STATISTICAL METHODS FOR THE BALTIC SEA POLLUTION LOAD COMPILATION

Task under HELCOM PLC-9 project

Technical Report from DCE – Danish Centre for Environment and Energy No. 401 2026

AARHUS
UNIVERSITY
DCE – DANISH CENTRE FOR ENVIRONMENT AND ENERGY

STATISTICAL METHODS FOR THE BALTIC SEA POLLUTION LOAD COMPILATION

Task under HELCOM PLC-9 project

Technical Report from DCE – Danish Centre for Environment and Energy No. 401 2026

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. 401
Title:	Statistical methods for the Baltic Sea Pollution Load Compilation
Subtitle:	Task under HELCOM PLC-9 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:	https://dce.au.dk/en
Year of publication:	2026
Editing completed:	August 2026
Referees:	Henrik Tornbjerg, Department of Ecoscience, Aarhus University
Quality assurance, DCE:	Henriette Hossy, DCE - Danish Centre for Environment and Energy; Aarhus University
External comments:	There are no comments
Financial support:	Helsinki Commission (HELCOM) Baltic Marine Environment Protection Commission has kindly provided funding
Please cite as:	Larsen, S.E, & Svendsen, L.M. 2026. Statistical methods for the Baltic Sea Pollution Load Compilation. Task under HELCOM PLC-9 project. Aarhus University, DCE – Danish Centre for Environment and Energy, 58 pp. Technical Report No. 401

Reproduction permitted provided the source is explicitly acknowledged

Abstract: HELCOM periodic pollution load compilation (PLC) assessments report the status and development in total annual runoff and total annual waterborne and airborne nutrient inputs to the Baltic Sea and assess the sources of those nutrient inputs. The PLC project also follows up on the progress towards achieving BSAP 2021 updated maximum allowable inputs (MIA) and nutrient input ceilings (NICs). This report presents statistical methods for assessing and evaluating time series of annual runoff and nutrient inputs. The methods include hydrological normalization of nutrient time series, trend analysis, changepoint analysis, and method for testing fulfilment of MAI and NIC. Furthermore, the report describes how to handle data gaps, manage observations below detection limits, estimate uncertainty in nutrient inputs, and account for future effects of climate change. These statistical methods are also included in the revised PLC guidelines.

Keywords:	Statistical methods, nutrients, runoff, hydrological normalization, climate change, trend estimation, uncertainty in nutrient inputs, data gaps, outliers, LOQ data, censoring methods, trend analysis, estimation of changes, evaluation of compliance with BSAP reduction targets (MAI and NIC), HELCOM
Front page photo:	Saulkrasti, Gulf of Riga, Latvia. Photo taken by Gints Gailis.
ISBN:	978-87-7648-085-1
ISSN (electronic):	2244-999X
Number of pages:	58

Contents

Sammenfatning **5**

Summary **6**

1 Introduction **7**

2 Data gaps and outliers **12**
Identification of data gaps 12
Methods for filling data gaps 12
Handling observations below detection limits (LOD/LOQ) 13
Outliers and dubious values 18
Summary 19

3 Uncertainty of inputs (yearly input from a specific country or sub-catchment) **20**
Summary 24

4 Hydrological normalization of nutrient inputs **25**

5 Trend analysis, change points and estimation of change **28**
Mann-Kendall Trend Analysis 28
Linear Regression Analysis 28
Trend Reversals and Change Points 29
Change Point Models 31
Estimation of trend magnitude 32
Examples 33
Summary 34

6 Testing fulfilment of BSAP reduction targets **35**
Trend based Target Assessment 36
Trend-Based Models with Change Points 37
Traffic-light system 38
Testing Whether the Estimated Input in the Final Year Differs from the Reference Period 39
Interpretation of the Traffic-Light Assessment 41
Summary 42

7 Step-by-step analysis illustrated by HELCOM data examples **43**

8 Recommendations **47**
Data quality and consistency 47
Estimation of uncertainty 47
Hydrological normalization 47
Trend analysis 48
Change point analysis 48
Estimation of trend magnitude 49
Assessment of compliance with HELCOM BSAP targets 49
Comparison with the reference period 50

9 References **51**

Annex 1: Mathematical description of the Mann-Kendall trend test **54**

Estimation of Trend Magnitude: The Theil-Sen Slope Estimator 55

Estimation of intercept 56

Seasonal Mann-Kendall test 56

Annex 2: List of 95% and 97.5% percentiles of the *t*-distribution for different combinations of degrees of freedom (df) **57**

Annex 3: Examples of programs for SAS and R **58**

Sammenfatning

HELCOM-projektet Ninth Baltic Sea Pollution Load Compilation (PLC-9) omfatter udarbejdelsen af en samlet opgørelse af de vandbårne og luftbårne tilførsler af næringsstoffer samt deres kilder til Østersøen i perioden 1995-2026. Projektet følger desuden op på implementeringen af HELCOMs næringsstofreduktionsskema og evaluerer (afhængigt af den tilgængelige finansiering) effekten af de tiltag, der er implementeret for at opfylde reduktionsmålene fastlagt i Baltic Sea Action Plan (BSAP).

Opgaverne i PLC-9 kræver anvendelse af standardiserede og fagligt velbegrundede metoder, herunder statistiske. De statistiske metoder anvendes til at teste og estimere udviklingstendenser i tidsserier for næringsstoftilførsler, identificere eventuelle ændringspunkter (*change points*) samt kvantificere usikkerheder knyttet til estimater af næringsstoftilførsler. Desuden understøtter de statistiske metoder vurderingen af fremskridt mod opfyldelse af HELCOMs BSAP-næringsstofreduktionsmål (henholdsvis de maksimalt tilladelige tilførsler (Maximum Allowable Inputs, MAI) og de nationale næringsstofslofter, national Nutrient Input Ceilings (NIC)), under hensyntagen til både usikkerheder og den mellemårlige variation i næringsstoftilførslerne.

Denne rapport udgør den fjerde udgave af samlingen af statistiske metoder og principper, der oprindeligt blev udviklet i forbindelse med HELCOM PLC-6-projektet. Rapporten giver en teoretisk og matematisk beskrivelse af de anbefalede statistiske metoder og illustrerer deres anvendelse gennem eksempler baseret på data for næringsstoftilførsler til Østersøen.

Rapporten omfatter metoder til:

- identifikation og håndtering af ekstreme observationer (*outliers*);
- estimering af manglende data (*gap filling*);
- håndtering af observationer rapporteret under detektionsgrænsen (LOD) og kvantificeringsgrænsen (LOQ);
- hydrologisk normalisering af tidsserier for næringsstoftilførsler;
- estimering af usikkerheder knyttet til vurderinger af næringsstoftilførsler;
- trendanalyse og modellering af hydrologisk normaliserede næringsstoftilførsler;
- identifikation af ændringspunkter (*change points*) i tidsserier; samt
- vurdering af opfyldelse af reduktionsmålene fastlagt som MAI og NIC.

Sammenlignet med den foregående udgave introducerer den aktuelle version en ny metode til imputation (erstatning) af værdier rapporteret under LOD- og LOQ-grænserne, indeholder en revideret metode til estimering af usikkerhed og giver opdaterede beskrivelser af samtlige øvrige statistiske metoder. Desuden er alle dataeksempler blevet opdateret, så de afspejler den seneste vurderingsperiode. Endelig har rapporten gennemgået en omfattende redaktionel revision med henblik på at forbedre konsistens, læsbarhed og klarhed, samtidig med at den videnskabelige og metodiske stringens i det oprindelige dokument er fastholdt.

Summary

The HELCOM project for the Ninth Baltic Sea Pollution Load Compilation (PLC-9) includes the preparation of a comprehensive assessment of waterborne and airborne nutrient inputs and their sources to the Baltic Sea over the monitoring period 1995-2026. The project also provides a follow-up on the implementation of the HELCOM nutrient reduction scheme and (pending on funding) evaluates the effectiveness of measures implemented to achieve the nutrient reduction targets defined in the Baltic Sea Action Plan (BSAP).

The tasks undertaken within PLC-9 require the application of standardized and appropriate methodologies, including robust statistical approaches. Statistical methods are used for testing and estimating temporal trends in nutrient input time series, identify potential change points, and quantify uncertainties associated with nutrient input estimates. Furthermore, statistical methods support the evaluation of progress towards achieving the HELCOM BSAP nutrient reduction targets (the Maximum Allowable Inputs (MAI) and national Nutrient Input Ceilings (NICs), while accounting for both uncertainty and interannual variability in nutrient inputs.

This report represents the fourth edition of the compilation of statistical methods and principles originally developed within the HELCOM PLC-6 project. It provides a theoretical and mathematical description of the recommended statistical methodologies and illustrates their application through examples based on nutrient input da-ta to the Baltic Sea.

The report covers methodologies for:

- identification and treatment of data outliers;
- gap-filling of missing data;
- handling of observations reported below the limits of detection (LOD) and quantification (LOQ);
- hydrological normalization of nutrient input time series;
- estimation of uncertainties associated with nutrient input assessments;
- trend analysis and modelling of normalized nutrient inputs;
- detection of change points in time series; and
- evaluation of compliance with nutrient reduction targets MAIs and NICs).

Compared with the previous edition, the present version introduces a new method for imputing values reported below the LOD and LOQ thresholds, includes a revised approach for uncertainty estimation, and provides updated descriptions of all other statistical methods. In addition, all data examples have been updated to reflect the most recent assessment period. Finally, the report has undergone a comprehensive editorial revision aimed at improving consistency, readability, and clarity while maintaining the scientific and methodological rigor of the original document.

1 Introduction

The key pressures related to the eutrophication and water quality of the Baltic Sea are waterborne and airborne nutrient inputs.

In the Baltic Sea Action Plan from 2007 (BSAP 2007), eutrophication targets were established and based on these, preliminary maximum allowable inputs (MAI) and country-allocated nutrient reduction targets were developed and adopted.

In HELCOM Copenhagen Ministerial Declaration of 3. October 2013, the Contracting Parties agreed on revised reduction targets for nitrogen and phosphorus input. In 2021, Ministers agreed on a BSAP 2021 update, retaining MAI, replacing nutrient reduction targets with nutrient input ceilings (NICs), and updating the inputs in the reference period 1997-2003 upon which NICs are calculated.

To implement these commitments:

- Article 3 and Article 16 of the Convention on the Protection of the Marine Environment of the Baltic Sea Area, 1992 (Helsinki Convention);
- The 2021 Baltic Sea Action Plan including the revised nutrient reduction scheme, HELCOM Ministerial Meeting, Lübeck, Germany, 20 October 2021;
- The Baltic Sea Action Plan (BSAP), HELCOM Ministerial Meeting, Copenhagen, Denmark, 3 October 2013 and BSAP 2021 update;
- The HELCOM Monitoring and Assessment Strategy, HELCOM Ministerial Meeting, Copenhagen, 3 October 2013, Denmark.

The HELCOM Head of Delegations approved the Ninth Baltic Sea Pollution Load Compilation (PLC-9) in correspondence procedure in January 2024.

The overall task of PLC-9 is to prepare a comprehensive assessment of the water- and airborne nutrient inputs and their sources to the Baltic Sea during the period 1995-2026, including tracking the implementation of the HELCOM nutrient reduction scheme (MAI and NIC). Further optionally assess the effectiveness of measures to reach the BSAP targets. Optionally PLC-9 will assess inputs of selected hazardous substances to the Baltic Sea.
The project will be based on annual data on water- and airborne data on nitrogen, phosphorus and heavy metals from 1995-2026, supplemented with periodical data from PLC source apportionment assessments (1995, 2000, 2006, 2014, 2017, 2021 and 2025).

The project is organized in working packages with the following tasks:

0. Project Management.

1. Data reporting and establishing datasets.
    - Monitoring, compilation and reporting of national annual and periodical data.
    - Annual updating of the PLC-Water database and data on atmospheric inputs (PLC-Air).

- Establishing the periodic assessment data sets.
- Update of background information including information on measures.

2. Regular assessment based on annual data.
    - Annual BSEFS on actual waterborne nutrient inputs.
    - Annual update of HELCOM indicator on inputs of nutrients to the BS sub-basins (MAI).
    - Bi-annual assessment of the progress towards national nutrient input ceilings (NIC).
3. Optional: Thematic PLC-9 assessment based on annual data.
    - Assessment of nutrient inputs of big rivers 1995-2026.
    - Assessment of inputs of selected hazardous substances 1995-2025.
4. Periodic assessment based on periodic 2025 data.
    - Assessment of sources of nutrients 1995-2025.
    - Assessment of the effectiveness of measures 1995-2025.
5. Methodological support.
    - Updating guidelines and a statistical methodology report.
    - Intercalibration on heavy metals and nutrients analysis.

These tasks and objectives require standardized and appropriate methodology, including statistical methods. Statistical methods are applied e.g., when analysing the effects of weather on flow and loads through normalization of time series, assessing potential trends in nutrient inputs, identifying change points, and estimating nutrient datasets uncertainty.

Furthermore, statistical methods support the evaluation of progress towards achieving HELCOM BSAP nutrient reduction targets (MAI and IC) considering the uncertainty of nutrient inputs. Statistical methods are supporting tools in the PLC assessments enabling scientifically based regarding possible trends and strengthening robustness of the assessment of achieving BSAP nutrient reduction targets.

The report "Statistical aspects in relation to Baltic Sea Pollution Load Compilation" (Larsen and Svendsen, 2013) was developed under the PLC-6 project. The methods have been further developed and updated in subsequent projects (PLC-7 (Larsen and Svendsen, 2019) and PLC-8 (Larsen and Svendsen, 2021), and this report represents an updated version based on these improvements.

This report describes the statistical methods - including theoretical considerations - to be applied in the PLC assessments and the BSAP nutrient reduction scheme including examples. The focus is on waterborne input and evaluation of Contracting Parties' progress achieving MAI and NIC to follow up the BSAP 2021 updated nutrient reduction scheme (e.g. (HELCOM, 2024) and (Svendsen et al., 2025).

The methods described include:

- "flow" normalization of nutrient inputs,
- gap filling and handling missing data,
- handling data below LOD/LOQ observations,
- testing for outliers,
- trends and change point analysis
- calculating changes in inputs in a time series
- estimation of dataset uncertainty
- statistical testing the fulfilment of the reduction targets considering the uncertainty on inputs.

The report includes examples of quantified uncertainty on riverine inputs.

The statistical methods are generally also applicable for the inputs of hazardous substances.

A summary of the statistical methods is included in the present PLC guidelines (HELCOM (2022) and will be updated in the updated PLC guidelines as an item under the PLC-9 project.

The statistical procedure for analysing trends in the normalized nutrient input values plays an important role in the pollution input compilation assessments. Preparation of data for trend analysis requires careful assessment of data quality, including identification of gaps, outliers, and censored data. Methods for handling these issues are presented in Chapter 2 including a proposal on how to fill in data gaps/missing data in input time series, handling observations below the detection limit, and how to test outliers in the data.

Assessment of the variability in time series is essential for evaluating degree of variance. By reducing some components, the trend analysis will be more powerful. Methods for estimating variance components and total uncertainty are described in Chapter 3.

Hydrological ("flow") normalization is a key final step in preparing data, reducing the influence of weather-related variation, smoothing the load time series. Climate change effects must be considered when analysing trends in runoff. Hydrological normalization and climate change effects are described in Chapter 4. Several 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 (originally on CART = country allocated reductions requirement) assessment under PLC 6 and 7. This method is known as the Mann-Kendall's trend test. Trend analysis methods, including the Mann–Kendall test and regression techniques, are described in Chapter 5. Since PLC-8, the Mann–Kendall method has primarily been used as a preliminary screening tool analysing on possible trends in TN, TP and flow time series, while final trend estimation and assessment on fulfilment of MAI and NIC rely on linear regression and parametric testing.

Chapter 6 presents the method used (as in HELCOM (2024) and Svendsen et al., 2025), for assessing and testing fulfilment of MAI and NIC. It includes a

traffic light system on inputs for determining whether MAI or NIC for a Baltic Sea basin or country by basin is:

1. achieved
2. it is not possible to judge if they are achieved due to statistical uncertainty.
3. not achieving the maximum allowable inputs (MAI) or the nutrient inputs ceilings per country and basin (NIC).

Chapter 7 provides step-by-step examples using real input data from the HELCOM PLUS database (HELCOM Pollution Load User system) to exemplify the practical use of some of the proposed methodologies.

Chapter 8 includes discussion on the different methods presented for normalizing, trend testing and estimating variance components, filling gaps, and testing the fulfilment of reduction targets. It provides recommendations for applying methods for the different statistical tasks in the PLC-9 assessment.

Annex 1 includes a description in-depth of the mathematical treatment of the Mann-Kendall trend test.

Annex 2 provides a table with selected percentiles of the t-distribution for different combinations of degrees of freedom.

Annex 3 provides a link to SAS and R scripts which can be applied to do the statistical analysis.

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

The statistical methods described in chapter 3-6 should be applied on calculated/assessed data as calculated runoff, loads and inputs. The methods in chapter 2 can be applied to raw data and calculated/assessed data, respectively.

The report includes several changes compared with the previous version (Larsen and Svendsen, 2021):

- Chapter 1: Minor updates to the text including the main aims of the PLC 9 project.
- Chapter 2: Several pages including an example of how to handle (and estimate values) observations below the detection/quantifications limit (LOD/LOQ) have been included.
- Chapter 3: A major revision has been performed, removing the text regarding the direct method for uncertainties and introducing an indirect method, which is much easier to apply.
- Chapter 4: Text concerning an unused normalization method for hydrological normalization has been deleted. The question of how to address climate change in relation to significant trends in runoff is included. examples have been updated.
- Chapter 5: Includes a suggestion on how to handle the effects of climate changes, otherwise there are no major changes besides updating figures.
- Chapter 6: No major changes besides updating figures.
- Chapter 7: No major changes besides removing first part of the chapter and updating examples and figures.

- Chapter 8: The chapter has been updated and revised.
- Chapter 9: Some references added, old references replaced or removed.
- Annex 3: Content has been removed, and the annex now only includes a link to R and SAS programs (script) for the statistical tests mentioned in the report allowing PLC IG to make some of these analyses on their national data.

The authors want to express their gratitude for the funding provided by the Helsinki Commission and Aarhus University. Furthermore, the authors would like to thank the participants of the PLC-9 implementation group, HELCOM Source to Sea and HELCOM RedCore DG, for comments and input, and HELCOM Secretariat for their help and support.

2 Data gaps and outliers

The reliability and power of all statistical methods and statistical analyses, such as normalization of time series of nutrient inputs and trend analysis of subsequent trend analysis 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, data handling errors, inconsistencies with previously reported data and discrepancies compared with data from comparable sources. Errors may also occur transferring data between databases.

The methods described in this chapter should primarily be applied to raw data but can also be used on processed and interpolated data.

Identification of data gaps

A key step in data quality assessment is the precise identification of gaps in the dataset. This includes determining:

- which variables are missing?
- the duration of missing periods
- the type of missing data (e.g. not measured, measured but not reported).

Data gaps in time series on nutrient input may occur for several reasons:

- Missing measurements in a sub-catchment over certain periods.
- Missing nutrient concentration measurements.
- Missing runoff measurements.
- Simultaneous absence of nutrient and runoff data over certain periods.
- External conditions preventing measurements (e.g. ice cover or flooding).
- Lack of reporting data for unknown reasons.
- Removal of concentrations and/or runoff and/or nutrient input data evaluated as suspicious without replacement values by the data provider.
- Measurements recorded as below the LOD (Limit of Detection) or the LOQ (Limit of Quantification).

Methods for filling data gaps

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

- **Mean value of a statistical distribution**: Missing values may be estimated using the mean of a statistical distribution derived either by including all relevant data on the given sub-catchment or from a subset of data (e.g. estimating missing data from point sources in the beginning or end of a time series).
- **Mean of adjacent values**: If x_a and x_c are perceived as two adjacent time series values with one x_b missing, then:

$$x_b = \frac{x_a + x_c}{2}. \quad (2.1)$$

- **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) and for $k = 1, \cdots, n$ can be estimated as:

$$x_k = x_a + k \cdot \frac{x_b - x_a}{n+1}. \quad (2.2)$$

- **Relationship based methods**:
    - If runoff (q) is known and the rating curve can reflect the relation between nutrient input and runoff
    - Use of q-q relationship to estimate missing runoff values; a good q-q relationship can often be established from a nearby river.
    - Use a load-load relationship with other rivers to which a high load correlation can be verified.
    - If possible, model estimations of unmonitored catchment inputs – otherwise, estimate inputs from comparable neighbouring catchments with similar conditions.
- **Method for LOD/LOQ observations**:
    - 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.

Handling observations below detection limits (LOD/LOQ)

Observations below the limit of detection (LOD) or limit of quantification (LOQ) require special treatment.

LOD is defined as a threshold value in which measured nutrient concentration values are not considered to be significantly different from a zero-value at a given level of significance probability (Helsel, 2012). Measurements above LOD are evidence of a non-zero value, indicating the presence of a substance e.g. in a water sample at a given probability (Helsel, 2012).

The LOQ (Limit Of Quantification) is defined as the threshold value above which single measurement can be reliably quantified and reported (Helsel, 2012).

In the HELCOM PLC guidelines (HELCOM, 2022) it is mentioned how to use the LOQ to assign a numeric value to all low-level concentrations: if an observed concentration is given as below LOQ, an estimated concentration should be calculated by the equation:

$$\hat{c} = \frac{(100 - A) \cdot LOQ}{100},$$

with A the percentage of observations below LOQ. If $A > 50\%$, then it is defined that:

$$\hat{c} = \frac{LOQ}{2}.$$

This is a substitution method with substituted values in the interval from one-half LOQ to LOQ. In the PLC guidelines there are no descriptions of what to do in situations where multiple detection limits exist for the same series of observations.

The suggested PLC-method of substituting the values below detection limits with a fraction of the detection limit is "fabrication" of data values according to Helsel (2006), and the method produces poor summary estimates and loads. Excluding such observations is an even more problematic handling of left-censored data, as it introduces positive bias (Helsel, 2006).

It has been a common practice in water chemistry to use one-half of the detection limit for substitution of the under the limit observations. This also produces poor estimates (Helsel, 2006). Imputation methods developed for censored data provide a more robust approach to handling censored data and out-perform the substitution methods to a very high statistical degree (Helsel, 2006; Helsel, 2012). These methods are viewed as more correct ways of handling LOD/LOQ values.

In this report, the handling of nutrient concentration measurements below a LOD/LOQ will focus on determining estimates for these measurements. Estimates that can be used for a more correct calculation of summary statistics and of nutrient loads. To handle LOD/LOQ for calculating summary statistics and loads, it is proposed, in this report, to use the robust imputation method described below.

The estimation method is from Helsel (2012) and is an imputation (replacement) method applying least squares regression on a probability plot ("Regression on Order Statistics", ROS).

Statistics with measurements/observations indicated as below a limit LOD/LOQ is called left censoring. The suggested ROS imputation method assumes that the nutrient concentrations (possibly after logarithmic transformation) follow a normal distribution. Before introducing the methodological details and formal equations, the proposed approach is described in qualitative terms below.

A ROS method estimates a simple linear regression using the uncensored data (raw or log-transformed) versus their z-scores (as the explanatory variable), i.e. the coordinates from a normal probability plot (Wilk and Gnanadesikan, 1968). The method fits a normal distribution by fitting a linear function to the data, and the estimates of intercept and slope are then estimates of expectation and standard deviation of the normal distribution.

Values for each individual censored nutrient observation are predicted using the simple regression model and using their normal scores. The predicted values are imputed and used with uncensored observations for subsequent calculations and analyses. The procedure can be summarized as follows:

1. Fit a regression model using uncensored observations.
2. Calculate plotting positions (probabilities).
3. Use the regression model to estimate values for censored observations.
4. Combine estimated and observed values to obtain a complete dataset.

If log transformation is applied, estimated values are back transformed using the exponential function. The method provides a full dataset to be used in load calculations or for calculating summary statistics of the full set of measured concentrations.

The presented ROS method is robust regarding back-transformation bias. The bias is avoided since the censored observations are imputed based on a parametric model followed by combining them with uncensored observations to calculate summary statistics. This avoids the transformation of means and variances. Issues with transformation of bias are eliminated by back-transforming the imputed values and uncensored observations rather than calculated summary statistics, and by calculating summary statistics after the back-transformation.

The robust imputation method is described in detail in Section 6.4 in Helsel (2012). In general, let us assume that we have multiple detection limits. First step is calculating the probability P_j of exceeding each given detection limit using the following equation for each value of *j*:

$$P_j = P_{j+1} + \frac{A_j}{A_j + B_j} \cdot \left(1 - P_j\right),$$

Where:

j: is indexing the different detection limits.

A_j: is the number of uncensored measurements detected between the *j'*th and *(j+1)'*th detection limit.

B_j: is the number of measurements (both censored and uncensored) below the *j'*th detection limit.

If *j* is the highest numerical detection limit, we define $P_{j+1} = 0$ and $A_j + B_j = n$, with n equal the total number of measurements. The number of censored measurements C_j below the *j'*th detection limit is calculated as:

$$C_j = B_j - B_{j-1} - A_{j-1}.$$

If *j* represents the lowest numerical detection limit, then $A_{j-1} = B_{j-1} = 0$. In the next step, so-called plotting positions must be calculated for both uncensored and censored measurements. For uncensored measurements these are calculated as:

$$pD_{ij} = \left(1 - P_j\right) + \frac{i}{A_j + 1} \cdot \left(P_j - P_{j+1}\right), i = 1,2, \dots, A_j$$

and for censored as:

$$pC_{ij} = \frac{i}{C_j + 1} \cdot \left(1 - P_j\right), i = 1,2, \dots, C_j.$$

Finally, fit a simple linear regression using only the uncensored measurements, i.e.:

$$y_{ij} = \alpha + \beta \cdot \Phi^{-1}\left(pD_{ij}\right) + \varepsilon_{ij}, i = 1,2, \dots A_j, for\ all\ j.$$

In the equation y_{ij} indicate observations, either raw or transformed and Φ is the probability distribution function for the standard normal distribution.

Let $\hat{\alpha}$ and $\hat{\beta}$ denote the parameter estimates. The y-values for the censored measurements are predicted as:

$$\widehat{y_{ij}} = \hat{\alpha} + \hat{\beta} \cdot \Phi^{-1}\left(pC_{ij}\right), i = 1,2, \dots C_j, for\ all\ j.$$

If data has been log-transformed the estimates for the censored measurements must be back-transformed by the exponential function, i.e.:

$$exp\left(\widehat{y_{ij}}\right), i = 1,2, \dots C_j, for\ all\ j.$$

Censored values are ranked in the method, see Table 2.1. To avoid the method introducing artificial trends in nutrient concentrations, the rank values must be chosen at random for each of the groups of censored values, for example if the j'th group has 5 censored measurements the 5 rank values must be chosen randomly and with equal probability of $1/5$. If the nutrient concentrations are used for load calculations, the censored measurements must be randomly mixed according to measurement date and time, but this is not required if only summary statistics are calculated.

The method is illustrated in Table 2.1 applying a constructed data example.

The regression estimates are $\hat{\alpha} = -0.6333$ and $\hat{\beta} = 2.3345$.

The above method does not use the information regarding runoff. Alternatively, a concentration-flow relationship (rating curve) could be established for all the uncensored measurements and use the relationship to estimate values for the censored measurements.

If all measurements are under one LOD/LOQ, it could be obvious to define that all these observations are zero. This is incorrect as the observations contain a substantial amount of information. However, interpreting the data requires certain assumptions. In the following we assume all observations are below one single LOD/LOQ. Summary statistics can be estimated assuming a known probability distribution for the observations by using the maximum likelihood method (Helsel, 2012). In many, if not all cases, it is reasonable to assume concentration of a substance follows a log-normal distribution (Helsel, 2012).

The maximum likelihood method produces very imprecise summary statistics. Thus, the available data is insufficient to produce reliable estimates of the mean and standard deviation. If ancillary information regarding the probability distribution – such as distribution type and distribution moments (e.g. expectation and variance) – is available from prior measurements or from comparable or nearby or similar streams and rivers; this information should be considered, as it can improve the precision of the estimates. However, substituting censored observations with zero or any value up to the quantification or detection limit is equally inadequate as applying the maximum likelihood method without incorporating such additional information.

Table 2.1. Applying ROS (least square regression on a probability plots) to handle measurements below LOQ/LOD. TN concentrations are constructed in this example. The total number of measurements is 20, and there is only 1 LOQ = 0.1 mg/l. There are in total 5 measurements below the LOQ. We assume that the 5 measurements below 0.1 mg/l have been ranked randomly from 1 to 5.

Measurement	Log(measurement)	Probability of detection	Plotting position	Rank	Predicted log-values	Measurements plus predicted
20.15	3.00		0.953	20		20.15
11.11	2.41		0.906	19		11.11
7.30	1.99		0.859	18		7.30
5.43	1.69		0.813	17		5.43
3.62	1.29		0.766	16		3.62
2.40	0.88		0.719	15		2.40
1.37	0.31		0.672	14		1.37
1.13	0.12		0.625	13		1.13
0.79	-0.24		0.578	12		0.79
0.51	-0.67		0.531	11		0.51
0.41	-0.89		0.484	10		0.41
0.35	-2.05		0.438	9		0.35
0.31	-1.17		0.391	8		0.31
0.22	-1.51		0.344	7		0.22
0.15	-1.90		0.297	6		0.15
<0.1		0.75	0.208	5	-2.53	0.080
<0.1			0.167	4	-2.89	0.056
<0.1			0.125	3	-3.32	0.036
<0.1			0.083	2	-3.87	0.021
<0.1			0.042	1	-4.67	0.009

If neither maximum likelihood estimation nor substitution methods provide an appropriate solution in cases where all observations fall below quantification or detection limit (i.e. "LOW" or "<DL"), then according to Helsel (2012, chapter 8), limited alternatives remain beyond reporting all the individual observations, as well as summary statistics (mean, median etc.), as "<LOQ" or "<DL". This approach reflects the actual limitation of the data and does not constitute "fabrication". However, it poses a significant challenge for substances with e.g. potential health or lethal effects when concentrations are close to the laboratory LOQ or DL.

The limitation of substitution methods can be illustrated by the following example: consider a heavy metal with a DL = 10 µg/L, where all observations are substituted with one-half the detection limit (i.e. 5 µg/L). Assuming the threshold for health concern for this heavy metal is 7 µg/L, such substitution may lead to the erroneous conclusion that the regulatory standard of this heavy metal is met. This conclusion is not justified, as the only valid inference is that the concentration is below 10 µg/L. To enable a more accurate assessment, measurements with a lower detection limit would be required.

Most methods for trend analysis, such as the non-parametric Mann-Kendall's trend method and linear regression (see chapter 5), can accommodate missing values, particularly when these occur within the interior of the time series rather than at its boundaries (e.g. the first 2 or the last 2 years). The trend test is only negligibly affected by a few missing values. However, the statistical power of trend tests decreases when the time series contains gaps, making it more difficult to detect a statistically significant trend. If a large proportion of missing values have been estimated and the imputed values are identical for many years, trend testing should be avoided, as the resulting variability will

be artificially reduced compared to the variability based on the actual observed data.

Above, various methods for filling in gaps have been described. In practice, the choice of methods will depend on circumstance, however, the following hierarchical ranking is suggested:

1. A model-based approach – e.g. a regression model type, to estimate nutrient load or runoff.
2. Linear interpolation
3. Imputation
4. Values derived from look-up tables or provided by experts
5. No gap filling: the time series is used as it is, and any assessments are carried out subsequently.

Outliers and dubious values

Outliers are data values that are extreme compared to other reported values for the same locality (e.g. country, basin, catchment, etc.) and should be identified and flagged using formal outlier detection (testing methods) such as:

- Dixon's Q test (Dean and Dixon, 1951).
- Z-score and interquartile range methods (Box and whisker plot).
- Grubbs test (Grubbs, 1950).
- Experience-based thresholds defining plausible maximum and minimum values) that are unlikely to be exceeded or fallen below.
- Water quality standards (interval values or limits) were available.

It is important to note that outliers are not necessarily erroneous data, however they require extra careful evaluation before being included in statistical analysis.

Suspect or dubious values are those that do not fulfil the criteria for formal outlier by the outlier tests but deviate significantly from the remaining observations in the time series, or are considered unreliable, e.g. a load value inconsistent with reported runoff or derived from a neighbouring sub-catchment. Such values may arise when measurements in a sub-catchment are available only for a limited period, when changes in laboratory procedures or standards occur, or when modifications in measurement methods lead to abrupt shifts in data values. They may also occur when identical values are reported over several consecutive measurements or years. In addition, calculation errors may arise due to incorrect units, faulty water samples, laboratory errors such as sample misidentification of samples or parameters.

Suspect or dubious values should be corrected and treated as formal outliers unless they can be verified as valid.

If a dubious value is determined and deemed incorrect it will be excluded from the assessments, and cannot be corrected by the Contracting Party, it should be removed from the PLC database by the Contracting Party. If a reported data value is classified as an outlier and excluded from assessments, it may be replaced using one of the methods listed above for handling gaps. However, gap-filling or replacement cannot fully substitute for measured

data; therefore, observed data or, where appropriate, reliable model-based estimates should always be preferred. It should be stressed that filled-in data gaps must be clearly marked in the PLC database.

To illustrate the applications of the Dixon outlier test, we consider the total phosphorus loads (TP) from the Pregolya River (Russia). Figure 2.1 presents a box-whisker plot of TP loads for the period 1995-2018, showing one apparent outlier at 880 tonnes TP per year. Application of the Dixon test in R confirms that the value in 2017 (880 tonnes TP) is an outlier within the series the annual TP load series (P=0.007). In contrast, the corresponding time series of total nitrogen (TN) loads for the same period do not contain any outliers.

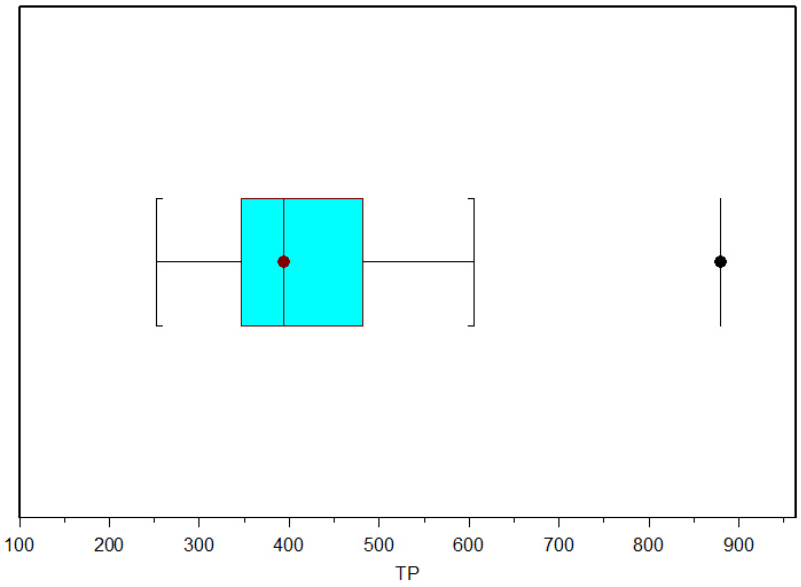


Figure 2.1. Box-and-whisker plot of annual total phosphorus (TP) riverine loads (tons) in the Pregolya River for the period 1995-2018. The Dixon outlier test identifies one statistically significant outlier (value 880 tonnes P).

Summary

The methods presented in this chapter improve data quality and enhance their suitability for statistical data analyses, which are addressed in the following chapters. The assumptions required for valid statistical analysis are more readily fulfilled, when gaps are filled, outliers are identified and evaluated, and censored data are appropriated imputed.

3 Uncertainty of inputs (yearly input from a specific country or sub-catchment)

Time series of nutrient inputs exhibit year-to-year variation due to the contributions from multiple sources. One such source is a potential trend in inputs over time. Consequently, time series are, by standard, detrended prior to analysis of variance components from other sources, as trend-induced variation are not of primary interest when estimating variance components. Detrending involves removing any significant trend from time series before calculating uncertainty. This can be achieved in several ways; for example, by calculating residuals between the observed and model-predicted values using a trend model, e.g. a linear regression model.

Variation is observed within the yearly values, and it is therefore assumed that the yearly inputs represent samples, sampled from the same statistical population, characterized by a given mean value and variation (after detrending). This variability represents an estimate of a component of the total uncertainty associated with a given annual input, i.e. the standard error of the mean. The remaining component is potential bias in the calculation of the annual loads, which in principle cannot be directly inferred from the values of the time series.

Total uncertainty (bias plus variation) is a very complex function (based on several assumptions) of multiple contributing components, including:

- Uncertainty related to the applied field sampling and field measuring methods (on nutrients/metals and other substances concentrations field sampling/measurements, on water stage and velocity measurements etc.).
- Laboratory uncertainty (e.g. applied chemical analysis methods in the laboratory, changing laboratories over time).
- Uncertainty associated with sampling design (frequency, location, timing) and methods used for calculating runoff (either stage-discharge relationship or other methods) and loads (based on combined concentrations and runoff).
- Variation induced by interannual weather variation (e.g. precipitation amount, type, and distribution of rainfall and changes in storage component such as snow/ice, soil moisture, and groundwater).
- Uncertainty in estimating inputs from unmonitored areas including bias from omitted unmeasured inputs and uncertainty in estimating methods for unmonitored inputs).
- Uncertainty in inputs from direct point sources, including sampling, analytical errors, etc.
- Additionally, other less well-characterized sources contribute to uncertainty.

Most countries have reasonable knowledge of laboratory uncertainty – particularly for nutrients (for HELCOM PLC analysis see e.g. Lassen and Larsen, 2025) - this component is generally well documented and may not be the dominant contributor to total uncertainty. Many other sources are complex and some of them very difficult to estimate in practice due to limited or absence

of empirical data. In general, uncertainty can be reduced by optimizing sampling design (time and location of sampling, more samples) and implementing monitoring programs considering variations in both concentrations and runoff. Strategic monitoring (optimized monitoring program), supported by improved and more precise analytical techniques as robust/optimized methods estimating inputs from unmonitored areas, is the most effective way to reduce uncertainty.

A detailed decomposition of individual uncertainty components is not required for testing trends or compliance with the agreed targets assessment. However, variance component analysis is used in statistics when optimizing sampling design in a hierarchical sampling regime and/or to test effects (e.g. of mitigations measures).

Within PLC assessments, the total uncertainty of detrended nutrient input time series (across countries, sub-basins, etc.) is relevant for evaluating comparability and identifying significant differences in time series uncertainty (for some countries, sub-basins, etc.,). Analysis of variance components can help explain such differences and to understand the factors causing such differences. The ultimate result of evaluating uncertainty is to improve data quality and consistency across Contracting Parties.

All Contracting Parties are required to report uncertainty estimates for annual TN, TP and runoff. However, reporting individual uncertainty components is not required.

A standardized methodology is required for estimating the uncertainties in the national datasets from monitored areas.

Total uncertainty reflects the closeness of the measurements (or estimates) to the true value and comprise both bias and precision. Theoretically, total uncertainty is the Root Mean Squared Error (*RMSE*) as defined by:

$$RMSE(\hat{X}) = \sqrt{(\sigma^2 + \beta^2)}. \quad (3.1)$$

Here

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

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

E_X denotes the theoretical mean value with respect to the random variable X, and $\hat{X}$ is the estimate of the true value X.

Uncertainties in input estimates from unmonitored areas must be accounted for to estimate the total sub-catchment scale uncertainty. Similarly, uncertainty in direct inputs can be estimated using the same formula (framework) as above.

Some uncertainty components remain poorly or not quantified. However, uncertainty in individual runoff estimates is relatively well understood and is typically less than ± 5% (Herschy, 2009; WMO,2008). The precision on daily discharges depends on measurement frequency and monitoring positions in the cross section: for examples in Denmark for open-channel gauging stations in streams channels, it is approx. 8% (given as standard deviation) with 10

annual measurements, decreasing to approx. 6% with 12 measurements and to less than 1% with more than 40 annual measurements (Kronvang et al. 2014). Model-based runoff estimates may be associated with higher uncertainty. For chemical analysis, Denmark standards require total (expanded) uncertainty for TN and TP to be below 15% (or at low concentration values: 0.1 mg N l^{-1} and 0.01 mg P l^{-1} in freshwater, respectively, and correspondingly 5 mg N l^{-1} and 1 mg P l^{-1} in wastewater).

As mentioned, precision can be estimated from the variance of detrended input time series (or from time series without trend). However, this approach does not account for systematic bias in e.g. runoff or inputs and instead reflects estimation of variability around the mean.

If strong serial correlation is present in a nutrient input or runoff time series to the Baltic Sea, the standard error may be underestimated, and the precision is accordingly underestimated. In this report, serial correlation in an annual time series of nutrient inputs is assumed to be weak or statistically insignificant; allowing standard error estimates to serve as reasonable approximation of to the true standard error value.

Bias (systematical under or over estimation) is inherently difficult to quantify. Earlier studies (Kronvang and Bruhn, 1996) indicate that linear interpolation estimations may systematically underestimate nutrient loads (negative bias), whereas more recent work (van't Veen et al., 2025) indicate that both positive and negative biases may occur, depending on the observed nutrient concentrations (sampling characteristics), stream/river type and on the hydrological regime. Bias is expected to be smaller in large rivers with rather stable flow regime, and larger in rivers systems with highly variability in, particularly for phosphorus due to higher concentration variability and risk of not sampling the high concentrations (Kronvang and Bruhn, 1996). However, a precise quantification of the size of the bias is impossible to get. Increasing sampling frequency reduces bias (to zero with infinite sampling) but does not eliminate it under typical monitoring regimes (12–24 samples per year).

Similarly, the variance of the load estimates is difficult to determine analytically. A few studies (Kronvang and Bruhn, 1996; van't Veen et al., 2025) have shown variance is relatively small, less than 15% for both total nitrogen and total phosphorous. The load calculations approximate a continuous integral (Riemann integral approximated by a Riemann sum) using discrete observations, often with complex autocorrelation structures. Limited sample sizes further constrain statistical inference. Increasing sampling effort reduces variance, but no widely accepted direct method exists for precise estimation.

Therefore, an indirect method/approach is recommended determining annual load variance. In summary, the variance can be estimated from detrended (normalized) time series of annual loads or runoff. Detrending removes most of the anthropogenic trends, while normalization (see Chapter 4) reduces interannual variability driven by meteorological factors. Residual variation is then primarily attributed to sampling, analytical uncertainty, and load estimation methods. The normalization methods are described in chapter 4 in this report, while detrending is achieved by calculating residuals from the trend models described in chapter 5.

This approach assumes stationarity and ergodicity of the normalized and detrended time series to be applicable. Stationarity guarantees that distributional properties regarding probability are the same regardless of time. Under the assumption of stationarity, ergodicity refers to the property that time averages (e.g., the sample mean and variance) computed from a single realization of a time series converges to the true expectation and variance of the underlying random variables. Under these assumptions, the sample variance provides an estimate of the theoretical variance of annual inputs.

In mathematical terms the method is: Let $\{L_{N_i}, i = 1,2, \dots, n\}$ denote the time series of normalized loads; *let* $\hat{g}(i)$ denote the deterministic trend component governing the trend of the normalized loads time series over time. For example, a simple linear trend:

$$\hat{g}(i) = \hat{\alpha} + \hat{\beta} \cdot i.$$

Detrended time series is calculated as:

$$dL_{N_i} = L_{N_i} - \hat{g}(i) + \hat{g}(\bar{\imath}).$$

Here

$$\bar{\imath} = \frac{1}{n}\sum_{i=1}^{n} i = \frac{n(n+1)}{2} = \frac{n+1}{2}.$$

For a simple linear trend, the detrended time series is calculated as

$$dL_{N_i} = L_{N_i} - \hat{\beta}\left(i - \frac{(n+1)}{2}\right).$$

The total uncertainty is then estimated as.

$$\sqrt{\frac{\sum_{i=1}^{n} dL_{N_i}^2}{n}}.$$

The used notation is:

$\hat{\alpha}$ and $\hat{\beta}$: Regression parameter estimates.

L_N: Normalized loads.

dL_N: Detrended normalized loads.

This estimate of variance may be slightly too conservative (too large) due to incomplete removal of variability from climate/meteorology and because anthropogenic impact might be described slightly too simplistic by a linear trend model. However, the estimate is considered reasonable. Chapter 7 provides examples of the indirect method.

For several streams/rivers the total uncertainty sum of aggregated loads can be propagated assuming independence as follows. Let $s_i, i = 1, \cdots, n$ denote the total uncertainties for n separate rivers. Total uncertainty s for the sum of the n loads is calculated as:

$$s = \sqrt{\sum_{i=1} s_i^2}.$$

Summary

Estimating uncertainty in annual nutrient loads enables identification of rivers or catchments with excessively uncertain load estimates, thereby highlighting where sampling effort is insufficient or data quality is inadequate to provide good estimates of the annual loads.

4 Hydrological normalization of nutrient inputs

The annual riverine inputs of nutrients exhibit substantial interannual variability. A primary driver of this variability is runoff, which is largely governed by meteorological conditions and hydrological processes including precipitation, accumulation and melting of snow/ice, evapotranspiration, and temperature. To reduce main part of the variability attributed to hydrological factors, annual nutrient inputs are flow normalized.

Flow normalization of riverine loads is a statistical method that produces a new time series of nutrient inputs in which a substantial part of the variability driven by hydrology has been removed. The resulting normalized time series exhibits reduced interannual variability, thereby improving the power of the subsequent trend analysis. Consequently, statistically significant trends in the normalized series can, probably, mainly largely be attributed to the anthropogenic influence.

A range of methods for normalizing inputs are described in Silgram and Schoumans (ed., 2004), Chapter 4. In this report, we focus on statistical methods based on empirical data. The empirical flow-normalization method is based on a regression of annual loads against annual discharge. Specifically, loads are normalized to a mean discharge level calculated over the entire time series period. This approach removes variability associated with the total annual discharge but does not remove the effect of differences in the intra-annual distribution of discharge.

In Silgram and Schoumans (eds., 2004), normalization is applied to untransformed load and discharge data. However, empirical evidence suggests that applying a natural logarithmic transformation to both variables prior to regression improves model performance, yielding a better fit and residuals that more closely satisfy the assumption of normality.

The hydrological normalization should be regarded as a prerequisite for trend analysis, which can be viewed as a two-step process including: 1) normalization and 2) trend estimation.

Following Silgram and Schoumans (eds., 2004), the empirical hydrological normalization method assumes a linear relationship between annual discharge (Q) and annual load (L) of a nutrient:

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

a and β = linear regression parameters

ε_i = linear regression model error.

The normalized load L_N is then defined as:

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

$\bar{Q}$ = Average discharge over the whole time series period

Q_i = Runoff in year i

β = Is the estimated regression coefficient

^ = Indicates that it is an estimated parameter.

Usually, the relationship is modelled after log-log transformation, which reduces the influence of large loads and discharge values, and as mentioned, provides a slightly more precise fit with residuals more likely to be Gaussian distributed, as statistical prerequisite for the regression method. Accordingly, normalization should be based on a log-log regression between load and discharge:

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

To avoid large negative values when transforming small load or discharge values (i.e.<1), both load and discharge may be multiplied by a scaling factor (e.g., 1000) prior to log transforming.

Formula 4.3 gives the following formula for normalized loads:

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

In formula (4.3) and (4.4), "log" is the natural logarithmic function, "exp" is the exponential function.

Because normalization based on transformed data (or nonlinear models) may introduce bias, the mean of the normalized series will generally differ from that of the original not normalized series, i.e.

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

A bias correction is therefore applied on the normalized values as:

$$L_{NB_i} = L_{N_i} \frac{\frac{1}{n}\sum_{i=1}^{n} L_i}{\frac{1}{n}\sum_{i=1}^{n} L_{N_i}}. \tag{4.5}$$

The described method assumes that the discharge time series is stationary and thereby ergodic and that the functional relationship between load and discharge remains stable over time. Violations of these assumptions reduce the effectiveness of normalization. If e.g. the discharge is stationary over time, but the load is decreasing, then the form of the relationship between load and discharge changes. In practice, many load series exhibit trends, implying that the load–discharge relationship may change over time, so the applied normalization method is less than optimal, which influence the normalized values.

Furthermore, climate change is expected to introduce trends in discharge series, further challenging these assumptions.

If discharge exhibits a statistically significant trend (see next chapter for trend analysis methods), this should be incorporated into the normalization procedure. Instead of using the transformed mean discharge $\overline{\log Q}$, a simple time-dependent trend model may be applied like the trend models described in the next chapter. While trends in discharge time series currently is not a major issue for the HELCOM dataset. It may become increasingly important in the future. Continuous research is therefore needed to develop normalization

methods that are robust to climate change. At present, temporal change in the load-discharge relationship constitutes the more critical limitation in most applications.

Hydrological normalization should be carried out at the river/sub-catchment level, i.e. nutrient loads should be normalized for each river/sub-catchment separately. Aggregating prior to normalization (e.g. performed country-wise or sub-basin-wise) yields results that are not equivalent to aggregating normalized river/sub-catchment level loads, although differences are typically small.

A commonly used alternative method is the ratio method. However, this method is a purely mechanistic method and lacks any kind of statistical model or scientific knowledge. Normalized values calculated from the ratio method do not preserve the mean value and therefore requires bias correction. The regression-based approach above is generally preferred, as it exploits the typically good correlation between load and discharge.

Flow-normalization enhances the statistical power of trend analyses of the load time series by reducing noise in the data. Moreover, the use of a regression-based framework incorporates additional information about the load–discharge relationship.

Figure 4.1 illustrates the benefit of the flow-normalization procedure. It shows observed and normalized total phosphorous waterborne loads to the Baltic Sea for the period 1995-2023. Observed loads are calculated as a sum of several monitored streams/rivers and estimated inputs from unmonitored areas. Normalized loads are computed at the sub-catchment level and subsequently aggregated.

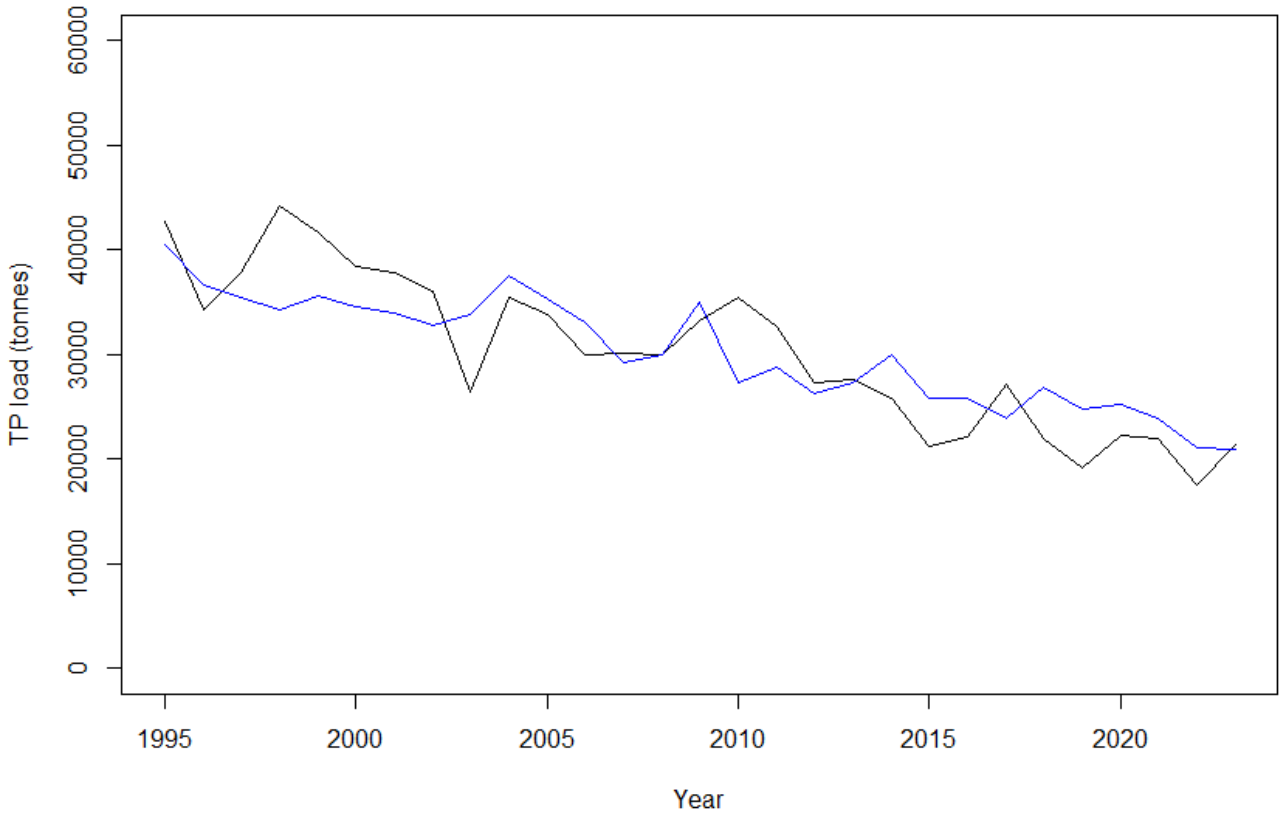


Figure 4.1. Time series (1995-2023) of annual actual (non-normalized) total phosphorus inputs to the Baltic Sea (black line) and of the corresponding flow-normalized series (blue line) using the methods described in this chapter.

5 Trend analysis, change points and estimation of change

Trend analysis of normalized nutrient input series to different parts of the Baltic Sea, including analysis of runoff trends, is an important component in the PLC assessments. It is used to evaluate whether nutrient inputs are reduced and to assess progress towards achieving the nutrient reduction targets (MAI and NIC) established under the Baltic Sea Action Plan (BSAP). Furthermore, trends analysis supports the evaluation of the effectiveness of implemented nutrient reduction measures.

The time series used in the trend analysis should always be normalized. However, the methods described below may also be applied to analyse trends in non-normalized nutrient input time series. Trend analysis can be conducted using a range of both parametric and non-parametric methods. Parametric methods include ordinary linear regression with year as the independent (explanatory) variable, as well as non-linear regression approaches, such as polynomial, exponential or more complex regression models. The most applied non-parametric methods are the Mann-Kendall trend test and the Theil-Sen estimator of annual nutrient input changes.

In addition to describing methods trend detection, this chapter addresses methods for estimating the magnitude of trends, particularly in situations where trends are not strictly linear.

Mann-Kendall Trend Analysis

The Mann-Kendall method (Hirsch et al., 1982) is a well-established method for testing the presence of a monotone trend in a time series. It is a non-parametric method and based on Kendall's tau, which is a measure of the correlation (strength and direction) between two different variables. The method is robust with respect to outliers and can accommodate a limited number of missing data/observations. When the underlining trend is linear, the Mann-Kendall test generally has slightly lower statistical power than ordinary regression analysis. A detailed mathematical description is provided in Annex 1. The R packages "trend", "rkt" and "Kendall" contain implementation of the Mann-Kendall trend test. The Kendall tau coefficient (see Annex 1) can also easily be computed in SAS.

The Mann-Kendall trend method is used in the PLC assessments for preliminary analysis of trends in the TN and TP load time series. It is also used for analysing trends in runoff time series. Subsequent trend modelling, including estimation of trend lines (e.g. slope, intercept) is based on linear regression and parametric hypothesis testing. In the first version of this report addressing statistical methods (Larsen and Svendsen, 2013), greater emphasis was placed on the use of the Mann-Kendall method.

Linear Regression Analysis

Ordinary regression analysis is a well-known statistical method for trend analysis (figure 5.1). However, it assumes a linear relationship between the response variable and time, as well as a normally distributed (Gaussian distributed) and statistical independent residuals (Snedecor and Cochran, 1989).

If a time series exhibits serial correlations both the Mann-Kendall test and ordinary linear regression require modification, as serial correlation affects the validity of the statistical test and associated p-values, and can therefore not be trusted.

Serial correlation in a time series may be assessed using the Durbin Watson test statistic (Durbin and Watson, 1971). The R package "lmtest" includes the function "dwtest()", whereas the R package "car" includes the function "Durbin-WatsonTest()". For annual time series of nutrient loads and runoff, autocorrelation generally appears to be weak and can therefore be ignored as an acceptable approximation. Consequently, the methods may normally be applied without adjustment.

A minimum time-series length of eight year is recommended for application of the Mann-Kendall test. This threshold may also serve as practical guideline for the application of ordinary linear regression.

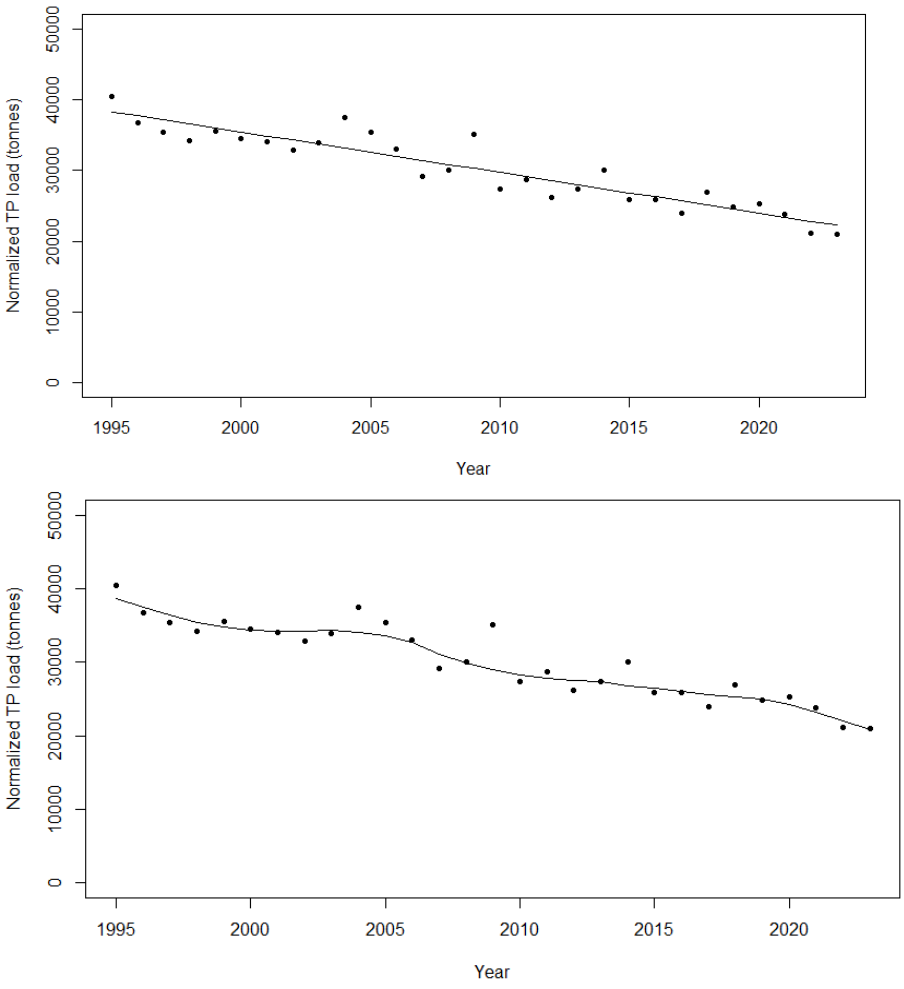


Figure 5.1. Upper figure (a): Annual normalized TP waterborne inputs (tonnes) to the Baltic Sea. The trend line is estimated applying linear regression model. Lower figure (b): The same time series as in the upper figure, but with the trend estimated using LOESS (locally weighted scatterplot smoothing) regression method.

Trend Reversals and Change Points

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

A time series plot may reveal one or more trend reversals, corresponding to change points in the series, e.g. the first part of the time series may show a

linear increase in nutrient inputs, whereas the latter part may show a linear decrease. In such cases, trend analysis may be performed using models consisting of two to four linear segments or, where sufficient observations are available, by applying separate Mann-Kendall trend tests to individual sections of the time series (example in figure 5.2).

The year of a trend reversal (i.e. the change point) may be identified either through visual inspection of the time series plot or through formal statistical method (Carstensen and Larsen, 2006). If the timing of a change in nutrient inputs is known a priori (e.g. due to implementation of major municipal wastewater treatment plants facilities or new significant management treatment interventions), this year should be used as the change point, and the time series should be divided accordingly.

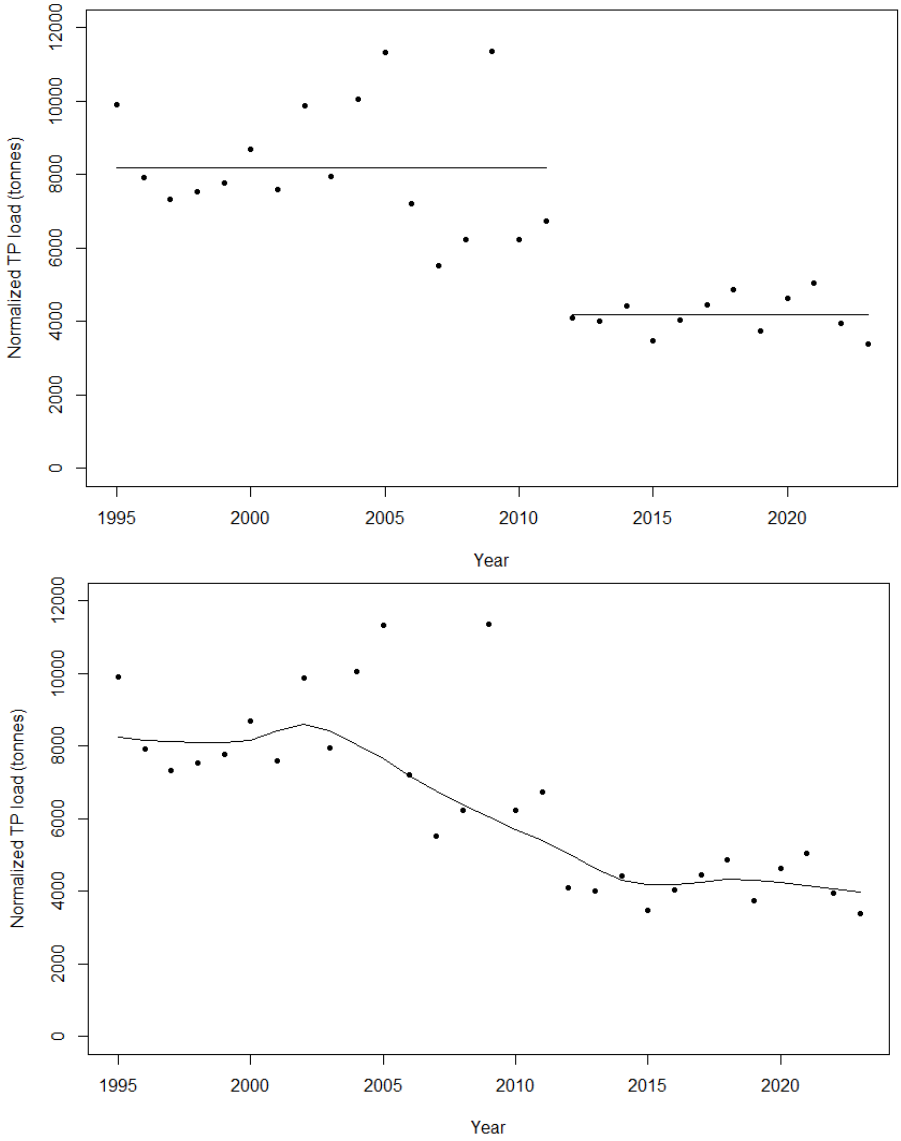


Figure 5.2. Upper figure (a): Annual normalized TP inputs (tonnes) to the Gulf of Finland 1955-2023. One change point has been identified in 2012, and trend lines are based on linear regression. Lower figure (b): The same time series analysed using the LOESS regression method.

Statistical estimation of change points in a time series is more complex and involves iterative models fitting procedures. The LOESS regression method (locally weighted scatterplot smoothing, Cleveland (1979)) may be used as a supplementary tool for detecting non-linear trends and identify change points or step changes, as illustrated in figure 5.1b and 5.2b. Similarly, generalized additive models (GAM) described by Hastie and Tibshirani (1990) provide flexible alternatives for modelling non—linear temporal patterns in the time

series. However, this flexibility may increase the risk of overfitting, particularly when applied to relatively short time series. Consequently, the present report focuses primarily on linear trend models.

Change Point Models

It is suggested to use models containing 1 to 4 linear segments for describing different sections of a time series. It should be noted that individual segments may still exhibit no statistically significant linear trends.

Change points are identified using iterative statistical procedures that determines the most significant change point based on the likelihood-ratio test statistic(-2logQ) (Carstensen and Larsen, 2006). Furthermore, each segment defined by a change point should ideally contain at least five years of observations.

Two alternative models are proposed for the identification of a single change point: Model 1: Continuous change point model, and Model 2: Discontinuous change point model.

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

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

L_N = Normalized input

α = Intercept parameter

β and δ = Slope parameters

Y = Change point year

i = Index indicating different years in the time series.

Model 1 The trend remains continuous across the change point (the two lines are connected), while allowing a change in slope, while model 2 allows discontinuity with different slopes and the level of the series may change at the change point, resulting in a step change (disconnected lines at the change point).

After identification of the first change point, a subsequent iterative procedure is conducted to determine whether an additional statistically significant change point exists.

Change point models support estimation of the normalized inputs in the final year of the time series and facilitate interpretation of the overall trend pattern across the assessment period.

Finally, in previous HELCOM BSAP MAI and NIC assessments, the significance of the slope in the final segment was evaluated. If the slope was not significantly different from zero the following model was applied:

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

c = Mean input for the final period (a constant).

For future assessments, it is recommended that all identified segments of a time series be tested for significant slopes. Segments with slopes that are not significantly different from zero should be represented by constant values where the slope can be accepted to be zero. For instance, a model like

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

Whereas segments with significant slopes should be represented by regressions lines.

Table 5.1 describes the modelling process that is recommended for fitting linear models with change points to the time series.

Table 5.1. The stepwise recommended modelling process for identifying change points, testing for significant slopes and fitting constants (where no statistically significant slopes) and linear regression parameters (with statistically significant slopes identified) in a time series.

1. step	2. step	3. step	4. step
Identify significant change point(s)	Tests for additional change points in each identified segment	Test for significant slopes in the identified segments	Fit a constant in segments with a non-significant slope. Fit linear regression parameters in the other identified segments.
No change point	Fit a constant for the entire time series		

Estimation of trend magnitude

The second component of trend analysis involves estimating the magnitude of the change over time e.g. expressed as change per year. Several methods are available, and the choice depends on the shape of the trend.

The Theil-Sen slope estimator (Hirsch et al., 1982 and Annex 1) is a non-parametric estimator that is resistant to outliers (suspicious) values. The method assumes a linear trend and estimates the change per year. If the trend is non-linear, and/or if one or more trend reversals occur within the time series, the estimator should be applied separately to individual segments.

The magnitude of a linear trend may also be estimated using regression analysis. This is the classical approach, but it is not flexible when trends deviate from linearity. The simplest method is to calculate change directly from the first and last observation in the flow-normalized input time series. However, this approach may be unreliable if either endpoint deviates substantially from the general trend.

To estimate the overall percentage change in nutrient inputs during the assessment period, normalized values predicted from the fitted trend model for the first and last year of the time series can be used:

$$100 \cdot \frac{(\widehat{L_{N_n}} - \widehat{L_{N_1}})}{\widehat{L_{N_1}}}, \tag{5.1}$$

1 = First year of the series

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

$\widehat{L_{N_1}}$ = Estimated normalized input in year 1, the first year of the time series

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

The same approach (as in 5.1) may be applied to individual segments of the time series separated by change points.

Examples

The trend analysis methods are illustrated below using the time series of normalized TN inputs to Kattegat. Figure 5.3 shows the normalized time series together with the fitted trend model. The model contains a single change point in 2012, represented by a continuous change. The linear trend fit before the change point and after, are statistically significant. Trend analysis should always begin with a visual inspection of the time series plot.

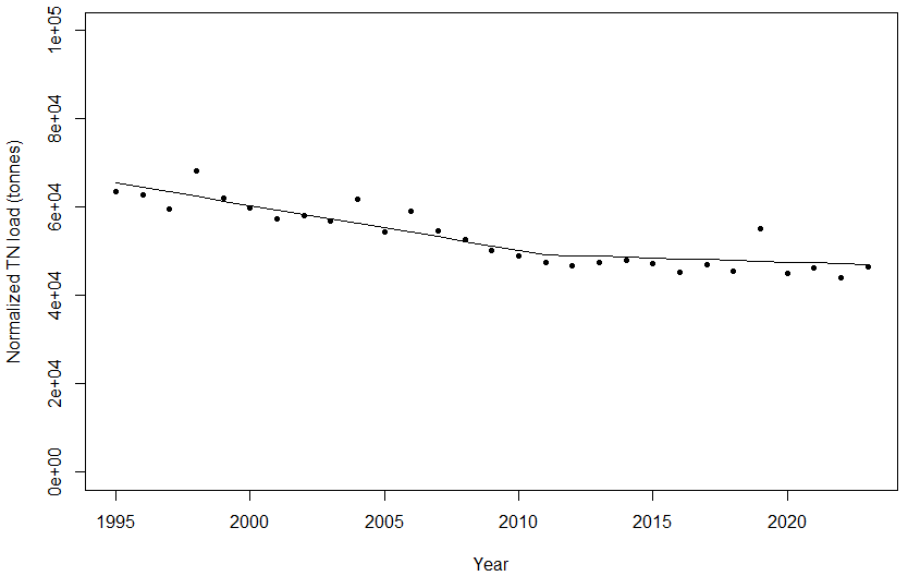
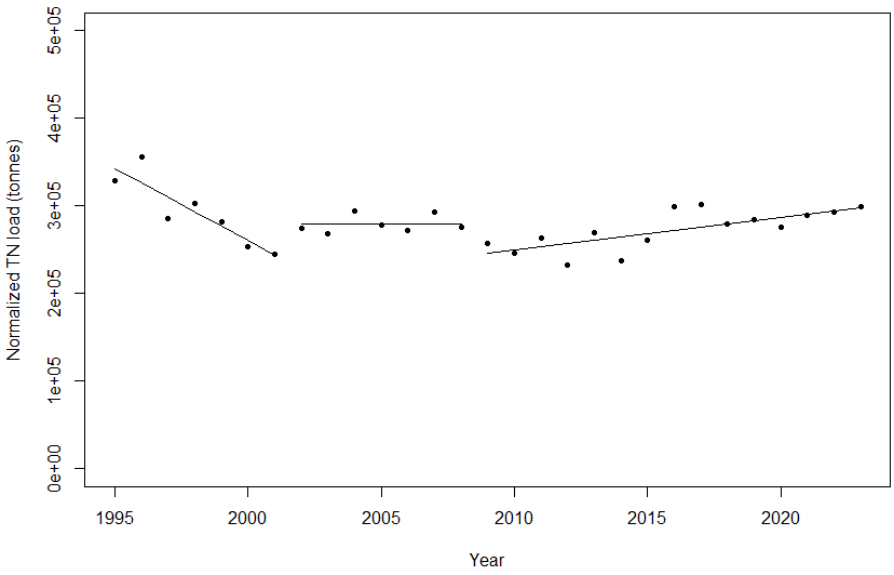


Figure 5.3. Fitted model with one change point for the normalized waterborne TN in-puts (tonnes) to the Kattegat 1995-2023.

The estimated change over the entire assessment period 1995-2023 for normalized TN inputs shown in figure 5.3 is -33.4%, calculated according to Equation 5.1.

Figure 5.4 presents a second example of the trend analysis method based on waterborne TN inputs to the Baltic Proper. Two change points were identified, occurring in 2002 and 2009. respectively. The intermediate segment is represented by a constant value, whereas the first and final segments exhibit statistically significant linear trends. The estimated total change over the assessment period 1995-2023 is -12.9%.

Figure 5.4. Fitted model with two change points for normalized waterborne TN inputs (tonnes) to the Baltic Proper in 1995-2023.



Summary

Trend analysis identifies normalized loads that exhibit statistically significant trend over time and thereby providing evidence of changes in anthropogenic nutrient inputs and pressures. The trend models support interpretation of long-term developments, estimation of total changes over time, and prediction of normalized nutrient loads in the final year of the assessment period.

6 Testing fulfilment of BSAP reduction targets

Progress in nutrient input reduction can be assessed using two complementary methods:

1) Trend analysis of normalized nutrient input time series, as described in chapter 5.

2) Statistical evaluation of whether country-specific nutrient input ceilings (NICs) and maximum allowable inputs per sub-basin (MAIs) under the HELCOM BSAP have been met by the Contracting Party.

This chapter proposes a statistical method for testing compliance with MAI and NIC and introduces a traffic-light system for illustrating progress towards achieving the targets. A statistical method is required to evaluate whether a normalized nutrient input time series has reached and/or moved closer to a predefined nutrient target. For this purpose, a parametric approach based on a simple hypothesis testing of the mean value in a sample of normally distributed data (Gaussian distributed data) is proposed. This approach is commonly referred to as the *fail-safe principle*.

In addition, a method is presented for testing whether the estimated normalized load for the final year of a time series differs significantly from the average value during a reference period. For nutrients input, the reference period is defined as 1997-2003.

Let us first assume that a normalized nutrient inputs time series is available. The time series is initially assumed to exhibit neither a statistically significant trend nor substantial serial correlation. Furthermore, the nutrient target *T,* whether defined as a nutrient input ceiling (NIC) or a maximum allowable input (MAI), is assumed to be known without uncertainty and therefore treated as a fixed value (certain amount of nitrogen/phosphorus given without any uncertainty). Finally, the observations are assumed to originate as a data sample from a normal (Gaussian) distribution with mean μ and variance σ^2.

The null hypothesis for the statistical test assumes target has not been achieved, i.e.:

$H_0: \mu \geq T,$

Indicating that the nutrient targets have not been met.

The alternative hypothesis is:

$$H_A: \mu < T$$

Indicating the nutrient target has been achieved.

Assuming a test significance level of 5% (α = 0.05), then calculate the following statistics:

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

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

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

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

n = number of observations in the time series

$t_{(n-1,0.05)}$ = 95% percentile of the *t*-distribution with *n*-1 degrees of freedom.

A test significance level of 5% implies a 5% probability of incorrectly rejecting the null hypothesis. The t-distribution is used because the true population standard deviation (σ) is unknown and the sample size (number of years) is relatively small.

The statistic defined in Equation (6.1) is referred to as the adjusted mean. If the adjusted mean is lower than the target value *T*, the reduction target is considered to have been met (achieved).

Trend based Target Assessment

When a nutrient input time series exhibits a statistically significant trend, an alternative statistical procedure is needed. Assuming that the trend is linear, a linear regression model with year as independent variable can be fitted to the data, yielding estimates of the intercept α and slope β. The model residuals are assumed to be normally (Gaussian) distributed.

The fitted linear regression model is then used to estimate the normalized nutrient input in the final year ($year_n$) in the time series:

$$\widehat{L_{N_n}} = \hat{\alpha} + \hat{\beta} \cdot year_n. \quad (6.2)$$

$\widehat{L_{N_n}}$ = Estimated normalized input in year *n* (final year of the time series)

$\hat{\alpha}$ = Estimated intercept

$\hat{\beta}$ = Estimated slope.

The standard error of the estimated values (of nutrient input) is calculated as:

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

MSE = Mean Squared Error of the regression model (defined in chapter 4, equation 4.5)

n = Number of years in the time series

$year_n$ = Final year of the time series (e.g. 2023)

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

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

The adjusted estimate is then calculated as:

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

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

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

The standard error SE of the prediction mathematically defined in Equation (6.3) is the standard error of the estimated mean response from an ordinary linear regression model (Snedecor and Cochran, 1989).

If the trend in the time series is non-linear, alternative models must be considered, and the corresponding standard-error calculations must be adapted accordingly. The choice of model should be based on the observed trend structure in the data, including the presence of change points as described in Chapter 5.

Trend-Based Models with Change Points

The following examples illustrate models with one change point, denoted by Y, where the final year of the time series is denoted by $year_n$. In general, we denote the method the "trend method".

Example 1 is a model with one change point, a linear regression model for the segment before and after the change point, and no level shift before and after the change point (a continuous change point). Example 2 is identical except for a level shift at the change point. Example 3 includes a constant level segment after the change point. The Equations for the three examples:

Example 1 (Continuous change point):

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

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

Example 2 (Step change):

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

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

Example 3 (Constant final segment):

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

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

The standard error *SE* of the estimated value input in the ($year_n$) is calculated from:

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

m = Number of years after the final change point Y ($\geq Y$)= $n - Y + 1$

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

MSE is calculated for the complete model fitted to the full-time series (all years included).

If the final segment is represented by a constant level, the standard error $SE_{\widehat{L_{N_n}}}$ for the final year ($year_n$) estimated input becomes:

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

Correction for calculating control value for $year_n$

$$\bar{x}_{AD} = \widehat{L_{N_n}} + k \cdot SE_{\widehat{L_{N_n}}}. \quad (6.5)$$

k = 95% percentile of 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 applied to classify whether a country or a Baltic Sea sub-basin has achieved its HELCOM BSAP nutrient reduction targets. The system categorizes the status into three levels:

- Achieved: The reduction target has been met.
- Close to achievement: The target is near fulfilment, but statistical validation is not possible due to uncertainty in the estimated inputs.
- Not achieved: The target has not been met.

The traffic-light system is originally described in Larsen and Svendsen (2013).

Statistically, the system distinguishes between:

Red:

$$\bar{x} \text{ or } \widehat{L_{N_n}} > T$$

The target value T is exceeded by either the estimated normalized input for the final year or – in the absence of a trend - the average normalized nutrient input over the considered period/segment.

Yellow:
If $\bar{x}$ or $\widehat{L_{N_n}} < T$ and $\bar{x}_{AD} > T$, then the null hypothesis of target test is accepted. This occurs when the estimated normalized input for the final year - or the average normalized input over the assessment period/segment in the absence of a trend - is lower than the target. With other words, the estimated load is

below the target, but the null hypothesis cannot be rejected because of the uncertainty estimation. Consequently, compliance with the target cannot be demonstrated with statistical confidence.

Green:
If $\bar{x}_{AD} < T$, the null hypothesis of the target test is rejected in favour of the alternative hypothesis, meaning the target has been achieved. Consequently, the estimated normalized input for the final year - or the average normalized input over the assessment period/segment in the absence of a trend - is lower than the target.

Testing Whether the Estimated Input in the Final Year Differs from the Reference Period

To evaluate whether the estimated normalized input in the final year of a time series $\widehat{L_{N_n}}$ differs significantly from the mean inputs during a reference period, the following procedure is applied.

The reference period is defined as the fixed period 1997-2003. The mean normalized nutrient input during the reference period is calculated as:

$$\overline{L_{N_{(ref)}}} = \frac{1}{7}\sum_{i=1997}^{2003} L_{N_i} \quad (6.6)$$

$\overline{L_{N(ref)}}$= average normalized nutrient input during the reference period

L_{N_i}= normalized nutrient input in year i

The corresponding 95% confidence interval for the reference period mean $\overline{L_{N_{(ref)}}}$ is calculated as:

$$\overline{L_{N_{(ref)}}} \pm k \cdot \mathrm{SE}_{\overline{L_{N_{(ref)}}}}. \quad (6.7)$$

$SE_{\overline{L_{N(ref)}}}$= standard error of the reference-period mean value

k = 97.5% percentile of the t-distribution with 6 degrees of freedom (k=2.447) (the reference period contains 7 observations (years) minus 1 degree of freedom).

The estimated normalized input in the final year of the time series $\widehat{L_{N_n}}$, may also be associated with a 95% confidence interval as well by calculating

$$\widehat{L_{N_n}} \pm k \cdot SE_{\widehat{L_{N_n}}}. \quad (6.8)$$

With

k = 97.5% percentile of the t-distribution with n-p degrees of freedom

p =Number of parameters in the final model

$SE_{\widehat{L_{N_n}}}$ = Standard error of estimated normalized inputs in the final year, calculated according to the selected model.

The procedure for calculating $SE_{\widehat{L_{N_n}}}$ is described in the previous sections of this chapter.

The hypothesis of no difference in input between the reference period and the estimated value in the final year value can be evaluated using:

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

Where:

k = 97.5% percentile of the t-distribution with n-p+6 degrees of freedom

|| denotes for absolute value

If the left-hand side of Equation (6.9) is greater than zero, the difference between the reference period and the estimated value for the final year is considered statistically significant at the 5% significance level.

Example: Danish Straits

To illustrate the method, normalized TN inputs to the Danish Straits were evaluated against the MAI input ceiling of 65,998 tonnes TN per year in 2023.

A trend model including two change points, identified in 2004 and 2014, was fitted to the normalized time series (Figure 6.1). Based on this model, the estimated normalized TN input in 2023 was 56,505 tonnes with a standard error $SE_{\widehat{L_{N_n}}}$ of 999 tonnes.

Applying Equation (6.5) gives an adjusted value (control value) of 58,214 tonnes. Since this value is below the MAI target of 65,998 tonnes, the assessment results in a green traffic-light classification, indicating that the target has been met with statistical confidence. This example is illustrated in figure 6.1.

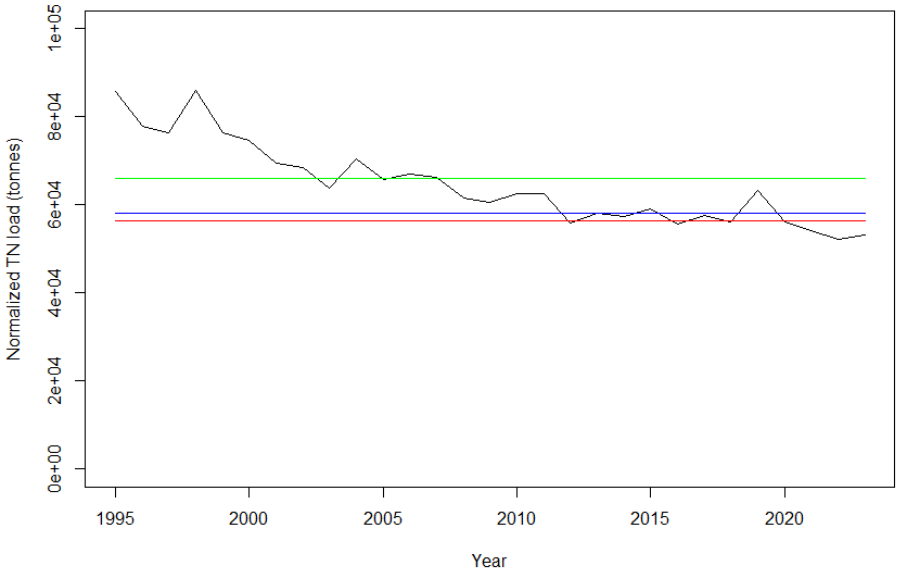


Figure 6.1. Illustration of the trend-based target assessment applied to the time series of annual TN inputs to the Danish Straits. The green line represents the MAI target. The red line represents the estimated TN input in 2023, and the blue line represents the adjusted value (control value) that accounts for estimation uncertainty on final year TN input according to Equation 6.5.

The average normalized TN inputs in the reference period (1997-2003) to the Danish Straits was 73,561 tonnes. The corresponding 95% confidence interval, calculating using Equation (6.7), was [66,938; 80,183] tonnes.

The estimated TN input in 2023 has a 95% confidence interval of: [54,444; 58,566] tonnes calculated using Equation (6.8). Applying Equation (6.9) yields a test statistic of 12,159 tonnes. Since this value exceeds zero, the difference is

statistically significant. It is therefore concluded that the normalized TN input to the Danish Straits in 2023 was statistical significantly lower than during the reference period 1997-2003. Overall, the normalized TN input in 2023 was estimated to be approximately 23% lower than the average input during the reference period.

Example: Evaluation of Finnish Phosphorus Input Ceilings

Table 6.1 provides another example of the statistical assessment method applied to evaluate compliance with Finish phosphorus input ceilings based on normalized TP inputs during 1995-2022.

The assessment demonstrates that the estimated TP inputs to Gulf of Finland in 2022 (605 tonnes P) are higher than the corresponding nutrient inputs ceiling of 315 tonnes P. When estimation uncertainty is considered, the adjusted assessment value (the test value) increases to 646 tonnes P. Consequently, the traffic-light classification for the Guld of Finland is red and the target has not been met. Based on the test value, a further reduction of 331 tonnes P would be required to achieve compliance with the TP input ceiling.

For the Bothnian Sea traffic-light classification changed from yellow in the 2017 assessment to green in 2022. This change occurred because the estimated TP inputs, including estimation uncertainty on the input, are below the TP input ceilings. Compliance with the target can therefore be demonstrated with statistical confidence.

The Bothnian Bay also received a green classification. After accounting for estimation uncertainty, phosphorus input remains below the TP input ceiling to an additional reduction margin of 159 tonnes P relative to the target level.

Table 6.1. Illustration of the traffic-light system. Evaluation of progress towards phosphorus inputs ceiling for Finland to Bothnian Bay (BOB), Bothnian Sea (BOS) and Gulf of Finland (GUF), based on normalized annual TP inputs during 1995-2022. Green: input ceiling is met with statistical confidence. Red: input ceilings are not met. Yellow: Compliance with the input ceiling cannot be demonstrated with statistical confidence because of uncertainty in the estimated inputs in 2022. TP in tonnes.

Finland TP	BOB	BOS	GUF
A: Input ceiling	1,683	1,246	315
B: Estimated input 2022	1,453	1,057	605
C: Inputs 2022 including uncertainty (test value)	1,524	1,137	646
Extra reduction (A-C)	159	109	
Remaining reduction to achieve MAI			331

Interpretation of the Traffic-Light Assessment

The traffic-light system provides a transparent framework for communicating progress towards BSAP nutrient reduction targets while explicitly accounting for uncertainty in the estimated nutrient inputs.

A **green** classification indicates that the adjusted assessment value remains below the target threshold after accounting for estimation uncertainty. In this situation, the target is considered to have been met with statistical confidence.

A **yellow** classification indicates that the estimated nutrient input is below the target value, but uncertainty is sufficiently large that compliance with the target cannot be demonstrated statistically. Additional monitoring data or further reductions may be required before a definitive conclusion can be reached.

A **red** classification indicates that the estimated nutrient input exceeds the target value and that further reductions are required to achieve compliance.

By incorporating uncertainty directly into the assessment, the traffic-light system provides a precautionary and statistically robust basis for evaluating progress towards BSAP nutrient reduction targets.

Summary

The statistical methods described in this chapter provide a framework for assessing compliance with HELCOM BSAP nutrient reduction targets while explicitly accounting for uncertainty in time series final year estimated nutrient inputs.

For time series without significant trends, compliance is evaluated using the adjusted mean derived from the fail-safe principle. For time series exhibiting significant trends, compliance is assessed using model-based estimates of the normalized nutrient inputs in the final year of the series together with their associated standard errors.

The estimated standard error is used to calculate an adjusted assessment value that accounts for uncertainty and supports objective evaluation of compliance with nutrient input ceilings (NICs) and maximum allowable inputs (MAIs): These adjusted values form the basis of the traffic-light classification system.

Furthermore, the estimated normalized load of the final year of the time series can be compared statistically with the reference period (1997-2003) to assess whether nutrient inputs have changed significantly over timer.

To reduce main part of the variability attributed to hydrological factors, annual nutrient inputs are flow normalized.

7 Step-by-step analysis illustrated by HELCOM data examples

This chapter provides a step-by-step illustration of the statistical methods described in the previous chapter using examples based on HELCOM PLC nutrient input data.

First, the application of the hydrological normalization method is demonstrated. The example is based on annual TN loads in River Ljungan, Sweden. The linear regression model (Equation 4.3, using logarithmic transformed values) provides a reasonable fit to the data, with an R^2value of 0.68. The estimated intercept (α) is 1.70 and the estimate slope (β) is 1.11. Furthermore,

$\overline{logQ}$=4.88

MSE=0.0368

$\bar{L} = 1.275$

$\overline{L_N} = 1.229.$

The normalized values were subsequently corrected for bias according to Equation (4.15). The results of normalization procedure are shown in figure 7.1. Notably, the hydrological normalization substantially reduces the year-to-year variability in the input timer series.

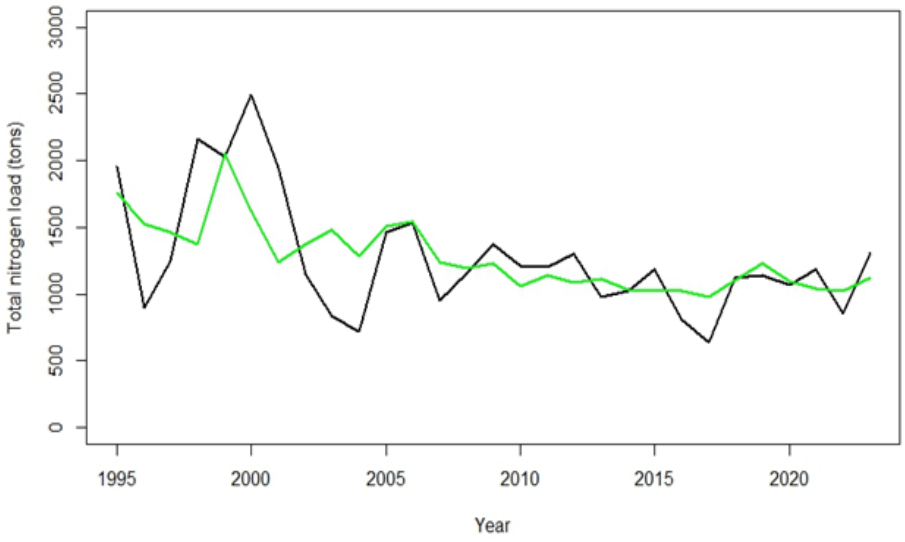


Figure 7.1. Normalized (green) and actual annual TN loads (tonnes) in river Ljungan.

The remainder of this chapter presents a complete statistical analysis, covering the entire process from normalization of annual nutrient inputs to assessment of compliance with nutrient reduction targets. The example is based on total nitrogen (TN) inputs to the Baltic Proper during 1995-2026.

It is assumed that the underlying data has been subjected to quality control procedures, including check for missing values, data gaps, outliers and errors, and have subsequently been accepted by all relevant HELCOM Contracting Parties.

Hydrological normalization is performed separately for each river with a full 1995-2026 time series discharging into the Baltic Proper.

Figure 7.2 shows the resulting normalized TN inputs aggregated all rivers discharging into the Baltic Proper, including the estimated normalized inputs from unmonitored areas and normalized airborne TN inputs, plotted together with the corresponding actual annual inputs. The figure shows the interannual variability is substantially reduced after normalization.

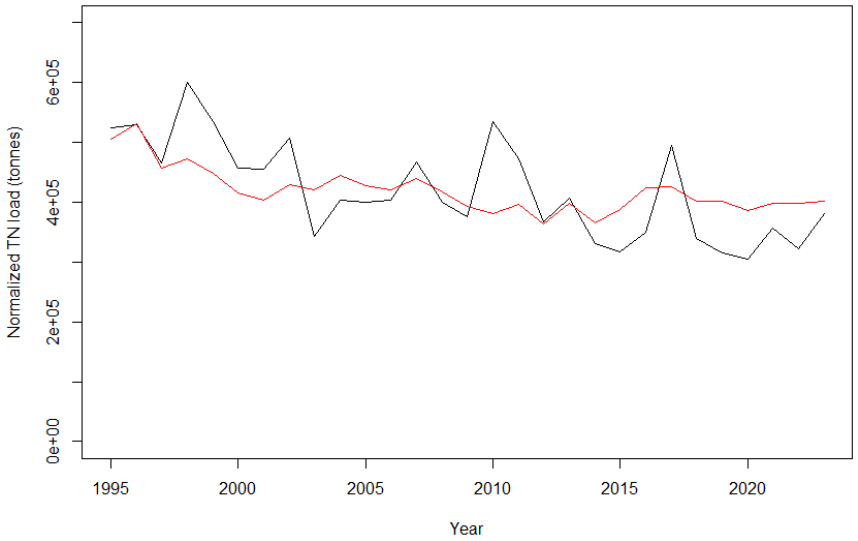


Figure 7.2. Normalized (red) and actual waterborne and airborne TN inputs (tonnes) to the Baltic Proper.

As described previously, normalization is performed for each river with complete 1995-2023 time series discharging into the Baltic Proper. The normalized riverine inputs are then aggregated and combined with estimated normalized inputs from unmonitored areas, actual inputs from direct point sources (point sources discharging directly into sea) and normalized atmospheric deposition on the Baltic Proper. The resulting total normalized inputs are subsequently used for trend analysis and assessment of compliance with BSAP reduction targets.

Figure 7.3 shows the fitted trend model for the normalized TN input time series. The selected model contains two change points, occurring in 2002 and the second in 2009, respectively. The two segments after the change points can be adequately represented by constant values, indication of no statistically significant trend within these periods (see figure 7.3). The first segment exhibits a statistically significant downward trend, estimated at approximately 20,315 tonnes TN per year.

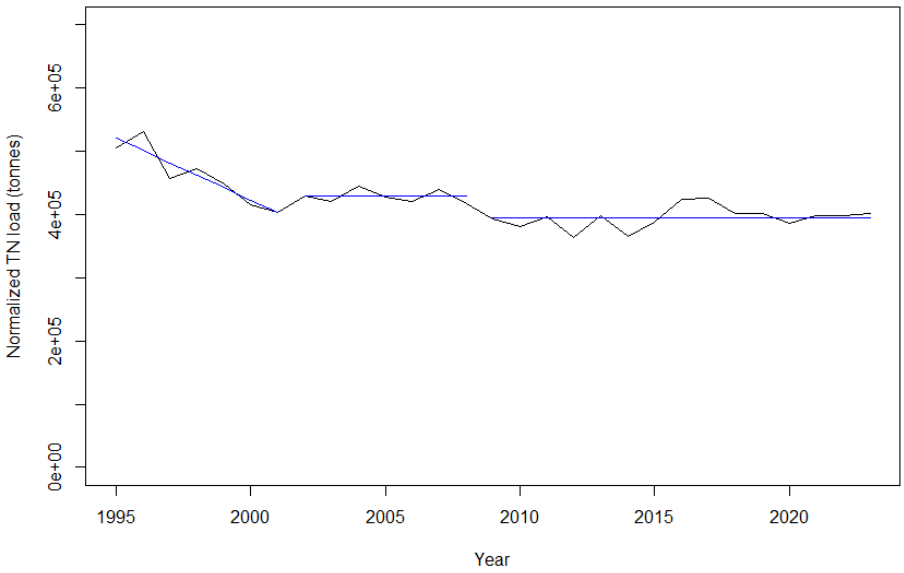


Figure 7.3. Linear trend model fitted to annual normalized water and airborne TN in-puts (tonnes) to the Baltic Proper during 1995-2023.

The maximum allowable total nitrogen input (MAI) to the Baltic Proper is 325,001 tonnes per year. The model-based estimated of the normalized TN input in 2023 is 394,934 tonnes. Applying Equation (6.5) yields and adjusted assessment value (test value) of 402,156 tonnes. Both values exceed the MAI target by a substantial margin. Consequently, the traffic-light classification for the Baltic Proper is red, indicated that the targets have not been achieved (Table 7.1 and figure 7.4).

Table 7.1. Assessment of compliance with the HELCOM BSAP maximum allowable input (MAI) target for total nitrogen (TN) inputs) to Baltic Proper (BAP) in 2023, based on normalized annual TN inputs during 1995-2023. Units in tonnes.

A: MAI	325,001
B: Estimated TN input in 2023	394,934
C: Standard error of estimated input (from formula 6.5)	4,228
D: Estimated TN input including uncertainty (B+C)	402,156
E: Exceedance of MAI (D-A)	77,156
Traffic light-classification	

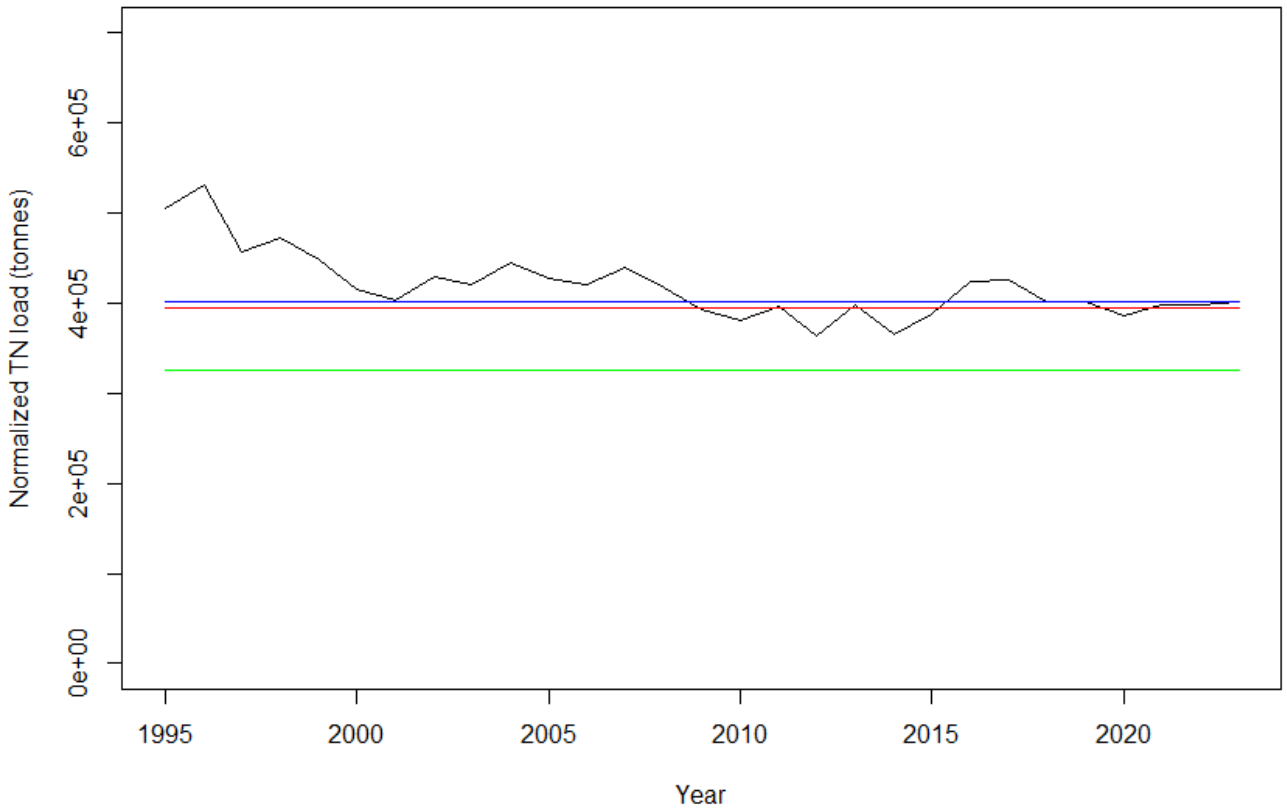


Figure 7.4. Assessment of compliance with the MAI target for total nitrogen inputs (water- and **airborne**) (in tonnes) to the Baltic Proper during 1995-2023. The green line represents the MAI target. The red line represents the estimated normalized TN input in 2023, while the blue line represents the adjusted assessment value (test value, according to Equation (6.5). including estimation uncertainty.

The estimated normalized inputs in 2023 (394,934 tonnes) have a 95% confidence interval of [386,226; 403,642 tonnes].

For the reference period 1997-2003, the mean normalized TN inputs were estimated at 435,342 tonnes with a corresponding 95% confidence interval of [412,654; 458,030 tonnes].

Application of Equation (6.9) yield a test statistic of 23,130 tonnes. Since this value is greater than zero, the difference is statistically significant at the 5% significance level. It can therefore be concluded that the estimated normalized TN inputs in 2023 was significantly lower than the average inputs during the reference period. The reduction is estimated to be approximately 9% relative to the mean input during 1997–2003.

The uncertainty associated with the detrended normalized TN input to the Baltic Proper is estimated at 32,779 tonnes, corresponding to approximately 8% of the estimated normalized input in 2023. This uncertainty estimate was calculated according to the procedures described in chapter 3.

8 Recommendations

This report addresses the statistical aspects of analyses conducted in relation to PLC data assessments and evaluations of compliance with HELCOM BSAP nutrient reduction targets (MAI and NIC). A range of topics are covered, including hydrological normalization, trend and change point analysis, and statistical significance tests of whether nutrient reduction targets have been met while accounting for uncertainty on nutrient input data.

The following section presents recommendations on which statistical methods are best suited for the PLC guideline chapter, "Data validation and assessment methodology for Maximum Allowable Inputs and Nutrient Input Ceilings".

Data quality and consistency

Good data quality and consistency are prerequisites for reliable statistical analyses of nutrient input time series. Time series may contain missing observations, data gaps and potentially erroneous or suspicious values. Chapter 2 describes methods for gap filling and procedures for determining whether an observation should be regarded as an outlier. In addition, methods for handling observations below analytical detection limit or limit of quantification through appropriated imputation techniques are presented.

Estimation of uncertainty

Total uncertainty in nutrient input data consists of both bias and precision. Estimating the exact total uncertainty associated with nutrient input data reported by the HELCOM Contracting Parties is challenging. The report therefore describes an indirect method for estimating the precision of annual nutrient input data. Although this estimate may slightly overestimate the true precision uncertainty, it provides a reasonable and practical approximation.

Information on the uncertainty associated with nutrient inputs from unmonitored areas should also be provided by the Contracting Parties. Such uncertainty estimates may be derived from model uncertainty analyses or expert assessment.

Hydrological normalization

Hydrological normalization of nutrient inputs should be carried out using the regression-based method employing transformed nutrient inputs (loads) and runoff, as described in Equations (4.3) and (4.4).

Because the normalization procedure involves back-transformation of the logarithmically transformed variables, the resulting normalized nutrient inputs should be corrected for retransformation bias according to Equation (4.5). Natural logarithmic transformation is recommended throughout the normalization procedure (Equation (4.3)).

Normalization should be performed separately for each sub-catchment or river. The resulting normalized nutrient inputs can subsequently be aggre-

gated at country level, country by basin, or at Baltic Sea sub-basin level together with the corresponding estimated inputs from unmonitored areas, direct point source discharges and atmospheric deposition on the sea.

Normalization is a necessary preprocessing step prior to trend analysis. The procedure substantially reduces variability in annual input due to hydrological conditions significantly and thereby improves the ability to detect statistically significant trends in nutrient inputs. Consequently, relatively small changes in nutrients may become detectable.

At present, the normalization method is not fully robust to long-term changes in the load and runoff relationship or systematic change in the runoff associated with climate change. Significant trends in the annual runoff have so far been observed only in a limited number of time series. In practice, changes in the load-runoff relationships may therefore represent a greater challenge than changes in runoff itself.

Should monthly nutrient input be adopted in future assessments, similar normalization methods can be applied (instance e.g. Silgram and Schoumans (ed., 2004).

Trend analysis

For initial trend assessment, the non-parametric Mann-Kendall trend test is recommended for identifying monotone trend in the normalized nutrient input time series. The method is robust and can be applied to "raw" nutrient data, normalized nutrient data, and runoff time series as well. As with all statistical trend test, autocorrelation may influence the power and significance of the test.

If monthly nutrient data are used in future assessments, the Seasonal Kendall trend test (Hirsch and Slack, 1984) should be considered.

For more detailed analysis of trends and trend magnitude, including estimation of slopes, intercepts, last-year inputs, ordinary linear regression (a parametric method) is recommended. Ordinary linear regression assumes a linear relationship between the response variable and time, normally (Gaussian) distributed residuals, and statistically (stochastic) independent observations.

Both Mann-Kendall test and ordinary linear regression should be modified if substantial serial correlated is present in the time series. Serial correlation can be assessed using the Durbin-Watson test statistics.

Change point analysis

If a time series exhibits one or more trend reversals (change points), the trend should be modelled using two or more linear trend segments.

Change point may be identified either through visual inspection of the time series plot or, preferably, using formal statistical methods such as proposed by Carstensen and Larsen (2006).

For time series of approximately 25-30 years of length, it is recommended that the number of change points be limited to a maximum of three or four. Furthermore, each segment before, between and after a change point should contain at least five years of observations.

Two alternative change points should be tested:

- Model 1: A continuous change point model in which the trend line remains connected (trend line connected) at the change point.
- Model 2: A discontinuous change point model allowing a step change in the level of the time series (disconnected trend lines at the change point).

Ordinary linear regression is recommended for estimating trend lines within each segment of a times series with change points.

Estimation of trend magnitude

Change in nutrient inputs may be estimated using the non-parametric Theil-Sen slope estimator. This method assumes a constant rate of change (i.e. a linear trend) and is therefore most appropriate when the trend can be represented by a single linear relationship.

Because change point analysis is recommended for PLC assessment and BSAP nutrient reduction targets evaluation, ordinary linear regression is generally preferred for estimating slope and intercepts within individual trend segments.

If the fitted trend model is non-linear, the overall change should be estimated from the modelled start and end values of the input or runoff time series.

Assessment of compliance with HELCOM BSAP targets

Evaluation of compliance with HELCOM BSAP nutrient reduction targets (MAI and NIC) requires statistical methods evaluating whether MAI and NIC have been achieved, that accounts for uncertainty in estimated nutrient inputs and allow quantification of remaining reduction requirements.

For time series without a significant trend, Equation (6.1) should be used to calculate the estimated final year nutrient input (the adjusted mean considering uncertainty) and evaluate compliance with the target value.

For time series with a significant linear trend, Equation (6.4) should be applied.

For time series containing one or more change points, Equation (6.5) should be used.

If no significant trend exists in the entire time series or in the final segment after a break point, use the average nutrient input of that period as the estimate for the final year.

To facilitate communication of assessment results on evaluating achieving nutrient reduction targets by Contracting Parties, Country-by-sub-basin or Baltic Sea sub-basin, a traffic-light system is recommended:

- Red = The estimated nutrient input for the final year exceeds the target value: The target has not been achieved.

- Yellow = The estimated nutrient input for the final year is below the target value, but the uncertainty-adjusted assessment value (final year estimated nutrient input including uncertainty on the estimate) exceeds the target: Compliance cannot be demonstrated with statistical confidence.
- Green: The uncertainty-adjusted assessment value (estimated final year nutrient input including uncertainty on the estimate) is below the target value: The target has been met with statistical confidence.

Comparison with the reference period

Finally; Equation (6.9) provides a statistical method for testing whether the estimated nutrient inputs in the final year of a time series differs significantly from the corresponding nutrients inputs during the HELCOM BSAP nutrient reduction scheme reference period (1997-2003). This procedure allows statistically significant changes in nutrient inputs since the reference period to be quantified and facilitate evaluation of long-term progress towards HELCOM BSAP nutrient reduction targets.

Together, the methods presented provide a transparent, consistent, and statistically robust approach for evaluating progress towards the nutrient reduction targets defined under the Baltic Sea Action Plan.

9 References

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

Carstensen, J. and 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.

Dean, R. B. and Dixon, W. J. (1951) Simplified statistics for small numbers of observations. Anal. Chem. 23, 636–638.

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

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

Grubbs, F. E. (1950) Sample criteria for testing outlying observations. Annals of Mathematical Statistics. 21 (1): 27–58.

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

HELCOM (2022) Guidelines for the annual and periodical compilation and reporting of waterborne pollution inputs to the Baltic Sea (PLC-Water). HELCOM,171 pp. https://helcom.fi/wp-content/uploads/2022/04/HELCOM-PLC-Water-Guidelines-2022.pdf

HELCOM (2024) Inputs of nutrients to the sub-basins. HELCOM core indicator report. 56 pp. Authors: Svendsen, L.M., Gustafsson, B. and Larsen, S.E. https://indicators.helcom.fi/wp-content/uploads/2023/04/HELCOM-Core-indicator-on-nutrients-1995-2022_PDF_October2025-2.pdf

Helsel, D.R. (2006) Fabricating data: How substituting values for nondetects can ruin results, and what can be done about it. Chemosphere, 65, p. 2434-2439.

Helsel, D.R. (2012) Statistics for censored environmental data using Minitab® and R. Wiley & Sons, New York.

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

Hirsch, R.M. and 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. and 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. and 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. and Søndergaard, M. (2014) Uncertainties in monitoring – how much do we tolerate. Presentation of NJF Workshop, 19-21 March 2014, Stavanger, Norway.

Larsen, S.E. and 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, and 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

Larsen, S.E, and 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

Lassen, P. and Larsen, M.M. (2025). Report on the HELCOM PLC-9 intercalibration on heavy metals and nutrients. Aarhus University, DCE – Danish Centre for Environment and Energy, 101 pp. Technical Report No. 362

Silgram, M. and 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. and 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., and Gustafsson, B. (2025) Progress towards national nutrient input ceilings https://helcom.fi/baltic-sea-action-plan/nutrient-reduction-scheme/national-nutrient-input-ceilings/

Van't Veen, S.G.M., Larsen, S.E., Andersen, P.M., Ovesen, N.B., Laugesen, J.R. and Kronvang, B. (2025) Uncertainty in nitrate load calculations evaluated using Monte Carlo simulations based on in-situ nitrate sensors in two Danish headwater streams. Journal of Hydrology, 662, 133816.

Wilk, M.B., and Gnanadesikan, R. (1968) Probability plotting methods for the analysis of data. Biometrika 55(1), 1–17.

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 n can be performed using the **Mann-Kendall trend test** (Hirsch et al., 1982), also known as **Kendall's tau test** (Kendall, 1975). The purpose of the test is to determine whether a statistically significant upward or downward trend exists over a period of n years, or whether the observed variation in the time series can be attributed to random fluctuations. The Mann-Kendall trend test has become a very effective and one of the most widely applied methods for trend analysis of water-quality data.

The Mann-Kendall trend test is a **non-parametric statistical method**, implying fewer assumptions than conventional parametric test methods. Unlike ordinary linear regression, the data is not required to follow a Gaussian (normal) distribution. However, observations should be free of substantial serial correlation. Furthermore, the test evaluates **monotonic trends**, and not necessarily linear trends, provided the overall direction of change remains consistently downward or upward throughout the period, and can therefore detect a broader range of trend patterns. The trend may be either increasing or decreasing without adhering to any specific functional form.

When data are normally distributed and the underlying trend is linear, ordinary linear regression generally has slightly greater statistical power because of its stronger assumptions. The R packages *trend* and *Kendall* both include implementations of the Mann-Kendall trend test, and Kendall's tau coefficient can also be calculated directly in SAS. Alternatively, the test can be implemented in any statistical software using the steps and equations outlined below.

Let x_i denote the annual observation in year i, where:

$$i = 1,2, \dots, n$$

The trend analysis null hypothesis (H_0) states that the n yearly observations are randomly ordered in time. The null hypothesis is tested against the alternative hypothesis (H_A) that the time series exhibits a monotonic trend.

The Mann-Kendall test statistic S is calculated as:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathrm{sgn}(x_j - x_i)$$

where the sign function *sgn* is defined as:

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

If either x_j or x_i is missing, then $sgn(x_j - x_i) = 0$ per definition, and the contribution to S is defined to zero.

The trend is tested by calculating the standardized test statistic Z:

$$Z = \begin{cases} \dfrac{S-1}{\sqrt{\mathrm{Var}(S)}}, & S > 0 \\ 0, & S = 0 \\ \dfrac{S+1}{\sqrt{\mathrm{Var}(S)}}, & S < 0 \end{cases}$$

Under the null hypothesis of no trend, the variance of S is given by:

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

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

A positive value of S indicates an upward trend, whereas a negative value indicates a downward trend. For a two-side test (when both a downward and an upward trend are of interest), the null hypothesis is rejected if the calculated value of Z falls below the -percentile or exceed the $\left(1 - {}^{\alpha}/_{2}\right)$-percentile in the standard normal (Gaussian) distribution with mean value 0 and variance 1. A one-sided test may also be applied when only increasing or decreasing trends are of interest.

The significance level is typically set to **5%**. The use of the standard normal distribution evaluating Z is justified because S is approximately normally distributed under the null hypothesis, with mean zero and variance $\mathrm{Var}(S)$, for sufficiently large sample sizes. Normal (Gaussian) approximation is generally considered good for $n \geq 10$ and acceptable for $5 \leq n \leq 10$.

Estimation of Trend Magnitude: The Theil-Sen Slope Estimator

An estimate of the annual trend magnitude (β) can be obtained if the trend is assumed to be approximately linear over the study period and the change is given as change per year. Hirsch et al. (1982) proposed the **Theil-Sen slope estimator**, which is calculated from all possible pairs of observations (x_i, x_j) with$1 \leq j < i \leq n$:

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

The Theil-Sen slope estimate, $\hat{\beta}$, is defined as the median of all the calculated d_{ij}-values. This estimator is robust and non-parametric and performs well for time series exhibiting non-normal distributions and moderate serial correlation. It is therefore widely used as a companion estimator to the Mann-Kendall trend test.

A $100(1-\alpha)$ % confidence interval for the slope β can be calculated following the procedure described by Gilbert (1987).

First select the desired significance level α (1%, 5% (typically) or 10%) and apply:

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

Calculate:

$$C_{\alpha} = Z_{1-\alpha/2}\sqrt{\mathrm{Var}(S)}$$

Then calculate:

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

and

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

where

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

is the total number of pairwise slope estimates.

The lower and upper confidence limits are obtained as theM_1th largest and the $(M_2 + 1)$th ranked value, respectively, among the ordered slope estimates d_{ij}.

Estimation of intercept

A non-parametric estimate of the intercept, α_0, may be calculated according to Conover (1980):

$$\alpha_0 = M_x - \beta M_i$$

where:

- $M_x$is the median of the observed values x_i,
- $M_i$is the median of the time indices i,
- $\beta$is the Theil-Sen slope estimate.

In practice, both slope and intercept may also be estimated using ordinary linear regression, which is the approach applied in the PLC assessment.

Seasonal Mann-Kendall test

If a time series consists of observations from multiple seasons (e.g. monthly nutrient loads), the **Seasonal Mann-Kendall Trend Test** (Hirsch & Slack, 1984) may be applied. The procedure involves calculating the Mann-Kendall statistic *S* for each season. The overall test statistics are then obtained by summing the seasonal test statistics across all seasons. The Seasonal Mann–Kendall test accounts for recurring seasonal patterns and is therefore particularly useful for analyses based on monthly or seasonal nutrient-input data.

For a detailed mathematical description of the Seasonal Mann-Kendall Trend Test, reference is made to Carstensen and Larsen (2006).

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

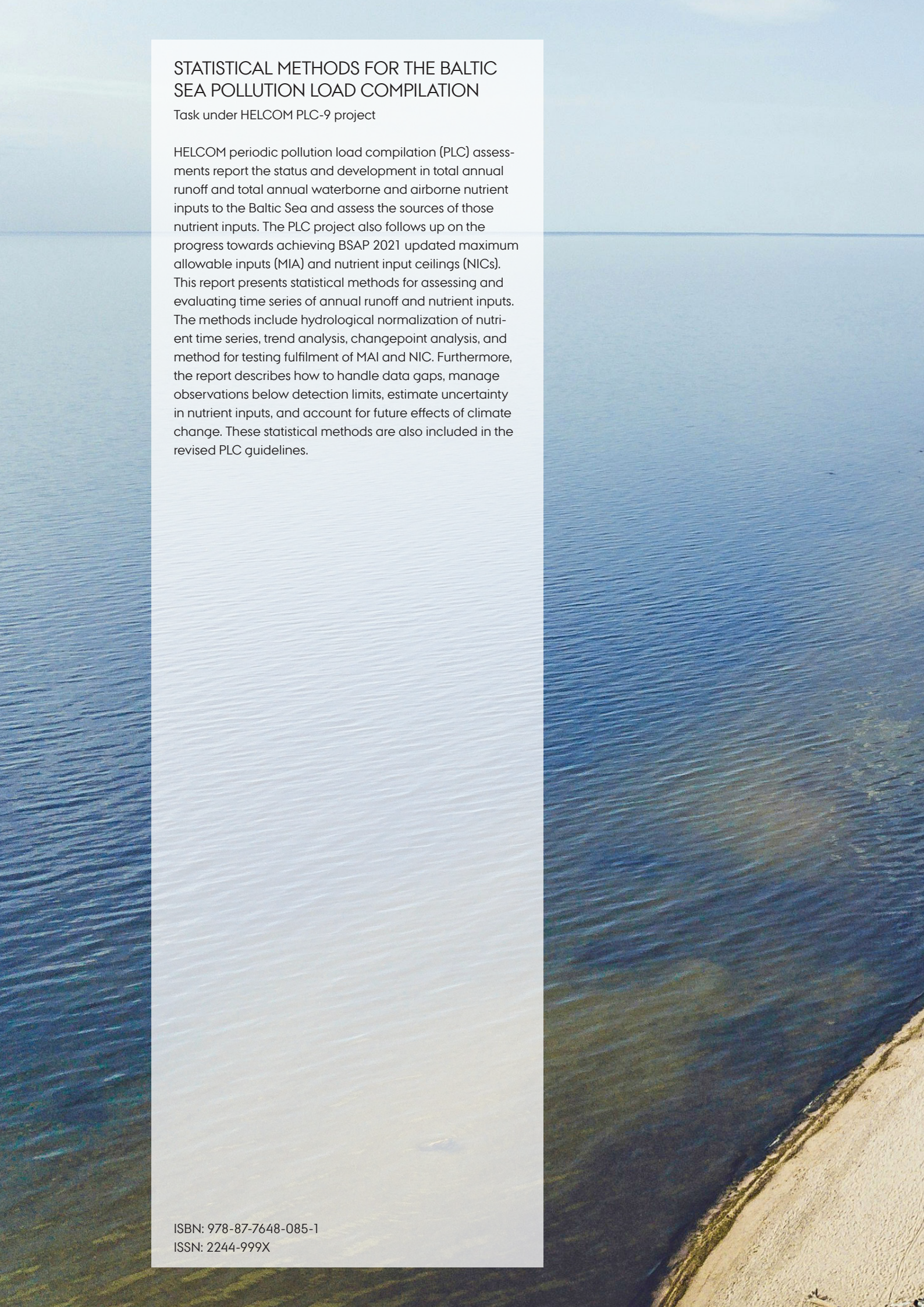
df	**95% percentile**
14	1.761
15	1.753
16	1.746
17	1.740
18	1.734
19	1.729
20	1.725
21	1.721
22	1.717
23	1.714
24	1.711
25	1.708
26	1.706
27	1.703
28	1.701
29	1.699
30	1.697
31	1.696
32	1.694
33	1.692
34	1.691
35	1.690

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
27	2.052
28	2.048
29	2.045
30	2.042
31	2.040
32	2.037
33	2.035
34	2.032
35	2.030

Annex 3: Examples of programs for SAS and R

Several SAS and R scripts implementing the statistical methods described in this report have been developed. These scripts can be used by the Contracting Parties to perform hydrological normalization and statistical analysis of their own nutrient input data.

The SAS and R scripts will be available on both the DCE homepage and the HELCOM website under the PLC Projects section.



STATISTICAL METHODS FOR THE BALTIC SEA POLLUTION LOAD COMPILATION

Task under HELCOM PLC-9 project

HELCOM periodic pollution load compilation (PLC) assessments report the status and development in total annual runoff and total annual waterborne and airborne nutrient inputs to the Baltic Sea and assess the sources of those nutrient inputs. The PLC project also follows up on the progress towards achieving BSAP 2021 updated maximum allowable inputs (MIA) and nutrient input ceilings (NICs). This report presents statistical methods for assessing and evaluating time series of annual runoff and nutrient inputs. The methods include hydrological normalization of nutrient time series, trend analysis, changepoint analysis, and method for testing fulfilment of MAI and NIC. Furthermore, the report describes how to handle data gaps, manage observations below detection limits, estimate uncertainty in nutrient inputs, and account for future effects of climate change. These statistical methods are also included in the revised PLC guidelines.

ISBN: 978-87-7648-085-1
ISSN: 2244-999X