Reparatie methodebreuken tijdreeksen gezondheid

09

Bob Lodder en Mohammed Kardal

Publicatiedatum CBS-website: 11 februari 2009

Centraal Bureau voor de Statistiek

Den Haag/Heerlen, 2009

Verklaring van tekens

.	= gegevens ontbreken
*	= voorlopig cijfer
x	= geheim
–	= nihil
–	= (indien voorkomend tussen twee getallen) tot en met
0 (0,0)	= het getal is kleiner dan de helft van de gekozen eenheid
niets (blank)	= een cijfer kan op logische gronden niet voorkomen
2005–2006	= 2005 tot en met 2006
2005/2006	= het gemiddelde over de jaren 2005 tot en met 2006
2005/'06	= oogstjaar, boekjaar, schooljaar enz., beginnend in 2005 en eindigend in 2006
2003/'04–2005/'06	= oogstjaar, boekjaar enz., 2003/'04 tot en met 2005/'06

In geval van afronding kan het voorkomen dat het weergegeven totaal niet overeenstemt met de som van de getallen.

Colofon

Uitgever
Centraal Bureau voor de Statistiek
Henri Faasdreef 312
2492 JP Den Haag

Prepress
Centraal Bureau voor de Statistiek - Facilitair bedrijf

Omslag
TelDesign, Rotterdam

Inlichtingen
Tel. (088) 570 70 70
Fax (070) 337 59 94
Via contactformulier: www.cbs.nl/infoservice

Bestellingen
E-mail: verkoop@cbs.nl
Fax (045) 570 62 68

Internet
www.cbs.nl

Inhoud

1. Inleiding

Het project Tijdreeksen Gezonde Levensverwachting is door het Expertisecentrum Lange Tijdreeksen uitgevoerd, in samenwerking met de afdeling SAH. Het project had onder andere als doel om een lange tijdreeks te construeren van de gezonde levensverwachting. Het belangrijkste probleem daarbij was dat de tijdreeks met gezondheidsdata gecorrigeerd diende te worden voor gevolgen van de wijzigingen in de enquête Permanent Onderzoek Leefsituatie (POLS) en de Gezondheidsenquête. In deze nota wordt verslag gedaan van de wijze waarop de reparatie van deze methodebreuken heeft plaatsgevonden.

In hoofdstuk 2 wordt ingegaan op de dataverzameling van prevalenties. Daarbij wordt het begrippenkader gegeven dat als uitgangspunt dient voor de analyses. In hoofdstuk 3 wordt de reparatiemethode in kort bestek, zoals is uitgewerkt door Lodder (Lodder, 2007), beschreven. In hoofdstuk 4 worden de de methodebreuken op een rijtje gezet ter voorbereiding van de feitelijke statistische analyses. Vervolgens worden in hoofdstuk 5, 6 en 7 de statistische analyses uitgevoerd die nodig zijn voor de reparatie van de reeksen. Daarbij worden drie verschillende schattingsmethoden doorgerekend. In hoofdstuk 8 wordt de analyse afgerond met de selectie van de keuze voor een schattingsmethode. In hoofdstuk 9 volgt een conclusie.

De gerepareerde reeksen worden gebruikt voor de berekening van de gezonde levensverwachting. In een ander rapport (Lodder, 2009) wordt dit verder beschreven. De berekeningswijze daarvan is complex en door een brede groep van onderzoekers geaccepteerd (Jagger, 2006).

2. Beschrijving reeksen prevalenties

2.1 Het begrip prevalentie

Met het begrip prevalentie wordt *het percentage aangeduid van een bepaalde doelgroep, dat ongezond is*. Daarbij wordt *ongezond* op drie verschillende manieren gedefinieerd:

1. Het ervaren van ongezondheid door de betrokken persoon zelf (EG);
2. Het hebben van één of meer langdurige aandoeningen (LA);
3. Het hebben van één of meer lichamelijke beperkingen (BP).

Deze begrippen worden in deze rapportage niet verder toegelicht. Zie voor een uitgebreide behandeling Stam, 2009. In het vervolg worden respectievelijk de termen *Ervaren gezondheid, Langdurige aandoeningen en Beperkingen* gebruikt.

Met betrekking tot de *doelgroep*, wordt onderscheid gemaakt naar geslacht en leeftijdsgroep. We onderscheiden de volgende leeftijdsgroepen:

– 0 jaar
– 1–4 jarigen
– 5–9 jarigen
– 10–14 jarigen

Enzovoort…….

– 75–79 jarigen
– 80 jaar of ouder

Een prevalentie wordt zodoende gerelateerd aan de volgende kenmerken:
Prev() = Prev(x, s, t, G)
x = leeftijdsgroep x (x= ”0”, ”1–4”, ”5–9”, …, ”75-80”, “80+”)
s= sekse (man, vrouw)
t = het jaar: t = 1981,, …., 2007.
G = de definitie van ongezondheid waar de prevalentie betrekking op heeft;
G = EG, LA of BP

We kunnen zodoende de volgende uitspraak doen:
Prev(man, “5–9”, 2001, LA)= 15%
De betekenis hiervan is:
Het percentage mannen in de leeftijdsgroep 5–9 jaar dat in 2001 een langdurige aandoening (LA) heeft, is 15%.

Er zijn zodoende ongeveer 2 (geslacht) x l8 (leeftijdsgroep) x 27 (jaren) x 3 (definitie gezondheid) ≈ drie duizend verschillende prevalenties. In totaal zijn er 2 (geslacht) x l8 (leeftijdsgroep) x 3 (definitie gezondheid) = 108 verschillende *reeksen* prevalenties.

2.2 De databronnen voor prevalenties

De prevalenties worden berekend op basis van het steekproefonderzoek dat bij het CBS is gedaan. Er wordt gebruik gemaakt van:

– Permanent Onderzoek Leefsituatie (POLS);
– Gezondheidenquête (GE).

De data van het Leefsituatieonderzoek (LSO) zijn in eerste instante wel gebruikt, maar er is geconstateerd dat deze beperkte hoeveelheid cijfers te veel variatie vertoonde om een betrouwbaar beeld te krijgen van de prevalenties.

Een compleet overzicht van de dataset wordt gegeven door Stam (Stam, 2009). Er zijn met name bij de Gezondheidenquête wijzigingen geweest in de vragen en de mogelijke antwoorden die de respondenten konden geven. De wijzigingen hebben er toe geleid dat er methodebreuken in de reeksen zijn (zie hoofdstuk 4). Er zijn verschillende reeksen voor mannen en vrouwen, maar de methodebreuken voor beide geslachten zijn identiek.

2.3 De reeksen

Ervaren gezondheid
De prevalentie reeks beslaat de periode 1981 t/m 2005.

Langdurige aandoeningen
De reeks bestrijkt de periode 1981 t/m 2007.

Beperkingen
Hier gaat het om een reeks lopend van 1983 t/m 2007. Voor kinderen zijn geen prevalenties beschikbaar.

3. De reparatie van methodebreuken

3.1 Het begrippenkader

In deze rapportage wordt het begrip *methodebreuk* gehanteerd. Daarmee wordt een wijziging in de onderzoeksmethode (concreet: de vragenlijst) bedoeld die een effect kan hebben op het verloop van de reeks. Aangezien het effect niet noodzakelijkerwijs hoeft te bestaan, is er niet altijd sprake van een breuk in de reeks (*reeksbreuk*) of een breuk in de trend (*trendbreuk*, Lodder, 2007). Wel is er *in ieder geval* sprake van een methodebreuk als de onderzoeksmethode is gewijzigd. De omvang van het effect van de methodebreuk op het verloop van de reeks wordt het *methode-effect* genoemd.

Opgemerkt moet worden dat de tijdreeks die gebruikt wordt in de reparatieslag in beginsel informatie bevat van verschillende (deel)onderzoeken. We spreken derhalve niet meer over één reeks, maar over een reeks cijfers die ontstaat door het *combineren* van reeksen. Daarbij wordt een verzameling reeksen aangemaakt die informatie verschaffen over de bron van de cijfers uit de reeks Combi. Dus we hebben:
– Een tijdreeks Combi die ontstaat door het combineren van de beschikbare reeksen.
– Een verzameling (technisch: matrix) tijdreeksen Dummy, die aangeeft uit welk (deel)-onderzoek het betreffende cijfer komt.

Het doel van de reparatie is het uitsplitsen van de reeks Combi in twee componenten:
Combi = Reële component + Methode-effect component.

Waarbij de Reële component weer onderverdeeld kan worden in:
Reële component = Trendmatige component + Toevallige fluctuaties.

De Methode-effect component is een functie van (de matrix) Dummy.

Het repareren van de reeks houdt in het construeren van een gerepareerde reeks, die we Repair noemen, die gelijk is aan de Reële component. Probleem is dat deze grootheid niet bekend is. Deze wordt daarom met behulp van een nader te kiezen methode geschat. Nauwkeuriger gezegd: het methode-effect wordt geschat en dat wordt verminderd op de waarden van de reeks:
Repair = Combi – E(Methode-effect component).

waarbij de notatie E(.) aangeeft dat het om de geschatte waarde van de genoemde grootheid gaat. Merk op dat voor de gerepareerde reeks *Combi* wordt gebruikt en niet E(Combi). De reden daarvan is dat we de beschikking hebben over de echte datareeks Combi, en er dus geen reden is om de (per definitie) slechtere reeks E(Combi) te gebruiken.

De grootheid E(Combi) is echter wel van belang omdat deze, in combinatie met Combi informatie verschaft over de kwaliteit van de schatting en daarmee ook over de kwaliteit van de reparatiemethode. Omdat E(Combi) de meest gebruikte schatting is en om te benadrukken dat het een schatting is, noemen we die kortweg Estimate:
Estimate = E(Combi).

Door de schatting komt er ook een schatting van de trendmatige component beschikbaar. Deze is ook interessant, omdat daarmee een cijferreeks wordt opgeleverd die niet alleen voor methodebreuken, maar ook de trendmatige ontwikkeling weergeeft. In de praktijk zijn zowel de gerepareerde reeks als de trend van belang.

Er geldt:
E(Trendmatige component) = Estimate – E(Methode-effect component);
We korten *E(Trendmatige component)* af tot *Trend*.

In de rest van dit hoofdstuk zal stil worden gestaan bij de constructie van de reeksen Combi, Estimate, Trend en Repair. Daarnaast volgt er een kort intermezzo over de Box Cox transformatie, die nodig is om de variantie van de reeks Combi te stabiliseren. In hoofdstuk 4 wordt besproken hoe de matrix Dummy er uit ziet voor de verschillende definities van gezondheid.

Er zal een wiskundige notatie worden gebruikt om bovenstaande concepten verder uit te werken. De gedachte achter de formules blijft echter hetzelfde als hierboven is geschetst.

3.2 De reeks Combi

Op basis van POLS en de Gezondheidenquête wordt er een reeks Combi gevormd. Deze reeks is een samenvoeging van de twee enquêtes tot een lange tijdreeks. Daarbij worden reeksen van prevalenties gevormd voor zowel mannen als vrouwen, diverse leeftijdsgroepen en 3 verschillende gezondheidsdefinities. We recapituleren nog even de notatie:

Prev(x, s, G, t) = De prevalentie van leeftijdsgroep x, geslacht s, gezondheidsdefinitie G, in jaar t.

Waarbij:
x = 0, 1–4, 5–9, 10–14, …, 75–79, 80+
s = man, vrouw
G = EG, LA of BP
t = 1981, …, 2007
Met behulp van deze notatie kunnen de 108 reeksen Combi gedefinieerd worden.

3.3 Intermezzo: de Box Cox transformatie

Aangezien de varianties niet constant zijn over de tijd, is er gewerkt met een Box Cox transformatie (Abraham, 1983). Dit is de meest gebruikte methode om varianties te stabiliseren. In algemene zin luidt deze:
$Z = (Y\lambda - 1) / \lambda$
In de praktijk bleek een transformatie met een wortelfunctie het best te voldoen ($\lambda = ½$).

Aangezien de getransformeerde variabele gebruikt zal worden als endogene in een lineair model, is het niet relevant of de waarde van $Y^{½}$ nog een lineaire transformatie ondergaat ($Z = 2Y^{½} - 2$), of direct als endogene wordt gebruikt. Voor de eenvoud gebruiken we daarom alleen de worteltransformatie voor de prevalenties:

$$\text{Combi} = Z = Z(x, s, G, t) = \text{Prev}(x, s, G, t)^{½} = \sqrt{\text{Prev}(x, s, G, t)}$$

Waarbij:
Prev(x ,s, G, t) = de prevalentie volgens één van de drie definities voor een bepaalde leeftijdsgroep *x* en geslacht *s* in jaar *t*.

Z(x, s, G, t) = de waarde van de getransformeerde prevalentie. Deze variabele wordt ook de endogene variabele genoemd.

3.4 De reeks Estimate

Met behulp van schattingsmodellen worden de methodebreuken gekwantificeerd en een schatting gemaakt voor de reële ontwikkeling. Het algemene model luidt:

$$Z = U + D\delta$$

Waarbij:
$Z = (Z_1 \ldots, Z_T)'$ = een kolomvector van de (Box-Cox getransformeerde) prevalenties. Daarbij is $Z_t = Z_t\ (x, s, G)$ de getransformeerde prevalentie voor een specifieke leeftijdsgroep (x), een geslacht (s), een gezondheidsdefinitie (G) en een jaar t (t= 1,..,T).

$U = (U_1 \ldots, U_T)'$ een kolomvector die de reële component met $U_t = U_t(x, s, G)$ de reële component voor een specifieke leeftijdsgroep (x), een geslacht (s) een gezondheidsdefinitie (G) en een jaar t (t=1,..,T).

$D = D_{t,k}$ (x,s,G) een matrix van dummy's die de informatie over de methodebreuken voor een specifieke leeftijdsgroep (x), geslacht (s) en een gezondheidsdefinitie (G) bevat. Daarbij is iedere kolom verbonden aan een bepaald jaar t (t=1,..,T) en iedere rij verbonden aan een bepaalde dummy (k=1,..,K). Daarbij kunnen ook vertraagde dummy's worden meegenomen (zie hoofdstuk 6).

D kan ook geschreven worden als K kolomvectoren: $D = (D_1,..,D_K)$ waarbij het element D_k (k=1,…,K) een kolomvector is met Tx1 elementen.

$\delta = (\delta_1(x,s,G), \ldots, \delta_K(x,s,G))'$ een (Kx1)-kolomvector van de parameters voor het methode-effect. Deze parameters geven de omvang van het effect weer. Deze kunnen verschillen per leeftijdsgroep, geslacht, jaar en methodebreuk.

In totaal zijn er 2 (geslacht) x l8 (leeftijdsgroep) x 3 (definitie gezondheid) = 108 verschillende *reeksen* Z_t prevalenties.

De keuze van de functionele vorm van U is afhankelijk van het soort model dat wordt gekozen. Deze schattingsmodellen zijn geclassificeerd in drie hoofdgroepen:
1. De Methode Reg Tijd
2. De Methode Reg ARIMA
3. De Methode State space

Deze methodieken worden achtereenvolgens in hoofdstuk 5, 6 en 7 besproken. Uiteindelijk is gekozen voor de Methode state space. In deze rapportage wordt daarom meer aandacht gegeven aan deze methode dan aan de andere methoden.
We krijgen dan voor ieder schattingsmodel een geschatte waarde van de endogene Z. We noteren deze waarde als $\hat{Z}$. Ook noteren we de andere geschatte variabelen door er een "dakje" boven te schrijven. Er geldt dus:

$$\text{Estimate} = \hat{Z} = \hat{U} + D\hat{\delta}$$

3.5 De reeks Trend

De Trend wordt verkregen door de component U_t te berekenen. Daarbij wordt dus het methode-effect uit de gewone reeks Z_t gehaald. Met behulp van de geschatte parameters voor de dummy's wordt de Trend als volgt berekend:

$$\text{Trend} = \hat{U} = \hat{Z} - D\hat{\delta}$$

De reeks Trend is van belang aangezien alle gewone reeksen niet alleen gekenmerkt worden door methodebreuken, maar ook door sterke tijdelijke schommelingen in de prevalenties. Deze schommelingen kunnen veroorzaakt worden door het eigen karakter van de reeks, maar ook door steekproeffouten. In het laatste geval is het wellicht wenselijk om uiteindelijk de trend te verkiezen boven de gerepareerde reeks. Bij de bespreking van de state space modellen wordt hier op teruggekomen (paragraaf 7.4) en bij de keuze voor een schattingsmethodiek (paragraaf 8.2).

3.6 De reeks Repair

De reeks Repair ontstaat door de getransformeerde prevalenties (de waarden Z) te verminderen met de geschatte waarden van de methodebreuken:
Repair = $Z^R = Z - D\hat{\delta}$

De dummy's zijn zo geconstrueerd dat de meest recente waarden van de reeksen (doorgaans het jaar 2005), alle dummy's gelijk aan nul zijn. Het meest recente jaar wordt daarom beschouwd als het basisjaar. Alle methodenbreuken zijn in feite gedefinieerd ten opzichte van het basisjaar. Gegeven deze constructie, dan is de gerepareerde reeks aan het eind altijd gelijk aan de waargenomen reeks:

$Z_t^R = Z_t - 0 = Z_t$, voor t = een basisjaar, omdat $D\hat{\delta} = 0$ in het basisjaar

In praktijk gaat het vaak om het grootste deel van de reeks. Dit betekent dat een groot gedeelte van de gerepareerde reeks gewoon gelijk is aan de reeks zelf (zie hoofdstuk 4).

3.7 De schattingsmethoden

Voor de schatting van de methodebreuken zijn een drietal methoden geformuleerd (Lodder, 2007):
– Methode tijd reg;
– Methode ARIMA;
– Methode state space.

De Methode tijd reg is in dit onderzoek wel gebruikt, maar er is geconstateerd dat er vaak sprake is van eerste orde autocorrelatie (Abraham, 1983), waardoor er niet meer is voldaan aan de voorwaarden van het model. Deze methode is daarom niet op grote schaal toegepast. Er wordt wel een voorbeeld gegeven van de methodiek.

Alle reeksen zijn geschat met zowel de Methode ARIMA als de Methode state space. Ze zullen in volgende hoofdstukken worden beschreven.

Het is mogelijk om op identieke wijze de BIC (Bayesian Information Criterion) te berekenen. Daarmee is er een maatstaf om de fit van beide modellen met elkaar te vergelijken. Daarnaast zijn er ook inhoudelijke argumenten om voor een bepaald model te kiezen.

Een belangrijk punt is of dummy-variabelen in het model gehandhaafd moeten worden als de parameter ä op basis van een t-toets niet significant van nul verschilt op het 95%-betrouwbaarheidsniveau. Wonnacot adviseert in dergelijke gevallen (Wonnacot, 1970) om de variabelen wel op te nemen omdat er redenen zijn om aan te nemen dat de variabele wel degelijk invloed heeft, maar dat door een te kleine steekproef dit niet aangetoond kan worden met een t-toets. Daarom zijn alle dummy's gehandhaafd in het model, ongeacht hun significantie.

4. De methodebreuken

4.1 Ervaren Gezondheid

Op basis van het onderzoek van Stam (2009) is geconstateerd dat er bij Ervaren gezondheid slechts één dummy nodig is voor de methodenbreuk in 1983. Dit wordt weergegeven in de volgende tabel waarbij de matrix Dummy slechts één kolom bevat, de vector D81/D82.

Tabel 1
De matrix Dummy bij de definitie Ervaren gezondheid

	D81/82
1981	1
1982	1
1983	0
1984	0
1985	0
1986	0
1987	0
1988	0
1989	0
1990	0
1991	0
1992	0
1993	0
1994	0
1995	0
1996	0
1997	0
1998	0
1999	0
2000	0
2001	0
2002	0
2003	0
2004	0
2005	0

Bron: CBS.

4.2 Beperkingen

Bij Beperkingen zijn er drie dummy's nodig. De matrix dummy bevat zodoende drie vectoren: D83, D86 en D84/85.

Tabel 2
De matrix Dummy bij de definitie Beperkingen

	D83	D86	D84/85
1983	1	0	0
1984	0	0	1
1985	0	0	1
1986	0	1	0
1987	0	1	0
1988	0	1	0
1989	0	0	0
1990	0	0	0
1991	0	0	0
1992	0	0	0
1993	0	0	0
1994	0	0	0
1995	0	0	0
1996	0	0	0
1997	0	0	0
1998	0	0	0
1999	0	0	0
2000	0	0	0
2001	0	0	0
2002	0	0	0
2003	0	0	0
2004	0	0	0
2005	0	0	0

Bron: CBS.

4.3 Langdurige Aandoeningen

Bij Langdurige Aandoeningen zijn er vier dummy's (D1, D2, D3 en D4) nodig om alle methodebreuken te modelleren. Er is tevens onderzocht of het wenselijk is om minder dummy's op te nemen om het aantal vrijheidsgraden in de schatting hoog te houden. Dit bleek niet het geval te zijn.

Tabel 3
De matrix Dummy bij de definitie Langdurige Aandoeningen

	D1	D2	D3	D4
1981	1	0	0	0
1982	1	0	0	0
1983	1	0	0	0
1984	1	0	0	0
1985	1	0	0	0
1986	0	1	0	0
1987	0	1	0	0
1988	0	1	0	0
1989	0	0	1	0
1990	0	0	1	0
1991	0	0	1	0
1992	0	0	1	0
1993	0	0	0	1
1994	0	0	0	1
1995	0	0	0	1
1996	0	0	0	1
1997	0	0	0	1
1998	0	0	0	1
1999	0	0	0	1
2000	0	0	0	0
2001	0	0	0	0
2002	0	0	0	0
2003	0	0	0	0
2004	0	0	0	0
2005	0	0	0	0

Bron: CBS.

5. De Methode tijd reg

Het algemene schattingsmodel luidt (zie paragraaf 3.4):

$$Z = U + D\delta$$

In het kader van een tijdreeksanalyse, kan men de tijd (t) als variabele opnemen voor het beschrijven van trends en golfbewegingen (Abraham, 1983):

$$U_t = \sum_j \beta_j t^j + \sum_j \left(\gamma_j \sin(f(j)t) + \lambda_j \cos(f(j)t)\right) + \varepsilon(t)$$

Met $f(i) = 2\pi i / s$ (s een frequentie parameter) en ε(t) is de storingsterm die normaal verdeel is (N(0, σ^2)).
Daarnaast wordt het methode-effect Dδ in jaar *t* als een lineaire functie in D_t gemodelleerd, dus zonder vertragingsstructuur:

$$(\mathrm{D}\delta)_t = \sum_{i=1}^{K} \delta_i D_{i,t}$$

Het algemene model kan daarom geschreven worden als:

$$Z_t = \sum_j b_j t^j + \sum_j \left(\gamma_j \sin(f(j)t) + \lambda_j \cos(f(j)t)\right) + \sum_{i=1}^{k} \delta_i D_{i,t} + \varepsilon(t)$$

De *eerste* term van het rechterlid van de bovenstaande formule betreft een trend, waarbij de variabele tijd in verschillende machten wordt opgenomen. De *tweede* term betreft periodieke golfbewegingen. De *derde* term betreft de dummy's voor de methodebreuken. De schatting van het bovenstaande model kan eenvoudig worden uitgevoerd met de kleinste kwadraten methode (OLS). Ter illustratie worden de schattingen weergegeven voor Ervaren gezondheid van mannen in de leeftijdsgroep 25–29 jaar. De geschatte uitkomsten staan in tabel 4. De *BIC* is gelijk aan –0,45. De fit is goed.

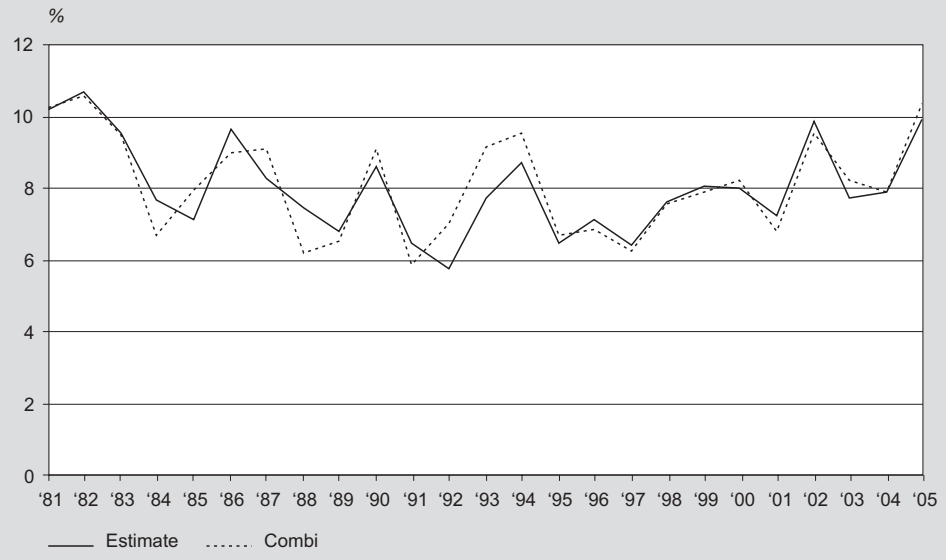
Probleem echter bij deze benadering is dat er sprake is van autocorrelatie. In dat geval is niet voldaan aan de aannames van het model. Het bleek bij de schatting met de Methode ARIMA dat er sterke aanwijzingen zijn dat er bij een groot aantal reeksen autocorrelatie is. Eén van de weinige reeksen waarbij geen autocorrelatie was, is de reeks van Ervaren gezondheid, mannen, 25–29 jaar. De resultaten worden hieronder weergegeven. Er is besloten vanwege autocorrelatie om deze methode niet verder te exploreren.

Tabel 4
Schattingen met Reg Tijd, Mannen, 25–29 jaarSchattingen met Reg Tijd, Mannen, 25–29 jaar

	Estimate	Combi
	%	
1981	10,20	10,27
1982	10,66	10,58
1983	9,54	9,50
1984	7,67	6,69
1985	7,11	7,95
1986	9,67	8,97
1987	8,27	9,12
1988	7,46	6,20
1989	6,77	6,52
1990	8,60	9,07
1991	6,44	5,86
1992	5,78	7,02
1993	7,71	9,13
1994	8,69	9,53
1995	6,47	6,68
1996	7,15	6,84
1997	6,40	6,23
1998	7,60	7,55
1999	8,07	7,90
2000	7,98	8,23
2001	7,23	6,77
2002	9,85	9,51
2003	7,73	8,23
2004	7,91	7,89
2005	9,92	10,35

Bron: CBS.

1. De geschatte reeks van EG, mannen, 25–29 jaar



Bron: CBS.

6. De Methode ARIMA

6.1 Een inleiding op het ARIMA-model

Methodebreuken in een datareeks kunnen onder andere door een ARIMA-model beschreven worden. De analyse van methodebreuken kan met de *TSMODEL* procedure van SPSS worden uitgevoerd. Het opsporen van de methodebreuken, via deze modellering, wordt gedaan door een regressie variabele toe te voegen aan het ARIMA-model. Als het zo geconstrueerde model een goede beschrijving voor data weergeeft, dan kunnen de uitkomsten van de significantieniveaus van de regressoren gebruikt worden. Daarmee kan worden beoordeeld of er wel of niet sprake is van een discontinuïteit in de reeks.

In formule vorm gaat het om het volgende model:

$$Z = U + D\delta$$

Waarbij

$Z = (Z_1 \ldots, Z_T)'$ en $Z_t = \sqrt{}\,\text{Prev}_t$, de Box Cox transformatie van de prevalenties.

$U = (U_t)_{1 \le t \le n}$ volgt een ARIMA-structuur met parameters (p,q) d.w.z.:

$$U_t = \theta_0 + \phi_1 U_{t-1} + \phi_2 U_{t-2} + \ldots + \phi_p U_{t-p} - \phi_1 \varepsilon_{t-1} - \phi_2 \varepsilon_{t-2} - \ldots - \phi_q \varepsilon_{t-q} + \varepsilon_t$$

$(\phi_i)_{1 \le i \le p}$: de autoregressieve (AR) parameters.
$(\phi_i)_{1 \le i \le q}$: de moving average (MA) parameters.
$(\varepsilon_t)_{1 \le t \le n}$ is een storingsterm.

Het identificeren van de ARIMA-parameters kan met behulp van het diagram van figuur 2 (Meeker, 2001).

De lengte van de reeks prevalenties is zo klein (minder dan 30) dat de voorkeur naar lage waarden van (p,q) gaat indien de data daarmee goed worden beschreven. Hoe groter de waarden van deze parameters (p,q) des te minder betrouwbaar de schattingen zijn.

Voor de drie definities van gezondheid zijn, nadat indien nodig de reeks stationair is gemaakt, de meest optimale modellen: ARIMA(1,0,0) en ARIMA(0,0,1). In minder mate komt het model ARIMA(1,0,1) voor.

De geschatte prevalentie reeks op basis van het meest voorkomende model ARIMA(1,0,0) heeft de volgende gedaante:

$$\hat{Z}_t = \hat{\phi}_1 Z_{t-1} + \hat{\lambda}(D_t - \phi_1 \hat{D}_{t-1})$$

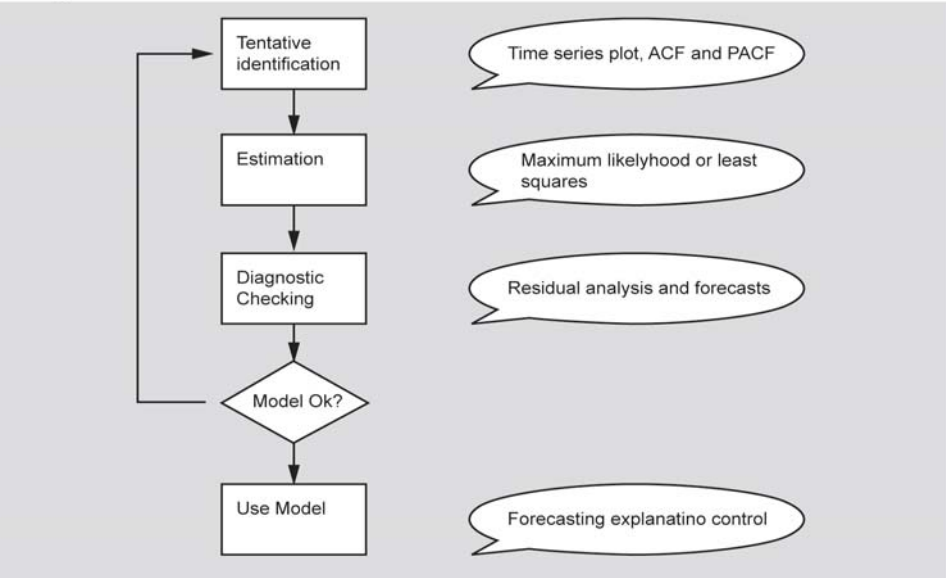
Daarbij is voor de eenvoud in de notatie uitgegaan van slechts één dummy voor een methodebreuk. Bij een ARIMA(1,0.0)-model met een dummy (of regressor) verschijnt in het rechterlid ook een vertraagde dummy (Abraham, 1983). Het methode-effect in jaar *t* heeft de gedaante:

$$(D\delta)_t = \hat{\lambda}(D_t - \hat{\phi}_1 D_{t-1})$$

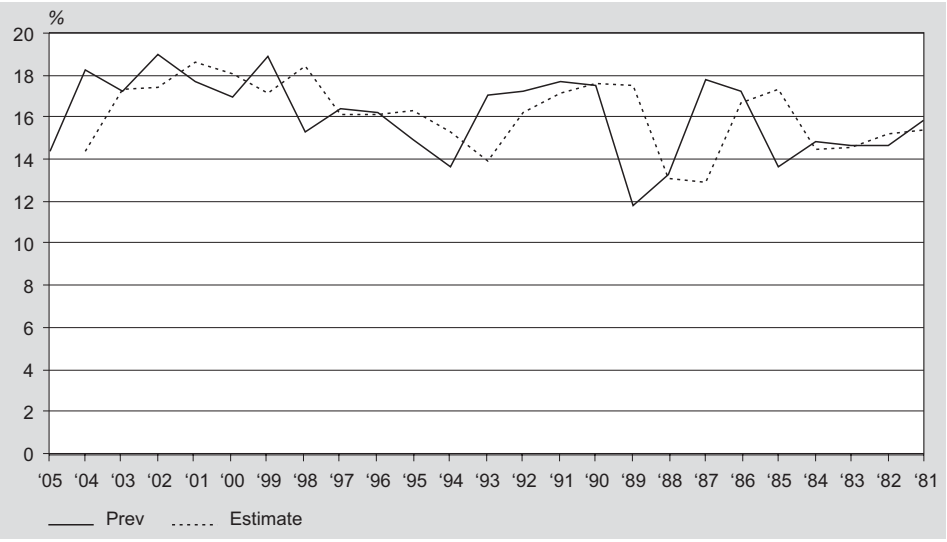
In matrix-notatie: $D = (D_{\cdot 1}, D_{\cdot 2})$, $D_{\cdot 1}$ en $D_{\cdot 2}$ zijn kolomvectoren met dimensie Tx1. $D_{\cdot 2} = \text{Lag}(D_{\cdot 1})$, waarbij Lag() de Lag-operator is. Deze neemt de vertraagde waarde op: $D_{t,2} = \text{Lag}(D_{t,1}) = D_{t-1,1}$; $\delta = (\lambda, -\lambda\varphi_1)'$, een 2x1 kolomvector. $D\delta = (D_{\cdot 1}, D_{\cdot 2})(\lambda, -\lambda\varphi_1)'$.

Bij de reeksen waarop een ARIMA(1,0,0) is toegepast lijkt de nieuwe (geschatte reeks) op een translatie in de tijd van de oorspronkelijke reeks (zie figuur 3).

2. Diagram identificatie van een ARIMA model



3. De met een ARIMA(1,0,0) geschatte reeks (reeks Estmate) van prevalentis) voor vrouwen leeftijdsgroep 35 bij Ervaren gezondheid (EG)



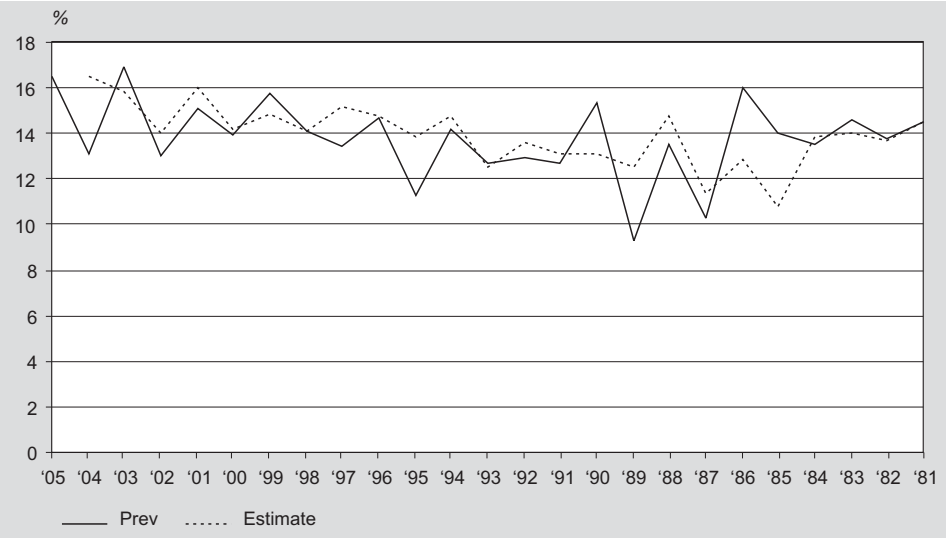
Bron: CBS.

Deze verschuiving van de oorspronkelijke data verdwijnt wanneer de reeksen stationair gemaakt worden door een éénmalige differentiatie van de datareeks (d.w.z. d =1 bij een ARIMA(p,d,q)) (zie figuur 4) of wanneer een constante in het arima model opgenomen kan worden.

6.2 Het toepassen van het ARIMA-model

Ter illustratie van het toepassen van het ARIMA-model nemen we de reeks van Langdurige aandoeningen bij vrouwen van leeftijdsgroep 25–29 jaar (kolom Prev in onderstaande tabel). Met behulp van het schema uit Figuur 2 kan een passend model bepaald worden. Het datavoorbeeld kan met een ARIMA(1,0,0) beschreven worden.

4. De met een ARIMA(1,1,0) geschatte reeks (reeks Estmate) van prevalentie data (reeks Prev) voor vrouwen leeftijdsgroep 30 bij Ervaren gezondheid (EG)



Bron: CBS.

Tabel 5
Schatting (kolom Estimate) van de reeks Langdurige aandoeningen leeftijdsgroep 25–29 jaar (Prev)

	Prev	Dummy	Schatting Dummy	Estimate	Repair
2005	33,6			–	33,63
2004	33,6			33,0	33,55
2003	40,2			32,9	40,15
2002	48,3			39,4	48,27
2001	39,9			47,3	39,88
2000	34,4	D4	–5,14	34,0	39,56
1999	29,7	D4	–5,14	33,6	29,80
1998	24,4	D4	–5,14	29,0	24,47
1997	28,4	D4	–5,14	23,8	28,54
1996	24,5	D4	–5,14	27,8	24,60
1995	28,5	D4	–5,14	23,9	28,58
1994	24,2	D3	–2,63	27,8	24,25
1993	26,4	D3	–2,63	26,1	24,02
1992	21,9	D3	–2,63	25,9	21,96
1991	25,3	D3	–2,63	21,4	25,35
1990	24,1	D3	–2,63	24,8	24,15
1989	23,0	D3	–2,63	22,6	23,03
1988	29,3	D2	–3,56	22,5	30,27
1987	21,5	D2	–3,56	28,6	21,52
1986	26,7	D2	–3,56	26,3	26,72
1985	13,4	D1	1,79	26,2	8,16
1984	12,5	D1	1,79	13,2	12,42
1983	13,7	D1	1,79	12,3	13,65
1982	7,6	D1	1,79	13,5	7,58
1981	7,9	D1	1,79	7,5	7,86

Bron: CBS.

De prevalentie reeks Prev uit bovenstaande tabel is gemodelleerd met een ARIMA(1,0,0) en een regressie component die de aanwezigheid van dummy's aangeeft. Bij dit voorbeeld gaat het om het model met de volgende gedaante:

$$\hat{Z}_t = \hat{\phi}_1 Z_{t-1} + \sum_{i=1}^{k} \hat{\lambda}_i (D_{i,t} - \hat{\phi} D_{i,t-1})$$

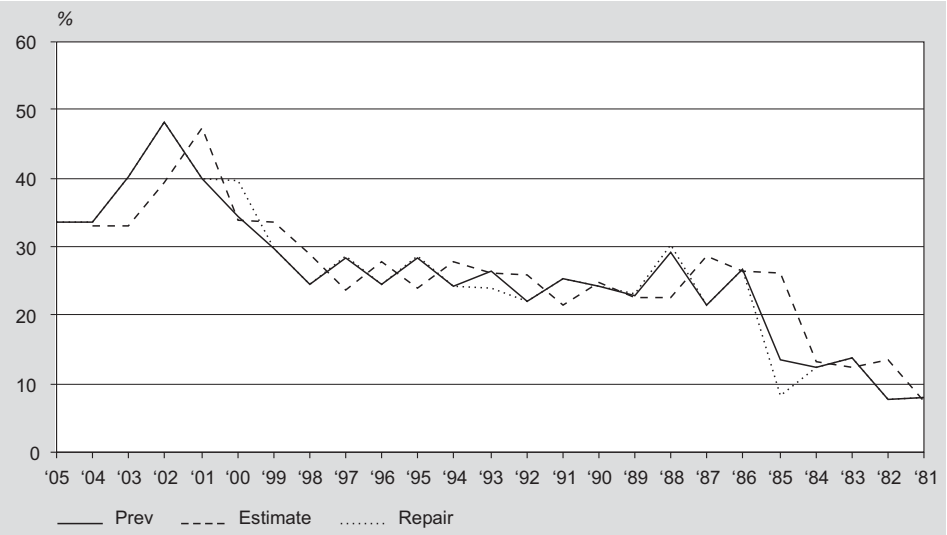
De uitdrukking na het plus-teken bevat het methode-effect:

$$D\hat{\delta}_t = \sum_{i=1}^{k} \hat{\lambda}_i (D_{i,t} - \hat{\phi} D_{i,t-1})$$

Door uitschakeling van dit effect kan de gerepareerde reeks Z_t^R geconstrueerd worden:

$$Z_t^R = Z_t - D\hat{\delta}_t$$

5. De met ARIMA(1,0,0) gerepareerde reeks (reeks Repair) van prevalentie data (reeks Prev) voor vrouwen leeftijdsgroep 25/29 bij Langdurige aandoeningen



Bron: CBS.

7. De Methode State space

7.1 Inleiding

De methode state space betreft het combineren van dummy's voor trendbreuken met een state space model.

Aangezien er verschillende soorten state space modellen zijn, moest er een keuze worden gemaakt. Er is gekozen voor het *local lineair trend model* (Durbin, 2001). Dit model is zowel flexibel als eenvoudig. Het wordt aangevuld met dummy's voor het methode-effect en een Box Cox transformatie voor de endogene variabele:

$$Z = U + D\delta$$

$$(D\delta)_t = \sum_{i=1}^{K} \delta_i D_{i,t}$$

$$U_t = \mu_t + \varepsilon_t \qquad ,met\ \varepsilon_t \sim N(0, \sigma^2\varepsilon)$$

$$\mu_t = \mu_{t-1} + V_{t-1} + \xi_t \qquad ,met\ \xi_t \sim N(0, \sigma^2{}_\xi)$$

$$V_t = V_{t-1} + \xi_t \qquad ,met\ \xi_\tau \sim N(0, \sigma^2{}_\xi)$$

Met:
$Z = (Z_1 \ldots, Z_T)'$ en $Z_t = \sqrt{}\ Prev_t$, de Box Cox transformatie van de prevalenties.
$D_{i,t}$ = de waarde van dummy nummer *i* in jaar *t*.
De variabele U_t in het algemene model is dus vervangen door $\mu_t + \varepsilon_t$ met de specifieke structuur zoals hierboven geformuleerd.

7.2 Resultaten

Aangezien er 108 reeksen zijn, is het niet wenselijk om alle reeksen in detail te beschrijven. Bij wijze van voorbeeld wordt de reeks van de prevalenties Langdurige aandoeningen van vrouwen tussen de 25 en 29 jaar behandeld. Er zijn 4 dummy's (zie tabel 3 hoofdstuk 4). De schattingsresultaten worden in de volgende tabel weergegeven:

Tabel 6
Schattingsresultaten prevalenties LA, vrouwen, 25–29 jaar

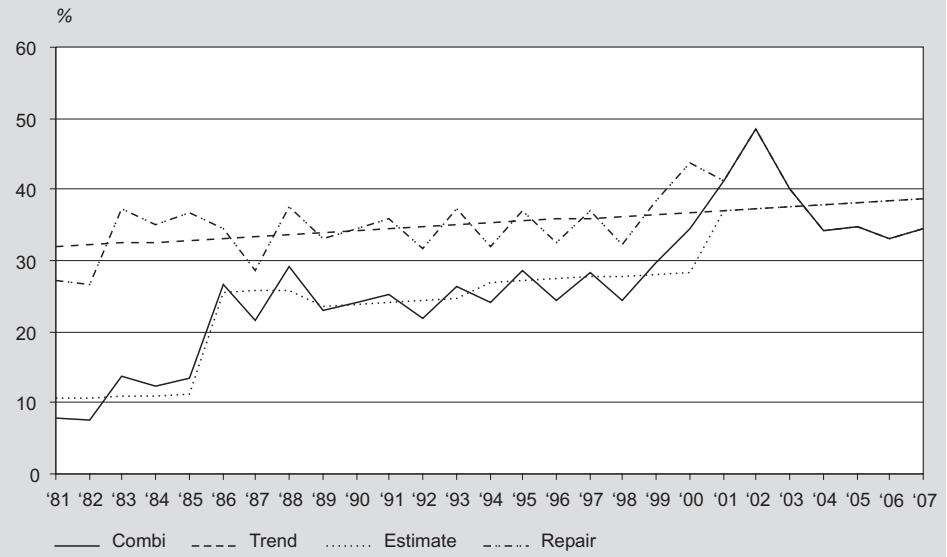
	Parameter	Standaardfout	T-waarde	P-waarden
Variabele				
D1	–2,40	0,95	–2,51	0,02
D2	–0,72	0,80	–0,89	0,37
D3	–0,96	0,62	–1,55	0,12
D4	–0,75	0,37	–2,01	0,05
m_{2005}	6,25	0,22	2,72	0,00
v_{2005}	0,02	0,04	0,50	0,62

Bron: CBS.

De BIC is 3,9 en de Mean Square Error 19,2. Uit de tabel blijkt dat niet alle methodenbreuken significant zijn. Wonnacott (zie paragraaf 3.7) adviseert echter om niet-significante variabelen wel op te nemen als daar inhoudelijke gronden voor zijn en het aantal waarnemingen beperkt is.

In de figuur 6 worden de resultaten weergegeven. Behalve de reeks Combi, worden ook de reeksen Estimate, Repair en Trend weergegeven.

6. De reeksen Combi ,Estimate, Repair en Trend bij prevelanties LA, vrouwen, 25–29 jaar



Bron: CBS.

Tabel 7
De reeksen Trend, Estimate, Repair en Combi van de prevalenties LA, vrouwen, 25–29 jaar

	Prev	Trend	Estimate	Repair
1981	7,9	31,9	10,6	27,1
1982	7,6	32,1	10,7	26,6
1983	13,7	32,4	10,8	37,2
1984	12,5	32,7	11,0	35,2
1985	13,4	32,9	11,1	36,8
1986	26,7	33,2	25,5	34,5
1987	21,5	33,4	25,7	28,6
1988	29,3	33,7	25,9	37,5
1989	23,0	33,9	23,7	33,1
1990	24,1	34,2	23,9	34,4
1991	25,3	34,4	24,1	35,9
1992	21,9	34,7	24,3	31,8
1993	26,4	35,0	24,6	37,2
1994	24,1	35,2	26,9	32,0
1995	28,5	35,5	27,2	37,0
1996	24,5	35,8	27,4	32,5
1997	28,4	36,0	27,6	37,0
1998	24,4	36,3	27,9	32,3
1999	29,7	36,6	28,1	38,4
2000	34,4	36,8	28,3	43,7
2001	41,3	37,1	37,1	41,3
2002	48,5	37,4	37,4	48,5
2003	40,2	37,6	37,6	40,2
2004	34,3	37,9	37,9	34,3
2005	34,6	38,2	38,2	34,6
2006	33,1	38,5	38,5	33,1
2007	34,5	38,7	38,7	34,5

Bron: CBS.

De schattingen van de andere reeksen geven vergelijkbare resultaten. In het algemeen geldt: Er is een goede overeenstemming tussen de originele data en de geschatte waarden. De *BIC*-scores liggen tussen de 0 en 5.

7.3 Het berekenen van de prevalenties op basis van getransformeerde waarden

Om de variantie te stabiliseren is er gekozen voor een Box Cox transformatie. Een transformatie met een wortelfunctie bleek het best te voldoen. Om de schatting van de Y(x,s,t) om te rekenen naar een schatting van de prevalenties(x,s,t) moet formeel de volgende berekening worden uitgevoerd:

$$E(\text{Prev}(t)) = E\{Z_t\}^2 = \{E(Z_t)\}^2 + \text{Var}(Z_t)$$

Doorgaans is de laatste term verwaarloosbaar klein ten opzichte van de eerste term. Daarom wordt in de praktijk alleen de eerste term berekend.

7.4 Trend versus gerepareerde reeks

Aangezien de oorspronkelijke reeksen gekenmerkt worden door sterke schommelingen, is het niet mogelijk om tijdreeksen met een glad verloop te verkrijgen door uitsluitend de methodebreuken te repareren. Een glad verloop ontstaat alleen door een correctie voor trendbreuken. Bij alle uitkomsten zijn daarom ook de trends berekend.

8. De keuze van de schattingsmethode

8.1 De keuze voor de Methode ARIMA of Methode State space

Aangezien er drie schattingsmethoden zijn geformuleerd, is het noodzakelijk om een keuze te maken voor een bepaalde aanpak. Een mogelijkheid is om de keuze te maken op grond van het model met de kleinste BIC.

Deze aanpak zou echter leiden tot de situatie dat voor een aantal leeftijdsgroepen ARIMA-modellen en voor andere leeftijdsgroepen een state space modellen zou worden gebruikt. Tevens kunnen dan voor sommige reeksen wel een trend worden bepaald en voor anderen niet. Dit is niet wenselijk. Er is daarom gezocht naar een globale aanpak die voor iedere reeks hetzelfde is.

De aanpak Methode tijd reg (klassieke regressie) werd niet gekozen omdat er volgens statistische tests sprake is van autocorrelatie, wat niet in overeenstemming is met de modelveronderstellingen.

Een probleem met de ARIMA-modellen is dat ze in bepaalde gevallen gekenmerkt worden door een soort “vertraging”. Het komt er op neer dat het lijkt alsof de schatting van jaar t beter aansluit bij de waarneming van jaar t–1 dan van het jaar t. Dergelijke problemen zouden wellicht niet bestaan als het verantwoord was om complexe autoregressieve structuren (AR-componenten) of voortschrijdende gemiddelden (MA-componenten) te modelleren. Maar dit was doorgaans niet aan de orde.

Een voordeel van de state space modellen is dat ze een goede schatting geven van de lange termijn trendmatige ontwikkeling (Durbin, 2001). Aangezien de interesse voor de gezonde levensverwachting voornamelijk ook de effecten op langere termijn betreft, en niet zozeer de tijdelijke schommelingen in de loop der jaren, is dit een belangrijk argument voor de state space modellen. Daarom is uiteindelijk gekozen voor de Methode state space.

8.2 De keuze voor de trend of de gerepareerde reeks

Aangezien de originele reeksen gekenmerkt worden door sterke schommelingen, wordt dit ook teruggevonden in de gerepareerde reeks. Er valt daarom veel voor te zeggen om in plaats van de gerepareerde reeks uiteindelijk de trend te gebruiken voor het berekenen van de gezonde levensverwachting. Toch moet het effect van deze schommelingen niet overschat worden. Aangezien de gezonde levensverwachting een functie is van vele leeftijdsspecifieke prevalenties, zal het effect van de schommelingen in de regel beperkt zijn. De reeks van de gezonde levensverwachting wordt beschreven in een aparte nota (Lodder, 2009).

9. Conclusie

Voor drie definities van (on)gezondheid (EG, BP en LA) zijn dummy's geformuleerd voor het optreden van methodebreuken in de reeksen van de gecombineerde steekproeven (Gezondheidsenquête en POLS). Er is een Box Cox transformatie toegepast om de variantie te stabiliseren.

Er zijn schattingen gemaakt met de Methode tijd reg, Methode ARIMA en de Methode state space. Vervolgens zijn de gerepareerde reeksen geconstrueerd door te corrigeren voor het kwantitatieve effect van de methodebreuken. Er is gekozen voor de Methode state space. Daarbij is niet zozeer het criterium "laagste BIC-score", maar de wens om de trendmatige ontwikkeling goed in beeld te krijgen van doorslaggevend belang geweest. Naast de gerepareerde reeks is ook de Trend per reeks berekend. Vooral bij Langdurige aandoeningen is deze reeks stijgend in de loop der jaren. In een andere nota wordt beschreven hoe de prevalenties gebruikt kunnen worden om de gezonde levensverwachting te berekenen (Lodder, 2009).

10. Literatuur

Abraham, B. Ledolter, J., 1983, *Statistical methods for forecasting*, Wiley & Sons., New York.

Brakel, J. van den, Smith, P., Compton, S., (2008), Quality procedures for survey transitions – experiments and discontinuities, in: *Journal for Survey Research Method*, in press.

Durbin, J. Koopman, S.J., (2001), *Time Series Analysis by State Space Methods*, Oxford University Press, Oxford.

Granger en Newbold, (1980), *Forecasting Economic Time Series*. Princeton University press, Princeton.

Jagger C., (2006), *Health Expectancy Calculation by the Sullivan Method: a practical guide*. Montpellier: Euro-REVES/INSERM.

Lodder, B.J.H., (2007), *Het repareren van trendbreuken.* CBS, interne nota, Den Haag.

Lodder, B.J.H., Kardal, M., (2009), *De gezonde levensverwachting in de periode 1981–2007*, Centraal Bureau voor de statistiek, Den Haag.

Meeker, W.O., (2001), *Graphical Tools for Exploring and Analyzing Data from Arima Time Series Models*. Department of Statistics. Iowa State University. Ames, IA 50011, Iowa.

Stam, S., Knoops, K., (2009), *Lange tijdreeksen gezonde levensverwachting. Beschikbaarheid van enquêtedata gezondheidsindicatoren*. Centraal Bureau voor de Statistiek, Den Haag.

Wonnacot, R., Wonnacot, T, (1970), *Econometrics*., Wiley & Sons, New York.