VLAAMSE MILIEUMAATSCHAPPIJ

inbo
Instituut voor natuur- en bosonderzoek

Een intuïtieve inleiding op de mixed model regressietechniek

Studie uitgevoerd in opdracht van MIRA, Milieurapport Vlaanderen

Onderzoeksrapport

MIRA/2013/03, april 2013

Een intuïtieve inleiding op de mixed model regressietechniek

De principes van *mixed models* aan de hand van simulatievoorbeelden en een analyse van twee concrete meetreeksen van het waterbodemmeetnet

Ivy Janssen, Paul Quataert

Afdeling Advies & Informatie

INBO

Studie uitgevoerd in opdracht van MIRA, Milieurapport Vlaanderen

MIRA/2013/03

April 2013

Documentbeschrijving

Titel
Een intuïtieve inleiding op de mixed model regressietechniek
De principes van *mixed models* aan de hand van simulatievoorbeelden en een analyse van twee concrete meetreeksen van het waterbodemmeetnet

Dit rapport verschijnt in de reeks MIRA Ondersteunend Onderzoek van de Vlaamse Milieumaatschappij. Deze reeks bevat resultaten van onderzoek gericht op de wetenschappelijke onderbouwing van het Milieurapport Vlaanderen. Dit rapport is ook beschikbaar via
www.milieurapport.be

Samenstellers
Ivy Janssen, Paul Quataert
Afdeling Advies & Informatie, INBO

Wijze van refereren
Janssen I. & Quataert P. (2013), Een intuïtieve inleiding op de mixed model regressietechniek, studie uitgevoerd in opdracht van de Vlaamse Milieumaatschappij, MIRA, MIRA/2013/03, INBO.

Vragen in verband met dit rapport
Vlaamse Milieumaatschappij
Milieurapportering (MIRA)
Van Benedenlaan 34
2800 Mechelen
tel. 015 45 14 61
mira@vmm.be

D/2013/6871/018
ISBN 9789491385223
NUR 973/943

Woord vooraf

De waterbodemkwaliteit wordt in Vlaanderen al sinds 2000 intensief opgevolgd met de triademethode. Die methode integreert de resultaten van chemische, biologische en ecotoxicologische analyses en laat toe waterbodems in te delen in kwaliteitsklassen, gaande van niet verontreinigd tot sterk verontreinigd. Het meetnet bestond aanvankelijk uit een 600-tal meetpunten die om de vier jaar bemonsterd werden. Recent is het meetnet ongeveer gehalveerd in omvang. De algemene evolutie van de waterbodemkwaliteit wordt reeds enkele jaren in beeld gebracht door de selectie te maken van de meetpunten die in elk van de beschouwde periodes bemonsterd werden en vervolgens de evolutie van de verdeling over de kwaliteitsklassen te bekijken. Die aanpak geeft een algemene indruk over de evolutie van de waterbodemkwaliteit in Vlaanderen, maar is niet statistisch onderbouwd.

Het doel van deze studie is een methode te introduceren die toelaat de trends van de waterbodemkwaliteit in Vlaanderen op een statistisch onderbouwde manier te analyseren. Daarbij doen zich een aantal specifieke problemen voor. Zo zijn de verschillende waarnemingen op eenzelfde meetplaats niet onafhankelijk van elkaar, werden niet alle meetplaatsen even vaak en met eenzelfde tussenperiode bemonsterd en worden vaak concentraties onder de detectielimiet teruggevonden. Bovendien wijzigen de detectielimieten in de tijd. De methode die hier voorgesteld wordt is die van de “mixed models” regressietechniek. Deze techniek blijkt zeer bruikbaar, maar de wijzigende detectielimieten blijven een aandachtspunt en de mogelijke effecten ervan verdienen verder onderzoek.

Inhoud

Lijst van figuren

Lijst van tabellen

Samenvatting

Dit rapport is een intuïtieve inleiding op de "*mixed models*" regressietechniek. Hiertoe leggen we de onderliggende principes uit aan de hand van een simulatiestudie en vervolgens illustreren we de techniek met twee voorbeelden uit het waterbodemmeetnet.

Net zoals bij klassieke regressie beschrijft de regressievergelijking van een *mixed effects model* of "gemengd" model het wiskundige verband tussen een uitkomstvariabele (bijvoorbeeld de concentratie van een verontreinigende stof in de waterbodem) en een reeks verklarende variabelen (het jaartal, de ecoregio, het type bodem, …). Maar bij *mixed effects* regressiemodellen zijn sommige coëfficiënten opgebouwd uit een som van een vast (*fixed*) en stochastisch (*random*) gedeelte. De toevalsterm modelleert individuele lokale afwijkingen ten opzichte van het globale effect zoals gekwantificeerd door een vaste waarde. Een voorbeeld hiervan is het *random trend* model dat de lokale trend van een verontreinigende stof in individuele waterlopen beschrijft als toevallige fluctuaties ten opzichte van een globale trend die het gemiddelde voorstelt over alle waterlopen heen.

Mixed models kunnen bijgevolg een situatie beschrijven waarbij het lokale patroon afwijkt ten opzichte van het gemiddelde. Omdat we de individuen onvoldoende kunnen karakteriseren, modelleren we het effect van een individu als een toevallige variatie. *Mixed effects* regressiemodellen (of "gemengde" modellen) combineren bijgevolg vaste parameters die de globale effecten van de verklarende variabelen op de uitkomstvariabele (*fixed effects*) kwantificeren en stochastische parameters die de specifieke individuele effecten van de verklarende variabele (*random effects*) modelleren. De mixed model regressietechniek geeft bijgevolg een grote flexibiliteit om naast een algemeen patroon ook recht te doen aan lokale variaties.

Mixed models zijn ook noodzakelijk om herhaalde metingen op eenzelfde object (meetplaats) te analyseren. De metingen van eenzelfde object zullen meer op elkaar gelijken dan metingen in een andere object. Als de concentratie van een stof in een bepaalde meetplaats hoog is, dan mogen we verwachten dat bij een volgend bezoek de concentratie ook hoog zal zijn of toch hoger dan een locatie waar oorspronkelijk de concentratie laag was. Klassieke regressiemodellen kunnen geen rekening houden met deze correlatie tussen de waarnemingen. De *random effects* in een mixed model karakteriseren daarentegen de specifieke kenmerken van de meetobjecten. Hierdoor bouwen we op een natuurlijke manier correlatie in tussen de waarnemingen van eenzelfde meetobject wat beter overeenstemt met de werkelijkheid.

Met deze tekst beogen we inzicht bij te brengen in de mogelijkheden en onderliggende principes van *mixed models* aan de hand van simulatiestudies. Stap voor stap maken we de modellen complexer en visualiseren we de implicaties. We starten met een eenvoudig lineair trendmodel waarbij een uitkomstvariabele (*response variable*) volgens een constante snelheid in de tijd verandert. We tonen aan hoe een regressiemodel meerdere regressierechten in één formule kan omvatten door extra vaste parameters in te voeren die kwantificeren hoe het effect van een verklarende variabele afwijkt onder invloed van een bepaalde categorische variabele of *factor* (bijvoorbeeld een andere trend naargelang ecoregio). *Mixed models* zijn hiervan een logische stap verder door de lokale afwijkingen te modelleren als toevallige variabelen. Hiermee kunnen we per meetpunt de (helling van een) regressierechte laten variëren zonder precies te moeten specifiëren wat de onderliggende oorzaak is. Deze theorie illustreren we aan de hand van de evolutie van de concentraties van twee metalen (cadmium en arseen) uit het waterbodemmeetnet.

Summary

This report is an intuitive introduction to the mixed models regression technique. We will explain the underlying principles by means of a simulation study and illustrate the technique on two examples from the watercourse sediment monitoring network.

As in classical regression, the regression equation of a mixed effects regression model describes the mathematical relationship between an outcome variable (eg concentration of a pollutant in the sediment) and a set of explanatory variables (year, ecoregion, soil type, ...). However, in mixed models some coefficients are composed of the sum of a fixed and a random (stochastic) part. The random term models individual local deviations from the overall effect as quantified by the fixed term. As an example, we can consider a trend model in which the local trend of a pollutant in individual streams is described as random fluctuations around the global trend, representing the average over all streams.

Therefore, mixed models can describe a situation in which the local pattern deviates from the average. Since individuals can be characterized insufficiently, the effect of an individual will be modelled as a random variation around the average. Mixed models then combine fixed parameters, quantifying the global effects of the explanatory variables on the outcome variable (fixed effects), with stochastic parameters, specifying individual effects of the explanatory variable (random effects). In this way, mixed models give great flexibility to add local variations to a general pattern.

To analyze repeated measurements on the same object, the use of mixed models is also necessary. Measurements of the same object will be more alike than measurements from other objects. When the concentration of a substance in a particular measurement site is high, we may expect it to be high again in a next visit, or at least higher than for a location where the original concentration was low. Classical regression models cannot take into account the correlation between repeated observations on the same object. In contrast, random effects in a mixed model characterize the specific characteristics of the measured objects. In this way, the correlation between observations on the same object is automatically accounted for.

This text aims to bring insight into the possibilities and underlying principles of mixed models using several simulation studies. Step by step the models are made more complex and the implications visualized. We start with a simple linear trend model in which an outcome variable changes constantly over time. We show how multiple regression lines can be combined in a single formula by entering an extra categorical variable or factor, allowing for a different effect of the explanatory variable on the outcome of interest (eg, a different trend depending on ecoregion). Mixed models regression is the next logical step. Adding random variables allows the (slope of a) regression line to vary across measurement locations without explicitely specifying the underlying cause. This theory will be illustrated on the evolution of concentrations of two metals (cadmium and arsenic) from the watercourse sediment monitoring network.

1 Inleiding

Dit rapport is een intuïtieve inleiding op de *mixed model* regressietechniek. Een *mixed effects* regressiemodel (een "gemengd" model) bevat zowel onveranderlijke parameters die het vast effect van een verklarende variabele op de uitkomstvariabele modelleert (*fixed effects*) en stochastische parameters die toevallige variaties ten opzichte van het vaste effect modelleren (*random effects*). Onder *fixed effects* verstaan we de algemene patronen in de populatie (een algemene trend, verschillen tussen ecoregio's in Vlaanderen), terwijl *random effects* de situatie van de meetobjecten (meetplaatsen) beschrijven, die we modelleren als toevalsvariabelen.

Om inzicht te geven in de onderliggende principes van *mixed models* zullen we werken met gesimuleerde gegevens. Dat heeft als voordeel dat we weten wat het werkelijke model is en bijgevolg kunnen we concreet vaststellen in hoeverre de statistische technieken inderdaad in staat zijn de onderliggende realiteit (die we anders niet kennen) op te pikken.

Voor deze introductie gaan we ervan uit dat de lezer vertrouwd is met lineaire regressiemodellen. Veel minder gekend is echter de veralgemening tot een factormodel met een of meerdere categorische verklarende variabelen. Als we bijvoorbeeld ecoregio als een (factor)variabele invoeren in het regressiemodel, kunnen we nagaan in hoeverre de trend regionaal verschilt. Hiermee kunnen we niet alleen toetsen of de regio's verschillen, maar kunnen we deze verschillen ook kwantificeren.

Met deze informatie als achtergrond zullen we vervolgens *mixed model* regressie introduceren. Het klassieke regressiemodel gaat ervan uit dat de waarnemingen onafhankelijk van elkaar zijn. Maar dat is niet meer het geval als we de meetplaatsen meerdere keren bezoeken. De meetresultaten op eenzelfde meetlocatie hebben iets gemeenschappelijks. Voor een meetplaats met een hoge concentratie in een bepaald jaar zal bij een volgend bezoek de kans groter zijn op een hoge waarde dan bij een meetplaats met een lage concentratie. Het verschil in uitgangssituatie kunnen we modelleren door het intercept van de regressievergelijking te laten variëren volgens een normale distributie. Ook kan de trend per meetplaats variëren. Ook die variatie in helling kunnen we modelleren via een normale distributie. Hierdoor krijgen we "*mixed models*" met zowel vaste regressiecoëfficiënten ("*fixed effects*") als de fluctuaties daaromheen ("*random effects*").

2 De basismodellen

Deze theoretische inleiding is abstract, maar de simulaties in het hierna volgende hoofdstuk zullen veel begrippen concreter maken. De mathematische inleiding moet helpen om beter de betekenis van de modellen te doorgronden.

2.1 Het lineaire basismodel

We starten met volgend additief lineair regressiemodel:

$$Y_j = \mu_j + \varepsilon_j \quad met: \begin{cases} \mu_j = E\left[Y_j\right] = \beta_0 + \beta_1 X_j \\ \\ \varepsilon_j \overset{iid}{\propto} N(0, \sigma_e{}^2) \end{cases} \quad (j = 1, ..., N) \qquad (1)$$

Hierbij modelleren we de waarnemingen *j* van een responsevariabele *Y* als een som een systematisch (voorspelbaar) gedeelte (gemiddelde of verwachte waarde *μ*) en een stochastisch (onvoorspelbaar) gedeelte (de ruis ε):

- Het *systematische* gedeelte drukt uit hoe de verwachte of gemiddelde waarde van Y (E[Y] = *expected value of Y*) verandert in functie van een X-variabele. De regressievergelijking is lineair: β_0 is het snijpunt met de Y-as (intercept) en β_1 is de helling (slope). De X-variabele stelt de (al dan niet causale) context voor van een waarneming die toelaat om de Y-variabele (gedeeltelijk) te voorspellen.

- Het *stochastische* gedeelte stelt de toevallige afwijkingen (de ruis) voor ten opzichte van de regressierechte. De ruis is *iid* verdeeld (*independent & identically distributed)*. De afwijkingen stammen af van dezelfde normale verdeling met een gemiddelde waarde 0 en een variantie σ_e^2. Ook variëren de afwijkingen totaal onafhankelijk van elkaar; m.a.w. de ruistermen zijn statistisch onafhankelijk.

Op basis van een steekproef van *N* waarnemingen (X_j, Y_j) kunnen we de onbekende parameters schatten: de regressiecoëfficiënten (β_0 & β_1) en de standaardafwijking van de ruis (σ_e). Bij deze puntschattingen wordt telkens de standaardfout bepaald die een indicatie geeft van de nauwkeurigheid van de schattingen. Betrouwbaarheidsintervallen (BI) gaan een stap verder en het (1 – α) % BI geeft het interval waarbinnen – indien het model correct is – de werkelijke waarden van de (onbekende) parameters in (1 – α) % van de gevallen zal liggen. Hierin stelt 1 – α het betrouwbaarheidsniveau voor. De lengte van een BI is een maat voor de nauwkeurigheid van de schattingen.

Zoals de simulaties straks zullen illustreren, bevatten BI heel belangrijke informatie ter aanvulling van de klassieke significantietoetsen die alleen nagaan of een bepaalde parameter significant verschillend is van nul of een andere belangrijke waarde. In tegenstelling hiermee geeft een BI het bereik van alle waarden die mogelijk zijn op basis van de steekproef.

Toch heeft de toets of de helling (β_1) significant verschillend is van nul een belangrijke waarde. Een helling gelijk aan nul stemt overeen met de nulhypothese (H_0) waarbij er geen (lineair) verband is tussen de Y- en X-waarde zodat we het model kunnen vereenvoudigen tot:

$$H_0: \quad Y_j = \beta_0 + \varepsilon_j \qquad (2)$$

Zeker wanneer er heel veel mogelijke verklarende variabelen zijn, is het belangrijk te onderzoeken welke variabelen significant bijdragen tot de voorspellingskracht van het model om het model te vereenvoudigen. Een belangrijk onderdeel van de statistische analyse is te bepalen welk model op een zo eenvoudig mogelijke wijze de ingezamelde gegevens voorstelt.

Maar, zoals hierboven al aangestipt, is het wel belangrijk te beseffen dat toetsen sterk misleidend kunnen zijn en dat het altijd noodzakelijk is om ook de BI te betrekken bij de interpretatie van de resultaten. Ook zijn er diagnostische testen zoals een residu-analyse nodig om na te gaan of de veronderstellingen van het model wel kloppen; bijvoorbeeld, of de relatie tussen de Y- en de X-waarde niet-lineair is (parabolisch) zoals in onderstaand model.

$$H_P: \quad Y_j = \beta_0 + \beta_1 X_j + \beta_2 X_j^2 + \varepsilon_j \tag{3}$$

2.2 Het veralgemeende lineaire model

We kunnen het basismodel heel beknopt als volgt voorstellen:

$$Y \sim X \tag{4}$$

Deze internationaal aanvaarde notatie laat ook toe heel vlot complexere modellen te noteren zonder alles in detail te moeten uitschrijven. Hieronder een paar voorbeelden:

Y ~ X + X² + X³ (derdegraadsmodel)
Y ~ X + Z + W (regressiemodel met drie onafhankelijke variabelen)
Y ~ poly(X,3) (derdegraadsmodel, maar nu geschreven als een polynoomfunctie)

Ook vermenigvuldigingen tussen variabelen zijn mogelijk. We noteren dat als interacties (W:Z in onderstaande formule) omdat de impact van de ene variabele beïnvloed wordt door de waarde van de andere variabele. Hieronder een voorbeeld:

$$Y \sim poly(X,3) + W + Z + W:Z \tag{5}$$

Het lineaire model kunnen we op meerdere manieren verder veralgemenen. Een eerste heel belangrijke maar minder gekende veralgemening is de toevoeging van niet-continue klassevariabelen aan het regressiemodel. De mogelijkheid om categorische variabelen (*factors*) op te nemen in het model, is heel interessant om na te gaan of regressievergelijkingen significant verschillen van elkaar. Bijvoorbeeld, voor het waterbodemmeetnet willen we weten of de trend verschillend is naargelang de ecoregio. Voor elke ecoregio apart kunnen we een regressievergelijking neerschrijven; bv. voor ecoregio A, B & C:

$$\mu_{j(R)} = \begin{cases} \beta_{0A} + \beta_{1A} X_j \\ \beta_{0B} + \beta_{1B} X_j \quad R = A, B, C \\ \beta_{0C} + \beta_{1C} X_j \end{cases} \tag{6}$$

Zoals de lezer eenvoudig kan nagaan, kunnen we deze vergelijkingen als volgt herschikken:

$$\mu_{j(R)} = \begin{cases} \beta_{0A} + \beta_{1A} X_j \\ (\beta_{0A} + \delta_{0B}) + (\beta_{1A} + \delta_{1B}) X_j \quad \delta_{0B} = (\beta_{0B} - \beta_{0A});\ \delta_{1B} = (\beta_{1B} - \beta_{1A}) \\ (\beta_{0A} + \delta_{0C}) + (\beta_{1A} + \delta_{1C}) X_j \quad \delta_{0C} = (\beta_{0C} - \beta_{0A});\ \delta_{1C} = (\beta_{1C} - \beta_{1A}) \end{cases} \tag{7}$$

of beknopter:

$$\mu_{j(R)} = (\beta_0 + \delta_{0R}) + (\beta_1 + \delta_{1R}) X_j \quad (R = A, B, C \ \& \ \delta_{0A} = \delta_{1A} = 0) \tag{8}$$

Modellen (7) of (8) zijn volledig equivalent met model (6), maar de parametervoorstelling is anders. Er is een referentierechte (in regio A) ten opzichte waarvan het model de andere regio's situeert. Het voordeel van deze (op het eerste zicht complexere) schrijfwijze is dat we nu kunnen toetsen of deze verschillen significant zijn en kwantificeren hoe groot de verschillen zijn tussen de regressiecoëfficiënten (met bijhorende BI).

In het meest algemene geval zijn de regressierechten totaal verschillend naargelang de regio. Twee nulhypothesen die het model vereenvoudigen zijn hier relevant:

$$\begin{cases} H_{01} \quad \leftrightarrow \quad \delta_{1B} = \delta_{1C} = 0 \\ H_{02} \quad \leftrightarrow \quad \delta_{1B} = \delta_{1C} = 0 \quad \& \quad \delta_{0B} = \delta_{0C} = 0 \end{cases} \tag{9}$$

Bij de eerste nulhypothese is de helling van de regressierechte in de drie regio's gelijk en hebben we bijgevolg evenwijdige rechten. De tweede nulhypothese gaat nog een stap verder en veronderstelt dat ook

de snijpunten met de Y-as voor elke regio gelijk zijn. In dat geval vallen de regressierechten volledig samen. In de verkorte notatie kunnen we de drie modellen (het algemene model + de twee nulhypothesen) als volgt schrijven:

$$\begin{cases} Y \sim X + Ecoregio + X:Ecoregio & (Full\ Model) \\ Y \sim X + Ecoregio & (H_{01}) \\ Y \sim X & (H_{02}) \end{cases} \tag{10}$$

In het eerste model is er een interactie tussen *Ecoregio* en *X* wat impliceert dat er minstens een regressierechte verschilt van de andere. In het tweede model is de interactie verdwenen, maar is er wel nog een effect van Ecoregio. De regressierechten zijn evenwijdig en minstens een van rechten valt niet samen met de andere. In het laatste model heeft Ecoregio helemaal geen invloed meer op de regressierechte.

Bovenstaande beschouwingen komen erop neer dat we in een regressiemodel zowel continue als categorische variabelen (of "factors") kunnen opnemen. Intern converteert de software de factor-variabelen naar numerieke waarden, maar het is niet nodig te begrijpen hoe deze conversie technisch in elkaar zit. De essentie is dat we binnen het veralgemeende lineaire model met zowel numerieke als factorvariabelen door elkaar kunnen werken en de interactie tussen deze variabelen kunnen bestuderen. Deze uitbreiding verruimt het lineaire model in een belangrijke mate.

Een tweede belangrijke punt is dat deze veralgemening ook een strategie aanbiedt om te toetsen welk model het beste aansluit bij de gegevens. Hierbij wordt gezocht naar het meest eenvoudige model dat toch nog goed aansluit bij de gegevens (*principle of parsimony*). We beginnen bij het volledige startmodel (*full model*) waarbij alle regressiemodellen van elkaar verschillen. We onderzoeken of de interactie met Ecoregio significant is. Indien niet, dan kunnen we de interactieterm laten vallen, waardoor het model vereenvoudigt tot evenwijdige rechten. Vervolgens toetsen we of Ecoregio zelf significant is. Indien niet, dan kunnen we ook Ecoregio als term laten vallen en houden we één gemeenschappelijke regressierechte over. Hoe de toetsing precies verloopt, zullen we uitleggen bij de simulaties en de concrete voorbeelden. In het bijzonder bespreekt sectie 3.1.4 enkele veel gebruikte criteria om de complexiteit van een model af te wegen ten opzichte van de precisie.

2.3 Het mixed model

Meestal zijn de waarnemingen niet onafhankelijk van elkaar. Het waterbodemmeetnet volgt een steekproef van waterlopen in een vierjaarlijkse cyclus op. De metingen van dezelfde waterloop zullen meer op elkaar gelijken dan deze van een andere waterloop. Ook is het onrealistisch te veronderstellen dat de trend in elke waterloop volledig identiek is. Lokale factoren zullen ervoor zorgen dat elke waterloop een eigen traject volgt.

We zouden voorgaand factormodel kunnen gebruiken om deze trajecten afzonderlijk te modelleren. Maar dat zou al heel snel leiden tot veel te veel parameters. Daarenboven zijn we op een regionale schaal niet geïnteresseerd in de evolutie van waterlopen apart, maar willen we een uitspraak kunnen doen over de waterlopen als geheel; wat de globale trend is en wat de variatie hierop is. Om al deze redenen is volgend model relevanter:

$$\mu_{jk} = (\beta_0 + b_{0k}) + (\beta_1 + b_{1k}) X_j \quad met: \quad \begin{cases} b_{0k} \overset{iid}{\propto} N(0, \sigma_0^2) \\ \\ b_{1k} \overset{iid}{\propto} N(0, \sigma_1^2) \end{cases} \tag{11}$$

In bovenstaand model stelt *k* een waterloop voor met elk een eigen regressierechte. De regressiecoëfficienten stellen we voor als een som van een gemiddelde vaste ("fixed") term (β_0 en β_1) en een toevallige ("random") afwijking (b_0 en b_1). Model (11) gelijkt heel sterk op model (8), maar de betekenis van de coëfficiënten is anders. De gemiddelde termen geven aan hoe de gemiddelde concentratie evolueert over de totale regio, de verschillen tussen de waterlopen worden nu gemodelleerd als afkomstig uit een normale verdeling met spreiding σ_0^2 (variantie op het intercept) en σ_1^2 (variantie op de helling). Zijn de variantietermen klein, dan evolueren alle waterlichamen op een gelijke manier. Zijn ze groot, dan zijn er grote individuele verschillen tussen de waterlichamen.

Meer abstract kunnen we twee modellen neerschrijven:

$$\begin{cases} Y \sim X \quad + \quad 1 \,|\, Waterloop \\ Y \sim X \quad + \quad X \,|\, Waterloop \end{cases} \tag{12}$$

In het eerste model varieert alleen het intercept (1 staat symbool voor het intercept), in het tweede model varieert ook de helling met de waterloop. In het eerste geval veronderstellen we dat $\sigma_1^2 = 0$ en bijgevolg is het eerste model een bijzonder geval van het tweede. Zoals we zullen aantonen met de simulatiestudie kunnen we statistisch toetsen in hoeverre een vereenvoudiging mogelijk is.

3 Enkele simulaties

Het voordeel van een simulatiestudie is dat we de werkelijkheid kennen. We kunnen op die manier direct – zonder al te veel theorie – vaststellen hoe de technieken "werken" en of de schattingen inderdaad de onbekende parameters op een correcte manier bepalen.

3.1 Het lineaire basismodel

3.1.1 De simulatiemodellen

Om te starten beschouwen we gegevens van één meetplaats waarvan we de concentratie van een bepaalde stof in de waterbodem in de tijd opvolgen. Hierbij gaan we uit van volgend lineair trend model:

$$M_{L1}: \quad C_j = \mu_j + \varepsilon_j \quad waarbij: \quad \begin{cases} \mu_j = 20 + 0.5\, mJaar \\ \varepsilon_j \overset{iid}{\propto} N(0,1) \end{cases} \tag{13}$$

Het model stelt de evolutie voor van de concentratie van een bepaalde stof (*C*) in een waterbodem in functie van de tijd. De concentratie verandert met een halve eenheid per jaar (β_1). De variantie op de ruis is gelijk aan 1 (σ_e^2).

We verzamelen gegevens over een periode van 12 jaar vanaf 2000 tot en met 2011 (N = 12). De X-variabele (***mJaar = Jaar – 2005.5***) is gecentreerd ten opzichte van het midden van deze periode en beschouwen we als **het referentiejaar**. Het intercept (β_0 = 20) stelt bijgevolg de concentratie voor halfweg 2005. Aangezien de trend lineair is, is het intercept ook gelijk aan de gemiddelde waarde voor de hele periode. Dat maakt het model goed vergelijkbaar met het nulmodel zonder trend (14).

$$M_{L0}: \quad C_j = 20 + \varepsilon_j \quad \& \quad \varepsilon_j \overset{iid}{\propto} N(0,1) \tag{14}$$

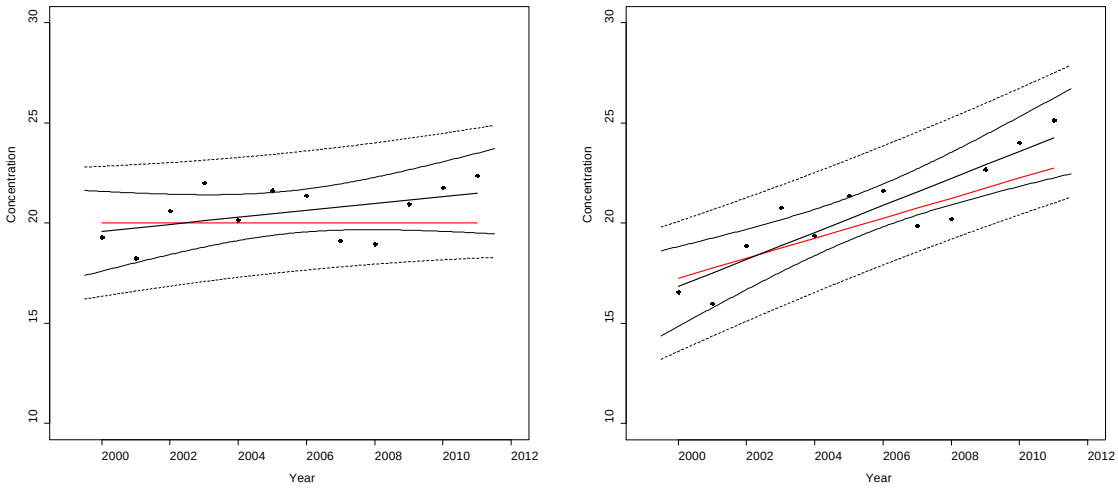
Een belangrijk doel van de statistische analyse is om na te gaan in hoeverre er aanwijzingen zijn voor een trend, m.a.w. om te onderzoeken op basis van de gegevens welk van de twee modellen geldig is.

3.1.2 De gesimuleerde waarnemingen

Figuur 1 toont het resultaat van twee simulaties telkens uitgaande van een ander model: links is gebaseerd op het model zonder trend (14), rechts op het model (13) met trend. De rechten in het rood stellen de (normalerwijze) onbekende regressielijn voor (die we hier wel kennen omdat we simuleren). De regressielijnen zijn geschat (met R) op basis van de gesimuleerde waarnemingen. In beide gevallen is het resultaat een (stijgende) regressierechte, maar links is het resultaat niet significant en rechts wel zoals grafisch aangegeven door de 95 % betrouwbaarheidsbanden (de binnenste volle lijnen in Figuur 1) die de modelonzekerheid karakteriseren. Als een betrouwbaarheidsband een horizontale rechte niet uitsluit (zoals in de figuur links), dan is de trend niet significant (zoals in de linkerfiguur).

Een 95 % betrouwbaarheidsband is zo berekend dat, als het model correct is, de werkelijke (normalerwijze onbekende) regressielijn in 95 % van de gevallen volledig binnen deze band ligt. In de figuur blijkt dat inderdaad het geval te zijn. De buitenste curven (in stippellijn) stellen het 95 % voorspellingsinterval voor. Een betrouwbaarheidsinterval bevat met 95 % zekerheid de gemiddelde concentratie, terwijl een voorspellingsinterval met 95 % zekerheid een individuele concentratie bevat. We verwachten dus dat 95 % van de waarnemingen binnen het voorspellingsinterval zullen liggen. Het voorspellingsinterval is breder dan de betrouwbaarheidsband. Dat is logisch. De onzekerheid op een (nieuwe) waarneming is een som van de onzekerheid op het model (het gemiddelde) en de ruis.

Figuur 1: Het lineaire basismodel (links: geen trend; rechts: wel een (lineaire) trend. Rode lijn = het werkelijke model waaruit de waarnemingen gesimuleerd werden; volle lijnen = geschatte regressierechte (in het midden) samen met de 95 % betrouwbaarheidsband (volle lijn) en voorspellingsband (stippellijn)



3.1.3 Bespreking van de R-output

R-output 1 en R-output 2 geven de output van de statistische analyse. Hiervan zullen we telkens (ook in de hierna volgende simulaties) volgende elementen bespreken:

- De selectie van het correcte model
- De bepaling van de regressieparameters (zowel de punt- als intervalschattingen)
- De schatting van de ruisterm (de standaardafwijking van de waarnemingen)

<u>De selectie van het correcte model</u>

De ANOVA-tabellen toetsen met een F-toets of de trend-term in het model statistisch significant is. In het eerste geval is de p-waarde 0.14 of verschilt de helling niet significant van 0 en zijn er dus geen aanwijzingen voor een trend. In het tweede geval verschilt de helling wel significant van nul ($p <<< 0.05$) en is er dus wel evidentie voor een trend.

<u>De regressiecoëfficiënten (puntschattingen)</u>

De tabel met de coëfficiënten geeft de schattingen, de standaardfout, t-waarde (verhouding geschatte waarde en standaardfout) en de corresponderende p-waarde (t-toets). In het geval van een enkelvoudige regressie is de t-toets voor de helling equivalent met een F-toets. Bij een meervoudige regressie (met meerdere variabelen) wordt voor elke parameter apart de significantie nagegaan (zie verder). We zien hier inderdaad dat voor beide modellen de t-toets voor de helling dezelfde p-waarde geeft als we kunnen afleiden uit de ANOVA-tabel en de veralgemeende F-toets. Het kwadraat van de t-waarde (1.62) is ook gelijk aan de F-waarde (2.63)

<u>De betrouwbaarheidsintervallen bij regressiecoëfficiënten (intervalschattingen)</u>

Uit het 95 % BI voor de regressiecoëfficiënten blijkt opnieuw dat de trend in het eerste geval niet significant is aangezien het BI de nulwaarde omvat (0.173; BI: -0.065, 0.412). In het tweede geval omvat het BI de nulwaarde niet (0.673; BI: 0.435, 0.912).

Maar BI geven tegelijk de onzekerheid aan van de schatting. We kunnen bijvoorbeeld in het eerste geval op basis van de gegevens niet uitsluiten dat in werkelijkheid de trend sterk positief is; de bovengrens van het BI is 0.412. Als deze precisie onvoldoende is voor een praktische toepassing, dan is een vervolgstudie nodig en moeten we de steekproefgrootte opdrijven (bijvoorbeeld door meerdere waterlopen op te volgen in de tijd, zie verder).

Bemerk dat in het tweede geval de BI even breed zijn. Op basis van de statistische analyse verwachten we de werkelijke waarde tussen 0.435 en 0.912. Maar we kunnen hier wel met een vrij hoge zekerheid

afleiden uit de gegevens dat de trend minstens 0.435 is en kunnen we het signaal geven dat er een stijging aan de hand is.

De ruisterm

Het laatste deel van de output geeft de standaardafwijking van de ruis (σ_e). De werkelijke waarde was 1 en de schatting was in beide gevallen 1.28 (BI: 0.625, 3.941). We bekomen dezelfde waarde omdat de ruis in beide modellen afkomstig is van dezelfde simulatie. Het toevoegen van een trend heeft bijgevolg nauwelijks invloed op de schatting van de ruis.

Het BI is heel breed. Dat is altijd zo. Variantiecomponenten zijn heel moeilijk precies te schatten omdat ze een chi-kwadraat verdeling volgen. Maar ook hier kan de schatting veel verbeteren door de steekproef te vergroten en door meerdere waterlopen op te volgen, waardoor we meer inzicht krijgen in de structuur van de variabiliteit (zie verder).

Het is echter belangrijk om ten volle te beseffen dat de standaardfouten op de andere coëfficiënten recht evenredig is met de grootte van de ruisterm. Het is dus belangrijk om inzicht te krijgen in de bronnen die bijdragen tot de ruisterm en bijvoorbeeld uit te zoeken of de ruis te maken heeft met de variabiliteit van de waterloop of met de onzekerheden van de metingen.

R-output 1: resultaten voor Figuur 1 – links (geen trend)

```
Analysis of Variance Table
          Df Sum Sq Mean Sq F value Pr(>F)
mJaar      1    4.3    4.30    2.63   0.14
Residuals 10   16.4    1.64

Generalised F-test of trend model (1) versus null model (2)
  Res.Df  RSS Df Sum of Sq    F Pr(>F)   AIC
1     10 16.4                          7.73
2     11 20.7 -1      -4.3 2.63   0.14  8.53

Coefficients (full model)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   20.546      0.369   55.62 8.56e-14
mJaar          0.173      0.107    1.62 1.36e-01

Confidence limits
             2.5 % 97.5 %
(Intercept) 19.723 21.369
mJaar       -0.065  0.412

Sigma: 1.28 (0.625,3.941)
```

R-output 2: resultaten voor Figuur 1 – rechts (positieve trend)

```
Analysis of Variance Table
          Df Sum Sq Mean Sq F value Pr(>F)
mJaar      1   64.8    64.8    39.6  9e-05
Residuals 10   16.4     1.6

Generalised F-test of trend model (1) versus null model (2)
  Res.Df  RSS Df Sum of Sq    F Pr(>F)    AIC
1     10 16.4                           7.73
2     11 81.2 -1     -64.8 39.6  9e-05  24.95

Coefficients
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   20.546      0.369   55.62 8.56e-14
mJaar          0.673      0.107    6.29 8.99e-05

Confidence limits
             2.5 % 97.5 %
(Intercept) 19.723 21.369
mJaar        0.435  0.912

Sigma: 1.28 (0.625,3.941)
```

3.1.4 Modellen vergelijken en het zuinigheidsprincipe (*parsimony*)

Een alternatieve manier, die hier op zich weinig toevoegt, maar die bij complexere modellen een algemene strategie aanreikt om (geneste) modellen met elkaar te vergelijken, is de veralgemeende F-toets die op

basis van de verandering van de *regression sum of squares* (SSR = variatie verklaard door het regressiemodel) t.o.v. de *residual sum of squares* (SSE = residuele ruis ten opzichte van het model) nagaat in hoeverre de toename in complexiteit zich vertaalt in een betere aansluiting van het model bij de gegevens.

$$F = \frac{\frac{SSR_1 - SSR_2}{df_1 - df_2}}{\frac{SSE_1}{df_1}} \propto F(df_1 - df_2; df_1) \qquad M_2 \subset M_1 \tag{15}$$

Hierin is M_1 het meest complexe model, M_2 het vereenvoudigde en wel zo dat het tweede model een bijzonder geval is van het eerste (technisch: "nested" models). Bijvoorbeeld, een model zonder trend is een bijzonder geval van een model met een trend: het model zonder trend bekomen we door in het model met trend de helling gelijk aan nul te stellen.

SSR is een maat voor de voorspellingskracht van het model en SSE een maat voor de onzekerheid van het model (de ruis). Bovenstaande formule vergelijkt dus de winst in voorspellingskracht ($SSR_1 - SSR_2$) door het model complexer te maken, met overblijvende ruis. Als SSR weinig verandert in vergelijking tot de ruis in het model, dan weegt de toename in complexiteit niet op tegen de winst in kwaliteit en kunnen we kiezen voor het meest eenvoudige model. Bijvoorbeeld, voor het constante model (R-output 1) daalt de SSR met 4.3 eenheden (p-waarde = 0.14) als we het eenvoudige model uitbreiden met een lineaire trend. We kunnen bijgevolg besluiten dat de extra complexiteit van de lineaire trend weinig bijdraagt in de reductie van de ruis, en onnodig is in dit regressiemodel. Voor het trend model (R-output 2) daalt de SSR met 64.8 eenheden (p-waarde < 0.001) als we een lineaire trend toevoegen, zodat dit een waardevolle uitbreiding van het model is.

Een alternatieve en meer algemene manier is het AIC-criterium (Akaike Information Criterion). Deze maat maakt een afweging tussen het aantal parameters en de ruis. In het geval van een lineair model geldt onderstaande formule:

$$AIC = SSE + 2p \tag{15}$$

Hierin stelt p het aantal parameters voor in het model. Hoe minder het aantal parameters, hoe minder complex en dus hoe beter. Hoe lager de ruis, hoe beter de voorspellingskracht en dus ook hoe beter. Voor het AIC criterium geldt dus hoe lager, hoe beter. In R-output 1 kunnen we aflezen dat door de trend weg te laten, AIC stijgt van 7.73 naar 8.53. Volgens het AIC criterium moet dus de trend in het model behouden worden wat in tegenspraak is met de F-test. Maar we zien wel dat de stijging slechts heel beperkt is. Zolang AIC niet meer dan 2 eenheden stijgt, wordt vaak het complexere model gekozen. In het tweede geval (R-output 2) is het signaal ondubbelzinnig. AIC stijgt van 7.73 naar 24.95. We moeten de trend behouden.

Daarnaast bestaat er nog het BIC-criterium (Bayesian Information Criterion) dat ook rekening houdt met het aantal observaties (n) in de dataset:

$$BIC = SSE + p\log(n) \tag{16}$$

Een uitdieping van de verschillende wiskundige achtergronden van deze criteria zou ons hier te ver leiden, maar het achterliggende beginsel is een wiskundige vertaling van een algemeen wetenschappelijk 'zuinigheidsprincipe': we proberen altijd een zo eenvoudig mogelijke verklaring te vinden dat zo goed mogelijk de waarnemingen voorspelt (*principle of parsimony*). Hierbij wordt vaak verwezen naar Ocham, een filosoof uit de middeleeuwen, die het principe niet uitvond maar wel heel frequent toepaste:

> *Occam's razor (also written as Ockham's razor, Latin lex parsimoniae) is a principle of parsimony, economy, or succinctness used in logic and problem-solving. It states that among competing hypotheses, the one that makes the fewest assumptions should be selected (Wikipedia, 30 januari 2013).*

3.2 Het factor model: vergelijken van regressiemodellen

3.2.1 De simulatiemodellen

Hier beschouwen we de situatie waarbij de trend naargelang de regio verschilt. We bekijken twee modellen. In het eerste model (17) zijn de regressierechten evenwijdig (gelijke helling), maar de intercepts ver-

schillen wel (de eerste nulhypothese in vorig hoofdstuk). In het tweede model (18) gaat het om kruisende rechten (verschillende helling). Bij het tweede model kruisen de regressierechten. In regio B is de trend sterker en in regio C is er helemaal geen trend (0 = 0.5 – 0.5). De ruisterm voor deze modellen is dezelfde als voorheen ($\sigma_e = 1$).

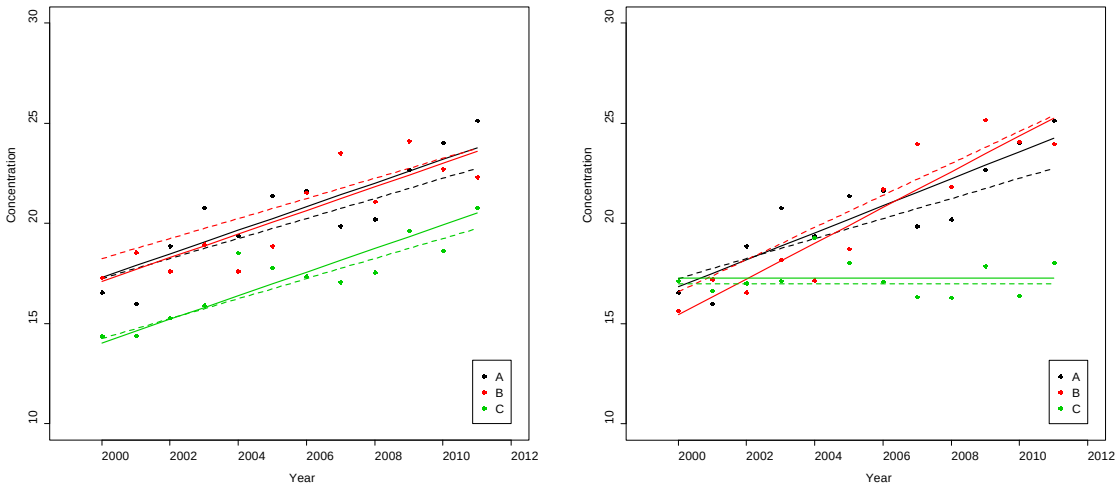
$$M_{f1}: \quad E\left[C_j\right] = \begin{cases} (20+0.0) + 0.5 \quad mJaar_j \\ (20+1.0) + 0.5 \quad mJaar_j \\ (20-3.0) + 0.5 \quad mJaar_j \end{cases} \tag{17}$$

$$M_{f2}: \quad E\left[C_j\right] = \begin{cases} (20+0.0) + (0.5+0.0) \quad mJaar_j \\ (20+1.0) + (0.5+0.3) \quad mJaar_j \\ (20-3.0) + (0.5-0.5) \quad mJaar_j \end{cases} \tag{18}$$

3.2.2 De gesimuleerde waarnemingen

Om regressierechten uit verschillende regio's onderling te vergelijken, kunnen we aparte regressievergelijkingen opstellen en vervolgens de geschatte parameters vergelijken op basis van een t-toets. Een veel directere aanpak is regio als een categorische variabele (een factor) in te voeren in een regressiemodel. We illustreren deze mogelijkheid aan de hand van de simulaties in Figuur 2. Zoals voorheen stellen de punten de gesimuleerde waarnemingen voor waarvoor we de regressierechten geschat hebben. In het geval links hebben we de software opgegeven dat de geschatte trend gelijk moest zijn wat leidde tot evenwijdige rechten, in het tweede geval was de optimalisatie van de parameters volledig vrij. In beide gevallen wordt het oorspronkelijke model vrij goed gereproduceerd.

Figuur 2: Factor model voor drie regio's (links: evenwijdige rechten; rechts: kruisende rechten). Stippellijnen = onderliggende regressiemodel waaruit de waarnemingen gesimuleerd zijn; volle lijnen = geschatte regressierechten



3.2.3 Bespreking van de R-output

<u>Selectie van het model</u>

Om statistisch te toetsen of de regressierechten significant verschillen, introduceren we regio als een factor in het model. Om te bepalen in hoeverre regio echt van belang is, zijn twee opeenvolgende (sequentiele) toetsen nodig. Eerst moeten we bepalen of de interactie tussen trend en regio significant is. Zo ja, dan bekomen we een verschillende helling per regio en kunnen we het model niet verder vereenvoudigen. Is de interactie tussen regio en trend niet significant, dan moeten we nagaan of regio op zich significant is. Indien wel, dan kunnen we het model vereenvoudigen tot evenwijdige rechten. Zo niet, dan vallen de regressierechten volledig samen en is een regressierechte voldoende ongeacht de regio.

Voor het eerste model links (R-output 3) is de interactie mJaar:Regio in de ANOVA-tabel niet significant (p = 0.47), maar Regio op zich wel (p <<< 0.05). We besluiten dat de rechten evenwijdig zijn en selecteren het model Conc ~ mJaar + Regio.

Voor het tweede model (R-output 4) is de interactieterm wel significant. We besluiten dat de rechten kruisen en selecteren het model Conc ~ mJaar + Regio + mJaar:Regio.

Voor een goed begrip is het nodig te weten dat de ANOVA-tabellen sequentieel zijn opgebouwd. De eerste lijn geeft de bijdrage van de eerste variabele in het model, de tweede lijn geeft de extra of marginale bijdrage van de volgende variabele, enzoverder. De laatste lijn gaat dus na in hoeverre de laatste term in het model nodig is. Om het model te vereenvoudigen moeten we "*backwards*" redeneren. Hier moeten we dus eerst kijken of de interactie significant is of niet. Indien wel, dan heeft het geen zin om de andere termen te evalueren. Het meest complexe model is nodig. Indien niet, dan kunnen we kijken naar de lijn erboven.

De schattingen voor de regio's voor de tweede simulatie (R-output 3)

De tabel met de coëfficiënten voor het model met evenwijdige rechten vergt enige toelichting. De eerste twee coëfficiënten zijn de klassieke regressiecoëfficiënten voor de referentieregio. De interpretatie ervan hangt af van het statistisch pakket. Tenzij we het anders bepalen, rangschikt R de regio's alfabetisch en kiest de eerste in de rangschikking als de referentie. De eerste twee regressiecoëfficiënten bepalen hier dus de regressievergelijking voor regio A: 20.5 + 0.588*cYear. De andere coëfficiënten geven het verschil met het intercept en drukken uit hoe de twee regio's verschillen van A. Voor regio B is dat -0.2 (BI: -1.2; 0.8) en voor regio C is dat -3.3 (BI: -4.2; -2.3). Het eerste BI is niet significant, het tweede BI wel. Uit de werkelijke modellen (17) en (18) blijkt dat het eerste verschil miniem is (+1) en dat het tweede verschil groter is (-3).

De aandachtige lezer zal bemerken dat het eerste BI de werkelijke waarde (+1) niet bevat (terwijl dat wel het geval is voor het tweede BI). Dat is echter niet in tegenspraak met de theorie. Met een 95 % BI zal in 5 % van de gevallen de werkelijke waarde buiten het interval vallen. Naarmate we meer intervallen berekenen zal de kans toenemen op een foutief interval dat de werkelijke waarde niet omvat. Willen we iets meer garanties dat de werkelijke waarde voor meerdere coëfficiënten binnen het interval liggen, dan is een correctie nodig.

Concreet hebben we hier simultane BI berekend volgens Tukey omdat we geïnteresseerd zijn in alle mogelijke verschillen tussen de regio's. De R-output maakt duidelijk dat de BI met een Tukey-correctie inderdaad breder zijn (meteen worden ook alle regio's met elkaar vergeleken en niet alleen met de referentie A). Bijvoorbeeld de BI voor het verschil tussen regio B en A is [-1.377,0.990] en de werkelijke waarde (1) valt er nu net buiten.

De schattingen voor de regio's voor de tweede simulatie (R-output 4)

Hier zijn er nu ook termen voor de afwijkingen ten opzichte van de helling. Zoals blijkt uit de t-toets voor de coëfficiënten, is de helling in Regio B niet significant (p-waarde = 0.13) hoger dan in Regio A (+0.218; BI: -0.0685,0.505); is de helling in Regio C significant lager dan in regio A (-0.674; BI: -0.961,-0.387). Samentellen van de helling voor A (0.673; BI: 0.4705, 0.876) geeft de schatting voor de helling. We stellen inderdaad vast dat voor regio C de totale helling ongeveer nul is: 0.673 - 0.674 = -0.001.

We kunnen analoog aan het voorgaande model de verschillen voor het intercept bekijken. Maar, aangezien de rechten kruisen, geldt het verschil alleen halfweg de waarnemingsperiode (2005.5). Door het verschil in trend hangen de verschillen tussen de regressierechten af van het Jaar. Aanvankelijk (in 2000) was de concentratie in alle regio's ongeveer even hoog, maar over de jaren heen is er een sterk verschil gegroeid tussen regio C en regio A en B. Halfweg verschillen A & B niet wezenlijk, maar het verschil met regio C is telkens significant.

De ruisterm

In vergelijking met de vorige analyse is het BI voor de ruisterm nauwer geworden (0.8 – 2.1). Dat komt omdat we in totaal drie keer meer gegevens hebben om de ruis te schatten (12 meetgegevens uit 3 regio's). Door alle regressierechten samen te berekenen, worden alle gegevens gecombineerd (gepooled) voor de schatting van de ruis. Hadden we de regressierechten apart geschat, dan waren er drie verschillende schattingen van de ruis geweest, met een veel bredere BI.

Indien de ruis naargelang de regio zou verschillen, is evenwel een aparte schatting per regio noodzakelijk. Maar de figuren geven geen indicatie dat de variatie rond de regressierechte verschilt naargelang de re-

gio. We kunnen dat ook formeel toetsen door een model te creëren met verschillende varianties voor de ruis en te vergelijken met een simpeler model met een variantie, maar dat zou ons hier te ver leiden.

R-output 3: resultaten voor de simulatie in Figuur 2 – links (evenwijdige rechten)

Analysis of Variance Table → mJaar + Regio

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
mJaar	1	148.4	148.4	105.17	2.5e-11
Regio	2	80.7	40.4	28.60	1.1e-07
mJaar:Regio	2	2.2	1.1	0.77	0.47
Residuals	30	42.3	1.4		

Coefficients after simplification (interaction dropped)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	20.546	0.3404	60.353	1.54e-34
mJaar	0.588	0.0569	10.329	1.02e-11
RegioB	-0.193	0.4814	-0.401	6.91e-01
RegioC	-3.268	0.4814	-6.789	1.13e-07

Confidence limits (uncorrected)

	Estimate	2.5 %	97.5 %
(Intercept)	20.546	19.852	21.239
mJaar	0.588	0.472	0.704
RegioB	-0.193	-1.174	0.787
RegioC	-3.268	-4.249	-2.288

Linear Hypotheses (simultaneous Tukey)

	Estimate	Std. Error	t value	Pr(>|t|)	lwr	upr
B - A == 0	-0.193	0.481	-0.40	0.92	-1.377	0.990
C - A == 0	-3.268	0.481	-6.79	<1e-04	-4.452	-2.085
C - B == 0	-3.075	0.481	-6.39	<1e-04	-4.259	-1.892

Sigma: 1.18 (0.763,2.063)

R-output 4: resultaten voor de simulatie in Figuur 2 – rechts (kruisende rechten)

Analysis of Variance Table → Conc ~ mJaar + Regio + mJaar:Regio

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
mJaar	1	116.6	116.6	82.7	4.0e-10
Regio	2	80.7	40.4	28.6	1.1e-07
mJaar:Regio	2	61.9	31.0	21.9	1.3e-06
Residuals	30	42.3	1.4		

Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	20.546	0.3429	59.919	8.65e-33
mJaar	0.673	0.0993	6.780	1.62e-07
RegioB	-0.193	0.4849	-0.398	6.93e-01
RegioC	-3.268	0.4849	-6.740	1.80e-07
mJaar:RegioB	0.218	0.1405	1.555	1.30e-01
mJaar:RegioC	-0.674	0.1405	-4.800	4.09e-05

Confidence limits (uncorrected; interactions only)

	Estimate	2.5 %	97.5 %
mJaar	0.673	0.4705	0.876
mJaar:RegioB	0.218	-0.0685	0.505
mJaar:RegioC	-0.674	-0.9612	-0.387

Linear Hypotheses (simultaneous confidence intervals ~ Tukey)

(lwr: lower limit & upr: upper limit)

	Estimate	lwr	upr
B - A == 0	-0.193	-1.389	1.003
C - A == 0	-3.268	-4.464	-2.073
C - B == 0	-3.075	-4.271	-1.879

Sigma: 1.19 (0.759,2.122)

3.3 Mixed model regressie

3.3.1 De simulatiemodellen

Wanneer we de concentratie van een bepaalde stof opvolgen in meerdere waterlopen, is het weinig realistisch te veronderstellen dat de concentratie overal een gelijk niveau heeft en/of een gelijke trend zal vol-

gen. Daarom is het basismodel (13) te beperkt. We zouden net als voor ecoregio voor elke meetplaats apart een regressierechte kunnen schatten, maar dat zou tot te veel parameters leiden. Daarenboven zijn we in een landelijk meetnet eerder geïnteresseerd in de globale evolutie en veel minder in wat zich lokaal afspeelt. De lokale metingen zijn daarenboven momentopnames op basis waarvan we weinig kunnen afleiden over de lokale evolutie. Om al deze redenen, is het zinvoller een ander model te formuleren.

Een eerste uitbreiding is het "*random intercept*" model (19) waarbij we toelaten dat elke waterloop een eigen intercept heeft. Het intercept varieert volgens een normale verdeling met spreiding 2.5 (de standaardafwijking van b_{0k}). We mogen bijgevolg verwachten dat het intercept (halfweg 2005) in 95 % van de gevallen zal liggen tussen 15 en 25 (= 20 ± 2 x 2.5).

$$M_{m1}: \quad C_{jk} = (20 + b_{0k}) + 0.5\, mJaar_j + \varepsilon_{jk} \qquad \begin{cases} b_{0k} \overset{iid}{\propto} N(0, 2.5) \\ \varepsilon_{jk} \overset{iid}{\propto} N(0,1) \end{cases} \tag{19}$$

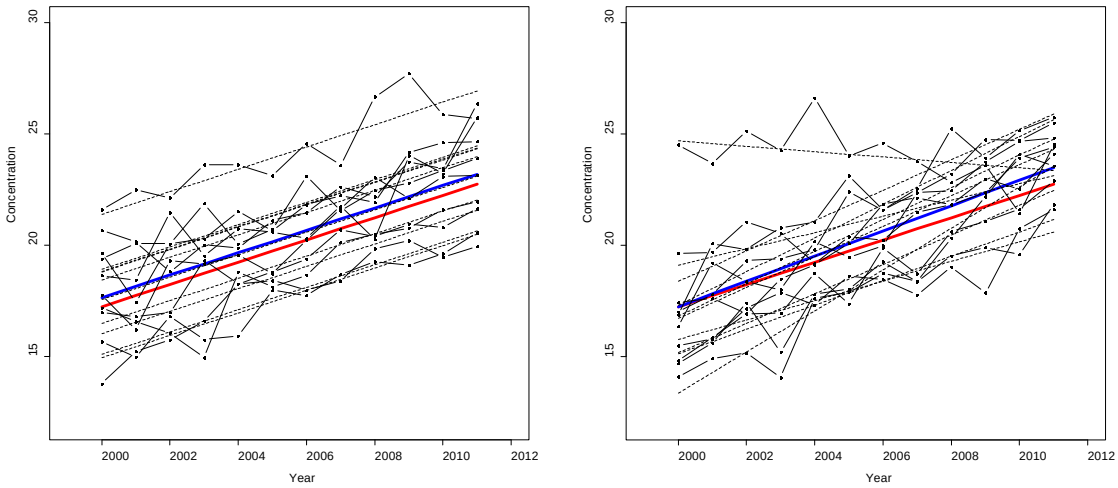
We kunnen het model verder veralgemenen door ook toe te laten dat de helling variabel kan zijn binnen bepaalde grenzen. Volgens het "*random slope*" model (20) verwachten we dat 95 % van de waarden zal liggen in het interval 0.5 ± 2 x 0.3 = -0.1 tot 1.1.

$$M_{m2}: \quad C_{jk} = (20 + b_{0k}) + (0.5 + b_{1k}) mJaar_j + \varepsilon_{jk} \qquad \begin{cases} b_{0k} \overset{iid}{\propto} N(0, 2.5) \\ b_{1k} \overset{iid}{\propto} N(0, 0.3) \\ \varepsilon_{jk} \overset{iid}{\propto} N(0,1) \end{cases} \tag{20}$$

3.3.2 De simulatie + geschatte rechten

Figuur 3 toont een situatie met meerdere waterlopen. De figuur links is afkomstig van het "*random intercept*" model. De volle lijnen verbinden de waarnemingen, de stippellijnen zijn de geschatte regressierechten. In de figuur links hebben we aan de software opgelegd de helling te fixeren op een gemeenschappelijke waarde. In de figuur rechts was er deze voorwaarde niet. Hierdoor kan niet alleen de uitzonderlijke trend eenvoudig gemodelleerd worden, maar krijgen ook de andere waterlopen een eigen rechte. De rechten lopen kriskras door elkaar.

Figuur 3: Het mixed model met random intercept (links: kleine verschillen, rechts: grote verschillen tussen meetlocaties). Rode lijn = globale trend (bekend omdat het om een simulatie gaat); blauwe lijn = globale trend geschat uit de waarnemingen; stippellijnen = geschatte (random) trend per meetplaats



3.3.3 De statistische analyse (R-output)

Mixed model regressie laat toe om deze patronen te analyseren door zowel het intercept als de helling als een normaal verdeelde variabele te modelleren. Net zoals bij modellen met een factor kunnen we testen of de introductie van deze termen significant zijn of niet. In de figuur links was het besluit dat een random intercept voldoende was (al rechten evenwijdig), in de figuur rechts was de uitkomst van de analyse een random slope model (kruisende rechten). We bespreken het verloop van de analyse aan de hand van R-output 5 en R-output 6.

Selectie van het optimale model (‘*comparison of the models’* & ‘analysis of variance’)

Het eerste luikje gaat na of het *random slope* model (het meest complexe model met een variabele helling per waterloop) significant verschilt van het *random intercept* model (het meer eenvoudige model met alleen een variabel intercept). Hierbij geeft R drie criteria: AIC, BIC en LRT.

- LRT (likelihood ratio test) is een statistische toets ingebed in de klassieke statistische theorie en is direct verwant met de veralgemeende F-toets (15) om modellen te vergelijken. LRT toetst of het eenvoudigere model significant verschilt van het meer complexere en geeft een p-waarde. Indien het verschil significant is, kiezen we voor het meest complexe model. Zo niet, kunnen we het model verder vereenvoudigen. Een belangrijke beperking van LRT is dat modellen “*genest*” moeten zijn, d.w.z. het ene model moet een bijzonder geval zijn van het andere model. Deze beperking is er niet voor de twee andere criteria die uit de informatietheorie stammen.

- AIC (Akaike Information Criterion) en BIC (Bayesian Information Criterion) maken in een getal een afweging tussen nauwkeurigheid en complexiteit. Hoe meer parameters in het model (als maat voor de complexiteit), hoe sterker de afstraffing want een complex model zal vanzelf beter bij de gegevens aansluiten. Bij AIC is de afstraffing twee keer het aantal parameters (2*npar); bij BIC wordt npar vermenigvuldigd met het de logaritme van het aantal waarnemingen (log(nobs)*npar). Naarmate het aantal waarnemingen toeneemt wordt het BIC criterium strenger voor complexe modellen. Meestal wordt gewerkt met AIC. Hoe lager AIC (en BIC) hoe beter het model. Omdat we bij statistische modelbouw naar eenvoud streven, is een algemene regel dat kleine verschillen voor AIC (minder dan 2) er weinig toe doen.

Uit alle criteria van R-output 5 blijkt dat het random intercept model het beste model is: AIC, BIC zijn er het kleinst en de p-waarde van LRT = 1. We kunnen het model dus vereenvoudigen en de random slope term in het model laten vallen, wat in overeenstemming is met het model (19).

Voor de tweede simulatie (R-output 6) gaat de voorkeur uit naar het meest complexe model. De AIC (BIC) van het random slope model ligt beduidend lager dan het random intercept model. We moeten bijgevolg de random slope in het model behouden wat opnieuw conform is met het simulatiemodel (20).

Net zoals bij de vorige modellen moeten we ook nagaan of de trend significant is. Een technische complicatie bij mixed models is dat de toets alleen geldig is indien de schatting van de parameters gebeurd is met de maximum likelihood (ML) methode. De standaardmethode is echter restricted maximum likelihood (REML) en we moeten expliciet de methode veranderen. Waarom REML niet tot goede resultaten leidt, zou ons hier te ver leiden, maar we vermelden het hier voor de volledigheid. Voor beide simulaties is mJaar hoog significant en we moeten bijgevolg de term in het model laten staan, geen vereenvoudiging is mogelijk.

De schatting van fixed effects (trend)

De intercept (waarde halfweg de waarnemingsperiode) is voor de eerste simulatie (R-output 5) 20.43 (BI: 19.1 – 21.7) en voor de tweede simulatie 20.37 (BI: 19.1 – 21.6). De twee schattingen zijn nagenoeg gelijk omdat we telkens met dezelfde ruis gewerkt hebben. Blijkbaar heeft het modeltype (random slope of niet) weinig invloed op de schattingen, wellicht omdat we de waarnemingen gecentreerd hebben in het midden van de observatieperiode.

De enige regressieterm is mJaar waarmee we de jaarlijkse trend schatten. Voor de eerste simulatie (R-output 5) valt de geschatte trend 0.504 (BI: 0.458,0.550) nagenoeg samen met de werkelijke waarde (0.5). Ook voor de tweede simulatie (R-output 6) sluit de waarde = 0.567 (BI: 0.375,0.759) goed aan bij het oorspronkelijk simulatiemodel.

De schatting van de ruistermen en random effects

Omdat de betekenis van de termen verschilt naargelang het model, bespreken we de resultaten per model. Telkens vermelden we tussen haakjes de werkelijke waarde (T = true value) en geschatte waarde (E = estimated value) samen met de BI zodat we vlotter kunnen inschatten in hoeverre het model in staat is de werkelijke situatie te reproduceren.

De schatting van de ruis en random effects voor het random intercept model (R-output 5)

Bij het random intercept model stelt de ruis (T: 1; E: 0.88; BI: 0.77,1.00) de fluctuatie voor van de waarnemingen ten opzichte van de trendlijn binnen elke waterloop. Het *random intercept* (T: 2.5; E: 2.04; BI: 1.27-3.25) modelleert de variatie van de evenwijdige rechten ten opzichte van de gemiddelde regressielijn. Er is bijgevolg een goede overeenstemming tussen het geschatte en werkelijke model.

De schatting van de ruis en random effects voor het random slope model (R-output 6)

Bij het random slope model kruisen de rechten. Het *random intercept* (T: 2.5; E: 1.97; BI: 1.23,3.15) modelleert bijgevolg alleen de variatie in het jaartal waarmee het intercept correspondeert (2005.5). De interpretatie van de ruis (T: 1; E: 0.85; BI: 0.74,0.98) blijft wel gelijk: de variatie rondom de individuele trendlijnen. Het *random slope* effect (T: 0.3; E: 0.298; BI: 0.183,0.486) drukt uit hoe sterk de trend varieert ten opzichte van de gemiddelde trend (T: 0.5; E: 0.57; BI:0.38,0.76).

R-output 5: Statistische analyse bij Figuur 3 – links (random intercept model)

Comparison of models (method='REML') → Random intercept model

Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
Random slope	5	364	377	-177			
Random intercept	4	362	373	-177	1 vs 2	7.88e-08	1

Analysis of Variance Table (method='ML') → keep mJaar in model

	numDF	denDF	F-value	p-value
(Intercept)	1	109	991	<.0001
mJaar	1	109	471	<.0001

Coefficients of simplified model without interaction

- Fixed effects: Response ~ mJaar

	Value	Std.Error	DF	t-value	p-value
(Intercept)	20.4	0.649	109	31.5	0
mJaar	0.5	0.023	109	21.7	0

- Random effects: Formula: ~1 | WL

	(Intercept)	Residual
StdDev:	2.04	0.877

Confidence limits

	lower	est.	upper
(Intercept)	19.143	20.429	21.71
mJaar	0.458	0.504	0.55
sigma	0.768	0.877	1.00
sd((Intercept))	1.274	2.036	3.25

R-output 6: Statistische analyse bij Figuur 3 – rechts (random slope model)

Comparison of models (method='REML') → Random slope model

Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
Random slope	5	383	397	-186			
Random intercept	4	452	463	-222	1 vs 2	71.1	<.0001

Analysis of Variance Table (method='ML') → keep mJaar in model

	numDF	denDF	F-value	p-value
(Intercept)	1	109	1152	<.0001
mJaar	1	109	37	<.0001

Coefficients of simplified model without interaction

- Fixed effects: Yv ~ mJaar

	Value	Std.Error	DF	t-value	p-value
(Intercept)	20.37	0.600	109	33.9	0
mJaar	0.57	0.093	109	6.1	0

- Random effects: Formula: ~ mJaar | WL (Structure: Diagonal)

```
        (Intercept) mJaar Residual
StdDev:        1.97 0.298    0.852
```

Confidence limits

```
                 lower    est.  upper
(Intercept)    19.129 20.372 21.616
mJaar           0.375  0.567  0.759
sigma           0.742  0.852  0.979
sd((Intercept)) 1.231  1.969  3.147
sd(mJaar)       0.183  0.298  0.486
```

4 Voorstelling en verkenning van de gegevens

4.1 Inleiding

In de twee hierna volgende hoofstukken zullen we twee tijdsreeksen van het waterbodemmeetnet op twee verschillende manieren analyseren. In het eerste geval zullen we de tijd beschouwen als een continue variabele, in het tweede geval als een categorische variabele (factor). De meetlocaties worden immers om de vier jaar bezocht en we kunnen bijgevolg de waarnemingen groeperen per meetcyclus en nagaan hoe de concentraties evolueren per cyclus.

Hieronder maken we een eerste verkennende grafische analyse van de meetgegevens. Deze EDA ("exploratory data analysis") is een wezenlijk onderdeel van elke statistische analyse. Een grondige kennismaking met de gegevens is noodzakelijk om voldoende voeling met de gegevens te krijgen. Dat is belangrijk voor de interpretatie van de resultaten. Een statistische toets of model reduceert de gegevens tot enkele kengetallen die misleidend kunnen zijn zonder kennis van de originele data. Uitbijters (al dan niet in groep) kunnen de resultaten sterk vertekenen. Een belangrijke functie van EDA is dan ook kwaliteitscontrole van de gegevens.

Voor de analyse hier zijn de waarden onder de bepaalbaarheidsgrens al aangepast door ze te vervangen door de helft ervan. Ook zijn onwaarschijnlijkheden al uit de dataset verwijderd. Nadere details zijn te vinden in een technisch rapport dat we als appendix hebben toegevoegd bij het huidige rapport (Jansen, 2012). We gaan hier bijgevolg uit van een opgeschoonde dataset. Toch is een verkennende analyse nog zinvol om de globale patronen te herkennen die we zullen aftoetsen met de statistische modelbouw.

Aangezien concentraties (en abundantiematen in het algemeen) vaak lognormaal verdeeld zijn, verkennen we de gegevens meteen in de logschaal. Ook verloopt de grafische verkenning van concentratiegegevens gemakkelijker in de logschaal dan in de originele schaal omdat het bereik van de concentraties meestal over meerdere grootteordes loopt en uitbijters minder het algemene beeld verstoren.

4.2 Cadmium

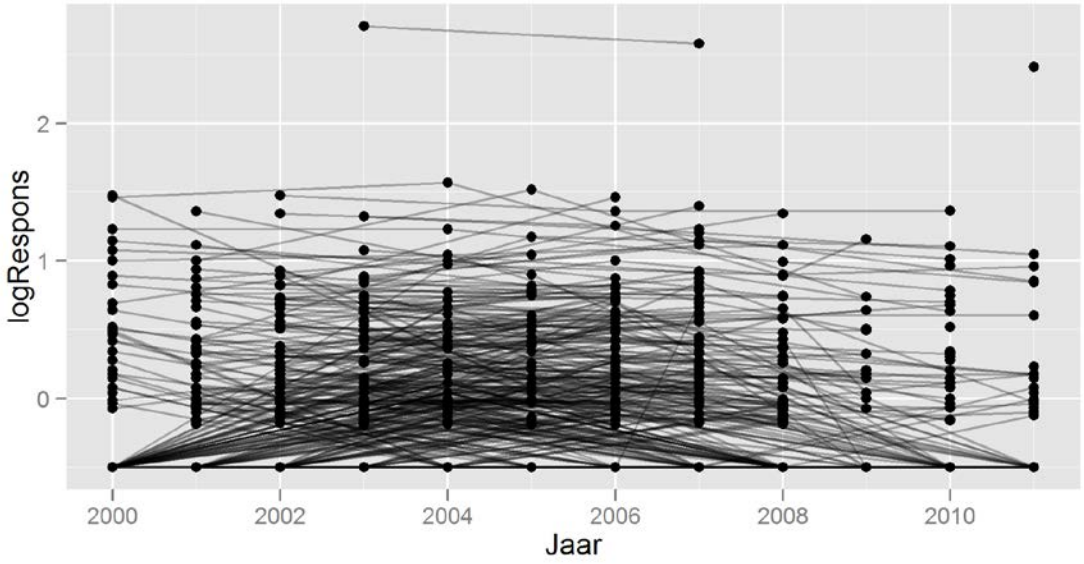
4.2.1 Grafische voorstelling met tijd als een continue variabele

Figuur 4 toont voor cadmium de jaarprofielen per meetplaats. We zien een grote variatie tussen de meetplaatsen. Het zal bijgevolg zeker nodig zijn een random effect voor de meetplaatsen in het model op te nemen.

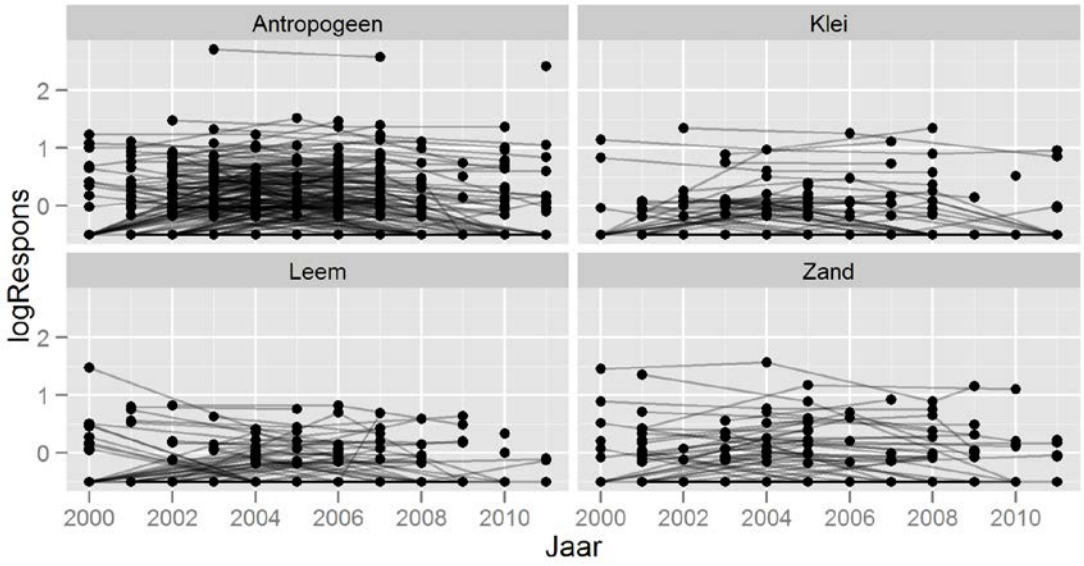
Ondanks de wirwar van lijnen tekent zich een globaal patroon af: in de beginjaren is er een stijging van de concentratie, vervolgens vlakt de curve af en begint te dalen. Een niet-lineair model zal wellicht nodig zijn dat we kunnen benaderen tot een tweedegraadsvergelijking of zelfs derdegraadsvergelijking.

Een uitsplitsing per ecoregio (Figuur 5) verandert het patroon niet. In alle ecoregio's tekent zich min of meer dezelfde trend af. Moeilijker in te schatten is of er systematische verschillen zijn tussen de regio's.

Figuur 4: Jaarprofielen per meetplaats voor cadmium



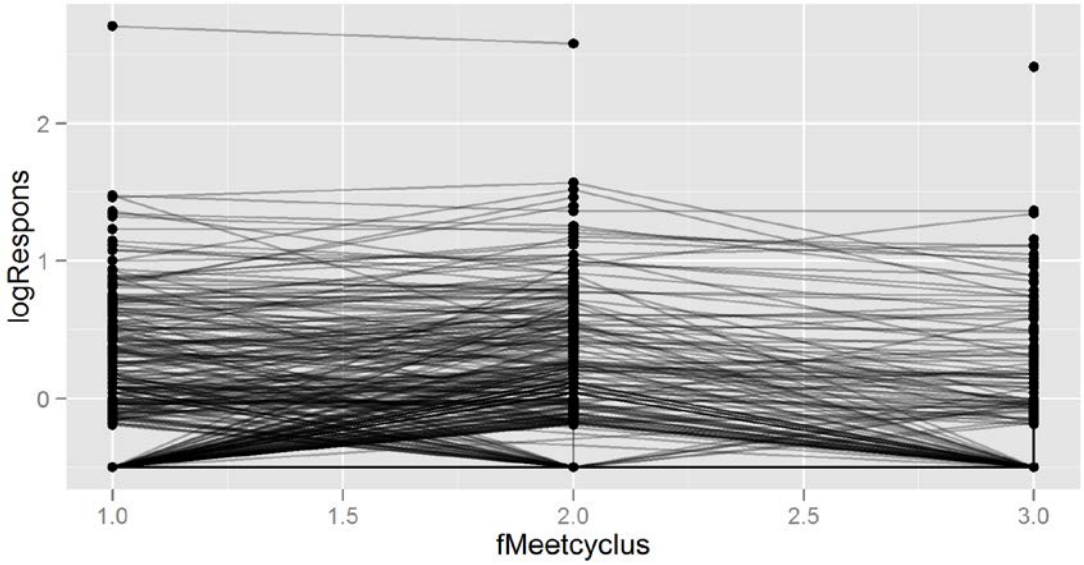
Figuur 5: Jaarprofielen per meetplaats voor cadmium, opgesplitst per ecoregio



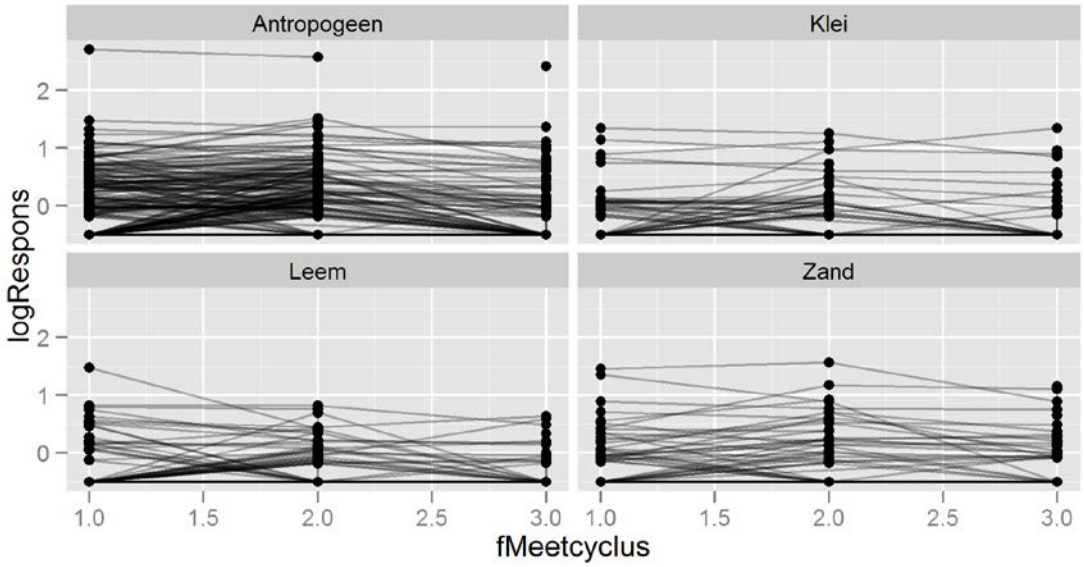
4.2.2 Grafische voorstelling met tijd als een categorische variabele

In Figuur 6 bekijken we de evolutie per meetcyclus. Hiermee verliezen we een stuk informatie, maar aan de andere kant kunnen we de globale trend beter zien. Binnen een cyclus worden (in principe) alle meetplaatsen precies een keer bezocht. De figuur bevestigt het globale patroon: van cyclus 1 naar cyclus 2 neemt de concentratie toe, het verschil tussen cyclus 2 en cyclus 3 is miniem; er is eerder een daling. Een opsplitsing per ecoregio verandert het patroon niet (Figuur 7).

Figuur 6: Cyclusprofielen per meetplaats voor cadmium



Figuur 7: Cyclusprofielen per meetplaats voor cadmium, opgesplitst per ecoregio



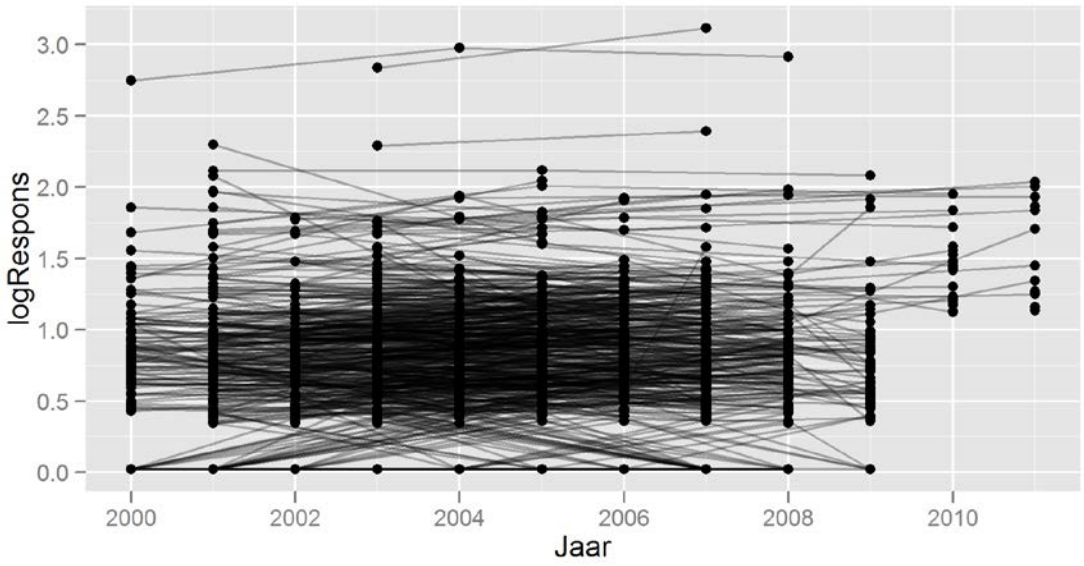
4.3 Arseen

4.3.1 Grafische voorstelling met tijd als een continue variabele

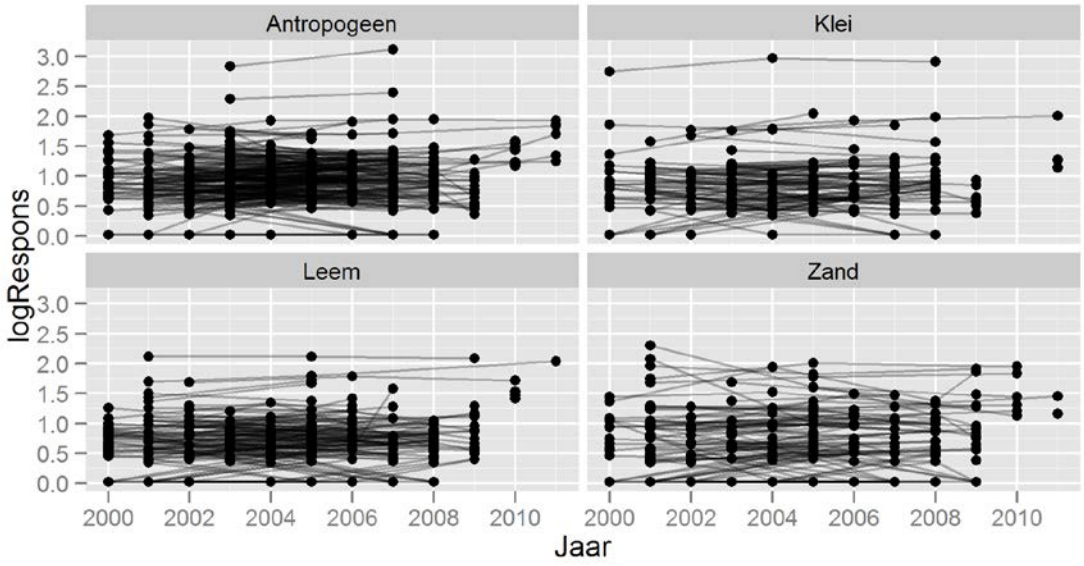
Figuur 8 toont voor arseen de jaarprofielen per meetplaats en Figuur 9 geeft de uitsplitsing per ecoregio. We zien een grote variatie tussen de meetplaatsen. Het zal bijgevolg zeker nodig zijn een random effect voor de meetplaatsen in het model op te nemen.

Een globale trend tekent zich hier niet duidelijk af, maar wellicht zal een lineair (eerste orde) model voldoende zijn.

Figuur 8: Jaarprofielen per meetplaats voor arseen



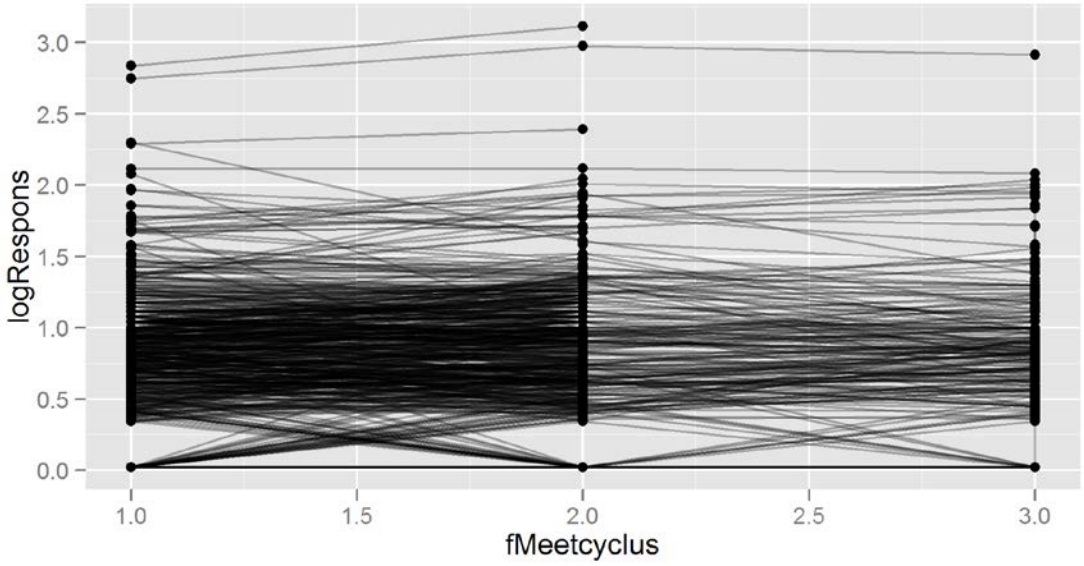
Figuur 9: Jaarprofielen per meetplaats voor arseen, opgesplitst per ecoregio



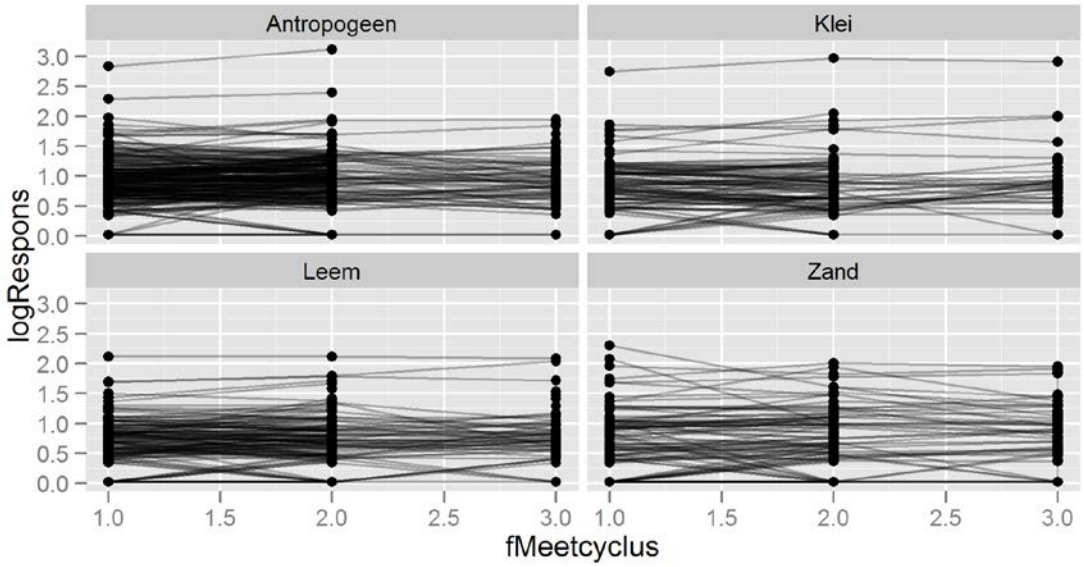
4.3.2 Grafische voorstelling met tijd als een categorische variabele

Figuur 10 en Figuur 11 geven de cyclusprofielen. Ten opzicht van de eerste meetcyclus lijken de waarden in meetcyclus 2 en 3 iets hoger te liggen.

Figuur 10: Cyclusprofielen per meetplaats voor arseen



Figuur 11: Cyclusprofielen per meetplaats voor arseen, opgesplitst per ecoregio



5 Voorbeeld 1: trendanalyse voor cadmium

5.1 Jaar van opname als tijdsvariabele

5.1.1 Het startmodel

Uit de verkennende analyse kwam naar voor dat de relatie met de tijd wellicht niet lineair is. Daarom introduceren we hier een derdegraadsvergelijking om de trend te modelleren (notatie: cJaar + cJaar² + cJaar³, waarbij **cJaar = Jaar – 2000**, zodat het **referentiejaar** samenvalt met de start van de studie in 2000).

Ook waren de verschillen tussen de meetplaatsen groot. Gezien de korte tijdsreeks (3 metingen per meetplaats) en de complexiteit van de trend (derdegraadsvergelijking) zijn de mogelijkheden voor het random effects gedeelte beperkt, en modelleren we deze verschillen tussen meetplaatsen enkel als toevallige normale fluctuaties ten opzichte van het intercept (notatie: 1 | Meetplaats). Een uitgebreidere random effects structuur illustreren we in het voorbeeld van arseen waar er een lineaire trend is (sectie 6).

Tenslotte onderzoeken we de rol van Ecoregio. De impact zou alleen op het intercept kunnen zijn (evenwijdige regressielijnen), maar kan ook de vorm van de relatie beïnvloeden. In het eerste geval is er alleen een verschuiving van het intercept (notatie: functie curve + Ecoregio), in het tweede geval is er een interactie tussen de curve en ecoregio (notatie: functie curve : Ecoregio).

Voegen we al deze potentiële factoren samen, dan krijgen we volgend startmodel.

logCd ~ 1|Meetplaats + (cJaar + cJaar² + cJaar³) + Ecoregio
+ (cJaar + cJaar² + cJaar³) : Ecoregio

De statistische modelbouw heeft als doel en biedt technieken aan om het startmodel te vereenvoudigen tot de essentie. Hierbij is het mogelijk dat tijdens de analyse nieuwe onvoorziene factoren naar boven komen die we in het model moeten opnemen. Maar het finale model is zo eenvoudig mogelijk, bevat alleen de significante factoren en sluit tegelijk zo nauw mogelijk aan bij de gegevens (*principle of parsimony*).

5.1.2 Keuze van het model (modelreductie en modelverfijning)

Een eerste belangrijke stap is na te gaan of we het model niet kunnen vereenvoudigen. We kunnen op basis van onderstaande ANOVA-tabel nagaan welke termen significant zijn. Niet significante termen mogen we elimineren. Hierbij is een stapsgewijze aanpak aanbevolen, omdat door eliminatie van een term de impact van de andere termen kan veranderen (stepwise regression).

R-output 7: ANOVA-tabel voor de trendanalyse van cadmium – cJaar

```
                          numDF denDF  F-value  p-value
(Intercept)                  1   586  3.333138  0.0684
poly(cJaar, 3)               3   586 17.953895  <.0001
Ecoregio                     3   441 12.484227  <.0001
poly(cJaar, 3):Ecoregio      9   586  2.139335  0.0247
```

Bekijken we de ANOVA tabel in R-output 7, dan zijn alle termen significant en lijkt het alsof we het model niet verder kunnen vereenvoudigen. De derdegraadsvergelijking zit echter volledig verscholen in de term poly(cJaar, 3) en het is niet rechtstreeks af te lezen of deze vereenvoudigd kan worden naar een tweedegraads- of zelfs eerstegraadsvergelijking. Daarvoor moeten we deze modellen expliciet fitten en onderling vergelijken.

R-output 8: Modelselectie voor cadmium – cJaar (MC0 = derdegraadsvergelijking, MC1 = tweedegraadsvergelijking en MC2 = eerstegraadsvergelijking)

```
    Model df      AIC      BIC    logLik  Test L.Ratio p-value
MC0     1 18 1101.650 1190.747 -532.8249
MC1     2 14 1102.006 1171.304 -537.0031 1 vs 2  8.35622  0.0794
MC2     3 10 1151.651 1201.150 -565.8258 2 vs 3 57.64544  <.0001
```

Modelselectie kan op 2 manieren gebeuren, enerzijds op basis van de AIC-waarden (model met de laagste AIC waarde geeft de beste fit), anderzijds op basis van de p-waarde bij de likelihood ratio test (L.Ratio). Beide criteria duiden nu echter een verschillend model aan. Volgens de AIC-waarden is het model met de derdegraadsvergelijking (MC0, AIC = 1101.650) het beste, volgens de p-waarde kan dit echter vereenvoudigd worden naar een tweedegraadsvergelijking (p-waarde = 0.0794). Verder vereenvoudigen

naar een eerstegraadsvergelijking is sowieso geen optie (p-waarde <.0001 en hogere AIC). Welk model er uiteindelijk gekozen wordt, blijft een subjectieve beslissing. Een regel die vaak gehanteerd wordt bij modelselectie is dat wanneer AIC-waarden minder dan 2 van mekaar verschillen, men best voor het eenvoudigste model kiest. Deze minimale verbetering in AIC weegt immers niet op tegen de groeiende complexiteit van het model, en de daarmee samenhangende moeilijkere interpretatie. Vandaar dat we verder zullen gaan met het model met de tweedegraadsterm (MC1).

R-output 9: ANOVA-tabel voor het vereenvoudigde model van cadmium – cJaar

```
                          numDFd enDF  F-value p-value
(Intercept)                  1   590  3.347891  0.0678
poly(cJaar, 2)               2   590 25.597614  <.0001
Ecoregio                     3   441 12.667262  <.0001
poly(cJaar, 2):Ecoregio      6   590  2.083408  0.0534
```

In een volgende stap worden we opnieuw geconfronteerd met hetzelfde probleem. Wat doen we met de interactieterm poly(cJaar, 2):Ecoregio? Deze term heeft een p-waarde gelijk aan 0.0534, en de AIC-waarde van het model zonder deze interactieterm bedraagt 1102.539 (t.o.v. 1102.006 voor het model MC1). Opnieuw weegt de complexiteit van het model met interactieterm niet op tegen de verbetering van de fit. Vandaar verwijderen we ook deze interactie.

R-output 10: ANOVA-tabel voor het finale model van cadmium – cJaar

```
                numDF denDF  F-value p-value
(Intercept)        1   596  3.35483  0.0675
poly(cJaar, 2)     2   596 25.26720  <.0001
Ecoregio           3   441 12.66399  <.0001
```

De resterende termen zijn significant (R-output 10), en zo bekomen we als finale model die alleen nog factoren bevat die significant zijn.

logCd ~ 1|Meetplaats + (cJaar + cJaar²) + Ecoregio

5.1.3 De parameterschattingen

We kunnen de parameterschattingen voor het finale model terugvinden in R-output 11 tot en met R-output 14. We bespreken eerst het parabolische trendmodel voor de referentieregio (Antropogeen) en onderzoeken vervolgens het effect van regio op het model. Tenslotte bekijken we de standaardafwijkingen in het model.

(a) <u>Het parabolische trendmodel</u>

Voor een interpretatie van de tweedegraadsterm werd de poly(cJaar, 2) term in R-output 11 nu expliciet uitgeschreven als cJaar en I(cJaar^2).

R-output 11: Parameterschattingen voor het fixed effects gedeelte van het finale model voor cadmium – cJaar

```
Fixed effects: logRespons ~ cJaar + I(cJaar^2) + Ecoregio
                    Value  Std.Error  DF   t-value p-value
(Intercept)  -0.0014095 0.03841839 596 -0.036689  0.9707
cJaar         0.0679803 0.01064056 596  6.388795  0.0000
I(cJaar^2)   -0.0070501 0.00100952 596 -6.983585  0.0000
EcoregioKlei -0.2408077 0.06335374 441 -3.801003  0.0002
EcoregioLeem -0.3184807 0.05833955 441 -5.459088  0.0000
EcoregioZand -0.2026120 0.05964304 441 -3.397077  0.0007
```

Uit R-output 11 kunnen we volgende kwadratische vergelijking voor de verwachte waarde van logconcentratie van cadmium (Cd) afleiden:

E[logCd] = -0.0014 + 0.0679 cJaar – 0.007 cJaar^2

Deze vergelijking stelt in de log-schaal een parabool voor met een top naar boven (aangezien de kwadratische term negatief is) en bereikt het maximum in cJaar = 4.85, want daar is de eerste afgeleide van de functie gelijk aan nul zoals blijkt uit onderstaande vergelijking:

d(E[logCd])/d(cJaar) = 0.0679 – 0.007 x 2 cJaar = 0 → cJaar = 0.0679/0.014 = 4.85

In 2000 (cJaar = 0) is E[logCd] -0.0014 (≈ 1.00 in de oorspronkelijke schaal); stijgt tot in 2005 (cJaar = 4.85) en bereikt daar de maximale waarde 0.16 (≈ 1.45). Hierna daalt de parabool tot -0.1015 (≈ 0.79) in

2011. Op het einde van de waarnemingsperiode ligt de concentratie dus ongeveer 20 % lager dan bij de start. Deze resultaten bevestigen het beeld dat al bij de verkennende analyse naar voren kwam, namelijk dat de concentratie de eerste jaren licht toeneemt, vervolgens afvlakt om daarna terug te dalen. We verwijzen hier ook naar Figuur 20 met de parabool getekend voor heel het bereik.

(b) Het effect van de regio

De (intercept) term kan geïnterpreteerd worden als de logconcentratie in het jaar 2000 voor Ecoregio Antropogeen (dat als referentiegroep genomen wordt). De schattingen in de laatste drie regels van R-output 11 geven het verschil in logconcentratie tussen ecoregio Antropogeen en ecoregio's respectievelijk Klei, Leem en Zand. We kunnen hieruit besluiten dat de concentratie cadmium (zowel in log-schaal als in gewone schaal) hoger is in de ecoregio Antropogeen dan in de andere regio's (vermits alle parameterschattingen negatief zijn).

Concreet kunnen we in het maximum (cJaar = 4.85) de maximale concentraties voor de andere regio's berekenen door bij 0.16 (de waarde in Antropogeen) -0.24 (Klei), -0.32 (Leem) en -0.20 (Zand) op te tellen. Hierdoor bekomen we respectievelijk: -0.08 (≈ 0.83 in de oorspronkelijke schaal), -0.16 (≈ 0.69) en -0.04 (≈ 0.91). Deze getallen kunnen we ook uit Figuur 20 afleiden. Ten opzichte van 1.45 (concentratie in de oorspronkelijke schaal in het antropogeen), liggen de concentraties bijgevolg 43 %, 52 % en 37 % lager in de Klei, Leem en Zandregio. Deze procentuele verschillen blijven gelijk ongeacht het jaartal waarin we ze berekenen.

R-output 12: Alle paarsgewijze verschillen in logconcentratie tussen de verschillende ecoregio's volgens het finale model voor cadmium (Tukey methode)

```
                          Estimate Std. Error z value Pr(>|z|)
Klei - Antropogeen == 0 -0.24081    0.06335  -3.801  < 0.001 ***
Leem - Antropogeen == 0 -0.31848    0.05834  -5.459  < 0.001 ***
Zand - Antropogeen == 0 -0.20261    0.05964  -3.397  0.00401 **
Leem - Klei == 0        -0.07767    0.07393  -1.051  0.71549
Zand - Klei == 0         0.03820    0.07495   0.510  0.95597
Zand - Leem == 0         0.11587    0.07075   1.638  0.35242

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

We kunnen ook expliciet alle andere paarsgewijze verschillen tussen de ecoregio's opvragen. Bij deze testen wordt de p-waarde gecorrigeerd voor het aantal hypotheses dat getest wordt. De Tukey correctie – gekend van de ANOVA-theorie voor de vergelijking van gemiddelden – is hier de aangewezen methode. Het gaat hier om de hypotheses dat elk van de paarsgewijze verschillen gelijk is aan 0 tegenover de alternatieve hypothese dat deze verschillen niet gelijk zijn aan 0. Uit R-output 12 concluderen we dat alleen de verschillen tussen Antropogeen en de andere ecoregio's (Leem, Klei, Zand) significant zijn.

(c) De standaardafwijkingen (random effect en de ruisterm)

Naast de parameterschattingen van de fixed effects is het ook van belang om de output voor de random effects in wat meer detail te bestuderen (R-output 13).

R-output 13: Parameterschattingen voor het random effects gedeelte van het finale model voor cadmium – cJaar

```
Random effects:
 Formula: ~1 | Meetplaats
        (Intercept)  Residual
StdDev:   0.4115555 0.2807594
```

De standaarddeviatie van het random intercept (σ_0=0.412) is groter dan de residuele standaard deviatie (σ_e=0.281). Een groot gedeelte van de variabiliteit in de gegevens is te wijten aan verschillen tussen de meetplaatsen. Binnen een meetplaats zijn de variaties minder groot.

R-output 14 geeft nog betrouwbaarheidsintervallen (BI) voor de schattingen van beide σ's. Door het grote aantal meetplaatsen zijn de BI vrij smal. We kunnen de standaardafwijkingen op minder dan 10 % nauwkeurig bepalen.

R-output 14: BI voor de random effects parameters van het finale model voor cadmium – cJaar

```
Random Effects:
  Level: Meetplaats
                    lower      est.     upper
sd((Intercept)) 0.3791538 0.4115555 0.4467261

 Within-group standard error:
    lower      est.     upper
0.2652083 0.2807594 0.2972223
```

5.1.4 Modeldiagnose

Bij de statistische analyse van gegevens hoort ook een toetsing van de modelveronderstellingen voor het finale model. Hierbij is een grafische analyse van de *residuals* een belangrijk instrument. De *residuals* zijn de afwijkingen van de waarnemingen t.o.v. het geschatte regressiemodel en ze zijn te interpreteren als een schatting van de ruisterm. Ze moeten bijgevolg in een goede benadering vergelijkbare kenmerken hebben als de ruis. Als dat niet het geval is, dan is dat een indicatie dat er iets fout is met het model. Uit het patroon kunnen we vaak afleiden wat er fout is.

Een eerste belangrijk kenmerk van de ruis is de statistische onafhankelijkheid. Als de residuals patronen vertonen wanneer we plotten ten opzichte van een variabele, dan kan dat een indicatie zijn dat de variabele niet in een goede vorm in het model opgenomen is. Een klassiek voorbeeld is dat het residupatroon parabolisch is, wat erop wijst dat een tweedegraadsfunctie nodig.

Een tweede kenmerk is homoskedasticiteit. Het model veronderstelt dat de *residuals* overal een gelijke variantie hebben. Omdat de variantie vaak de neiging heeft om te stijgen naarmate de response toeneemt, worden de residuals geplot in functie van de fitted values. Ook kan de variantie verschillen naargelang een verklarende variabele zoals ecoregio. Dat kunnen we nagaan met boxplots.

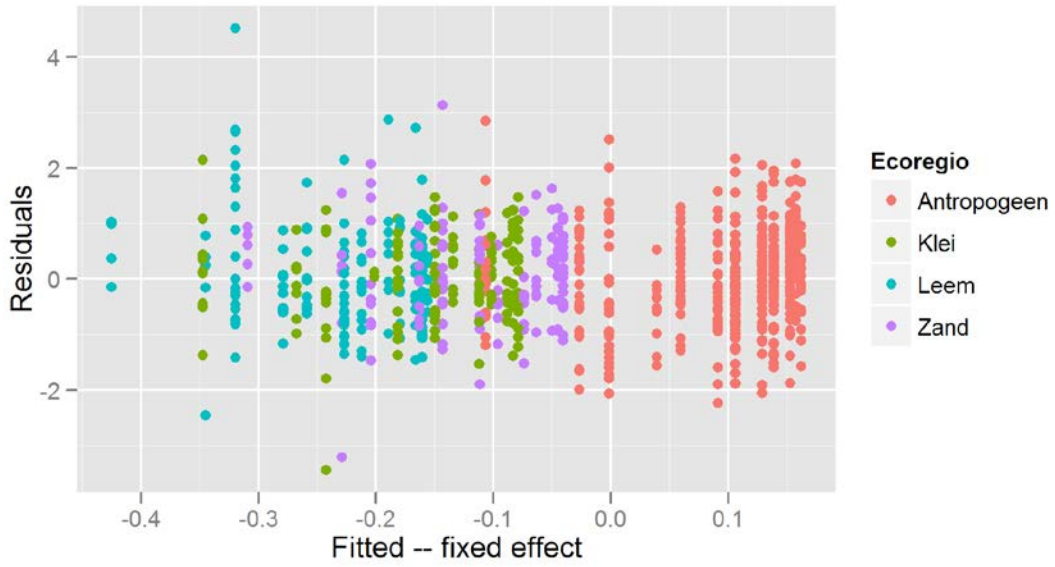
Ten derde moeten de *residuals* in een eerste benadering normaal verdeeld zijn. Dat kunnen we nagaan met QQ-plots die de *residuals*, gerangschikt van klein naar groot, vergelijkt met de kwantielen van een normale verdeling. Ook kunnen we een histogram van de residuals maken om de normaliteit te testen.

Voor alle figuren gebruiken we de gestandaardiseerde residuals. In een eerste goede benadering verwachten we ongeveer 95 % van de waarden binnen het interval ± 2. Op die manier kunnen we ook uitbijters gemakkelijk opsporen. Want niet alleen met het model kan iets fout zijn, maar ook met de waarnemingen zelf.

Tenslotte moeten we ook nog de normaliteit van de random effecten controleren.

(a) Residu-plots ten opzichte van de gefitte waarden.

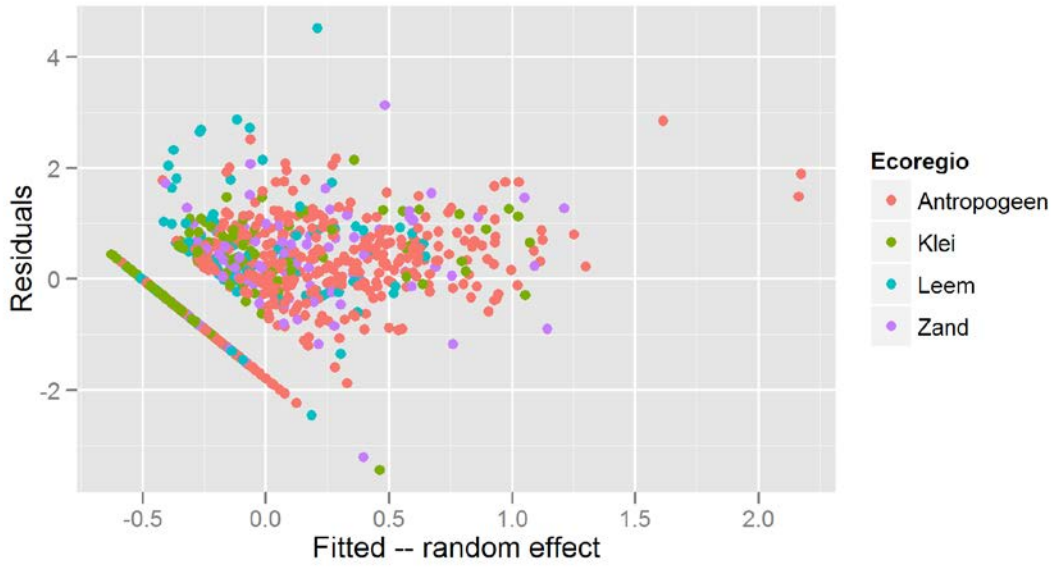
Figuur 12: Modeldiagnose voor cadmium – cJaar: Fitted values (fixed effects) versus residuals



In Figuur 12 worden enkel de fixed effects (cJaar, cJaar², ecoregio) meegenomen om de fitted values te berekenen. Buiten enkele uitschieters (waarden buiten de range -2 tot 2) vertoont deze figuur geen opmerkelijke patronen.

Wanneer we ook de random effects (meetplaats specifiek intercept) mee in rekening brengen bij de berekening van de fitted values, dan krijgen we een heel andere figuur te zien (Figuur 13). Links onderaan kunnen we duidelijk twee lijnen onderscheiden, te wijten aan de detectielimiet. De eerste lijn bevat alle waarden van observaties die onder de detectielimiet lagen, en als waarde de helft van de maximale detectielimiet meekregen. De tweede lijn ("ondergrens" van de wolk) geeft de scheidingslijn van deze maximale detectielimiet weer. Dit geeft duidelijk aan dat er een probleem is met de "continuïteit" van de gegevens.

Figuur 13: Modeldiagnose voor cadmium – cJaar: Fitted values (fixed + random effects) versus residuals



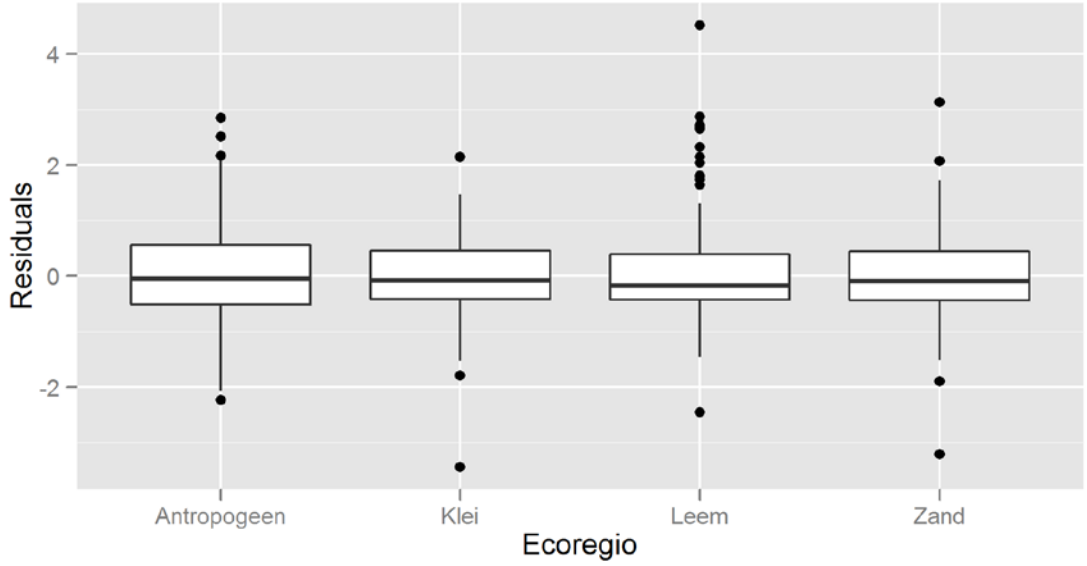
(b) Residu-plots ten opzichte van de verklarende variabelen

De opgesplitste boxplots per ecoregio (Figuur 14) laten zien dat de variabiliteit per regio constant is, en dat de voorwaarde van homogeniteit van de residuals geen probleem vormt.

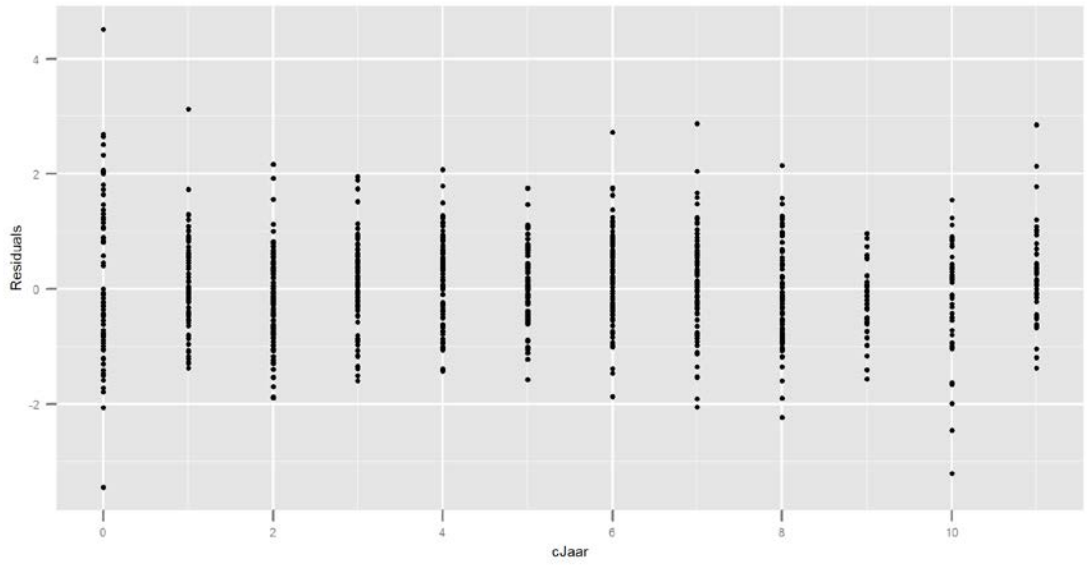
Ook de variabiliteit over de verschillende jaren blijft nagenoeg constant (Figuur 15).

Vermits er voor elk van de 445 meetplaatsen slechts 3 metingen zijn, werden in Figuur 16 geen boxplots getekend, maar wel de individuele punten. Op een beperkt aantal uitschieters na (~5 %) lijkt de variabiliteit binnen de meetplaatsen een constante (voor zover de drie punten per meetplaats goed te onderscheiden zijn).

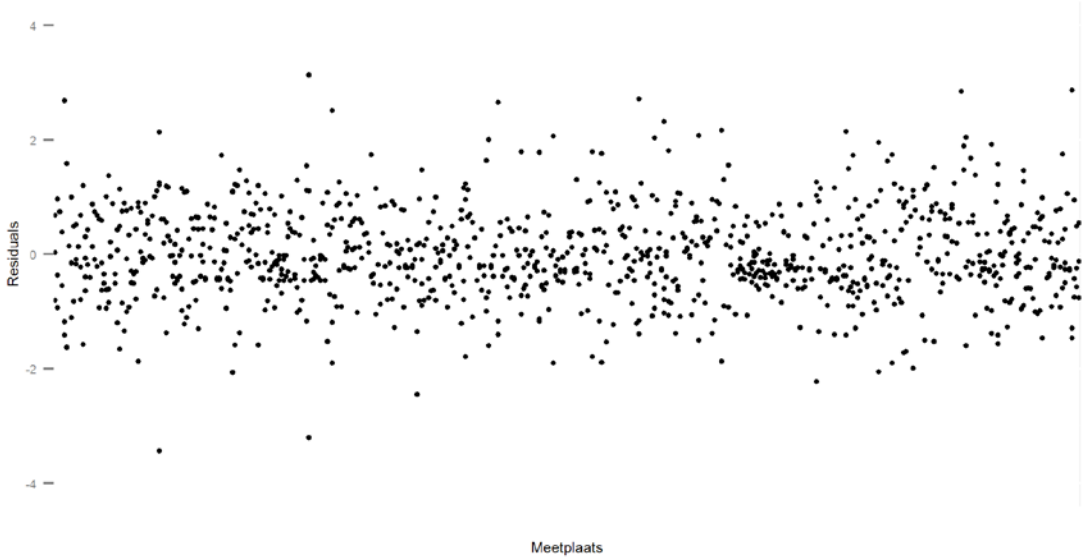
Figuur 14: Modeldiagnose voor cadmium – cJaar: Residuals opgesplitst per Ecoregio



Figuur 15: Modeldiagnose voor cadmium – cJaar: Residuals opgesplitst per cJaar



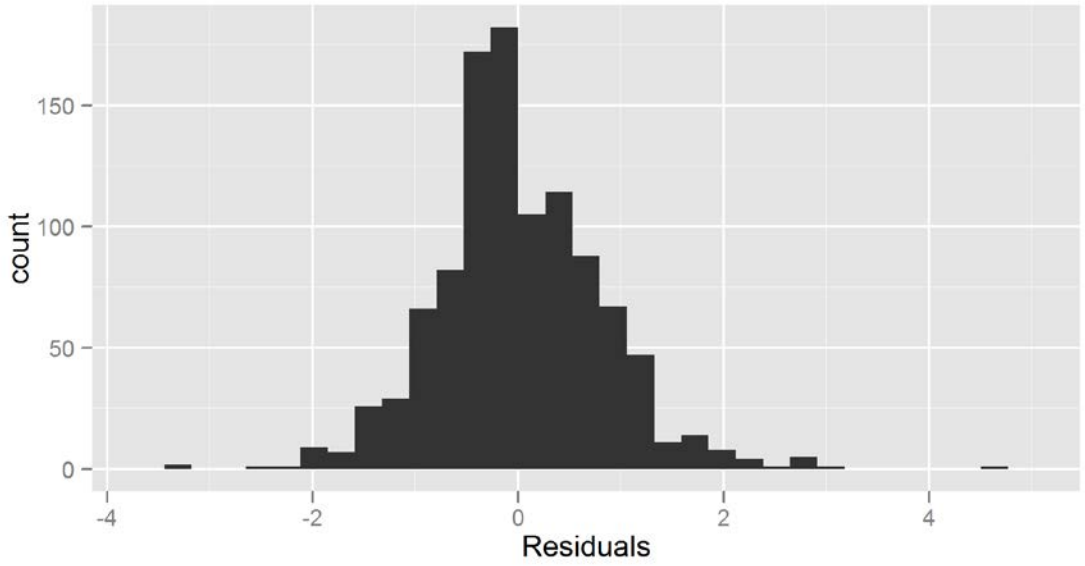
Figuur 16: Modeldiagnose voor cadmium – cJaar: Residuals opgesplitst per meetplaats



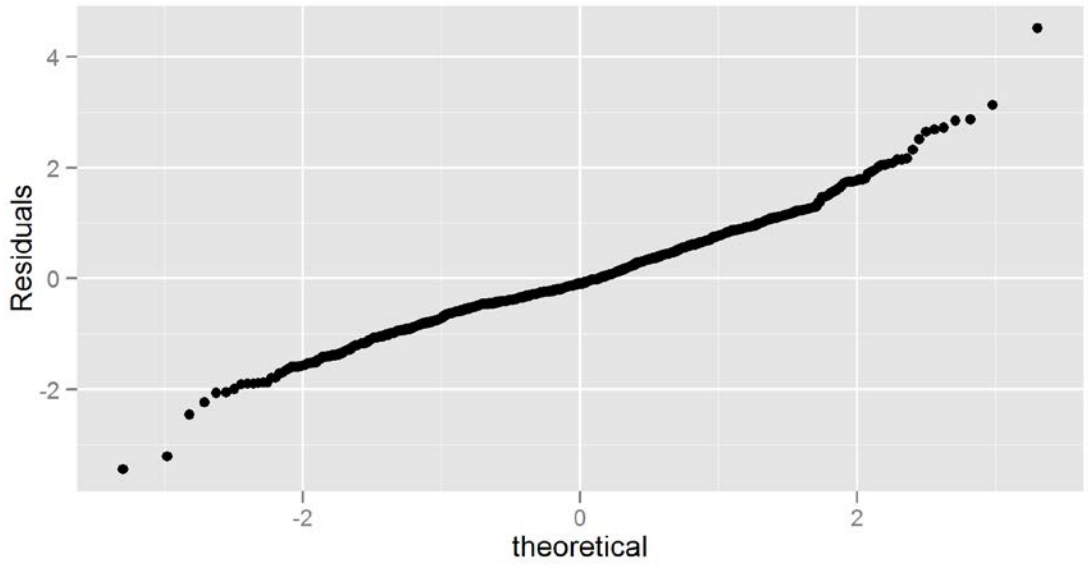
(c) Normaliteit van de residuals en het random intercept

Zowel het histogram (Figuur 17) als de QQ-plot (Figuur 18) laten geen al te grote afwijkingen van normaliteit zien voor de residuals. De normaliteit van de random effects (Figuur 19) is twijfelachtig omwille van de onderste staart, maar deze is vermoedelijk te wijten aan de waarnemingen onder de detectielimiet, die allemaal dezelfde waarde (halve maximale detectielimiet) gekregen hebben. Merk ook op dat de lijn die door de punten getrokken kan worden, niet samenvalt met de eerste bissectrice (door de oorsprong en helling gelijk aan 1), zoals wel het geval is voor de residuals, maar een helling heeft gelijk aan σ_O, omdat de random effects niet gestandaardiseerd werden zoals de residuals.

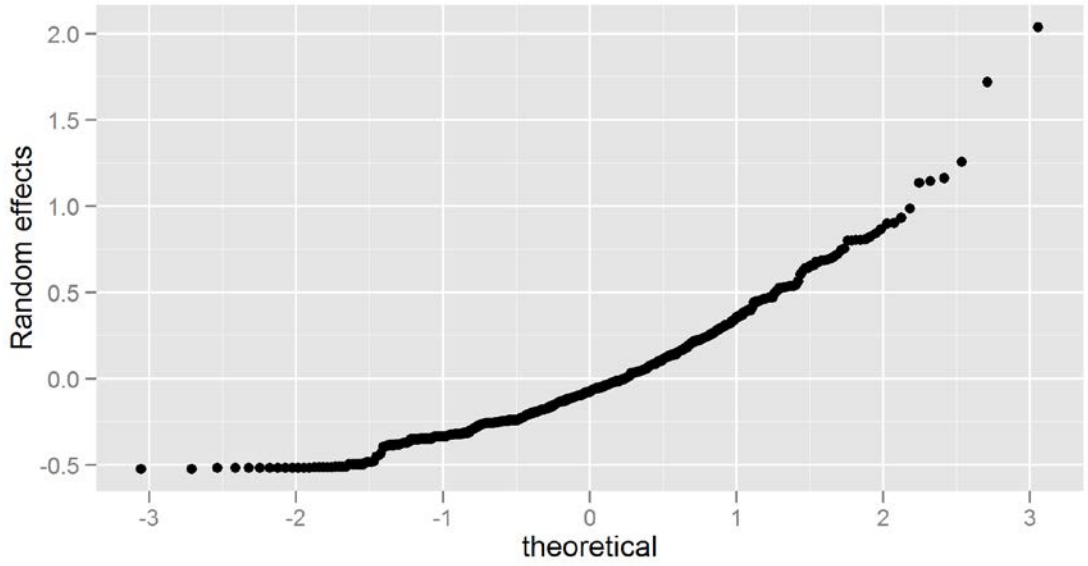
Figuur 17: Modeldiagnose voor cadmium – cJaar: Histogram van de residuals



Figuur 18: Modeldiagnose voor cadmium – cJaar: QQ-plot van de residuals



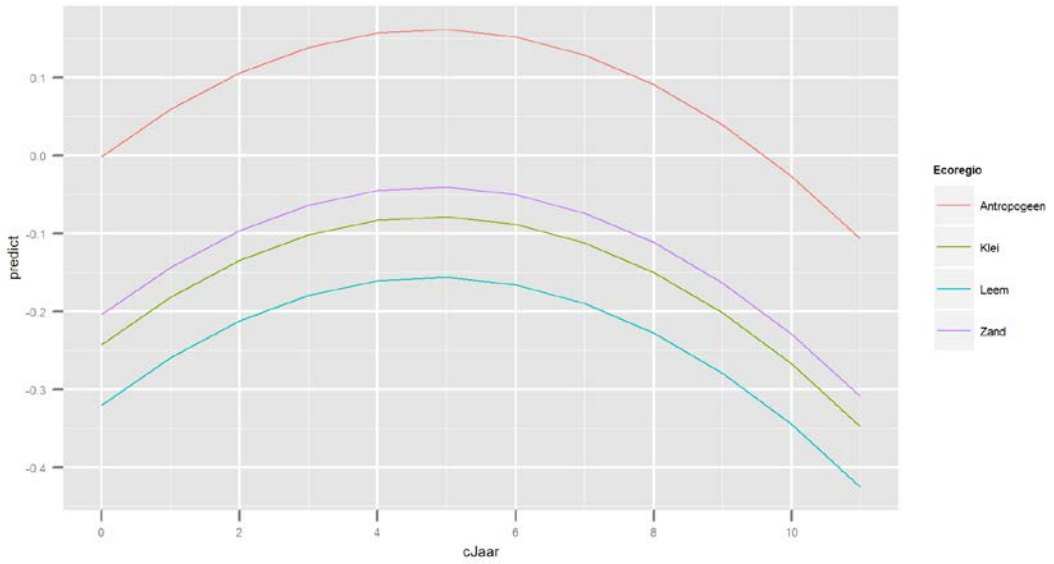
Figuur 19: Modeldiagnose voor cadmium – cJaar: QQ-plot van de random effects



5.1.5 Grafische voorstelling van het finale model

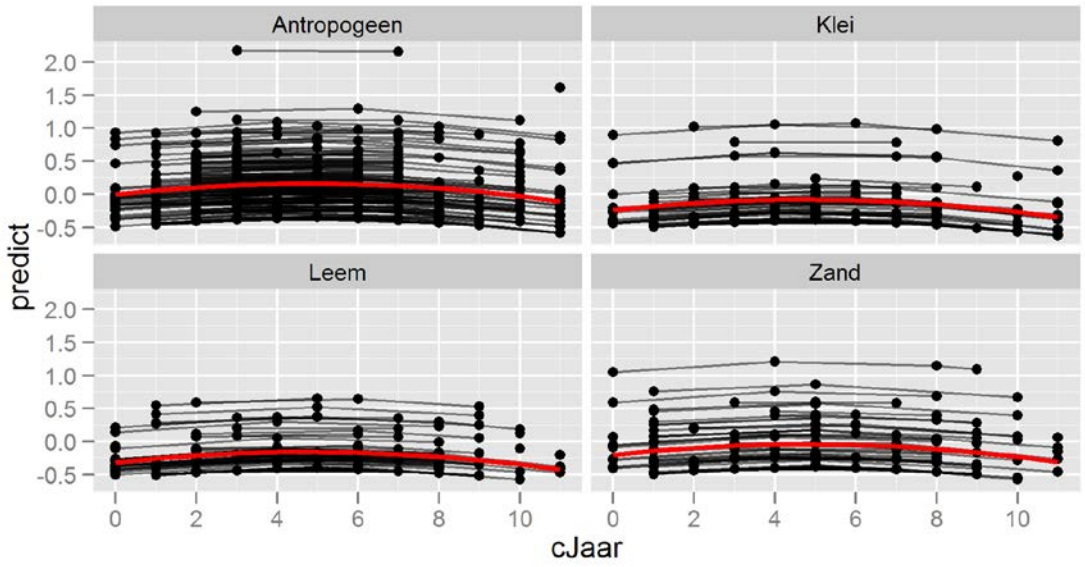
Op basis van de parameterschattingen kunnen we nu het finale model grafisch voorstellen waarbij we de effecten van de verschillende factoren visualiseren. Het model geeft aan dat de evolutie in de tijd parabolisch is en dat er significante verschillen zijn tussen de ecoregio's. We beginnen met het fixed effect gedeelte voor te stellen. Vervolgens voegen we ook de random effecten van de meetplaatsen toe aan de figuur.

Figuur 20: Grafische voorstelling van het finale model voor cadmium – cJaar



Figuur 20 laat duidelijk het verschil in logconcentratie cadmium tussen de ecoregio's zien, en dat de concentratie in de ecoregio Antropogeen beduidend hoger ligt dan in de andere ecoregio's. Ook de kwadratische vorm van de curve is duidelijk zichtbaar, met het maximum net voor cJaar = 5.

Figuur 21: Grafische voorstelling van het finale model voor cadmium – cJaar, opgesplitst per ecoregio



In Figuur 21 werden naast de 4 curves voor de verschillende ecoregio's (rode lijnen, enkel fixed effect voor de desbetreffende ecoregio en trend) ook de individueel voorspelde punten per meetplaats weergegeven (fixed effect van ecoregio, trend in de jaren waarin een opmeting gebeurde, en random effect van de meetplaats), verbonden met zwarte lijnen. Hieruit wordt nog maar eens duidelijk dat de grootste variabiliteit te wijten is aan het verschil tussen de meetplaatsen.

5.2 Meetcyclus als tijdsvariabele

5.2.1 Het startmodel

Een andere manier om de trend te modelleren, is de gegevens per meetcyclus te analyseren. We modelleren tijd dan als een categorische variabele of factor (fMeetcyclus).

Opnieuw introduceren we meetplaats als een random effect (notatie: 1 | Meetplaats) en we kunnen nagaan of Ecoregio een rol speelt.

Voegen we al deze modeltermen samen, dan krijgen volgend globaal model.

logCd ~ 1|Meetplaats + fMeetcyclus + Ecoregio + fMeetcyclus : Ecoregio

5.2.2 Keuze van het model (modelreductie en modelverfijning)

We kunnen nu op basis van de ANOVA-tabel in R-output 15 nagaan of al deze termen significant zijn. We elimineren de niet-significante interactie. De overblijvende termen blijven significant, zodat we dit finale model bekomen:

logCd ~ 1|Meetplaats + fMeetcyclus + Ecoregio

R-output 15: ANOVA-tabel voor de trendanalyse van cadmium – fMeetcyclus

```
                     numDF denDF   F-value p-value
(Intercept)              1   590  3.424117  0.0648
fMeetcyclus              2   590 24.398069  <.0001
Ecoregio                 3   441 12.828306  <.0001
fMeetcyclus:Ecoregio     6   590  1.242943  0.2824
```

5.2.3 De parameterschattingen

Voor het finale model zijn de parameterschattingen en de bijhorende betrouwbaarheidsintervallen terug te vinden in R-output 16.

R-output 16: Het finale model voor de trendanalyse van cadmium – fMeetcyclus

Anova

```
            numDF denDF   F-value p-value
(Intercept)     1   596  3.429992  0.0645
fMeetcyclus     2   596 24.315590  <.0001
Ecoregio        3   441 12.839178  <.0001
```

Summary

```
Linear mixed-effects model fit by REML
 Data: AllData
       AIC      BIC    logLik
  1132.636 1172.189 -558.3179

Random effects:
 Formula: ~1 | Meetplaats
        (Intercept)  Residual
StdDev:   0.4068611 0.2827068

Fixed effects: logRespons ~ fMeetcyclus + Ecoregio
                  Value  Std.Error  DF   t-value p-value
(Intercept)   0.0687796 0.03295156 596  2.087293  0.0373
fMeetcyclus2  0.0992331 0.02023395 596  4.904286  0.0000
fMeetcyclus3 -0.0662752 0.02589494 596 -2.559388  0.0107
EcoregioKlei -0.2432922 0.06282814 441 -3.872345  0.0001
EcoregioLeem -0.3171363 0.05784569 441 -5.482453  0.0000
EcoregioZand -0.2008818 0.05915040 441 -3.396119  0.0007

Number of Observations: 1043
Number of Groups: 445
```

Approximate 95 % confidence intervals

```
 Fixed effects:
                    lower        est.       upper
(Intercept)   0.004064284  0.06877957  0.13349486
fMeetcyclus2  0.059494587  0.09923311  0.13897163
fMeetcyclus3 -0.117131607 -0.06627519 -0.01541877
EcoregioKlei -0.366772005 -0.24329224 -0.11981247
EcoregioLeem -0.430823766 -0.31713629 -0.20344881
EcoregioZand -0.317133461 -0.20088176 -0.08463007

 Random Effects:
  Level: Meetplaats
                    lower      est.     upper
sd((Intercept)) 0.3747698 0.4068611 0.4417002
```

```
 Within-group standard error:
    lower      est.     upper
0.2670949 0.2827068 0.2992313

Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Tukey Contrasts
Linear Hypotheses:
                       Estimate Std. Error z value Pr(>|z|)
Klei - Antropogeen == 0 -0.24329    0.06283  -3.872  < 0.001 ***
Leem - Antropogeen == 0 -0.31714    0.05785  -5.482  < 0.001 ***
Zand - Antropogeen == 0 -0.20088    0.05915  -3.396  0.00378 **
Leem - Klei == 0        -0.07384    0.07330  -1.007  0.74125
Zand - Klei == 0         0.04241    0.07433   0.571  0.93969
Zand - Leem == 0         0.11625    0.07018   1.657  0.34211
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

(a) De fixed effects

In het gedeelte onder Fixed effects zien we opnieuw dat de logconcentratie cadmium het hoogst is voor ecoregio Antropogeen (referentieklasse), vermits alle parameterschattingen voor de 3 andere ecoregio's (die het verschil in logconcentratie aangeven tussen Antropogeen en die andere ecoregio) negatief zijn, dus een lagere logconcentratie geven. Deze verschillen tussen Antropogeen en de andere ecoregio's zijn allemaal significant. In het gedeelte Simultaneous Tests zien we dat de verschillen tussen de andere ecoregio's niet significant zijn. Voor de trend kijken we nu naar de parameterschattingen van de factorvariabele fMeetcyclus. Deze geven de verschillen weer met de logconcentratie in de eerste meetcyclus. Voor meetcyclus 2 is de logconcentratie significant hoger dan in meetcyclus 1, voor meetcyclus 3 significant lager dan in meetcyclus 1. Dit geeft dus opnieuw eerst een stijgende trend, en daarna terug een daling (wat ook al bleek uit de kwadratische trend bij de analyse met cJaar).

(b) De random effects

Ook de schatting voor de random effects variabiliteit en de bijhorende BI zijn terug te vinden in R-output 16 onder Random effects. De standaard deviatie van het random intercept (σ_0=0.407) is duidelijk groter dan de residuele standaard deviatie (σ_e=0.283), waaruit we kunnen concluderen dat de variabiliteit tussen de meetplaatsen groter is dan de resterende variabiliteit binnen de meetplaatsen die niet verklaard kan worden door het model.

5.2.4 Modeldiagnose

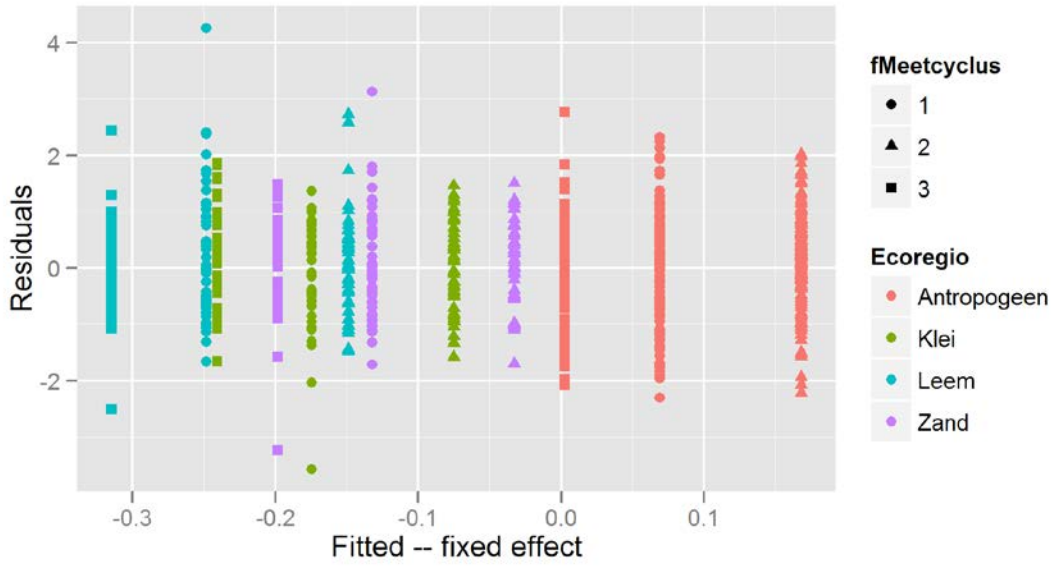
Ook voor deze analyse controleren we de gemaakte modelveronderstellingen voor het finale model (onafhankelijkheid, homogeniteit en normaliteit van de residuals, en normaliteit van de random effects).

(a) Residuplots in functie van de fitted values

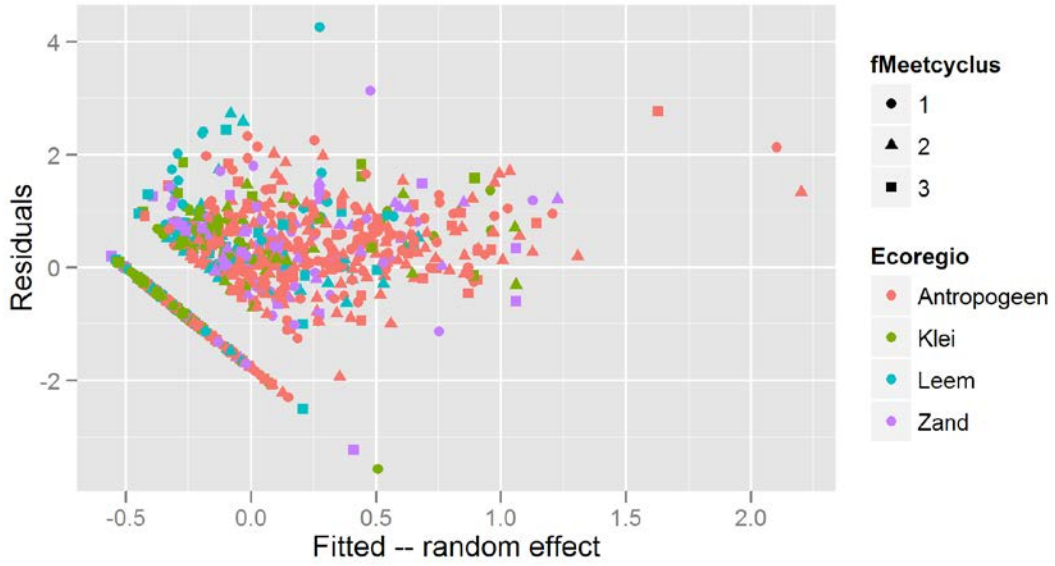
In Figuur 22 worden enkel de fixed effects (fMeetcyclus, ecoregio) meegenomen om de fitted values te berekenen. Buiten enkele uitschieters (waarden buiten de range ±2, vermoedelijk dezelfde waarnemingen als in Figuur 12) vertoont deze figuur geen opmerkelijke patronen.

In Figuur 23 worden ook de random effects toegevoegd aan de fitted values om het profiel per waterloop te schatten. We zien hetzelfde patroon als in Figuur 13. Links onder een lijn die veroorzaakt wordt door de waarnemingen die herleid werden naar de halve maximale detectielimiet, en daarnaast een lijnvormige ondergrens van de wolk, veroorzaakt door de maximale detectielimiet die als minimale waarde voorkomt voor de "echte" waarnemingen. Verder zijn er geen onverklaarde patronen zichtbaar in de figuur.

Figuur 22: Modeldiagnose voor cadmium – fMeetcyclus: residuals versus fitted values (fixed effects)



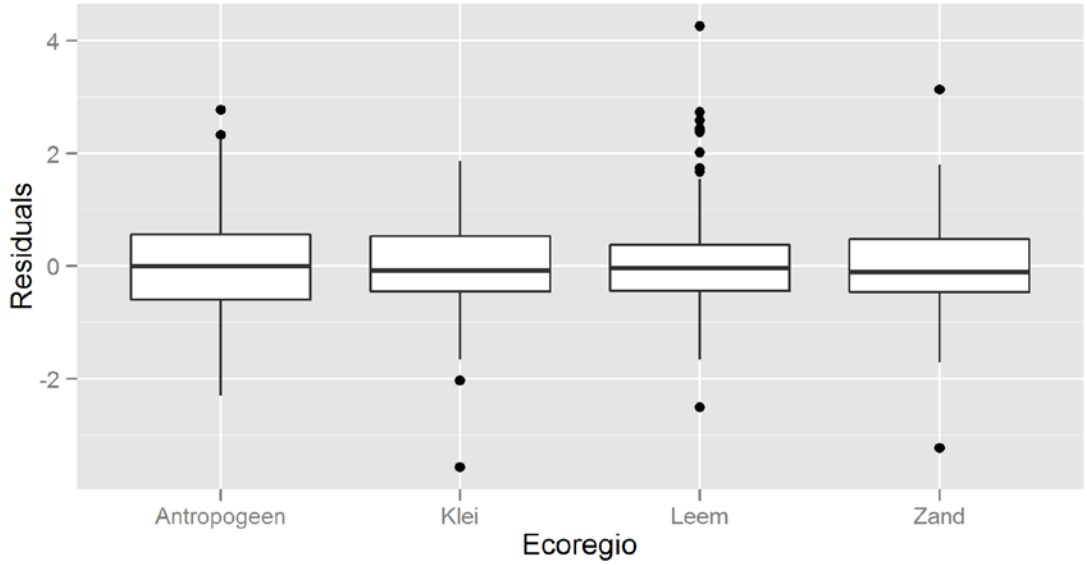
Figuur 23: Modeldiagnose voor cadmium – fMeetcyclus: residuals versus fitted values (fixed + random effects)



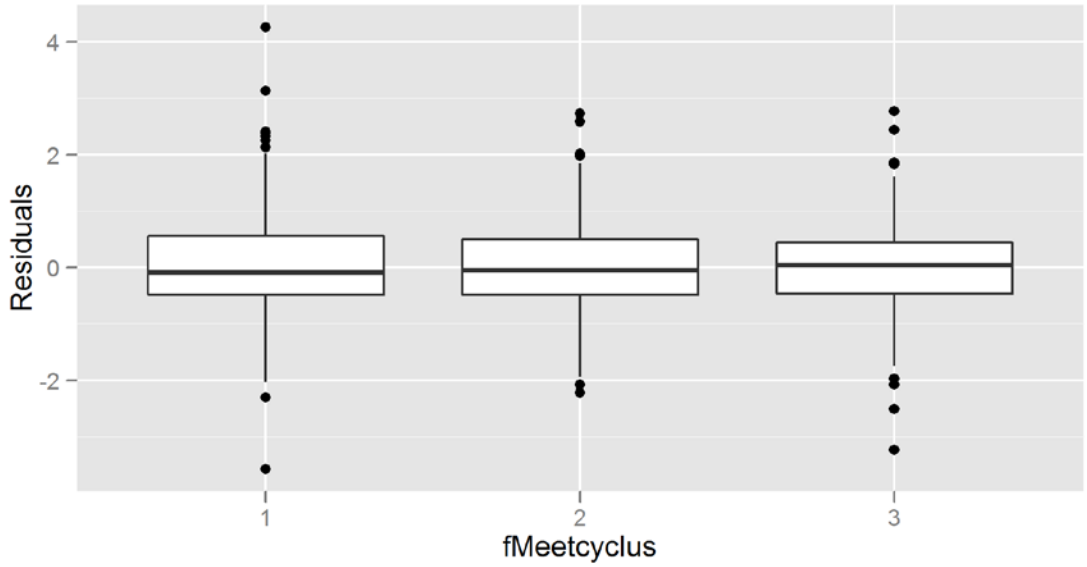
(b) Residuplots in functie van verklarende variabelen

De opgesplitste boxplots per ecoregio (Figuur 24) laten zien dat de variabiliteit per regio redelijk constant is. De voorwaarde voor homoskedasticiteit is vervuld. Ook een opsplitsing van de residuals over de 3 meetcycli laat geen duidelijke verschillen zien in variabiliteit (Figuur 25). Voor de opsplitsing van de residuals over de meetplaatsen gebruiken we opnieuw de individuele punten i.p.v. boxplots (Figuur 26). Op een beperkt aantal uitschieters na (~5 %) lijkt de variabiliteit binnen de meetplaatsen opnieuw een constante (voor zover de 3 punten per meetplaats te onderscheiden zijn).

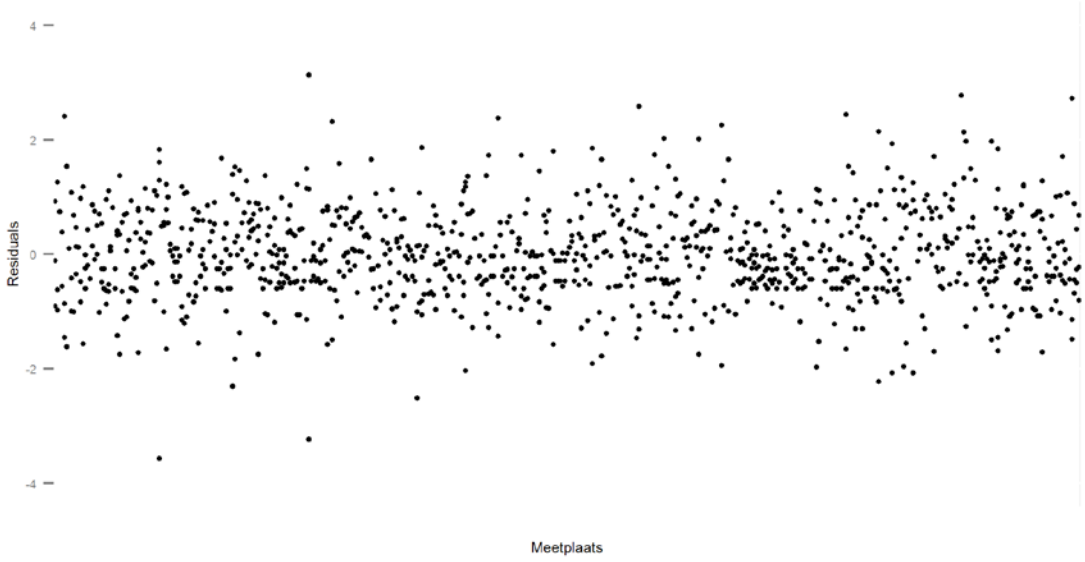
Figuur 24: Modeldiagnose voor cadmium – fMeetcyclus: Residuals opgesplitst per Ecoregio



Figuur 25: Modeldiagnose voor cadmium – fMeetcyclus: Residuals opgesplitst per Meetcyclus



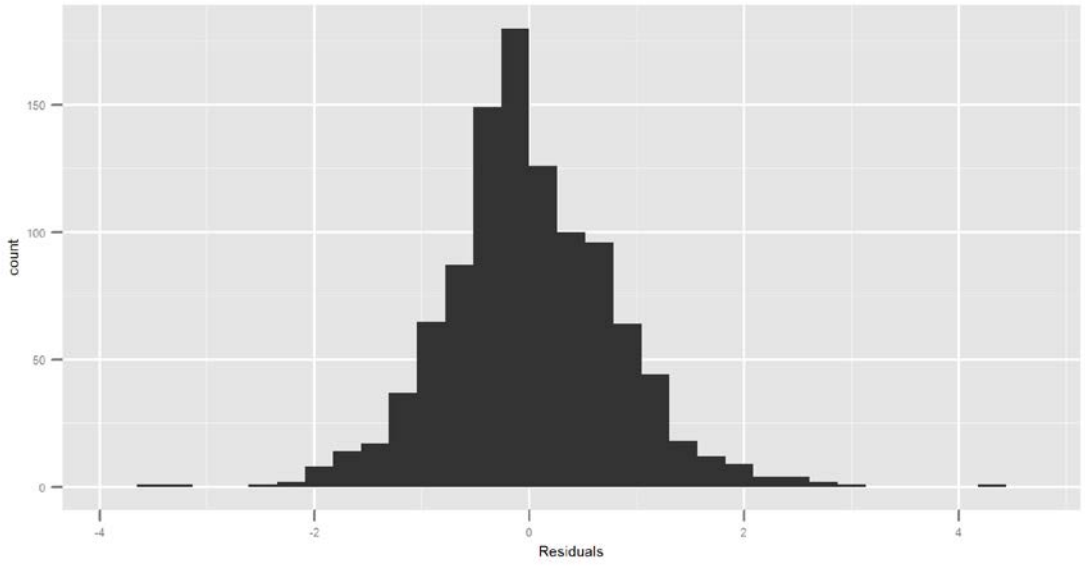
Figuur 26: Modeldiagnose voor cadmium – fMeetcyclus: Residuals opgesplitst per meetplaats



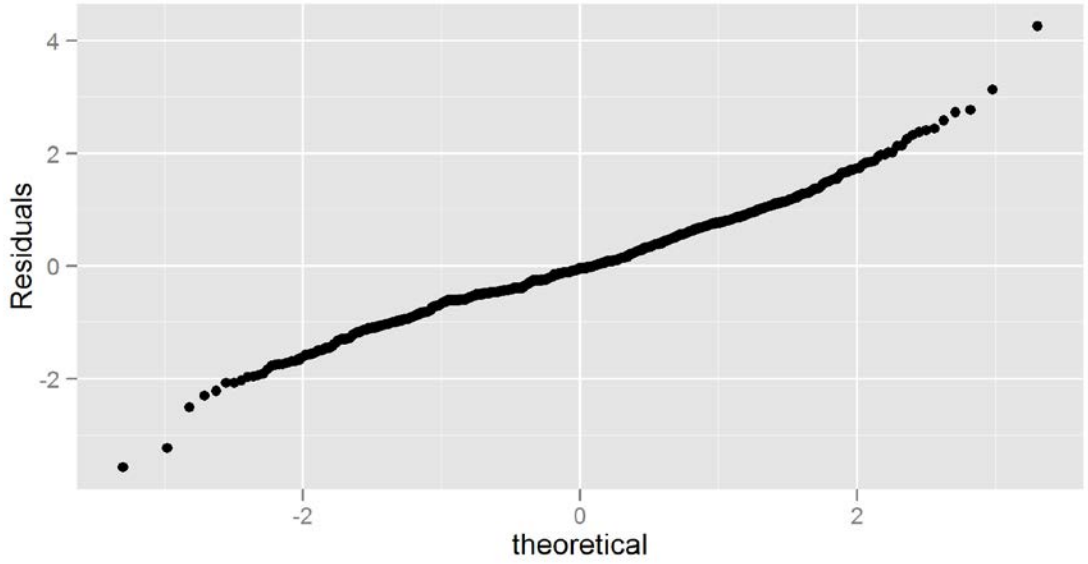
(c) Normaliteit van de residuals en het random intercept

Zowel het histogram (Figuur 27) als de QQ-plot (Figuur 28) laten geen al te grote afwijkingen van normaliteit zien voor de residuals. De normaliteit van de random effects (Figuur 29) is opnieuw twijfelachtig omwille van de onderste staart (die weer naar boven “krult”), maar ook nu vermoedelijk te wijten is aan de waarnemingen onder de detectielimiet.

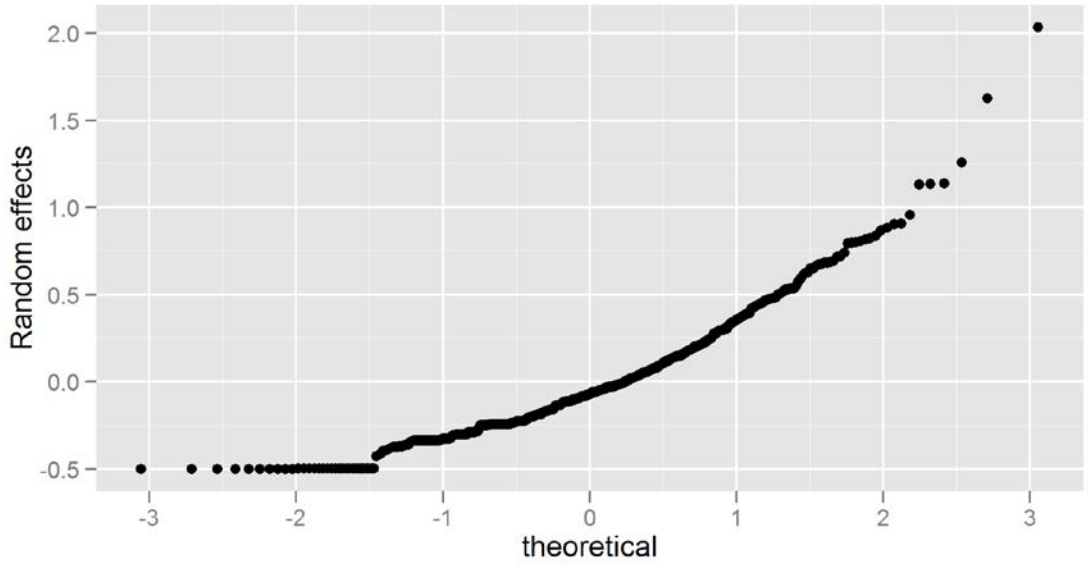
Figuur 27: Modeldiagnose voor cadmium – fMeetcyclus: Histogram van de residuals



Figuur 28: Modeldiagnose voor cadmium – fMeetcyclus: QQ-plot van de residuals



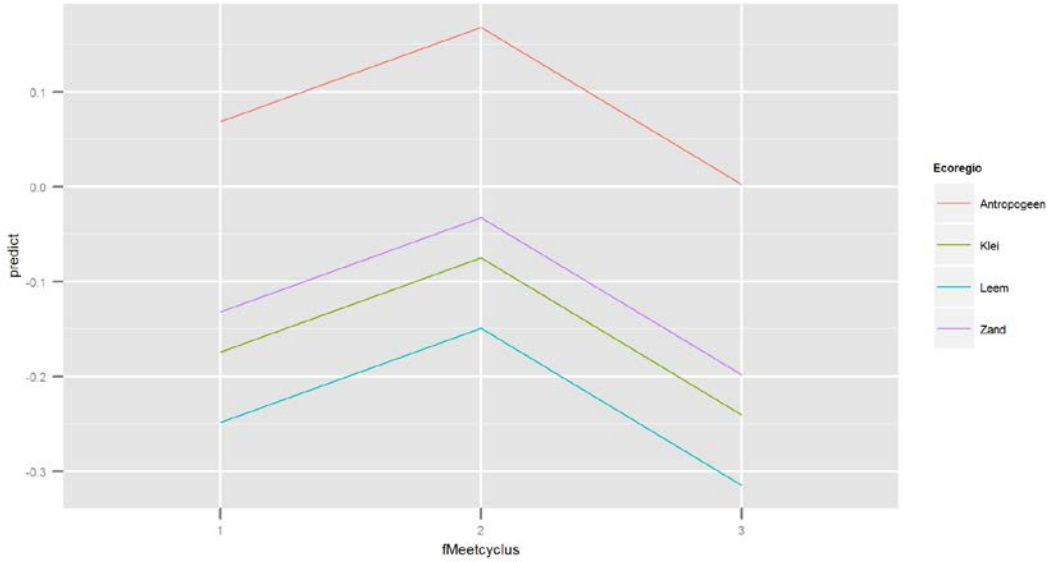
Figuur 29 Modeldiagnose voor cadmium – fMeetcyclus: QQ-plot van de random effects



5.2.5 Grafische voorstelling van het finale model

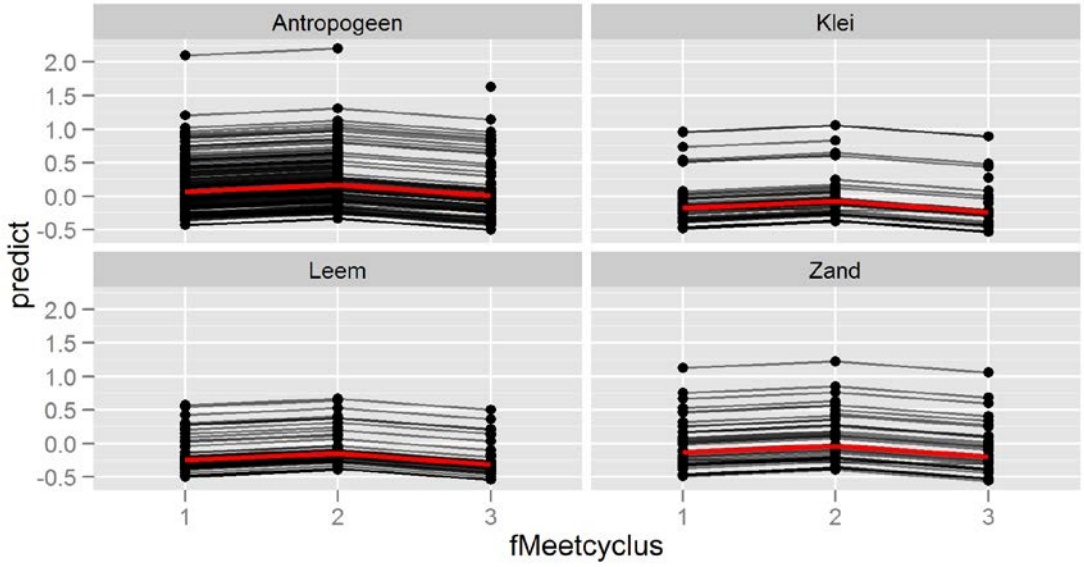
Opnieuw kunnen we op basis van de parameterschattingen het finale model grafisch voorstellen Figuur 30 laat weer duidelijk het verschil in logconcentratie cadmium tussen de ecoregio's zien, en dat de concentratie in de ecoregio Antropogeen beduidend hoger ligt dan in de andere ecoregio's. Vermits meetcyclus als een factorvariabele meegenomen werd in het model, krijgt elke meetcyclus een afzonderlijke geschatte waarde, en dit resulteert in de gebroken lijnen. Echter, de "kwadratische" trend van voorheen (met cJaar) is er nog in te herkennen (eerst een stijging en daarna een sterke daling).

Figuur 30: Grafische voorstelling van het finale model voor cadmium – fMeetcyclus



In Figuur 31 werden naast de 4 curves voor de verschillende ecoregio's (rode lijnen, enkel fixed effect voor de desbetreffende ecoregio en trend) ook de individueel voorspelde punten per meetplaats weergegeven (fixed effect van ecoregio, fixed effect van meetcyclus waarin een opmeting gebeurde, en random effect van de meetplaats), en verbonden (parallelle lijnen omwille van het random intercept). Hieruit wordt nog maar eens duidelijk dat de grootste variabiliteit te wijten is aan het verschil tussen de meetplaatsen.

Figuur 31: Grafische voorstelling van het finale model voor cadmium – fMeetcyclus, opgesplitst per ecoregio



5.3 Besluiten

Voor de analyse van de gegevens van cadmium hebben we twee modellen verkend. Het eerste model behandelde tijd als een continue variabele, het tweede model beschouwde tijd als een factor waarbij de jaren gegroepeerd werden per cyclus. Beide modellen blijken goed aan te sluiten bij de data. De modelvalidatie laat alleen problemen zien voor de waarnemingen onder de detectielimiet die vervangen werden door de halve maximale detectielimiet. Hierdoor gaat de variabiliteit in de metingen verloren waardoor er patronen in de residu's ontstaan.

Voor beide modellen komen we ook tot dezelfde conclusie. Tijdens de eerste jaren van de studie (van meetcyclus 1 naar 2) stijgt de gemiddelde concentratie cadmium, waarna deze nog sterker daalt (tussen meetcyclus 2 en 3). Ook de impact van ecoregio is in beide modellen gelijk. De hoogste concentratie komt voor in de ecoregio Antropogeen. De concentraties in de andere ecoregio's verschillen niet significant van mekaar.

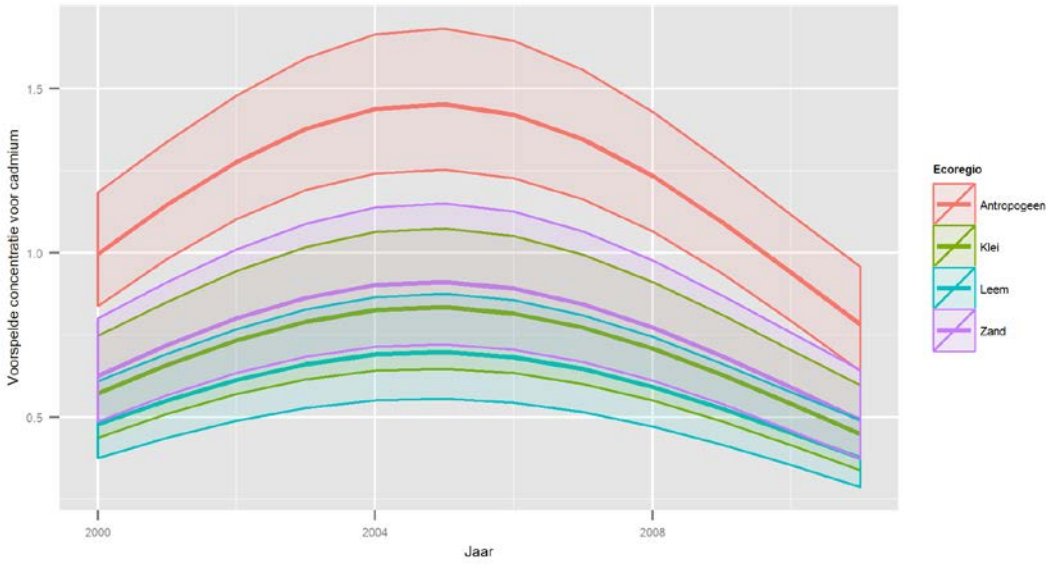
Figuur 32 balt alle conclusies nog eens samen in een figuur. In tegenstelling tot eerdere figuren staan de gegevens nu in de originele schaal (concentratie in mg/kg ds) i.p.v. in de log-schaal. De figuur bevat ook betrouwbaarheidsintervallen voor elk van de trendlijnen zodat de verschillen tussen de ecoregio's onmiddellijk kunnen beoordeeld worden. Zo kunnen we direct aflezen dat de concentratie in het Antropogeen significant verschilt van de andere regio's die overlappen.

De hoogste concentraties werden gemeten rond 2005 (cJaar = 5), en sindsdien is er een sterke daling opgetreden. We moeten echter voorzichtig zijn met extrapolaties en voorspellingen voor de toekomst, en mogen er niet zomaar vanuit gaan dat deze daling zich op dezelfde manier zal verder zetten. Met regressiemodellen modelleren we niet de onderliggende dynamiek, maar zoeken we alleen naar de beste aanpassing van een curve van de data. Doordat we geen inzicht hebben in het onderliggende mechanisme, mogen we niet zomaar aannemen dat ook in de toekomst deze trend zich zal doorzettten.

Uit de betrouwbaarheidsbanden in Figuur 32 kunnen we ook nog aflezen dat voor antropogene gebieden de gemiddelde concentratie cadmium in 2000 met 95 % zekerheid lag tussen 0.85 en 1.17 mg/kg ds, dat dit in 2005 steeg tot [1.25;1.65] en daarna terug daalde tot [0.65;0.95] in 2011. Analoge conclusies kunnen we afleiden voor de andere ecoregio's.

Voor de interpretatie van de figuur is het ook nodig te onthouden dat de regressielijnen de gemiddelde evolutie voorstellen van de waterlopen in de ecoregio. In random effects modellen gelden alle interpretaties van fixed effects voor een "gemiddelde" meetplaats, d.w.z. voor een meetplaats met random effect gelijk aan 0 (= het gemiddelde random effect).

Figuur 32: Het finale model voor cadmium in de originele schaal (met 95 % betrouwbaarheidsbanden)



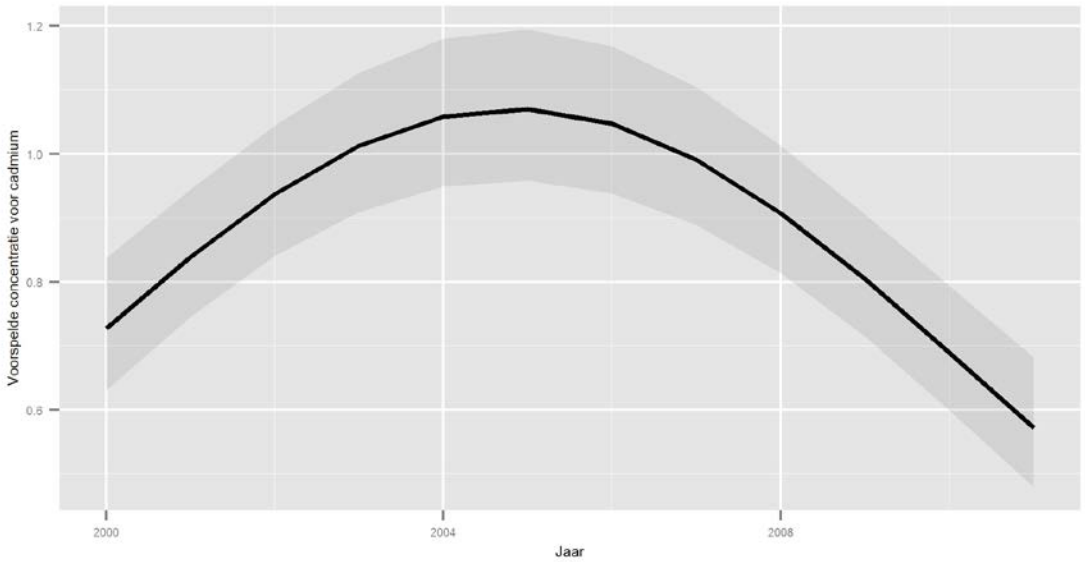
In het kader van de rapportering is het interessant om een uitspraak te doen over de globale trend voor het hele meetnet, en niet afzonderlijk over de verschillende ecoregio's. Een eenvoudig antwoord bekomen we door het finale model te herhalen, maar de variabele Ecoregio niet meer mee op te nemen. De kwadratische trend en het random intercept blijven gewoon behouden. Voorwaarde is wel dat de steekproef representatief is voor Vlaanderen, en dat dus de verdeling van de ecoregio's binnen de steekproef overeenkomt met die in de populatie. Het meetnet is zo opgezet dat aan deze voorwaarde voldaan is, maar door het verwijderen van een hele reeks observaties (Jansen 2012) moet deze voorwaarde in principe eerst opnieuw onderzocht worden. We bespreken de resultaten in de veronderstelling dat de resterende steekproef nog steeds representatief is voor Vlaanderen.

R-output 17: Trendanalyse over heel Vlaanderen – Cadmium

```
Linear mixed-effects model fit by REML
 Data: AllData
       AIC      BIC    logLik
  1160.844 1185.579 -575.4219

Random effects:
 Formula: ~1 | Meetplaats
        (Intercept)  Residual
StdDev:   0.4306609 0.2808509

Fixed effects: logRespons ~ cJaar + I(cJaar^2)
                  Value   Std.Error  DF   t-value p-value
(Intercept) -0.13843668 0.031262419 596 -4.428214       0
cJaar        0.06948618 0.010662333 596  6.516977       0
I(cJaar^2)  -0.00717792 0.001012292 596 -7.090758       0
 Correlation:
            (Intr) cJaar
cJaar       -0.627
I(cJaar^2)  0.513 -0.959

Number of Observations: 1043
Number of Groups: 445
```

R-output 17 laat de parameterschattingen zien van de algemene trend over Vlaanderen. Voor de *fixed effects* is het enige verschil met de resultaten in R-output 11 de schatting van het intercept (de ligging van de curve) die nu een “gewogen” gemiddelde is evenredig met steekproefgrootte van de vier ecoregio’s, en niet meer specifiek is voor de ecoregio Antropogeen. De standaarddeviatie van het random intercept is nu groter dan voorheen (σ_O=0.431) vermits deze nu ook de variabiliteit tussen de ecoregio’s moet opvangen. De residuele standaard deviatie (σ_e=0.281) blijft gelijk. De globale trend over Vlaanderen wordt voorgesteld in Figuur 33. Hierbij willen we benadrukken dat alleen als de steekproefgrootte per ecoregio de werkelijke verdeling in Vlaanderen weerspiegelt, de figuur representatief is voor Vlaanderen. Anders moeten we manueel de gewichten per ecoregio aanpassen.

Figuur 33: De trend over Vlaanderen voor cadmium in de originele schaal (met 95 % betrouwbaarheidsinterval)



6 Voorbeeld 2: trendanalyse voor arseen

De analyse verloopt totaal analoog als voor cadmium. Daarom bespreken we de meeste output slechts heel summier. Alleen bij nieuwigheden gaan we dieper in op de materie.

6.1 Jaar van opname als tijdsvariabele

6.1.1 Het startmodel

Uit de verkennende analyse was geen duidelijke trend zichtbaar. Voor de zekerheid starten we de analyses met een derdegraadsvergelijking om de trend flexibel te modelleren (notatie: cJaar + cJaar² + cJaar³, waarbij cJaar = Jaar – 2000, zodat de start van de studie in 2000 het referentiejaar is). We gaan ook na of Ecoregio een rol speelt, en of de trend verschilt tussen de ecoregio's. Ook nu waren de verschillen tussen de meetplaatsen groot. We modelleren deze verschillen als toevallige normale fluctuaties ten opzichte van een intercept (notatie: 1 | Meetplaats).

Voegen we al deze modeltermen samen, dan krijgen we volgend startmodel.

logAs ~ 1|Meetplaats + (cJaar + cJaar² + cJaar³) + Ecoregio
+ (cJaar + cJaar² + cJaar³) : Ecoregio

6.1.2 Keuze van het model (modelreductie en modelverfijning)

(a) Modelreductie

R-output 18 geeft de anova-tabel voor de modelselectie van arseen. Er is nu minder twijfel dan voorheen of een vereenvoudiging gerechtvaardigd is of niet. De p-waarde voor de reductie van poly(cJaar, 3) naar poly(cJaar, 2) is 0.1234 met een iets lagere AIC voor het model met de tweedegraadsvergelijking, verdere vereenvoudiging naar poly(cJaar, 1) levert een p-waarde van 0.7038 en een duidelijk lagere AIC, en voor de overblijvende interactie tussen poly(cJaar, 1) en Ecoregio bedraagt de p-waarde 0.9428. We eindigen nu dus met dit eenvoudigere model.

logAs ~ 1|Meetplaats + cJaar + Ecoregio

R-output 18: Modelselectie voor arseen – cJaar (MJ = derdegraadsvergelijking, MJa = tweedegraadsvergelijking, MJb = eerstegraadsvergelijking, MJc = eerstegraadsvergelijking zonder interactie)

```
    Model df      AIC      BIC    logLik   Test  L.Ratio p-value
MJ      1 18 745.7732 839.8057 -354.8866
MJa     2 14 745.0208 818.1571 -358.5104 1 vs 2 7.247553  0.1234
MJb     3 10 739.1950 791.4352 -359.5975 2 vs 3 2.174175  0.7038
MJc     4  7 733.5823 770.1505 -359.7912 3 vs 4 0.387348  0.9428
```

De parameterschattingen van dit vereenvoudigde model voor arseen zijn weergegeven in R-output 19. Voorlopig geven we hier echter nog geen interpretatie aan, vermits we nog een kleine uitbreiding aan het model zullen toevoegen.

R-output 19: Het gereduceerde model voor de trendanalyse van arseen – cJaar

Anova

```
                numDF denDF   F-value p-value
(Intercept)         1   735 2932.3726  <.0001
poly(cJaar, 1)      1   735    8.9040  0.0029
Ecoregio            3   632   14.9966  <.0001
```

Summary

```
Linear mixed-effects model fit by REML
 Data: AllData
       AIC      BIC    logLik
  764.3446 800.8872 -375.1723

Random effects:
 Formula: ~1 | Meetplaats
        (Intercept)  Residual
StdDev:   0.3683467 0.1912658

Fixed effects: logRespons ~ cJaar + Ecoregio
```

```
                Value Std.Error  DF t-value p-value
(Intercept)   0.9321019 0.02651642 735 35.15187  0.0000
cJaar         0.0059637 0.00200547 735  2.97372  0.0030
EcoregioKlei -0.0955671 0.04551839 632 -2.09953  0.0362
EcoregioLeem -0.2256750 0.03963657 632 -5.69361  0.0000
EcoregioZand -0.2317663 0.04394434 632 -5.27409  0.0000

Number of Observations: 1372
Number of Groups: 636
```

(b) Modeluitbreiding met een random slope

Aangezien de trend nu lineair is, overwegen we ook om een random slope toe te voegen aan het model (notatie: cJaar | Meetplaats), zodat de trend kan verschillen per meetplaats. Voor cadmium was dit eerder niet mogelijk, omdat er een kwadratische trend gedetecteerd werd, en er zijn slechts 3 metingen per meetplaats, zodat de random (kwadratische) trend perfect door elk punt zou gaan, en een gesatureerd model zou opleveren. Het model met random slope voor arseen noteren we als volgt

logAs ~ cJaar|Meetplaats + cJaar + Ecoregio

R-output 20: Toevoeging van een random slope voor arseen – cJaar (MJd = random intercept, MJe = random intercept en random slope)

```
    Model df      AIC      BIC    logLik   Test L.Ratio p-value
MJd     1  7 764.3446 800.8872 -375.1723
MJe     2  9 761.4928 808.4762 -371.7464 1 vs 2 6.85176  0.0325
```

Uit R-output 20 blijkt dat het model met random slope een betere fit geeft, en we zullen dan ook de parameterschattingen van dit uitgebreidere model verder bespreken.

6.1.3 De parameterschattingen

(a) De fixed effects

In R-output 21 onder Fixed effects zien we dat er een significant stijgende lineaire trend is doorheen de tijd. De gemiddelde logconcentratie arseen stijgt jaarlijks met 0.006 eenheden per jaar. In het gedeelte Simultaneous Tests zien we dat er een hogere logconcentratie arseen gemeten werd voor ecoregio Antropogeen in vergelijking met de andere ecoregio's, maar dat het verschil tussen Antropogeen en Klei nu niet significant is (met Zand en Leem wel). Nu is wel nog de logconcentratie in ecoregio Klei significant hoger dan in Zand en Leem. Tussen deze 2 laatste ecoregio's is er dan weer geen significant verschil waar te nemen is. Er valt op dat het toevoegen van een random slope geen effect heeft op de parameterschattingen van de fixed effects.

R-output 21: Het finale model voor de trendanalyse van arseen – cJaar

Anova

```
            numDF denDF   F-value p-value
(Intercept)     1   735 2943.0844  <.0001
cJaar           1   735    7.9899  0.0048
Ecoregio        3   632   14.9861  <.0001
```

Summary

```
Linear mixed-effects model fit by REML
 Data: AllData
       AIC      BIC    logLik
  761.4928 808.4762 -371.7464

Random effects:
 Formula: ~cJaar | Meetplaats
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev     Corr
(Intercept) 0.37585382 (Intr)
cJaar       0.02054066 -0.199
Residual    0.17840309

Fixed effects: logRespons ~ cJaar + Ecoregio
                Value  Std.Error  DF  t-value p-value
(Intercept)   0.9314981 0.02652635 735 35.11596  0.0000
cJaar         0.0060157 0.00213856 735  2.81298  0.0050
EcoregioKlei -0.0941162 0.04535676 632 -2.07502  0.0384
EcoregioLeem -0.2256681 0.03947978 632 -5.71604  0.0000
EcoregioZand -0.2300014 0.04378621 632 -5.25283  0.0000
```

```
Number of Observations: 1372
Number of Groups: 636

Approximate 95 % confidence intervals
 Fixed effects:
                    lower         est.        upper
(Intercept)   0.87942169  0.931498134  0.983574575
cJaar         0.00181731  0.006015734  0.010214158
EcoregioKlei -0.18318440 -0.094116219 -0.005048041
EcoregioLeem -0.30319550 -0.225668084 -0.148140670
EcoregioZand -0.31598548 -0.230001424 -0.144017365

 Random Effects:
  Level: Meetplaats
                              lower        est.       upper
sd((Intercept))         0.34790118  0.37585382 0.406052360
sd(cJaar)               0.01347908  0.02054066 0.031301736
cor((Intercept),cJaar) -0.38672718 -0.19878972 0.005001535

 Within-group standard error:
    lower      est.     upper
0.1655314 0.1784031 0.1922757

Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Tukey Contrasts

Linear Hypotheses:
                          Estimate Std. Error z value Pr(>|z|)
Klei - Antropogeen == 0 -0.094116   0.045357  -2.075   0.1592
Leem - Antropogeen == 0 -0.225668   0.039480  -5.716   <0.001 ***
Zand - Antropogeen == 0 -0.230001   0.043786  -5.253   <0.001 ***
Leem - Klei == 0        -0.131552   0.048876  -2.692   0.0354 *
Zand - Klei == 0        -0.135885   0.052416  -2.592   0.0464 *
Zand - Leem == 0        -0.004333   0.047422  -0.091   0.9997
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

(b) De random effects

Ook de schatting voor de random effects variabiliteit en de bijhorende betrouwbaarheidsintervallen zijn terug te vinden in R-output 21 onder Random effects. De standaard deviatie van het random intercept (σ_0=0.376) is opnieuw duidelijk groter dan de residuele standaard deviatie (σ_e=0.178), waaruit we kunnen concluderen dat de variabiliteit tussen de meetplaatsen groter is dan de resterende variabiliteit binnen de meetplaatsen die niet verklaard kan worden door het model.

Deze resterende variabiliteit σ_e is nu echter nog kleiner dan in het model met enkel een random intercept (R-output 19, σ_e=0.191) omdat er nog een extra random slope toegevoegd werd aan het model, die nu een gedeelte van de eerder onverklaarde variabiliteit wel verklaart, met een significante standaard deviatie (σ_1=0.021).

Er wordt nu eveneens een correlatie berekend tussen het random intercept en de random slope (ρ = -0.199), en die geeft aan dat de logconcentratie van meetplaatsen met een lagere concentratie in de beginjaren sneller zal toenemen dan voor meetplaatsen met een hogere concentratie in de beginjaren (deze zullen trager toenemen of misschien zelfs afnemen). Het betrouwbaarheidsinterval voor deze correlatie gaat echter van -0.387 tot +0.005, zodat deze relatie niet significant is.

6.1.4 Modeldiagnose

Ook voor het finale model voor arseen voeren we een validatie van de modelveronderstellingen uit, en gebruiken hiervoor de gestandaardiseerde residuals.

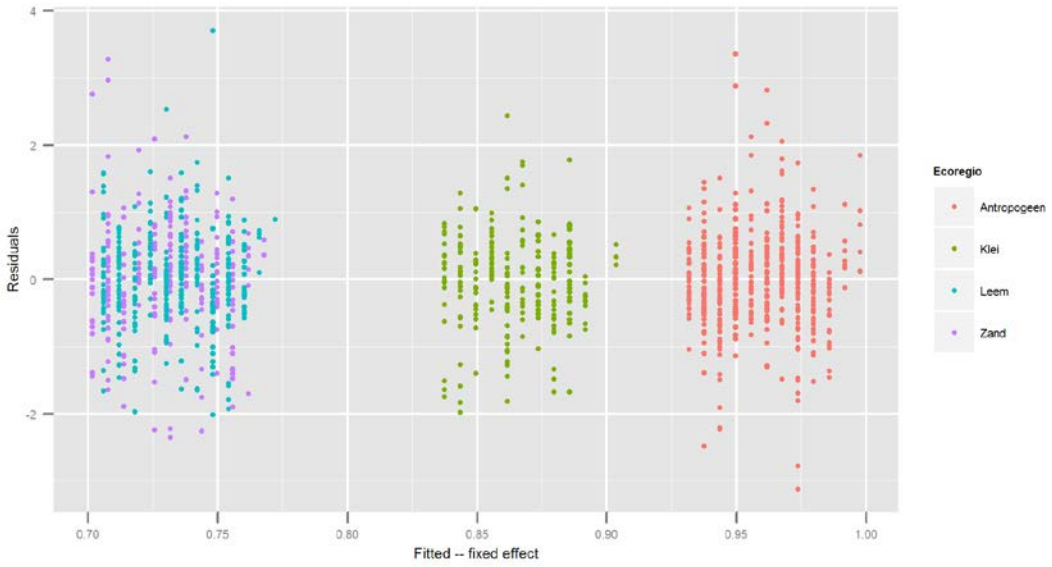
(a) Residu-plot in functie van de fitted values.

Figuur 34 vertoont, uitgezonderd enkele uitschieters, geen duidelijke onverklaarde patronen tussen de fitted values (met de fixed effects cJaar en Ecoregio) en de residuals.

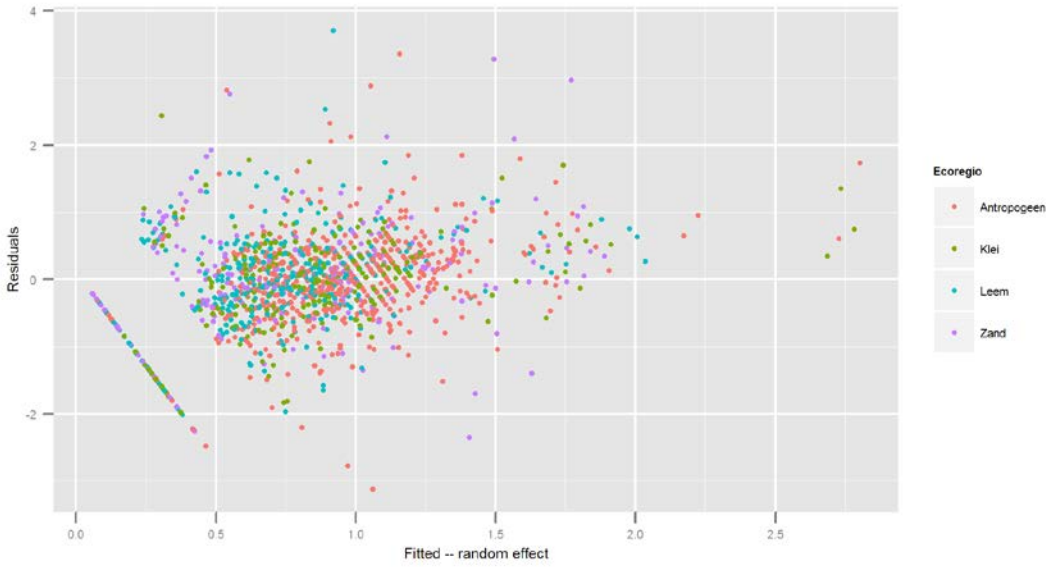
Wanneer ook de random effects mee in rekening gebracht worden om de fitted values te bepalen (Figuur 35), dan zien we weer een gelijkaardige figuur zoals voor cadmium, met linksonderaan een schuine lijn voor alle waarnemingen die herleid zijn naar de halve maximale detectielimiet, en evenwijdig daarmee een

ondergrens voor de puntenwolk, bepaald door deze maximale detectielimiet. Verder zien we dat er veel minder waarnemingen zijn met een fitted value groter dan 1.5 dan met een kleinere fitted value.

Figuur 34: Modeldiagnose voor arseen – cJaar: residuals versus fitted values (fixed effects)



Figuur 35: Modeldiagnose voor arseen: residuals versus fitted values (fixed + random effects)



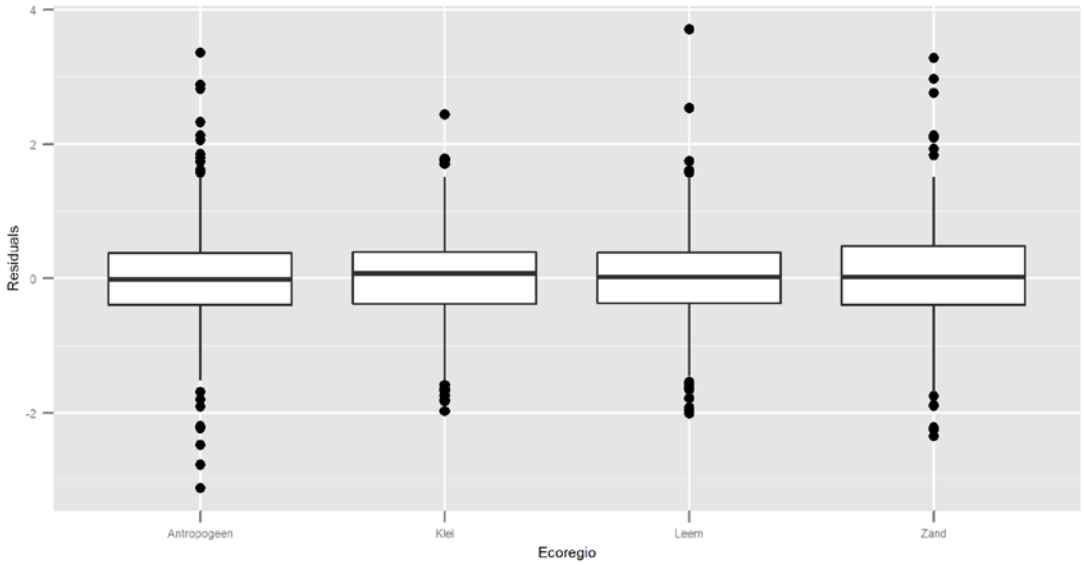
(b) <u>Residu-plot in functie van de verklarende variabelen.</u>

De variabiliteit van de residuals in de verschillende ecoregio's is heel gelijkaardig (Figuur 36) en er is geen reden om aan te nemen dat de voorwaarde van homogeniteit niet voldaan is.

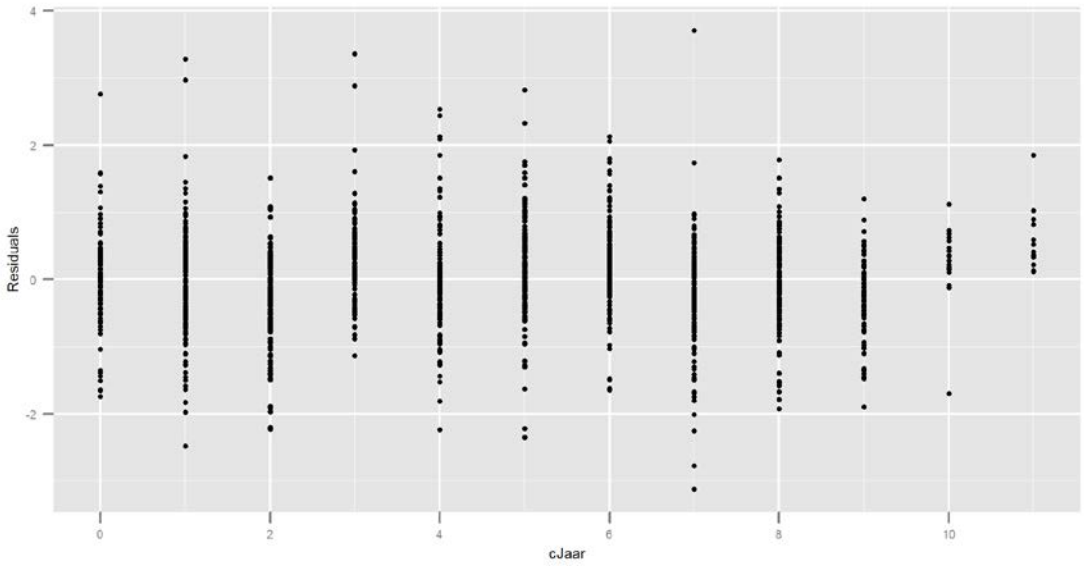
Opvallend in Figuur 37 is dat de spreiding van de residuals voor cJaar gelijk aan 10 of 11 (jaren 2010 en 2011) aanzienlijk kleiner is dan in voorgaande jaren. Dit kan verklaard worden door het beperkter aantal observaties in deze laatste 2 jaren (19 en 13 respectievelijk, in vergelijking met > 100 voor de andere jaren), vermits er veel van de observaties in deze periode verwijderd werden omwille van detectielimiet issues. Ook is de spreiding van de residuals niet overal symmetrisch rond 0. Hiervoor zou mogelijks nog gecorrigeerd kunnen worden door de heterogeniteit expliciet te gaan modelleren, maar dit valt buiten de scope van dit project.

Figuur 38 laat opnieuw een puntenwolk zien met de residuals per meetplaats. Ondanks de overvloed aan meetplaatsen (445) en het beperkt aantal metingen per meetplaats (3) lijkt, op een beperkt aantal uitschieters na (~5 %), de variabiliteit binnen de meetplaatsen een constante.

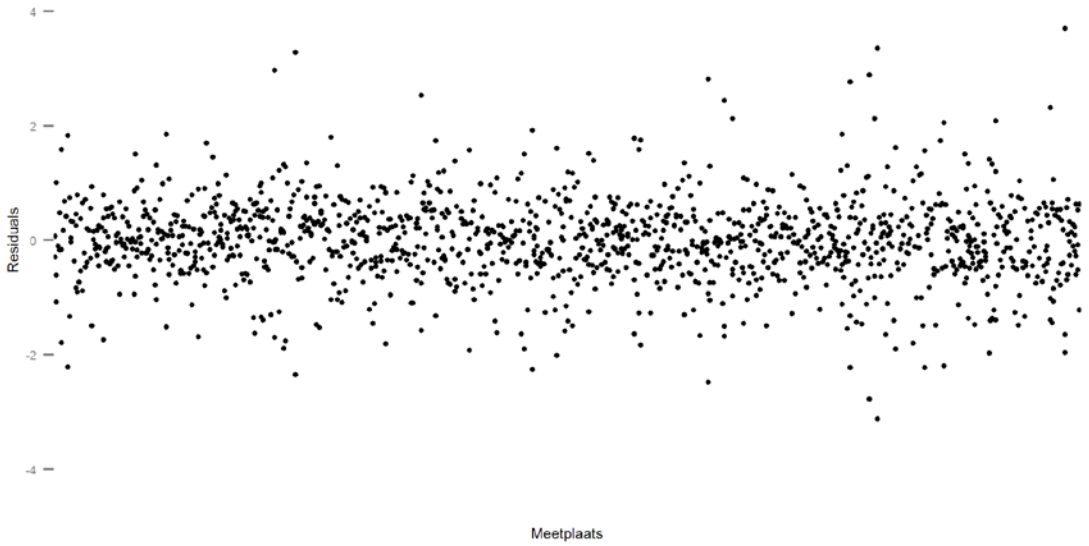
Figuur 36: Modeldiagnose voor arseen – cJaar: Residuals opgesplitst per Ecoregio



Figuur 37: Modeldiagnose voor arseen – cJaar: Residuals opgesplitst per cJaar



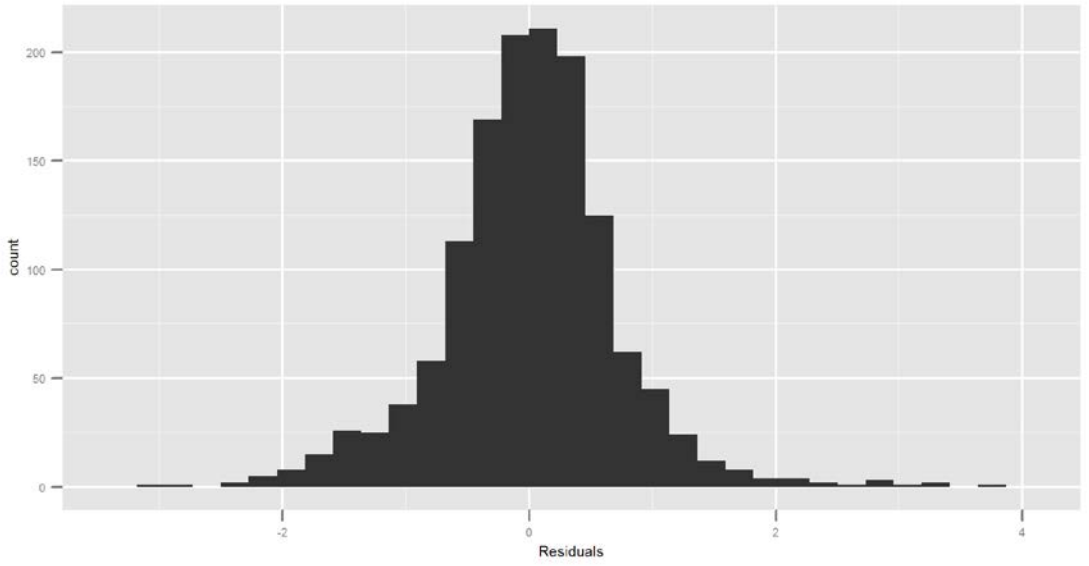
Figuur 38: Modeldiagnose voor arseen – cJaar: Residuals opgesplitst per meetplaats



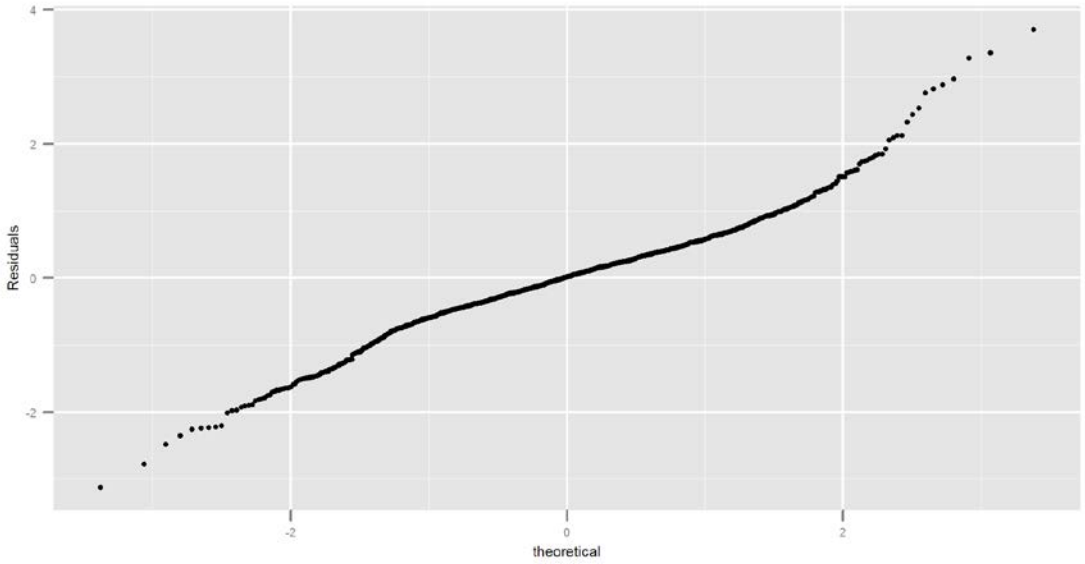
(c) Normaliteit van de *residuals*, het random intercept en de random slope

Figuur 39 en Figuur 40 onderzoeken de normaliteit van de residuals en vertonen geen duidelijke afwijkingen. De staarten in de QQ-plot voor het random intercept buigen beiden naar boven (Figuur 41), zodat de normaliteit in twijfel getrokken kan worden. De lijn die door de punten getrokken kan worden, valt ook nu niet meer samen met de eerste bissectrice (door de oorsprong en helling gelijk aan 1) maar moet een helling hebben gelijk aan σ_0. Ook voor de random slope kan de normaliteit in twijfel getrokken worden (Figuur 42), maar vermoedelijk zijn deze afwijkingen opnieuw te wijten aan de waarnemingen onder de detectielimiet. Hier moet de lijn door de punten een helling hebben gelijk aan σ_1.

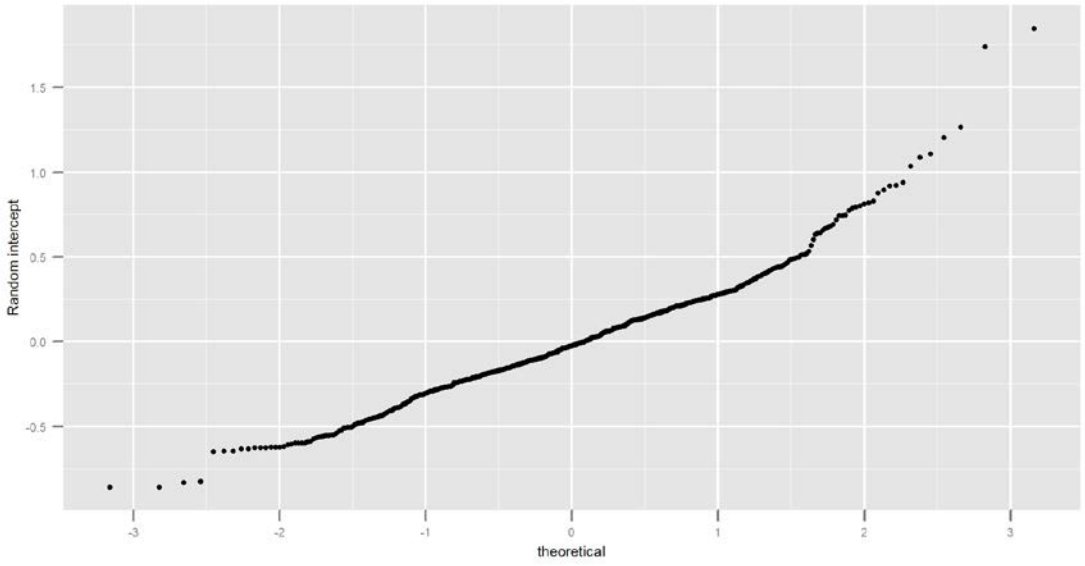
Figuur 39: Modeldiagnose voor arseen – cJaar: Histogram van de residuals



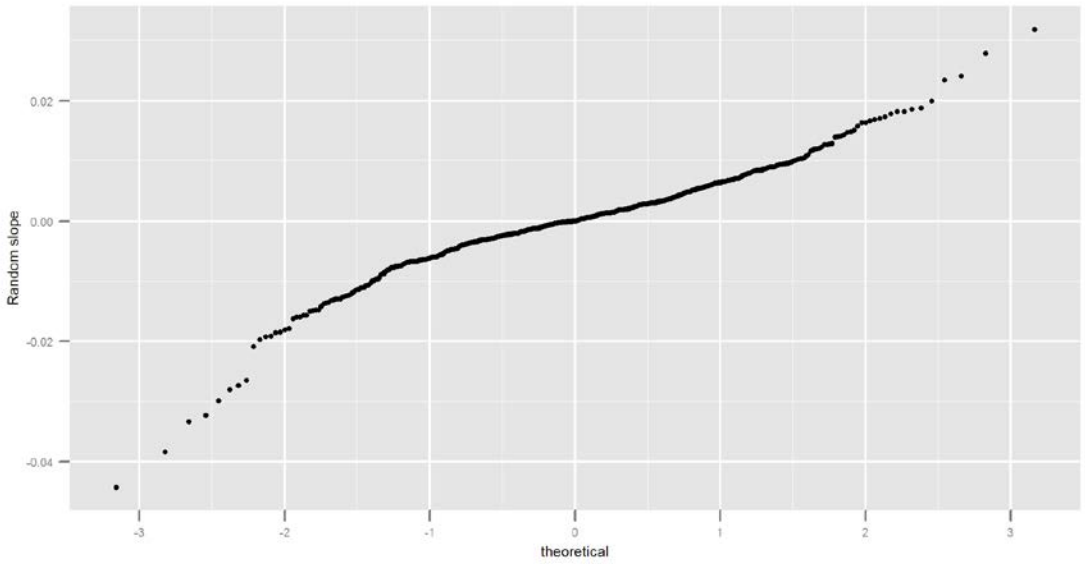
Figuur 40: Modeldiagnose voor arseen – cJaar: QQ-plot van de residuals



Figuur 41: Modeldiagnose voor arseen – cJaar: QQ-plot van het random intercept



Figuur 42: Modeldiagnose voor arseen – cJaar: QQ-plot van de random slope

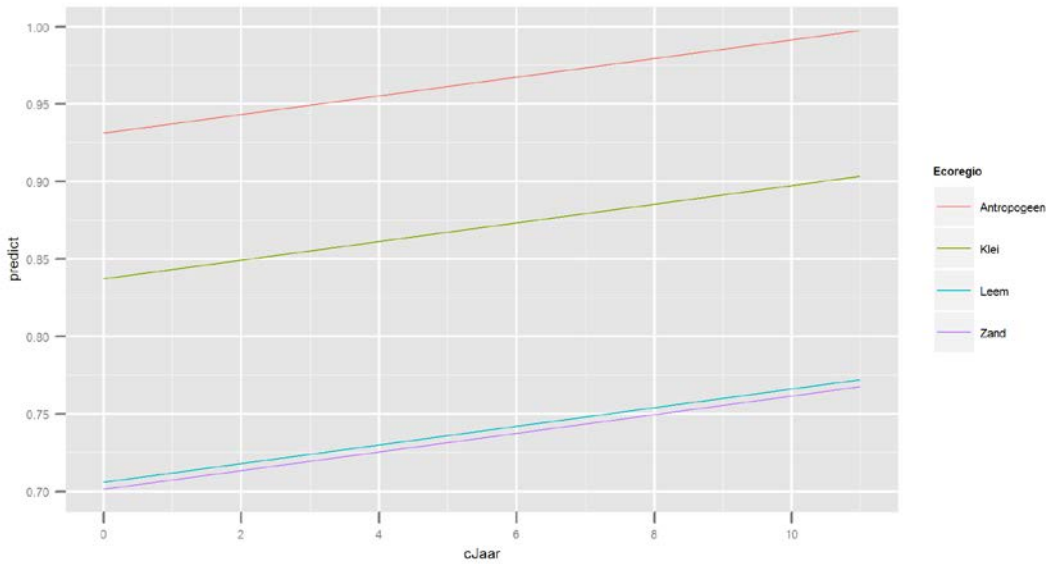


6.1.5 Grafische voorstelling van het finale model

Op basis van de parameterschattingen kunnen we nu het finale model grafisch voorstellen om de impact van de verschillende factoren beter te begrijpen. Het model geeft aan dat er een significante lineaire trend is in de tijd en dat er significante verschillen zijn tussen de ecoregio's. We beginnen met het fixed effect gedeelte voor te stellen. Vervolgens voegen we ook de random effecten van de meetplaatsen toe aan de figuur.

Figuur 43 laat duidelijk zien dat de logconcentratie arseen in de ecoregio Antropogeen beduidend hoger ligt dan in de andere ecoregio's, en dat ook in de ecoregio Klei een significant hogere logconcentratie arseen gemeten werd. Deze keer bevatte het model een lineair effect van cJaar, en deze trend is duidelijk stijgend.

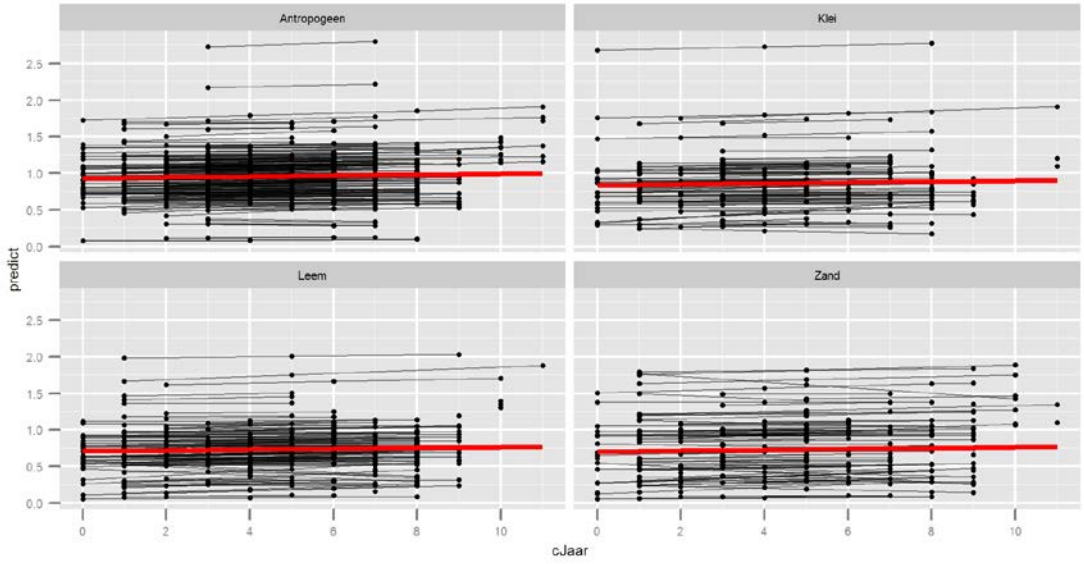
Figuur 43: Grafische voorstelling van het finale model voor arseen – cJaar



In Figuur 44 werden naast de 4 curves voor de verschillende ecoregio's (rode lijnen, enkel fixed effect voor de desbetreffende ecoregio en trend) ook de individueel voorspelde punten per meetplaats weergegeven (fixed effect van ecoregio, trend in de jaren waarin een opmeting gebeurde, en random effect van de meetplaats) en verbonden met zwarte lijnen. De toevoeging van een random slope aan het model laat toe

dat de lineaire trends verschillen tussen de meetplaatsen. Soms stijgt de concentratie sneller, soms trager dan de gemiddelde trend, en voor sommige meetplaatsen is er zelfs een afname vast te stellen.

Figuur 44: Grafische voorstelling van het finale model voor arseen – cJaar, opgesplitst per ecoregio



6.2 Meetcyclus als tijdsvariabele

6.2.1 Het startmodel

Als alternatief gebruiken we opnieuw meetcyclus om de trend in de gegevens te modelleren, en starten met de categorische variabele fMeetcyclus. Ook nu introduceren we meetplaats als een random effect (notatie: 1 | Meetplaats) en we kunnen nagaan of Ecoregio een rol speelt. Voegen we al deze modeltermen samen, dan krijgen we volgend globaal model.

logAs ~ 1|Meetplaats + fMeetcyclus + Ecoregio + fMeetcyclus : Ecoregio

6.2.2 Keuze van het model (modelreductie en modelverfijning)

(a) Modelreductie

Een eerste vereenvoudiging die we uitvoeren, is de interactie tussen fMeetcyclus en Ecoregio (R-output 22).

R-output 22: Modelselectie voor arseen – fMeetcyclus (MM0 = met interactie, MM1 = zonder interactie)

```
     Model df      AIC      BIC    logLik   Test  L.Ratio p-value
MM0      1 14 743.7630 816.8993 -357.8815
MM1      2  8 736.9627 778.7549 -360.4813 1 vs 2 5.199708  0.5185
```

R-output 23: Het gereduceerde model voor arseen – fMeetcyclus

```
Linear mixed-effects model fit by maximum likelihood
 Data: AllData
       AIC      BIC    logLik
  736.9627 778.7549 -360.4813

Random effects:
 Formula: ~1 | Meetplaats
        (Intercept) Residual
StdDev:   0.3677745 0.191043
```

```
Fixed effects: logRespons ~ fMeetcyclus + Ecoregio
                   Value  Std.Error  DF  t-value p-value
(Intercept)   0.9422127 0.02569247 734 36.67272  0.0000
fMeetcyclus2  0.0281338 0.01167841 734  2.40905  0.0162
fMeetcyclus3  0.0354177 0.01730325 734  2.04688  0.0410
EcoregioKlei -0.0962499 0.04555795 632 -2.11269  0.0350
EcoregioLeem -0.2278274 0.03968014 632 -5.74160  0.0000
EcoregioZand -0.2328420 0.04398974 632 -5.29310  0.0000

Number of Observations: 1372
Number of Groups: 636
```

(b) Meetcyclus als een continue variabele

Bekijken we nu de resultaten van het gereduceerde model (R-output 23), dan stellen we vast dat er een significante stijging is van meetcyclus 1 naar meetcyclus 2, en een verdere stijging naar meetcyclus 3. Vandaar (en ook een beetje geleid door de resultaten in sectie 6.1) zullen we nagaan of de verschillen tussen de meetcycli kunnen wijzen op een lineaire stijging d.m.v. het volgende model waarbij Meetcyclus geen factor variabele meer is, maar wel continu:

logAs ~ 1|Meetplaats + Meetcyclus + Ecoregio

R-output 24: Verdere modelselectie voor arseen – Meetcyclus (MM1 = fMeetcyclus als factor, MM2 = Meetcyclus continu)

```
    Model df      AIC      BIC    logLik   Test  L.Ratio p-value
MM1     1  8 736.9627 778.7549 -360.4813
MM2     2  7 735.7378 772.3060 -360.8689 1 vs 2 0.775156  0.3786
```

(c) Toets of random slope zinvol

Uit R-output 24 is nu duidelijk dat de vereenvoudiging naar een lineaire trend volgens meetcyclus het model verbetert. Zoals in sectie 6.1 kunnen we, omwille van die lineaire trend, de random effects structuur nu ook uitbreiden met een random slope, zodat we voor het huidige model voorlopig geen verdere interpretatie zullen geven, maar nagaan of de uitbreiding naar het onderstaande model een verbetering oplevert:

logAs ~ Meetcylus|Meetplaats + Meetcylus + Ecoregio

R-output 25: Toevoeging van een random slope voor arseen – Meetcyclus (MM2 = random intercept, MM3 = random slope)

```
    Model df     AIC     BIC   logLik   Test L.Ratio p-value
MM2     1  7 763.7281 800.2707 -374.8641
MM3     2  9 766.4723 813.4556 -374.2361 1 vs 2 1.255853  0.5337
```

R-output 25 geeft echter aan dat een random slope deze keer geen verbetering van het model betekent, en we bespreken nu toch de parameterschattingen van model MM2 (Meetcyclus als continue variabele, Ecoregio en een random intercept).

6.2.3 De parameterschattingen

(a) De regressiecoëfficiënten

De resultaten zijn weergegeven in R-output 26. Uit de fixed effects output kunnen we afleiden dat de gemiddelde logconcentratie arseen tussen de meetcycli significant stijgt met 0.02 eenheden. Uit het gedeelte Simultaneous Tests volgt opnieuw dat er een hogere logconcentratie arseen gemeten werd voor ecoregio Antropogeen in vergelijking met de andere ecoregio's, maar dat het verschil tussen Antropogeen en Klei nu niet significant is (met Zand en Leem nog wel), en dat de logconcentratie in ecoregio Klei significant hoger is dan in Zand en Leem. Tussen deze 2 laatste ecoregio's is er opnieuw geen significant verschil waar te nemen.

R-output 26: Het finale model voor de trendanalyse van arseen – Meetcyclus

```
Summary
Linear mixed-effects model fit by REML
 Data: AllData
       AIC      BIC    logLik
  763.7281 800.2707 -374.8641
```

```
Random effects:
 Formula: ~1 | Meetplaats
        (Intercept)  Residual
StdDev:   0.3688674 0.1913337

Fixed effects: logRespons ~ Meetcyclus + Ecoregio
                 Value  Std.Error  DF  t-value p-value
(Intercept)   0.9241037 0.02840288 735 32.53556  0.0000
Meetcyclus    0.0206329 0.00797860 735  2.58603  0.0099
EcoregioKlei -0.0968472 0.04558050 632 -2.12475  0.0340
EcoregioLeem -0.2282923 0.03970045 632 -5.75037  0.0000
EcoregioZand -0.2337188 0.04400515 632 -5.31117  0.0000

Number of Observations: 1372
Number of Groups: 636
```

Approximate 95 % confidence intervals

```
 Fixed effects:
                    lower        est.        upper
(Intercept)   0.868343260  0.92410371  0.979864157
Meetcyclus    0.004969373  0.02063294  0.036296498
EcoregioKlei -0.186354716 -0.09684716 -0.007339596
EcoregioLeem -0.306253016 -0.22829226 -0.150331497
EcoregioZand -0.320132756 -0.23371876 -0.147304772

 Random Effects:
  Level: Meetplaats
                    lower      est.     upper
sd((Intercept)) 0.3465057 0.3688674 0.3926723

 Within-group standard error:
    lower      est.     upper
0.1818648 0.1913337 0.2012957
```

Simultaneous Tests for General Linear Hypotheses

```
Multiple Comparisons of Means: Tukey Contrasts

Linear Hypotheses:
                         Estimate Std. Error z value Pr(>|z|)
Klei - Antropogeen == 0 -0.096847   0.045581  -2.125   0.1433
Leem - Antropogeen == 0 -0.228292   0.039700  -5.750   <0.001 ***
Zand - Antropogeen == 0 -0.233719   0.044005  -5.311   <0.001 ***
Leem - Klei == 0        -0.131445   0.049108  -2.677   0.0366 *
Zand - Klei == 0        -0.136872   0.052653  -2.599   0.0454 *
Zand - Leem == 0        -0.005427   0.047648  -0.114   0.9995
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

(b) Het random intercept en de ruisterm

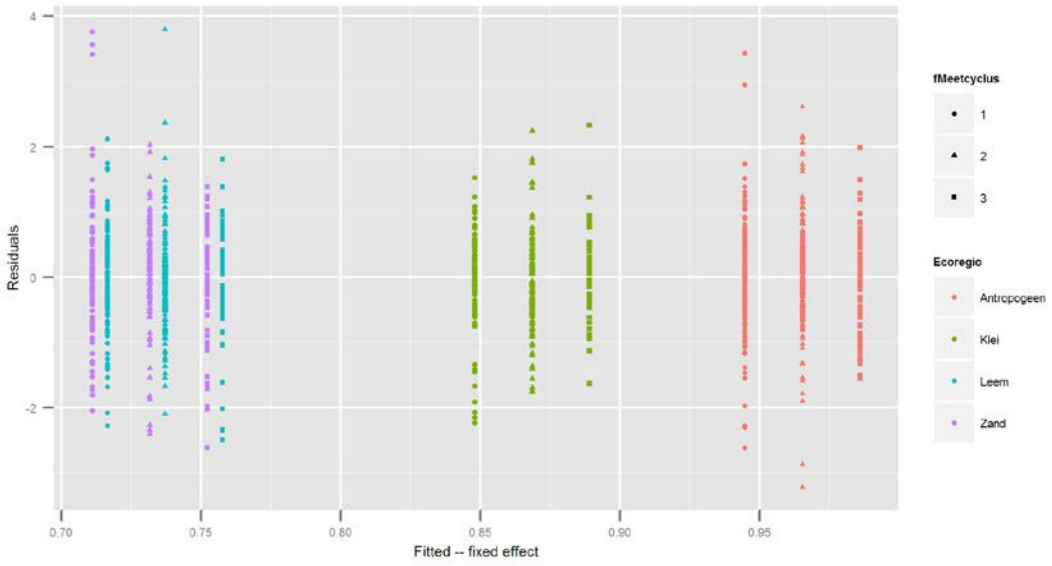
Voor het random intercept bestuderen we de variabiliteit en de bijhorende betrouwbaarheidsintervallen in R-output 26 onder Random effects. De standaard deviatie van het random intercept (σ_O=0.369) is opnieuw duidelijk groter dan de residuele standaard deviatie (σ_e=0.191), waaruit we kunnen concluderen dat de variabiliteit tussen de meetplaatsen groter is dan de resterende variabiliteit binnen de meetplaatsen die niet verklaard kan worden door het model. De bijhorende BI zijn heel smal, zodat deze schattingen vrij nauwkeurig zijn.

6.2.4 Modeldiagnose

(a) Residuplots ten opzichte van de fitted values

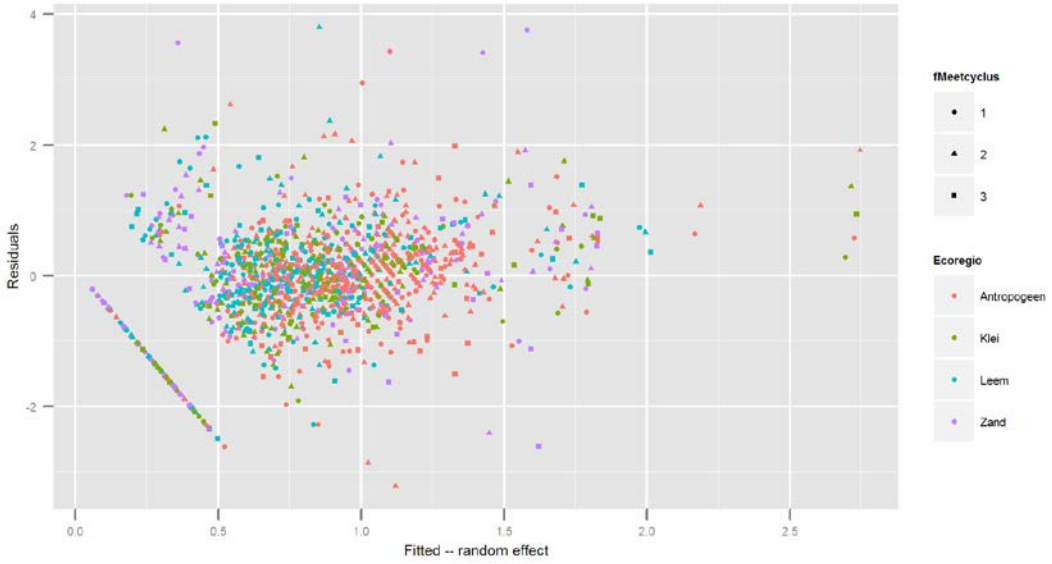
Figuur 45 vertoont, uitgezonderd enkele uitschieters, weer geen duidelijke onverklaarde patronen tussen de fitted values (met de fixed effects Meetcyclus en Ecoregio) en de residuals.

Figuur 45: Modeldiagnose voor arseen – Meetcyclus: Fitted values (fixed effects) versus residuals



Wanneer ook het random intercept mee in rekening gebracht worden om de fitted values te bepalen (Figuur 46), dan zien we weer een gelijkaardige figuur zoals in alle voorgaande situaties, namelijk linksonderaan een schuine lijn voor alle waarnemingen die herleid zijn naar de halve maximale detectielimiet, en evenwijdig daarmee een ondergrens voor de puntenwolk, bepaald door deze maximale detectielimiet.

Figuur 46: Modeldiagnose voor arseen – Meetcyclus: Fitted values (fixed + random effects) versus residuals



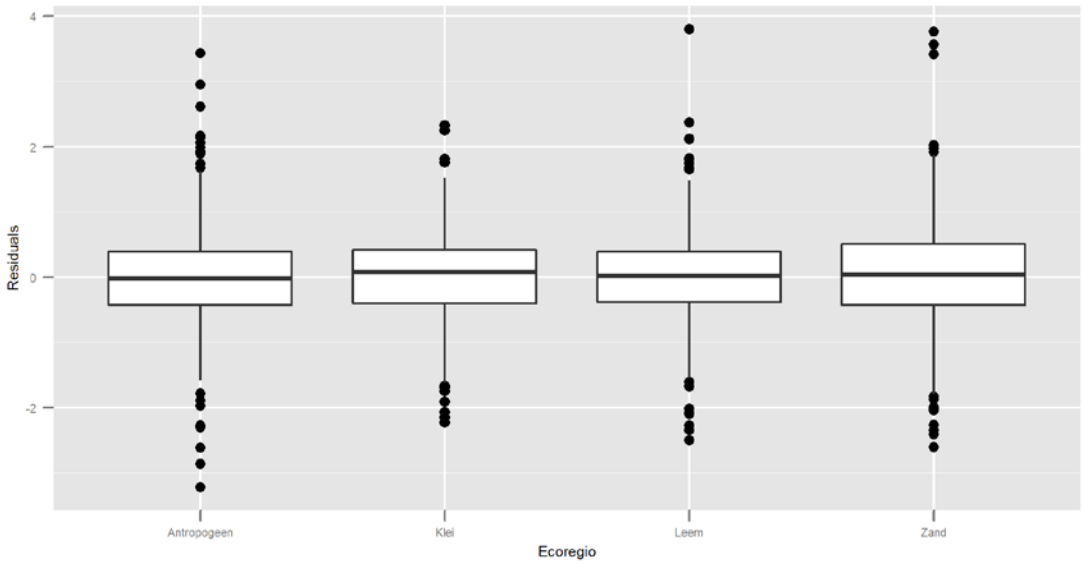
(b) <u>Residuplots ten opzichte van de verklarende variabelen</u>

De variabiliteit van de residuals vertoont geen duidelijke verschillen tussen de verschillende ecoregio's (Figuur 47), dus er is geen reden om de homogeniteit van de residuals in twijfel te trekken.

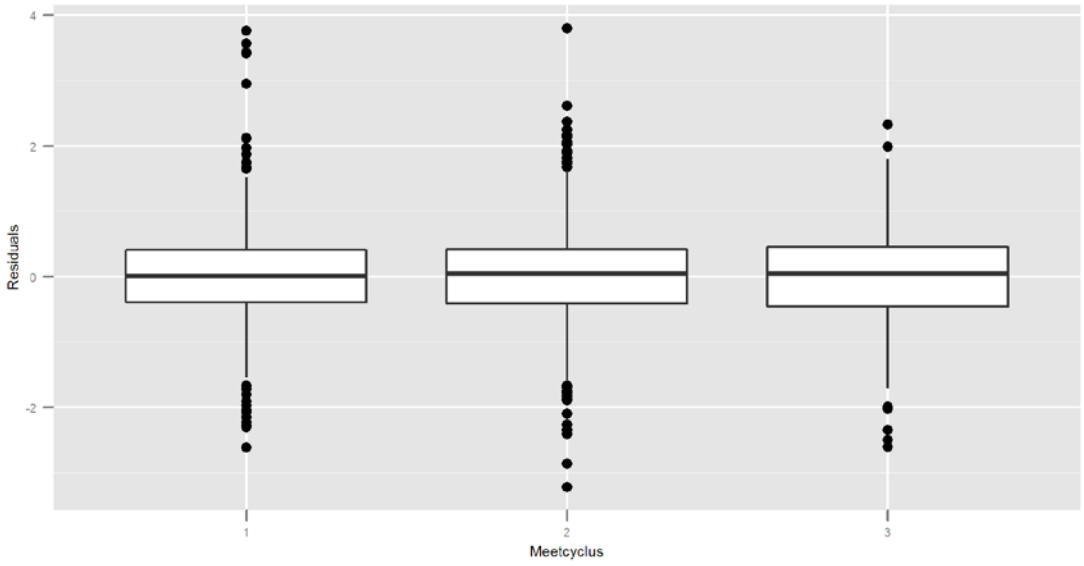
De spreiding van de residuals tussen de verschillende meetcycli is heel gelijkaardig (Figuur 48), dus opnieuw lijkt de homogeniteit van de residuals in orde te zijn.

In de puntenwolk in Figuur 49 (residuals per meetplaats) lijkt ook nu op een beperkt aantal uitschieters na (~5 %), de variabiliteit binnen de meetplaatsen een constante (voor zover de 3 punten per meetplaats te onderscheiden zijn).

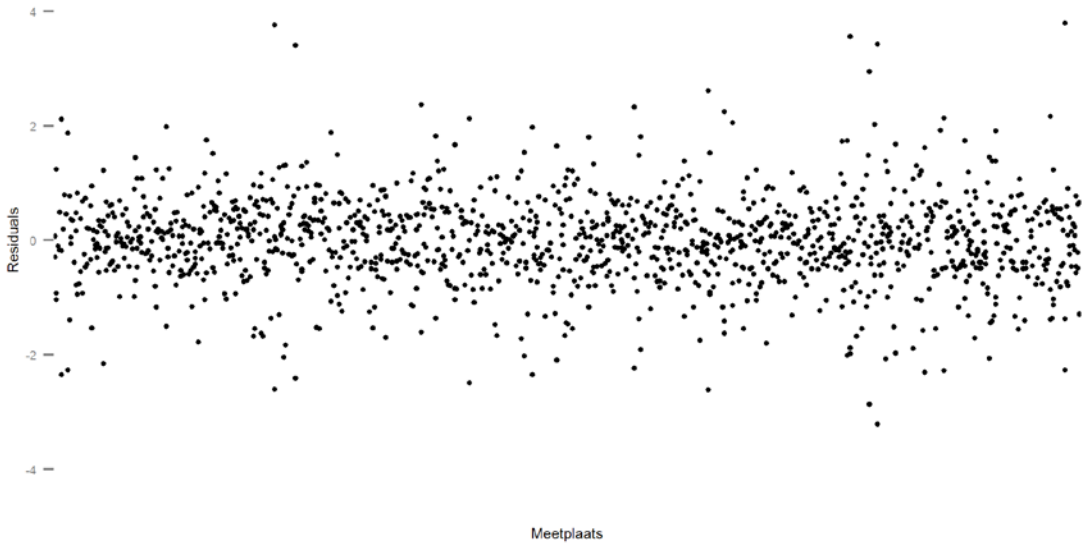
Figuur 47: Modeldiagnose voor arseen – Meetcyclus: Residuals opgesplitst per Ecoregio



Figuur 48: Modeldiagnose voor arseen – Meetcyclus: Residuals opgesplitst per Meetcyclus



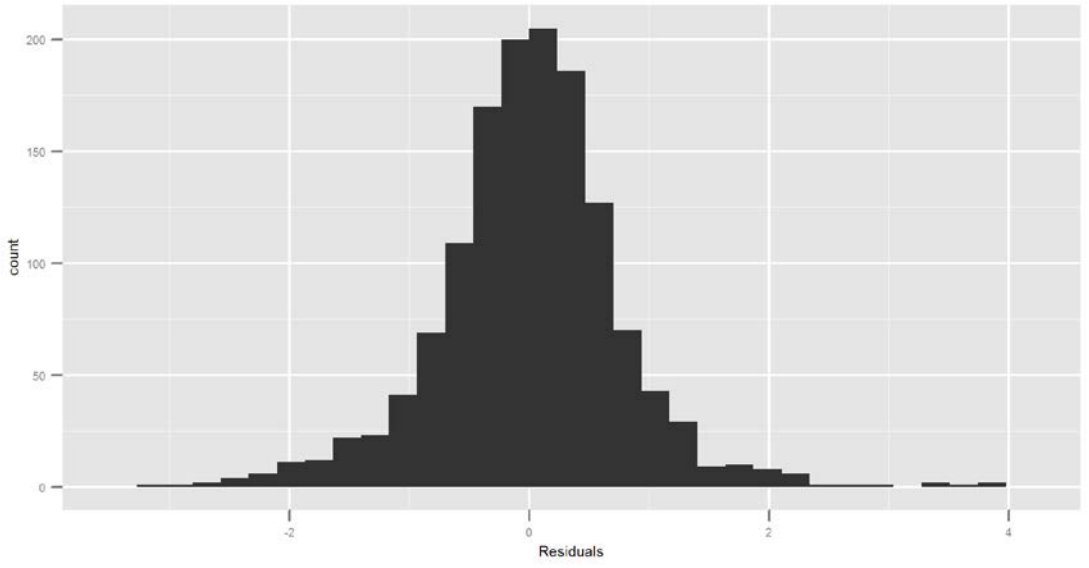
Figuur 49: Modeldiagnose voor arseen – Meetcyclus: Residuals opgesplitst per meetplaats



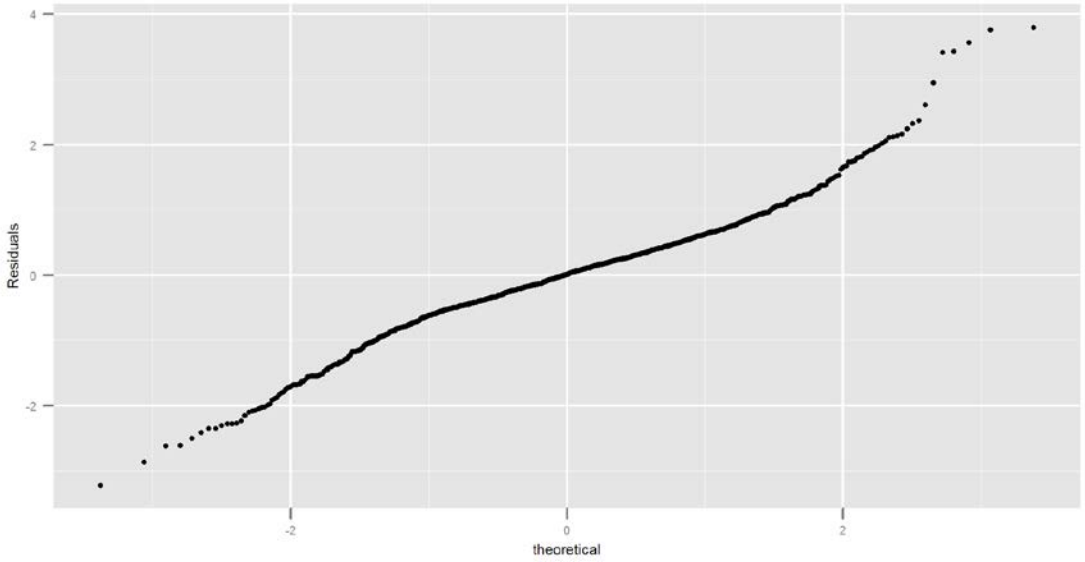
(c) Normaliteit van de *residuals* en het random intercept

Figuur 50 en Figuur 51 onderzoeken de normaliteit van de residuals en vertonen geen duidelijke afwijkingen. Voor het random intercept daarentegen kan de normaliteit weer in vraag getrokken worden. Beide staarten in de QQ-plot (Figuur 52) buigen naar boven. Vermoedelijk is dit opnieuw te wijten aan de problemen met de detectielimiet.

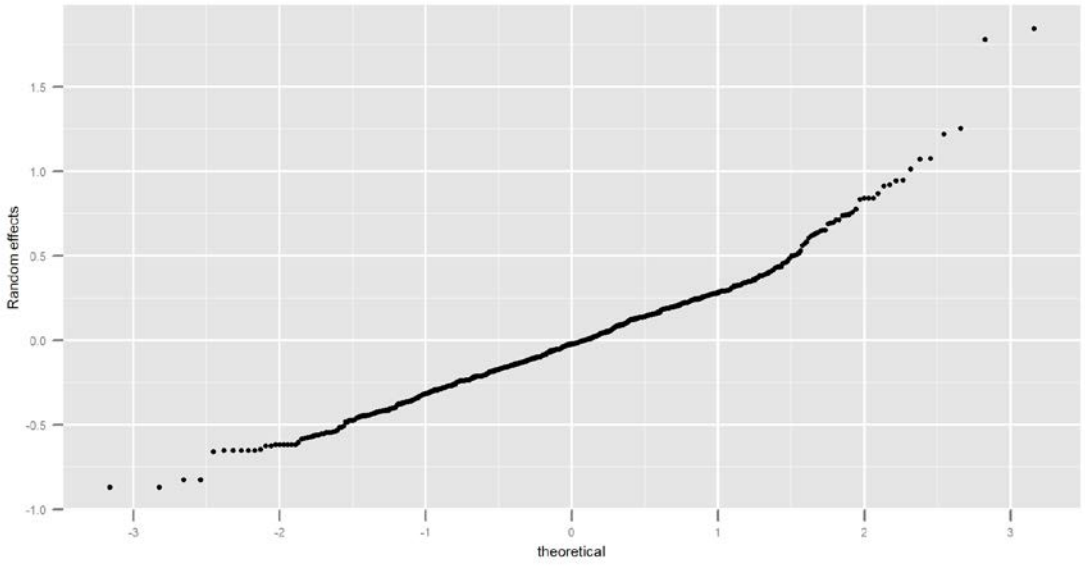
Figuur 50: Modeldiagnose voor arseen – Meetcyclus: Histogram van de residuals



Figuur 51: Modeldiagnose voor arseen – Meetcyclus: QQ-plot van de residuals



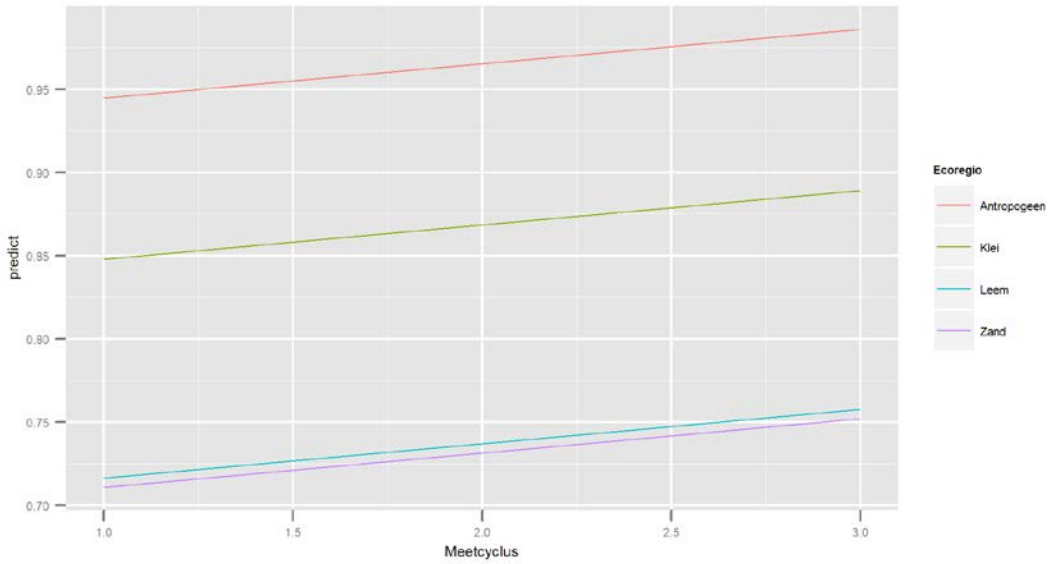
Figuur 52: Modeldiagnose voor arseen – Meetcyclus: QQ-plot van het random intercept



6.2.5 Grafische voorstelling van het finale model

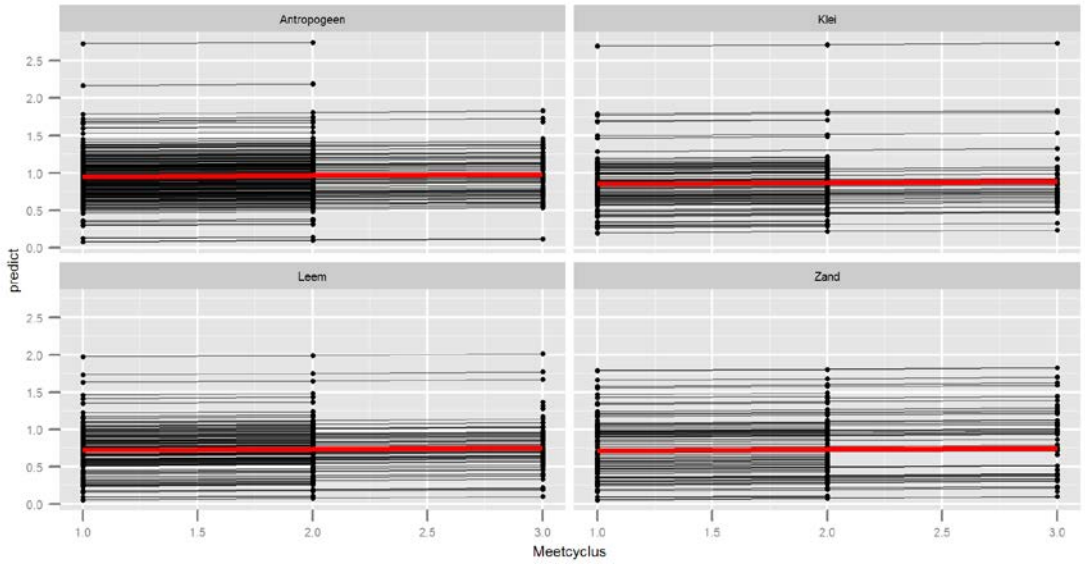
We visualiseren nu het model op basis van de parameterschattingen. Figuur 53 laat duidelijk zien dat de logconcentratie arseen in de ecoregio Antropogeen beduidend hoger ligt dan in de andere ecoregio's, en dat ook in de ecoregio Klei een significant hogere logeconcentratie arseen gemeten werd. De lineair stijgende trend is duidelijk zichtbaar.

Figuur 53: Grafische voorstelling van het finale model voor arseen – Meetcyclus



In Figuur 54 werden naast de 4 curves voor de verschillende ecoregio's (rode lijnen, enkel fixed effect voor de desbetreffende ecoregio en trend) ook de individueel voorspelde punten per meetplaats weergegeven (fixed effect van ecoregio, effect van meetcyclus waarin een opmeting gebeurde, en random effect van de meetplaats) en verbonden met zwarte lijnen. Vermits er nu enkel een random intercept toegevoegd werd aan het model en geen random slope, lopen de lineaire trends voor de meetplaatsen terug parallel (in tegenstelling tot Figuur 44).

Figuur 54: Grafische voorstelling van het finale model voor arseen – Meetcyclus, opgesplitst per ecoregio



6.3 Besluiten

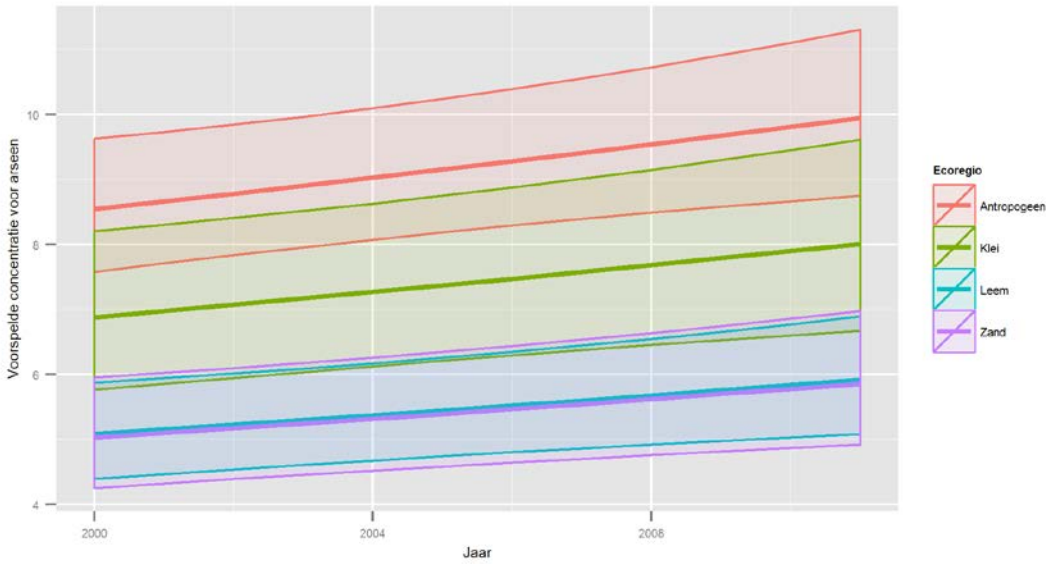
Voor arseen kunnen we eveneens besluiten dat de concentratie significant verschillend is tussen de ecoregio's, met de hoogste concentratie voor ecoregio Antropogeen, gevolgd door ecoregio Klei. De ecoregio's Leem en Zand verschillen niet significant van mekaar.

Bij beide manieren om de tijdsvariabele te modelleren (enerzijds met cJaar, anderzijds met Meetcyclus) bevatte het resulterende model geen interactie tussen tijd en ecoregio, en werd een lineaire stijging van de concentratie vastgesteld.

Figuur 55 laat bovenstaande conclusies duidelijk zien. De lineair stijgende trend per ecoregio in de log-schaal vertaalt zich in een exponentiële stijging in de originele schaal (concentratie in mg/kg ds). Doordat de trend zeer klein is (10^0.006 = 1.014, iets meer dan 1 % per jaar) is de trend nagenoeg lineair. De trend loopt voor alle ecoregio's gelijk.

Uit de betrouwbaarheidsbanden in Figuur 55 kunnen we aflezen dat voor de waterlopen in een antropogene ecoregio de gemiddelde concentratie arseen in 2000 met 95 % zekerheid lag tussen 7.6 en 9.6 mg/kg ds en stijgt tot [8.7;11.3] in 2011. De banden worden (exponentieel) breder bij hogere concentraties. Analoge conclusies kunnen afgelezen worden voor de andere ecoregio's. Voor ecoregio's met minder meetplaatsen zijn de banden breder dan voor ecoregio's met veel meetplaatsen. Het aantal waarnemingen heeft immers een invloed op de schatting van de standaardfout, die gebruikt wordt bij de berekening van betrouwbaarheidsintervallen.

Figuur 55: Het finale model voor arseen in de originele schaal (met 95 % betrouwbaarheidsinterval)



Het random effects gedeelte is verschillend naargelang de tijdsvariabele die gebruikt werd. Dit kan verschillende oorzaken hebben. De studie is zo opgezet dat elke meetplaats om de 4 jaar gemeten wordt, zodat dit exact 1x per meetcyclus is. In de praktijk echter zijn hierop enkele uitzonderingen. Niet alle geplande metingen werden uitgevoerd in het jaar waarin ze gepland waren, maar vroeger of later, zodat de verschillen in cJaar voor eenzelfde meetplaats niet altijd exact gelijk zijn aan een veelvoud van 4 (dit is het geval voor 68 waarnemingen). Voor meetcyclus is dit echter wel altijd een veelvoud van 1. Dit kan een invloed hebben op de beschikbare informatie om een random slope effect als significant te kunnen bestempelen. Ook zijn er 20 meetplaatsen meermaals in 1 cyclus gemeten, maar in verschillende jaren. Voor het model met cJaar als tijdsvariabele leverde dit extra informatie op (2 verschillende punten in de tijd), voor het model met Meetcyclus als tijdsvariabele niet, zodat een random slope effect gemakkelijker gedetecteerd kan worden.

De modelvalidatie laat vooral problemen zien voor de waarnemingen onder de detectielimiet die vervangen werden door de halve maximale detectielimiet. Dit is immers een verlies aan informatie, en een verdwijning van alle variabiliteit in deze metingen.

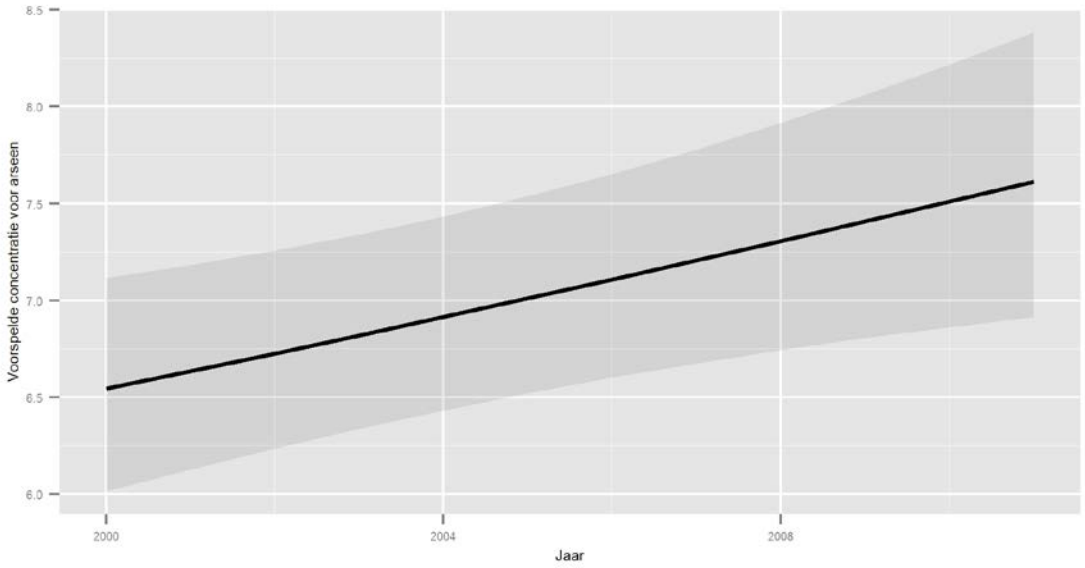
Zoals voor cadmium is het ook voor arseen in het kader van de rapportering interessant om een uitspraak te doen over de globale trend voor het hele meetnet, en niet afzonderlijk over de verschillende ecoregio's. Hier werden minder observaties uit de oorspronkelijke dataset verwijderd (Jansen 2012), maar dezelfde voorwaarde blijft geldig, namelijk dat de resterende steekproef representatief moet zijn voor Vlaanderen. We gaan ervan uit dat de verwijderde observaties geen afbreuk doen aan de representativiteit van de steekproef en herhalen het finale model met een lineaire trend en een random slope, maar nemen we de variabele Ecoregio niet meer op.

R-output 27: Trendanalyse over heel Vlaanderen – Arseen

```
Linear mixed-effects model fit by REML
 Data: AllData
       AIC      BIC    logLik
  785.4803 816.8157 -386.7401

Random effects:
 Formula: ~cJaar | Meetplaats
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev     Corr
(Intercept) 0.38820423 (Intr)
cJaar       0.02055352 -0.187
Residual    0.17855744

Fixed effects: logRespons ~ cJaar
                Value   Std.Error  DF  t-value p-value
(Intercept) 0.8158055 0.018611126 735 43.83429  0.0000
cJaar       0.0059821 0.002142705 735  2.79186  0.0054
 Correlation:
      (Intr)
cJaar -0.509
```

R-output 27 laat de parameterschattingen zien van de algemene trend over Vlaanderen. Voor de fixed effects is het enige verschil met de resultaten in R-output 21 de hogere schatting van het intercept (de ligging van de curve) die nu een “gewogen” gemiddelde is evenredig met steekproefgrootte van de vier ecoregio’s, en niet meer specifiek is voor de ecoregio Antropogeen. De standaarddeviatie van het random intercept is ook nu groter dan voorheen (σ_0=0.388) vermits deze opnieuw de variabiliteit tussen de ecoregio’s moet opvangen. De standaarddeviatie van de random slope (σ_1=0.021) en de residuele standaard deviatie (σ_e=0.179) blijven gelijk. De correlatie tussen het random intercept en de random slope is nu iets kleiner (ρ = -0.187) met betrouwbaarheidsinterval van -0.375 tot +0.015, zodat deze correlatie opnieuw niet significant is. De globale trend over Vlaanderen wordt voorgesteld in Figuur 56. Alleen als de steekproefgrootte per ecoregio de werkelijke verdeling in Vlaanderen weerspiegelt, is de figuur representatief is voor Vlaanderen. Anders moeten we manueel de gewichten per ecoregio aanpassen om een onvertekende schatting te bekomen.

Figuur 56: De trend over Vlaanderen voor arseen in de originele schaal (met 95 % betrouwbaarheidsinterval)



7 Tot besluit

Mixed model regressie is een zeer krachtig instrument onontbeerlijk bij de statistische analyse van meetnetgegevens om recht te doen aan de “complexiteit” ervan.

Deze regressietechniek behandelt geclusterde gegevens afkomstig van eenzelfde meetplaats statistisch correct door expliciet het meetplaatseffect in het model in te bouwen. Elke meetplaats kan afwijkende eigenschappen hebben ten opzichte van de algemene, gemiddelde relatie. De *fixed effects* termen in het model beschrijven de algemene relatie, de afwijkingen van de individuele meetplaatsen worden gemodelleerd als *random effects*, toevallige variaties ten opzichte van de algemene trend. Deze opsplitsing in twee componenten verhoogt de precisie van de schattingen van het algemene model en bijgevolg zal ook het onderscheidend vermogen om een trend te detecteren stijgen (Jansen, 2012).

In dit rapport hebben we de werking van mixed model regressietechnieken en het belang ervan uitgelegd aan de hand van een simulatiestudie en twee tijdsreeksen afkomstig van het VMM waterbodemmeetnet. Het eenvoudigste geval, het *random intercept* model, beschrijft de trend per meetplaats als evenwijdige lijnen parallel met de algemene trend. Maar de modellen laten ook toe om de trend te laten variëren per meetplaats (*random slope*). Naarmate de complexiteit van het model toeneemt, moeten we echter meer parameters schatten. Om die reden zijn de huidige tijdsreeksen te kort om een random effect te modelleren voor niet-lineaire (vb. parabolische) trends. We bevelen in dat geval aan alleen het random intercept in het model op te nemen en zeker niet het model te vereenvoudigen tot een lineaire trend om kost wat kost een random slope in te voeren. Om misleidende resultaten te voorkomen, is het belangrijker om de fixed effects correct te modelleren en een random intercept ondervangt meestal al veel van de verschillen tussen meetplaatsen.

Bij modelbouw in een mixed model context vertrekt men in principe van het model met de meest uitgebreide *fixed* en (indien zinvol) *random effects* structuur. Men gaat van start met het reduceren van de fixed effects structuur (algemene relatie), om daarna na te gaan of het mogelijk is de random effects structuur verder te reduceren.

Modelselectie is echter nooit volledig te objectiveren want zowel de keuze van de criteria en de toepassing ervan zijn nooit helemaal hard te maken. Waar wordt de grens van wel/niet significant gelegd (klassiek bij een p-waarde < 0.05)? Welk criterium wordt gebruikt voor modelselectie (p-waarde, AIC,…)? Het is eveneens belangrijk dat het finale model eenvoudig te interpreteren is en realistische resultaten oplevert. Het is dus cruciaal om bij elke stap goed na te denken over de implicaties ervan. Het grote voordeel van modelbouw is echter wel dat de beslissingen gemotiveerd moeten worden wat bijdraagt tot de transparantie.

Naast modelselectie is ook modelvalidatie een belangrijk aspect van een statistische analyse van gegevens. Het gaat enerzijds na of alle modelveronderstellingen (onafhankelijkheid, homoskedasticiteit en normaliteit) voldaan zijn, maar kan anderzijds ook helpen bij de modelselectie (ontbreken er nog cruciale verklarende variabelen? wat is het effect van het toevoegen/weglaten van een variabele op de modelvalidatie? …). Het is tevens een manier om tekortkomingen of problemen in de gegevens te detecteren, zoals bvb. uitschieters, of artefacten van de data (zoals de detectielimiet). Ook hier is het belangrijk te beseffen dat geen elk model perfect is en dat er een (subjectieve) inschatting nodig is of bepaalde anomalieën het resultaat wezenlijk beïnvloeden of niet. Voor de beschouwde tijdsreeksen is het voornaamste knelpunt de waarden onder de detectielimiet. Wij denken niet dat het probleem de resultaten hier sterk beïnvloedt, maar het blijft wel een belangrijke uitdaging om de aanpak ervan verder op punt te zetten.

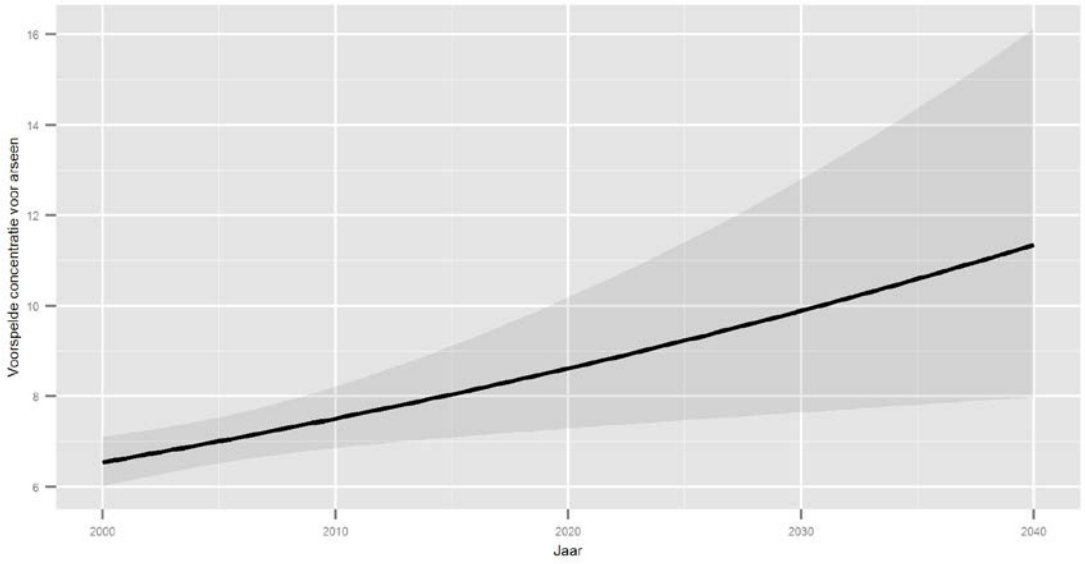
Ten slotte willen we nog enkele kanttekeningen plaatsen over de extrapolatie van de finale modellen naar de toekomst (Figuur 32, Figuur 33, Figuur 55 en Figuur 56).

Eerst en vooral moeten we goed voor ogen houden dat regressiemodellen de trend modelleren, maar niet het onderliggende mechanisme ervan. Bij interpolatie hebben modelfouten een relatief geringe impact, maar bij extrapolatie kan een foutief model aanleiding geven tot grote afwijkingen. Daarenboven kan het onderliggende mechanisme in de toekomst veranderen. Daarom moeten we extrapolaties naar de toekomst altijd heel voorzichtig interpreteren.

Hierbij willen we ook opmerken dat de statistische betrouwbaarheidsintervallen deze fundamentele modelonzekerheid niet weerspiegelen. Ze geven alleen de onzekerheid aan van de trend in de toekomst in de veronderstelling dat het model correct is (en zo blijft in de toekomst). Extrapolaties zijn zinvol als projecties. Stel dat alles gelijk blijft, wat zijn dan de gevolgen en moeten we hierop anticiperen?

Bij wijze van voorbeeld extrapoleert Figuur 57 de schatting voor Arseen (vergelijk met Figuur 56). Hieruit kunnen we afleiden dat de betrouwbaarheidsintervallen heel breed worden. Maar deze band omvat alleen de statistische onzekerheid, maar niet de onzekerheid van het model. We moeten bijgevolg toekomstvoorspellingen heel voorzichtig hanteren. Anderzijds toont de figuur wel degelijk aan dat de trend positief is. We kunnen binnen de band geen horizontale regressierechte tekenen. Dat is niet altijd even duidelijk binnen een kort interval: in Figuur 56 omvat de band wel nog een horizontale rechte. Maar op termijn dus niet meer. Als de helling significant verschillend is van nul, zal de regressieband een horizontale rechte uitsluiten. We moeten hiervoor wel het volledige interval]-∞,+∞[bekijken.

Figuur 57: Extrapolatie van de trend voor Vlaanderen – Arseen



Literatuurlijst

Jansen I. (2012). Mixed model regressietechnieken voor de trendanalyse van de waterbodemgegevens. Tussentijds technisch rapport. Brussel, Belgium: Instituut voor Natuur- en Bosonderzoek (INBO). 1-100 p.

Zuur A.F. et al. (2009). *Mixed Effects Models and Extensions in Ecology with R.* Statistics for Biology and Health, DOI 10.1007/978-0-387-87458-6 1, Springer.

Lijst met afkortingen

AIC: Akaiki Information Criterion

ANOVA: analysis of variance

BI: betrouwbaarheidsinterval

BIC: Bayesian Information Criterion

cJaar: Jaar – 2000

EDA: exploratory data analysis

Iid: independent and identically distributed

LRT: likelihood ratio test

mJaar: Jaar – 2005.5

ML: maximum likelihood

QQ: quantile-quantile

REML: restricted maximum likelihood

SSR: residual sum of squares