Stockholms universitet

# The problem of trend detection in log–linear regression models when switching from single specimen to pooled sample design

Sebastian Nordlund

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2023:15
Matematisk statistik
Juni 2023

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2023:15**
http://www.math.su.se/matstat

# The problem of trend detection in log–linear regression models when switching from single specimen to pooled sample design

Sebastian Nordlund[*]

June 2023

## Abstract

Concentrations of contaminants in animal populations are usually assumed to be log–normally distributed. Contaminant concentrations can either be measured in tissue taken from individuals or by measuring the concentration in a mixture of tissue taken from several different individuals. When individual concentrations are measured, they can be transformed into log–concentrations and their sample mean is often used as a yearly input in linear trend models. When mixtures of tissue taken from different individuals is introduced as the reported concentration in an ongoing observational study where earlier concentrations were reported individually, a bias will propagate and give rise to biased estimations for the coefficient parameters of log–linear trend models (for as long as both types of input are used). This thesis derives a second–order Taylor series approximation for the distribution of the logarithm of such log–normal averages, and uses this to approximate the size of the biased trend estimator as a function of two sources of variance in log–concentration. When a trend estimator is biased, the probability of Type I error will not correspond to the nominal significance levels chosen for analysis. In order to approximate the true probability of Type I error, a parametric bootstrap simulation study is conducted, and a logistic regression model is then fitted to the outcomes of the simulations. A case study on mercury concentration in Baltic herring is also conducted, as a means to illustrate the problem in a practical setting and exemplify what the logistic regression model would predict regarding increased probability of Type I error and how the biased trend estimates differ from trend estimates of bias–corrected data.

[*] Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige. E-post: davidsebastiannordlund@gmail.com. Supervisor: Martin Sköld.

## Acknowledgements

I am thankful for the patience of all the different people I have spoken to during the course of writing this thesis. I have a tendency to become very interested in minor details, and this is naturally a double-edged sword since it is on the one hand a necessary criterion for gaining a deeper understanding of a topic, but it can also become very distracting and lead to an inefficient use of time. I am very glad that my supervisor Martin Sköld has helped me steer towards the main path during this semester, and that the two other students of my supervisor-student group, Max Brehmer and David Helms, have been as interested in my monologues as they have been. This gratitude is also extended to my friend Aron Benedicto and my older brother Max Nordlund, the latter who read one of my drafts and gave some frank feedback regarding the lengths of sentences and the disposition of the text. AI-chatbots have not been used during the work with this thesis, except for on one occasion when I described my current draft's content for Chat-GPT and asked it for an alternative disposition — an alternative disposition that was then discussed with my supervisor as inspiration for how the different parts of the thesis could best be combined. Finally, I would like to thank my fellow student Gustav Thorzén who very enthusiastically shared his interest in the log–normal distribution with me and gave his view of an article of Caudill et al that clarified some things for me.

# Contents

# 1 Introduction

This thesis focuses on how biased trend estimates arise in log–linear regression models when old data were collected as individual observations and new data are collected as averages of individual concentrations. Log–linear regression models are common across the sciences when effects are better modelled as multiplicative rather than additive, which is revealed visually as curved data where the empirical dispersion about the median response $M(y_i)$ is highly asymmetric and positively skewed at each setting $i$ of an explanatory variable (in the case of simple linear regression). In the prototypical case we would then have that the original response variable $Y$ is *log–normally distributed* at each setting $i$, and a log–transformation of $Y$ would allow us to fit a linear regression model where the effects on $\ln(Y)$ are linear. Hopefully such a log–transformation would mitigate the problems of curvature, heteroskedasticity and auto–correlation, so that the obtained linear model for $\ln(Y)$ is a convincing fit to data. Thereafter, we can use the estimated coefficient $\hat{\beta}$ of the log-linear regression model to approximate a non–linear model for the original variable $Y$. Simple log–linear regression models where *time* is the explanatory variable occur frequently in biomonitoring programmes of contaminant concentrations in animal and human populations alike. Monitoring of possible *trends* in contaminant concentrations is very important since the presence of contaminants could impact entire ecosystems as well as the health of individual persons, and it is therefore of interest that such trends are estimated with unbiased estimators.

An example of such a biomonitoring programme is *The Swedish National Monitoring Programme for Contaminants in Marine Biota* (**SNMPCMB**), which monitors 20 different contaminants (among them PFAS, PCB and Mercury) at currently 29 different *collection stations* around the Swedish West and East coasts. The programme fits both long–term trend models and short–term trend models for the different combinations of contaminant and collection station, and the trend estimates of these log–linear regression models are published in an report every two years. Trends are regarded as significant if P–values calculated for the t-statistic of $\hat{\beta}$ *under the null hypothesis that* $\beta = 0$ are less than five percent. The short–term models are estimated from intervals of ten years of data, and the thesis focuses its investigation of what the trend bias will be in such a setting where log–linear trend models are fitted to ten years of data, and how trend bias relates to the likelihood of making a *Type I error*.

For biomonitoring programmes that study the distribution of some particular contaminant concentration, one question of interest has been whether to measure concentrations in each individual specimen (*an individual specimen could for instance be an individual human*) or to make *homogenates* which are mixtures from the tissue collected in different specimen of the same population (*such as humans aged 0-3*) and then measure the concentration of the homogenate. In this thesis we refer to this practice of measuring contaminant concentration in each individual member of a sample as **Method I**. The alternative practice of measuring the concentration of a contaminant in a homogenate is correspondingly referred to as **Method II**. If a homogenate mixture is perfectly *homogenous* with regards to the concentration of some contaminant, and this mixture has been created by equally sized contributions from each included individual specimen, we can interpret it as a measure of *average concentration.* Homogenate mixtures are sometimes referred to as *pooled samples.* Gradually during the 2000's the SNMPCMB has begun applying Method II for some combinations of contaminant, specie and location. A very notable benefit of choosing Method II has been that the programme could spend less financial resources on chemical analysis and instead use these resources to fund an expansion of locations monitored (roughly doubling the number of sites where species are fished each year). For a detailed description of the SNMPCMB, see Sörensen and Faxneld [10].

The main focus of this thesis is in how a change from measuring individual concentrations (**Method I**) to homogenate concentrations interpreted as sample averages (**Method II**) yields a bias for the estimator of the coefficient parameter $\beta$ in log-linear regression models based on log–normal time series data, *if the data of the trend model includes measurements of both types.* This bias arises as a consequence of another bias, namely the positive bias that occurs when expected log–concentration $E[X] = \mu$ is estimated with $\ln(\bar{Y})$ instead of average log–concentration $\bar{X}$. Biased coefficient parameter estimates implies that *in hypothetical scenarios where there is no linear trend in the response variable with respect to time* a change from Method I to Method II should likely result in more rejections of the null hypothesis $H_0 : \beta = 0$ than the chosen significance level of our test statistics would suggest.

The thesis is outlined in the following manner. It begins by providing some brief background on the pecularities of the log–normal distribution, its role in biomonitoring of contaminant concentrations and a review of the literature regarding benefits of Method I compared to Method II in a biomonitoring context. In the *Theoretical framework* derivations of analytical expressions are provided for the bias and variance of the coefficient parameter estimator, and goes on to describe a method of how to correct the bias. After doing this, parametric bootstrap simulations are used to explore the question of how the bias seems to affect the likelihood of Type I error — and a logistic regression model is fitted to the outcomes of the simulations so that the probability of Type I error can be estimated. In the *Case study* the method for bias correction is applied on a time series of mercury concentrations in six geographically distinct herring populations dispersed around the Swedish coasts. Based on predictions given by a logistic regression model that was fitted to the outcomes of the aforementioned simulations, the case study suggests that the risk of Type I error — under the assumption of a chosen significance level threshold of 0.05 — is increased to reach a maximum ranging 0.054–0.115 for the six different stations when five years has passed since changing from Method I to Method II.

# 2 Literature review and background

## 2.1 Overview of the log-normal distribution

The log–normal distribution is defined by an anti-log transformation of a normally distributed variable. Whereas the normal distribution is symmetric about its median and mean (they both equal $\mu$), the log–normal distribution is asymmetric and positively skewed.

**Definition 1** *A random variable $Y$ is said to be log-normally distributed if*

$$Y := e^X$$

*where*

$$X \sim N(\mu, \sigma^2),$$

*and we may then write $Y \sim LN(\mu, \sigma^2)$.*

For a log–normal variable $Y$, we have the following expressions for the expected value, distributional median and the variance of such a variable:

$$E[Y] = e^{\mu + \frac{\sigma^2}{2}} \tag{1}$$

$$M(Y) = e^{\mu} \tag{2}$$

$$\mathrm{Var}(Y) = (e^{\sigma^2} - 1) \cdot e^{2\mu + \sigma^2} \tag{3}$$

There are some peculiarities to the log-normal distribution, in particular we have that sums of independent log-normals often have a rather slow rate of convergence towards a normal distribution (see Mitchell [8]) as well as a the fact that Moment Generating Functions are undefined in the neighborhood of zero for log-normal distributions (implying that these functions are unable to generate the sought-for moments) as described by Allan Gut in the textbook *An intermediate course in Probability.*[4] The second property does not present a problem if we are limited to estimating moments from a sample of i.i.d. log-normal random variables, because such moments have been derived by other methods, but if we instead wish to find the moments for a sum of non-identical (but still independent) log-normals difficulties will arise. Even greater difficulties arise when the log-normal terms are correlated. As a matter of fact, the problem also extends itself to the characteristic function $\varphi(t)$ for log-normals because even though it exists it does not have a convergent Taylor series representation when $t$ is complex according to Holgate [6]. The remaining option of instead relying on repeated application of the convolution formula is deemed practically unworkable by authorities on the subject matter, according to Dufresne.[3]

### 2.1.1 Why the log-normal distribution is (often) a suitable choice for modeling the distribution of contaminant concentrations

*The following paragraphs are intended for those readers who wish to gain an understanding of why the log-normal distribution is as commonplace as it is within the context of biomonitoring programmes for contaminants in animal populations, and can be skipped without loss of ability to understand later sections of the paper.*

Concentrations of contaminants in biological populations are usually not suitably described by normal distributions since they tend to exhibit positive skew when studied empirically. The reason that this positive skew arises can in part be attributed to the fact that a concentration of a substance is measured in strictly non–negative values — for example it could be measured in parts of millions — and does therefore not have a support across the entire range of real numbers as the normal distribution does. Contributing to the problem of bounded support is the circumstance that many monitored pollutants tend to do lethal harm in very low concentrations, which effectively means that the expected fraction of pollutant A in a tissue taken from species B will be tremendously much closer to a value of 0 (no detected occurrence) than to 1 (entire tissue mass consisting of pollutant). Therefore, observed concentrations are very likely to reside either directly at the lower bound (due to chemical-analytical measurements being unable to detect concentrations below some given threshold) or at least very close to the lower bound.

Under the circumstances where data is realized as right-tailed with a low positive mean in comparison to both the sample variance as well as the support of data (for example a range from 0 to a $10^6$ when the unit is micrograms per gram), Limpert et al [7] suggest that the log–normal distribution may serve as the most suitable model for capturing (*at least approximately*) the underlying data generating process. Log–normal models are very common across the sciences — see Limpert et al [7] or Dufresne [3] for examples — owing in part to their property that *the logarithm of original data is normally distributed.* When log–transforming concentration measurements we have to decide how to handle zero–valued observations (which originate from true non–zero concentrations being below the detection limit), as the limit of $\ln(0)$ diverges to negative infinity, but the details of these imputation considerations are beyond the scope of this thesis. If zero–valued measurements do not occur in data, all observations of *log–concentrations* will be real-valued numbers and their support will be the half-closed interval of $(-\infty, \ln(N)]$ instead of the fully-closed interval of $[0, N]$ for the original concentrations. In this context $N$ denotes the theoretical maximum concentration that could be detected (such as one million if concentration is measured in micrograms per gram), but as mentioned we can expect the highest observed concentration in a sample to be very much closer to zero than $N$, and therefore the theoretical problem of right-truncation is of no practical importance.

## 2.2 Overview of biomonitoring of contaminant concentrations

One reason for measuring the concentration in each individual specimen is that it will provide us with information that readily allows us to estimate variance among individuals of the same population. A second reason to measure individual concentrations is that *if the observations of the collected sample are assumed to be log-normally distributed* then a simple log-transform of the observations will yield a corresponding group of normally distributed *log-concentrations*, which could be preferred in some contexts.

However, there are also reasons to instead consider making homogenates and measure their concentration. One such reason is that if the cost of collecting an individual specimen is much lower than the cost of sending it for chemical analysis (and thereby measure its concentration), then a study design where *many* different homogenates formed for the same target population and where each homogenate is created from the tissue of *even more* individual specimen *could very well result in better point estimates of concentration and yet allow for inference about the variance among individuals.* Individual variance can be estimated from the between-homogenates variance and the Law of Large Numbers implies that the average homogenate concentration is an unbiased estimator of the expected value for the concentration. Caudill et al [2] mention another reason to choose the homogenate approach (referred to as *pooled samples* in their paper), namely that homogenates tend to have a larger volume than individual mixtures of tissue (*or saliva in Caudill's case*), and larger volumes can sometimes increase the span that concentrations can be detected in by decreasing the *Limit of Detection (LOD).* Bignert et al [1] estimate the cost-effectiveness of the two approaches and makes suggestions about when either one is preferable. The study does not, however, consider what consequences there will be in situations where the purpose is to fit log-linear regression models to time series data where concentrations are reported both from individual specimen (for some years) and from homogenates (for other years). In fact, it seems that there has been no published study that examines this particular problem.

# 3 Theoretical framework

One can imagine a scenario when a sample of $n$ individuals has been collected and the distribution of a contaminant in their target population is log-normal, and that one could choose to apply both Method I and Method II. If the homogenate has been perfectly mixed so that the the contaminant is homogenously distributed within the mixture, and also that each individual sample member has contributed with an equal amount of tissue, then the homogenate concentration can be interpreted as a measure of the sample average concentration $\bar{y}$. In theory the homogenate concentration would therefore equal the arithmetic average of the $n$ individual measurements, and we would consider the former a unbiased point estimate of $E[Y] = e^{\mu + \frac{1}{2}\sigma^2}$. In the context of biomonitoring of contaminants it is more common that the primary parameter of interest is $\mu$, the expected value of the normally distributed log-concentration, and an unbiased estimator for this quantity is the average log-concentration $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} \ln(Y_i)$.

Caudill et al [2] observe that a first–order Taylor series approximation for *the logarithm of* $\bar{Y}$ (in the neighborhood of $E[\bar{Y}]$) will let us approximate $E[\ln \bar{Y}] \approx \mu + \frac{1}{2}\sigma^2$, and therefore a bias of (approximately) $\frac{1}{2}\sigma^2$ arises when $\ln(\bar{Y})$ is used as an estimator of $\mu$. Higher precision for the approximation of this bias can be obtained by deriving a second–order Taylor series, and such a Taylor series is derived *but with additional layers of realism* that requires a more thorough treatment than the approximation used by Caudill et al. One source of additional complexity is introduced when there is as a random group effect $b_t \sim N(0, s^2)$ which is i.i.d. across all years $t$ included in the time series, that affects all members of a collected sample. Such an effect is included in the assumed model (see model 4 in section 3.1.1). Another detail that is accounted for is the fact that several *replicate* homogenates are used, and so it is not an individual $\ln(\bar{Y})$ that is chosen as an estimator of $\mu_t$ (the expected log-concentration for year $t$) but *the average* of the log-homogenate concentrations, and this estimator is introduced as $\tilde{X}$ in Equation 9.

The first part of the Theoretical framework introduces the model for individual log-concentration and how $Bias(\hat{\beta})$ relates to $Bias(\tilde{X})$. In the second part a method is proposed for correcting $Bias(\hat{\beta})$ where one first attempts to correct $Bias(\tilde{X})$ by calculating bias-correction terms from the derived formula for this bias with relevant parameter estimates as input, and then proceeding with fitting corresponding log-linear trend models to bias-corrected datasets. The theoretical framework concludes with the third part, which presents a simulation study that investigates how the risk for Type I error can modelled in different scenarios.

## 3.1 The bias incurred to $\hat{\beta}$ when switching from Method I to Method II

### 3.1.1 The assumed log-linear model

In order to answer the question of what consequences there will be for trend monitoring based on log-normal data *and in which circumstances there will be any consequences of practical importance*, when switching from Method I to Method II we shall begin by first assuming that the log-concentration of year $t$ is generated by a simple linear regression model.

**Model 1** *We assume the following model for log-concentration of some particular contaminant in an individual specimen for a given year:*

$$X_{it} = \ln(Y_{it}) = \tilde{\alpha} + \beta(t - \bar{t}) + b_t + \epsilon_{it}, \tag{4}$$

*where*

$$b_t \sim N(0, s^2) \tag{5}$$

*represents yearly variation attributable to heterogeneity between individuals collected from different samples. We also assume that*

$$\epsilon_{it} \sim N(0, \sigma^2), \tag{6}$$

*and this term is interpreted as a source of variance between individuals collected simultaneously.*

The $\beta$ parameter of Equation 4 leads to an approximate yearly change of concentrations which is

$$\textbf{"Yearly percentual change"} \approx (\exp(\beta) - 1) \cdot 100.$$

As an illustrating example for one source of variation between samples, consider an individual fish that feeds on seaweed in some geographically limited habitat, and assume that we can monitor the concentration of mercury in the fish's liver in real–time. Let us also assume that there is a specific source for mercury contamination in the fish's habitat, for example a broken thermometer, and that the fish can therefore feed on its seaweed at different distances from this source. It seems reasonable to expect that the likelihood of measuring a large increase in the concentration of mercury in the fish is drastically heightened when the fish is feeding on seaweed one meter away from a broken thermometer compared to when it is residing a hundred meters away from aforementioned mercury source. The example illustrates why we should not be too surprised to find a random effect between different samples of a specie taken from the same geographical area, such as a variance between different *fish shoals* (groups of fish swimming together), because animals that move together in groups are likely to have been near the same contaminant sources. This phenomenon contributes to the random group effect $b_t$ for each group of collected specimen, and is included in Model 1.

The question of what the consequences of the method shift would be if the data is better described as being generated from a non–linear distribution is beyond the scope of this paper. However, as mentioned in section 6.2 of the appendix, formal tests for non–linearity are limited to longer time series for the SNMPCMB, so the assumption of linearity seems uncontroversial for the Case study (section 4) given that the focus of this thesis is short–term models.

### 3.1.2 Estimating time trends when switching observational study design

When individual specimen concentrations are measured (Method I), the mean log-concentration at (a particular year $t$) will be a normally distributed variable denoted $\bar{X}_t$. It is defined as

$$\bar{X}_t = \frac{\sum_{i=1}^{n} X_{it}}{n}, \tag{7}$$

and because it is the arithmetic average of $n$ i.i.d. normally distributed random variables it follows that

$$\bar{X}_t \sim N(\tilde{\alpha} + \beta(t - \bar{t}), s^2 + \frac{1}{n}\sigma^2). \tag{8}$$

Hence, when Method I is used, variance is decreased by increasing $n$ (the number of individuals per sample) and by improving the sample design so that the characteristics of the samples resembles each other more throughout the years (thereby decreasing the variance of $b_t$).

On the other hand, when homogenates are used (Method II), the relevant variable is defined as

$$\tilde{X}_t = \frac{\ln(\bar{Y}_{1t}) + ... + \ln(\bar{Y}_{mt})}{m}, \tag{9}$$

where each $\bar{Y}_{ht} = \frac{1}{k}\sum_{i=1}^{k} Y_{iht}$ represents concentration measured in homogenate $h$ in year $t$. We have that

$$E[\bar{Y}_{ht}] = E[Y_{iht}] = \exp\{\tilde{\alpha} + \beta(t - \bar{t}) + \frac{1}{2}(s^2 + \sigma^2)\},$$

but the task of finding a second–order Taylor approximation for the distribution of $\tilde{X}_t$ (and thereby be able to approximate its expected value and variance) requires careful consideration and its derivation is presented in the appendix (section 6.1). Here it is given as the following lemma.

**Lemma 1** *A second–order Taylor series expansion for $\tilde{X}_t$ asserts that*

$$\tilde{X}_t \approx N\left(\mu_t + \frac{s^2+\sigma^2}{2} - \frac{e^{s^2}-1}{2} - \frac{e^{\sigma^2}-1}{2k}, \frac{Var(\ln(\bar{Z}_k))}{m} + \frac{Var(\ln(V))}{m}\right), \tag{10}$$

*and hence*

$$Bias(\tilde{X}_t) = \mu_t - E[\tilde{X}] = \frac{s^2+\sigma^2}{2} - \frac{e^{s^2}-1}{2} - \frac{e^{\sigma^2}-1}{2k}, \tag{11}$$

*where **m** is the number of homogenates and **k** the number of individuals used to form each homogenate.*

Now assume that we will estimate a short–term trend for a period of 10 years, where $R$ years of data are gathered by Method I (measuring concentrations in individual specimen) and $10-R$ years of data are gathered by Method II. Let

$$(\omega_1, ..., \omega_r, \omega_{r+1}, ..., \omega_{10}) = (\bar{X}_1, ..., \bar{X}_r, \tilde{X}_{r+1}, ..., \tilde{X}_{10})$$

be a vector of yearly estimators for yearly log-concentrations $\mu_t$ ($\mu_t = \tilde{\alpha} + \beta(t - \bar{t})$). We shall now proceed to determining the analytical expressions for $E[\hat{\tilde{\alpha}}]$ and $E[\hat{\beta}]$ when $Bias(\tilde{X})$ is left neglected. Applying the expressions for least-squares estimators given to us by Sundberg [11], we have that the intercept is estimated by

$$\hat{\tilde{\alpha}} = \bar{\omega}$$

and the trend-coefficient will in its turn be estimated by

$$\hat{\beta} = \frac{\sum_{t=1}^{10}(t-\bar{t})\omega_t}{\sum_{t=1}^{10}(t-\bar{t})^2}.$$

Linearity of expectation, and the assumption that each year's estimator $\omega_t$ of $\mu_t$ is independent from the other years', yields the following to expressions for the expectations of the least-squares estimators:

$$E[\hat{\tilde{\alpha}}] = \frac{\sum_{t=1}^{10} E[\omega_t]}{10} = \frac{\sum_{t=1}^{10} \mu_t}{10} + \frac{\sum_{t=r+1}^{10} Bias(\tilde{X}_t)}{10} = \tilde{\alpha} + \frac{\sum_{t=r+1}^{10} Bias(\tilde{X}_t)}{10},$$

$$E[\hat{\beta}] = \beta + Bias(\hat{\beta}) \tag{12}$$

where

$$Bias(\hat{\beta}) = \frac{\sum_{t=r+1}^{10}(t-\bar{t})Bias(\tilde{X}_t)}{\sum_{t=1}^{10}(t-\bar{t})^2} \tag{13}$$

We have that under the assumed model for log-concentration (Model 1), $Bias(\tilde{X}_t)$ will be identical for all years $[R+1, \ldots, 10]$ when the number of individuals making up a homogenate is constant throughout the years. This has the consequence that for positive (negative) values of $Bias(\tilde{X}_t)$ the bias of $\hat{\beta}$ will reach a maximum (minimum) when $r = 5$ and decrease (increase) symmetrically about the maximum (minimum) until reaching zero when $r \in \{0, 10\}$ (i.e. when all years apply the same method). Along with the expression for $Bias(\hat{\beta})$ one may also be interested in finding an expression for $Var(\hat{\beta})$ that holds even when the usual assumptions of linear regression are violated.

With an approximate distribution for $\tilde{X}$ one can also approximate $Var(\hat{\beta})$ when switching from Method I to Method II after $R$ consecutive years of using the first measurement method. Since we have that yearly measurements $\omega_i$ are independent from other yearly measurements and given that they are approximately normally distributed regardless of whether they are of the form of $\bar{X}$ or $\tilde{X}$, the fact that $\hat{\beta}$ is a linear combination of 10 $\omega_t$:s lets us approximate its variance by the following formula

$$Var(\hat{\beta}) \approx \frac{1}{(\sum_{i=1}^{10}(t_i-\bar{t})^2)^2}\sum_{i=1}^{10}(t_i-\bar{t})^2 Var(\omega_i). \tag{14}$$

The necessary derivations for a second-order Taylor series approximation of the distribution of $\tilde{X}$ are presented in the appendix (see section 6.1).

## 3.2 A proposed method for correcting $Bias(\hat{\beta})$

The outcome of a control that our assumptions about $\epsilon_{it}$ (Equation 6) and $b_t$ (Equation 5) are supported by data is *ideally* what should guide our choice of estimators for $\sigma^2$ and $s^2$. If data does not seem to support the notion that $\epsilon_{it}$ and/or $b_t$ are i.i.d. across the years investigated, the question of how an alternative model should be specified is a topic worth considering in its own right. Since we have limited our investigation to Model 1 and because it is under the assumption of such a data generating process that the simulation study of section 3.3 is designed, it is assumed that $\epsilon_{it}$ and $b_t$ are not serially correlated.

In Equation 12 one can see that the expected value of $\hat{\beta}$ can be viewed as the sum of the underlying parameter $\beta$ and a bias-term, where the latter is a function of $Bias(\tilde{X}_t)$ for each yearly $\mu_t$–estimator $\omega_t$. When Method I is used for a particular year $t$, the corresponding $\omega_t = \bar{X}_t$ is an unbiased estimator of that particular $\mu_t$, but when Method II is used $\omega_t = \tilde{X}_t$ and the estimator will instead be biased. If one solves the problem of obtaining unbiased estimators for the variances of $\epsilon_{it}$ (which is $\sigma^2$) and $b_t$ (which is $s^2$), then one can either choose to

i) compute $\mathbf{Bias}(\hat{\beta})$ directly and subtract this bias from the uncorrected $\hat{\beta}$ estimate **or**

ii) compute $\mathbf{Bias}(\tilde{X}_t)$ and subtract it from each log-concentration measurement $\omega_t$ (originating in a year when Method II was used), and thereafter fit a linear model.

Whereas both approaches are equivalent with regard to obtaining an unbiased estimate of $\beta$, the second alternative has the benefit of letting a statistical software calculate a confidence interval and a p-value for $\hat{\beta}$ when working with data in an applied setting. Such p-values and confidence intervals could still be inefficient, since the bias-corrected estimators $(\tilde{X}_t - Bias(\tilde{X}_t))$ might have a different variance than the unbiased estimators $(\bar{X}_t)$, implying heteroscedasticity. A final word of caution is that since our expressions for various biases are based on second-order Taylor series approximations, *bias-corrected estimators* are therefore only *approximately unbiased.* We shall now turn our attention to how $\sigma^2$ and $s^2$ can be estimated in a scenario where the number of replicate homogenates is deemed to low for inference about variance between replicates to be meaningful. In such a scenario, only data gathered by Method I is useful when seeking to estimate $\sigma^2$ and $s^2$.

Given that we have assumed that $\epsilon_{it}$ is i.i.d. across both individuals $i$ and years $t$, an unbiased estimator for $\sigma^2$ is the pooled sample variance $\hat{\sigma}_p^2$ which is calculated as the *weighed average* of yearly sample log-variances $\frac{1}{n-1}\sum_{i=1}^{n}[X_{it} - \bar{X}_t]^2$ for all years when Method I was used. If the variance of the random group effect is negligible, and Model 1 is indeed the correct specification for the distribution of individual log-concentration, then a linear regression model for $\bar{X}_t$ can be expected to produce an estimated residual variance that is very close to $\frac{1}{n}\hat{\sigma}_p^2$. On the other hand, if the variance of the random group effect is non-negligible (but Model 1 is still the correct specification), then one can expect the residual variance to be somewhat greater than $\hat{\sigma}_p^2$. In such a case, we can estimate the size of the variance of $b_t$ as

$$\widehat{s^2} = Var(\mathbf{Residual}) - \frac{\widehat{\sigma^2}_p}{n}.$$

On the other hand, if $\widehat{Var}(\mathbf{Residual}) \leq \frac{\widehat{\sigma^2}_p}{n}$, we could instead decide to neglect the random group effect $b_t$ since in that case there does not seem to be much evidence for its existence.

Once $\widehat{s^2}$ and $\widehat{\sigma^2}_p$ have been estimated, we may use our second order Taylor expansion for the expected value of $\tilde{X}$ and from each observed $\tilde{x}_t$ subtract an estimate of its associated bias. This estimate is calculated as

$$\widehat{Bias}(\tilde{X}) = \frac{\widehat{s^2} + \widehat{\sigma^2}_p}{2} - \frac{e^{\hat{s^2}} - 1}{2} - \frac{e^{\hat{\sigma^2}_p} - 1}{2k}, \tag{15}$$

which except for the neglecting of higher order terms would mean that we now have unbiased estimates for $\mu_t$.

A special circumstance that could motivate a modification of described approach is when data seem to indicate a clear point in time when estimated sample variance is altered, for example if there is knowledge of a change of chemical analysis methods that greatly increased the range of detectable values and sample variances are decidedly higher after such a chemical analysis method has been put in production. In such cases, it might seem more reasonable to base the pooled variance estimate on the available yearly sample variances from (and including) the year when the new chemical analysis method was first used.

## 3.3 Estimation of Type I error probability by simulations

As elaborated on in section 3.1, a bias occurs for $\hat{\beta}$ when the first series of consecutive years $[1, R]$ report the unbiased $\mu_t$–estimator of $\bar{X}$ and the remaining years $[R+1, 10]$ report the biased $\mu_t$–estimator of $\tilde{X}$ (assuming that $R \in [1, 9]$). Because $\hat{\beta}$ is a biased estimator for $\beta$ under these circumstances, the *assumed* distribution for the t–statistic associated with $\hat{\beta}$ under $H_0 : \beta = 0$ is incorrect (as in *assumed by statistical software packages* when a linear regression model is specified). In such a setting it is extremely unlikely that the chosen significance level (which is 0.05 in this case) will coincide with the true probability of rejecting the null hypothesis, even though data is generated from Model 1 with $\beta = 0$. Knowing that the assumed distribution for the t–statistic of $\hat{\beta}$ under the null hypothesis is wrong does not imply that deriving a correct specification for that statistic is a trivial task. Therefore, simulations are utilized as a means for estimating how often $H_0 : \beta = 0$ is rejected when we vary $R$ across $[1, 10]$ as well as the size of the standard deviations associated with $\epsilon_{it}$ and $b_t$.

The number of simulated observations for $\bar{Y}$ at each year when Method II is ”applied” is set to $m = 2$ (the typical number of homogenates used per location and year by the SNMPCMB), and the number of individuals per ”homogenate” (the interpretation of each $\bar{Y}$) is $k = 12$. The number of simulated individual log-concentrations per site and year when Method I is ”applied” is also set to $n = 12$. For each setting of $R$, 300 trials are simulated for each chosen combination of random inter–sample variance $s^2$ (*where homogenates created for the same year and location are viewed as originating in the sample*) and intra–sample variance $\sigma^2$. In the conclusion of each trial a simple linear regression model is fitted to the simulated outcomes for log–concentration estimates, and depending on if the reported p–value for $\hat{\beta}$ is below 0.05 the trial is categorized as a *rejection* (and otherwise as a *non–rejection*). Each trial can therefore result in either a success (if the trial is classified as a non–rejection) or a failure (if classified as a rejection).

Based on estimates of $\sigma$ and $s$ in the Case study of section 4, it seems motivated that simulations are to be carried out at 42 equally spaced points in the ($s$, $\sigma$)–rectangle of $[0, 0.6] \times [0.1, 0.6]$ (our estimates from the Case study suggest that $\sigma \in [0.25, 0.51]$ and $s \in [0.20, 0.49]$). The outcomes from all simulations within the rectangle are then used to fit logistic regression model, $M_{smooth}$, by applying the *glm()* procedure of the **R** programming language [9]. $M_{smooth}$ is a hierarchical model which has as its highest-order term a mixed-effect between all predictors. The predictors of the model are $Bias(\hat{\beta})$, $Var(\hat{\beta})$, $R$ and $R^2$ (the squared number of consecutive years), and all were treated as continuous. $M_{smooth}$ should be regarded as an attempt at fitting smoothed lines that predict outcomes within a small area of the ($s$, $\sigma$)-space, and so these lines can be interpreted as qualified guesses about what the Type I-error frequency would be if we were to simulate a very large number of ten-year time series at each point in the ($s$, $\sigma$, $R$)-space.

Since there is no bias if one chooses to rely *either* exclusively on Method I (setting $R = 10$) *or* exclusively on Method II (setting $R = 0$), consequently the observed percentage of rejections (our point estimate of Type I error risk) will converge in probability to the nominal significance level when the number of trials $N$ tends to infinity. When both methods are used (meaning that $R$ is in the interval of $[1, 9]$) there will be a bias that is expected to increase the probability Type I error above and beyond the chosen significance level, but this effect could possibly be offset by increased heteroscedasticity when there is a relatively large difference between $Var(\bar{X})$ and $Var(\tilde{X})$. The predicted percentage of rejections given by the logistic regression model $M_{smooth}$ can therefore be regarded as rough approximations of what the Type I error risk is in settings closely related to the ones illustrated in the Case study, and how by how much increases in heteroscedasticity does offset a biased estimator for $\beta$.

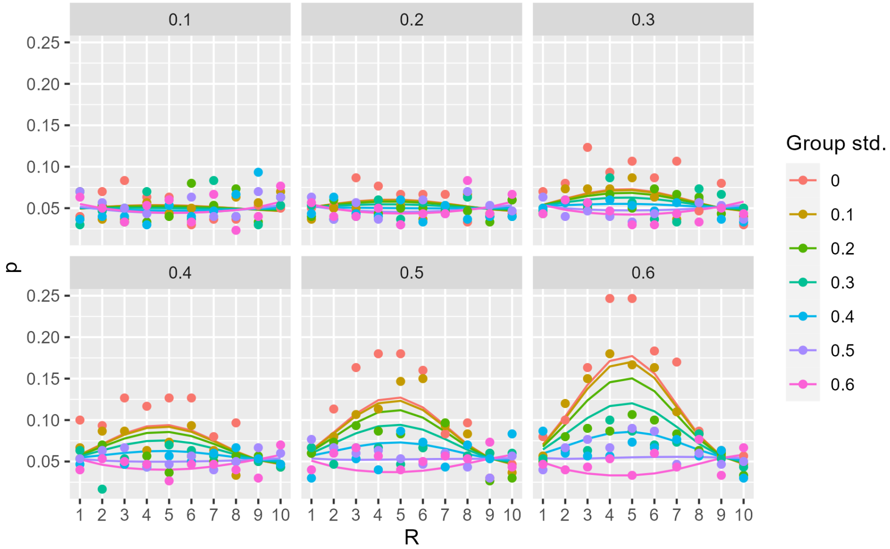

Figure 1: Outcomes of simulations in the $[0, 0.6] \times [0.1, 0.6]$–rectangle

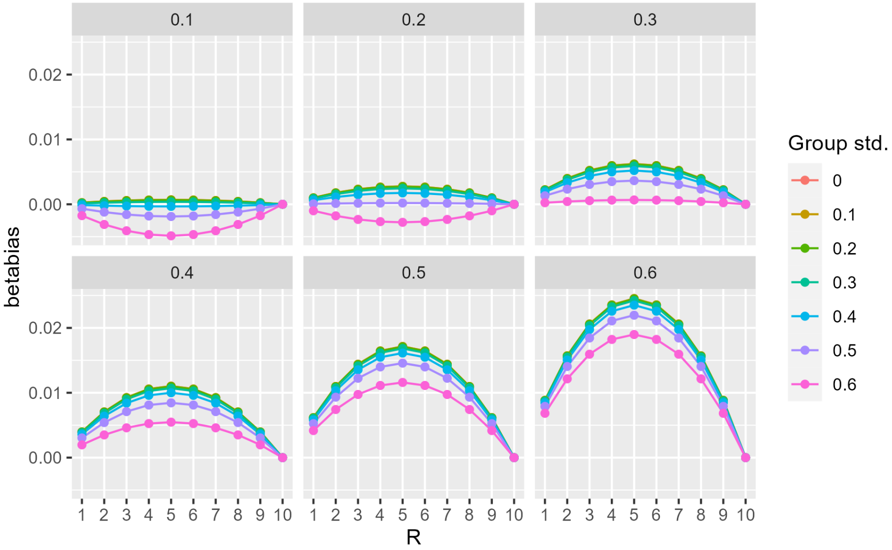

Figure 2: Calculated values of $\hat{\beta}$ in the $[0, 0.6] \times [0.1, 0.6]$–rectangle

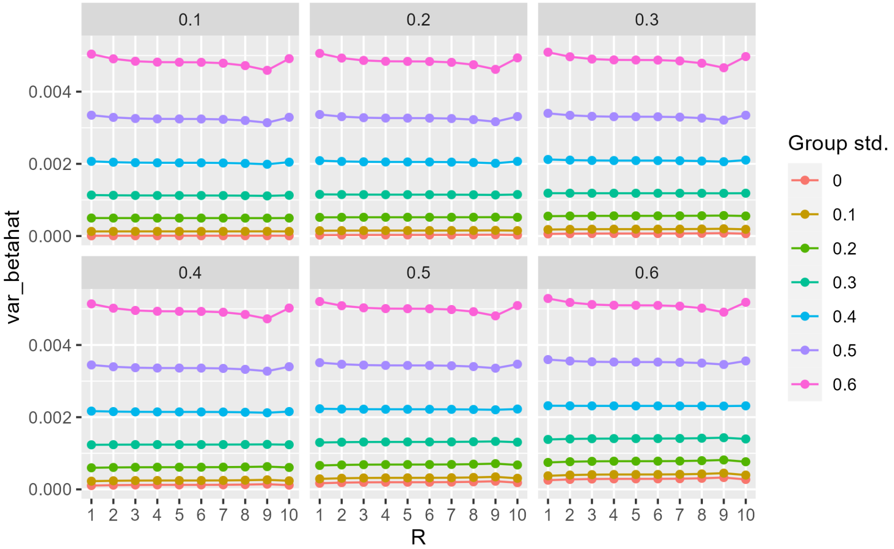

Figure 3: Calculated values of $Var(\hat{\beta})$ in the $[0, 0.6] \times [0.1, 0.6]$–rectangle

In figure 1 it is shown how *percentage of rejections* (denoted $p$ in the left–hand side of the figure) varies across different combinations of $s$ (denoted *group std.* and indicated by the seven different colors), $\sigma$ and $R$. The uppermost left facet shows the settings where $\sigma = 0.1$ whereas the lower facet on the right side shows the settings where $\sigma = 0.6$. When making a comparison with the figures of 2 and 3, it seems rather clear that for a given combination of $s$ and $\sigma$, $Var(\hat{\beta})$ is almost constant across the settings of $R$ while $Bias(\hat{\beta})$ can vary somewhat, and that fitted values for estimated percentage of rejections (seen in the smoothed curves suggested by model $M_{smooth}$ ) allows us to conclude that the lower the variance of $\hat{\beta}$ for a given $Bias(\hat{\beta})$ the higher the probability of Type I error.

# 4 Case study: mercury concentrations in herring

As an application of the bias correction method presented in section 3.2, log-linear models are fitted to data from the Swedish biomonitoring programme SNMPCMB. The studied specie is *Clupea harengus* (Baltic herring) and the contaminant of interest is *mercury*. Under the assumption that Model 1 is a correct specification for individual yearly log-concentration, estimates are calculated for the variance of the inter-sample random group effect $b_t$ and intra-sample individual variance of $\epsilon_{it}$. Since the number of homogenates per location and year is rather small (typically 2-3), inference about $Var(\tilde{X})$ from differences between homogenates was deemed to not be meaningful and therefore only the collection stations where Method I has been used (at least historically) are investigated. The case study concludes with predictions about the probability for Type I error in hypothetical scenarios where $R$ varies from $R = 1$ to (and including) $R = 10$ and the distributions of $\epsilon_{it}$ and $b_t$ correspond to the calculated estimates for the six different collection stations. These predictions are based on the logistic regression model that was used for plotting smoothed curves for the simulated outcomes of the preceding simulation study.

## 4.1 Data

Mercury concentrations are reported in ng/g in dry liver tissue. The earliest observations are from 1990 while the latest ones are from 2020. Prior to 2007 all observations are individual measurements, but sample sizes range from 25 to 10 with 12 being the most common sample size (with 119 out of 164 individual-measurement samples having this size). Out of 20 fish collection stations, just 6 have data on individual concentrations with the remaining 14 stations only reporting pooled sample concentrations (typically with two such pooled samples per year). The stations reporting measurements of individual concentrations are Harufjärden, Landsort, Utlängan, Väderöarna, Ängskärsklubb and Fladen.

No missing concentration values are present in the dataset for the six stations of interest for all years *when concentrations are reported*, but some years lack observations for particular stations. We will not impute values for those combinations of year and station when an observation is present, and given that we will base our variance parameter estimates on series of data that span 1990–2020 (for all stations except Väderöarna which has a span of 1995–2020), the issue of individual missing years does not seem to be detrimental. Since the biomonitoring programme does not remove suspected outliers the same practice will be adhered to in this illustrative study (see section 6.2 in the appendix for details about current practices of the programme). Out of the six stations, Utlängan never reports pooled sample concentrations whereas the other five stations report pooled concentrations to varying degree. The list below summarises the distribution of individual and pooled sample measurements respectively, for the six stations.

- Harufjärden has individual measurements from 1990 until 2017, no data for 2018, and pooled sample measurements for 2019 and 2020.
- Landsort has individual measurements from 1990 until 2020 with the exception of 2015 when there are no data, and additionally this station also has a pooled sample measurement for 2020.
- Utlängan does only report individual measurements, and with the exception of 1999, 2006 and 2010 does so for the entire period 1990-2019.
- Väderöarna reports individual measurments from 1995 until 2017 (except for 2006 when there are no data) and thereafter two pooled sample measurements (of 12 individuals each) for the period 2018-2020.
- Ängskärsklubb reports individual measurements only in the periods of 1990-2007, 2009-2010, 2012-2014 and 2017-2018. However it also reports two pooled sample concentrations 2008 and three pooled sample concentrations in 2015, and one pooled sample 2011 along with individual concentrations for that same year.
- Fladen has exclusively individual concentration data for 1990-2018 as well as 2020, but also both pooled sample concentration data (one measurement) together with 12 individual concentrations for 2019.

Since Väderöarna and Harufjärden are the only two stations where median log-concentration estimation by Method II is the only option after a particular year $R$ (prior to which Method I was the only option) we shall later on concentrate our comparison of specific outcomes for these two stations, by constructing time series from 10 measurements for each of these two stations. The other four stations of interest will still be subjected to estimations of the variance parameters $s^2$ and $\sigma^2$, since this will yield insight into the magnitude of Type I-error inflation in the trend monitoring of mercury concentrations in Baltic herring. With *inflation* we here refer to the situation when in expected Type I-error frequency is higher than our chosen significance level.

## 4.2 Estimation of $\widehat{s^2}$ and $\widehat{\sigma^2}_p$

After filtering out pooled sample observations (that is, *homogenate concentrations*) from our dataset, the individual variance $\sigma^2$ was first estimated for each of the six aforementioned stations and calculated as the weighed average of each year's sample variance.

An interpretation of the $\sigma^2$ estimates in Table 1 is that given a variance of for example 0.186 (the estimate for Harufjärden), 95 percent of observations of mercury concentration are expected to have relative sizes around the median concentration ranging from a factor of $\exp(-1.96 \cdot \sqrt{0.186}) \approx 0.43$ to $\exp(1.96 \cdot \sqrt{0.186}) \approx 2.33$.

Figure 4 shows how yearly sample variances are distributed around their respective weighed pooled variance estimates. The volatility of these estimates

Table 1: Pooled variance estimates weighed for sample size

| Station name | $\widehat{\sigma^2}_p$ |
|---|---|
| Ängskarsklubb | 0.24 |
| Fladen | 0.09 |
| Harufjärden | 0.19 |
| Landsort | 0.26 |
| Utlängan | 0.22 |
| Väderöarna | 0.06 |

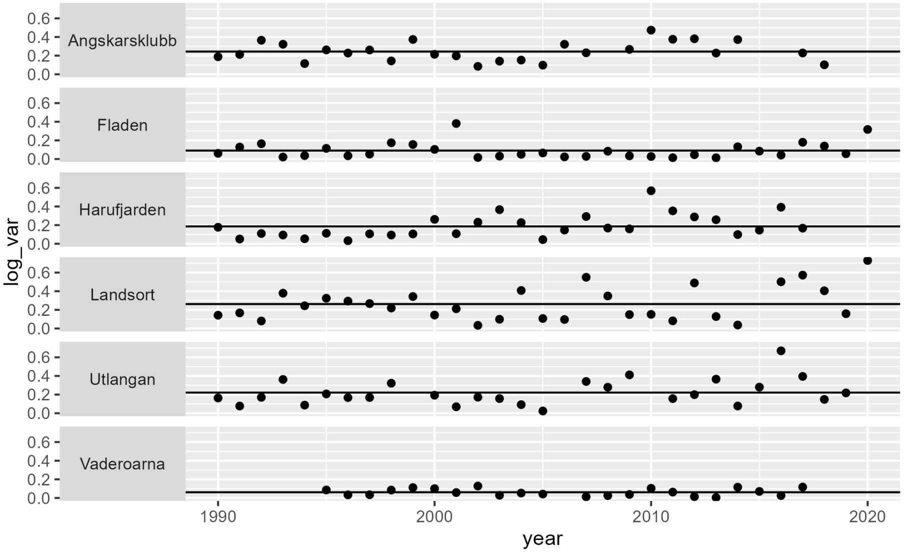

Figure 4: Distribution of yearly sample variances around $\widehat{\sigma^2}_p$

(measured as distance to $\widehat{\sigma^2}_p$) seems to increase in the second half of the period (from 2005 and onwards) for Landsort and Harufjärden, but this could possibly be caused by a reduction of sample size from (typically) 20 individuals prior to 1997 to 12 individuals from 1997 and beyond. For Harufjärden there is also another indication that the pooled variance approach might be problematic, because all sample variances prior to 2000 are below $\widehat{\sigma^2}_p$ whereas most of the sample variances thereafter are above this pooled estimate. Otherwise, the assumption that $\sigma^2$ is the same across the years seems to hold reasonably well.

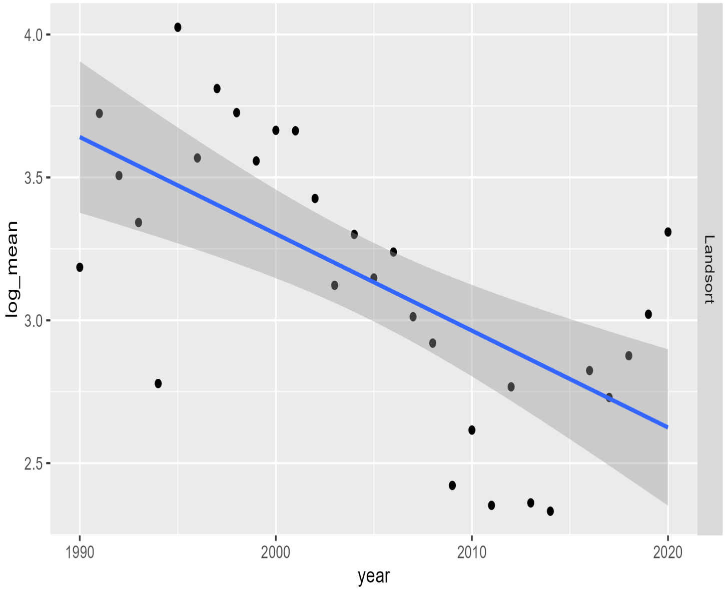

Figure 5: Log-linear regression fit for Landsort

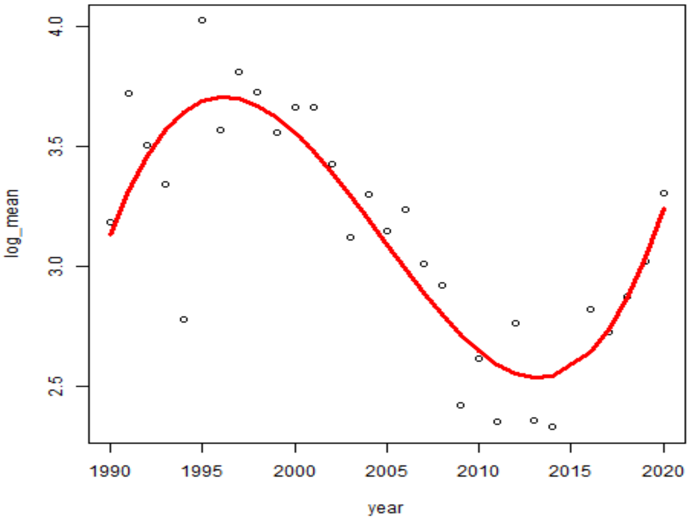

Figure 6: Third-order polynomial regression fit for Landsort

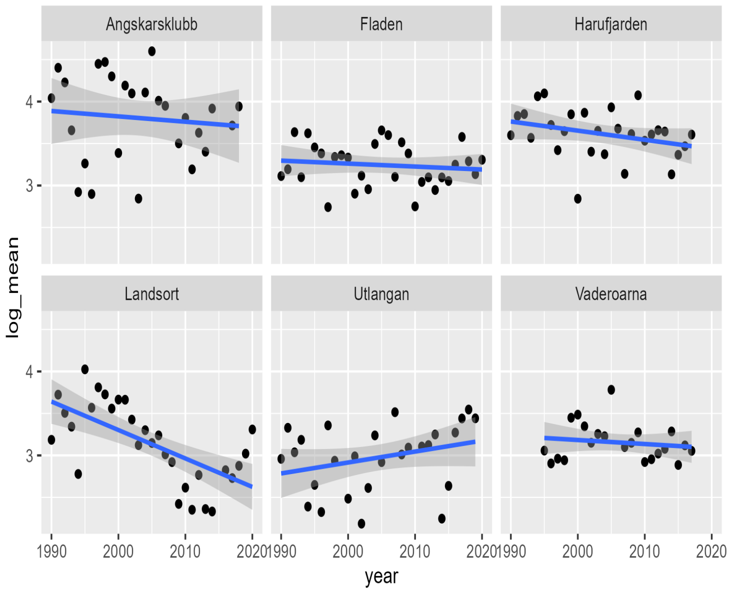

Figure 7: Fitted log-linear regression models for the six collection stations

A visual inspection of the distribution of average log-concentrations for the six stations (see Figure 7) seems to lend support to the notion of linearity with respect to time and therefore linear regression seems like a decent model specification with the notable exception of the model for Landsort (see Figure 5) where a non-linear pattern is present. In order to test if the assumptions of homoscedasticity and no autocorrelation hold for the fitted models Breusch-Pagan tests and Durbin-Watson tests were applied, with the result that the null hypothesis of no autocorrelation was rejected for Landsort under a 5 % significance level (p-value = 0.002). Non-linearity in a long-term time series is not of too much concern as long as these two assumptions seem to be met for the short-term time series of interest. The reason for this is that we are only seeking to estimate the residual variance around a regression curve of historical data so that we can then use it to estimate the between-group variance $s^2$ (as outlined in section 4.3 **A proposed method for correcting Bias($\hat{\beta}$)**), and if it seems likely that the historical data has been generated from a non-linear process then residual variance is better estimated from a non-linear model. A third-order polynomial regression model was therefore fitted for Landsort (see figure 6), and the residual variance estimate of this model was then used for further analysis.

## 4.3 Results

Based on the estimates for residual variances in the six aforementioned regression models, the method described in section 3.2 was applied to estimate $\hat{s^2}$

under the assumption that pooled samples will be limited in size to 12 (as they are for mercury in our dataset). Results are reported as squared roots of $\hat{s^2}$ and $\hat{\sigma^2}$ and can be found in Table 2. The logistic regression model fitted to the simulated outcomes was then used to make predictions about the probability of falsely rejecting a null hypothesis of $\beta_0 = 0$ under circumstances where $s$ and $\sigma$ are set to the values presented in Table 2 and are presented in figures 8 and 9. These results may therefore be interpreted as a qualified guess of what the Type I error percentage *would be* if the data generating process for individual log-concentration is specified as in Model 1 with $\beta = 0$ and one could repeat the experiment an unlimited number of times, for each one of the six stations.

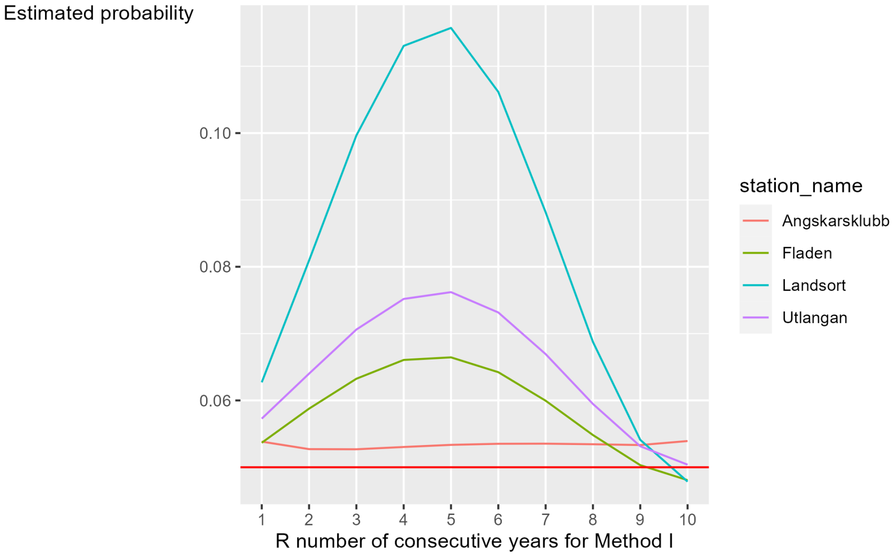

Figure 8: Predicted probability of making a Type I error in hypothetical scenarios corresponding to Ängskärsklubb, Fladen, Landsort and Utlängan

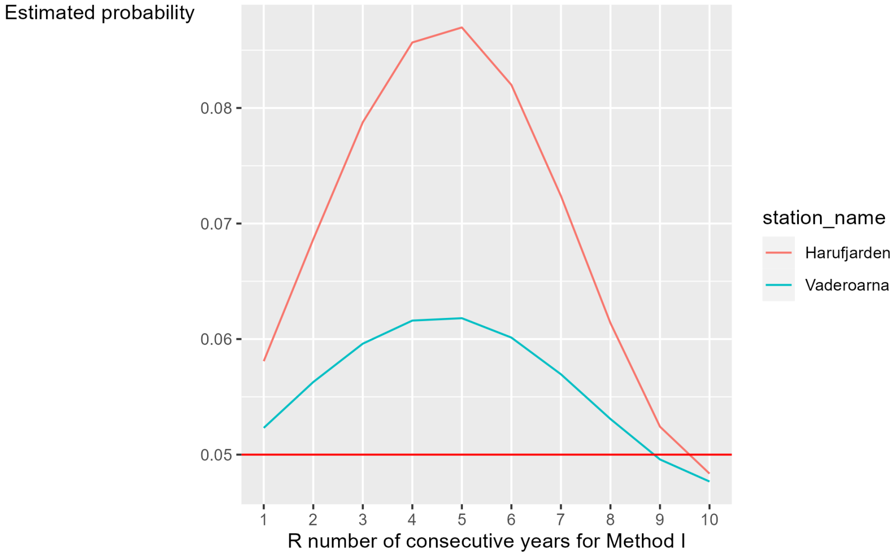

Figure 9: Predicted probability of making a Type I error in hypothetical scenarios corresponding to Harufjärden and Väderöarna

Table 2: Estimates of $s$ and $\sigma$ for the six stations

| Station name | $\sqrt{\hat{s^2}}$ | $\sqrt{\hat{\sigma^2}}$ |
|---|---|---|
| Angskarsklubb | 0.49 | 0.49 |
| Fladen | 0.24 | 0.30 |
| Harufjarden | 0.25 | 0.43 |
| Landsort | 0.20 | 0.51 |
| Utlangan | 0.37 | 0.47 |
| Vaderoarna | 0.21 | 0.25 |

As can be seen in both figures, the logistic regression model predicts that the probability of Type I error will increase with R until reaching a maximum when $R = 5$ (as we predicted in the Theoretical framework), and thereafter decrease and eventually come very close to the nominal significance level of five percent when $R = 10$. For the two stations that have made a shift from single-specimen measurements to pooled-sample measurements, Harufjärden and Väderöarna, the model guides us into believing that the (hypothetical) risk of Type I-error would roughly be the same as the chosen significance level in the ten-year time-series ending in 2020 (when $R = 8$) but that it will increase and reach a maximum of about a factor twice the significance level in year 2023 (when $R = 5$) with Harufjärden having its estimated maximum just below nine percent whereas Väderöarna would only have the Type I error risk increased to just above six percent. The predictions do also suggest that hypothetically, if Landsort was to make a shift from Method I to Method II, the maximum Type I error risk would be as high as about 11.5 percent.

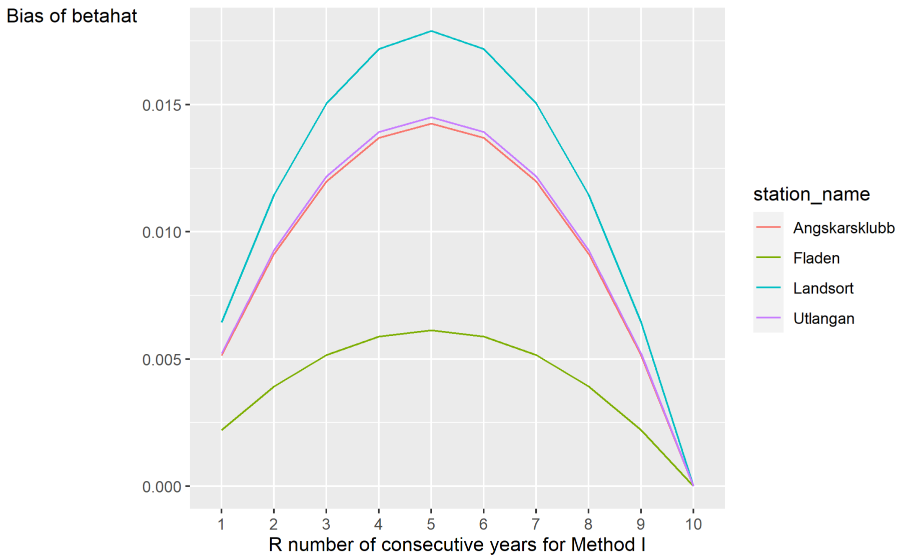

Figure 10: Estimated $Bias(\hat{\beta})$ corresponding to Ängskärsklubb, Fladen, Landsort and Utlängan

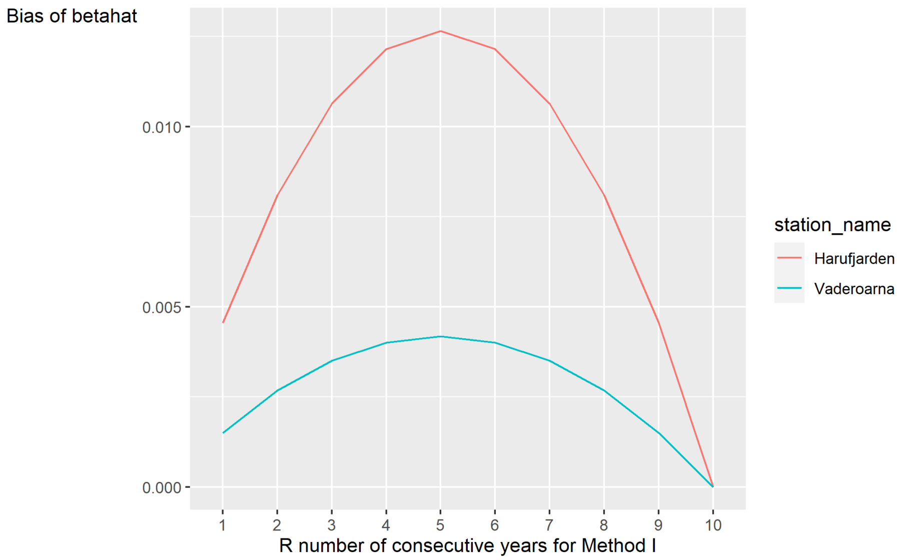

Figure 11: Estimated $Bias(\hat{\beta})$ corresponding to Harufjärden and Väderöarna

In figures 10 and 11 estimates for $Bias(\hat{\beta})$ based on $\hat{s}$ and $\hat{\sigma}$ are presented. The highest bias is seen when $R = 5$ for Landsort at $\approx 0.018$ which corresponds to overestimating the yearly percentual change in mercury concentration by approximately 1.82% per year five years after a *counterfactual* change of measurement method had been initiated. The effect may seem small, but if one would accidentally reject the null hypothesis in such a scenario, one would falsely conclude that there has been an approximate 19.7 % increase in mercury concentrations during the last ten years when in fact there has been no such true underlying increase. Also, even if there has been a positive trend, we should expect to estimate it by the same percentage since we have seen in equation 12 that $E[\hat{\beta}]$ can be expressed as the sum of the true parameter $\beta$ and its bias term $Bias(\hat{\beta})$. For the two stations where a change of measurement method has occurred, Harufjärden and Väderöarna, the corresponding overestimations of yearly percentual increases in concentration at $R = 5$ would be about 1.26 percent a year for Harufjärden and 0.4 percent a year for Väderöarna.

Table 3: Comparison of models based on bias corrected estimates of $\mu_t$ vs models without bias correction

| Station and type | $R^2$ | $\hat{\beta}$ | % yearly change in concentration |
|---|---|---|---|
| Harufjärden uncorrected | 0.0252 | 0.0090 | 0.9021 |
| Harufjärden corrected | 0.0186 | 0.0076 | 0.7663 |
| Väderöarna uncorrected | 0.0167 | 0.0087 | 0.8756 |
| Väderöarna corrected | 0.0088 | 0.0063 | 0.6312 |

In Table 3 features a comparison on model output from log-linear regression models based on ten years' data. For both collection stations, the last two available years' mercury concentrations have been measured by Method II, which implies that the model for Harufjärden spans 2010–2020 (due to a missing value in 2018) whereas Väderöarna spans 2011–2020. Since both models are instances of simple linear regression models, the $R^2$ statistic can be interpreted as the squared correlation coefficient between log-concentration and year, and even though this measure exhibits even lower values when biased–corrected estimates of $\mu_t$ are used (for Väderöarna the $R^2$ value is reduced by half) our conclusion would in all four situations be that there is no detectable trend present. Analogously the p-values are in all four cases much above the chosen significance level of 0.05. In the last column of the table, we can see how the point estimates of annual percentage change in mercury concentrations differ between models based on corrected and uncorrected data respectively — for Harufjärden and Väderöarna the estimated difference would in this case amount to merely 0.13 and 0.25 percentage points respectively per year.

# 5 Discussion

In this paper we have seen how the choice to switch from individual measurements of log–normal data to pooled–sample measurements yields biased trend estimates in log–linear regression models. We have derived analytical expressions for this bias in Equation 12 and shown how it can be expressed as a linear combination of biased estimators for $\mu_t$, and similarly derived an expression for the variance of this biased $\beta$ estimator in Equation 14. These two expressions were then used (together with $R$ *the number of consecutive years when individual measurements were used*) to fit a logistic regression model to the binary outcomes of rejection or non–rejection of the null hypothesis $H_0 : \beta = 0$ in 42 equally spaced settings of an area chosen to be somewhat near the obtained estimates of inter–sample variance $s^2$ and intra–sample variance $\sigma^2$ from the *Case study.*

As an illustration of the problem, a proposed solution for estimating inter–sample variance $s^2$ and intra-sample variance $\sigma^2$ (both defined as variance in *log-concentration*) based on historical data was applied to a dataset of mercury concentrations in Baltic herring collected at six different locations on the Swedish coasts. The fitted logistic regression model was used on the estimated parameters, and showed that if $H_0 : \beta = 0$ is true and we switch to estimating trends on time series that have both pooled sample measurements and individual measurements, we shall expect there to be a somewhat increased risk of falsely rejecting the null hypothesis after 5 years has passed (assuming a significance level of 0.05) compared to when adhering to one method of measurements, and the risk of Type I error at this setting seems to vary between 5.4 percent to 11.5 percent.

If one is only interested in a method for calculating unbiased $\beta$ estimates in linear regression models for log–transformed data — where original data is log–normally distributed and a shift of measurement methods from individual to pooled-sample measurements has occurred — one could also decide to instead calculate sample mean *concentrations* for earlier years' data, log–transform these values and then fit a linear regression model to this dataset. In that case, we see that since the expected value of $\tilde{X}_t$ can be approximated as a sum of $\mu_t$ and a constant, then the difference between the models should mainly be found in the intercept parameter $\alpha$ whereas the slope coefficient $\beta$ should be approximately the same for both models. The benefit of such an approach would be that it should in theory give us unbiased $\tilde{X}_t$ estimates without requiring us estimate between–group log–variance $s^2$ or between–individual log–variance $\sigma^2$ from historical data.

A potential topic for further investigation is how to estimate inter–sample variance $s^2$ and intra–sample variance $\sigma^2$ if one assumes a model for the data generating process of $X_{it}$ where these parameters are allowed to be serially correlated. For example, one could possibly be interested in simulating $N$ number of time series from a distribution for $X_{it}$ where again $\beta = 0$ but both $\epsilon_{it}$ and $b_t$ are dependent on time $t$, and then make a comparison of how effective two different methods for correcting the bias of $\hat{\beta}$ are (measured as the percentage of rejected null hypotheses). Especially an inquiry into how a bias–correction method for $Bias(\tilde{X}_t)$ which utilizes the distribution of differences between *logarithms of log–normal averages* (constructed from samples of the same year $t$) could be used to infer intra–sample variance $\sigma_t^2$, and under what circumstances such an approach would be more beneficial than merely forecasting $\sigma_t^2$ from earlier years' observed intra–sample variance.

# 6 Appendix

## 6.1 Derivation of an approximate distribution for $\tilde{X}$

In order to derive a second-order Taylor series approximation for the distribution of $\tilde{X}$ *which is necessary for approximating the* $Bias(\tilde{X})$ *term*, and with the approximated $Bias(\tilde{X})$ carry on and compute approximations of $Bias(\hat{\beta})$ and $Var(\hat{\beta})$, we shall begin by factorizing our log-normal variable $Y_{it}$ into its two constitutive factors. In section 3.1.1 we have formulated a model that allows for both a *group effect* in variance at the log-scale ($b_t \sim N(0, s^2)$) and an *individual effect* for the variance between individuals drawn from the same sample ($\alpha + \beta(t - \bar{t}) + \epsilon_{it} \sim N(\alpha + \beta(t - \bar{t}), \sigma^2)$). We shall first derive expressions for the variances of the two log-normal variables of interest, and thereafter combine these expressions to find $Var(\bar{Y}_t)$.

The reason we are interested in $\bar{Y}_t$ is because we can interpret it as the distribution for the concentration of a *perfectly mixed* homogenate of $k$ individual specimen (assuming that the homogenate is made from equal amounts of tissue from each individual). Under the assumption that all homogenates have been created in this way, $\bar{Y}_t$ describes the distribution of the concentration measured in one such homogenate created from a sample of $k$ specimen. In those years and at those places where Method II is used, there are $m$ samples of $k$ specimen and therefore $m$ reported *homogenate concentrations*, and the average of the $m$ *log-homogenate concentrations* is regarded as an observation of $\tilde{X}$. It is the observed values of $\tilde{X}$ that occur in the log-linear regression models for years when Method II was used, and the following tedious calculations will eventually let us derive a second-order Taylor series approximation for the distribution of $\tilde{X}$.

### 6.1.1 Derivation of the $Var(\bar{Y}_t)$

Let $V_t = e^{b_t}$ denote the log-normal factor that accounts for randomness in sample composition for year $t$ and let $Z_{it} = e^{\tilde{\alpha}+\beta(t-\bar{t})+\epsilon_{it}}$ denote the factor that accounts for the individual variation in the same year. With these two factors we may now write $Y_{it} = V_t \cdot Z_{it}$, and further we can write $\bar{Y}_t = V_t \cdot \bar{Z}_t$. If one conditions on $V_t$, one can see that

$$E[Y_{it}|V_t = e^C] = e^C \cdot E[Z_{it}] = e^C \cdot e^{\tilde{\alpha}+\beta(t-\bar{t})+\frac{1}{2}\sigma^2},$$

and since we have that $E[\bar{Y}_i] = E[Y_{it}]$ we now get

$$E[\bar{Y}_i|V_t = e^C] = e^C \cdot E[\bar{Z}_t] = e^C \cdot e^{\tilde{\alpha}+\beta(t-\bar{t})+\frac{1}{2}\sigma^2}.$$

For the variances the following holds:

$$Var(Y_{it}|V_t = e^C) = e^{2C} \cdot Var(Z_{it}) = e^{2C} \cdot (e^{\sigma^2} - 1) \cdot e^{2(\tilde{\alpha}+\beta(t-\bar{t}))+\sigma^2}$$

$$Var(\bar{Y}_t|V_t = e^C) = e^{2C} \cdot \frac{1}{k} Var(Z_{it}) = \frac{e^{2C}}{k} \cdot (e^{\sigma^2} - 1) \cdot e^{2(\tilde{\alpha}+\beta(t-\bar{t}))+\sigma^2}$$

We can now provide an expression for the variance of $\bar{Y}_t$. Because of the property that $\bar{Y}_t$ is a product of two independent variables, it follows that

$$Var(\bar{Y}_t) = Var(V_t \cdot \bar{Z}_t) = E[(V_t \cdot \bar{Z}_t)^2] - E[(V_t \cdot \bar{Z}_t)]^2,$$

is equivalent to stating that

$$\begin{aligned} Var(\bar{Y}_t) &= E[V_t^2 \cdot \bar{Z}_t^2] - (E[V_t] \cdot E[\bar{Z}_t])^2 \\ &= (Var(V_t) + E[V_t]^2)(\frac{1}{k} Var(Z_{it}) + E[Z_{it}]^2) - (E[V_t] \cdot E[\bar{Z}_t])^2. \end{aligned}$$

The reason that the two propositions about $Var(\bar{Y}_t)$ are equivalent is that $E[\bar{Y}_t] = E[V_t] \cdot E[\bar{Z}_t]$. From equation 6.1.1 we may conclude that whenever $k > 1$ and $s^2 = \sigma^2$ the variance of $\bar{Y}_t$ will to a larger extent be determined by the yearly variation factor $V_t$. Even though we do not yet have an expression for $Var(\tilde{X}_t)$, we can now make the prediction that at equally large increases of $s^2$ and $\sigma^2$ (for example an increase of $\delta\sigma^2 = \delta s^2 = 0.01$), the resulting increase of the yearly variation factor $V_t$ will make a larger contribution to the increase of $Var(\bar{Y}_t)$ than what $\bar{Y}_t$ contributes, and therefore increases in $s^2$ should also yield greater increases in $Var(\tilde{X}_t)$ than what equal increases in $\sigma^2$ yields. Thus we will expect there to be lower frequencies of Type I errors when $s^2 = C$ and $\sigma^2 = 0$ compared to when the situation is reversed (i.e. $\sigma^2 = C$ and $s^2 = 0$), if the size of $Bias(\hat{\beta})$ is approximately equal in the two scenarios. The last conclusion follows from the fact that the more our computer-generated measurements are scattered, the greater the estimated residual variance will become and hence make it less likely that the null hypothesis of $\beta_0 = 0$ is rejected.

### 6.1.2 Derivation of a second-order Taylor approximation for $\tilde{X}$

For both $\bar{V}_t$ and $\bar{Z}_t$ it is clear that the Delta method is applicable, and since both variables are independent from one another while it also holds that both are log-normals, it will suffice to provide general expressions for second-order approximation of the logarithm of a log-normal. For the sake of simplicity we will use the same notation as the one used by Held and Bové [5] (see page 357). Let $\bar{T}_n := \frac{\sum_{i=1}^{n} T_i}{n}$ denote the sample mean of $n$ i.i.d. log-normal random variables $T_1, T_2, ..., T_n$ with finite mean $E[T]$ and finite variance $Var(T)$. By the Central Limit Theorem we have that $\sqrt{n}(\bar{T} - E[T])$ converges in distribution towards a $N(0, Var(T))$-distribution, for $\bar{T}_n$ converges in probability to $E[T]$ and is thus a consistent estimator for the log-normal mean. Further, since $\ln(\cdot)$ is a continuously differentiable function for positive real numbers and the log-normal mean is guaranteed to be strictly positive, it follows that the Delta method is applicable. In order to gain more accurate approximations a second order Taylor expansion is used instead of merely first order expansion. Hence we have that

$$\ln(\bar{T}_n) \approx \ln(E[T]) + \frac{(\bar{T}_n - E[T])}{E[T]} - \frac{(\bar{T}_n - E[T])^2}{2E[T]^2}$$

and taking the expectation on both sides yields (after some algebra)

$$\begin{aligned} E[\ln(\bar{T}_n)] &\approx \mu + \frac{\sigma^2}{2} - \frac{Var(\bar{T}_n)}{2E[T]^2} \\ &= \mu + \frac{\sigma^2}{2} - \frac{(e^{\sigma^2} - 1)}{2 \cdot n} \end{aligned}$$

Regarding the variance, derivation of an approximation is made easier by first introducing the following notation:

$$\begin{aligned} \tau &:= \frac{\sqrt{n}(\bar{T}_n - E[T])}{Var(T)} \\ a &:= \frac{Var(T)^2}{2 \cdot n \cdot E[T]^2} \\ b &:= \frac{\sqrt{Var(T)}}{\sqrt{n} \cdot E[T]} \end{aligned}$$

By the Central Limit Theorem, we have that $\tau \xrightarrow{a} N(0,1)$, and so it follows that $\tau^2 \xrightarrow{a} \chi_1^2$ and $b \cdot \tau \xrightarrow{a} N(0, \frac{Var(T)}{n \cdot E[T]^2})$. Hence, the following approximation of $Var(\ln(\bar{T}_n))$ arises:

$$\begin{aligned} Var[\ln(\bar{T}_n)] &\approx Var(\frac{\bar{T}_n - E[T]}{E[T]} - \frac{(\bar{T}_n - E[T])^2}{2E[T]^2}) \\ &= Var(b \cdot \tau - a \cdot \tau^2) \\ &= Var(b \cdot \tau) + Var(a \cdot \tau^2) - 2 \cdot Cov(b \cdot \tau, a \cdot \tau^2) \\ &= b^2 \cdot Var(\cdot \tau) + a^2 \cdot Var(\tau^2) - 2 \cdot (E[ab \cdot \tau^3] - E[b\tau] \cdot E[a\tau^2]) \\ &= \frac{Var(T)}{n \cdot E[T]^2} + \frac{Var(T)^4}{2 \cdot n^2 \cdot E[T]^4} \end{aligned} \tag{16}$$

The covariance term in equation is 16 is zero-valued, which follows from the fact that i) $E[\tau^3] = 0$ due to $\tau$ being a zero-centred normal RV, and ii) that $E[b\tau] \cdot E[a\tau^2] = 0$ since we have that $E[b\tau] = 0$.

Let us now consider forming a second-order Taylor approximation for $\tilde{X}_t$ at a given setting of $m$, $k$, $\alpha$, $\beta$, $s^2$ and $\sigma^2$. Since we have that $\bar{Y}_{.st} := V_{st} \cdot \bar{Z}_{.st}$ with the two variables assumed independent, and $\tilde{X}_t := \frac{1}{m} \sum_{s=1}^{m} \ln(\bar{Y}_{.st})$ one can approximate the distribution of $\tilde{X}_t$ (which can be re-written as $\frac{\sum_{s=1}^{m}(\ln(V_{st})+\ln(\bar{Y}_{.st}))}{m}$) as *a normally distributed sample mean* of the sum of the two second-order Taylor approximations for $\ln(V_{st})$ and $\ln(\bar{Z}_{.st})$ respectively. The resulting approximation when this method is applied is the following:

$$\tilde{X}_t \approx N\left(\mu_t + \frac{s^2+\sigma^2}{2} - \frac{e^{s^2}-1}{2} - \frac{e^{\sigma^2}-1}{2k}, \frac{Var(\ln(\bar{Z}_k))}{m} + \frac{Var(\ln(V))}{m}\right) \tag{17}$$

This relies on the assumption that the two Taylor approximations (which both are assumed to be normally distributed by construction) are decently good at describing the behavior of the two empirical sample log-transformed distributions, as is usually the case when applying the Delta method. Whether or not this assumption is deemed to be too strong for the problem at hand will be determined by the rate of convergence for the underlying log-normal distributions given the constraints on sample size, but this question is beyond the scope of the thesis.

## 6.2 Current methods employed in trend monitoring by the SNMPCMB

*In the appendix of the article by Soerensen and Faxneld* [10] *current diagnostic checks of the monitoring program are presented and compared to alternatives. As a means for assessing the strengths and weaknesses of the conclusions reached by this thesis regarding the risk for Type I error, a summarising account for these diagnostic checks and their usefulness has been included in this appendix.*

The program fits linear regression models for each combination of specie, location and contaminant and does so for both the full time span (from the current year until the first year when that combination was recorded) as well as for the short-term period (the latest ten years recorded). The default choice of a linear model is motivated by the simplicity of interpretation and as a means to avoid overfitting - but the assumption of linearity is yet tested against a larger model that includes non-linear components.

The larger model is a so called *Locally Estimated Scatterplot Smoothing* (LOESS) model, which is estimated by forming a local polynomial for each data point based on a subset of other points that are deemed to reside "close" to the current point of focus, and then combining these locally weighed polynomial models to a non-linear model for the entire period. Thereafter, an ANOVA-test compares the linear model with the non-linear LOESS model to check if the reduction of variance resulting from choosing the latter model over the former is statistically significant. However, the program only reports non-linear models when the null hypothesis has been rejected, and does not fit non-linear models for the short-term trends.

A second model diagnostic that is *mentioned (but not reported in the results)* in the appendix for the monitoring program's documentation is a *non-parametric test* in the form of the Mann-Kendall trend test. The motivation behind why one should consider applying the test is that linear regression parameter estimates are sensitive to leverage points, and this could lead us into overestimating the magnitude of a trend when observations near the edges of the time series randomly deviate sufficiently (but not necessarily *in an extreme manner*) in the direction indicated by the underlying trend $\beta$ from their expectation $\mu_t$. The Mann-Kendall (MK) statistic is formed by first deciding the sign (which can be 1, 0 or -1) of each possible difference between pairs of observations $(\bar{x}_j - \bar{x}_k)$, where the observations are ordered chronologically in time (such that $j > k$), and then one counts the difference between the total number of positive signs and the total number of negative signs. A quick overview of the MK statistic can be found at `https://vsp.pnnl.gov/help/vsample/design_trend_mann_kendall.htm`, and since it effectively only looks at the proportion of positive signs to negative signs (and ignores the magnitude of the differences) this trend test will be much more conservative than corresponding trend tests of the log-linear model. However as mentioned before, the program does not present any outcomes from MK tests, so therefore it seems like the authors have concluded that the inherent limitations of the study design makes the Mann-kendall test too conservative to be practically useful in evaluating goodness–of–fit — where these limitations are attributable to the circumstance that the datasets are rather small in relation to the true underlying coefficients of variation (assuming that the data generating process is log-linear).

Even though the MK-statistic is not reported by the program, another non–parametric test is reported instead, namely the absolute value of Kendall´s Tau. Kendall's Tau is defined as

$$\tau = \frac{\text{"Number of concordant pairs"} - \text{"Number of discordant pairs"}}{\text{"Total number of pairs"}}$$

where a pair of observations $(x_i, y_i)$ and $(x_j, y_j)$ are *concordant* if either both inequalities $x_i > x_j$ and $y_i > y_j$ are jointly true or equivalently when both $x_i < x_j$ and $y_i < y_j$ are jointly true (and *discordant* otherwise). Hence, the absolute value of Kendall's Tau ranges from 0 (no relationship) to 1 (perfect correlation). As explained by Soerensen and Faxneld in the aforementioned appendix of their report, this statistic can be viewed as analogous to the traditional correlation coefficient (Pearson's $\rho$), but its numerical values indicating (in the words of the authors) *somewhat strong linear correlations* are lower than that of Pearson's $\rho$. They mention a rule of thumb that suggests that a $\tau$ of 0.7 should be interpreted as a $\rho$ of 0.9.

In order to evaluate the effect of policy changes (e.g. bans on a particular chemical) on log-linear trends, the program sometimes apply a method called *change point detection*. The method consists of forming two models, A and B, where A allows for different slopes at two adjacent subsets of the time series and B is the usual linear model with a single slope for the entire set. At each step, the method proceeds to picking a new change point that defines the two subsets of Model A (requiring that each set includes atleast four points) and thereafter performs a likelihood ratio test for the null hypothesis that Model B holds while Model A does not hold.

Currently, the program constructs confidence intervals for its trend estimates by estimating covariance matrices that are robust against heteroskedasticity and auto-correlation, so called HAC consistent covariance matrices. The authors refer to an article by Zeileis (2004)[12] where this topic is introduced for the applied statistician and one is shown how an R-package called *sandwich* can be used to aquire such HAC robust estimates. Unfortunately, it is unclear from the program's appendix exactly which HAC consistent estimator is chosen (more specifically the details of how the weights are computed are not explained).

# 7 References

## References

[1] Anders Bignert et al. "Consequences of using pooled versus individual samples for designing environmental monitoring sampling strategies". In: *Chemosphere* 94 (2014), pp. 177–182.

[2] Samuel P Caudill, Wayman E Turner, and Donald G Patterson Jr. "Geometric mean estimation from pooled samples". In: *Chemosphere* 69.3 (2007), pp. 371–380.

[3] Daniel Dufresne. "Sums of lognormals". In.

[4] Allan Gut. "Transforms". In: *An Intermediate Course in Probability.* New York, NY: Springer New York, 2009, pp. 57–99. ISBN: 978-1-4419-0162-0. DOI: 10.1007/978-1-4419-0162-0_3. URL: https://doi.org/10.1007/978-1-4419-0162-0_3.

[5] Leonhard Held and Daniel Sabanés Bové. *Likelihood and Bayesian Inference.* Springer, 2020.

[6] P Holgate. "The lognormal characteristic function". In: *Communications in Statistics-Theory and Methods* 18.12 (1989), pp. 4539–4548.

[7] Eckhard Limpert, Werner A Stahel, and Markus Abbt. "Log-normal distributions across the sciences: keys and clues: on the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—normal or log-normal: that is the question". In: *BioScience* 51.5 (2001), pp. 341–352.

[8] Richard L Mitchell. "Permanence of the log-normal distribution". In: *JOSA* 58.9 (1968), pp. 1267–1272.

[9] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria, 2023. URL: https://www.R-project.org/.

[10] Anne L Soerensen and Suzanne Faxneld. *The Swedish National Monitoring Programme for Contaminants in Marine Biota (until 2019 year's data)-Temporal trends and spatial variations.* 2020.

[11] Rolf Sundberg. *Kompendium i Lineära Statistiska Modeller.* Matematiska Institutionen vid Stockholms Universitet, 2022.

[12] Achim Zeileis. "Econometric computing with HC and HAC covariance matrix estimators". In: (2004).