Baltic Marine Environment Protection Commission

Assessment methodology for contaminants in biota, sediment and water

The assessment protocol is structured in three main parts:

1) changes in log concentrations over time are modelled,
2) check for compliance against threshold value and evidence for temporal change of contaminant concentration per station and
3) a spatial aggregation of status per assessment unit.

Contents

It should be noted that the assessment protocol makes the assumption that monitoring data stems from the same monitoring stations during consecutive years. The stations used by the protocol are defined in the ICES Station Dictionary. Stations with similar station name are grouped together, but it is also possible to define a group of stations with different names to be defined as the same station in the Station Dictionary. Usually a station is defined in the Station Dictionary with coordinates and a valid box around these coordinates, but coordinates outside of the box will only give a warning when reporting the data, and are not used in the actual data extraction.

Overview

Time series of contaminant concentrations are assessed in three stages:

1. For sediment, the concentrations are normalised prior to the assessment to account for changes in the bulk physical composition of the sediment such as particle size distribution or organic carbon content. The concentrations are log transformed and changes in the log concentrations over time are modelled using linear mixed models. The type of temporal change that is considered depends on the number of years of data:
    1. 1-2 years: no model is fitted because there are insufficient data
    2. 3-4 years: concentrations are assumed to be stable over time and the mean log concentration is estimated
    3. 5-6 years: a linear trend in log concentration is fitted
    4. 7+ years: more complex (smooth) patterns of change over time are modelled
2. The fitted models are used to assess status against available threshold value and evidence of temporal change in contaminant levels in the last twenty years
3. The fitted models are also used for spatial aggregation to assess regional status against available threshold value and evidence of temporal change in contaminant levels on a scale 4 level HELCOM assessment unit.

These stages are described in more detail below. There is also information on how the methodology is adapted when there are 'less-than' measurements, i.e. some concentrations are reported as below the detection limit, and missing uncertainties, i.e. the analytical variability associated with some of the concentration measurements was not reported.

Normalisation

In most sub-regions, the concentrations are first normalised to account for changes in the bulk physical composition of the sediment such as particle size distribution or organic carbon content. Normalisation requires pivot values, estimates of the concentrations of contaminants and normalisers in pure sand. A normalised concentration is given by:

$$c_{ss} = c_x + \frac{(c_m - c_x)(n_{ss} - n_x)}{n_m - n_x}$$

where

- *Css* is the normalised concentration of the contaminant
- *Cm* is the measured concentration of the contaminant
- *cx* is the pivot concentration for the contaminant
- *nss* is the reference concentration of the normaliser
- *nm* is the measured concentration of the normaliser
- *nx* is the pivot concentration for the normaliser

The analytical standard deviation *u* of the normalised concentration is estimated from:

$$u^2 = \left(\frac{n_{ss} - n_x}{n_m - n_x}\right)^2 \left(u_c^2 + \left(\frac{c_m - c_x}{n_m - n_x}\right)^2 u_n^2\right)$$

where *uc* and *un* are the analytical standard deviations of the contaminant and normalised concentration measurements respectively. These are submitted with the data where they are known as ‘uncertainties’.

Metal concentrations are normalised to a standard sediment with 5% aluminium. The pivot values *cx* and *nx* and reference concentration *nss* depend on the digestion method used in the chemical extraction and can be found here. Organic concentrations are normalised to a standard sediment with 5% organic carbon content and, regardless of the digestion method, *nss* = 5. For organics, the contaminant and normaliser pivot values are both 0, so the formulae above simplify to:

$$c_{ss} = \frac{c_m n_{ss}}{n_m}$$

and

$$u^2 = \left(\frac{n_{ss}}{n_m}\right)^2 \left(u_c^2 + \left(\frac{c_m}{n_m}\right)^2 u_n^2\right)$$

Modelling changes in log concentration over time

The log concentrations are modelled by a linear mixed model of the form:

- response: log concentration
- fixed: f(year)
- random: year + sample + analytical

The fixed effects model describes how log concentrations change over time (year), where the form of f(year) depends on the number of years of data (described in the next paragraph). The random effects model has three components:

- year: random variation in log concentration between years. Here, year is treated as a categorical variable
- sample: random variation in log concentration between samples within years. When there is only one sample each year, this term is omitted and implicitly subsumed into the between-year variation
- analytical: random variation inherent in the chemical measurement process. This is assumed known and derived from the the ‘uncertainties’ reported with the data. Specifically, if *ui*, *i* =1...*n*, are the uncertainties associated with concentrations *ci* (expressed as the standard deviations of the concentration measurements), then the standard deviations of the log concentration measurements log *ci* are taken to be *ui*/*ci*. Measurements with *ui*>*ci* (i.e. an analytical coefficient of variation of more than 100%) are omitted from the time series.

The model is fitted by maximum likelihood assuming each of the random effects are independent and normally distributed (on the log concentration scale)[1]

The form of f(year) depends on the number of years of data:

1-2 years

no model is fitted as there are too few years for formal statistical analysis

3-4 years

mean model f(year)=μ

there are too few years for a formal trend assessment, but the mean level is summarised by μ and is used to assess status 5-6 years

linear model f(year)=μ+βyear

log concentrations are assumed to vary linearly with time; the fitted model is used to assess status and evidence of temporal change

7+ years

smooth model f(year) = s(year)

log concentrations are assumed to vary smoothly over time; the fitted model is used to assess status and evidence of temporal change

The last case requires more explanation. When there are 7-9 years of data, both a linear model and a smoother (thin plate regression spline) on 2 degrees of freedom (df) are fitted to the data. Of these, the model chosen to make inferences about status and temporal trends is the one with the lower Akaike's Information Criterion corrected for small sample size (AICc)[2]. When there are 10-14 years of data, a linear model and smoothers on 2 and 3 df are fitted, with the chosen model that with the lowest AICc. And when there are 15+ years of data, a linear model and smoothers on 2, 3, and 4 df are fitted, with model selection again based on AICc. Effectively, the data determine the amount of smoothing, with AICc providing an appropriate balance between model fit and model parsimony[3].

[1] Such models cannot be readily fitted in the R statistical environment because the **analytical** variance is assumed know. Instead, the likelihood is maximised directly using the optimal function. Ideally, the models should be fitted by restricted maximum likelihood (apart from when being used for likelihood ratio tests), but this has not been implemented yet.

[2] AICc is a model selection criterion that gives greater protection against overfitting than AIC when the sample size is small. For contaminant time series, small sample sizes correspond to few years of data. AICc is not formally defined for mixed models, but the usual definition is adapted to give a sensible criterion for the models considered here. The usual definition of AICc is

AICc = - 2 log likelihood+2*kn*/(*n*−*k*−1)

where ***n*** is the sample size and ***k*** is the number of parameters in the model. For a contaminant time series, the natural definition of the sample size is the number of years of data, ***N*,** say. The number of parameters in the number of fixed effects parameters, k_{fixed}, plus the number of (unknown) variance parameters, k_{random}. For example, the linear model has k_{fixed} = 2 and k_{random} = 2 (or 1 if the sample variance component is subsumed into the year variance component). This suggests using

AICc = - 2 log likelihood+2(k_{fixed}+k_{random})*N*/(*N*−k_{fixed}−k_{random}−1)

However, the denominator now overly penalises models because the 'sample size' is the number of years and, whilst subtracting k_{random} correctly corrects for the year variance component, it also corrects for the sample variance component which measures within-year variation. (Indeed, the denominator = 0 if ***N*** = 5 and the linear model is fitted, or ***N*** = 3 or 4 and the mean model is fitted). It therefore makes sense to take k_{random} in the denominator to be 1, corresponding to the year variance component, giving

AICc = - 2 log likelihood+2(k_{fixed}+k_{random})*N*/(*N*−k_{fixed}−2)

The denominator is now analogous to that used in a linear model with a single normally distributed error term. The AICc is still undefined when $\boldsymbol{N}$ = 3 and the mean model is fitted, but this doesn't matter in practice.

[3] Methods for estimating the smoothing degrees of freedom as part of the fitting process, for example by treating the amount of smoothing as an extra variance component, are available for several classes of models. However, such methods are not implemented in R for the case when the residual variance (the **analytical** variance) is known. This is a topic for future development.

Assessing environmental status and temporal trends

Environmental status and temporal trends are assessed using the model (preceding paragraph) fitted to the concentration data.

Environmental status is assessed by:

- calculating the upper one-sided 95% confidence limit on the fitted mean log concentration in the most recent monitoring year[4]
- back-transforming this to the concentration scale
- comparing the back-transformed upper confidence limit to the available assessment criteria

For example, if the back-transformed upper confidence limit is below the threshold value, then the median concentration in the most recent monitoring year is significantly below the threshold value and the status at the specific station are considered good. For an example, see Fryer & Nicholson (1999).

No formal assessment of status is made when there is only 1 or 2 years of data. However, an ad-hoc assessment is made by:

- calculating the median of the log concentration measurements in each year
- back-transforming these to the concentration scale
- comparing the back-transformed median log concentration (1 year) or the larger of the two back-transformed median log concentrations (2 years) to the assessment criteria.

Temporal trends are assessed for all time series with at least five years of data. When a linear model has been fitted (i.e. when there are 5-6 years of data, or if there are 7+ years of data and no evidence of nonlinearity), the statistical significance of the temporal trend is obtained from a likelihood ratio test[5] that compares the fits of the linear model $f(\text{year})=\mu+\beta\text{year}$ and the mean model $f(\text{year})=\mu$. The summary maps show a downward or upward trend if the trend is significant at the 5% significance level.

When a smooth model has been fitted, a plot of the fitted model is needed to understand the overall pattern of change. However, the summary map in the core indicator, presenting results per station, focusses on just one aspect of the change over time: the change in concentration in the most recent twenty monitoring years; i.e. between 1996 and 2015 (the assessment only includes data up to 2015). For this, the fitted value of the smoother in 2015 is compared to the fitted value in 1996 using a t-test, with significance assessed at the 5% level. The correlation between the two fitted values is accounted for by the t-test. If the time series does not extend to 2015, then the fitted value in the last monitoring year is used instead. Similarly, if the time series starts after 1996, the fitted value in the first monitoring year is used.

[4] Approximate standard errors on the fixed effects parameter estimates are obtained from the Hessian matrix. These are used to estimate standard errors on the fitted values, with confidence intervals based on a t-distribution with $\boldsymbol{N}-\boldsymbol{k}_{\text{fixed}} - 1$ degrees of freedom. One-sided t-tests of whether the fitted value in the last monitoring year is below the

assessment criteria can be found on the Statistical analysis page on the right hand side of the summary map under Graphics. The standard errors can be computed analytically (i.e. without using the Hessian), but this hasn't been implemented yet. The degrees of freedom for the t-tests is a sensible approximation because, for time series models, the natural definition of the 'sample size' is ***N***, the number of years of data (see discussion on AICc above). However, if the year variance is small compared to the other variances, the degrees of freedom might be too small leading to a loss of statistical power. This is a topic for future development.

[5] These tests have a type 1 error that is larger than the nominal value. For example, tests conducted at the 5% significance level will find 'significant' trends in more than 5% of time series, even when there are no trends. Using the standard error of the estimate of *β* from a restricted maximum likelihood fit of the linear model would be one way to improve the situation. Better still would be to use the Kenward Roger modification of F tests for linear mixed models (Kenward MG & Roger JH, 1997; Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood, Biometrics 53: 983-997).

Methods for estimating regional status

Contaminants

The status of each time series was summarised by the difference between the estimated mean log concentration in the final monitoring year and the log assessment concentration. This ensures that status is always measured on the same scale, even though the assessment criterion might vary between determinands and time series.

For determinand groups with more than one determinand (metals in biota and sediment, PAHs in biota and organo-bromines in biota), the following linear mixed model was fitted by restricted maximum likelihood:

response: status (mean log concentration - log assessment concentration)

fixed model: region : determinand

random model: station + status estimation variation + residual variation

The fixed model means that a separate regional status is estimated for each determinand.

The random model has three terms:

- station allows for variation in trend between stations common to all determinands
- status estimation variation is the variance of the status estimates from the individual time series analysis, assumed known and fixed
- residual variation is the variation that cannot be explained by any of the fixed effects or the other random effects

For determinand groups with only one determinand (POPs and CBs in biota), a simpler mixed model was fitted:

response: status (mean log concentration - log assessment concentration)

fixed model: region

random model: status estimation variation + residual variation

Imposex

The status of each time series is summarised by the ratio between the estimated mean VDS in the final monitoring year and the assessment critrion. This ensures that status is always measured on the same scale, even though the assessment criterion might vary between time series (because VDS is measured in a different species). The ratio is transformed to the square root scale (to better satisfy modelling distributional assumptions) and the following mixed model was fitted:

response: status (sqrt(mean VDS / assessment criterion))

fixed model: region

random model: status estimation variation + residual variation

Treatment of ‘less-than’ measurements

Overview

Measurements reported as below the detection limit are often known as ‘less-than’ measurements. Less-thans are examples of left-censored data. Provided there are not too many less-thans, the same contaminant time series models can be fitted provided the likelihood is adjusted accordingly. Further refinements are required to prevent over-fitting if there are many less-thans or if the less-thans are unevenly distributed across the time series. When most of the data are less-thans, a non-parametric test is used to compare levels with assessment criteria.

Adjustments to the likelihood

The likelihood when some measurements are less-thans is straightforward when there is only one measurement each year, because the measurements are then (assumed to be) statistically independent. Let *yi*be the logarithm of the reported concentration in year *ti*,*i*=1...*N*, and let *A* be {*i*: *yi* is a non-censored measurement} and *A*¯ be {*i*: *yi*is a less-than}. The likelihood of the data is then:

$$\prod_{i\in A}\phi\left(\frac{y_i-\mathrm{f}(t_i)}{\omega_i}\right)\prod_{i\in\bar{A}}\Phi\left(\frac{y_i-\mathrm{f}(t_i)}{\omega_i}\right)$$

where *ϕ* is the density function of a standard normal distribution (with zero mean and unit variance), Φ is the corresponding cumulative density function, f(*ti*) is the expected value of *yi*, and *ωi* is the standard deviation of *yi* given by:

$$\omega_i^2=\sigma^2_{\mathrm{year}}+\sigma^2_{\mathrm{sample}}+\sigma^2_{\mathrm{analytical},i}$$

where *σ*year, *σ*sample, *σ*analytical,*i* are the between-year, between-sample and analytical standard deviations respectively. Note that the analytical standard deviations are measurement specific and are based on the uncertainties reported with the data.

The likelihood is more complicated when there are several measurements in a year, because these measurements are dependent. Extending the previous notation, let *yij* be the logarithm of the *j*th reported concentration in year *ti*, and let *Ai* be {*j*: *yij* is a non-censored measurement} and *A*¯*i* be {*j*: *yij* is a less-than}. Then the likelihood of the data is

$$\prod_i\int_{-\infty}^{\infty}\phi\left(\frac{z-\mathrm{f}(t_i)}{\sigma_{\mathrm{year}}}\right)\prod_{j\in A_i}\phi\left(\frac{y_{ij}-z}{\omega_{ij}}\right)\prod_{j\in\bar{A}_i}\Phi\left(\frac{y_{ij}-z}{\omega_{ij}}\right)\mathrm{d}z$$

where *ωij*is the within-year standard deviation of *yij* given by:

$$\omega_{ij}^2 = \sigma_{\text{sample}}^2 + \sigma_{\text{analytical},ij}^2$$

Refinements

Less-than measurements contain less information about changes in concentration over time than non-censored measurements. Therefore, the form of f(t) fitted to the data is based on N+, the number of years of data with at least one non-censored measurement, rather than N, the total number of years of data (although N is also considered for short time series). Specifically:

N+≤1

no model is fitted

N+=2 and N=2

no model is fitted

2≤N+≤4 and N≥3

mean model f(t)=μ

5≤N+≤6

linear model f(t)=μ+βt

N+≥7

smooth model f(t)=s(t)

Smoothers on 2 degrees of freedom (df) are considered when 7≤N+≤9, on 2 and 3 df when 10≤N+≤14 and on 2, 3, and 4 df when N+≥15.

For consistency, N+ is also used instead of N in the calculation of AICc and residual degrees of freedom.

When N+ is relatively small compared to N, the model fits can become environmentally implausible, particularly if there are changes in the limit of detection over time, or if a linear or smooth model is fitted and the years at the start and end of the time series only have less-than measurements. To protect against this behaviour, three additional constraints are placed on the time series.

1. The time series is truncated from the left (i.e. early years are omitted) until N+≥N/2. For example, if there are ten years of data (each with a single measurement) and the measurements in years 6, 7, and 9 are non-censored, then the time series assessed comprises the data from years 5 through 10.

2. If a linear or smooth model is fitted (i.e. N+≥5), then the first year of the time series is taken to be the first year with a non-censored measurement (i.e. all earlier years, which only contain less-thans, are omitted). For example, if there are ten years of data and the measurements in years 3, 4, 6, 8, 9, and 10 are non-censored, then the time series assessed comprises the data from years 3 through 10.

3. If a linear or smooth model is fitted (i.e. N+≥5), and the measurements in the most recent year(s) of the time series are all less-thans, then the expected concentration in the most recent year(s) is assumed to be constant. Specifically, if $t$last is the last year with a non-censored measurement, then f(t) is adjusted to:

$$\mathrm{f}(t) = \begin{cases} \alpha + \beta t, & \text{if } t < t_{\text{last}} \\ \alpha + \beta t_{\text{last}}, & \text{if } t \geq t_{\text{last}} \end{cases}$$

for the linear model and similarly for the smooth model.

Non-parametric assessment of environmental status

If the length of the truncated time series is 2 years of less, then there are insufficient years to fit a parametric model and make a formal assessment of environmental status. However, if the original time series has more than five years of data, a one-sided sign test is used instead to provide a non-parametric test of status. The median log concentration measurement each year is first calculated (with less-thans treated as if they were non-censored measurements) and then back transformed to the concentration scale. These can be thought of annual contaminant indices. The indices in the last twenty years (the same period used to assess recent trends for the summary maps) are then used to test the null hypothesis: H_0: median concentration ≥AC against the alternative: H_1: median concentration < AC, where AC is the assessment criterion[1].

[1] This approach might lack power, particularly for longer time series where there are non-censored measurements at the start of the time series, but all recent measurements are less-thans. In such cases, a better approach might be to model how the probability that the annual index is below the AC changes with year and to use the upper one-sided 95% confidence limit on the fitted value in the final monitoring year to assess status. This is a topic for future development.

Missing uncertainties

Sometimes uncertainties are not reported with the concentration measurements, usually when the data were submitted before the reporting of uncertainties became mandatory. In such cases, the uncertainties are estimated using fixed and relative standard deviations derived from the uncertainties that have been reported to the ICES data base (extraction date: 14 December 2015).

Suppose *c* is a concentration reported with missing uncertainty. Then the uncertainty *u* is estimated from:

$u^2=s^2+f^2v^2c^2$

where:

- *S* is the fixed standard deviation, taken to be one third of the detection limit when reported, and otherwise the value estimated from the uncertainties in the ICES database.
- *v* is the relative standard deviation, estimated from the uncertainties in the ICES database
- *f* is an inflation factor, based on available quality assurance information, that increases the relative error if analytical performance is poor

The inflation factors are based on QUASIMEME Z-scores supplied to ICES on CD by analytical laboratories and Certified Reference Material (CRM) concentrations held in the ICES database. The CRM concentrations (for the relevant contaminant, matrix and year) are converted to Z-scores

$$Z_i = \frac{M_i - A_i}{0.125 A_i}$$

where *Mi* are the measured concentrations and *Ai* are the certified reference values. These are then combined with the QUASIMEME Z-scores (for the contaminant, matrix and year) to give

$$f^2 = \frac{1}{n} \sum_i Z_i^2$$

where *n* is the total number of Z-scores. In practice, Z-scores that are large (possibly due to unit misreporting) can unduly influence the inflation factors, so are truncated at 3, leading to a maximum inflation factor of 3. The inflation factors are also truncated below at 1, since this corresponds to good analytical performance. If there is no quality assurance information (for the contaminant, matrix and year), the inflation factors are set to 3.

References

Fryer RJ & Nicholson MD, 1999. Using smoothers for comprehensive assessments of contaminant time series in marine biota. ICES Journal of Marine Science 56: 779-790.