual loads of TN (a – plot to the left) in the Aalbek river and TP (b – plot to the right) in the Vistula river against runoff and the linear regressions (transformation based on natural logarithmic function). Plotted values are differences between the annual values and the annual values one year before as modeled in formula 4.6. Data represent the loads of TN and TP in the two rivers during 1995-2018.

The next figure illustrates the improvement in the distribution of the model residuals over time when applying the new introduced normalizing method on TN loads measured in the Danish river Langvad. The plot in figure 4.4a shows the distribution of the model residuals over time applying the model in (4.3). The residuals show an almost linearly trend over time from large positive values to large negative values. This pattern in the residuals illustrate a poor model fit to the data. In figure 4.4b the residuals, from applying the model in (4.7), look to be randomly distributed over time, both according to sign and size and therefore this model is a much-improved model for the data.

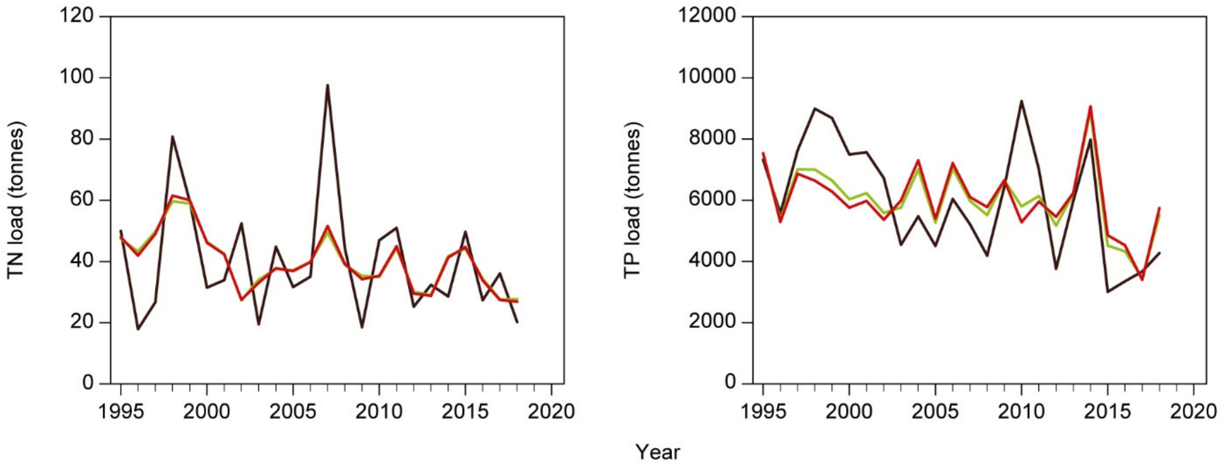


Figure 4.3. Time series plot of annual actual (not normalized) time series (black), the until now used method (red) and of normalized time series (green) with the difference method in formula 4.6 of annual TN (a, Aalbek, figure to the left) and TP (b, Vistula, figure to the right) in tonnes 1995-2018.

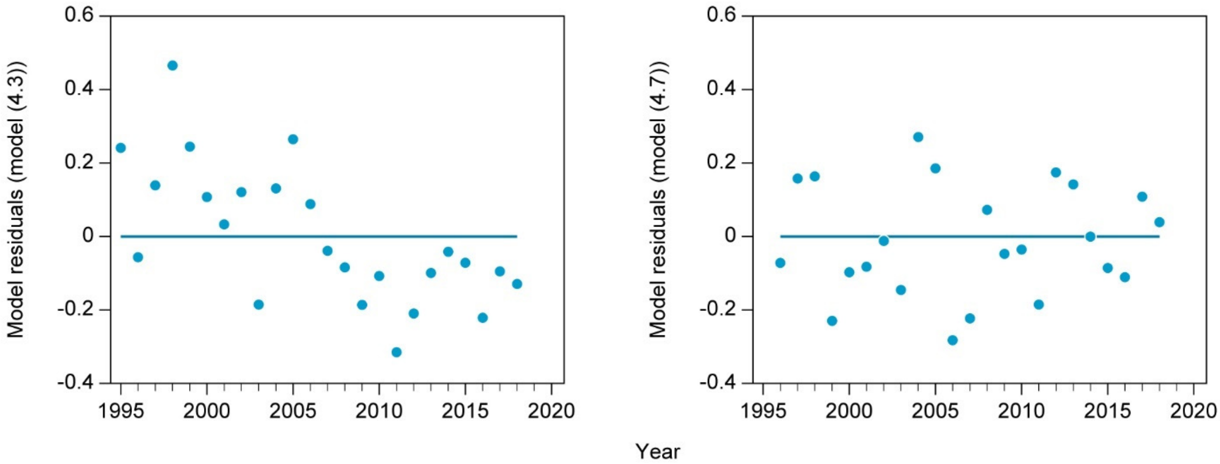


Figure 4.4. Model residual plots using TN loads measured in River Langvad (Denmark). a (figure to the left): residuals applying the model in (4.3). b (figure to the right): residuals applying the model in (4.7). Notice that there is no model residual for the first year in the time series when applying the model in (4.7).

5. Trend analysis, change points and estimation of change

Trend analysis on normalized nutrient input series to different parts of the Baltic Sea, including trend analysis of the water runoff is an important tool in the PLC assessments, when evaluating if nutrient inputs are reduced and when evaluating progress towards fulfilling BSAP nutrient reduction targets (MAI and NIC). Further, it supports evaluation of the effects of implemented measures. The time series used in the trend analysis should always be normalized, but the methods described below may, of course, be used to analyze trends in un-normalized nutrient inputs as well. Trend analysis can be performed using a range of different both parametric and non-parametric methods. Parametric methods comprise ordinary regression with year as the independent variable and linear and non-linear regression methods, such as polynomial, exponential or more complex regression methods. The most well known non-parametric method is the Mann-Kendall trend test and the Theil-Sen estimator for the yearly change in nutrient input. Apart from describing trend analysis methods, this chapter also addresses methods for estimating the size of the trend when it is not linear.

The Mann-Kendall method (Hirsch et al., 1982) is a well-established method for testing for a monotone trend in a time series. It is non-parametric and based on Kendall's tau, which is a measure of the correlation between two different variables. The method is robust towards outliers and a few missing data. If the trend is linear, Mann-Kendall's method has slightly less power than ordinary regression analysis. Annex 1 gives a detailed mathematical description of the method, and the R packages "trend", "rkt" and "Kendall" includes the Mann-Kendall trend test. The Kendall tau coefficient (see Annex 1) can easily be computed in SAS, see the code in Annex 3.

The Mann-Kendall trend method is used for a preliminary analysis of possible trends in the TN and TP load time series. Furthermore, the Mann-Kendall method is used for analyzing possible trends in runoff time series. The remaining trend analysis, as estimating trend line (slope, intersect etc.), is based on linear regression and parametric testing. In the first version of this report (Larsen & Svendsen, 2013), concerning statistical methods more focus was placed on using the Mann-Kendall method.

Ordinary regression analysis is also a well-known statistical method (figure 5.1), but demands a linear relationship with Gaussian distributed residuals, which are stochastic independent as well (Snedecor & Cochran, 1989). In the case where the time series is serially correlated, both the Mann-Kendall test and ordinary regression must be modified, since the tests will be impacted by this, and the probabilities of statistical test values can therefore not be trusted. Serial correlation in a time series can be tested by the Durbin Watson test statistic (Durbin & Watson, 1971). The R package "lmtest" includes the function "dwtest", additionally the R package "car" includes the function "durbinWatsonTest", and see Annex 3 for a SAS code computing the Durbin Watson test. It appears that the autocorrelation for annual time series of either loads or runoff is small and can be ignored; thus, the methods can be used without modifications as a good approximation. The minimum time series length for application of the Mann-Kendall test is 8 years and this can also be used as a guideline for applying ordinary linear regression.

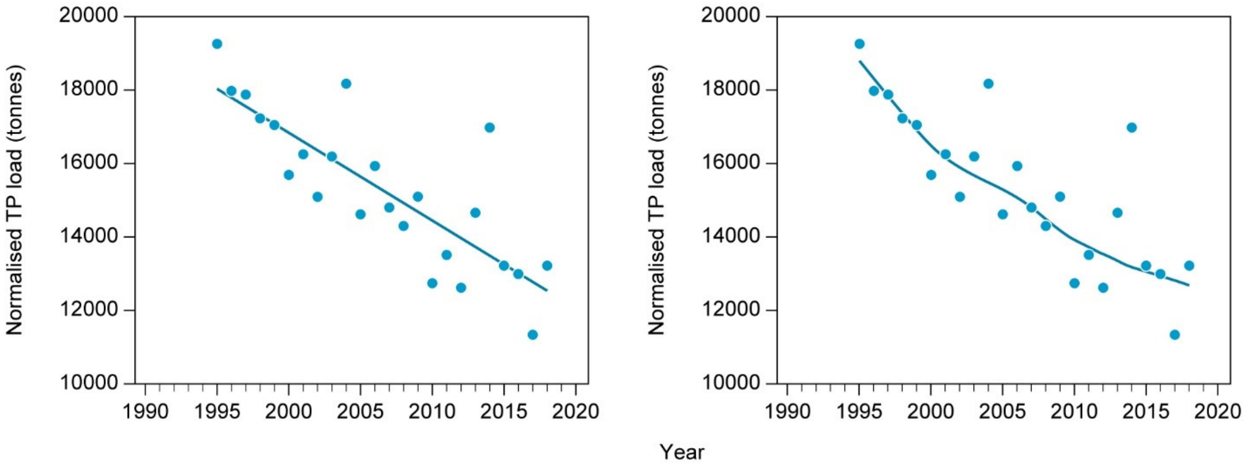


Figure 5.1 a (left figure). Annual normalized TP waterborne inputs (tonnes) to the Baltic Sea. Trend line estimated with linear regression model. B (right figure). As figure a, but the trend line is estimated with LOESS (locally weighted scatterplot smoothing) regression method.

Both Mann-Kendall's trend analysis and ordinary linear regression allow performance of a one-sided trend test if focus is on testing for a downward or increasing development in a time series. This is of relevance in the PLC assessments and when evaluating progress towards HELCOM BSAP MAI and NIC.

A time series plot can show one or two clear trend reversals (also called change points in time), e.g. when the first part of the time series shows a linear increase and the second part shows a linear decrease in nutrient inputs. The trend analysis can then be carried out by using a model with two or three linear curves or by applying two or three Mann-Kendall trend tests if time series sections include a sufficient number of years (example in figure 5.2).

Year of trend reversal (the change point) can either be determined by inspecting the time series plot or by applying a statistical method (Carstensen & Larsen, 2006). If an exact year of change in the inputs is known (e.g. reduced inputs due to implementation of new municipal wastewater treatment plants or new treatment methods, etc.), this year should be applied as the change point, and the time series should be divided accordingly. Statistical estimation of the time when a change occurs in a time series is complex and involves a calculation procedure with iterative estimations. The LOESS (locally weighted scatterplot smoothing, Cleveland (1979)) regression method can be used as a supplement for detecting non-linear trends and for helping detecting change points / step trends as shown in figure 5.1b and 5.2b. LOESS and Generalized additive models (GAM) as described by Hastie & Tibshirani (1990) are valid alternatives for modelling the time series, although there is a risk of overfitting the trend models using these two methods. In this report, focus is on linear models.

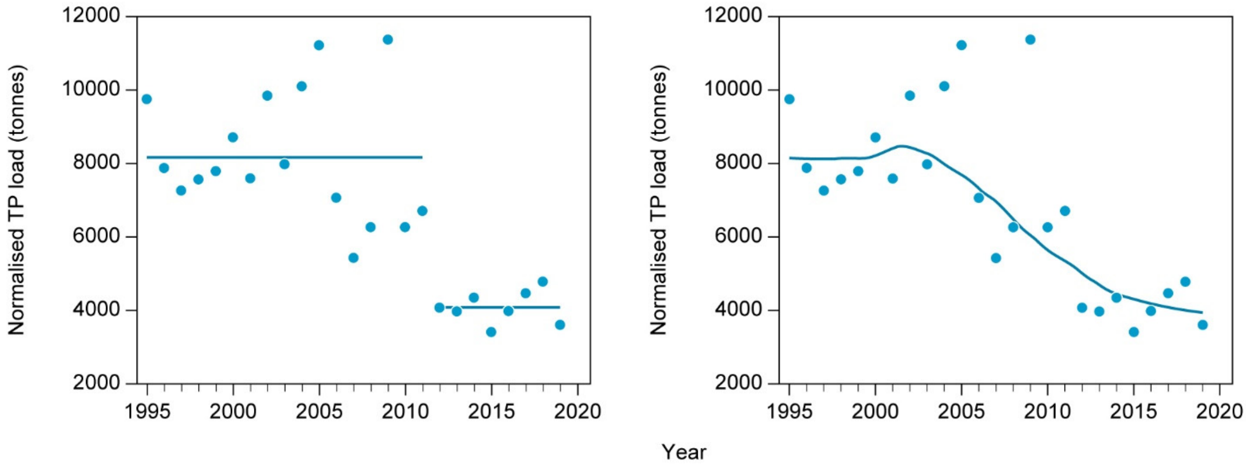


Figure 5.2. a (left figure)**.** Annual normalized TP inputs (tonnes) 1995-2018 to Gulf of Finland. One change point in the time series is detected, and the trend lines are based on linear regression. b (right figure). As figure a, but the trend line is estimated with LOESS regression method.

It is suggested to use models with 1, 2 or 3 linear segments for different sections of the time series (it is still possible that no part of the time series includes significant linear trends). Determination of breakpoints will be statistically analyzed by using an iterative statistical process, which will determine the most significant breakpoint (the significance of the breakpoint is evaluated by the change in -2logQ) – or automatically, where -2logQ is the result from testing a statistical hypothesis with likelihood-ratio test (Carstensen & Larsen, 2006). In addition, each part of time series before or after a change point, should at least be five years or more. It is proposed to investigate two different breakpoint models, here described with two linear parts:

Model 1: $$L_{Ni} = \begin{cases} \alpha + \beta \cdot i, \; for\; i < Y \\ \alpha + \beta \cdot i + d \cdot (i - Y), \; for\; i \geq Y \end{cases}$$

Model 2: $$L_{Ni} = \begin{cases} \alpha_1 + \beta_1 \cdot i, & for\; i < Y \\ \alpha_2 + \beta_2 \cdot i, & for\; i \geq Y \end{cases}$$

L_N = Normalized input

α = Intercept parameter

β and d = Slope parameters

Y = Breakpoint year

i = Index indicating different years in the time series.

Model 1 is continuous at the breakpoint (the two lines are connected) while model 2 has disconnected lines at the breakpoint (a step).

After the first breakpoint is determined, another iterative process looking for a second breakpoint is performed.

Change-points models are an aid for calculating an estimate for the last year value in a time series. Furthermore to get an idea of the overall shape of the trend during the full time series period.

Finally, in former assessments of HELCOM BSAP MAI and NIC significance of the slope in the last segment has been tested, and if not significant different from zero then following model was used:

$$L_{Ni} = \begin{cases} \alpha + \beta \cdot i, & for\ i < Y \\ c, & for\ i \geq Y \end{cases}$$

c = Estimated input (a constant)

In future assessments, it is suggested to test the significance of slopes in all identified parts of different segments of a time series, and fit a model with constant values in the segments where the slope can be accepted to be zero. For instance, it could end up with a model like

$$L_{Ni} = \begin{cases} c_1, & for\ i < Y \\ c_2, & for\ i \geq Y \end{cases}.$$

Table 5.1 describes the modelling process that is suggested for fitting linear models with breakpoints to the time series.

Table 5.1: The modelling process for identifying breakpoints, testing for significant slopes and fitting constants (no significant slopes) and regression parameters (significant slopes) in a time series.

step	step	step	step
A significant breakpoint	Test for additional breakpoints in each segment	Test for significant slopes in the segments	Fit a constant in segments with a non-significant slope. Fit regression parameters in the rest of the segments.
No breakpoint	Fit a constant for the whole time series		

The second part of trend analysis is the task of estimating the size of the trend or the change per year. Again, several different methods exist, and the specific use of these depends on the shape of the trend. The Theil-Sen slope estimator (Hirsch et al., 1982) is a non-parametric estimator that is resistant towards outliers (suspect) values. The method assumes a linear trend and estimates the change per year, and the estimator fails if the trend is non-linear, and if the time series shows time reversal, it is necessary to split the time series into e.g. two parts.

The size of a linear trend can also be estimated by regression. This is the classical approach, which is, however, not flexible with regard to all shapes of trend. The simplest method is using the start and end values in the time series of flow-normalized inputs, but if start and / or end values are too distant from the general trend, this method is not reliable.

If we seek to identify the total change in nutrient inputs over the whole time series, expressed as a percentage, we can apply the following method. Using, the fitted trend model, we estimate the normalized values at the first year and the last year in the time series and simply calculate the change as

$$100 \cdot \frac{(\widehat{L_{Nn}} - \widehat{L_{N1}})}{\widehat{L_{N1}}}, \tag{5.1}$$

1 = First year in the series

n = Length (number of years) in the series (last year)

$\widehat{L_{N1}}$ = Estimated normalized input in year 1, the first year in the time series

$\widehat{L_{Nn}}$ = Estimated normalized input in year n, the last year in the time series.

The same method as given in (5.1) can be used for segments of the time series, i.e. parts before, between and after breakpoints.

The trend analysis methods are illustrated below based on the time series of normalized TN inputs to the Kattegat. In figure 5.3, the normalized time series are shown together with a model fit of the trends. The model fit consists of one change point (a step change in the level) in year 2011 and the linear fit before the change point is significant, the linear fit after is not significant, so a constant value is fitted. A trend analysis should always be initiated with a time series plot of the data series.

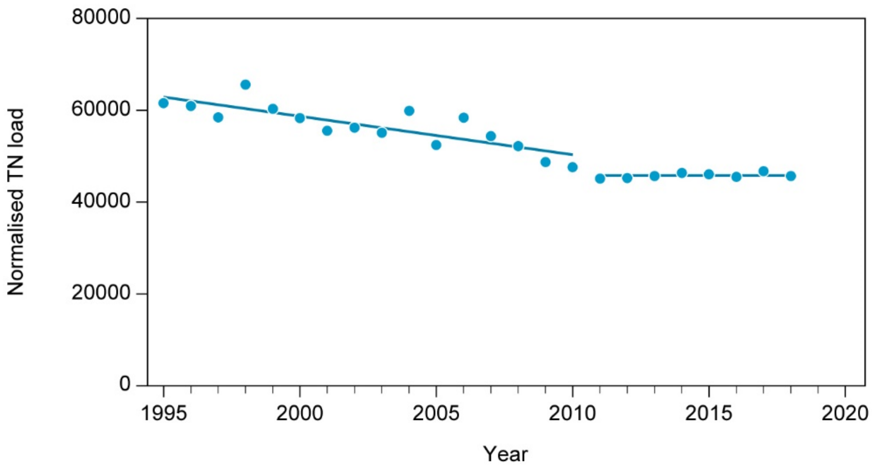
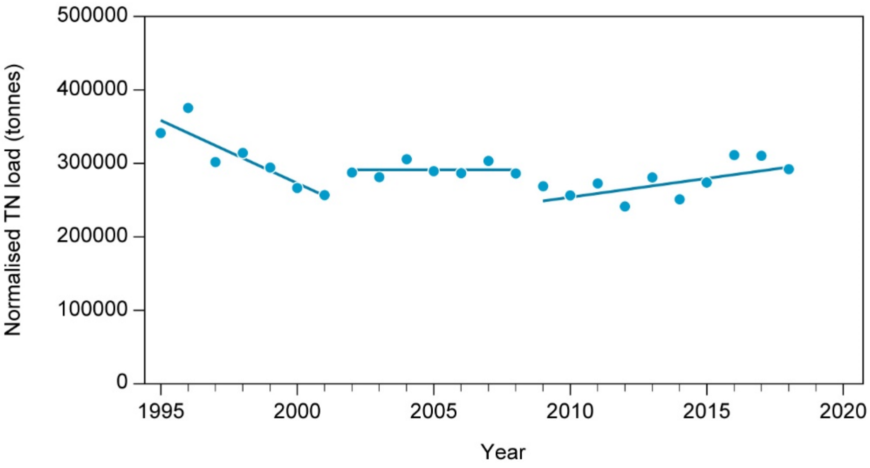


Figure 5.3. The fitting of a model with one change point to the normalized waterborne TN inputs (tonnes) to the Kattegat.

The estimated change over the whole period for the normalized TN inputs in figure 5.3 is -27.2% according to formula 5.1.

Figure 5.4 shows another example of the trend analysis method. Data are waterborne TN inputs to the Baltic Proper. In this example, two change points are identified, the first in 2002, and the second in 2009. The middle part is fitted with a constant value; the other two lines have significant slopes. The total estimated change is estimated to -17.7%.

Figure 5.4. The fitting of a model with two change points to the normalized waterborne TN inputs (tons) to the Baltic Proper.



6. Testing fulfilment of BSAP reduction targets

The progress in nutrient input reduction can be tested by two different methods:

1) Trend analysis of time series of normalized nutrient inputs, as discussed in chapter 5

2) Statistical analysis of whether the country-wise nutrient input ceilings (NIC) and maximum allowable inputs per sub-basin (MAI) under HELCOM BSAP have been significantly met by a Contracting Party.

In this chapter, a statistical method for testing fulfillment of MAI and NIC reduction targets is proposed, and a traffic light system is introduced to illustrate a country's progress towards fulfilling the targets. A statistical method for testing if a normalized nutrient time series has moved relative to a defined nutrient target is needed. For this purpose, a parametric method based on the simple test of the mean value in a sample of Gaussian distributed data is suggested – a method that is often referred to as the fail-safe principle.

In addition, a method for testing, if the estimated load for the last year in the time series is significantly different from the average value in a reference period (which for nutrients input is 1997-2003), is presented.

Let us first assume that we have a time series of normalized inputs. The time series is initially assumed to be without a statistical significant trend and without a significantly large serial correlation. We assume further that the nutrient target T (any kind of target such as, for instance, nutrient input ceilings for a country basin – NIC, or maximum allowable input for a Baltic Sea Sub-basin - MAI) is defined without error, i.e. is a fixed value (certain amount of nitrogen / phosphorus given without any uncertainty). Let us finally assume that the data are sampled from a Gaussian distribution with mean value α and variance σ^2.

As null hypothesis for the statistical test, we assume that the target has not been fulfilled, i.e.:

$$H_0 : \mu \geq T ,$$

The alternative hypothesis H_A: $\mu < T$ follows from this, i.e. the target has been fulfilled. Now assume that the test significance level α is defined to be 5% (0.05), and then calculate the statistic

$$\bar{x}_{AD} = \bar{x} + t_{n-1,0.05} \cdot SE_{\bar{x}}, \qquad (6.1)$$

$\bar{x}_{AD}$ = Adjusted mean

$\bar{x}$ = Mean of all values in the time series

SE = Standard error (SE = standard deviation divided by square root of n), standard error of $\bar{x}$

n = number of observations in the time series

$t_{n-1,0.05}$ = 95% percentile in a t-distribution with n-1 degrees of freedom.

A test significance level of 5% means that we have a 5% probability of incorrectly rejecting the null hypothesis. For hypothesis testing and estimation of confidence intervals we use the t-distribution table since the true population standard deviation (σ) is not known and since the number of years (the sample size) is still small (less than 30 years).

The statistic in (6.1) is called the adjusted mean, and if the statistic is less than the target T, the reduction target is fulfilled.

In the case of a time series on nutrient inputs with a significant trend, another statistical method is needed for testing if a HELCOM BSAP target is fulfilled. Let us assume that the trend is linear, a linear regression model with year as independent variable can be fitted to the time series, estimates for α and β can be calculated, and the residuals are Gaussian distributed. The linear model is then used to predict a normalized nutrient input for the last year ($year_n$) in the time series. This estimate is calculated as:

$$\widehat{L_{Nn}} = \hat{\alpha} + \hat{\beta} \cdot year_n. \tag{6.2}$$

$\widehat{L_{Nn}}$ = Estimated normalized input in year n, the last year in the time series

$\hat{\alpha}$ = Estimated intercept

$\hat{\beta}$ = Estimated slope

Next, we need the standard error of the prediction (predicted nutrient input) and this is defined as:

$$SE_{\widehat{L_{Nn}}} = \sqrt{MSE} \cdot \sqrt{1/n + (year_n - \overline{year})^2 / \sum_{i=1}^{n} (year_i - \overline{year})^2} \tag{6.3}$$

MSE = Mean Squared Error as defined in chapter 4 (formula 4.5)

n = Number of years in the time series

$year_n$ = Last year in the time series (e.g. 2019)

$year_i$ = A given year in the time series (e.g. 1997)

$\overline{year} = \sum_{i=1}^{n} year_i / n$

Then the statistic is calculated as:

$$\bar{x}_{AD} = \widehat{L_{Nn}} + t_{n-2,0.05} \cdot SE_{\widehat{L_{Nn}}}, \tag{6.4}$$

$t_{n-2,0.05}$ = 95% percentile in a t-distribution with n-2 degrees of freedom.

A list with the 95% percentiles for different degrees of freedom is given in Annex 2.

The mathematical definition of the standard error of the prediction SE given in (6.3) is a well-known statistic from ordinary linear regression (Snedecor & Cochran, 1989).

If the trend is not linear, other models have to be used for the time series, and the formula for the standard error needs to be revised. The form of the trend in the data will dictate the methods to be applied. These methods are based on the assumption of the existence of one or two change points in the time series (see chapter 5).

"Trend method"

A few examples are given below, they are based on models with one change-point Y, and we assume that the last year in the time series is denoted by $year_n$. In general, we denote the method the "trend method". The first example is a model with one change point and a linear model before and a linear model after and no change in level before and after the change point. The second example is equal to the first example but with a change in level at the change point. The last example (example 3) is with a constant level after the change point.

<u>Example 1</u>:

$$L_{Ni} = \begin{cases} \alpha + \beta \cdot i, & for\ i < Y \\ \alpha + \beta \cdot i + d \cdot (i - Y), & for\ i \geq Y \end{cases}$$

$$\widehat{L_{Nn}} = \hat{\alpha} + \hat{\beta} \cdot year_n + \hat{d} \cdot (year_n - Y)$$

<u>Example 2</u>:

$$L_{Ni} = \begin{cases} \alpha_1 + \beta_1 \cdot i, & for\ i < Y \\ \alpha_2 + \beta_2 \cdot i, & for\ i \geq Y \end{cases}$$

$$\widehat{L_{Nn}} = \widehat{\alpha_2} + \widehat{\beta_2} \cdot year_n.$$

<u>Example 3</u>:

$$L_{Ni} = \begin{cases} \alpha + \beta \cdot i, for\ i < Y \\ c, \qquad for\ i \geq Y \end{cases}$$

$$\widehat{L_{Nn}} = \hat{c}$$

The SE for the estimated input for the last year ($year_n$) has the general form of

$$SE_{\widehat{L_{Nn}}} = \sqrt{MSE} \cdot \sqrt{1/m + (year_n - \overline{year})^2 \Big/ \sum_{i=Y}^{year_n} (i - \overline{year})^2}$$

m = Number of years after Y ($\geq Y$)

$\overline{year} = \sum_{i=Y}^{year_n} i \Big/ m.$

MSE is calculated for the full model i.e. including all years in the time series.

If the model include a constant level after the last change point the $SE_{\widehat{L_{Nn}}}$ for the estimated input for the last year ($year_n$) has the form

$$SE_{\widehat{L_{Nn}}} = \sqrt{MSE} \cdot \sqrt{1/m}.$$

Correction for calculating control value for *year*$_n$

$$\bar{x}_{AD} = \widehat{L_{Nn}} + k \cdot SE_{\widehat{L_{Nn}}}. \tag{6.5}$$

k = 95% percentile in a t-distribution with n-p degrees of freedom.

p = Number of parameters in the final model.

Traffic light system

Finally, a traffic light system can be defined to obtain a status of whether a country or a sub-basin has met the set HELCOM BSAP target, whether it is close to fulfilling the target (when it is not possible to statistical judge if targets are fulfilled due to uncertainty on the estimated inputs), or whether the target has not been fulfilled. This is originally described in Larsen & Svendsen (2013). Statistically, we define the system as:

Red:

If $\bar{x}$ or $\widehat{L_{Nn}} > T$, i.e. the estimated normalized input for the last year or the average normalized nutrient input over the considered period (when there is no trend) is above the target value T.

Yellow:

If $\bar{x}$ or $\widehat{L_{Nn}} < T$, and if $\bar{x}_{AD} > T$, i.e. the null hypothesis of target test is accepted, but the estimated normalized input for the last year or the average normalized input over the considered period (when there is no trend) is lower than the target value.

Green:

If $\bar{x}_{AD} < T$, i.e. the null hypothesis of the target test is rejected, i.e. the alternative hypothesis is accepted meaning the target has been fulfilled, and the estimated normalized input for the last year or the average normalized input over the considered period (when there is no trend) is lower than the target value.

Testing whether estimated last year input is lower than input in the reference period

For testing whether the estimated last year value $\widehat{L_{Nn}}$ is significantly different from the mean value in the reference period we apply the following procedure.

The reference period is defined to be the fixed period 1997-2003. First, calculate the mean value in the reference period

$$\overline{L_{N\,(ref)}} = \frac{1}{7}\sum_{i=1997}^{2003} L_{Ni} \tag{6.6}$$

and calculate the 95% confidence interval for $\overline{L_{N\,(ref)}}$ by

$$\overline{L_{N\,(ref)}} \pm k \cdot \mathrm{SE}_{\overline{L_{N\,(ref)}}}. \tag{6.7}$$

The k factor is the 97.5% percentile in a t-distribution with 6 degrees of freedom (k=2,447) (7 years in the reference period minus 1). The $\mathrm{SE}_{\overline{L_{N\,(ref)}}}$ is the standard error of the mean value.

For the estimate of the last year $\widehat{L_{Nn}}$ we can calculate the 95% confidence interval as well by calculating

$$\widehat{L_{Nn}} \pm k \cdot SE_{\widehat{L_{Nn}}} \tag{6.8}$$

where the k factor is the 97.5% percentile in a t-distribution with n-p degrees of freedom. The number p is the number of parameters in the final model. $SE_{\widehat{L_{Nn}}}$ is the standard error for the specified model, used for the time series. How to calculate the $SE_{\widehat{L_{Nn}}}$ is given above in this chapter.

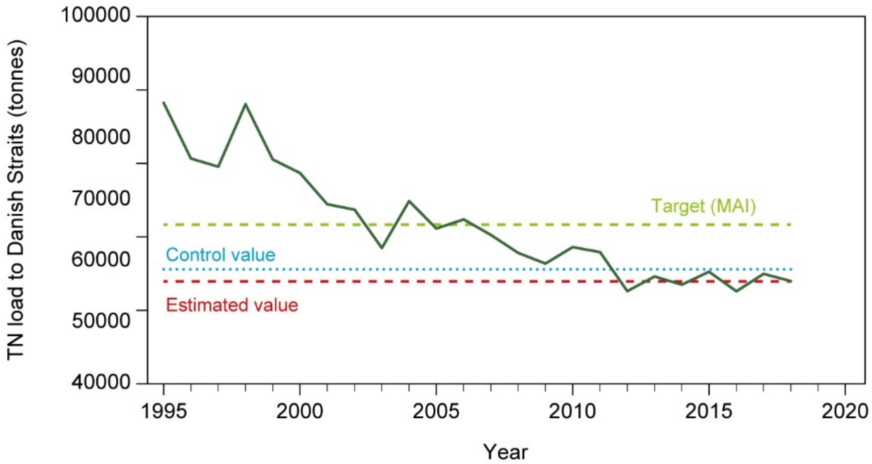
Testing the hypothesis of no difference between the reference period and the last year value can simply be done by determining if

$$\left|\overline{L_{N\,(ref)}} - \widehat{L_{Nn}}\right| - k \cdot \sqrt{\mathrm{SE}_{\overline{L_{N\,(ref)}}}{}^{2} + SE_{\widehat{L_{Nn}}}{}^{2}} > 0 \tag{6.9}$$

where k is the 97.5% percentile in a t-distribution with n-p+6 degrees of freedom, and | | stands for absolute value.

To illustrate the principles, we tested if the normalized TN inputs to the Danish Straits met the provisional MAI input ceiling of 65,998 tonnes TN per year in 2018. Using the model with one change point in 2003 (se figure 5.3) the estimating input in year 2018 is 56,719 tonnes with a $SE_{\widehat{L_{Nn}}}$ of 1,082 tonnes. According to formula (6.5) the control value becomes 58,683 tons, which is less than 65,998 tonnes, so in the example the traffic light evaluation results in a green light. This example is illustrated in figure 6.1.

Figure 6.1. Principles on time series with trend created annual TN input to the Danish Straits. Dotted green line is the target (MAI), "-----" red line is the estimated value (TN input) in 2018, and "…." blue line is the control value (TN input taking into account uncertainty) according to formula 6.5.



The average TN inputs in the reference period to Danish Straits was 73,167 tonnes and the confidence interval according to formula (6.7) is [66,280; 80,054] tonnes. The confidence interval for the 2018 estimate is [54,343; 59,094] tonnes according to formula (6.8). Using formula (6.9) we calculate that the left side of the inequality sign is 11,280 tonnes, which is larger than zero, so we conclude that TN input in 2018 to Danish Straits is statistical significantly reduced (with 22%) since the reference period 1997-2003.

Table 6.1 includes another examples of applying of the statistical analysis described to evaluate fulfilment of Finish phosphorus input ceilings based on data from 1995-2014. TP inputs to Gulf of Finland (647 tonnes P) are higher than the inputs ceiling to GUF (322 tonnes P), and the traffic light is then red and taking into account uncertainty the remaining reduction to fulfill reductions targets was 351 tons. The traffic light for Bothnian Sea is yellow, because the estimated TP inputs in 2014 when including uncertainty on the input estimate are higher than the input ceilings to BOS. For Bothnian Bay meet the input ceiling (green) with 137 tonnes P (extra reduction) taking into account uncertainty.

Table 6.1. Illustration of the traffic light system. Evaluation of progress towards reductions targets (nutrient inputs ceiling) of TP for Finland to Bothnian Bay (BOB), Bothnian Sea (BOS) and Gulf of Finland (GUF) based on normalized annual TP inputs from Finland during 1995-2017. Green: input ceiling are meet. Red: input ceilings are not fulfilled. Yellow: It cannot be judge with statistical certainty if input ceilings are fulfilled taking into account uncertainty on estimated inputs in 2017.

Finland TP	**BOB**	**BOS**	**GUF**
A : Input ceiling	1668	1255	322
B: Estimated input 2017	1545	1292	634
C: Inputs 2017 including uncertainty (test value)	1608	1357	668
Extra reduction (A-C)	60		
Remaining reduction to fulfill MAI		103	346

7. Step by step analysis illustrated by HELCOM data examples

Firstly, it is demonstrated in details how the new normalization method is applied. The example regarding the normalization is the same as in chapter 4, TN loads in River Albeck. The relationship between differenced (natural) logarithmic transformed TN loads and differenced logarithmic transformed flows is shown in figure 4.2a. The linear fit (formula (4.7), with logarithmic transformed values) is good with a R^2=0.76 and the parameter α can be accepted to be zero with a P-value of 0.59, so the regression model has only one parameter, the slope β. The estimate of β is 1.391, $log\bar{Q}$=-0.9796, MSE=0.0480, $\bar{L} = 40.08$ and $\overline{L_N} = 38.22$. Using formulas (4.11) and (4.12), we get the following values:

Year	I	Q	L	log(L)	log(Q)	r	L_N	L_{NB}
1995	1	0.409	50	3.912023	-0.89404	5.155551	45.46574	47.32679
1996	2	0.209	17.9	2.884801	-1.56542	5.062157	41.4118	43.27285
1997	3	0.251	26.71	3.285038	-1.3823	5.207693	47.89935	49.7604
1998	4	0.486	80.77	4.391606	-0.72155	5.395211	57.77864	59.63969
1999	5	0.394	59.5942	4.087558	-0.9314	5.383056	57.08061	58.94167
2000	6	0.2973	31.47	3.449035	-1.21301	5.136225	44.59552	46.45658
2001	7	0.3352	33.9	3.523415	-1.09303	5.043716	40.65512	42.51617
2002	8	0.6426	52.4196	3.959281	-0.44223	4.574386	25.42656	27.28761
2003	9	0.266	19.47	2.968875	-1.32426	4.810797	32.20771	34.06876
2004	10	0.45	44.81	3.802431	-0.79851	4.913082	35.67644	37.53749
2005	11	0.3533	31.67	3.45537	-1.04044	4.902523	35.30171	37.16277
2006	12	0.36	35.0615	3.557104	-1.02165	4.978127	38.07413	39.93518
2007	13	0.64	97.62	4.581082	-0.44629	5.201827	47.61916	49.48021
2008	14	0.4347	44.3741	3.792656	-0.8331	4.95142	37.07077	38.93182
2009	15	0.2501	18.5604	2.92103	-1.38589	4.848682	33.45129	35.31235
2010	16	0.4908	46.9368	3.848802	-0.71172	4.838737	33.12027	34.98133
2011	17	0.4348	50.9471	3.930788	-0.83287	5.089232	42.54833	44.40939
2012	18	0.3524	25.2515	3.228886	-1.04299	4.679586	28.24722	30.10827
2013	19	0.4327	32.3819	3.4776	-0.83771	4.642778	27.22639	29.08744
2014	20	0.3005	28.5873	3.352963	-1.20231	5.025262	39.91173	41.77278
2015	21	0.429	49.7162	3.906331	-0.8463	5.083454	42.30319	44.16424
2016	22	0.3375	27.3257	3.307828	-1.08619	4.818618	32.46057	34.32163
2017	23	0.4869	36.1017	3.58634	-0.7197	4.587372	25.7589	27.61995
2018	24	0.3203	20.2534	3.008323	-1.1385	4.591868	25.87497	27.73602

The values of L_{NB} are the bias corrected normalized values according to (4.12). The result of the normalization can be seen in figure 4.3a. The year-to-year variation is markedly reduced.

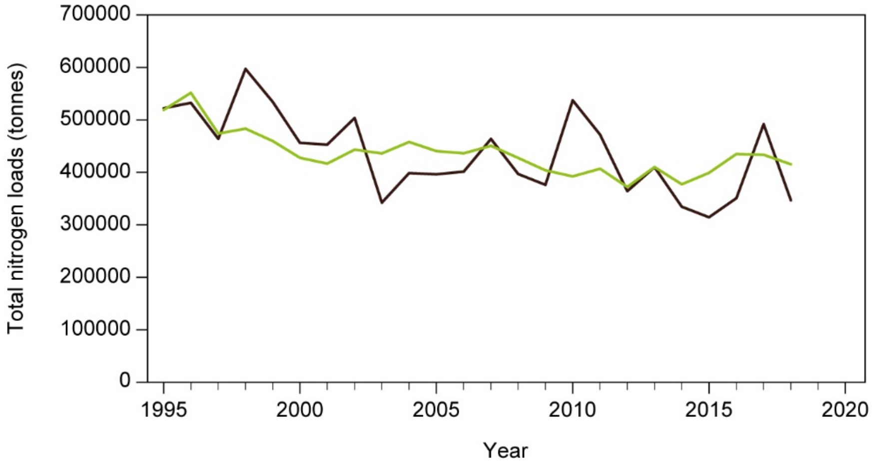
The rest of this chapter will present a full statistical analysis of time series from normalization in order to test whether a target has been fulfilled. We use data from the input of TN to the Baltic Proper.

We assume that the data have been evaluated for data gaps and outliers and thus are without missing values and errors – in other words, the data have been accepted by all relevant Contracting Parties.

Firstly hydrological normalization is performed individually for all rivers that discharge into the Baltic Proper.

Figure 7.1 shows the normalized inputs of TN summed up for all rivers discharging into the Baltic Proper, plotted together with the measured actual inputs. As can be seen, the variation between years is significantly reduced.

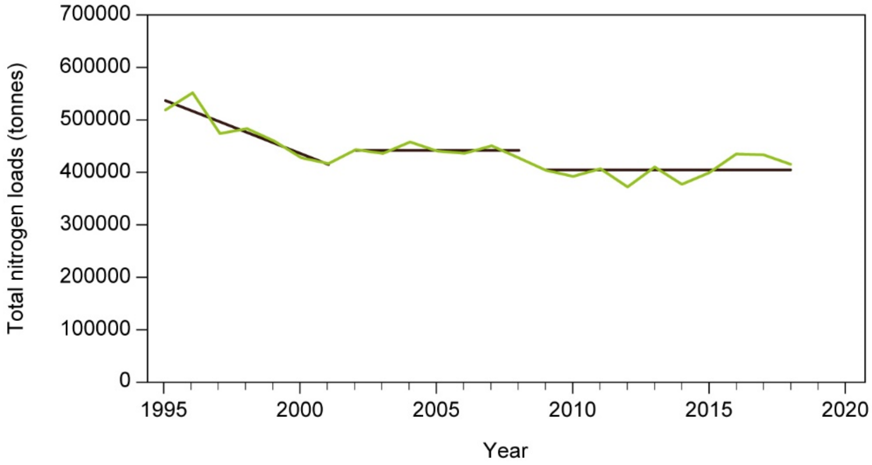
Figure 7.1 Normalized (green) and actual water and airborne TN inputs (tonnes) to the Baltic Proper.



As mentioned, the normalization is carried out for individually for all rivers discharging into the Baltic Proper, and these normalized inputs summed for all the rivers together with inputs from direct point sources and atmospheric deposition are used for the trend analysis and the target testing.

Figure 7.2 shows the trend model estimated for the time series. The model consists of two change-points, one in 2002 and the second in 2009. The two segments after the change-points can be accepted to have constant values, see figure 7.2. The first segment has a significant downward trend, estimated to be 20,315 tonnes per year.

Figure 7.2. Linear trend fit to annual normalized water + airborne TN inputs (tons) to the Baltic Proper during 1995-2018.

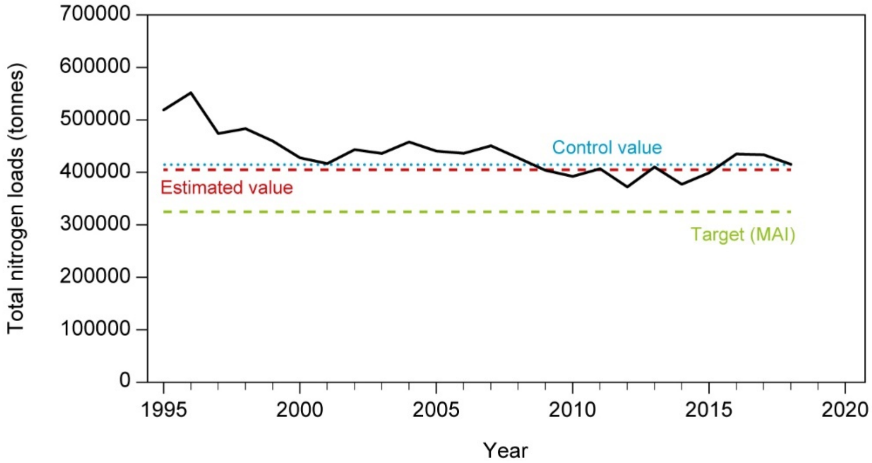


The nitrogen input ceiling (target) for the Baltic proper is set to 325,000 tons. The model estimated normalized load in 2018 is 404,613 tons and the test value using formula 6.5 is calculated to be 414,590 tons, both are well above the target, so the traffic light would be red for the basin (see table 7.1 and figure 7.3).

Tabel 7.1. Assessment of fulfilment of HELCOM BSAP maximum allowable total nitrogen (TN) input (MAI target) to Baltic Proper (BAP) in 2018 based on normalized annual TN input data from 1995-2018. Units in tons.

A: MAI	325,000
B: Estimated TN input	404,613
C: Statistical uncertainty (from formula 6.5)	9,977
D: Estimated TN input including uncertainty (B+C)	414,590
E: Exceedance of MAI (D-A)	89,590
Traffic light colors	

Figure 7.3. Testing the target value for water and airborne TN inputs (tonnes) to the Baltic Proper for the period 1995-2018. Dotted green line is the target (MAI), "-----" red line is the estimated value (TN input) in 2018, and "…." blue line is the control value (TN input taking into account uncertainty) according to formula 6.5.



The 95% confidence interval for the estimated normalized TN input in 2018 (404,613 tons) is [392,547; 416,680 tons]. For the reference period 1997-2003 the mean normalized TN inputs is 448,714 tons with a 95% confidence interval of [426,049 ; 471,379 tons]. Since the left hand side of the inequality in formula (6.9) equals 25,474 which is larger than zero, the estimated TN inputs for 2018 is statistical significantly lower (14%) than the average TN input during the reference period.

8. Recommendations

This report deals with the statistical aspects of analyses in relation to PLC data assessments, and evaluations of fulfilment of HELCOM BSAP reduction targets (MAI and NIC) etc. A number of different topics are covered, as for instance hydrological normalization, trend and change point analysis, and significance tests for whether targets have been met or not taking into account statistical uncertainty on nutrient input data. In the following, we have listed recommendations for which statistical method is best suited for the preparation of the PLC guideline.

- Good data quality and consistency are necessary to conduct reliable statistical analyses of the available time series. Time series may include gaps and / or suspicious / dubious values. In chapter 2 of this report, methods for filling in gaps and how to determine if a dubious value is an outlier are described.

- Regarding total uncertainty in country nutrient input data. Total uncertainty consist of uncertainty due to bias and precision, respectively. It is a difficult task to calculate the exact total uncertainty for the data provided by the contracting parties. One potential method may be to apply the simpler method DUET-H / WQ described in Harmel et al. (2009), which gives an approximation to the total uncertainty in monitored catchments. The report includes some standard values for bias and precision on annual nutrient loads from different catchments sizes dominated by bed rock and soils, respectively, values which could be used when information on uncertainty otherwise are not available. Information on the uncertainty of nutrient inputs in unmonitored areas has also to be given by the Contracting Party – either by model uncertainty or as an expert evaluation.

- Normalization of nutrient inputs should be performed using the method based on transformed inputs and runoff. It is recommended to apply an adjusted method compared with former edition of the report by applying method of differencing the time series relating the difference in loads between following years with the corresponding difference in flow (formula 4.6 in this report). Transformation should be undertaken using the natural logarithmic function (see formula 4.9 in this report). Normalization is carried out for each catchment (river) separately, and normalized inputs can be summed up at country, country by basin or at Baltic Sea sub-basin level. Normalization is a necessary step before conducting trend analysis. The method ensures that variation in annual inputs is significantly reduced, contributing to an improved test for a significant trend in inputs by allowing identification of minor trends as being statistically significant. If a decision is made to use monthly input time series in the future, similar normalization methods can be applied to the monthly data (see for instance Silgram & Schoumans (ed., 2004)).

- Concerning trend analysis, the Mann-Kendall non-parametric trend method is recommended for a first analysis testing possible monotone trend in the normalized time series. The method is robust although autocorrelation can deflate the power of the test, as it will for all statistical test methods. This non-parametric method can be used on both “raw” nutrient time series, normalized time series and runoff (climate) time series. If it is decided to use monthly input time series in the future, the Kendall trend test has been extended to a seasonal version (Hirsch & Slack, 1984). Ordinary regression (a parametric method) analysis is recommended for the more detailed analysis of trends and their magnitude (as for the slope, intercept, including sections of a time series, last year inputs, changes in inputs etc.). The ordinary regression analysis demands a linear relationship with Gaussian distributed residuals, which are stochastic independent. Mann-Kendall test and ordinary regression should be modified when the time series is serial correlated. Durbin Watson test statistic can test for serial correlation in a time series.

- If the time series show one or two trends reversal (change points), two or more linear trend segments / sections should be applied to model the time series. The change-point can either be determined by visual inspection of the time series plot or best by a statistical method (Carstensen & Larsen, 2006). It is suggested to limit the number of change points to two for time series of 20-25 years of length. Each part of time series before or after a change point should at least be five years or more. Two different change point models should be tested, model 1 where the time series is continuous in the changes point (trend line connected in the change point) and model 2 with disconnected trend lines in the change point. We recommend to apply ordinary linear regression for estimating trend lines for each segment of a times series with change points.

- Estimating the change in nutrient inputs can be done by the non-parametric Theil-Sen slope estimator. The method assumes a constant change, i.e. a linear trend. As change point analysis is applied, it is recommended to use linear regression for estimating slope and intercept in PLC assessment and for evaluation progress towards MAI and NIC reductions targets. If the trend model is not linear, the estimated start-end difference in load or flow values should be used.

- To assess and follow up on HELCOM BSAP nutrient reduction targets (MAI and NIC) a statistical method is needed evaluating whether MAI and NIC have been fulfilled, and to quantify remaining reduction requirements / distance to the targets. For time series with a non-significant trend, the equation in formula 6.1 can be used to calculate the estimated latest year nutrient input (the trend method) and evaluate this value against the target value taking into account uncertainty. For time series with a significant linear trend, the equation in 6.4 should be used. Regarding time series with linear trend segments, the equation in 6.5 should be applied. For times series without trend or without trend in the part after a break point including the latest year the average of inputs in the part should be used as the estimated input. We have defined a traffic light system allowing evaluation of nutrient inputs from varying catchments / Contracting Parties to the Baltic Sea according to defined targets in three categories:

    - Red = estimated input for the last year is higher than the target: target not fulfilled;

    - Yellow = estimated input for the last year is lower than the target but the estimated input for the last year including uncertainty is higher than the target: due to statistical uncertainty it cannot be judged whether the target is fulfilled;

    - Green: estimated input for last year including uncertainty on the estimated input is lower than the target: target fulfilled

- Finally in formula 6.9 we indicate how it can be tested if the estimated nutrient inputs in the last year of a time series is significant different from the corresponding inputs in the reference period in HELCOM BSAP nutrient reduction scheme (reference period = 1997-2003), allowing for calculating significant changes since the reference period.

9. References

BSAP (2007) HELCOM Baltic Sea Action Plan. HELCOM Ministerial Meeting, Krakow, Poland 15 November 2007.

Box, E.P., Jenkins, G. M., Reinsel, G. C. & Ljung, G. M. (2015) Time Series Analysis: Forecasting and Control. John Wiley and Sons, New York.

Carstensen, J. & Larsen, S. E. (2006) Statistisk bearbejdning af overvågningsdata – Trendanalyser. NOVANA. (Statistical assessment of monitoring data. Trend analyses. NOVANA). Danmarks Miljøundersøgelser. 38 s. – Teknisk anvisning fra DMU nr. 24. http://www.dmu.dk/Pub/TA24.pdf (in Danish)

Cleveland, W. S. (1979) Robust Locally Weighted Regression and Smoothing Scatterplots. Journal of the American Statistical Association, 74, p. 829-836.

Conover, W.J. (1980) Practical Nonparametric Statistics. Second Edition. John Wiley and Sons, New York.

Durbin, J. & Watson, G. S. (1971) Testing for serial correlation in least squares regression. III. Biometrika, 58 (1): 1–19.

Ferguson, R. I., (1986) River loads underestimated by rating curves: Water Resources Research, 22, 1, p. 74-76.

Gilbert, R.O. (1987) Statistical Methods for Environmental Pollution Monitoring. New York: Van Nostrand Reinhold.

Gustafsson, Bo (2019) Flow normalization notes. Baltic Nest Institute, Baltic Sea Centre, Stockholm University, SE-106 91 Stockholm, Sweden.

Harmel, R. D., Smith, D. R., King, K. W. & Slade, R. M. (2009) Estimating storm discharge and water quality data uncertainty: A software tool for monitoring and modeling application. Environmental Modeling & Software, 24, 832-842.

Hastie, T. J. & Tibshirani, R. J. (1990) Generalized Additive Models. Chapman & Hall/CRC, New York

HELCOM (2020) Inputs of nutrients to the sub-basins of the Baltic Sea. HELCOM core indicator report. Online. December 2020. https://helcom.fi/media/core%20indicators/HELCOM-core-indicator-on-inputs-of-nutrients-for-period-1995-2018_final.pdf

HELCOM (in prep). Guidelines for the annual and periodical compilation and reporting of waterborne pollution inputs to the Baltic Sea (PLC-water).

Herschy, R.W. (2009) Streamflow measurements. Third edition, Routledge Taylor & Francis, 507 p.

Hirsch, R.M. & Slack, J.R. (1984) A nonparametric trend test for seasonal data with serial dependence. Water Resources Research, 20, 727-732.

Hirsch, R.M., Slack, J.R. & Smith, R. A. (1982) Techniques of trend analysis for monthly water quality data. Water Resources Research, 18, 107-121.

Kendall, M.G. (1975) Rank Correlation Methods. Charles Griffin, London.

Kronvang, B. & Bruhn, A.J. (1996) Choice of sampling strategy and estimation for calculating nitrogen and phosphorus transport in small lowland streams. Hydrological Processes, 10, 1483-1501.

Kronvang, B., Larsen, S.E., Windolf, J. & Søndergaard, M. (2014) Uncertainties in monitoring – how much do we tolerate. Presentation of NJF Workshop, 19-21 March 2014, Stavanger, Norway.

Larsen, S.E. & Svendsen, L.M. (2013) Statistical aspects in relation to Baltic Sea Pollution Load Compilation. Task 1 under HELCOM PLC-6. Aarhus University, DCE – Danish Centre for Environment and Energy, 34 pp. Technical Report from DCE – Danish Centre for Environment and Energy No. 33. http://dce2.au.dk/pub/TR33.pdf

Larsen, S.E, & Svendsen, L.M. (2019) Statistical aspects in relation to Baltic Sea Pollution Load Compilation. Task under HELCOM PLC-7 project. Aarhus University, DCE Danish Centre for Environment and Energy, 42 pp. Technical Report No. 137. http://dce2.au.dk/pub/TR137.pdf

Li, D. & Edwards, E. (2001) Automatic Estimation of Dixon's Test for Extreme Values Using a SAS® Macro Driven Program. https://www.lexjansen.com/pharmasug/2001/Proceed/Stats/SP08_li.pdf

Silgram, M. & Schoumans, O. F. editors (2004) Modeling approaches Model parameterization, calibration and performance assessment methods in the EUROHARP project. EUROHARP 8-2004.

Snedecor, G. W. and Cochran, G. W (1989) Statistical Methods. Iowa State University Press, Ames, Iowa.

Svendsen, L.M., Staaf, H., Gustafsson, B., Pyhälä, M., Bartnicki, J., Knuuttila, S. & Durkin, M. (2013) Review of the Fifth Baltic Sea Pollution Load Compilation for the 2013 HELCOM Ministerial Meeting. Balt. Sea Environ. Proc. No. 141, 49 p.

Svendsen, L.M., Larsen S.E., Gustafsson, B., Sonesten L. & Frank-Kamenetsky D. (2020) Progress towards national targets for input of nutrients: How much is left to reach the HELCOM nutrient input targets set for a clean Baltic Sea? Online. http://www.helcom.fi/baltic-sea-action-plan/nutrient-reduction-scheme/progress-towards-country-wise-allocated-reduction-targets/

WMO (2008) Guidance to Hydrological Practices. Volume 1. Hydrology – From Measurement to Hydrological Information. World Meteorological Organization, WMO No. 168 Sixth ed. 2008. WNT, Warszawa, p. 296.

Annex 1: Mathematical description of the Mann-Kendall trend test

Trend analysis of a time series of length T and yearly loads of nutrients can be done by applying Mann-Kendall's trend test (Hirsch et al., 1982). This test method is also known as Kendall's tau (Kendall, 1975). The aim of this test is to show if a downward or upward trend over the period of T years is statistically significant, or if the time series merely consists of a set of random observations of a certain size. The Mann-Kendall's trend test has become a very effective and popular method for trend analysis of water quality data.

The Mann-Kendall's trend test is a non-parametric statistical method, which means that the method has fewer assumptions than a formal parametric test method. The data do not need to follow a Gaussian distribution as in ordinary linear regression but should be without serial correlation. Furthermore, the method tests for monotone trends and not necessarily linear trends, and it thus tests for a wider range of possible trend shapes. The direction of the monotone trends may be either downward or upward without any specific form. The power of the Kendall trend method is slightly lower than ordinary linear regression if the time series data are Gaussian distributed and the trend is actually linear, as this will encompass the slightly less restrictive assumptions. The R packages "trend" and "Kendall" includes both the Mann-Kendall trend test. The Kendall tau coefficient can easily be computed in SAS, see the code in Annex 3. Otherwise, the Mann-Kendall trend test can be computed by the formulas and steps described in the following. It is straight forward to implement this in any statistical software.

Let $x_1, x_2, \ldots, x_n$ be yearly loads of total nitrogen (TN) or total phosphorus (TP) for the years $1, 2, \ldots, n$. The null hypothesis of the trend analysis is: the n yearly data values are randomly ordered. The null hypothesis is tested against the alternative hypothesis that the time series has a monotone trend. The Kendall statistic is calculated as:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{sgn}\left(x_j - x_i\right),$$

where

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

If either x_j or x_i is missing, then $\operatorname{sgn}\left(x_j - x_i\right) = 0$ per definition.

The trend is tested by calculating the test statistic:

$$Z = \begin{cases} \dfrac{S-1}{(\mathrm{var}(S))^{\frac{1}{2}}} & S > 0 \\ 0 & S = 0 \\ \dfrac{S+1}{(\mathrm{var}(S))^{\frac{1}{2}}} & S < 0 \end{cases}.$$

The variance of S under the hypothesis of no trend is calculated as:

$$\mathrm{var}(S) = \frac{n(n-1)(2n+5)}{18},$$

where n is the number of loads in the time series.

A positive S-value indicates an upward trend and a negative value indicates a downward trend. When both a downward and an upward trend are of interest (a two-sided test), the null hypothesis of randomly ordered data is rejected when the numerical value of Z is less than the $\left(\alpha/2\right)$-percentile or greater than the $\left(1-\alpha/2\right)$-percentile (two-sided test) in the Gaussian distribution with mean value 0 and variance 1. A one-sided test can be carried out as well. The significance level α is typically 5%. The reason for evaluating Z in the standard Gaussian distribution is the fact that S under the null hypothesis is Gaussian distributed with mean value 0 and variance $var(S)$ for $n \to \infty$. The Gaussian approximation is very good if $n \geq 10$, and fair for $5 \leq n \leq 10$.

It is possible to calculate an estimate of the trend β (the slope estimate) if one assumes that the trend is constant (linear) during the period and the estimate is change per year. Hirsch et al. (1982) introduced the Theil-Sen slope estimator, which can be calculated in the following way for all pair of observations $\left(x_i, x_j\right)$ with $1 \leq j < i \leq n$:

$$d_{ij} = \frac{x_i - x_j}{i - j}.$$

The slope estimator is the median value of all the d_{ij} -values and is a robust non-parametric estimator and will generally work for time series with serial correlation and non-Gaussian distributed data. A $100(1-\alpha)$ % confidence interval for the slope can be obtained by undertaking the calculations below (Gilbert, 1987).

Select the desired confidence level α (1, 5 or 10 %) and apply:

$$Z_{1-\alpha/2} = \begin{cases} 2{,}576 & \alpha = 0{,}01 \\ 1{,}960 & \alpha = 0{,}05 \\ 1{,}645 & \alpha = 0{,}10 \end{cases},$$

in the following calculations. It is standard to use a confidence level of 5%.

Calculate:

$$C_\alpha = Z_{1-\alpha/2} \cdot (\mathrm{var}(S))^{\frac{1}{2}}.$$

Calculate:

$$M_1 = \frac{N - C_\alpha}{2},$$

and

$$M_2 = \frac{N + C_\alpha}{2},$$

where

$$N = \frac{1}{2}n(n-1).$$

Lower and upper confidence limits are the M_1th largest and the $(M_2 + 1)$th largest value of the N ranked slope estimates d_{ij} .

A non-parametric estimate for the intercept α can be calculated according to Conover (1980). The estimator is calculated as:

$$\hat{\alpha} = M_x - \hat{\beta} \cdot M_i,$$

where M_x is the median value of all the data in the time series, and M_i is the median value of $1,2,\ldots,n$. Intercept and slope can also be determined from linear regression, which is the method used in the PLC assessment.

If the time series consists of data from different seasons (i.e. monthly loads), it is possible to apply Mann-Kendall's seasonal trend test (Hirsch & Slack, 1984). This is done by calculating the test statistic S for every season separately. Subsequently, the test statistic for the whole time series is equaled to the sum of each of the seasonal test statistics. We refer to Carstensen & Larsen (2006) for a detailed mathematical description of the seasonal trend test.

Annex 2: List of 95% percentiles and 97.5 percentiles of the *t*-distribution for the different possible combinations of degrees of freedom (df)

df	**95% percentile**
14	1.761
15	1.753
16	1.746
17	1.740
18	1.734
19	1.729
20	1.725
21	1.721
22	1.717
23	1.714
24	1.711
25	1.708
26	1.706

df	**97.5% percentile**
14	2.145
15	2.131
16	2.120
17	2.110
18	2.101
19	2.093
20	2.086
21	2.080
22	2.074
23	2.069
24	2.064
25	2.060
26	2.056

Annex 3: Examples of programs for SAS and R

This annex includes programming codes and suggestions for the presented statistical methods in chapters 2 to 6. The programs will only suggest how to perform the methods and users will need to substitute their own variable names in the coding. R codes will not be presented for most of the methods, but names of the R functions and packages will be listed for the separate methods.

It is possible to download an Excel spreadsheet that perform the Dixon test (see chapter 2). This spreadsheet can be found here:

https://otscweb.tamu.edu/Risk/Tools/DixonOutlierCalculator.xls

R-code: normalizing, model (4.3):

```
log.tn <- log(tn)

log.q <- log(q)

model.43 <- lm(log.tn ~ log.q)

beta <- model.43$coefficients[2]

mse <- sum(model.43$residuals**2)/(length(log.tn)-2)

mean.log.q <- mean(log.q)

tn.n <- exp(log.tn-beta*(log.q-mean.log.q))*exp(0.5*mse)

mean.tn <- mean(tn)

mean.tn.n <- mean(tn.n)

tn.n.b <- tn.n+(mean.tn-mean.tn.n)
```

Note, that if the intercept α can be accepted to equal zero then use the command:

```
model.43 <- lm(log.tn ~ log.q-1)
```

and change the calculation of MSE to

```
mse <- sum(model.43$residuals**2)/(length(log.tn)-1)
```

R-code: normalizing, model (4.7):

```
log.tn <- log(tn)

log.q <- log(q)
```

```
model.47 <- lm(diff(log.tn) ~ diff(log.q))

beta <- model.47$coefficients[2]

mse <- sum(model.47$residuals**2) / (length(diff(log.tn))-2)

mean.log.q <- mean(log.q)

tn.n <- exp(log.tn-beta*(log.q-mean.log.q))*exp(0.5*mse)

mean.tn <- mean(tn)

mean.tn.n <- mean(tn.n)

tn.n.b <- tn.n+(mean.tn-mean.tn.n)
```

Note, that if the intercept α can be accepted to equal zero then use the command:

```
model.47 <- lm(diff(log.tn) ~ diff(log.q)-1)
```

and change the calculation of MSE to

```
mse <- sum(model.47$residuals**2) / (length(log.tn)-1)
```

R-code: trend model:

Trend models can be estimated in several different ways in R. The following R-code is one simple way of doing this using the package segmented. This R-code illustrates a trend model with two linear segments

```
library(segmented)

x <- c(1:10, 13:22)

y <- numeric(20)

## Create first segment

y[1:10] <- 20:11 + rnorm(10, 0, 1.5)

## Create second segment

y[11:20] <- seq(11, 15, len=10) + rnorm(10, 0, 1.5)

## Plot it

par(mar=c(4,4,1,1)+0.2)

plot(x,y, ylim=c(5, 20), pch=16)

lin.mod <- lm(y~x)

segmented.mod <- segmented(lin.mod, seg.Z = ~x, psi=14)
```

```
plot(segmented.mod, add=T)
```

SAS codes:

Dixon 4 sigma: We have not programmed this (we use the R function), but it is described in Li & Edwards (2001) how to program this test in SAS.

Kendall's tau:

```
PROC CORR DATA=X KENDALL;
```

Normalisation:

Normalisation for one station (a program for several stations can be written easily):

```
data dk;

  set helcom.plc_data;

  if source_code="SCDE00001";

run;



data aar_sum;

  set dk;

  log_tn=log(tn);

  log_q=log(flow);

run;



proc sort; by source_code year;



proc univariate data=aar_sum noprint;

  by source_code;

  var log_q;

  output out=q_m mean=qmean;

run;
```

```
proc sort data=aar_sum; by source_code year;

data differens;
  set aar_sum;
  by source_code year;
  aar_s=year-1994;
  lag_tn=lag1(tn);
  lag_q=lag1(flow);
  lagl_tn=lag1(log_tn);
  lagl_q=lag1(log_q);
  d_tn=tn-lag1(tn);
  d_q=flow-lag1(flow);
  dl_tn=log_tn-lag1(log_tn);
  dl_q=log_q-lag1(log_q);
  if first.source_code then lag_tn=.;
  if first.source_code then lag_q=.;
  if first.source_code then lagl_tn=.;
  if first.source_code then lagl_q=.;
  if first.source_code then d_tn=.;
  if first.source_code then d_q=.;
  if first.source_code then dl_tn=.;
  if first.source_code then dl_q=.;
run;

proc model data=differens;
  by source_code;
   parms a b;
   log_tn=a+lagl_tn+b*(log_q-lagl_q);
```

```
  fit log_tn  / fiml;
run;
```

If a is not significant then use this:

```
proc model data=differens;
  by source_code;
   parms b;
   log_tn=lagl_tn+b*(log_q-lagl_q);
   fit log_tn  / fiml;
run;
```

```
data helcom.tidsserie_norm_tn_aalbek;
  set differens;
```

This need to be typed in, but SAS-code can be written so it's not neccesary.

```
  qmean=-0.979641718;
  mse=0.0480;
  beta=1.390908;
  residual=log_tn-(beta*log_q);
  tn_norm=exp(beta*qmean+residual)*exp(0.5*mse);
```

Biascorrection:

```
  tn_norm_k=tn_norm+(40.0763083-38.2152548);
run;
```

```
proc univariate;
var tn tn_norm tn_norm_k;
run;
```

Trend model:

```
data bap;
  set helcom.raadata_waterborne;
  if basin="BAP";
```

One linear regression:

```
proc model data=bap;
   parms a b;
   tnnorm=a+b*year;
   fit tnnorm  / fiml;
run;
```

One step change point and two linear regressions:

```
proc model data=bap;
   parms a1 b1 a2 b2;
   if year <2010 then tnnorm=a1+b1*year;
   if year >= 2010 then tnnorm=a2+b2*year;
   fit tnnorm  / fiml;
run;
```

One change point and two linear regressions:

```
proc model data=bap;
   parms a b d;
   tnnorm=a+b*year;
   if year >= 2010 then tnnorm=a+b*year+d*(year-2010);
   fit tnnorm  / fiml out=pred_diffus outpredict outest=par_diffus outcov;
   test 'Zero slope in the last segment' b+d=0, / lr;
run;
```

One change point and constant in the last segment:

```
proc model data=bap;
   parms a b c;
   tnnorm=a+b*year;
   if year >= 2010 then tnnorm=c;
   fit tnnorm  / fiml;
run;
```

Two change points and three linear regressions:

```
proc model data=bap;
   parms a b d c;
   tnnorm=a+b*year;
   if 2002 <= year < 2009 then tnnorm=a+b*year+d*(year-2002);
   if year >= 2009 then tnnorm=a+b*year+d*(year-2002)+c*(year-2009);
   fit tnnorm  / fiml;
   test 'Zero slope in the middle' b+d=0,  / lr;
   test 'Zero slope in the last segment' b+d+c=0,  / lr;
run;
```

Two change points, middle and last segment constant:

```
proc model data=bap;
   parms a b d c;
   tnnorm=a+b*year;
   if 2002 <= year < 2009 then tnnorm=d;
   if year >= 2009 then tnnorm=c;
   fit tnnorm  / fiml;
run;
```

Two step change points and three linear regressions:

```
proc model data=bap;
   parms a1 b1 a2 b2 a3 b3;
   tnnorm=a1+b1*year;
   if 2002 <= year < 2009 then tnnorm=a2+b2*year;
   if year >= 2009 then tnnorm=a3+b3*year;
   fit tnnorm  /fiml;
run;
```

In addition, many more models can be constructed using this setup.

Durbin Watson (chapter 5):

```
PROC AUTOREG DATA=X;
MODEL TNnorm = Year / dw=n dwprob;
```

n: is the order of autocorrelations you want to test

Loess:

```
PROC LOESS DATA=X;
MODEL TNnorm = Year;
```

STATISTICAL ASPECTS IN RELATION TO BALTIC SEA POLLUTION LOAD COMPILATION

Task under HELCOM PLC-8 project

HELCOM periodic pollution load compilation (PLC) assessments reports status and development in total annual runoff and total annual waterborne and airborne nutrient inputs to the Baltic Sea. This report deals with statistical methods for assessing and evaluating time series of annual runoff and nutrient inputs. Methods included are hydrological normalization of nutrient time series, trend analysis, change point analysis and a method for testing fulfilment of HELCOM Baltic Sea Action Plan (BSAP) nutrient reduction targets (MAI and NIC). Further is described how to fill in data gaps and to estimate the total uncertainty in nutrient inputs. These statistical methods are also included in the revised PLC guidelines.

ISBN: 978-87-7156-642-0
ISSN: 2244-999X