Stockholms universitet

Linjär regressionsanalys av miljögifterna i marin biota vid tillämpning av en blandad modell

Liming Huang

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2021:21
Matematisk statistik
September 2021

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2021:21**
http://www.math.su.se/matstat

Linjär regressionsanalys av miljögifterna i marin biota vid tillämpning av en blandad modell

Liming Huang*

September 2021

Sammanfattning

I denna studie är vi intresserad av att analysera koncentration av miljögifter, vid olika platser längs kusten i Sverige, genom att tillämpa linjära blandade modeller med MLE-metoden. Övervakningsprogrammet som genomförs i Sverige analyserar idag varje plats separat med en enkel linjär regressionsmodell. Datasetet är dock heteroskedastisk då residualvariansen varierar mellan platser och observationerna som erhålls från en plats kan vara beroende sinsemellan. Vi kommer att undersöka tre linjära blandade modeller med fixerade och/eller slumpmässiga effekter där man tar hänsyn till variation mellan platserna. Genom att illustrera de skattade parametrarna får vi att variationen mellan platserna är faktisk signifikant. De skattade trenderna för olika nivåer är i stort sett negativa. Det finns ett positivt proportionellt samband mellan de slumpmässiga intercepten och de slumpmässiga trenderna. Vi hittar ingen rumslig trend i koncentration.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: alicehbb@gmail.com. Handledare: Martin Sköld.

Förord

Jag vill uttrycka min enorma tacksamhet till min handledare Martin Sköld för hans vänligaste och mest tålmodiga vägledning. Dessutom vill jag tacka min vän Jie Wen som jag har haft många fantastiska och värdefulla diskussioner med och såväl som Peter Yuanzong Li för hans hjälp och stöd.

Innehåll

1 Introduktion **5**

1.1 Bakgrund 5

1.2 Databeskrivning 5

1.2.1 Databehandling 7

2 Teori **7**

2.1 Linjär regressionsmodell 7

2.1.1 Enkel linjär regressionsmodell 8

2.1.2 Linjär blandad modell - LMM 8

2.1.3 Heteroskedasticitet 9

2.2 Modellanpassning 9

2.2.1 Maximum Likelihood Estimation - MLE 9

2.2.2 MLE för den blandade modellen 10

2.2.3 Akaike's Information Criterion - AIC 11

2.2.4 Summering av AIC-värden 12

2.3 Programvara och paket 12

3 Modellering och analys **12**

3.1 Nuvarande metod 13

3.2 Kandidatmodell 13

3.3 Modell analys 15

3.3.1 Modell 1 15

3.3.2 Modell 2 16

3.3.3 Modell 3 18

3.3.4 Modell 4 20

4 Resultat **22**

5 Diskussion **23**

Bilaga **25**

Referenser **37**

1 Introduktion

1.1 Bakgrund

Med tiden börjar människor gradvis uppmärksamma vikten av miljöföroreningar. Avdelning för miljöforskning och övervakning vid Naturhistoriska Riksmuseet i Sverige genomför ett övervakningsprogram i ett gemensamt arbete med Institutionen för miljövetenskap och analytisk kemi (ACES) vid Stockholms universitet, Institutionen för kemi vid Umeå universitet och IVL Svenska Miljöinstitutet. Övervakningsprogrammet finansieras av Naturvårdsverket. Programmets huvudsakliga syfte är att genomföra övervakning av miljögifter i marin biota och analysera detta.

För nuvarande analysen av miljögifterna använder programmet en enkel linjär regressionsmodell (modell 1) för att detektera tidstrender vid varje provtagningsplats separat. Analysen bygger på geometriska medelvärden årsvis och variationen mellan provtagningsplatser tas inte hänsyn till. Om vi exempelvis intresserar oss för trend över de senast 10 åren har vi bara 10 datapunkter att basera analysen på. Modellen leder också till heteroskedasticitet hos feltermen, då residualvarians för olika platser är ej konstant. Observationerna uppfyller inte oberoende antagande för vanlig linjär regression på grund av observationer som kommer från samma plats är egentligen korrelerade. Ett sätt att förbättra eventuella skattningar i analysen är att ta med flera platser i samma modell genom att använda linjär blandad modell. Därmed antas exempelvis att residualvariansen eller trenden inte skiljer sig så mycket för ingående platser.

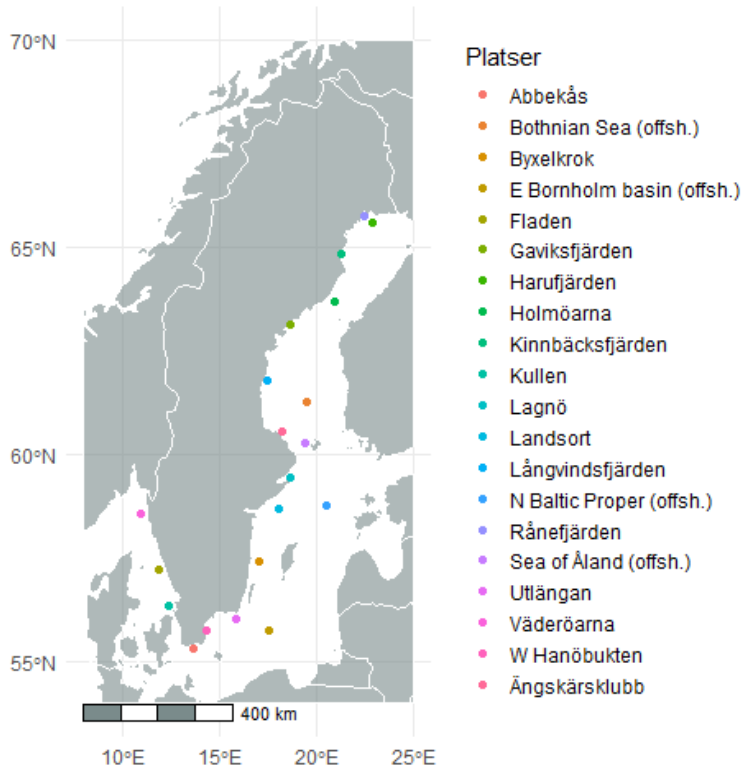
I denna fallstudie undersöker vi huvudsakligen följande tre linjära blandade modeller med MLE-metoden: modell 2 med både intercept och trend slumpmässiga; modell 3 med slumpmässigt intercept och fixerad trend samt modell 4 med fixerat intercept och slumpmässig trend, då man tar hänsyn till variation mellan provtagningsplatserna som anses vara en slumpmässig effekt. Genom att basera analys på dessa modeller samt visualisering av dessa skattade parametrar förväntar vi oss att förstå exempelvis hur parameterskattningar skiljer sig inom och även mellan modellerna. Dessutom finns det något samband mellan intercept och trend och med mera. På så sätt får en bild av hur trender och nivåer för miljögifterna fördelar sig i landet.

1.2 Databeskrivning

Datamaterialet som används i denna uppsats förvärvades av Martin Sköld från Naturhistoriska Riksmuseet. Data mäts årligen ett antal individer på olika geografiska platser, framför allt längs kusten i hela Sverige. Provtagningsplatserna är belägna i områden som betraktas som icke kontaminerad och opåverkad av större flodutlopp så långt som möjligt och även inte för nära tätbefolkade områden. Design för provtagning ändras över tiden. För nuvarande genomförs 10 – 12 enskilda mätningar per art från var och en av de platserna vid Östersjön och västkusten. Det är framför allt mätningar från

de subadulta som samlats in. För många arter representerar de subadulta en nyare bild av föroreningsbelastning än de vuxna på grund av bioackumulering av många föroreningar. Samtidigt för att undvika eventuella influenser mellan variansen på grund av könsskillnader, är huvudsakligen honfisk som samlats in och analys görs baserad på samma köns typ. För mer information och detalj om dataprovtagningen hänvisar till årsrapport från Bignert et al. (2017)[1]. I figur 1 visas de geografiska lägena för provtagningsplatser för strömming i Sverige.

Den uppmätta koncentrationen av ett miljögift i en individ antas ofta vara log-normalfördelat. Vid vissa platser mäter man i stället koncentrationer i homogenat för att hålla analyskostnaderna nere. Med homogenat innebär det att blanda ett antal, i allmänhet 5 - 12, individer. Koncentrationen i ett homogenat kan antas vara medelvärdet av koncentrationen i de mixade individerna.



Figur 1: Provtagningsplatserna för strömming i Sverige.

1.2.1 Databehandling

Mätningarna som noterat med 'offsh.', som står för 'offshore', och 'spring'. Dessa innebär bland annat att mätningar görs vid plasterna som inte ligger vid kusten, respektive mätningarna som görs på våren. Normalt görs mätningarna på hösten. I denna studie bortser vi från mätningar som görs på våren medan sparar 'offsh.' som övriga. Årtalen i datamaterialen behandlas från kalenderår till siffror med storleksordning, det vill säga 1 till 12 då vårt datamaterial täcker 12 års detektering (2007 - 2018).

Det finns antalet 1088 uppmätt observationer i 20 provtagningsplatser hos strömmingar. I Tabell 2 visar exempel på data. Denna studie baserar dock analysen på de geometriska medelvärden årsvis, det vill säga en datapunkt per år per plats. Således blir det 213 antal observationer i datamängden.

2 Teori

Detta avsnitt framgår en genomgång av de teoretiska delarna som tillämpas i uppsatsen. Här nämner även lite om det programspråket som har använt.

2.1 Linjär regressionsmodell

Anta ett dataset som involverar i kluster av observationer, var och en med n_i observationer där $i = 1, ..., m$. Observationerna antas vara oberoende av varandra, detta är nödvändigt för att uppskattningarna ska vara konsekventa och för att de ska ha bra fördelningsegenskaper. Oberoende antagandet tyder på att modeller har endast fixerade effekter vilket i sin tur innebär att alla observationer behandlas som om de är på samma hierarkiska nivå.

En linjär regressionsmodell kan uttryckas som

$$Y = X\beta + \epsilon, \tag{1}$$

där

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, \quad X = \begin{bmatrix} 1_{n_1} & x_{1,1} & \cdots & x_{1,p-1} \\ 1_{n_2} & x_{2,1} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1_{n_m} & x_{m,1} & \cdots & x_{m,p-1} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{bmatrix}.$$

Där responsvariabel Y är $(n \times 1)$ vektor av observationer med $n = \sum_{i=1}^{m} n_i$; X en $(n \times p)$ designmatris med p antalet förklarande variabler och 1_{n_i} är vektor av ettor med längd n_i; och β en $(p \times 1)$ vektor som består av modellparametrar som relaterar till Y;

felterm ϵ en $(n \times 1)$ vektor som står för de slumpmässiga variationerna och antas i vanliga fall oberoende sinsemellan och normalfördelade med väntevärde 0 och konstant standardavvikelse σ, $N(0, \sigma^2)$. Antagandet är bra och kan på så sätt lätt erhålla statistiska tester på modellparametrarna, konfidensintervall på parametrarna och konfidens- och prediktionsintervall för medelrespons. (Raymond H. Myers et al. 2010[5])

2.1.1 Enkel linjär regressionsmodell

En enkel linjär regressionsmodell är då man har en responsvariabel Y och endast en förklarande variabel X. I så fall är $p = 2$ och designmatrisen X kommer att bestå av två kolumner, ettor och $x_{i,1}$; och β inkluderar ofta en intercept- och trendparameter, det vill säga β_0 och β_1.

2.1.2 Linjär blandad modell - LMM

Det är dock mycket vanligt att data är grupperade i den meningen att observationer inte är oberoende längre eftersom de till exempel uppmättes i samma kluster eller är grupperade efter någon annan egenskap. I sådana fall anses observationerna hierarkiska.

Till skillnad från en linjär regressionsmodell, tillåter en linjär blandad modell (*Linear mixed model* - LMM) både fixerade- och slumpmässiga effekter representerade. Att lägga till slumpmässiga effekter betyder att integrera hierarkiska strukturer i modeller. En linjär blandad modell kan ha ett separat intercept och/eller trend för varje hierarkisk nivå.

En linjär blandad modell kan uttryckas, i matrisform, som

$$Y = X\beta + Z\gamma + \epsilon, \tag{2}$$

där

$$Z = \begin{bmatrix} z_1 & 0_{n_1 \times q} & \cdots & 0_{n_1 \times q} \\ 0_{n_2 \times q} & z_2 & \cdots & 0_{n_2 \times q} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{n_m \times q} & 0_{n_m \times q} & \cdots & z_m \end{bmatrix} \quad \text{och} \quad \gamma = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_{mq} \end{bmatrix}.$$

Här är Y, X, β och ϵ som definierade i (1) och Z är en $(n \times mq)$ diagonalmatris av förklarande variabler för de q slumpmässiga effekterna; z_i och nollmatriser $0_{n_i \times q}$ har dimension $(n_i \times q)$; och γ en $(mq \times 1)$ vektor som består av m nivåer för var och en de slumpmässiga effekterna. Vanligtvis, under *Gaussiska linjära blandade modeller*, antar man $\epsilon \sim N(0, R)$, $\gamma \sim N(0, G)$, och $\mathrm{Cov}(\epsilon, \gamma) = 0$, det vill säga att de är normalfördelade och oberoende av varandra. Där $R = \sigma_\epsilon^2 I$ är $(n \times n)$ kovariansmatris med enhetsmatris I. Om man antar att slumpmässiga effektkoefficienterna är oberoende sinsemellan är

kovariansmatris G en enkel diagonalmatris med diagonala element lika med variansen av de slumpmässiga effektkoefficienterna. Med andra ord är $G = \sigma_\gamma^2 I$ en $(mq \times mq)$ kovariansmatris. (Raymond H. Myers et al. 2012 [5])

2.1.3 Heteroskedasticitet

Om felterm är heteroskedastisk innebär att variansen för feltermen skiljer sig mellan olika värde på x. Det finns många anledningar till att heteroskedasticitet existerar. En vanlig förklaring är att residualvarians förändras proportionellt med en faktor, i så fall kan denna faktor gissningsvis vara en variabel i modellen.

2.2 Modellanpassning

2.2.1 Maximum Likelihood Estimation - MLE

Maximum likelihood-metod (*Maximum likelihood estimation*) är en användbar metod som används för att härleda parameterskattningar. Låt $Y = \mathbf{y}$ vara en stokastisk variabel med sannolikhetsfunktion eller täthetsfunktion $f(\mathbf{y};\theta)$ beror på om variabeln är diskret eller kontinuerligt. Där θ är de okända parametrar som indexerar sannolikhetsfördelningen som kan ta värden i parameterrum Θ. I den här metoden spelar *likelihoodfunktionen* en stor roll och kan definieras som

$$L(\theta;\mathbf{y}) = f(\mathbf{y};\theta), \quad \theta \in \Theta.$$

Alltså är likelihoodfunktionen täthetsfunktion eller sannolikhetsfunktion för observerade data y, uttryckas som en funktion av de okända parametrarna. Den maximum likelihood-skattning av θ är det θ-värde som maximerar likelihoodfunktionen, $L(\theta)$, vilket kan exempelvis betecknas som $\hat{\theta}_{MLE}$. Definitionen kan formuleras som

$$\hat{\theta}_{\text{MLE}} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, L(\theta).$$

I praktiken är det ofta bekvämt att arbeta med den naturliga logaritmen för likelihood-funktionen, vilket kallas för *log-likelihoodfunktion*

$$l(\theta) = \ln\left(L(\theta)\right).$$

Eftersom den naturliga logaritmen är en monoton funktion uppträder det maximalt värde av $l(\theta)$ samma värde som det maximalt värde av $L(\theta)$ och ger uttryck

$$\hat{\theta}_{\text{MLE}} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, l(\theta).$$

Likelihoodfunktion för en normallfördelning $N(\mu, \sigma^2)$ för n oberoende och likafördelade normal slumpmässiga variabler ges

$$L(\mu, \sigma^2; y_1, ..., y_n) = f(y_1, ..., y_n; \mu, \sigma^2) =$$

$$= \prod_{i=1}^{n} f(y_i; \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right\}.$$

Log-likelihoodfunktion ges

$$l(\mu, \sigma^2) = \ln\left(L(\mu, \sigma^2; \mathbf{y})\right) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Genom att sätta derivatan[1] av log-likelihoodfunktionen lika med 0 får man de två skattade parametrarna som maximerar log-likelihoodfunktionen, det vill säga $\hat{\theta}_{\text{MLE}}$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x},$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Notera att den MLE under normalantagande för feltermen är identisk med OLS[2]-skattaren för modellen. Med andra ord OLS-skattaren maximerar MLE för modellen.

2.2.2 MLE för den blandade modellen

Låt $Y = \mathbf{y}$ vara stokastiska variabler i en blandad modell med täthetsfunktion eller sannolikhetsfunktion $f(\mathbf{y}; \beta, \theta)$, där β är en vektor av parametrar för de fixerade effekterna och θ är vektor av parametrar för de slumpmässiga effekterna från kovariansmatriserna G och R definierade i avsnitt 2.1.2. Av samma anledning som tidigare illustrerar vi här bara log-likelihoodfunktion som

$$l(\beta, \theta) = \ln(L(\beta, \theta; \mathbf{y})) = \ln(f(\mathbf{y}; \beta, \theta)).$$

Med MLE ger parameteruppskattningarna av

$$(\hat{\beta}, \hat{\theta})_{\text{MLE}} = \underset{(\beta,\theta)}{\operatorname{argmax}}\, l(\beta, \theta).$$

Formulering av en kovariansmatris V i följande som beskriver kovarianser mellan två observationer i datasetet

$$\begin{aligned}
V &= \text{var}(\mathbf{y}) = \text{var}\left(X\beta + Z\gamma + \epsilon\right) \\
&= \text{var}\left(X\beta\right) + \text{var}\left(Z\gamma\right) + \text{var}(\epsilon) \quad (\text{oberoende för } \mathbf{y}) \\
&= \text{var}\left(Z\gamma\right) + \text{var}(\epsilon) \quad (\text{var}(\text{fixerade effekter}) = 0) \\
&= Z\text{var}\left(\gamma\right)Z' + \text{var}(\epsilon) \quad (Z \text{ är konstant}) \\
&= ZGZ' + R.
\end{aligned}$$

[1] $\frac{\partial}{\partial\mu} l(\mu, \sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0$ och $\frac{\partial}{\partial\sigma^2} l(\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0.$

[2] *The Oridinary Least Squares.* Kallas för den minstakvadratmetod på svenska, används vid bland annat regressionsanalys för att hitta den mest anpassande regressionslinje funktion genom att minimera summan av fel i kvadraten.

Fördelning för $\mathbf{y}$ ges då

$$\mathbf{y} \sim N\left(X\beta, V(\theta)\right).$$

Likelihoodfunktion[3] ges

$$L(\beta, \theta; \mathbf{y}) = f(\mathbf{y}; \beta, \theta) =$$
$$= \frac{1}{(2\pi)^{n/2}|V|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y} - X\beta)'V^{-1}(\mathbf{y} - X\beta)\right\},$$

och log-likelihoodfunktion

$$l(\beta, \theta) = -\frac{1}{2}\left\{n\ln(2\pi) + \ln(|V|) + (\mathbf{y} - X\beta)'V^{-1}(\mathbf{y} - X\beta)\right\}.$$

Genom att sätta derivatan[4] av log-likelihoodfunktionen lika med 0 får man de två skattade parametrarna som maximerar log-likelihoodfunktionen, det vill säga $(\hat{\beta}, \hat{\theta})_{\text{MLE}}$

$$\hat{\beta} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}\mathbf{y}, \tag{3}$$

där $\hat{V} = V(\hat{\theta})$, det vill säga V består av MLE av dess varianskomponenterna. Således så fort man hittar $\hat{\theta}_{\text{MLE}}$ får man $\hat{\beta}_{\text{MLE}}$ vilket kan beräknas ut med ekvationen(3).

2.2.3 Akaike's Information Criterion - AIC

Akaikes informationskriterium (*Akaike's information criterion*) är ett mått som används i stor utsträckning för att uppskatta den så kallade *goodness of fit* av modeller. Goodness of fit av en statistisk modell beskriver hur väl den passar en uppsättning observationer, och dessa uppskattas med MLE. Kriteriet hanterar både risk för överanpassning och underanpassning. AIC definieras som följande

$$\text{AIC} = 2k - \ln\left(\hat{L}\right),$$

där k är antalet uppskattade parametrar i modellen och $\hat{L}$ är den maximerade likelihoodfunktionen för modellen. AIC är negativt orienterat vilket innebär modell med lägst AIC-värde ska väljas. Termen $2k$ är en ökande funktion fungerar som ett straff som avskräcker överanpassning med för många parametrar i modellen. Om två modeller har liknande AIC-värden, kan man i de flesta fall välja den enklare modellen, det vill säga modellen med mindre antal faktorer. (Held L. 2014, 7.1.1[3])

[3] för mer detaljer se Jiming J. 2007, 1.3.1[4]

[4] $\frac{\partial}{\partial\beta}l(\beta,\theta) = X'V^{-1}\mathbf{y} - X'V^{-1}X\beta = 0$ och se Appendix B.2 i Jiming J. 2007, [4] ger $\frac{\partial}{\partial\theta_r}l(\beta,\theta) = \frac{1}{2}\{(\mathbf{y} - X\beta)'V^{-1}\frac{\partial V}{\partial\theta_r}V^{-1}(\mathbf{y} - X\beta) - tr(V^{-1}\frac{\partial V}{\partial\theta_r})\} = 0$ för $r = 1, ..., q$, där θ_r är den $r$te komponent av θ som har dimension q.

2.2.4 Summering av AIC-värden

Summering av AIC-värden av n linjära modeller med lika många parametrar kan ges formel som

$$\text{AIC}_{sum} = n2k - 2\sum_{i=1}^{n} \ln(\hat{L}_i),$$

med hjälp av matrisformeln för linjär regressionsmodell i 2.1. Då betraktas Y som en vektor av modeller Y_i, X en matris av designmatriser X_i, β en vektor av parametervektorer β_i och ϵ en vektor av ϵ_i för $i = 1, ..., n$.

2.3 Programvara och paket

Alla statistiska beräkningar, modellering, grafisk presentation av statistik och även databehandling genomfördes med hjälp av programmeringsspråk `R` (version 4.0.5) i den integrerade utvecklingsmiljön `RStudio` (version 1.3.1093). I uppsatsen användes funktion `lm` från paket `stat` men framför allt funktion `lmer` från paket `lme4`. Dessa två funktionerna kan enkelt passa linjära modeller respektive linjära blandade modeller och används för att exempelvis utföra regression, variansanalys och analys av kovarians.

3 Modellering och analys

I detta avsnitt beskriver vi den nuvarande metoden övervakningsprogrammet använder och ställer upp tre kandidatmodeller med användning av blandad modell samt analys som baserade på dessa.

Låt Y vara logaritm av de observerade koncentrationer av miljögifterna. För enkelhets skull bortser vi individuell variation från denna studie och analysen baseras endast på, med årliga geometriska medelvärden, mätningarna från strömmingar (herring). Strömmingsmätdata samlas in i totalt 20 provtagningsplatser som visas i Figur 1. Dessutom baserar vår analys på koncentrationen av det kemiska ämnet kvicksilver (HG), med anledning av den observerade kvicksilver har minst saknad värden i mätdata i jämförelse med de andra, vilket är lämpligt att undersöka. I inferensteori antar faktiskt att observationerna följer normalfördelning, dock finns det en hel del praktiska situationer som inte uppfyller inte detta antagande. Genom att illustrera ett histogram och en QQ-plot som visas i Figur 7 observerar vi att Y approximerar en normalfördelning. Då histogrammet följer den täthetsfunktion väl och Q-Q plot visar en linjär trend med väldigt få avvikande observationer.

3.1 Nuvarande metod

Den nuvarande metoden i övervakningsprogrammet analyserar varje provtagningsplats separat, för detektering av en viss årig trend passar en enkel linjär regressionsmodell

$$Y_i = X_i\beta_i + \epsilon_i, \quad i = 1, \ldots, 20.$$

Med definitionen i avsnitt 2.1 kan uttrycket skrivs i matrisform som

$$\begin{bmatrix} y_{i1} \\ \vdots \\ y_{ij} \end{bmatrix} = \begin{bmatrix} 1_{i1} & x_{i1} \\ \vdots & \vdots \\ 1_{ij} & x_{ij} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{ij} \end{bmatrix}, \quad j = 1, \ldots, k_i,$$

för varje plats i, vilka är oberoende av varandra. Här är Y_i en $(k_i \times 1)$ vektor av observationer, där k_i är antalet år för detektering och y_{ij} det aritmetiskt medelvärdet av individernas logkoncentrationer för detekteringsår j; X_i en $(k_i \times 2)$ matris; β_i en (2×1) vektor och feltermen $\epsilon_i \sim N(0, \sigma_i^2)$ en $(k_i \times 1)$ vektor i i-te plats. Feltermen är heteroskedastisk då residualvariansen inte är konstant för olika platser, se Figur 8 och 9. Notera att feltermen är oberoende sinsemellan. Vi formulerar följande **modell 1** med det aritmetiska medelvärdet av logkoncentrationer på plats i år j

$$y_{ij} = \beta_{i0} + \beta_{i1}x_j + \sigma_i\epsilon_{ij}, \quad i = 1, \ldots, 20, \quad j = 1, \ldots, k_i. \tag{4}$$

Sätt $x_j = j - k_i$ så att $x_{k_i} = 0$, vilket ger logkoncentration av HG är densamma som intercept det senaste året. På så sätt ger intercept β_{i0} som betecknar den logkoncentrationen för det senaste året.

Eftersom observationer som kommer från samma plats är korrelerade, uppfyller dessa observationer inte oberoende antagande för vanlig linjär regression. Vi använder således i nästa avsnitt linjär blandad modell för att ta med effekt av provtagningsplatserna.

3.2 Kandidatmodell

Vi listar nu ut tre kandidatmodeller vid tillämning av linjär blandad modell. En av dem har både slumpmässigt intercept och slumpmässig trend medan de andra två har den ena slumpmässigt intercept och fixerad trend, och tvärt om för den andra slumpmässig trend och fixerat intercept.

Antar nu att platserna skulle kunna vara ett slumpmässigt urval av platser i Sverige. På så sätt och med definitionen i avsnitt 2.1.2 kan vi uttrycka denna slumpmässighet i en gemensam modell som

$$Y = X\beta + Z\gamma + \epsilon,$$

där $Y = \{y_{ij}\}$ betecknar den log-koncentration för plats i år j, här $i = 1, \ldots, 20$ och $j = 1, \ldots, k_i$. Här har X, β och ϵ samma definition som den nuvarande modellen.

Vi sätter först upp **modell 2** då vi lägger till ett slumpmässigt intercept på nivå och antar också att förhållandet mellan koncentration av HG och tid är olika för olika nivåer. Matriser Z och γ för modell 2 med både parametrarna slumpmässiga

$$Z = \begin{bmatrix} 1_{n_1} & 0_{n_1} & \cdots & 0_{n_1} & x_1 & 0_{n_1} & \cdots & 0_{n_1} \\ 0_{n_2} & 1_{n_2} & \cdots & 0_{n_2} & 0_{n_2} & x_2 & \cdots & 0_{n_2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0_{n_{20}} & 0_{n_{20}} & \cdots & 1_{n_{20}} & 0_{n_{20}} & 0_{n_{20}} & \cdots & x_{20} \end{bmatrix} \quad \text{och} \quad \gamma = \begin{bmatrix} \gamma_{0,1} \\ \vdots \\ \gamma_{0,20} \\ \gamma_{1,1} \\ \vdots \\ \gamma_{1,20} \end{bmatrix}.$$

Omskrivning av modellen ger

$$y_{ij} = \beta_0 + \gamma_{0i} + (\beta_1 + \gamma_{1i})x_j + \sigma\epsilon_{ij}, \quad i = 1, \ldots, 20, \quad j = 1, \ldots, k_i, \tag{5}$$

där β_0 är ett gemensamt intercept, den nuvarande log-koncentrationen i Sverige, och $\gamma_{0i} \sim N(0, \sigma^2_{\gamma_0})$, så att $\beta_0 + \gamma_{0i}$ är den nuvarande log-koncentrationen i plats i. På motsvarande sätt betecknar β_1 trend för log-koncentration i Sverige och $\beta_1 + \gamma_{1i}$, där $\gamma_{1i} \sim N(0, \sigma^2_{\gamma_1})$, den nuvarande trenden i plats i. Parametrarna (β_0, β_1) kallas för de fixerade effekterna medan $(\gamma_{0i}, \gamma_{1i})$ de slumpmässiga effekterna. Alltså istället för att uppskatta $20 \cdot 2 = 40$ regressionskoefficienter till modellen lägger man bara till två variansparameter. Antag att residualvariansen σ^2 är konstant och lika vid alla platser. Feltermen $\epsilon_{ij} \sim N(0, 1)$ antas vara oberoende av de slumpmässiga effekterna.

Matrisen Z och γ i för sig kan dela upp på två delar som följande

$$Z = \begin{bmatrix} Z_1 & \vdots & Z_2 \end{bmatrix} \text{ och}, \quad \gamma = \begin{bmatrix} \gamma_1 \\ \cdots \\ \gamma_2 \end{bmatrix},$$

där

$$Z_1 = \begin{bmatrix} 1_{n_1} & 0_{n_1} & \cdots & 0_{n_1} \\ 0_{n_2} & 1_{n_2} & \cdots & 0_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{n_{20}} & 0_{n_{20}} & \cdots & 1_{n_{20}} \end{bmatrix} \quad , \quad Z_2 = \begin{bmatrix} x_1 & 0_{n_1} & \cdots & 0_{n_1} \\ 0_{n_2} & x_2 & \cdots & 0_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{n_{20}} & 0_{n_{20}} & \cdots & x_{20} \end{bmatrix}$$

och

$$\gamma_1 = \begin{bmatrix} \gamma_{0,1} \\ \vdots \\ \gamma_{0,20} \end{bmatrix} \quad , \quad \gamma_2 = \begin{bmatrix} \gamma_{1,1} \\ \vdots \\ \gamma_{1,20} \end{bmatrix}.$$

Formulering av **modell 3** med slumpmässigt intercept och fixerad trend ger

$$y_{ij} = \beta_0 + \gamma_{0i} + \beta_1 x_j + \sigma\epsilon_{ij}, \quad i = 1, \ldots, 20, \quad j = 1, \ldots, k_i, \tag{6}$$

med $Z = Z_1$ som endast består av ettor och nollor, och $\gamma = \gamma_1$.

Modell 4 med slumpmässig trend och fixerat intercept formuleras som

$$y_{ij} = \beta_0 + (\beta_1 + \gamma_{1i})x_j + \sigma\epsilon_{ij}, \quad i = 1, \ldots, 20, \quad j = 1, \ldots, k_i, \tag{7}$$

med $Z = Z_2$ och $\gamma = \gamma_2$.

3.3 Modell analys

Med hjälp av `lm`- och `lmer`-funktionen i `RStudio` anpassar vi en enkel linjär modell för varje plats och tre linjära blandade modeller enligt modellerna som nämndes tidigare. Modellerna är anpassade med metod MLE för både lm- och lmer-funktionen, vilket är en av de vanligaste metoderna som passar linjära modeller. Tabell 3 visar skattningsvärde för intercept och trend för modell 1 vid olika platser.

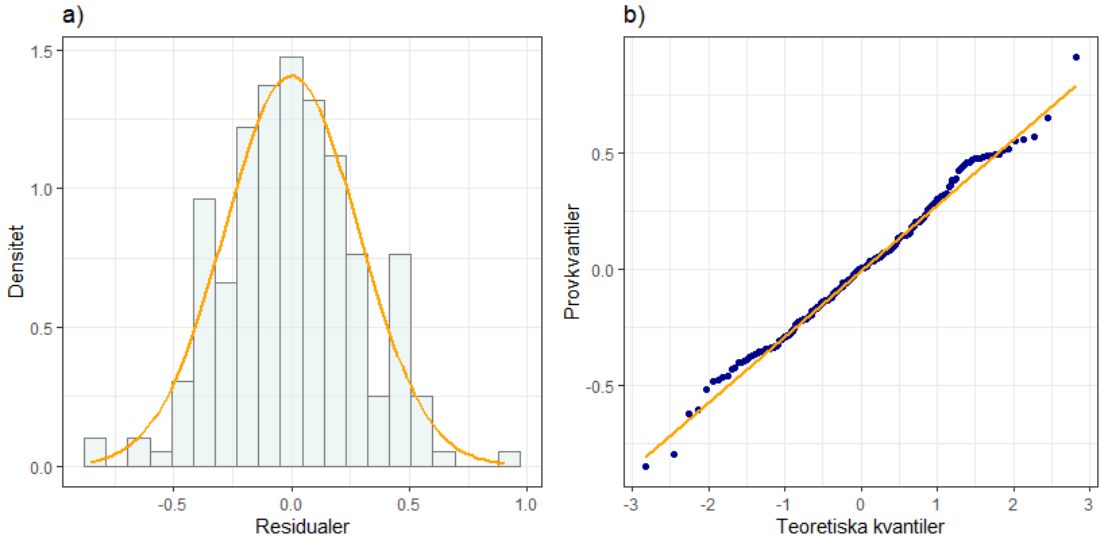
Figur 10 visar resultat av linjär blandade modellanalys av förhållandet mellan koncentration av HG och tid, med både ett slumpmässigt intercept och en trend för tid på platsnivå. Figur 11 visar resultat av modellanalys av förhållandet med ett slumpmässigt intercept på platsnivå samt Figur 12 visar resultat av modellanalys av förhållandet med en trend för tid på platsnivå. Ur dessa resultat integrerar vi skattningsvärden i Tabell 1 samt AIC-värden. Där AIC-värden för originalmodellen är summan av AIC-värden från enkel linjär regressionsmodell för varje plats. Vanligtvis är mindre AIC-värde bättre, men det finns inget absolut 'bra' värde. Resultaten ger dessutom information om att alla fixerade intercept fixerade trenderna är signifikanta i modellerna på en 5%-signifikansnivå. Notera att $\sigma^2_{\hat{\gamma}_0}$ är variansen mellan platserna och $\hat{\sigma}^2$ är variansen inom platserna samt $\sigma^2_{\hat{\gamma}_1}$ är variansen av regressionskoefficienterna för tid för de olika platserna.

Modell	AIC	$\hat{\beta}_0$	$\hat{\beta}_1$	$\sigma_{\hat{\gamma}_0}$	$\sigma_{\hat{\gamma}_1}$	$\hat{\sigma}$
1	128.79	——	——	——	——	——
2	161.19	3.3347	-0.0233	0.3445	0.0128	0.2991
3	157.80	3.3331	-0.0232	0.3626	——	0.3021
4	195.80	3.3435	-0.0242	——	0.0421	0.3370

Tabell 1: AIC- och skattningsvärden för parametrar för fyra modellerna: den originella enkel linjär regressionsmodell och de tre blandade modeller med olika egenskaper hos parametrarna.

3.3.1 Modell 1

Originalmodellen har ett relativt litet AIC-värde i jämförelse med alla andra. Detta kan bero på att vissa provtagningsplatser har väldigt högt värde för log-likelihoodfunktion. Den straffade termen $n2k$ i AIC blir $20 \cdot 2 \cdot 2 = 80$ då vi summerar 20 linjära modeller med 2 parametrar, vilken är uppenbarligen större än de andra kandidatmodeller. Trots



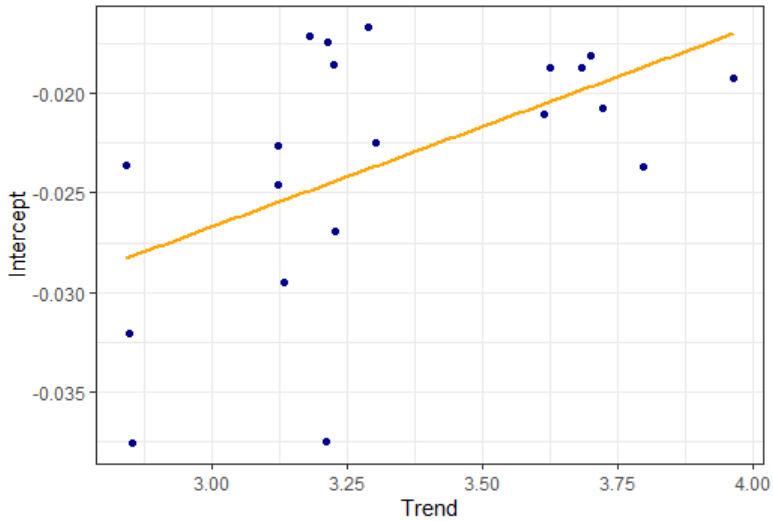
Figur 2: a) Histogram och normal kurvan för residualerna för modell 2 och b) Normal QQ-plot med residualerna från modell 2.

detta påverkar termen ändå väldigt lite AIC-värden för modellen på grund av effekten av den avtagande likelihoodfunktionstermen. I Tabell 3 visar AIC-värdena för varje plats. Vi observerar exempelvis att provtagningsplatsen E Bornholm basin har extremt litet AIC-värde med liten residualvarians som antyder dess höga log-likelihoodfunktionsvärden. Dess regressionslinje passerar alla fyra datapunkterna som illustreras i Figur 9 vilket resulterar ett väldigt litet OLS-värde.

3.3.2 Modell 2

Vi kontrollerar först våra antaganden, att residualerna från modellen följer en normalfördelning. Figur 2 illustrerar den teoretiska normalfördelningen i ett histogram och en Q-Q plot med de skattade residualerna som erhålls från modell 2. Histogrammet följer den täthetsfunktion (den orange linjen) väl och Q-Q plot visar en linjär trend med små avvikande svans på bägge ändarna men är i stort sett bra. Detta tyder på att det är rimligt att anta residualerna följer en normalfördelning. Sedan observerar vi figuren längst upp i Figur 13 som illustrerar två Q-Q plot för det slumpmässigt intercept respektive trend för modellen. Både visar en fin linjär trend för effekt-kvantilerna. Alltså är normalfördelad antagande för slumpmässiga effekterna rimlig i modell 2.

Figuren till vänster i Figur 14 visar de skattade regressionslinjerna för respektive provtagningsplats för modell 2, med andra ord funktioner av den förklarande variabeln y_{ij} med avseende på prediktorn x_j. Eftersom modellen är en slumpmässig intercept- och trendmodell är regressionslinjes intercept och trend annorlunda i olika platser. Vi kan intuitivt



Figur 3: Spridningsdiagram (scatter plot) för de skattade intercepten och trenderna från modell 2.

se att alla platserna har negativ uppskattade trend över tiden. Därmed i spridningsdiagrammet 3 visar samband mellan intercepten och trenderna för modellen. Sambandet säger i stort sett ju större intercept är desto större trend är. Vi kan exempelvis passa en linjär regressionslinje i diagrammet. Den orange linjen påvisar ett positivt proportionellt samband mellan intercepten och trenderna med en funktion

$$\text{Intercept} = -0.0568 + 0.0100 \cdot \text{trend}.$$

Undersöker närmare med de slumpmässiga effekterna illustrerar man Figur 15 som visualiserar variationer för slumpmässigt intercept och trend. Denna figur är identisk med Figur 13, vid kontroll av normalitet av de slumpmässiga effekterna, men har variationerna på x-axeln och nivåerna på y-axeln. Där de 'neutrala' linjerna, det vill säga de vertikallinjer vid noll[5] innebär att då de vertikala effekterna inte indikerar någon effekt. Värdena som visas i figuren säger hur mycket de varierar från det gemensamt intercept eller trend som erhålls från modellen. Figuren visualiserar dessutom också konfidensintervall för effekterna. Vi observerar även att neutrallinjen går igenom de flesta konfidensintervallen för trend dock inte för intercept.

Generellt skriver man modellen som följande

$$y_{ij} = 3.3347 + \hat{\gamma}_{0i} + (-0.0233 + \hat{\gamma}_{1i})x_j + 0.2991\epsilon_{ij}, \quad i = 1, \ldots, 20, \quad j = 1, ..., k_i, \quad (8)$$

med uppskattade parametrar intercept och trend som följer normalfördelningar ger

$$\hat{\gamma}_0 \sim N(0, 0.3445^2),$$

[5]x-axelposition 0 för de flesta linjära modeller.

$$\hat{\gamma}_1 \sim N(0, 0.0128^2),$$

och feltermen

$$\epsilon_{ij} \sim N(0, 1).$$

Modellen säger att för provtagningsplats i följer skattning av log-koncentration av HG en regressionslinje med funktion y_{ij} i (8) med avseende på tid med slumpmässigt intercept och slumpmässig trend. Där intercepten varierar med medelvärde 3.3347 och varians $\sigma^2_{\hat{\gamma}_0} = 0.3445^2$; och trenderna varierar med medelvärde -0.0233 och varians $\sigma^2_{\hat{\gamma}_1} = 0.0128^2$ för plats i. Observera att vi har en väldig liten varians för slumpmässiga trenden, dock innebär detta inte nödvändigtvis att den slumpmässiga trenden för tid inte är viktig.

Exempelvis i platsen Utlängan skulle det uppskattade intercept vara $\hat{\beta}_0 + \hat{\gamma}_{0\text{Utlängan}} = 3.3347 - 0.1523 = 3.1824$ och på motsvarande sätt för den uppskattade trenden $\hat{\beta}_1 + \hat{\gamma}_{1\text{Utlängan}} = -0.0233 + 0.0062 = -0.0171$ samt residualvarians $\hat{\sigma}^2 = 0.2991^2$. Vi får alltså en regressionslinje för nivå Utlängan med funktion

$$y_{\text{Utlängan},j} = 3.1824 - 0.0171x_j + \epsilon_{\text{Utlängan},j}.$$

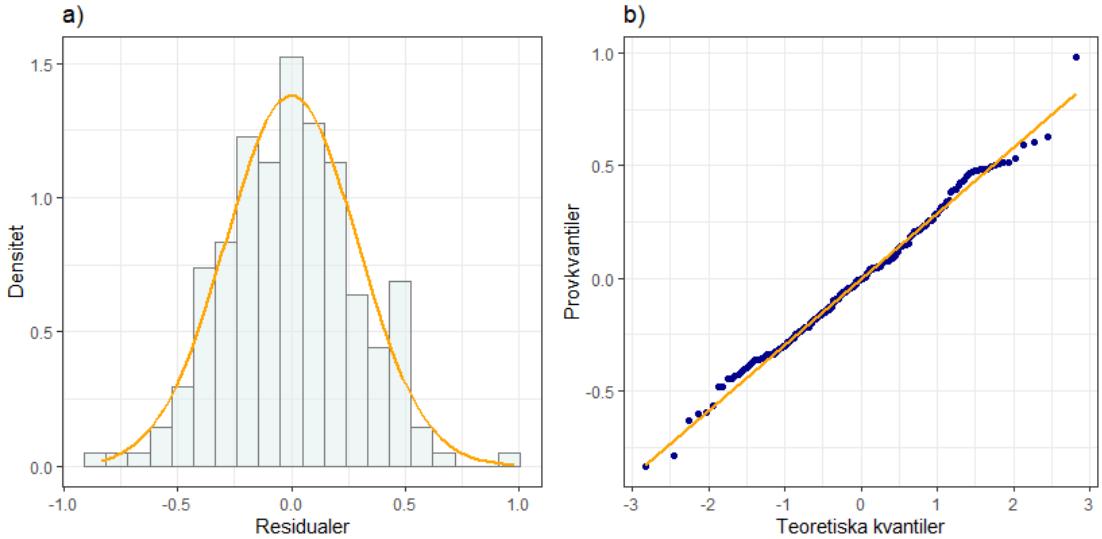
Funktionen innebär bland annat att log-koncentration av miljögift HG är positiv proportionell mot tiden i Utlängan. Alltså en regressionslinje med intercept 3.1824 och trend -0.0171.

3.3.3 Modell 3

På samma sätt som ovan gör vi först kontroll av antaganden för modell 3 i Figur 4. Modellen har liknande resultat som modell 2, att histogrammet följer täthetsfunktionen väl och Q-Q plot visar en linjär trend med små avvikande svans på bägge ändarna, men är i stort sett bra. Vilket tyder på att det är rimligt att också anta att residualerna följer en normalfördelning för modell 3. Figuren längst ner till vänster i Figur 13 visar en Q-Q plot för slumpmässigt intercept i modell 3. Den liknar utseende för slumpmässigt intercept i modell 2 och approximerar också en normalfördelning vilket uppfyller vårt antagande.

Figuren i mitten i Figur 14 visar de skattade regressionslinjer för respektive provtagningsplats för modell 3. Eftersom modellen är en slumpmässig interceptmodell skiljer regressionslinjerna endast intercept mellan olika platser. Med andra ord alla regressionslinjer är parallella vilket visas i figuren. Således vi undersöker inget samband mellan intercepten och trenderna.

Undersöker man närmare med den slumpmässiga effekten illustrerar man figur till vänster i Figur 16 som visualiserar variationer och konfidensintervall för slumpmässigt intercept för modell 3. Värdena som visas i figuren säger hur mycket det varierar från det gemensamt intercept från modellen. Vi observerar att figuren ser liknande ut som skattningsvärde för intercept för modell 2, med samma resultat att neutrallinjen inte går igenom de flesta konfidensintervallen.



Figur 4: a) Histogram och normal kurvan för residualerna för modell 3 och b) Normal QQ-plot med residualerna från modell 3.

Generellt skriver man modellen som följande

$$y_{ij} = 3.3331 + \hat{\gamma}_{0i} - 0.0069x_j + 0.3021\epsilon_{ij}, \quad i = 1, \ldots, 20, \quad j = 1, ..., k_i, \tag{9}$$

med uppskattade parameter intercept som följer en normalfördelning ger

$$\hat{\gamma}_0 \sim N(0, 0.3626^2),$$

och feltermen

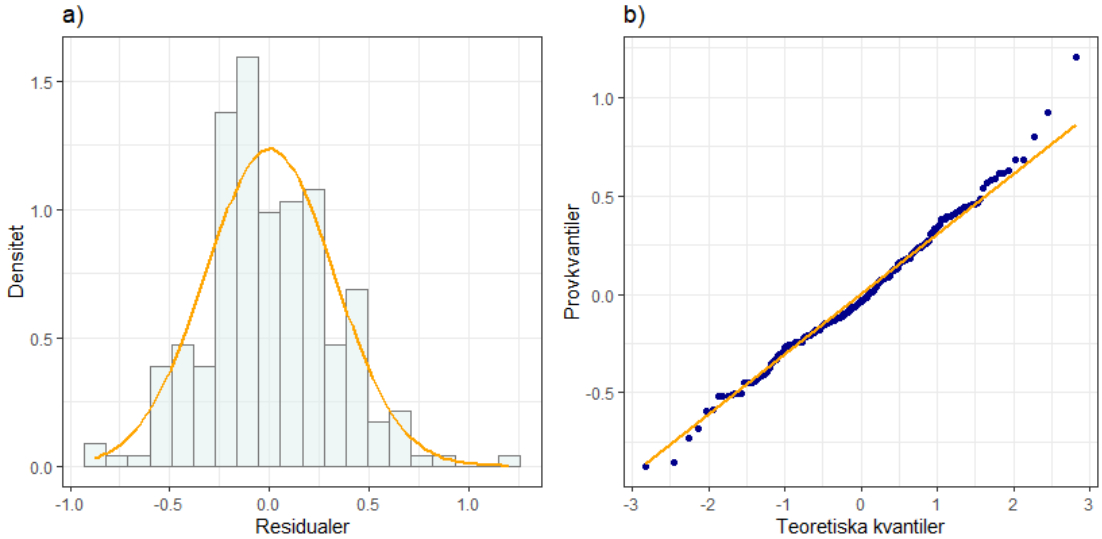
$$\epsilon_{ij} \sim N(0, 1).$$

Modellen säger att skattning av log-koncentration av HG följer regressionslinjer med respektive funktion y_{ij} i (9) med avseende på tid med slumpmässigt intercept och en gemensam trend för varje i. Där intercept varierar med medelvärde 3.3331 och varians $\sigma^2_{\hat{\gamma}_0} = 0.3626^2$ samt trenden är densamma med värde -0.0232 för alla platserna.

Återigen ett exempel på plats Utlängan kommer det uppskattat intercept vara $\hat{\beta}_0 + \hat{\gamma}_{0\text{Utlängan}} = 3.3331 - 0.1147 = 3.2184$ och en gemensam trend $\hat{\beta}_1 = -0.0069$ samt residualvarians $\hat{\sigma}^2 = 0.3021^2$. Vi får alltså en modell för nivå Utlängan med

$$y_{\text{Utlängan},j} = 3.2184 - 0.0232x_j + 0.3021\epsilon_{\text{Utlängan},j}.$$

Funktionen innebär bland annat att log-koncentration av miljögift HG är negativ proportionell mot tiden i Utlängan. Alltså en regressionslinje med intercept 3.2184 och trend -0.0232.



Figur 5: a) Histogram och normal kurvan för residualerna för modell 4 och b) Normal QQ-plot med residualerna från modell 4.

3.3.4 Modell 4

Kontrollering av antagandet för modell 4 i Figur 5. Figuren visar att histogrammet följer täthetsfunktionen väl och Q-Q plot visar en linjär trend med lite avvikande svans uppe till höger men är i stort sett bra. Vilket tyder på att det är rimligt att anta residualerna följer en normalfördelning för modell 4 också. Figuren längst ner till höger i Figur 13 visar en Q-Q plot för slumpmässig trend i modell 4. Den visar en väldig fin linjär trend med få avvikande punkter. Vi säger här att de slumpmässiga trenderna approximerar också en normalfördelning och antagandet är uppfylld.

Figuren till höger i Figur 14 visar de skattade regressionslinjerna för respektive provtagningsplats för modell 4. Eftersom modellen är en slumpmässig trendmodell varierar skattningarna beroende på varje nivås slumpmässiga trender, men intercept är detsamma. Detta visas också i figuren där att regressionslinjerna är radiella vid ett gemensamt intercept. Således vi undersöker inget samband mellan intercepten och trenderna.

Undersöker man närmare med den slumpmässig effekt illustrerar man figur till höger i Figur 16 som visualiserar variationer och konfidensintervall för slumpmässig trend för modell 4. Värdena som visas i figuren säger hur mycket det varierar från den gemensamma trenden från modellen. Till skillnad från modell 2 observerar vi här att figuren skiljer sig från de variationerna för trend för modell 2, då vi har både figurerna med samma skala på x-axeln. Detta förklarar skillnad mellan varianserna för de skattade trenderna i Tabell 3. Neutrallinjen går därmed inte heller igenom de flesta konfidensintervallen.

Generellt skriver man modellen som följande

$$y_{ij} = 3.3435 + (-0.0242 + \hat{\gamma}_{1i})x_j + 0.3370\epsilon_{ij}, \quad i = 1, \ldots, 20, \quad j = 1, ..., k_i, \tag{10}$$

med uppskattad parameter trend som följer en normalfördelning ger

$$\hat{\gamma}_1 \sim N(0, 0.0421^2),$$

och feltermen

$$\epsilon_{ij} \sim N(0, 1).$$

Modellen säger att skattningarna för log-koncentration av HG följer regressionslinjer med respekterande funktion y_{ij} i (9) med avseende på tid med slumpmässiga trender och ett gemensamt intercept för varje i. Där trenderna varierar med medelvärde -0.0242 och varians $\sigma^2_{\hat{\gamma}_1} = 0.0421^2$ samt intercept är detsamma med värde 3.3435 för alla platserna.

Återigen ett exempel på plats Utlängan kommer den uppskattade trenden vara $\hat{\beta}_1 + \hat{\gamma}_{1\text{Utlängan}} = -0.0242 - 0.0062 = -0.0304$ och ett gemensamt intercept $\hat{\beta}_0 = 3.3435$ samt residualvarians $\hat{\sigma}^2 = 0.3370^2$. Vi får alltså en modell för nivå Utlängan med

$$y_{\text{Utlängan},j} = 3.3435 - 0.0304x_j + 0.3370\epsilon_{\text{Utlängan},j}.$$

Funktionen innebär bland annat att log-koncentration av miljögift HG är negativ proportionell mot tiden i Utlängan. Alltså en regressionslinje med intercept 3.3435 och trend -0.0304.

4 Resultat

Sammanfattningsvis får vi följande tre modeller

Modell 2: en slumpmässig intercept- och trendmodell

$$y_{ij} = 3.3347 + \hat{\gamma}_{0i} + (-0.0233 + \hat{\gamma}_{1i})x_j + 0.2991\epsilon_{ij},$$

$$\hat{\gamma}_0 \sim N(0, 0.3445^2),$$

$$\hat{\gamma}_1 \sim N(0, 0.0128^2);$$

Modell 3: en slumpmässig interceptmodell

$$y_{ij} = 3.3331 + \hat{\gamma}_{0i} - 0.0232x_j + 0.3021\epsilon_{ij},$$

$$\hat{\gamma}_0 \sim N(0, 0.3626^2);$$

Modell 4: en slumpmässig trendmodell

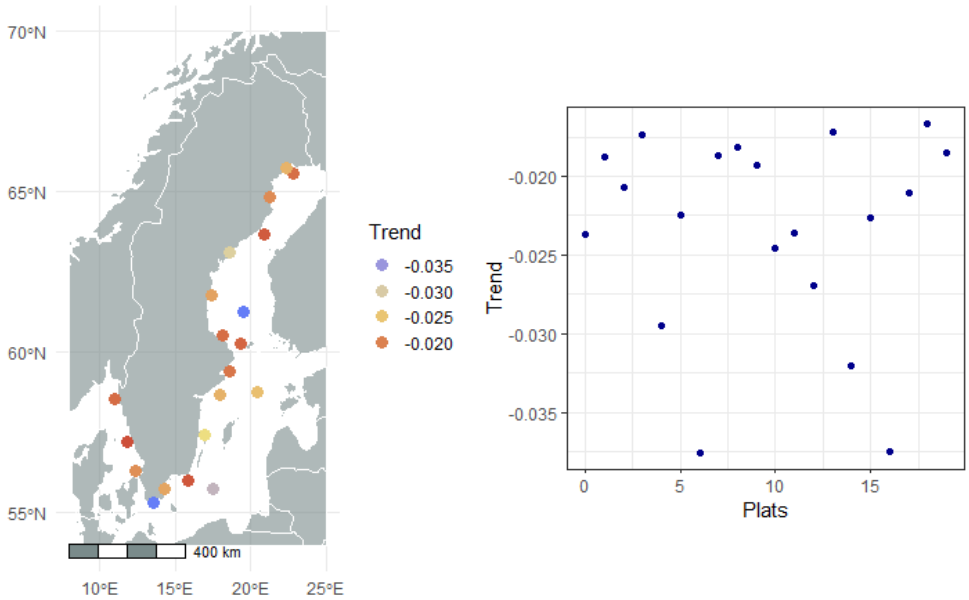
$$y_{ij} = 3.3435 + (-0.0242 + \hat{\gamma}_{1i})x_j + 0.3370\epsilon_{ij},$$

$$\hat{\gamma}_1 \sim N(0, 0.0421^2),$$

för $i = 1, \ldots, 20$, $j = 1, ..., k_i$ och $\epsilon_{ij} \sim N(0, 1)$. Där $\hat{\gamma}_0$ och $\hat{\gamma}_1$ är skattningarna för intercept respektive trend som följer dess normalfördelning för olika modeller. Normalantaganden för alla modeller är uppfyllda genom illustrering av figurerna som nämndes i avsnitt 3.3. Provtagningsplatserna är signifikanta då den neutrallinje inte går igenom de flesta konfidensintervallen i både Figur 15 och 16. De uppskattade parametrarna för modellerna är listade i Tabell 4. De skattade trenderna för modell 2 är negativa och är i sina tur positiv proportionell mot de skattade intercepten; för modell 3 en gemensam negativ trend med värde -0.0232 samt för modell 4 är situation relativ jämn, där en del platser har positiv trend och en del negativ. Resultat från modell 2 innebär att koncentration av HG minskar med tiden vid alla platser.

Med analys av varianserna från modellerna kan det sägas att för modell 2 är både variationerna mellan respektive inom provtagningsplatserna stora, och variationen av regressionskoefficienterna för tid för olika platserna är relativ liten. För modell 3 utan variationen av regressionskoefficienterna för tid är det liknande situation som modell 2. I fallet för modell 4 utan variation mellan platserna har variationen av regressionskoefficienterna för tid ökat enormt såväl som variationen inom platserna.

Figur 6 illustrerar en karta med markerat platser och dess skattade årlig trend, samt ett spridningsdiagram som visar hur de geografiska trender följer strömmen igenom Östersjön. Där vi behandlar platserna i data genom att markera den nordligaste platsen som 0 och på så sätt ordnar alla platser längs strömmen i Östersjön. X-axlen i spridningsdiagrammen visar markeringsnummer för respektive plats med ett intervall från 0 till 19. Ju högre värde på x-axeln, desto längre söderut befinner detekteringsplats vid Östersjön. På



Figur 6: Figur till vänster är en karta med markerat platserna och dess skattade årlig trend, figur till höger är ett spridningsdiagram som visar hur de geografiska trenderna relaterar till strömmen i Östersjön.

motsvarande sätt illustreras de skattade intercepten i Figur 17. Spridningsdiagram visar att varken intercept eller trender har något tydligt samband med strömmen i Östersjön. Dock kan vi intuitivt se att intercept är relativt stort för platser vid kust nära Stockholm och Norrbotten på kartan för de skattade intercepten.

5 Diskussion

Syftet med denna studie är att få en bild av hur trender och nivåer fördelar sig i landet. Därför har jämförelse mellan modellerna en mindre vikt under fallstudien. Vi kommer dock diskutera detta i detta avsnitt. Dessutom är modell 4 en ointressant modell att undersöka på, då den inte tar hänsyn till variation mellan platserna utan endast variation inom platserna. Därtill är AIC-värden för modell 4 relativt hög i jämförelse med modell 2 och modell 3.

Slumpmässiga trendsmodeller ger modellen mycket mer flexibilitet för att passa data, men kräver också mer data för att få bättre approximativa uppskattningar av separata trender för varje nivå. Vårt datamaterial är obalanserad då vid vissa platser täcker

mindre detekteringsår. Vilket kan också synas i Figur 9. Resultaten kan förväntas bättre om vi har ett mer komplett datamaterial. I modell 2 fick vi en relativ liten varians $\sigma^2_{\hat{\gamma}_1}$ som är lika med 0.0128^2. Dock är standardavvikelsen $\sigma_{\hat{\gamma}_1}$ i samma storleksordning som den skattade trenden, det vill säga $\hat{\beta}_1 = -0.0233$. Därför kan man inte på samma vis direkt dra en slutsats om att variationen av förhållandet mellan koncentration av HG och tid på nivåer är betydelselös.

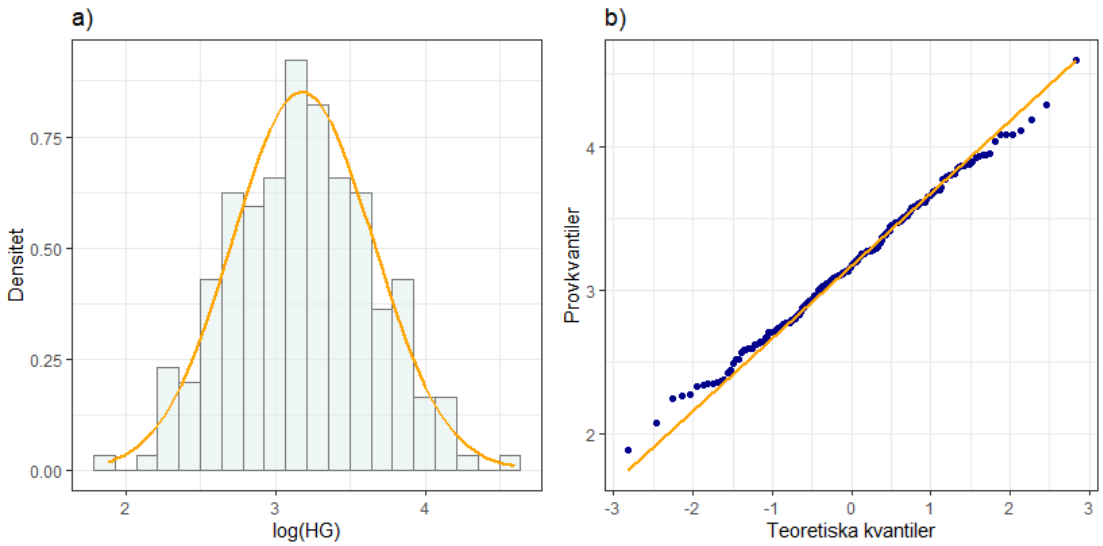
Om man vill avgöra huruvida att ta med variation av förhållandet är nödvändigt, med andra ord huruvida en slumpmässig trend ska läggas till i modellen kan man, förutom AIC, även med hjälp av likelihood ratio test (LRT)[6]. LRT jämför $-2\cdot$log-likelihood mellan modellen med och utan en slumpmässig trend, det vill säga modell 2 respektive 3. Uttrycket ger $-2 \cdot (-74.9 - (-74.6)) = 0.6$, vilket är mycket mindre än det kritiska värdet för en Chi-tvåfördelning med 2 frihetsgrader som är 5.991. I detta fall visar det att modell 3 är betydligt mer exakt på en 5%-signifikansnivå. Vi har också ett AIC-värde för modell 3 som är mindre än för modell 2. Skillnaderna mellan modellerna är dock små.

Enligt vårt resultat har koncentration av HG negativ trend över tid vid alla platser, vilket är gissningsvis inte rimlig i tanken på faktiska situationer och ytterligare undersökning är nödvändigt. Det skulle vara intressant att ta hänsyn till den individuella variationen i koncentration av miljögifter. Med individuell variation innebär variation mellan individuella mätningar. Genom att ta med hänsyn till att egenskaper hos varje individfisk är annorlunda, kan man undersöka närmare i huruvida individuell variation indikerar någon effekt.

Vårt datamaterial är rikligt och innehåller data om många andra faktorer, så som arter, fettprocent, ålder och så vidare. Vår fallstudie utforskade med ett förenklat dataset med endast en art av strömming, vilket är en relativ enkel studie. Linjära blandade modeller kan vara mer komplexa på denna basis. Att exempelvis lägga till arter som en annan nivå i modellen kan också vara en intressant undersökning, då man tänker att observationer som tillhör samma art är relevant. Eftersom att bara undersöka mätningar från strömmingar leder det möjligtvis inte till det förväntade resultatet i andra undersökningar som inte är strömming. Då manifestation av ackumulering av miljögifter kan vara olika hos olika arter. Dessutom med tanke på påverkan av fett på ackumuleringen av miljögifter, kan vi lägga till fettprocent som en annan faktor i modellen. På så sätt kan vi förvänta oss ett mer exakt resultat.

[6] Också kallas för *sannolikhetsförhållandestestet* på svenska. Testet används för att jämföra och avgöra hur bra 'goodness of fit' är för två konkurrerande statistiska modeller baserat på förhållandet mellan deras likelihood.

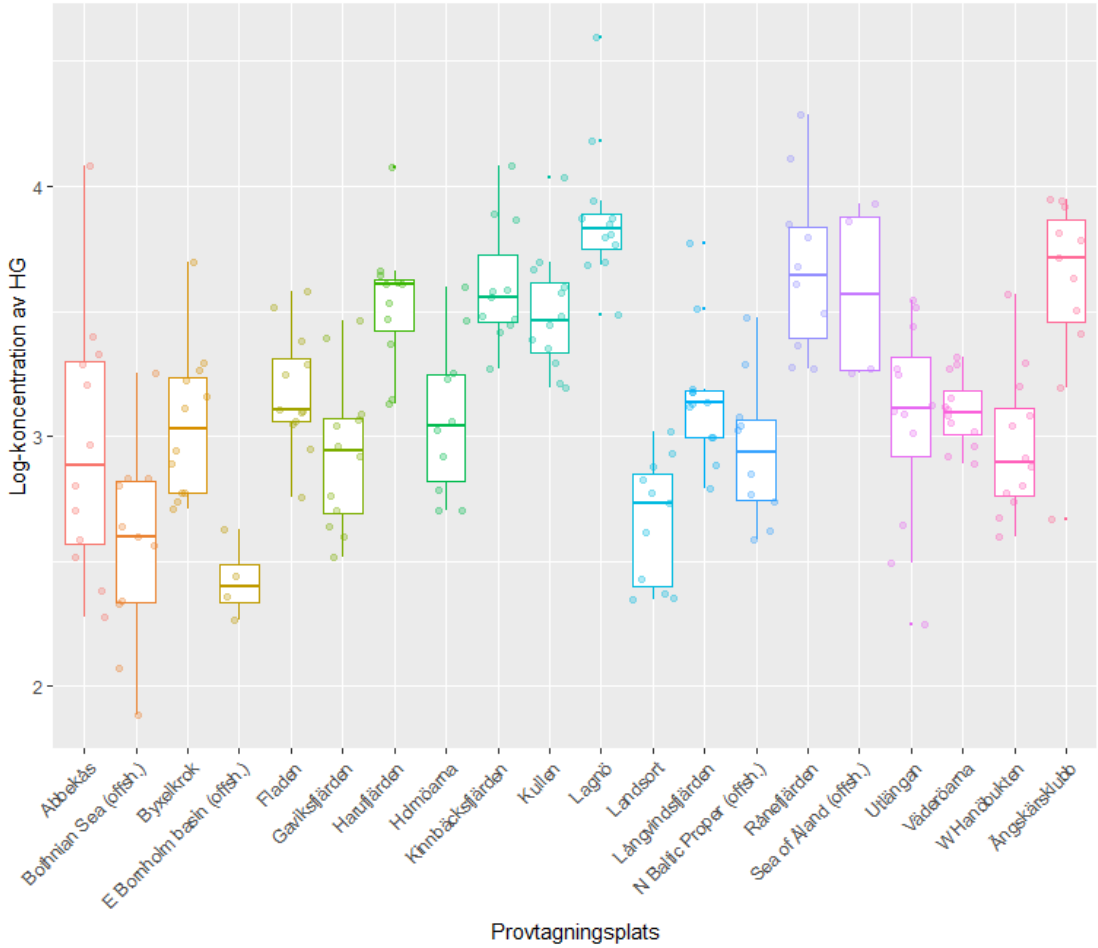
Bilaga



Figur 7: a) Histogram och normal kurvan för logaritm av koncentrationen för strömming och b) Normal QQ-plot.

	ACCNR	YEAR	LOC	LAT	LONG	ALDR	TOTL	TOTV	FPRC	HG
1	C16/03576	10	Utlängan	56.01	15.83	4	17.00	30.40	1.20	266.00
2	C07/06456	1	Utlängan	56.01	15.83	4	17.80	36.30	1.74	125.00
3	C10/11632	4	Ängskärsklubb	60.53	18.16	5	18.10	40.30	1.35	125.00
4	C07/01846-01855	1	Lagnö	59.43	18.63	4	15.20	23.70	2.97	121.00
5	C14/03509	8	Ängskärsklubb	60.53	18.16	5		47.90	1.80	121.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1084	C16/07226	10	Landsort	58.69	18.00	4	18.70	41.90	3.60	6.00
1085	C17/13732	11	Landsort	58.69	18.00	4			2.70	6.00
1086	C18/05124	12	Landsort	58.69	18.00	5	18.60	48.30	3.29	5.02
1087	C13/04810	7	Landsort	58.69	18.00	4	17.20	31.20	3.90	5.00
1088	C16/07228	10	Landsort	58.69	18.00	4	18.40	42.00	2.80	5.00

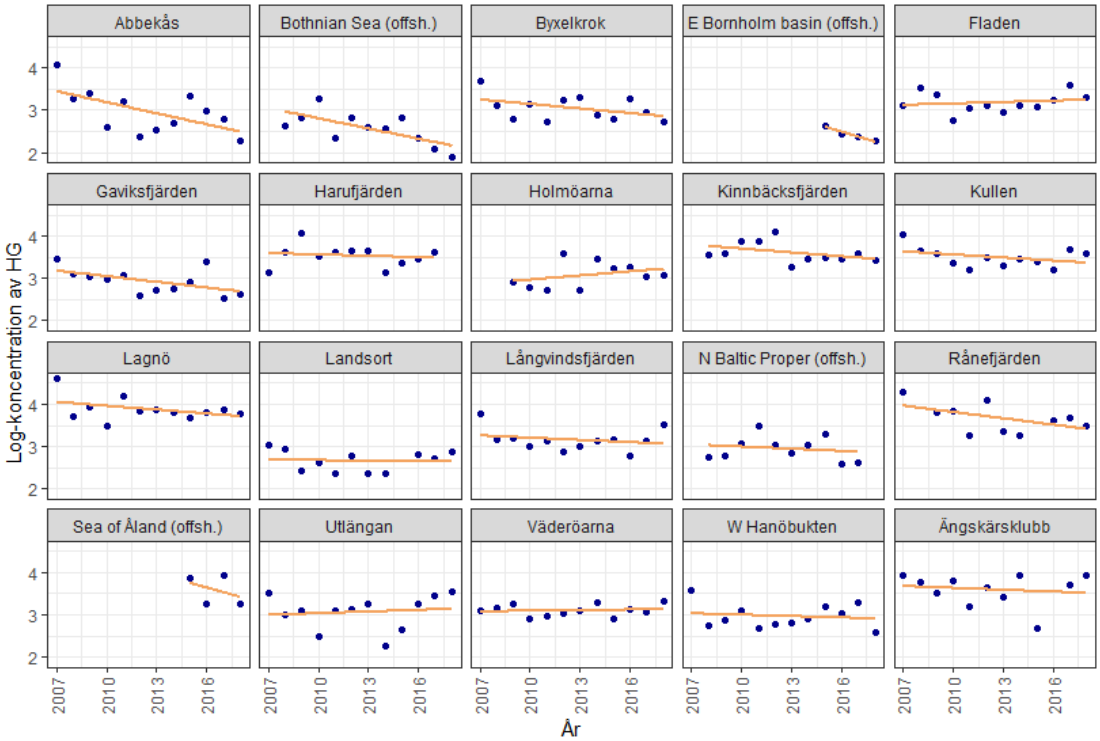
Tabell 2: Exempeldata för strömmingar.



Figur 8: Låddiagram för logkoncentrationer av HG med avseende på provtagningsplatser för strömming.

Plats	Intercept	Trend	Residualvarians	AIC
Abbekås	3.5269	-0.0869	0.4363	17.9597
Bothnian Sea (offsh.)	3.1230	-0.0806	0.2944	8.1094
Byxelkrok	3.2822	-0.0361	0.2801	7.3202
E Bornholm basin (offsh.)	3.6609	-0.1179	0.0374	-11.7160
Fladen	3.1057	0.0111	0.2432	3.9380
Gaviksfjärden	3.2204	-0.0446	0.2681	6.2703
Harufjärden	3.5974	-0.0107	0.2731	6.4551
Holmöarna	2.8334	0.0323	0.3114	8.8152
Kinnbäcksfjärden	3.8156	-0.0304	0.2335	3.0128
Kullen	3.6512	-0.0242	0.2326	2.8690
Lagnö	4.0761	-0.0303	0.2699	6.4319
Landsort	2.6818	-0.0033	0.2634	5.6566
Långvindsfjärden	3.2713	-0.0177	0.2688	6.3392
N Baltic Proper (offsh.)	3.0571	-0.0171	0.2995	8.0366
Rånefjärden	4.0093	-0.0502	0.3108	8.7765
Sea of Åland (offsh.)	4.7279	-0.1094	0.4151	7.5445
Utlängan	2.9747	0.0133	0.4255	17.3565
Väderöarna	3.0686	0.0046	0.1467	-8.1987
W Hanöbukten	3.0350	-0.011	0.2963	8.6723
Ängskärsklubb	3.6846	-0.0148	0.4052	15.1365

Tabell 3: Skattningsvärde för intercept och trend, residualvarians och AIC-värden för modell 1 vid olika platser.



Figur 9: De aritmetiska medelvärden av log-koncentrationer av HG under tidsperiod 2007-2018 för strömming vid olika provtagningsplatser, samt deras linjär regressionslinjen.

```
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's method
['lmerModLmerTest']
Formula: log_y ~ YEAR + (YEAR | LOC)
   Data: new_data

     AIC      BIC   logLik deviance df.resid
   161.2    181.4    -74.6    149.2      207

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.83588 -0.67257  0.00525  0.60368  3.03810

Random effects:
 Groups   Name        Variance  Std.Dev. Corr
 LOC      (Intercept) 0.1186589 0.34447
          YEAR        0.0001649 0.01284  0.06
 Residual             0.0894294 0.29905
Number of obs: 213, groups:  LOC, 20

Fixed effects:
              Estimate Std. Error        df t value Pr(>|t|)
(Intercept)  3.334723   0.090989 18.158318  36.650  < 2e-16 ***
YEAR        -0.023346   0.006862 17.573633  -3.402  0.00327 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
     (Intr)
YEAR -0.413
optimizer (nloptwrap) convergence code: 0 (OK)
Model failed to converge with max|grad| = 0.0024 (tol = 0.002, component 1)
```

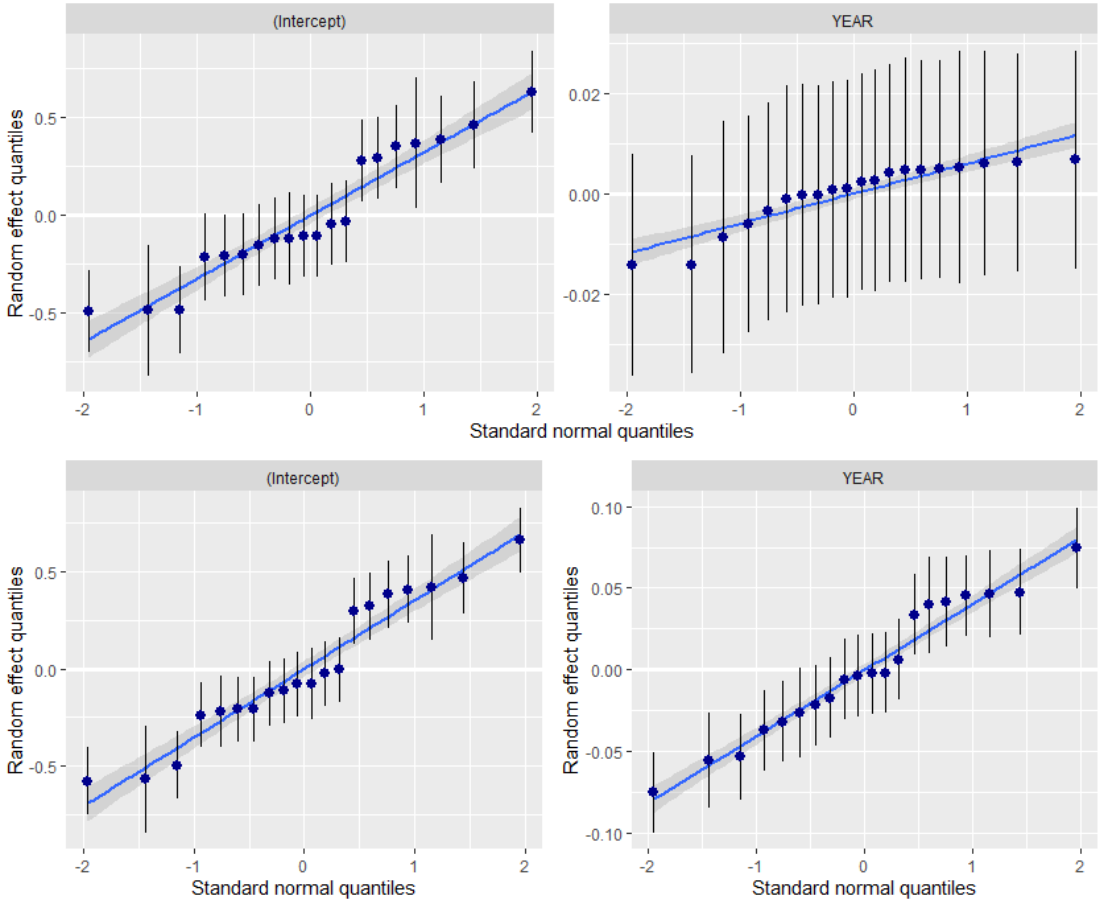
Figur 10: `lmer`-funktionens output för modellen med båda parametrarna slumpmässiga.

```
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's method
['lmerModLmerTest']
Formula: log_y ~ YEAR + (1 | LOC)
   Data: new_data

     AIC      BIC   logLik deviance df.resid
   157.8    171.2    -74.9    149.8      209

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.7637 -0.6701 -0.0087  0.6327  3.2483

Random effects:
 Groups   Name        Variance Std.Dev.
 LOC      (Intercept) 0.13145  0.3626
 Residual             0.09129  0.3021
Number of obs: 213, groups:  LOC, 20

Fixed effects:
              Estimate Std. Error         df t value Pr(>|t|)
(Intercept)   3.333077   0.094427  30.850122  35.298  < 2e-16 ***
YEAR         -0.023216   0.006253 195.277806  -3.713 0.000267 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
     (Intr)
YEAR -0.458
```

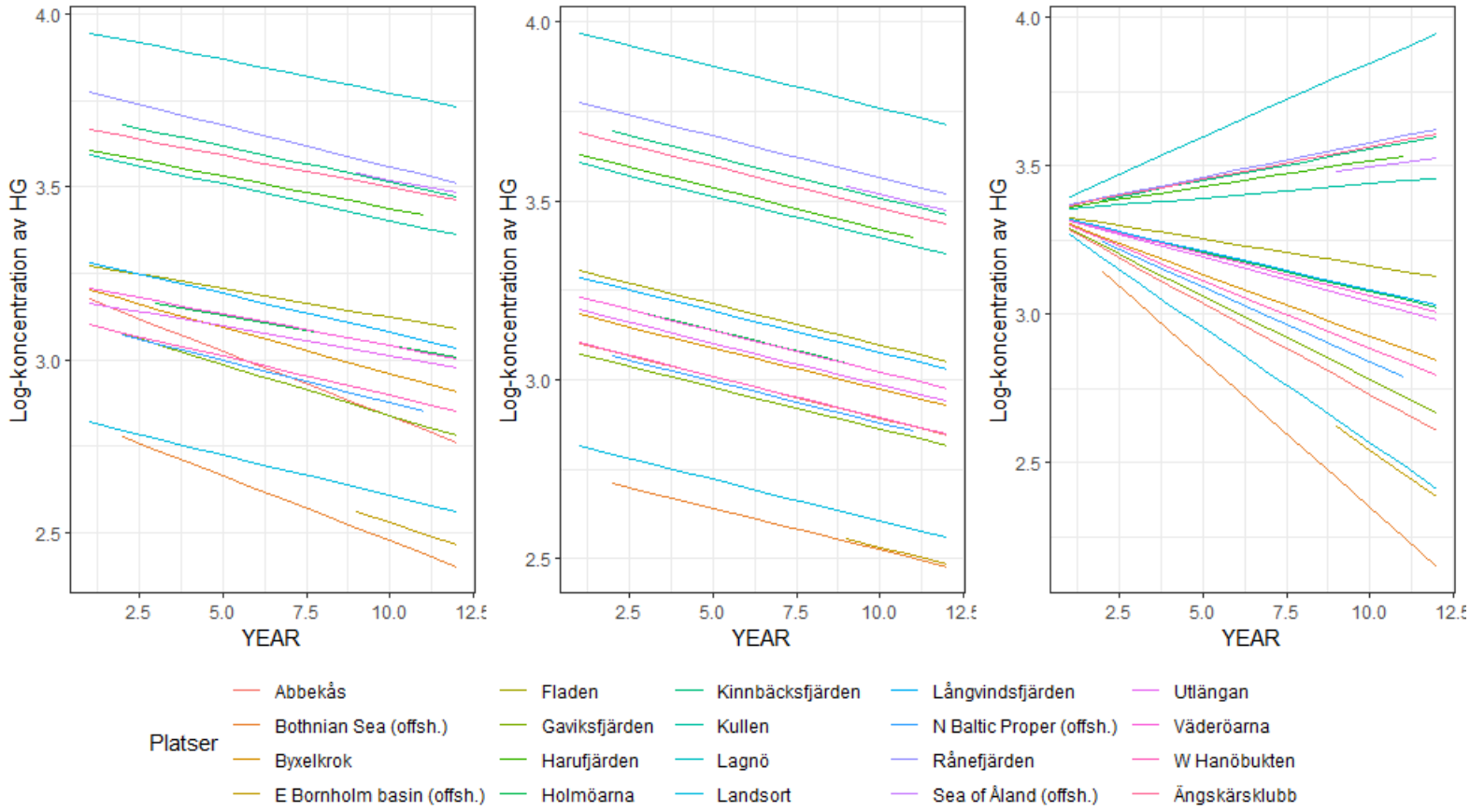
Figur 11: `lmer`-funktionens output för modellen med slumpmässigt intercept.

```
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's method
['lmerModLmerTest']
Formula: log_y ~ YEAR + (0 + YEAR | LOC)
   Data: new_data

     AIC      BIC   logLik deviance df.resid
   195.8    209.2    -93.9    187.8      209

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.5898 -0.6130 -0.1055  0.6084  3.5650

Random effects:
 Groups   Name Variance Std.Dev.
 LOC      YEAR 0.001772 0.04209
 Residual      0.113569 0.33700
Number of obs: 213, groups:  LOC, 20

Fixed effects:
             Estimate Std. Error        df t value Pr(>|t|)
(Intercept)   3.34349    0.05146 193.63884  64.973   <2e-16 ***
YEAR         -0.02416    0.01163  37.19889  -2.078   0.0447 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
     (Intr)
YEAR -0.523
```

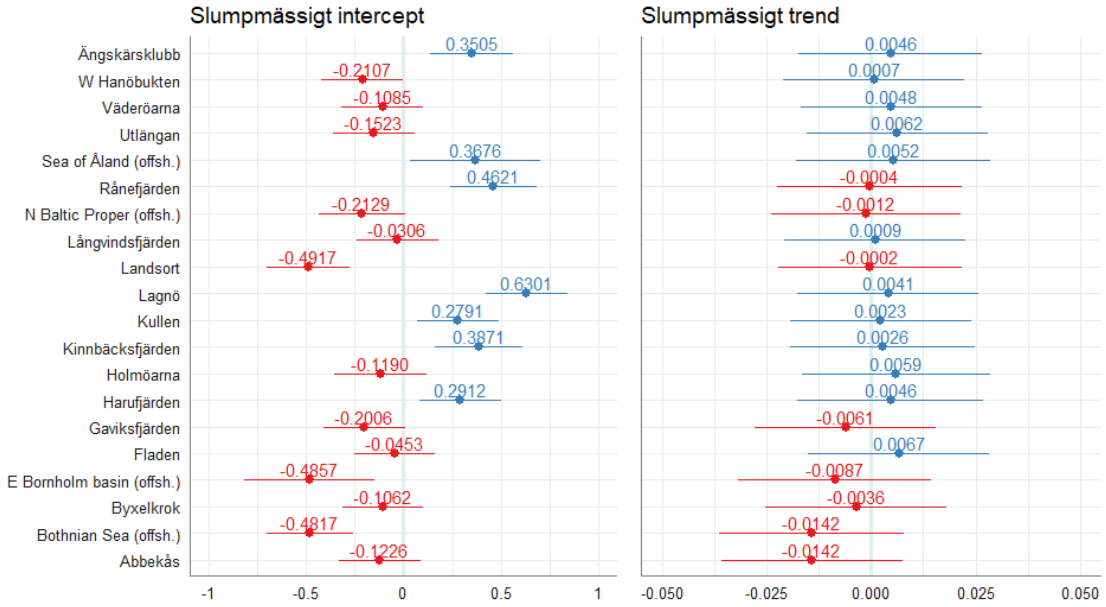
Figur 12: `lmer`-funktionens output för modellen med slumpmässig trend.



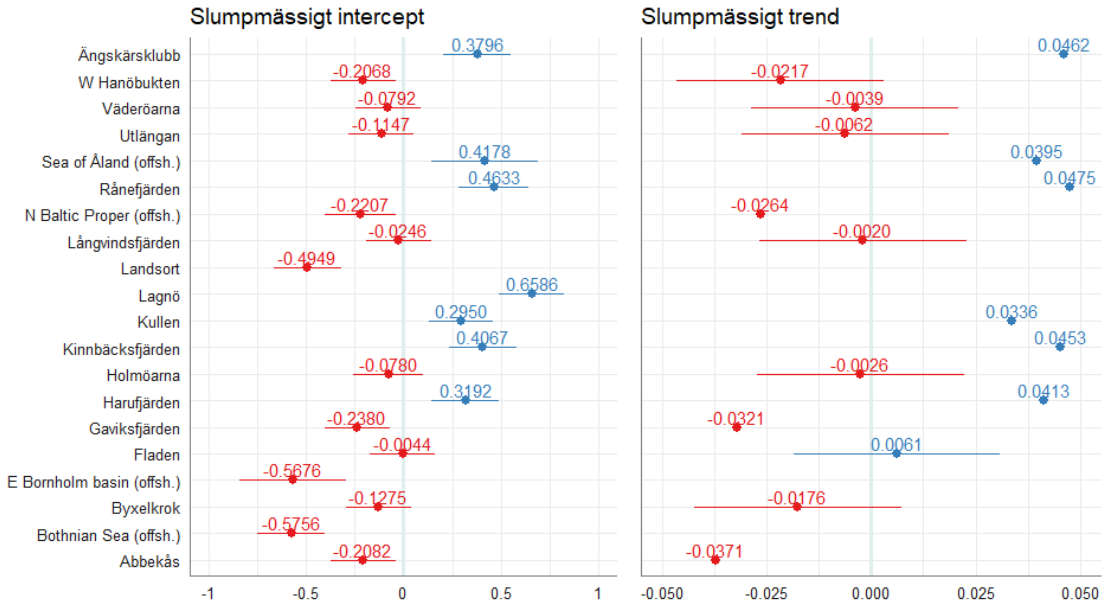
Figur 13: Figur längst upp är Q-Q plot för de teoretiska kvantiler mot slumpmässiga effkterna, intercept respektive trend, för modell 2; längst ner till vänster för modell 3 och längst ner till höger för modell 4.



Figur 14: De uppskattade regressionslinjerna för modell 2, 3 och 4 respectivt.



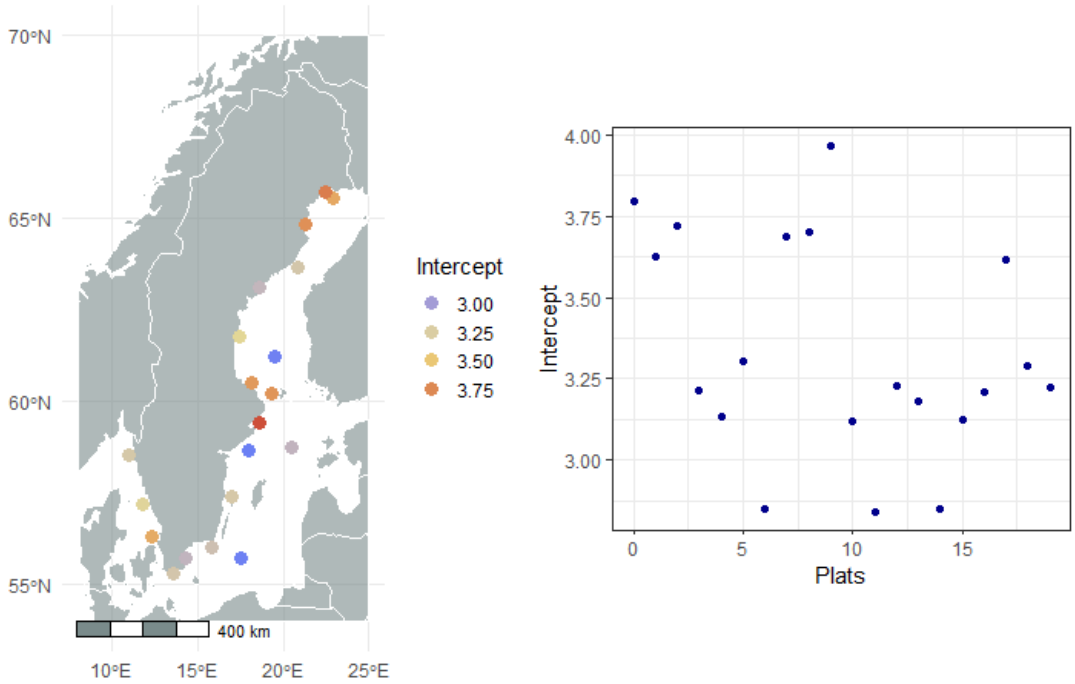
Figur 15: Skattning av slumpässigt intercept och trend för modell 2.



Figur 16: Skattning av slumpässigt intercept för modell 3 och slumpässig trend för modell 4.

	Modell 2		Modell 3		Modell 4	
Plats	Intercept	Trend	Intercept	Trend	Intercept	Trend
Abbekås	3.2121	-0.0375	3.1249	-0.0232	3.3435	-0.0613
Bothnian Sea (offsh.)	2.8531	-0.0376	2.7574	-0.0232	3.3435	-0.0994
Byxelkrok	3.2285	-0.0269	3.2056	-0.0232	3.3435	-0.0418
E Bornholm basin (offsh.)	2.8491	-0.0321	2.7655	-0.0232	3.3435	-0.0800
Fladen	3.2894	-0.0167	3.3287	-0.0232	3.3435	-0.0181
Gaviksfjärden	3.1341	-0.0295	3.0951	-0.0232	3.3435	-0.0562
Harufjärden	3.6260	-0.0187	3.6522	-0.0232	3.3435	0.0172
Holmöarna	3.2158	-0.0174	3.2551	-0.0232	3.3435	-0.0268
Kinnbäcksfjärden	3.7218	-0.0207	3.7398	-0.0232	3.3435	0.0212
Kullen	3.6138	-0.0211	3.6281	-0.0232	3.3435	0.0095
Lagnö	3.9648	-0.0193	3.9917	-0.0232	3.3435	0.0502
Landsort	2.8431	-0.0236	2.8382	-0.0232	3.3435	-0.0775
Långvindsfjärden	3.3041	-0.0225	3.3085	-0.0232	3.3435	-0.0262
N Baltic Proper (offsh.)	3.1218	-0.0246	3.1124	-0.0232	3.3435	-0.0506
Rånefjärden	3.7968	-0.0237	3.7963	-0.0232	3.3435	0.0234
Sea of Åland (offsh.)	3.7023	-0.0182	3.7509	-0.0232	3.3435	0.0153
Utlängan	3.1825	-0.0171	3.2184	-0.0232	3.3435	-0.0304
Väderöarna	3.2262	-0.0185	3.2539	-0.0232	3.3435	-0.0280
W Hanöbukten	3.1240	-0.0227	3.1263	-0.0232	3.3435	-0.0459
Ängskärsklubb	3.6853	-0.0187	3.7127	-0.0232	3.3435	0.0220

Tabell 4: Skattningsvärde för intercept respektive trend för olika kandidatmodeller.



Figur 17: Figur till vänster är en karta med markerat platserna och dess skattade årligt intercept, figur till höger är ett spridningsdiagram som visar hur de geografiska intercepten relaterar till strömmen i Östersjön.

Referenser

[1] Bignert A., Danielsson S., Ek C., Faxneld S., Nyberg E.(2017). Comments Concerning the National Swedish Contaminant Monitoring Programme in Marine Biota, 2017, http://nrm.diva-portal.org/smash/get/diva2:1090746/FULLTEXT01.pdf.

[2] Rolf Sundeberg(Oktober 2016). *Lineära Statistiska Modeller*, Matematisk statistik Stockholms universitet

[3] Held L., Sabanés Bové D.(2014). *Applied Statistical Inference, Likelihood and Bayes*. 1st ed. Springer.

[4] Jiming Jiang(2007) *Linear and Generalized Linear Mixed Models and Their Applications*. 1st ed. Springer.

[5] Raymond H. Myers, Douglas C. Montgomery, G. Geoffrey Vining(January 2012). *Generalized Linear Models: with applications in engineering and the sciences*. 2nd ed. John Wiley & Sons, Inc.

[6] Ben Bolker and others (07 April 2021). *GLMM FAQ*, https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html.

[7] Bates D., Mächler M., Bolker B., Walker S. (07 October 2015). *Fitting Linear Mixed-Effects Models Using lme4*, https://cran.r-project.org/web/packages/lme4/vignettes/lmer.pdf.

[8] Daniel Lüdecke (03 Febuari 2021). *Plotting Estimates (Fixed Effects) of Regression Models*, https://strengejacke.github.io/sjPlot/articles/plot_model_estimates.html.

[9] Jos W. R. Twisk(2019) *Applied Mixed Model Analysis: A Practical Guide*. 2nd ed. Cambridge University Press. https://www-cambridge-org.ezp.sub.su.se/core/books/applied-mixed-model-analysis/16BB3849827F848579608B8C788A51F8.