

calibratie_handleiding

January 20, 2021

0.1 Kalibratie van een tijdreeksmodel

In algemene zin wordt onder kalibratie verstaan het vinden van de modelparameters zodat de uitkomst van het tijdreeksmodel zo goed mogelijk overeenkomt met de gemeten waarden. De gevonden parameters worden de optimale parameters genoemd. De meest gebruikte methode om de optimale parameters te vinden is de minimalisatie van de som van de kwadraten van de verschillen tussen de gemeten stijghoogte en de gemodelleerde stijghoogte, ook wel de kleinste kwadraten methode genoemd. De som van de kwadraten is een niet-lineaire functie van de modelparameters. Om de optimale modelparameters te vinden wordt gebruikt gemaakt van een zoekmethode.

Er bestaan een aantal verschillende zoekmethoden die gebruikt kunnen worden om de optimale modelparameters te vinden. Iedere zoekmethode is gebaseerd op een ander zoek algoritme, maar het idee is altijd hetzelfde: het model wordt een aantal keer gedraaid met verschillende waarden van de modelparameters. Gebaseerd op de verschillende modeluitkomsten (en de daaruit voortvloeiende som van de kwadraten) bepaald het zoekalgoritme een volgende mogelijke set modelparameters die een kleinere som van de kwadraten zou kunnen opleveren. Het zoekalgoritme gaat door totdat het besluit dat de beste parameters gevonden zijn. De efficiëntie van het zoekalgoritme wordt bepaald door het aantal keer dat het model gedraaid moet worden om de optimale modelparameters te vinden. Het is vanzelfsprekend aan te bevelen om de initiele modelparameters zo dicht mogelijk bij de optimale modelparameters te kiezen. Daartoe heeft de meeste software geprobeerd om slimme keuzes te maken voor de initiele waarden van de modelparameters, maar soms heeft de modelleur kennis van het systeem en kan het handig zijn om de initiele waarden handmatig aan te passen.

Sommige zoekalgoritmes zoeken de optimale modelparameters door de modelparameters steeds een beetje aan te passen zodat de som van de kwadraten kleiner wordt. Dit kan tot gevolg hebben dat het zoekalgoritme in een lokaal minimum belandt. Andere zoekalgoritmes gaan ook set modelparameter draaien die juist om grotere kwadraten som geven met als doel om uiteindelijk het globale minimum te vinden. Dit kost vanzelfsprekend (veel) meer rekentijd.

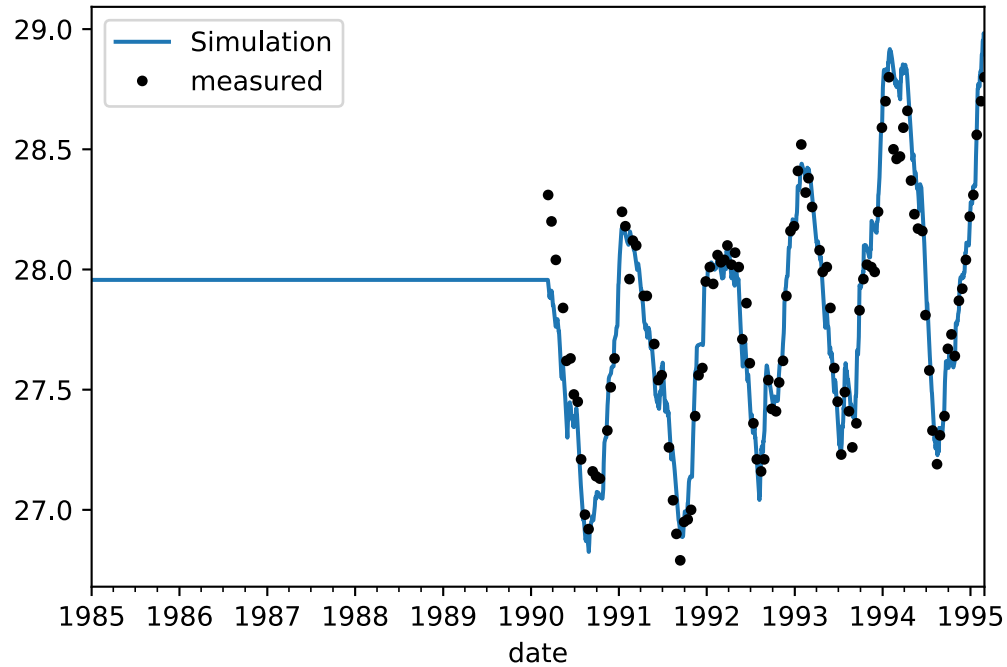
Naast de keuze van de modelstructuur, het zoekalgoritme en de initiele parameters dienen nog een aantal keuzes gemaakt te worden:

1. Hoe lang moet een stijghoogte reeks zijn om een tijdreeksmodel te kunnen maken?
2. Wanneer moet er een ruismodel gebruikt worden?

0.1.1 Hoe lang moet een stijghoogte reeks zijn om een tijdreeksmodel te kunnen maken?

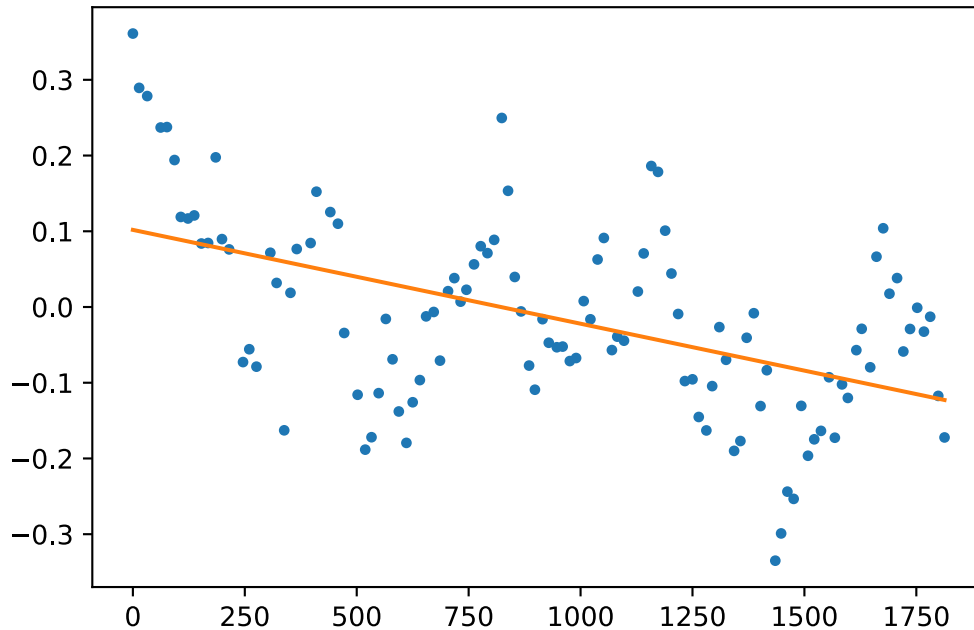
Zaken die besproken worden:

1. Verklarende reeksen moeten beginnen voor de gemeten stijghoogte.
2. Een stijghoogtereeks moet ten minste een aantal keer het geheugen van een systeem zijn.
3. De resultaten kunnen afhangen van het deel van de stijghoogtereeks dat gebruikt wordt voor kalibratie.



p value: 7.658940795075122e-09

[<matplotlib.lines.Line2D at 0x11c746100>]



0.2 Het gebruik van een ruismodel

0.2.1 Moet er een ruismodel gebruikt worden?

Een ruismodel kan gebruikt worden om te proberen er voor te zorgen dat de overgebleven verschillen tussen het model en de metingen aan een aantal statistische toetsen voldoen zodat er statistische uitspraken gedaan kunnen worden met het model (bijvoorbeeld: wat is de kans dat het effect van een onttrekking boven een bepaalde grenswaarde uitkomt).

De verschillen tussen de gemeten stijghoogte en de gemodelleerde stijghoogte (de residuën) is vrijwel altijd gecorreleerd in de tijd. Simpel gezegd: als het model vandaag iets hoger is dan de gemeten stijghoogte, dan is de kans groot dat het model volgende week ook iets hoger is dan de gemeten stijghoogte. Dat is op zich prima te verklaren. Het kan bijvoorbeeld zo zijn dat er een regen reeks gebruikt wordt van een weerstation dat 20 km verderop ligt. Als het nu bij het regenstation wel geregend heeft maar bij de peilbuis niet, dan zal het model waarschijnlijk een te hoge stijghoogte simuleren. En aangezien het een tijdje duurt voordat de stijghoogte weer gezakt is, zal het model waarschijnlijk voor een aantal dagen of weken te hoge waarden geven. Dat is op zich geen probleem, behalve als de gebruiker het model wil gebruiken om statistische uitspraken te doen. Als er statistische uitspraken gedaan moeten worden, dan moeten de residuën aan bepaalde statistische voorwaarde voldoen. Eén van de belangrijkste voorwaarden is dat de residuën niet gecorreleerd zijn. Een manier om te trachten dit te bewerkstelligen is om een ruismodel te gebruiken. Het doel van het ruismodel is om de gecorreleerde residuën te transformeren naar ruis die ongecorreleerd is. De eenvoudigste manier om te trachten dit te doen is door het gebruik van een AR1 ruismodel. In dit model is het residu $\varepsilon(t)$ op tijdstip t gelijk aan een factor ρ keer het residu op

tijdstip $t - \Delta t$ plus een random (ongecorreleerde) fout $n(t)$, of in formule vorm

$$\varepsilon(t) = \rho\varepsilon(t - \Delta t) + n(t)$$

Als de tijd Δt tussen twee residuën varieert, dan kan de factor ρ een functie zijn van Δt .

In onderstaande voorbeeld worden synthetische reeksen gegenereerd en geanalyseerd om te laten zien dat bij tijdreeksmodellen zonder ruismodel de parameters wel goed geschat worden maar de betrouwbaarheidsintervallen onderschat worden.

Fit report head		Fit Statistics	
=====			
nfev	17	EVP	90.61
nobs	644	R2	0.91
noise	True	RMSE	0.13
tmin	1985-11-14 00:00:00	AIC	3.17
tmax	2015-06-28 00:00:00	BIC	21.04
freq	D	Obj	2.14
warmup	3650 days 00:00:00	---	
solver	LeastSquares	Interpolated	No

Parameters (4 were optimized)

	optimal	stderr	initial	vary
recharge_A	797.719332	±5.57%	215.674528	True
recharge_a	161.413117	±5.92%	10.000000	True
constant_d	27.532182	±0.09%	27.900078	True
noise_alpha	66.013704	±12.86%	15.000000	True

Parameter correlations |rho| > 0.5

=====		
recharge_A	recharge_a	0.93
	constant_d	-0.80
recharge_a	constant_d	-0.75

```

1985-11-14    -0.147761
1985-11-28     0.018985
1985-12-14     0.049225
1985-12-28     0.053405
1986-01-13     0.079575

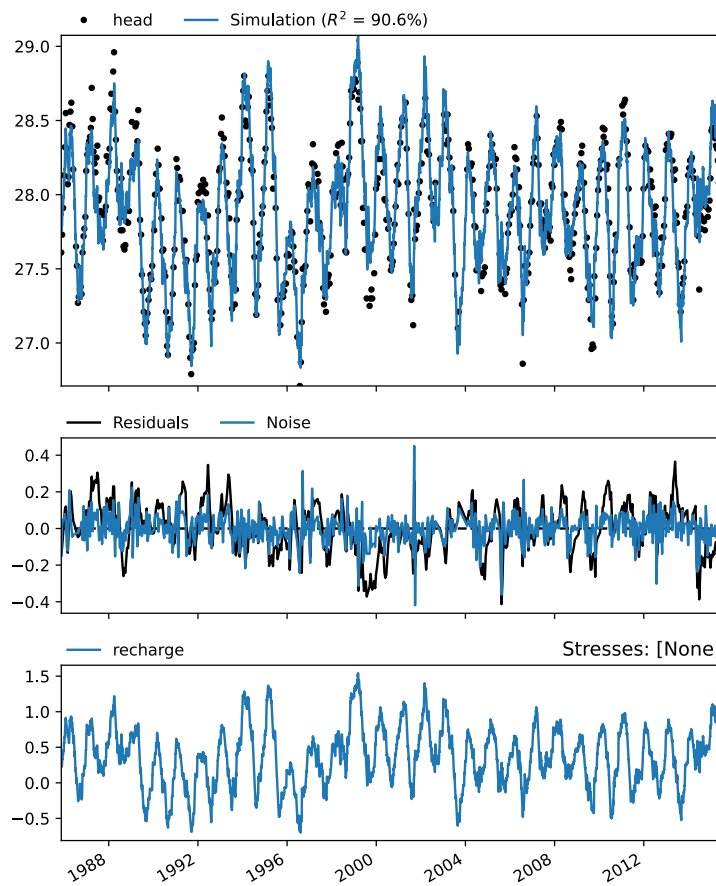
```

```

...
2015-04-28     0.081696
2015-05-14     0.012572
2015-05-28    -0.073930
2015-06-14     0.120032
2015-06-28    -0.092199

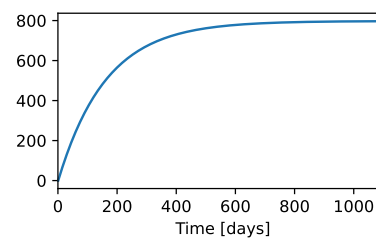
```

Name: Noise, Length: 644, dtype: float64

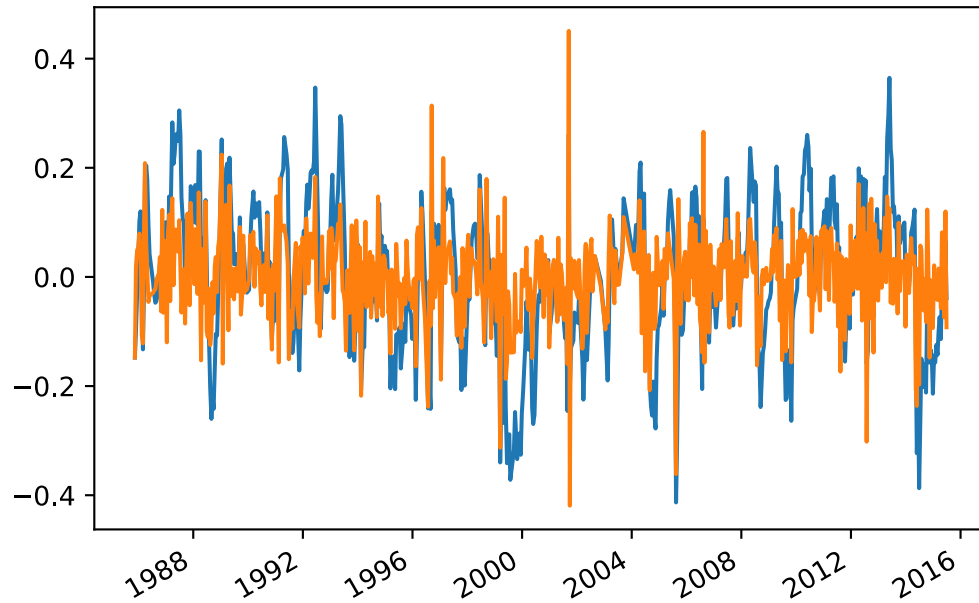


Model Parameters

name	optimal	stderr
recharge_A	797.72	5.57%
recharge_a	161.41	5.92%
constant_d	27.53	0.09%
noise_alpha	66.01	12.86%

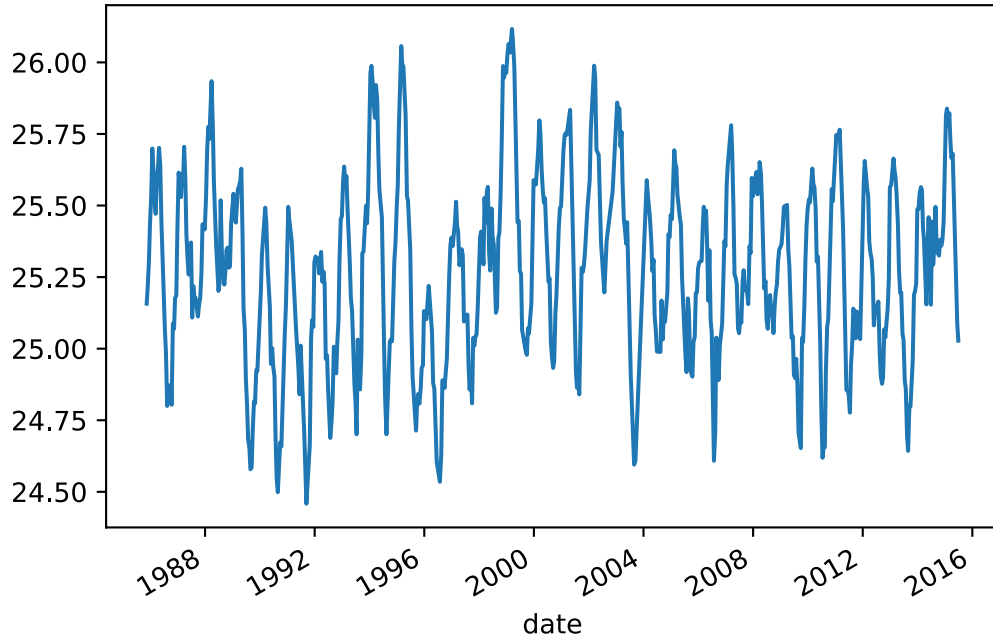


<AxesSubplot:>



Maak een model waarbij de dagelijkse grondwateraanvulling gelijk is aan de gemeten regen min de gemeten potentiële verdamping. De response functie van de grondwateraanvulling is een exponentiële functie met $A = 600$ en $a = 150$ d. Het basisniveau is $d = 25$ m. Een synthetische stijghoogte reeks wordt gesimuleerd op dezelfde tijdstippen waarop gemeten was. Deze synthetische reeks bevat geen fouten en een tijdreeksmodel met een exponentiële response functie kan de parameters vrijwel exact terugvinden.

```
<AxesSubplot:xlabel='date'>
```

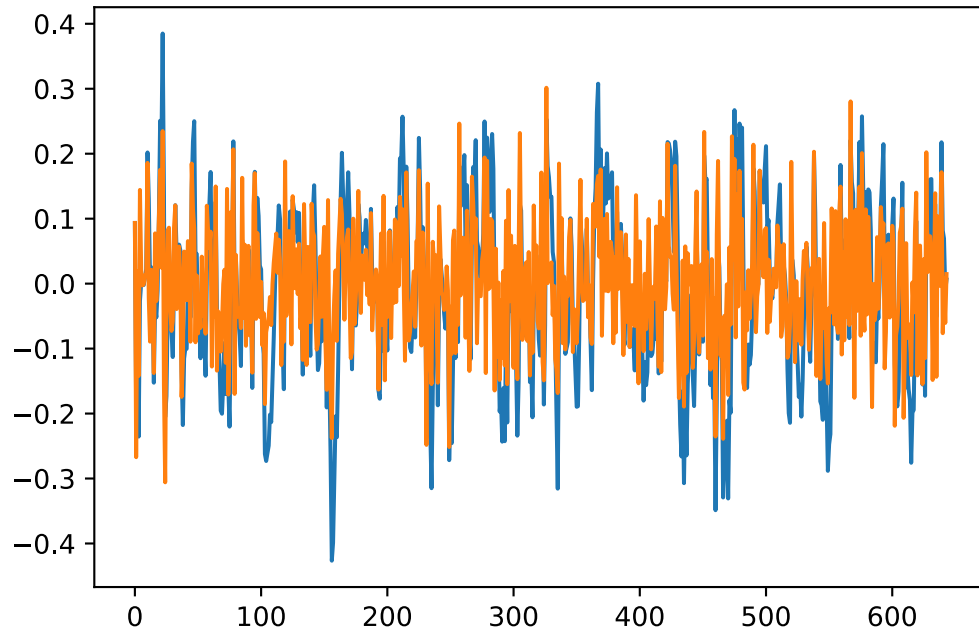


Vervolgens wordt er een residu (een fout) opgeteld bij de synthetische reeks. Het residu op tijdstip t , $\epsilon(t)$, is gecorrleerd met het residu bij de vorige meeting op tijdstip $t - \Delta t$ volgens de formule

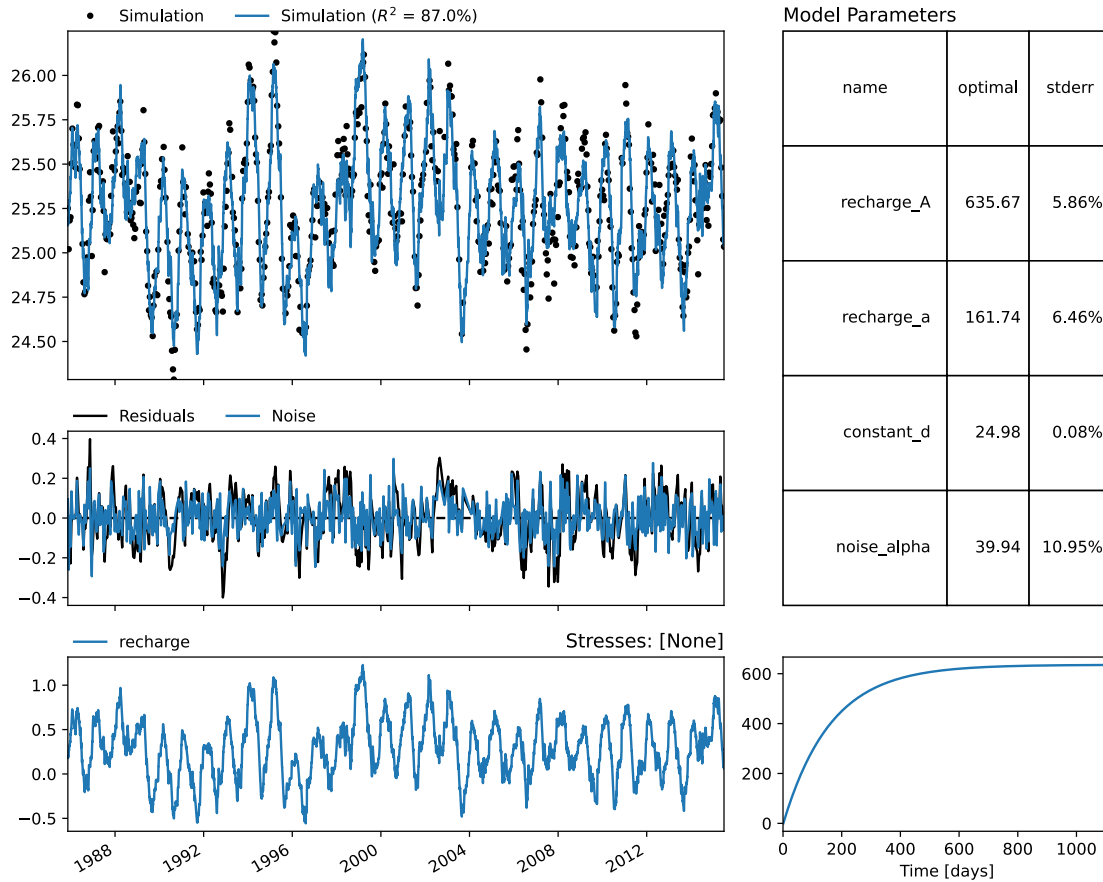
$$\epsilon(t) = e^{-\Delta t/\alpha} \epsilon(t - \Delta t) + n(t)$$

waarbij α een parameter is die aangeeft hoe veel correlatie er tussen residuen is en n de noise is, een willekeurige fout die getrokken wordt uit een normale verdeling met gemiddelde nul en standaard afwijking σ . De standaardafwijking van de noise is gekozen als $\sigma = 0.1$ m. Hoe groter de correlatie tussen de residuen is, hoe groter de de standaardafwijking van de residuen is.

```
std residuals 0.12842200175482732
std noise 0.09478825754856138
```



Vervolgens wordt er een tijdreeksmodel gemaakt voor de synthetische reeks inclusief de gecorreleerde residuen. Het tijdreeksmodel zal ongeveer de parameters terugvinden die opgegeven zijn, maar niet exact omdat er een fout opgeteld is bij de werkelijke stijghoogte. Hieronder is een voorbeeld gegeven.



Vervolgens gaan we een experiment doen. We creëren 100 keer een synthetische reeks en gaan daar 100 keer een tijdreeksmodel van maken. Daar de synthetische reeks elke keer andere (gecorrleerde) redisuen heeft komt er elke keer wat anders uit. Maar komt er toch ongeveer de opgegeven waarden voor de parameters uit. Het experiment is twee keer gedaan. De eerste keer is er 100 keer een tijdreeksmodel gemaakt zonder een ruismodel en vervolgens is er 100 keer een tijdreeksmodel gemaakt met een ruismodel.

0, 20, 40, 60, 80,

Results without a noise model

Mean value of parameter A: 604.6617309824148

Mean value of parameter a: 152.2590997834025

Mean value of parameter d: 24.997033127005306

0, 20, 40, 60, 80,

Results with a noise model

Mean value of parameter A: 601.4909834713767

Mean value of parameter a: 150.8283550961852

Mean value of parameter d: 24.998188602381564

Mean value of parameter alpha: 46.90115706632419

Het blijkt dat zowel het tijdreeksmodel zonder het ruismodel als het tijdreeksmodel met het ruismodel gemiddeld een goede schatting geeft van de parameters. Dus waarom moeten we dan een ruismodel gebruiken? De reden is dat de geschatte onzekerheid van de parameters te klein is als er geen ruismodel gebruikt wordt. De statistische manier waarop de onzekerheid geschat wordt gaat er namelijk van uit dat de overgebleven ruis niet gecorreleerd is. Als de ruis echter toch gecorreleerd is, dan is de geschatte onzekerheid te klein. Dit kan gemakkelijk nagegaan worden met het hierboven uitgevoerde experiment. Tijdens het experiment zijn de parameters 100 keer geschat inclusief een schatting van de standaard fout van de parameters. Onder enkele statistische aannames (inclusief de aannames van normaliteit en ongecorreleerde ruis), varieert het 95% betrouwbaarheidsinterval van min twee keer de standaard fout tot plus twee keer de standaard fout (eigenlijk is het 1.96 i.p.v. 2, maar laten we dan hier maar afronden). Er kan dus eenvoudig nagegaan worden hoe vaak de echte waarde van de parameters ligt binnen het 95% betrouwbaarheidsinterval van de geschatte parameters. En aangezien we het experiment 100 keer gedaan hebben, zou dit ongeveer 95 keer het geval moeten zijn. Zoals te zien is hieronder, geeft het tijdreeksmodel zonder ruismodel slechts zo'n 60-65 keer een betrouwbaarheidsinterval dat de echte waarde bevat. Het tijdreeksmodel met een ruismodel doet het veel beter: meer dan 95 keer liggen de echte parameters binnen het 95% betrouwbaarheidsinterval, behalve voor de parameter α van het ruismodel.

```
results without noise model
number of times true A within estimated confidence interval: 58
number of times true a within estimated confidence interval: 69
number of times true d within estimated confidence interval: 64

results with noise model
number of times true A within estimated confidence interval: 98
number of times true a within estimated confidence interval: 98
number of times true d within estimated confidence interval: 96
number of times true alpha within estimated confidence interval: 82
```

In bovenstaand experiment is uitgelegd wat het voordeel van het gebruik van een ruismodel is. Als er geen ruismodel gebruikt wordt, dan worden de parameters goed geschat, maar de betrouwbaarheidsintervallen van de parameters is te klein. Als het ruismodel wel meegenomen wordt, dan wordt het betrouwbaarheidsinterval veel beter geschat.

0.2.2 Wat als een eenvoudig ruismodel niet werkt?

1. Voorbeeld waarbij het ruismodel werkt als de tijdsperiode tussen metingen vergroot wordt
2. Voorbeeld ARMA(1,1) ruismodel?