

STATISTICAL ASPECTS IN RELATION TO BALTIC SEA POLLUTION LOAD COMPILATION

Task under HELCOM PLC-7 project

Technical Report from DCE - Danish Centre for Environment and Energy No. 137

2019



[Blank page]

STATISTICAL ASPECTS IN RELATION TO BALTIC SEA POLLUTION LOAD COMPILATION

Task under HELCOM PLC-7 project

Technical Report from DCE – Danish Centre for Environment and Energy No. 137

2019

Søren E. Larsen¹ Lars M. Svendsen²

¹ Aarhus University, Department of Bioscience ² 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. 137
Title: Subtitle:	Statistical aspects in relation to Baltic Sea Pollution Load Compilation Task under HELCOM PLC-7 project
Authors: Institutions:	Søren E. Larsen ¹ & Lars M. Svendsen ² Aarhus University, ¹ Department of Bioscience & ² DCE - Danish Centre for Environment and Energy
Publisher: URL:	Aarhus University, DCE – Danish Centre for Environment and Energy © http://dce.au.dk/en
Year of publication: Editing completed:	February 2019 February 2019
Referees: Quality assurance, DCE:	Henrik Tornbjerg, Department of Bioscience Jesper R. Fredshavn
Financial support:	Helsinki Commission (HELCOM) Baltic Marine Environment Protection Commission has kindly provided funding
Please cite as:	Larsen, S.E, & Svendsen, L.M. 2019. Statistical aspects in relation to Baltic Sea Pollution Load Compilation. Task under HELCOM PLC-7 project. Aarhus University, DCE – Danish Centre for Environment and Energy, 42 pp. Technical Report No. 137 <u>http://dce2.au.dk/pub/TR137.pdf</u>
	Reproduction permitted provided the source is explicitly acknowledged
Abstract:	HELCOM periodic pollution load compilation (PLC) assessments reports status and development in total annual runoff and total annual waterborne and airborne nutrient inputs to the Baltic Sea. This report deals with statistical methods for evaluating time series of annual runoff and nutrient inputs. Methods included are hydrological normalization of nutrient time series, trend analysis, change point analysis and a method for testing fulfilment of HELCOM Baltic Sea Action Plan (BSAP) nutrient reduction targets. Further is described how to fill in data gaps and to estimate the total uncertainty in nutrient inputs. These statistical methods are also included in the revised PLC guidelines.
Keywords:	Statistical methods, nutrients, runoff, hydrological normalization, trend estimations, uncertainty of nutrient inputs, data gaps, outliers, trend analysis, estimation of changes, evaluation off fulfilling BSAP reduction targets, HELCOM
Layout: Front page photo:	Graphic Group, AU Silkeborg Colourbox
ISBN: ISSN (electronic):	978-87-7156-389-4 2245-019X
Number of pages:	42
Internet version:	The report is available in electronic format (pdf) at <u>http://dce2.au.dk/pub/TR137.pdf</u>

Contents

1.	Introduction	5
2.	Data gaps and outliers	8
3.	Uncertainty of inputs (yearly input from a specific country or area)	11
4.	Hydrological normalization of nutrient inputs	17
5.	Trend analysis, change points and estimation of change	21
6.	Testing fulfilment of BSAP reduction targets	26
7.	Step by step analysis illustrated by HELCOM data examples	31
8.	Discussion and recommendations	34
9.	References	36
Anr	nex 1: Mathematical description of the Mann-Kendall trend test	38
Anr	nex 2: List of 95% percentiles and 97.5 percentiles of the t-distribution for the different possible combinations of degrees of freedom (df)	41

1. Introduction

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

In order to implement the following commitments:

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

HELCOM HOD 51-2016 approved the HELCOM project for the Seventh Baltic Sea Pollution Load Compilation (PLC-7) (Outcome HOD 51-2016 item 6.79)

The overall task of the PLC-7 project is to prepare a comprehensive assessment of the water- and airborne inputs of nutrients and selected hazardous substances and their sources to the Baltic Sea during the period 1995-2017 including follow up implementation of the HELCOM nutrient reduction scheme and assessment of the effectiveness of measures to reach the BSAP targets. The PLC-7 project is organized in working packages with the following tasks:

- 1. Establishing datasets and update of MAI (Maximum Allowable Inputs) and CART (Country Allocated Reduction Targets):
 - a) Monitoring and reporting of national annual/periodical data
 - b) Updating PLC-Water database and data on atmospheric inputs (PLC-Air)
 - c) Establishing the periodic assessment data set
- 2. Periodic assessment:
 - a) Assessment of sources of nutrients
 - b) Assessment of the effectiveness of measures
 - c) Assessment of inputs of selected hazardous substances
 - d) Compilation of the executive summary including policy messages
- 3. Methodologies
 - a) Updating guidelines and a statistical methodology report
 - b) Intercalibration on nutrient and heavy metals.

All these tasks and objectives call for a standardized and appropriate methodology, including statistical methods. Statistical methods are used e.g. when assessing trend in nutrient input time series including test for change points, to identify the extent of trends, and estimate uncertainty in nutrient inputs datasets. Further statistical methods are needed for the evaluation of progress fulfilling HELCOM BSAP nutrient reduction targets (MAI and CART) taking into account the uncertainty on nutrient inputs. The statistical methods are tools in the PLC assessments to allow the most qualified decisions to be made regarding possible trends and acceptance of fulfilling reduction requirements. The report "Statistical aspects in relation to Baltic Sea Pollution Load Compilation" (Larsen & Svendsen, 2013) was developed as a part of the former HEL-COM PLC-6 project. During the first follow-ups of progresses towards fulfilling BSAP nutrient reduction scheme and the PLC-6 assessment, the statistical methods has been further developed. The revised statistical methodology and new methods are included in this revised version of the 2013 report.

This report describes and includes a theoretical treatment of the statistical methods to be applied in the PLC assessments and BSAP nutrient reduction scheme. Focus points are assessment of waterborne input – and for the follow up of BSAP nutrient reduction scheme evaluating Contracting Parties fulfilment of MAI and CART. The described methods include flow normalization of nutrient inputs, filling in data gaps, testing for outliers, trends and change points, calculating changes in inputs in a time series, estimation of dataset uncertainty, and, finally, how to test whether reduction targets are fulfilled taking into account the uncertainty on inputs. Examples of quantified uncertainty on riverine inputs are also included. The statistical methods are included in the PLC guidelines. The statistical methods are also applicable for hazardous substances inputs.

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

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

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

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

In chapter 6, a method for testing the fulfilment of BSAP nutrient reduction targets (fulfilment and MAI and CART) is presented. The chapter also includes a definition of a traffic light system for inputs to determine for which Baltic Sea main basins Contracting Parties (or catchments) MAI or CART are 1) fulfilled, 2) not possible to judge if they are fulfilled due to statistical uncertainty or 3) do not fulfill nutrient input ceiling (or MAI) (reduction requirement). In chapter 7, we illustrate the proposed methods by a step-by-step analysis of real input data from the PLC water database to exemplify the practical use of some of the proposed methodologies.

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

The report includes Annex 1 with an in-depth mathematical treatment of the Mann-Kendall trend test and Annex 2 with selected percentiles of the t-distribution for different combinations of degrees of freedom.

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

The authors want to express their gratitude for the funding provided by Helsinki Commission and Aarhus University. Further, we want to thanks participants of the PLC-7 implementation group, HELCOM PRESSURE and HEL-COM RedCore DG for comments and inputs, and HELCOM Secretariat for help and support.

2. Data gaps and outliers

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

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

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

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

- The mean value of a statistical distribution. The distribution is determined either by including all relevant data on the given catchment or from a shorter time series, for instance when estimating missing data from point sources in the beginning or end of a time series.
- The mean of adjacent values. If *x*_a and *x*_c are perceived as two time series values with *x*_b missing, then:

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

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

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

- If runoff is known and a good relationship can be established between nutrient input and runoff, this can be used to estimate missing values.
- A q-q relationship can be used to estimate missing runoff values; a good qq relationship can often be established for a nearby river.(q is the discharge)

- A load-load relationship for another river for which high correlation can be verified.
- Model estimations of unmonitored catchment inputs, if possible otherwise, inputs can be estimated from reference data.
- Assignment of a real value in the interval between zero and the limit of detection (LOD)/limit of quantification (LOQ) to observations below a limit of detection/limit of quantification. The PLC guidelines (see chapter 1) describe how to handle concentrations under LOD/LOQ when calculating loads.

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

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

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

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

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

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

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

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

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

Time series of nutrient inputs demonstrate a certain amount of year-to-year variation due to the contributions from a large number of different components. One such component is a possible trend in inputs over time, and time series are therefore, by standard, detrended before analysis of variance components since trend-induced variations are not of basic interest in estimating the variance components.

In the case of a time series with a constant mean value, i.e. no trend present, the time series will either be detrended or tested to avoid a significant upward or downward trend. Variation appears within the yearly values – and it is thus assumed that the yearly inputs are sampled from the same population of inputs with a given mean value and a given variation. This variation is, in fact, an estimate of the total uncertainty of a given yearly input, i.e. the standard error of the mean.

Total uncertainty is a complex sum (based on certain assumptions) of a number of different uncertainty components:

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

Awareness exists in most countries of analysis (laboratory) uncertainty, at least regarding nutrients. This is relatively well documented but may be one of the components contributing the least to total uncertainty. Most other components are complex, and some of them are very difficult to estimate in practice due to unavailability of empirical data. Uncertainty can be diminished by optimizing, for instance, time and location of sampling and implementation of a monitoring program taking into account variations in concentrations and runoff. An optimized monitoring program may introduce more strategic monitoring and more precise and modern techniques as well as an optimized methodology for estimating inputs from unmonitored areas, strategic measuring being most important factor to decrease uncertainty. Knowing the size of the different uncertainty components is not necessary when - as will be discussed later – testing for trends and for compliance with set targets. Variance component analysis is used in statistics when the researcher seeks to optimize the sampling design in a hierarchical sampling regime and/or to test for effects (treatment, emission reducing measures or other factors) using the correct sums of squares.

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

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

For this purpose, we need a standardized methodology for estimating the uncertainties in the national datasets. One such methodology for estimating the uncertainty of data from monitored rivers has been described in a paper by Harmel et al. (2009). The method is called DUET-H/WQ (software is available at the HELCOM web pages), which is based on the so called RMSE (root mean square error) propagation method. It is a fair approximation to the true value, which is often very complicated to derive.

In DUET-H/WQ, the uncertainty of individual measured inputs is estimated by the formula:

$$EP = \sqrt{E_Q^2 + E_C^2 + E_{PS}^2 + E_A^2 + E_{DPM}^2},$$
(3.1)

where according to Harmel et al. (2009):

 E_0 =Uncertainty of the discharge measurement (±%)

 E_c =Uncertainty of sample collection (±%)

 E_{PS} =Uncertainty of sample preservation/storage (±%)

 E_A =Uncertainty of laboratory analysis (±%)

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

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

$$EP_{total} = \frac{100}{\sum_{i=1}^{n} x_i} \sqrt{\sum_{i=1}^{n} \left(x_i \cdot \frac{EP_i}{100} \right)^2}$$
(3.2)

where EP_{total} is given as ±%.

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

 x_i = Monthly input from a catchment or a country.

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

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

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

As mentioned in the beginning of this chapter, total uncertainty may also be estimated from the variance of a time series of inputs without trends or a detrended time series. It is the standard error of the mean input throughout the period. The two estimates of total uncertainty can then be compared. Both of the methods described here do not detect a systematical measurement bias, e.g. in runoff or in phosphorus inputs. Rather, they estimated the variation around an average value.

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

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

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

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

In Example 1 (table 3.1) *EP* is 11% and in Example 2 *EP* is 129% when using formula 3.1. Total uncertainty of assuming a constant monthly input of 2500 tons (x_i) is 3% for Example 1 and 36% for Example 2. Total uncertainties were calculated using formula 3.2.

A third example of calculating the uncertainty is from using Danish data for total nitrogen (TN) inputs to the marine areas around Denmark. The total input to the Danish marine environment is a sum of two components. One component is from the monitored catchment area and the other is from the unmonitored area. The inputs from the unmeasured area is estimated by using a model.

Example: Uncertainty on total nitrogen inputs for Danish monitored areas:

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

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

$$X = \sum_{i=1}^{n} X_i. \tag{3.3}$$

The variance of *X* can be calculated as:

$$\sigma_X^2 = Var(X) = \sum_{i=1}^n \sigma_{X_i}^2. \tag{3.4}$$

The standard deviation is then calculated as:

$$\sigma_X = \sqrt{\sum_{i=1}^n \sigma_{X_i}^2}.$$
(3.5)

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

$$100 \cdot \frac{\sigma_X}{x} = \frac{100}{\sum_{i=1}^n X_i} \sqrt{\sum_{i=1}^n \sigma_{X_i}^2}.$$
 (3.6)

The calculation of the total inputs from the monitored areas constitute of measurements from 169 stations in streams. These stations cover approximately 55% of the total Danish catchment area. Bias and precision can then be calculated as

$$bias (\%) = \frac{100}{\sum_{i=1}^{169} X_i} \sum_{i=1}^{169} bias_i \cdot X_i,$$
(3.7)

precision (%) =
$$\frac{100}{\sum_{i=1}^{169} X_i} \sqrt{\sum_{i=1}^{169} (precision_i \cdot X_i)^2}$$
. (3.8)

Here $bias_i$ and $precision_i$ are the individual biases and precisions (given in decimal notation) for each river indexed by *i*. The total uncertainty can then be calculated as

uncertainty (%) =
$$\frac{100}{\sum_{i=1}^{169} X_i} \sqrt{\sum_{i=1}^{169} (bias_i \cdot X_i)^2 + (precision_i \cdot X_i)^2}$$
. (3.8)

A Monte Carlo study (Kronvang & Bruhn, 1996) based on daily samples has shown that for Danish streams categorized by their catchment area, the following values for bias and precision are valid for TN load:

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

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

Using the formulae, it can be calculated that the total bias is -1% to -3%, the total precision is 0.7% to 1.2% and the total uncertainty is 0.7% to 1.3%. For an average stream station the bias is -1% to -3%, the precision is 3% to 5% and the uncertainty is 3.2% to 5.8%.

Example: Uncertainty on total nitrogen inputs for Danish unmonitored areas The TN input from the unmonitored areas is based on model estimates for 1286 very small catchments covering the rest of the Danish area (45%). The year load from each small catchment is calculated using the formula

$$L = Ndiffuse_{model} + R_{lake} + R_{stream} + N_{waste} - R_{total}, \qquad (3.10)$$

*Ndiffuse*_{model} = the estimated nitrogen inputs from the model

 R_{lake} = Estimated nitrogen retention in lakes

R_{stream} = Estimates nitrogen retention in streams

N_{waste} = Nitrogen inputs from waste water

 R_{total} = Total nitrogen retention.

In table 3.2 are shown bias and precision for the components in formula (3.10) based on both numerical calculations, the study by Kronvang & Bruhn (1996) and estimates.

Table 3.2. Bias and precision for nitrogen inputs in formula (3.10) based on both numerical calculations, estimates and Kronvang and Bruhn (1996).

Components	Bias (%)	Precision (%)
Model	-15 to 25	12 to 15
Retention lake	-5 to 5	40
Retention stream	-5 to 10	40
Retention total	-5	40
Point source: industry	-1 to -3	1 to 10
Point source: waste water	-1 to -3	1 to 10
Point source: fishfarms	-1 to -3	1 to 20
Point source: rain water	-5	40

Using the formulae (3.3) to (3.9) and the bias and precision indicated in table 3.2 the total bias for the unmonitored area is calculated to 20% to 28%, the total precision is 0.8% to 2.0% and the total uncertainty is 1.2% to 2.2%. For an average small unmonitored catchment the bias is 27%, precision 15% to 20% and the uncertainty 31% to 34%.

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

Example: Uncertainty on total phosphorus inputs from Denmark

With respect to total phosphorus (TP), calculations show that for the measured area the bias is -6 to -3%, the precision is 1 - 2% and the uncertainty is then 1 - 2.5%. For the unmeasured area the bias is between -5 and 30%, the precision is 1 - 3% and the uncertainty is 1 - 4%.

Hydrological normalization of nutrient inputs

The annual riverine inputs of nutrients show large variations between the reported years. Variation in runoff is a major reason behind this and is mainly caused by climate effects on hydrological factors such as precipitation, including accumulation and melting of snow/ice, and evapotranspiration, but also by temperature, etc. To remove the main part of the variation introduced by hydrological factors, the annual nutrient inputs are flow-normalized. Care should be taken when normalizing data if point sources have a large impact on calculated inputs, especially during periods with low water flow. Normalization should therefore not be applied on inputs from point sources discharging directly to the sea.

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

Different methods for normalizing inputs are described in Silgram and Schoumans (ed., 2004), chapter 4. In this report, we focus on methods based on empirical data. The empirical hydrological normalization method is based on the regression of annual loads and annual runoff; thus, the method normalizes the loads to an average runoff (averaged over the time series period). In this way, the variation attributable to the annual amount of runoff is removed, whereas the effect of differences in the distribution of runoff over the year is not removed. In Silgram and Schoumans (ed., 2004), the normalization is based on un-transformed loads and runoffs. In our experience, the regression explains slightly more of the variation if both annual input and annual runoff values are transformed by the natural logarithmic function before normalizing.

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

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

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

a and β = Parameters associated with linear regression

 ϵ_i = Residual error in the linear regression.

Then, the normalized load is calculated as:

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

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

Q_i = Runoff in year i

^ = Indicates that it is an estimated parameter.

To avoid possible negative loads, the below formula should be used:

$$L_{iN} = L_i \cdot \frac{\hat{\alpha} + \hat{\beta} \cdot \bar{Q}}{\hat{\alpha} + \hat{\beta} \cdot Q_i}.$$
(4.3)

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

$$\log L_i = \alpha + \beta \cdot \log Q_i + \varepsilon_i. \tag{4.4}$$

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

Formula 4.4 gives the following formula for normalized loads:

$$L_{iN} = \exp\left(\log L_i - \left(\log Q_i - \overline{\log Q}\right) \cdot \hat{\beta}\right) \cdot \exp(0.5 \cdot \text{MSE}), \quad (4.5)$$

In the above formula (4.4) and (4.5), "log" is the natural logarithmic function, "exp" is the exponential function, and MSE stands for Mean Squared Error and is derived by the regression analysis (Snedecor and Cochran, 1989). MSE is calculated in all standard statistical software programs and is in general defined as:

$$MSE = \frac{1}{n-2} \sum_{i=1}^{n} (x_i - \hat{x}_i)^2, \qquad (4.6)$$

n = Number of observations in the time series

$x_{i=}$ Observed value

 \hat{x}_i = Modeled value from linear regression.

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

The factor "exp($0.5 \cdot MSE$)" in the formulae is a bias correction factor and is derived as described by Ferguson (1986). The factor is needed in order to back-transform to a mean value and not to a geometric mean, whose calculation does not require this factor. If exp($0.5 \cdot MSE$) > 1.25, this indicate that the fit in formula (4.4) is not very good and it is probably better to use formula (4.1) and (4.2). The main reason for using the natural logarithmic function for transformation is stabilization of the variance among residuals. Without the transformation, residuals are often distributed with a heavy tail to the right.

In the PLC-5 assessment, the following method was used:

$$\log_{10}L_{iN} = \log_{10}L_i \cdot \frac{\widehat{\alpha} + \widehat{\beta} \cdot \log_{10}\overline{Q}}{\widehat{\alpha} + \widehat{\beta} \cdot \log_{10}Q_i},\tag{4.7}$$

after which the power function was used to back-transform formula (4.7). This method gives normalized loads, which are a bit too low. Use of the natural logarithmic function has a more solid foundation in statistics than the base 10 logarithmic function.

In principle, the presented methods can be applied even with a significant trend in the runoff time series, as long as the relationship between runoff and load is unchanged. Usually, the relationship changes with a significant change in the runoff over time. This implies that a trend analysis of the runoff time series is needed in order to determine whether an upward or downward trend in the flow is present. *It is our recommendation to divide the time series in two or more parts of equal length, and do the normalization for the each part separately. Each part of the time series should at least have a length of 8-10 years. Impact of this procedure need to be investigated further by data examples.*

In general, the differences between the methods, mentioned in formulas 4.1 to 4.6, are small, but especially for time series with a large year-to-year variation, methods without a correction term will give biased values with an underestimation of the normalized loads. This can have an unwanted effect when testing fulfillment of targets.

Hydrological normalization should be carried out catchment-wise, i.e. nutrient loads should be normalized for each catchment separately. If the normalization is performed country-wise or sub-basin-wise, the result will not be the same as the catchment-wise normalized nutrient loads summed to country or sub-basin level.

To illustrate the method, we used data from the Vistula River, Poland, to normalize both the load of TN and TP. Figure 4.1 shows scatter plots and the linear relation between loads and flow. Figure 4.2 shows the normalized time series together with the unnormalized loads. Note the large reduction in between-year variation in the normalized time series.

Figure 4.3 demonstrates the importance of normalizing TN and TP inputs. The example compares actual (not normalized) annual inputs of water + airborne inputs of nitrogen and phosphorus to the Bothnian Sea with the corresponding normalized TN and TP inputs.



Figure 4.1. Scatter plots of annual inputs of TN (a) and TP (b) against runoff and the linear regressions (transformation based on natural logarithmic function). Data represent the inputs of TN and TP to the Baltic Sea from the Vistula River in Poland during 1994-2010.

Figure 4.2. Time series plot of annual actual (not normalized) time series (dark green) and of normalized time series (green) of annual TN (a) and TP (b) in tonnes of Vistula River, Poland 1995-2016.



Figure 4.3. Actual (green) and normalized (black) annual total (water and airborne) nitrogen (a) and phosphorus (b) inputs to Bothnian Bay 1995-2016. Airborne Inputs are climate normalize as described in PLC guidelines annex 7 (HELCOM, in prep.).



5. Trend analysis, change points and estimation of change

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

The Mann-Kendall method (Hirsch et al., 1982) is a well-established method for testing for a monotone trend in a time series. It is non-parametric and based on Kendall's tau, which is a measure of the correlation between two different variables. The method is robust towards outliers and a few missing data. If the trend is linear, Mann-Kendall's method has slightly less power than ordinary regression analysis. The Annex gives a detailed mathematical description of the method, and software can be downloaded free at http://en.ilmatieteenlaitos.fi/makesens.

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

Ordinary regression analysis is also a well-known statistical method (figure 5.1), but demands a linear relationship with Gaussian distributed residuals, which are stochastic independent as well (Snedecor and Cochran, 1989). If the time series is serially correlated, both the Mann-Kendall test and ordinary regression must be modified, since the tests will be impacted by this, and the probabilities of statistical test values can therefore not be trusted. On the other hand, it appears that the autocorrelation for annual time series of either loads or runoff is small and can be ignored; thus, the methods can be used without modifications as a good approximation. The minimum time series length for application of the Mann-Kendall test is 8 years.

Both Mann-Kendall's trend analysis and ordinary linear regression allow performance of a one-sided trend test if focus is on testing for a downward or increasing development in a time series. This is of relevance in the PLC assessments and when evaluating progress towards HELCOM BSAP reduction targets. A time series plot can show one or two clear trend reversals (also called change points in time), e.g. when the first part of the time series shows a linear increase and the second part shows a linear decrease in nutrient inputs. The trend analysis can then be carried out by using a model with two or three linear curves or by applying two or three Mann-Kendall trend tests if time series sections include a sufficient number of years (example in figure 5.2).

Year of trend reversal (the change point) can either be determined by inspecting the time series plot or by applying a statistical method (Carstensen and Larsen, 2006). If an exact year of change in the inputs is known (e.g. reduced inputs due to implementation of new municipal wastewater treatment plants or new treatment methods, etc.), this year should be applied as change point, and the time series should be divided accordingly. Statistical estimation of the time when a change occurs in a time series is complex and involves a calculation procedure with iterative estimations. The LOESS (locally weighted scatterplot smoothing) regression method can be used as a supplement for detecting non-linear trends and for helping detecting change points/step trends as shown in figure 5.1b and 5.2b.



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

Figure 5.1. a. Total annual normalized air and waterborne TN inputs (tonnes) to the Baltic Sea. Trend line estimated with linear regression model.

b. As a but the trend line is estimated with LOESS (locally weighted scatterplot smoothing) regression method. **Figure 5.2.** a. Annual normalized TP inputs (tonnes) 1995-2015 to Gulf of Riga. One change point in the time series are detected, and the trend line is based on Mann-Kendal trend test and linear regression.

b. As a but the trend line is estimated with LOESS regression method.



Model 1:

$$L_{Ni} = \begin{cases} \alpha + \beta \cdot i, & \text{for } i < Y \\ \alpha + \beta \cdot i + d \cdot (i - Y), & \text{for } i \ge Y \end{cases}$$
Model 2:

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

 $L_N = Normalized input$

 α = Intercept

 β and d = Slopes

Y = A given year

i = Different years in the time series.

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

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

Change-points models are an aid for estimating the last year value, and to get an idea of the overall trend during the full time series period.

Finally, significance of the slope in the last segment is tested, and if not significant different from zero then we use the model:

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

c = *Estimated input (a constant)*

Summarizing the procedure: Analyzing for one change-point: No: Is there an overall significant trend? Yes: Fit a linear model No: Estimate a constant throughout time series Yes: Analyzing for a second change-point Yes: Testing the significance of the last segment Yes: Fit model No: A constant in the last segment Yes: Fit model No: Testing the significance of the last segment Yes: Fit model No: A constant in the last segment.

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

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

If we seek to identify the total change in nutrient inputs over the whole time series expressed as a percentage, we can use the two methods below. Estimated linear slope:

$$100 \cdot \frac{(n-1)\cdot\hat{\beta}}{\hat{\alpha}},\tag{5.1}$$

n = Length (number of years) of the series

 $\hat{\alpha}$ = Estimated intercept (input at start year minus one year)

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

In the case of change-points use the formula for each segment and add up the percentages. Remember to add the step trend if such one is detected at the change point. If the slope is not significant in one or more of the segment use a slope estimate equal zero in the formula.

For the evaluation of BSAP reduction targets and PLC assessments intercept and slope in formula (5.1) is based on linear regression estimates. Alternatively slope can be determined with Theil Sen slope estimator and intercept using the estimator suggested by Conover (1980). When using start and end values we have the formula:

(5.2)

In the case of change-points use the formula for each segment and add up the percentages. Remember to add the step trend if such one is detected at the change point. If the slope is not significant in one or more of the segment use the average value in the segment for starting and ending values.

For some times series, the start value, the end value or both can deviate too much from the general trend; if so, an approach using the average value of, for instance, the first 3 years and the last 3 years would reduce the influence of single years.

The trend analysis methods are illustrated below based on the time series of normalized TN inputs to Danish Straits. In figure 5.3, the normalized time series are shown together with a model fit of the trends. The model fit consists of one change point in year 2003 and the linear fits before and after the change point are significant. A trend analysis should always be initiated with a time series plot of the data series. The Mann-Kendall trend test is also significant with Z=-5.24 (P<0.0001).



The estimated change over the whole period for the normalized TN inputs is -32% according to formula 5.1 and -36% according to formula (5.2). Using the model with one change point the estimated change over the whole period is - 33%.



Testing fulfilment of BSAP reduction targets

The progress in nutrient input reduction can be tested by two different methods: 1) trend analysis of time series of normalized nutrient inputs, as discussed in chapter 5; and 2) statistical analysis of whether the country-wise nutrient reduction targets under BSAP have been significantly met by a Contracting Party. In this chapter, a statistical method for testing fulfillment of reduction targets is proposed, and a traffic light system is introduced to illustrate a country's progress towards fulfilling the targets. A statistical method for testing if a normalized nutrient time series has moved relative to a defined nutrient target is needed. For this purpose, a parametric method based on the simple test of the mean value in a sample of Gaussian distributed data is suggested – a method that is often referred to as the fail-safe principle.

Let us assume that we have a time series of normalized inputs. The time series is initially assumed to be without a statistical significant trend and without a significantly large serial correlation, and we assume that the reduction target *T* (or any kind of target such as, for instance, nutrient input ceilings for a country) is defined without error, i.e. is a fixed value (certain amount of nitrogen/phosphorus given without any uncertainty). Let us finally assume that the data is sampled from a Gaussian distribution with mean value μ and variance σ^2 .

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

$$H_0: \mu \ge T$$
 ,

H = Hypothesis

T = Target

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

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

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

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

SE = Standard error (SE = standard deviation divided by square root of *n*)

n = number of observations in the time series

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

A test probability of 5% means that we have a 5% probability of incorrectly rejecting the null hypothesis.

This statistic is called the adjusted mean, and if the statistic is less than the target *T*, the reduction target is fulfilled.

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

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

 $\widehat{L_{nN}}$ = Estimated normalized input in year n

 $\hat{\alpha}$ = Estimated intercept

 $\hat{\beta}$ = Estimated slope

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

$$SE = \sqrt{MSE} \cdot \sqrt{1/n + year_n^2 / \sum_{i=1}^n year_i^2}$$
(6.3)

MSE = Mean Squared Error as defined in chapter 4

n = Number of years in the time series

*year*ⁿ = Last year in the time series (i.e. 2014)

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

As reference year for the time series of years is used *year=0*. Then the statistic is calculated as:

$$\bar{x}_{AD} = \widehat{L_{nN}} + t_{n-2,0.05} \cdot \text{SE},$$
 (6.4)

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

A list with the 95% percentiles for different values of *n*-2 is given in annex 2.

The mathematical definition of the standard error of the prediction *SE* given in (6.3) is a well-known statistic from ordinary linear regression (Snedecor and Cochran, 1989). If the trend is not linear, other models have to be used for the time series, and the formula for the standard error needs to be revised. The form of the trend in the data will dictate the methods to be applied. These methods are based on the assumption of the existence of one or two change points in the time series (see chapter 5).

"Trend method"

A few examples are given here, the examples are based on models with one change-point *Y*, and we assume that the last year in the time series is denoted by *year*_n. In general, we denote the method the "trend method". The first example is a model with one change point and a linear model before and a linear

model after and no change in level before and after the change point. The second example is equal to the first example but with a change in level at the change point. The last example (example 3) is with a constant level after the change point.

Example 1:

$$L_{Ni} = \begin{cases} \alpha + \beta \cdot i, & \text{for } i < Y \\ \alpha + \beta \cdot i + d \cdot (i - Y), & \text{for } i \ge Y \end{cases}$$
$$\widehat{L_{nN}} = \hat{\alpha} + \hat{\beta} \cdot year_n + \hat{d} \cdot (year_n - Y)$$

Example 2:

$$L_{Ni} = \begin{cases} \alpha_1 + \beta_1 \cdot i, & \text{for } i < Y \\ \alpha_2 + \beta_2 \cdot i, & \text{for } i \ge Y \end{cases}$$
$$\widehat{L_{nN}} = \widehat{\alpha_2} + \widehat{\beta_2} \cdot year_n.$$

Example 3:

$$L_{Ni} = \begin{cases} \alpha + \beta \cdot i, \text{ for } i < Y \\ c, & \text{ for } i \ge Y \end{cases}$$
$$\widehat{L_{nN}} = \hat{c}$$

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

$$SE = \sqrt{MSE} \cdot \sqrt{\frac{1}{m} + \frac{year_n^2}{\sum_{i=Y}^{year_n} i^2}}$$

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

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

Correction for calculating control value for year_n

$$\bar{x}_{AD} = L_{\bar{n}N} + k \cdot \text{SE.}$$
(6.5)

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

p Number of parameters in the final model.

Traffic light system

Finally, a traffic light system can be defined to obtain a status of whether a country or a sub-basin has met the set BSAP target, whether it is close to fulfilling the target, or whether the target has not been fulfilled. This is described in HELCOM LOAD 4/2012 doc 5/2.2. Statistically, we define the system:

Red:

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

Yellow:

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

Green:

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

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

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

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

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

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

$$\overline{L_N(ref)} \pm k \cdot SE_{ref}.$$
(6.7)

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

For the estimate of the last year $\widehat{L_{nN}}$ 1 we can calculate the 95% confidence interval as well by calculating

$$\widehat{L_{nN}} \pm k \cdot SE \tag{6.8}$$

where the k factor is the 97.5% percentile in a t-distribution with n-p degrees of freedom. The number p is the number of parameters in the final model.

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

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

where *k* is the 97.5% percentile in a t-distribution with n-p+6 degrees of freedom.

To illustrate the principles, we tested if the normalized TN inputs to the Danish Straits met the provisional MAI input ceiling of 65,998 tonnes TN per year. Using the model with one change point in 2003 (se figure 5.3) the estimating input in year 2016 is 55,442 tonnes with a SE of 1,082 tonnes. According to formula (6.5) the control value becomes 57313 tons, which is less than 65,998 tonnes, so in the example the traffic light evaluation results in a green light. This example is illustrated in figure 6.1.

Figure 6.1. Principles on time series with trend created annual TN input to the Danish Straits. Full green line is the target (MAI), "------" line is the estimated value (TN input) in 2016, and "...." line is the test value (TN input taking into account uncertainty) according to formula 6.5.



The average TN inputs in the reference period to Danish Straits were 71,728 tonnes and the confidence interval according to formula (6.6) is [65,789; 77,667] tonnes. The confidence interval for the 2016 estimate is [53,177; 57,707] tonnes according to formula (6.7). Using formula (6.9) we calculate that the left side of the inequality sign is 11,288 tonnes, which is larger than zero, so we conclude that TN input in 2016 to Danish Straits is statistical significantly reduced (with 23%) since the reference period 1997-2003.

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

Table 6.1. Illustration of the traffic light system. Evaluation of progress towards reductions targets (nutrient inputs ceiling) of TP for Finland to Bothnian Bay (BOB), Bothnian Sea (BOS) and Gulf of Finland (GUF) based on normalized annual TP inputs from Finland during 1995-2014. Green: input ceiling are meet. Red: input ceilings are not fulfilled. Yellow: It cannot be judge with statistical certainty if input ceilings are fulfilled taking into account uncertainty.

Finland TP	BOB	BOS	GUF
A : Input ceiling	1668	1255	322
B: Estimated input 2014	1483	1248	647
C: Inputs 2014 including uncertainty (test value)	1531	1311	673
Extra reduction (A-C)	137		
Remaining reduction to fulfill MAI		56	351

7. Step by step analysis illustrated by HELCOM data examples

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

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

Firstly hydrological normalization is performed individually for all rivers that discharge into the Baltic Proper. The normalization of the River Vistula (Poland) TN loads is given as an example.

The relationship between log-transformed inputs of TN and runoff in River Vistula is shown in figure 7.1.



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

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

Figure 7.3 shows the model estimated for the time series. The model consist of one change point in year 2000 and the linear components before and after 2000 are both statistical significant.

Figure 7.1. Linear regression on annual TN inputs (tonnes) and runoff for River Vistula in Poland.





Figure 7.3. Linear trend fit to annual normalized water + airborne TN inputs (tonnes) to the Baltic Proper during 1995-2016.

The result of the Mann-Kendall trend test is a highly significant downward trend (two-side test, Z=-4.06, P<0.0001) in both section of the time series as there is a break point in year 2000. The slope in the section from 1995 to 2000 is estimated to be -16.630 tons per year and from 2000 to 2016 it is -2.946 t per year (linear regression, Theil-Sen slope). The total change in input over the period is estimated to be -21% (using formula (5.1) Theil-Sen slope), -25% using formula (5.2) (linear regression) consisting of -16.1% from 1995 to 2000 and -10.9% from 2000 to 2016.

The nitrogen input ceiling (target) for the Baltic proper is set to 325,001 tons. The model estimated normalized load in 2016 is 386,869 tons and the test value using formula 6.5 is calculated to be 399,163 tons, both are well above the target, so the traffic light would be red for the basin (see figure 7.4).

The 95% confidence interval for the estimated normalized TN input in 2016 s (386,869 tons) is [371,989; 401,750]. For the reference period 1997-2003 the mean normalized TN inputs is 437,010 tons with a 95% confidence interval of [417,213 ;456,807], since the left hand side of the inequality in formula (6.9) equals 27,953 which is larger than zero so the estimated TN inputs for 2016 is statistical significantly lower (11%) than the average TN input during the reference period.

Figure 7.4. Testing the target value for water and airborne TN inputs (tonnes) to the Baltic Proper for the period 1995-2016. Full line is target (MAI), "......" line is estimated value (inputs) in 2016, and "-----" line is test value (inputs including statistical uncertainty) according to formula 6.5.



8. Discussion and recommendations

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

- Good data quality and consistency are necessary to conduct reliable statistical analyses of the available time series. Time series may include gaps and/or suspect/dubious values. In chapter 2 of this report, methods for filling in gaps and how to determine if a dubious value is an outlier are described.
- Regarding total uncertainty in country data: It is a difficult task to calculate the exact uncertainty for the data provided by the contracting parties. One potential method may be to apply the simpler method DUET-H/WQ described in Harmel et al. (2009), which gives an approximation to the total uncertainty in monitored catchments. Information on the uncertainty of nutrient inputs in unmonitored areas has to be given by the Contracting Party either by model uncertainty or as an expert evaluation.
- Normalization of nutrient inputs should be performed using the method based on transformed inputs and runoff. Transformation should be undertaken using the natural logarithmic function (see formula 4.6 in this report). Normalization is carried out for each catchment (river) separately, and normalized inputs can be summed up at country or at Baltic Sea subbasin level. Normalization is a necessary step before conducting trend analysis. The method ensures that variation in annual inputs is significantly reduced, contributing to test for a significant trend in inputs by allowing identification of minor trends as being statistically significant. If a decision is made to use monthly input time series in the future, similar normalization methods can be applied to the monthly data (see Silgram and Schoumans (ed., 2004)).
- Concerning trend analysis, the Mann-Kendall non-parametric trend method is recommended for testing a significant monotone trend in the normalized time series. The method is robust although autocorrelation can deflate the power of the test as it will for all statistical test methods. We assume that the autocorrelation in the yearly time series of nutrient inputs is of minor importance and therefore see the Mann-Kendall trend test as very good approximation. This non-parametric method can be used on both "raw" nutrient time series, normalized time series and runoff (climate) time series. If it is decided to use monthly input time series in the future, the Kendall trend test has been extended to a seasonal version (Hirsch and Slack, 1984). It is suggested to use the Mann-Kendall trend method for a first analysis of possible trends. Parametric methods like regression is suggested for a more detailed analysis of trends and their magnitude (as for the slope, intercept, last year inputs, changes in inputs etc.).

- If the time series show two or more distinct trends (trend reversal), two or more linear trends should be applied to model the time series. The change-point can either be determined by visual inspection of the time series plot or by a statistical method (Carstensen and Larsen, 2006). It is suggested to limit the number of change points to two for time series of 20 years of length. And each part of time series before or after a change point should at least be five years or more.
- Estimating the change in nutrient inputs can be done by the non-parametric Theil-Sen slope estimator. The method assumes a constant change, i.e. a linear trend. As change point analysis is applied, it is recommended to use linear regression for estimating slope and intercept in PLC assessment and for evaluation progress towards reductions targets. If the trend is not linear, a non-linear model or start-end difference should be used.
- To follow up HELCOM BSAP nutrient reduction targets a statistical method is needed in order to evaluate whether the reductions requirements have been fulfilled, and to quantify remaining reduction/distance to the targets. For time series with a non-significant trend, the equation in formula 6.1 can be used to calculate the estimated latest year nutrient input (trend method) and evaluate this value against the target value taking into account uncertainty. For time series with a significant linear trend equation in 6.4 should be used. For times series without trend or without trend in the part after a break point including the latest year the average of input in part should be used as the estimated input. We have defined a traffic light system allowing evaluation of nutrient inputs from varying catchments/Contracting Parties to the Baltic Sea according to defined targets.

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)

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

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

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

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

HELCOM, 2013 by 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.

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

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

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

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) Choise of sampling strategy and estimation for calculating nitrogen and phosphorus transport in small lowland streams. Hydrological Processes, 10, 1483-1501.

Kronvang, B., Larsen, S.E., Windolf, J. & Søndergaard, M. 2014. Uncertainties in monitoring – how much do we tolerate. Presentation of NJF Workshop, 19-21 March 2014, Stavanger, Norway. Larsen, S.E. & Svendsen, L.M. 2013. Statistical aspects in relation to Baltic Sea Pollution Load Compilation. Task 1 under HELCOM PLC-6. Aarhus University, DCE – Danish Centre for Environment and Energy, 34 pp. Technical Report from DCE – Danish Centre for Environment and Energy No. 33. http://dce2.au.dk/pub/TR33.pdf

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.

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

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

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

The Mann-Kendall's trend test is a non-parametric statistical method, which means that the method has fewer assumptions than a formal parametric test method. The data do not need to follow a Gaussian distribution as in ordinary linear regression but should be without serial correlation. Furthermore, the method tests for monotone trends and not necessarily linear trends, and it thus tests for a wider range of possible trend shapes. The direction of the monotone trends may be either downward or upward without any specific form. The power of the Kendall trend method is slightly lower than ordinary linear regression if the time series data are Gaussian distributed and the trend is actually linear, as this will encompass the slightly less restrictive assumptions.

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

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

where

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

If either x_i or x_i is missing, then $sgn(x_i - x_i) = 0$ per definition.

The trend is tested by calculating the test statistic:

$$Z = \begin{cases} \frac{S-1}{\left(\operatorname{var}(S)\right)^{\frac{1}{2}}} & S > 0\\ 0 & S = 0\\ \frac{S+1}{\left(\operatorname{var}(S)\right)^{\frac{1}{2}}} & S < 0 \end{cases}$$

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

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

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

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

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

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

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

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

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

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

Calculate:

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

Calculate:

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

and

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

where

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

Lower and upper confidence limits are the M_1 th largest and the $(M_2 + 1)$ th largest value of the *N* ranked slope estimates d_{ii} .

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

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

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

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

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

df	95% percentile
14	1.761
15	1.753
16	1.746
17	1.740
18	1.734
19	1.729
20	1.725
21	1.721
22	1.717
23	1.714
24	1.711
25	1.708
26	1.706
df	97.5% percentile
df 14	97.5% percentile 2.145
df 14 15	97.5% percentile 2.145 2.131
df 14 15 16	97.5% percentile 2.145 2.131 2.120
df 14 15 16 17	97.5% percentile 2.145 2.131 2.120 2.110
df 14 15 16 17 18	97.5% percentile 2.145 2.131 2.120 2.110 2.101
<i>df</i> 14 15 16 17 18 19	97.5% percentile 2.145 2.131 2.120 2.110 2.101 2.093
<i>df</i> 14 15 16 17 18 19 20	97.5% percentile 2.145 2.131 2.120 2.110 2.101 2.093 2.086
df 14 15 16 17 18 19 20 21	97.5% percentile 2.145 2.131 2.120 2.110 2.101 2.093 2.086 2.080
df 14 15 16 17 18 19 20 21 22	97.5% percentile 2.145 2.131 2.120 2.110 2.101 2.093 2.086 2.080 2.074
df 14 15 16 17 18 19 20 21 22 23	97.5% percentile 2.145 2.131 2.120 2.110 2.101 2.093 2.086 2.080 2.074 2.069
df 14 15 16 17 18 19 20 21 22 23 24	97.5% percentile 2.145 2.131 2.120 2.110 2.101 2.093 2.086 2.080 2.074 2.069 2.069
df 14 15 16 17 18 19 20 21 22 23 24 23	97.5% percentile 2.145 2.131 2.120 2.110 2.101 2.093 2.086 2.080 2.074 2.069 2.064 2.060
df 14 15 16 17 18 19 20 21 22 23 24 25 26	97.5% percentile 2.145 2.131 2.120 2.110 2.101 2.093 2.086 2.080 2.074 2.069 2.064 2.060 2.060 2.056

STATISTICAL ASPECTS IN RELATION TO BALTIC SEA POLLUTION LOAD COMPILA-TION

Task under HELCOM PLC-7 project

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

ISBN: 978-87-7156-389-4 ISSN: 2245-019X

