Schatten van maandcijfers over de beroepsbevolking

Jan van den Brakel en Sabine Krieg

Publicatiedatum CBS-website: 6 mei 2010

Centraal Bureau voor de Statistiek

Den Haag/Heerlen

Verklaring van tekens

.	=	gegevens ontbreken
*	=	voorlopig cijfer
**	=	nader 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
2008–2009	=	2008 tot en met 2009
2008/2009	=	het gemiddelde over de jaren 2008 tot en met 2009
2008/'09	=	oogstjaar, boekjaar, schooljaar enz., beginnend in 2008 en eindigend in 2009
2006/'07–2008/'09	=	oogstjaar, boekjaar enz., 2006/'07 tot en met 2008/'09

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 - Grafimedia

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

Schatten van maandcijfers over de beroepsbevolking

Jan van den Brakel en Sabine Krieg

Samenvatting:

Het CBS publiceert cijfers over de beroepsbevolking op basis van de Enquête Beroepsbevolking (EBB). De steekproefomvang van dit onderzoek is te klein om voldoende betrouwbare cijfers op maandbasis te publiceren. Om toch betrouwbare cijfers te schatten worden iedere maand cijfers gepubliceerd die gebaseerd zijn op de steekproefinformatie die in de voorgaande drie maanden is waargenomen. In dit rapport wordt een nieuwe schattingsmethodiek beschreven die gebaseerd is op een structureel tijdreeksmodel. Via dit model wordt gebruik gemaakt van informatie uit het verleden om de nauwkeurigheid van de schattingen te verbeteren. Het model houdt expliciet rekening met het roterende panelontwerp van de EBB. Met deze methodiek is het mogelijk om betrouwbare maandcijfers over de beroepsbevolking te produceren.

Trefwoorden: Enquête beroepsbevolking, Gegeneraliseerde regressieschatter, Roterend panel, Structurele tijdreeksmodellen.

1. Inleiding

Het CBS publiceert cijfers over de werkloze, werkzame en totale beroepsbevolking op basis van de Enquête Beroepsbevolking (EBB). Deze cijfers worden gemaakt op jaarbasis, waarbij zeer gedetailleerde uitsplitsingen worden gemaakt naar verschillende sociaaldemografische kenmerken. Daarnaast worden iedere maand actuele cijfers gepubliceerd. Deze schattingen worden gemaakt voor heel Nederland en voor een uitsplitsing naar leeftijd en geslacht in 6 categorieën (mannen 15-24 jaar, mannen 25-44 jaar, mannen 45-64 jaar, vrouwen 15-24 jaar, vrouwen 25-44 jaar, vrouwen 45-64 jaar).

Schattingen van het CBS zijn vaak gebaseerd op de gegeneraliseerde regressieschatter (Särndal et al, 1992). De steekproefomvang van de EBB is echter te klein om met deze schattingsmethodiek op maandbasis voldoende nauwkeurige schattingen te maken. Daarom zijn de maandelijkse cijfers gebaseerd op de waarnemingen van de afgelopen drie maanden. Een consequentie van het gebruik van het driemaandsgemiddelde is dat het cijfer vooral betrekking heeft op de middelste maand. Omdat het cijfer betrekking heeft op de gemiddelde situatie van drie maanden, worden ontwikkelingen in de trend en het maandelijkse seizoenspatroon uitgemiddeld en dus afgevlakt. Het voortschrijdende driemaandsgemiddelde is daarom geen actueel maandcijfer.

Idealiter zouden iedere maand cijfers worden gepubliceerd die uitsluitend betrekking hebben op de situatie op de arbeidsmarkt van die betreffende maand. Daarom is een schattingsmethodiek ontwikkeld, waarbij gebruik wordt gemaakt van een tijdreeksmodel, om iedere maand cijfers te publiceren die uitsluitend betrekking hebben op de situatie op de arbeidsmarkt van de betreffende maand. Deze methodiek maakt het wel mogelijk om voldoende nauwkeurige cijfers op maandbasis te publiceren. In dit rapport wordt deze nieuwe schattingsmethodiek beschreven.

Het rapport is als volgt opgebouwd. In paragraaf 2 wordt de onderzoeksopzet van de EBB beschreven. In paragraaf 3 wordt de nieuwe schattingsmethodiek, gebaseerd op structurele tijdreeksmodellen, beschreven. In paragraaf 4 worden schattingsresultaten gegeven. Het rapport wordt in paragraaf 5 afgerond met een conclusie. Voor een technische toelichting van de oude en de nieuwe methodiek wordt verwezen naar de bijlage.

2. Enquête Beroepsbevolking

2.1 Steekproefontwerp

De Enquête beroepsbevolking (EBB) heeft tot doel informatie te verschaffen over de werkzame en werkloze beroepsbevolking. De doelpopulatie van de EBB bestaat uit alle personen van 15 jaar of ouder, die woonachtig zijn in Nederland, exclusief bewoners van inrichtingen, instellingen en tehuizen. Het steekproefkader is een lijst van adressen gebaseerd op de Geografisch Basis Administratie. Uit dit kader wordt maandelijks een gestratificeerde tweetrapssteekproef van adressen getrokken. Hierbij wordt gestratificeerd naar een kruising tussen de COROP (COmité voor coordinatie Regionale OnderzoeksProgramma) gebieden en de interviewregio's.

In de eerste trap wordt een systematische steekproef van gemeenten getrokken met een insluitkans die evenredig is aan het aantal adressen per gemeente. In de tweede trap wordt uit iedere geselecteerde gemeente een steekproef van adressen getrokken. Om nauwkeurigere schattingen over de beroepsbevolking te kunnen maken, worden diverse bevolkingsgroepen onder- en oververtegenwoordigd in de steekproef. Adressen waar uitsluitend personen van 65 jaar en ouder wonen, zijn ondervertegenwoordigd in de steekproef. Vanaf 2009 zijn adressen waar personen wonen die bij het CWI zijn ingeschreven, oververtegenwoordigd. Ook zijn vanaf 2008 adressen waar jongeren en niet-westerse allochtonen wonen, oververtegenwoordigd. Hierdoor kunnen nauwkeurigere schattingen over de werkloze beroepsbevolking gemaakt worden. Alle huishoudens, met een maximum van drie, die op een steekproefadres worden aangetroffen, worden in de steekproef geselecteerd. Voor alle personen in het huishouden van 15 jaar en ouder wordt via een vragenlijst de arbeidspositie vastgesteld. Indien een of meerdere huishoudleden niet geïnterviewd kunnen worden, zijn proxi-interviews toegestaan. Dat wil zeggen dat een ander huishoudlid, bijvoorkeur een kernlid, de vragen voor de afwezige personen beantwoordt. Huishoudens waarvan een of meer van de geselecteerde personen niet direct of via een proxi-interview responderen, worden behandeld als een niet-responderend huishouden.

2.2 Panelopzet EBB

Van 1987 t/m september 1999 is de EBB uitgevoerd als een doorlopend cross-sectioneel onderzoek. Vanaf oktober 1999 is de EBB gebaseerd op een roterend panelontwerp. De huishoudens die op deze adressen worden aangetroffen worden de eerste keer door een interviewer thuis benaderd, die de elektronische vragenlijst via een persoonlijk vraaggesprek afneemt. Deze manier van data verzamelen wordt aangeduid met Computer Assisted Personal Interviewing (CAPI). Vervolgens worden de huishoudens nog vier keer telefonisch

herbenaderd en wordt een verkorte versie van de vragenlijst afgenomen. Deze manier van data verzamelen wordt aangeduid met Computer Assisted Telephone Interviewing (CATI). Gedurende de periode van 2000 tot 2009 is de maandelijks uitgezette steekproefomvang afgenomen van circa 8000 adressen per maand naar circa 6500 adressen per maand.

In het rotatieschema worden huishoudens vijf keer benaderd met een interval van drie maanden. Dit rotatieschema is gebaseerd op de verslagperiode van de maandelijkse publicaties. Het rotatieschema zorgt ervoor dat het voortschrijdende driemaandsgemiddelde gebaseerd is op 15 onafhankelijke maandsteekproeven. Tabel 1 illustreert dat op ieder tijdstip de data die in het afgelopen kwartaal zijn waargenomen, gebaseerd zijn op 15 verschillende steekproeven.

Tabel 1: Rotatieschema van het EBB panel

	Maand					
Peiling	t-5	t-4	t-3	t-2	t-1	t
1	t-5	t-4	t-3	t-2	t-1	t
2	t-8	t-7	t-6	t-5	t-4	t-3
3	t-11	t-10	t-9	t-8	t-7	t-6
4	t-14	t-13	t-12	t-11	t-10	t-9
5	t-17	t-16	t-15	t-14	t-13	t-12

Voor iedere maand staat in de kolom aangegeven wanneer de steekproef getrokken is die in de vijf peilingen wordt waargenomen.

Doordat de herbenaderingen telefonisch plaats vinden en doordat gebruik wordt gemaakt van een sterk verkorte vragenlijst is met betrekkelijk weinig kosten de steekproefomvang vervijfvoudigd ten opzichte van het cross-sectionele ontwerp zoals dat tot en met 1999 werd gehanteerd.

2.3 Schattingsmethodiek driemaandsgemiddelden

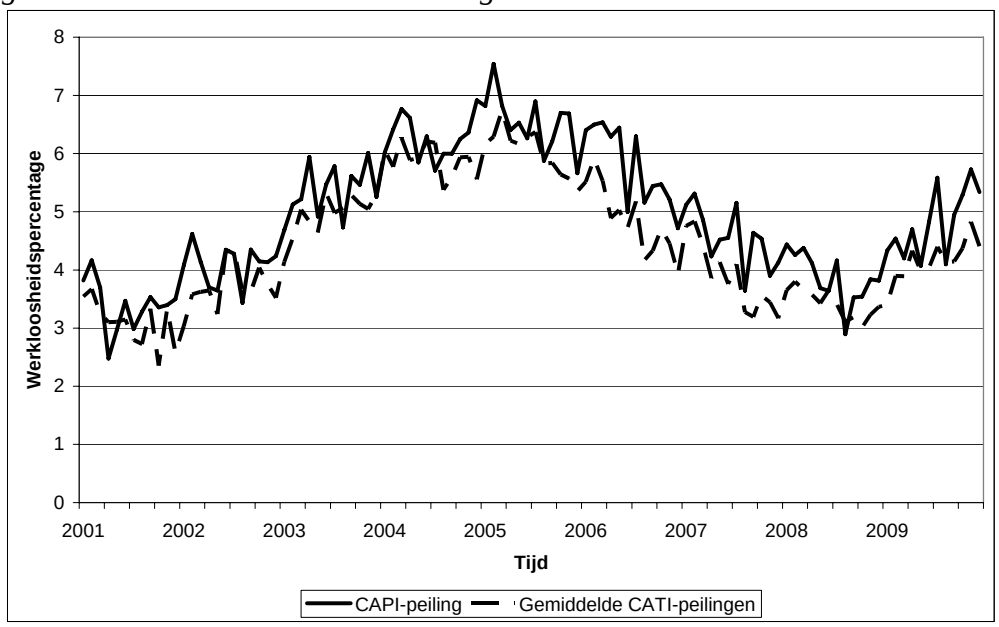
De schattingsmethodiek van de EBB is gebaseerd op de gegeneraliseerde regressieschatter, zie Särndal et al. (1992). Deze schatter behoort tot een klasse van schatters die kan worden geschreven als de som over de gewogen waarnemingen in de steekproef. De gewichten die aan de steekproefelementen worden toegekend, zijn gebaseerd op het steekproefontwerp en beschikbare hulpinformatie over de doelpopulatie. Hierbij gaat het om sociaaldemografische kenmerken waarvoor de populatietotalen exact bekend zijn. Een beschrijving van de schattingsmethodiek voor het driemaandsgemiddelde op basis van de gegeneraliseerde regressieschatter is opgenomen in Bijlage A.1.

Veel statistische bureau’s baseren hun schattingsmethodiek op de gegeneraliseerde regressieschatter. Desalniettemin kent deze methodiek een aantal belangrijke beperkingen. Ten eerste resulteert de gegeneraliseerde regressieschatter bij een kleine steekproefomvang in grote varianties. Bij de EBB doet deze situatie zich onder andere voor indien maandelijkse werkloosheidscijfers worden gemaakt. Om deze reden worden iedere maand voortschrijdende driemaandsgemiddelden gepubliceerd.

Ten tweede treedt er vertekening op in de uitkomsten van de verschillende herbenaderingen van het panel. Dit is een bekend fenomeen dat zich vaak voordoet bij roterende panelontwerpen en wordt in de literatuur aangeduid met de term rotation group bias (RGB),

Bailar (1975). Om een indruk te krijgen van de omvang van de vertekening worden in Figuur 1 twee reeksen voor het maandelijkse werkloosheidspercentage met elkaar vergeleken. Beide reeksen zijn geschat met de gegeneraliseerde regressieschatter. De reeks van de ononderbroken lijn is gebaseerd op de data die zijn verzameld in de eerste peiling. De andere reeks is gebaseerd op de data uit de vier telefonische herbenaderingen. Uit de figuur blijkt dat op basis van de herbenaderingen het werkloosheidspercentage systematisch lager wordt geschat dan op basis van de eerste peiling, en dat de vertekening vrij groot is. De belangrijkste oorzaken voor deze vertekening worden in bijlage B beschreven.

Figuur 1: Maandelijkse werkloosheidspercentage op basis van de eerste peiling versus het gemiddelde van de vier herbenaderingen



Uit de beschrijving van de factoren die de vertekening veroorzaken, blijkt dat de schattingen gebaseerd op de eerste (CAPI) peiling betrouwbaarder zijn dan de schattingen gebaseerd op CATI. Daarom wordt zowel bij de voortschrijdende driemaandsgemiddelden als ook in het tijdreeksmodel aangenomen dat de schattingen gebaseerd op de eerste peiling niet vertekend zijn. De gegeneraliseerde regressieschatter van het driemaandsgemiddelde kan niet rechtstreeks corrigeren voor deze vertekening. Daarom wordt bij deze schattingsmethodiek achteraf op tabelniveau een correctie uitgevoerd voor deze vertekening, zodanig dat de schattingen op het niveau van de uitkomsten van de eerste peiling komen. Dit is beschreven in Bijlage A.2.

3. Schatten van maandcijfers

Zoals aangegeven in de inleiding heeft het voortschrijdende driemaandsgemiddelde een aantal nadelen. Deze cijfers hebben betrekking op de gemiddelde situatie op de arbeidsmarkt voor een periode van drie maanden en zijn daarom gecentreerd rondom de middelste maand. Dit heeft tot gevolg dat de structurele ontwikkeling en het maandelijkse seizoenspatroon binnen de verslagperiode van drie maanden wordt uitgemiddeld en afgevlakt. Het cijfer kan daarom

niet worden geïnterpreteerd als een actueel maandcijfer. Idealiter zouden iedere maand cijfers worden gepubliceerd die uitsluitend betrekking hebben op de situatie op de arbeidsmarkt van die betreffende maand. Daarom is een schattingsmethodiek ontwikkeld waarbij op basis van een structureel tijdreeksmodel voldoende nauwkeurige schattingen worden gemaakt voor maandelijkse cijfers over de beroepsbevolking.

Dankzij het roterende panelontwerp worden iedere maand voor vijf opeenvolgende peilingen data verzameld. Voor iedere afzonderlijke peiling kan een reeks van gegeneraliseerde regressieschattingen voor een doelvariabele op maandbasis worden geconstrueerd. Deze tijdreeksen zijn de input voor een multivariaat structureel tijdreeksmodel. Omdat bij deze schattingsmethodiek het tijdreeksmodel corrigeert voor de vertekening tussen de opeenvolgende herbenaderingen, wordt bij de berekening van deze gegeneraliseerde regressieschattingen een eenvoudige weging toegepast waarbij veel minder gepoogd wordt om te corrigeren voor de vertekening tussen de peilingen. In Bijlage C wordt beschreven hoe voor iedere afzonderlijke peiling gegeneraliseerde regressieschatters op maandbasis worden verkregen.

Het tijdreeksmodel wordt beschreven in paragraaf 3.1. Vervolgens wordt beschreven hoe via dit model maandcijfers over de beroepsbevolking worden geschat (paragraaf 3.2) en dat trendcijfers als alternatief voor seizoensgecorrigeerde cijfers worden berekend (paragraaf 3.3).

3.1 Tijdreeksmodel voor het schatten van maandcijfers

Voor het schatten van maandelijkse cijfers over de beroepsbevolking is een methodiek ontwikkeld die gebaseerd is op een multivariaat structureel tijdreeksmodel. Deze methode is ontwikkeld door Pfeffermann (1991) en in Van den Brakel (2005), Van den Brakel and Krieg (2008a, 2009a, 2009b) verder uitgewerkt en toegepast op de EBB. Voor iedere afzonderlijke peiling wordt een reeks van gegeneraliseerde regressieschattingen geconstrueerd. Deze vijf reeksen worden gemodelleerd via een tijdreeksmodel dat rekening houdt met alle aspecten van het panelontwerp. Dit model bestaat uit de volgende drie componenten:

1. een tijdreeksmodel voor de onbekende populatieparameter,
2. een tijdreeksmodel voor de vertekening tussen de opeenvolgende peilingen,
3. een tijdreeksmodel voor de steekproeffout.

Deze componenten worden in de volgende drie subparagrafen beschreven. In Bijlage D is een technische beschrijving van het model opgenomen.

3.1.1 Tijdreeksmodel voor de populatieparameter

Via deze component van het tijdreeksmodel wordt efficiënter gebruik gemaakt van de steekproefinformatie die is waargenomen in voorgaande periodes. De gegeneraliseerde regressieschatter, afkomstig uit de klassieke steekproeftheorie, beschouwt de populatieparameter in maand t als een vaste maar onbekende waarde die geschat wordt op basis van de steekproefdata. Onder dit paradigma kan voor het schatten van bijvoorbeeld het maandelijkse werkloosheidcijfer alleen gebruik worden gemaakt van de data die in de desbetreffende maand zijn waargenomen. Het werkloosheidscijfer hangt echter sterk samen

met de werkloosheidscijfers uit voorgaande periodes. Het ligt daarom voor de hand om de precisie van de gegeneraliseerde regressieschatter voor het maandelijkse werkloosheidscijfer te verbeteren door gebruik te maken van steekproefinformatie uit voorgaande periodes. In het verleden gebeurde dit door iedere maand cijfers te publiceren die zijn gebaseerd op de afgelopen drie maanden. Hierbij wordt een gemiddelde waarde voor de verslagperiode berekend. Deze berekening gaat voorbij aan de ontwikkeling die binnen deze periode plaatsvindt zoals de structurele veranderingen in de trend en het seizoenspatroon. Op deze manier kan daarom slechts gebruik worden gemaakt van een beperkt aantal periodes uit het nabije verleden. Het is efficiënter om de onbekende populatieparameters in de opeenvolgende maanden op te vatten als een realisatie van een stochastisch proces dat kan worden gemodelleerd aan de hand van een tijdreeksmodel. Aan de hand van dit model wordt het mogelijk om alle beschikbare steekproefinformatie uit het verleden te gebruiken om de precisie van de gegeneraliseerde regressieschatters voor de maandcijfers te verbeteren.

In het multivariate tijdreeksmodel worden de vijf gegeneraliseerde regressieschattingen meegenomen als onafhankelijke schattingen voor de onbekende doelvariabele in maand *t*. Deze doelvariabele wordt gemodelleerd met een tijdreeksmodel dat is opgebouwd uit een trendcomponent, een seizoenscomponent en een storingsterm. Aan de hand van deze componenten wordt naast de steekproefinformatie waargenomen in maand *t*, ook gebruik gemaakt van steekproefinformatie die is waargenomen in voorgaande perioden om de doelvariabele in maand *t* nauwkeuriger te schatten. De trendcomponent is bepalend voor het niveau van de reeks. Via de seizoenscomponent wordt gebruik gemaakt van informatie uit het verleden om de systematische afwijkingen in de afzonderlijke maanden te bepalen. Via de storingsterm worden alle andere veranderingen beschreven, die niet door de trend en de seizoenscomponent verklaard worden.

De trend- en de seizoenscomponent worden gemodelleerd met stochastische modellen waardoor deze componenten tijdsafhankelijk zijn. In deze modellen worden de trend en het seizoenpatroon opgevat als een functie van een aantal onbekende parameters. Voor ieder tijdstip wordt verondersteld dat de waarden voor deze parameters gelijk is aan de waarde uit de voorgaande periode met een kleine afwijking. Hierdoor kunnen deze parameters door de tijd heen geleidelijk van waarde veranderen en daarmee ook de trend en het seizoenspatroon. De flexibiliteit van deze componenten wordt bepaald door de varianties van de afwijkingen tussen de opeenvolgende perioden. Deze varianties worden op basis van de waargenomen tijdreeks geschat via de methode van de grootste aannemelijkheid (maximum likelihood). Naarmate de schattingen voor deze variantietermen groter zijn, worden de trend en het seizoenspatroon flexibeler en heeft informatie uit het verleden minder invloed op de schattingen voor een bepaalde maand.

3.1.2 Tijdreeksmodel voor de rotation group bias

De tweede component van het tijdreeksmodel beschrijft de systematische afwijking tussen de gegeneraliseerde regressieschattingen die zijn gebaseerd op de vijf afzonderlijke peilingen. Het is niet mogelijk om uitsluitend op basis van de beschikbare steekproefinformatie schattingen te maken van de absolute vertekening van de doelvariabelen. Het is wel mogelijk

om de systematische verschillen tussen de vijf peilingen te schatten. Er wordt aangenomen dat de reeks op basis van de eerste peiling niet vertekend is. Dit is een plausibele aanname omdat aan de hand van de oorzaken van de vertekening (zie Bijlage B) kan worden gemotiveerd dat de uitkomsten van de eerste peiling het meest betrouwbaar zijn. Deze veronderstelling wordt ook gehanteerd bij de driemaandsgemiddelden.

Vervolgens kan via het tijdreeksmodel de systematische vertekening van de opeenvolgende herbenaderingen ten opzichte van de eerste peiling worden gemodelleerd. Uit analyses blijkt dat zowel het niveau als ook het seizoenspatroon van de vervolgpeilingen systematisch verschilt ten opzichte van de eerste peiling, Van den Brakel and Krieg (2008a, 2009a, 2009b). Omdat voor het modelleren van de vertekening in het seizoenspatroon veel extra modelparameters nodig zijn en het effect op de puntschattingen betrekkelijk klein is, is uiteindelijk gekozen voor een model dat alleen rekening houdt met niveauverschillen tussen de eerste peilingen en de vervolgpeilingen, Van den Brakel en Krieg (2008b). Door deze vertekening expliciet te modelleren wordt voorkomen dat deze vertekening in de schattingen voor de maandcijfers terecht komt.

3.1.3 Tijdreeksmodel voor de steekproeffout

In de derde en laatste component van het tijdreeksmodel worden de steekproeffouten gemodelleerd. In het tijdreeksmodel dat op deze manier ontstaat wordt de gegeneraliseerde regressieschatting van de eerste peiling opgevat als de som van de echte populatievariabele (gemodelleerd via het model beschreven in subparagraaf 3.1.1) en een steekproeffout. De gegeneraliseerde regressieschatting van iedere vervolgpeiling wordt opgevat als de som van de echte populatievariabele, de vertekening ten opzichte van de eerste peiling (gemodelleerd via het model beschreven in subparagraaf 3.1.2) en een steekproeffout.

Het roterende panel heeft tot gevolg dat de steekproeffouten van de verschillende peilingen op verschillende tijdstippen met elkaar samenhangen. Omdat de steekproef van de eerste peiling voor het eerst wordt waargenomen, hangt de steekproeffout van deze peiling niet samen met steekproeffouten uit het verleden. De steekproeffout uit de tweede peiling in maand *t* hangt samen met de steekproeffout uit de eerste peiling van maand *t-3*, omdat beide peilingen betrekking hebben op dezelfde steekproef. Om dezelfde reden hangt de steekproeffout uit de derde peiling in maand *t* samen met de steekproeffout uit de tweede peiling van maand *t-3* en de steekproeffout uit de eerste peiling van maand *t-6*, etc. De derde component van het tijdreeksmodel modelleert deze autocorrelatie in de steekproeffouten. Dit resulteert in een verdere reductie van de standaardfout van de schattingen voor de maandelijkse werkloosheidscijfers.

De varianties en de autocorrelaties van de steekproeffouten worden geschat op basis van steekproefdata. Deze schattingen worden aan het tijdreeksmodel als priorinformatie meegegeven. Variantieschattingen voor de steekproeffouten worden gemodelleerd via de methodiek ontwikkeld door Binder and Dick (1990). De autocorrelaties van de steekproeffouten worden geschat via de procedure die is ontwikkeld door Pfeffermann et al. (1998). De autocorrelatiestructuur van de steekproeffouten in de tweede tot en met de vijfde peiling wordt gemodelleerd aan de hand van een AR(1) model, zie Van den Brakel and Krieg

(2009a, 2009b). Dit betekent dat de correlatie tussen de gegeneraliseerde regressieschattingen voor maandelijkse cijfers over de beroepsbevolking op basis van de opeenvolgende peilingen wordt meegenomen in de modelschattingen voor de maandelijkse cijfers over de beroepsbevolking.

3.2 Schatten van maandcijfers

Het tijdreeksmodel, beschreven in paragraaf 3.1, wordt geanalyseerd met het zogenaamde Kalmanfilter. Hiermee wordt voor iedere maand een optimale schatting gemaakt voor de doelvariabele en de modelparameters op basis van de informatie die beschikbaar is tot en met deze periode. Dit zijn de zogenaamde gefilterde schattingen. In Bijlage E wordt deze methode en de daarvoor benodigde software kort toegelicht.

De maandelijkse cijfers over de beroepsbevolking bestaan uit de gefilterde schatting voor de trend plus de seizoenscomponent. Zoals opgemerkt in paragraaf 3.1.1, is de trend bepalend voor het niveau van de reeks en de seizoenscomponent voor de systematische afwijking hiervan in de afzonderlijke maanden. Beide componenten gebruiken informatie uit het verleden om tot een optimale schatting te komen. De mate waarin informatie uit het verleden wordt gebruikt bij het schatten van het maandcijfer hangt af van de flexibiliteit van de trend- en de seizoenscomponent. Naarmate de componenten flexibeler worden ingesteld is de bijdrage van informatie uit het verleden geringer en wordt de invloed van de gegeneraliseerde regressieschattingen uit de betreffende maand groter. De schattingsmethodiek bepaalt aan de hand van de waargenomen tijdreeks welke mate van flexibiliteit voor beide componenten optimaal is. Dit gebeurt door de variantiecomponenten voor de trend en het seizoensmodel te schatten via de methode van de grootste aannemelijkheid (maximum likelihood).

De beschikbare informatie uit de vijf peilingen wordt door het tijdreeksmodel geïntegreerd tot één gefilterde schatting voor de doelvariabele. Door de verschillen tussen de vijf reeksen van gegeneraliseerde regressieschattingen expliciet te modelleren, wordt voorkomen dat deze afwijkingen de schattingen voor de doelvariabele vertekenen.

Aan de hand van de hierboven beschreven methodiek worden iedere maand schattingen gemaakt voor:

1. Werkloze beroepsbevolking,
2. Werkzame beroepsbevolking,
3. Totale beroepsbevolking.

Voor deze drie variabelen worden schattingen gemaakt voor Nederland en een uitsplitsing naar leeftijd en geslacht in zes categorieën. In een eerste stap wordt voor elke afzonderlijke doelvariabele het tijdreeksmodel toegepast om een maandcijfer te schatten. De consequentie van deze werkwijze is dat de som van de werkloze en werkzame beroepsbevolking niet meer exact gelijk is aan de totale beroepsbevolking, zowel voor heel Nederland als ook voor de uitsplitsingen naar leeftijd en geslacht. Ook komt de som over deze zes categorieën niet exact overeen met de schatting voor heel Nederland. Daarom worden in een tweede stap de schattingen voor bovengenoemde doelvariabelen consistent gemaakt. Dit gebeurt door de

consistentie-eisen als restricties op te leggen via een Lagrangefunctie. Via deze methode worden de schattingen minimaal aangepast, waarbij de omvang van de aanpassing evenredig is met de variantie van de modelschattingen. Dat wil zeggen, hoe onbetrouwbaarder de schatting voor een doelvariabele, hoe groter de toegestane aanpassing van de variabele. Deze methode is in detail beschreven in Krieg en Van den Brakel (2008b) en Van den Brakel, Krieg en Souren (2009a). Ten slotte worden de maandelijkse werkloosheidspercentages berekend als quotiënt van de maandelijkse schattingen voor de werkloze en totale beroepsbevolking.

Over het algemeen zijn statistische bureau's terughoudend om expliciet gebruik te maken van statische modellen bij het publiceren van cijfers. Voor zover bekend wordt alleen bij het US Bureau of Labor Statistics gebruik gemaakt van structurele tijdreeksmodellen voor het samenstellen van maandelijkse werkloosheidscijfers, Tiller (1992, 2006).

3.3 Trendcijfers als alternatief voor seizoensgecorrigeerde cijfers

Van de voortschrijdende driemaandsgemiddelden worden seizoensgecorrigeerde cijfers gepubliceerd. Het tijdreeksmodel genereert gefilterde trendschattingen voor de maandcijfers over de beroepsbevolking. Deze vervangen de seizoensgecorrigeerde voortschrijdende driemaandsgemiddelden. De gefilterde trendschattingen voor de werkloze, werkzame en totale beroepsbevolking en de uitsplitsingen naar leeftijd en geslacht in zes categorieën worden consistent gemaakt via een Lagrangefunctie die ook voor de maandelijkse schattingen wordt gebruikt. Vervolgens worden de gefilterde trendschattingen voor de maandelijkse werkloosheidspercentages berekend door het quotiënt van de gefilterde trendschattingen voor de maandelijkse werkloze en totale beroepsbevolking te bepalen.

De standaard methodiek op het CBS voor het berekenen van seizoensgecorrigeerde cijfers is gebaseerd op X-12-ARIMA, Findley *et al*. (1998). Deze methodiek wordt ook toegepast om seizoensgecorrigeerde voortschrijdende driemaandsgemiddelden te berekenen. Aan de hand van een set van digitale filters wordt voor een reeks van, in dit geval voortschrijdende driemaandsgemiddelden, een seizoenspatroon berekend. Het seizoensgecorrigeerde cijfer wordt vervolgens bepaald door dit seizoenspatroon van de originele reeks af te trekken.

Het gebruik van een gefilterde trendschatting als alternatief voor de seizoensgecorrigeerde maandcijfers leidt tot een andere interpretatie van de cijfers. Het seizoensgecorrigeerde cijfer zoals dat berekend werd met X-12-ARIMA bestaat uit een trend, cyclische patronen met een periode die afwijkt van de jaarcyclus van het seizoenspatroon, witte ruis in de reeks van de doelvariabele en steekproeffouten. De gefilterde trendschatting bevat geen witte ruis en steekproeffouten. Afhankelijk van de flexibiliteit van het trendmodel worden cyclische bewegingen met een periode langer dan een jaar opgenomen in het seizoensgecorrigeerde cijfer.

Het publiceren van een trendcijfer heeft een aantal voordelen ten opzichte van seizoensgecorrigeerde cijfers. De seizoensgecorrigeerde voortschrijdende driemaandsgemiddelden worden na publicatie een aantal malen gereviseerd omdat met de informatie die na publicatie van het cijfer beschikbaar komt betere schattingen voor het seizoensgecorrigeerde cijfer kunnen worden gemaakt. De gefilterde trend op basis van het tijdreeksmodel is stabiel genoeg om zonder revisiestrategie niveauschattingen en

ontwikkelingen te publiceren. Zweden en Finland publiceren een trend in plaats van een seizoensgecorrigeerd cijfer over de beroepsbevolking.

Tot slot wordt opgemerkt dat X-12-ARIMA geen rekening houdt met de autocorrelatie die geïntroduceerd wordt door het roterende panelontwerp van de EBB. Omdat een deel van de autocorrelatie in de steekproeffouten ten onrechte wordt opgenomen in de trend kan de methodiek van X-12-ARIMA in deze situatie leiden tot vertekende schattingen voor de seizoensgecorrigeerde cijfers, Pfeffermann et al. (1998), Tiller (2006). Om deze reden wordt bij het US Bureau of Labor Statistics een structureel tijdreeksmodel toegepast voor het schatten seizoensgecorrigeerde cijfers, Tiller (2006).

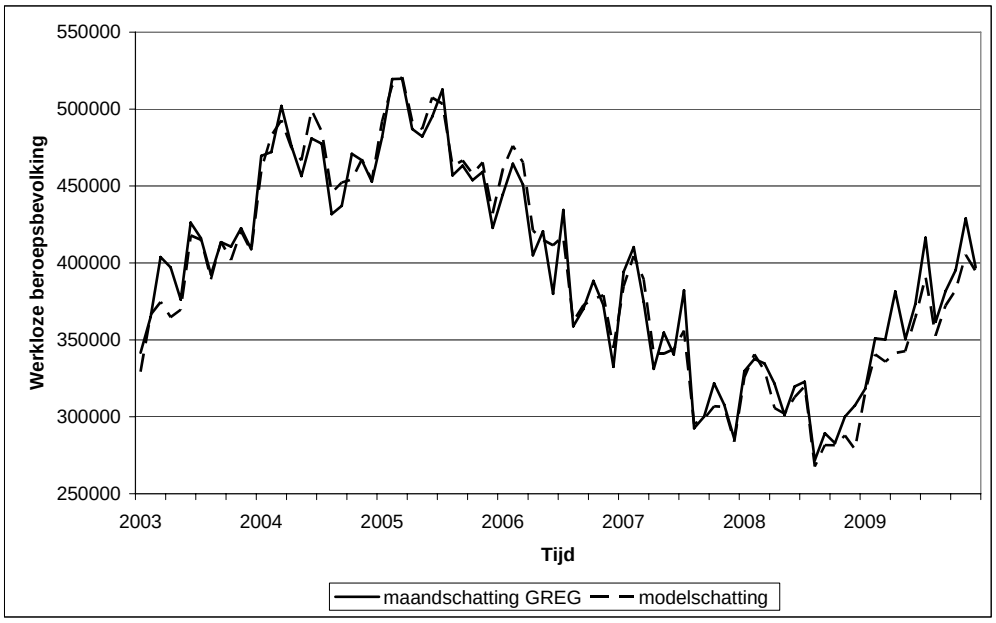
4. Resultaten

In deze paragraaf is het tijdreeksmodel gebruikt om schattingen te maken voor de omvang van de werkloze beroepsbevolking op maandbasis voor de periode van januari 2003 tot en met juli 2009. Alle peilingen van het roterende panelontwerp worden vanaf januari 2001 waargenomen. Het Kalmanfilter dat gebruikt wordt om de tijdreeksen te analyseren heeft een opstartperiode nodig om betrouwbare schattingen voor de onbekende parameters van het tijdreeksmodel te genereren. Omdat gedurende deze periode geen betrouwbare schattingen voor de doelvariabelen worden gegenereerd, worden de resultaten vanaf januari 2003 gepresenteerd. Voor deze periode zijn ook schattingen gemaakt voor de omvang van de werkloze beroepsbevolking op basis van de gegeneraliseerde regressieschatter op maandbasis. Hierbij is dezelfde methodiek toegepast om te corrigeren voor de vertekening tussen de opeenvolgende peilingen als bij de voortschrijdende driemaandsgemiddelden. Deze laatste reeks wordt in de onderstaande figuren aangeduid met de term “maandschatting GREG”.

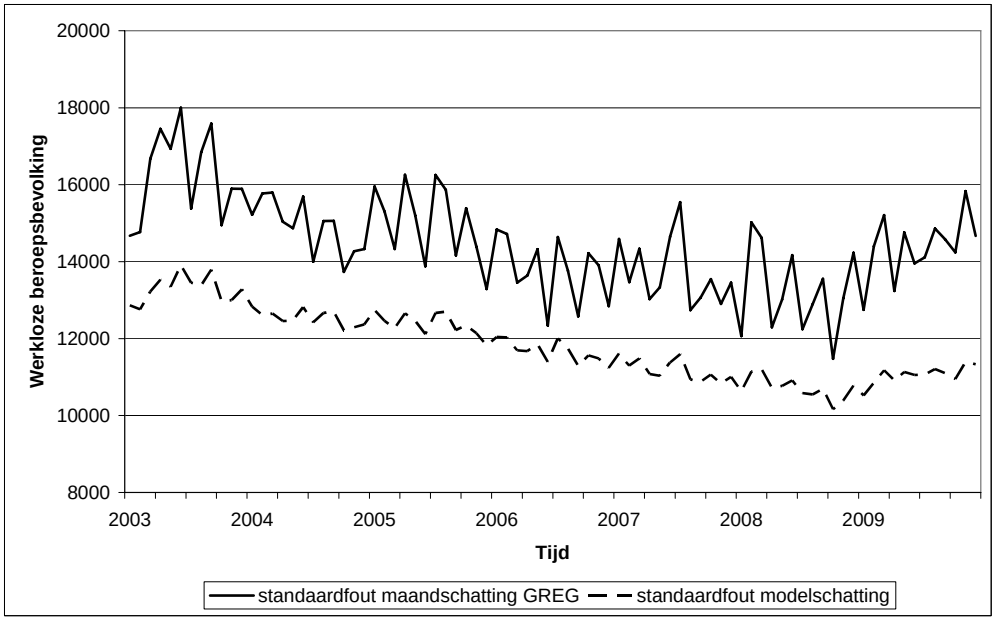
In Figuur 2 worden de reeksen van de schattingen voor de omvang van de werkloze beroepsbevolking op basis van het tijdreeksmodel en de gegeneraliseerde regressieschatter, beide op maandbasis, met elkaar vergeleken. Beide reeksen liggen op hetzelfde niveau. Dit impliceert dat het tijdreeksmodel en de tabelcorrectie (Bijlage A.2) die bij de gegeneraliseerde regressieschatter wordt toegepast op een vergelijkbare wijze corrigeren voor de vertekening tussen de opeenvolgende peilingen. Dit komt omdat bij beide methoden de uitkomsten van de herbenaderingen worden bijgesteld naar de uitkomsten van de eerste peiling. Verder is het verloop van de reeks van de gegeneraliseerde regressieschatter wat onregelmatiger ten opzichte van de schattingen op basis van het tijdreeksmodel. Dit komt omdat een aantal pieken en dalen in de reeks van de gegeneraliseerde regressieschatter door het tijdreeksmodel worden opgevat als steekproeffouten en daardoor uit de schatting voor de omvang van de werkloze beroepsbevolking worden weg gefilterd.

In Figuur 3 worden de standaardfouten van beide reeksen met elkaar vergeleken. De schattingen op basis van het tijdreeksmodel zijn preciezer, vooral omdat bij deze schattingen gebruik wordt gemaakt van steekproefinformatie uit het verleden.

Figuur 2: Schattingen omvang werkloze beroepsbevolking op maandbasis op basis van de gegeneraliseerde regressieschatter en het tijdreeksmodel

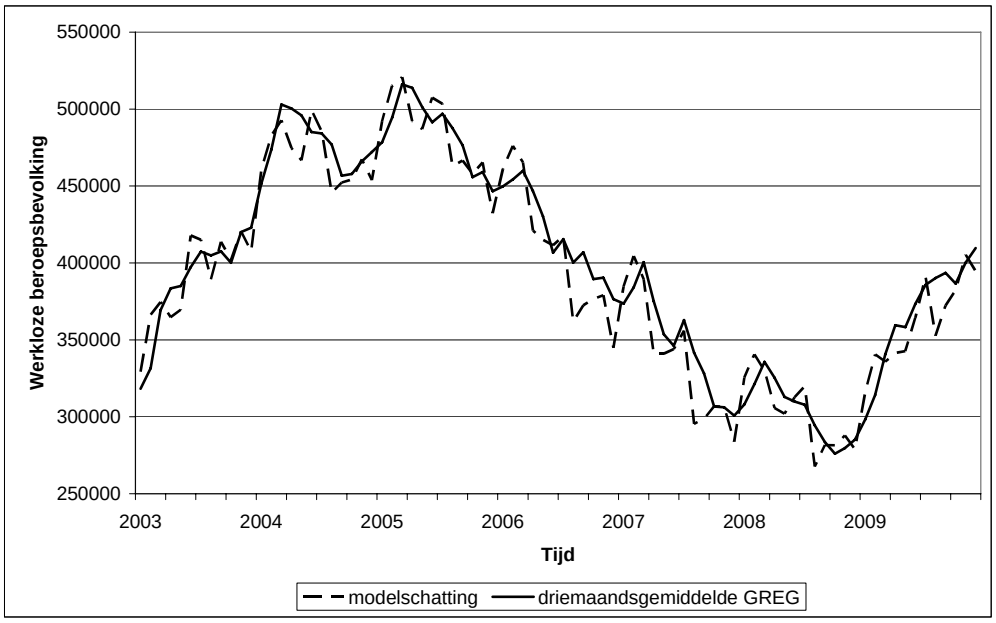


Figuur 3: Standaardfout van de schattingen omvang werkloze beroepsbevolking op maandbasis op basis van de gegeneraliseerde regressieschatter en het tijdreeksmodel



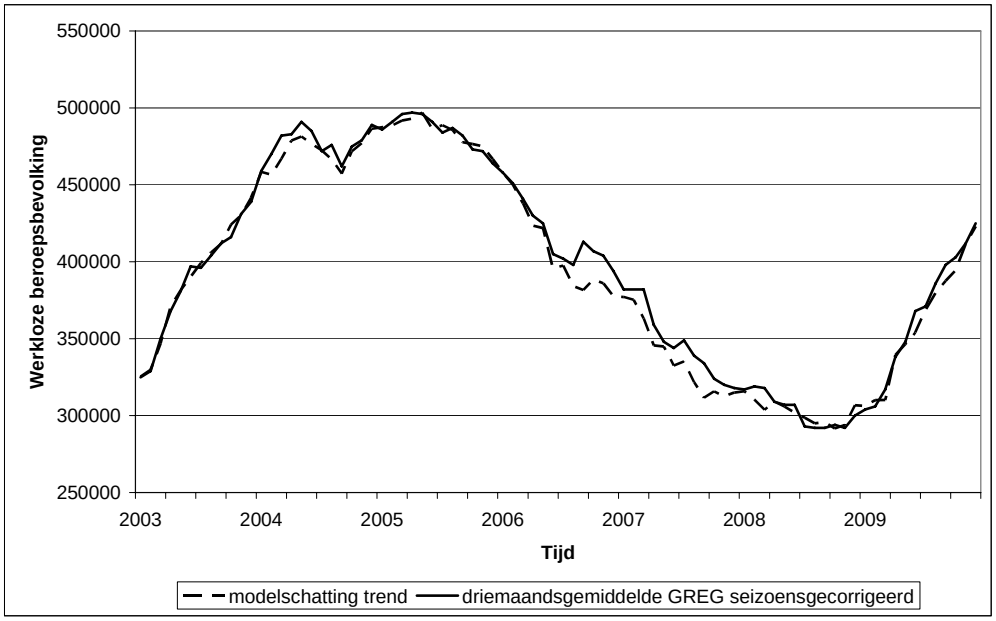
In Figuur 4 wordt de reeks van de omvang van de werkloze beroepsbevolking op maandbasis geschat via het tijdreeksmodel, vergeleken met het voortschrijdend driemaandsgemiddelde. Zoals blijkt uit Figuur 4 is het seizoenspatroon in het voortschrijdende driemaandsgemiddelde uitgevlakt ten opzichte van de maandcijfers op basis van het tijdreeksmodel.

Figuur 4: Schattingen omvang werkloze beroepsbevolking op maandbasis op basis van het tijdreeksmodel en het voortschrijdend driemaandsgemiddelde



In Figuur 5 worden de reeksen van de seizoensgecorrigeerde voortschrijdende driemaandsgemiddelde en de gefilterde trendschattingen op basis van het tijdreeksmodel met elkaar vergeleken. Het seizoensgecorrigeerde driemaandsgemiddelde is verkregen via X-12-ARIMA. Opvallend zijn de verschillen tussen beide reeksen gedurende de periode van een dalende trend in 2006 en 2007. De seizoensgecorrigeerde reeks van het voortschrijdende driemaandsgemiddelden laat een tweetal kortstondige stijgingen van de werkloosheid zien terwijl de gefilterde trend op basis van het tijdreeksmodel sterker vasthoudt aan de dalende werkloosheid. Er zijn geen inhoudelijke verklaringen voor deze kortstondige stijging van de werkloosheid. De voortschrijdende driemaandsgemiddelden zijn gevoeliger voor uitschieters in de afzonderlijke maandschattingen dan de gefilterde trendschattingen van het tijdreeksmodel. Janssen en Souren (2009) concluderen daarom dat deze kortstondige stijgingen eerder kunnen worden verklaard door samenvoeging van drie maanden met duidelijk verschillende seizoenspatronen in combinatie met meet- en steekproeffouten. Voor een uitgebreidere vergelijking van de voortschrijdende driemaandsgemiddelden en maandcijfers geschat op basis van het tijdreeksmodel wordt verwezen naar Jansen en Souren (2009) en Van den Brakel, Krieg en Souren (2009b).

Figuur 5: Schattingen omvang werkloze beroepsbevolking op maandbasis op basis van het tijdreeksmodel en het voortschrijdend driemaandsgemiddelde gecorrigeerd voor seizoensinvloeden



5. Conclusie

Veel statistische bureau's zijn terughoudend in het gebruik van schattingsmethodieken waarbij expliciet gebruik gemaakt wordt van statistische modellen. Veelal worden schattingsmethodieken gehanteerd afkomstig uit de traditionele steekproeftheorie zoals de gegeneraliseerde regressieschatter. Deze methodiek kent een aantal belangrijke beperkingen. Ten eerste resulteert de gegeneraliseerde regressieschatter bij een kleine steekproefomvang in grote varianties. Bij de EBB doet deze situatie zich onder andere voor indien maandelijkse werkloosheidscijfers worden gemaakt. Ten tweede kan de gegeneraliseerde regressieschatter slechts in beperkte mate omgaan met de vertekening in de uitkomsten van de verschillende herbenaderingen van een roterend panelontwerp.

Voor de EBB is een schattingsmethodiek ontwikkeld waarbij gebruik wordt gemaakt van een multivariaat structureel tijdreeksmodel. Aan de hand van deze methodiek kunnen voldoende nauwkeurige schattingen over de beroepsbevolking op maandbasis worden gemaakt. Dit komt omdat via het tijdreeksmodel op efficiënte wijze gebruik wordt gemaakt van steekproefinformatie uit voorgaande perioden. Daarnaast houdt het model rekening met de vertekening tussen de opeenvolgende peilingen en de autocorrelatie ten gevolge van het panelontwerp.

Om bovengenoemde redenen is deze methodiek zeer geschikt om officiële cijfers over de beroepsbevolking op maandbasis te publiceren. In plaats van seizoensgecorrigeerde cijfers kunnen trendcijfers worden gepubliceerd. Het belangrijkste voordeel hiervan is dat deze stabiel genoeg zijn om zonder revisiestrategie niveauschattingen en ontwikkelingen te publiceren.

Referenties

Bailar, B.A. (1975). The Effects of Rotation Group Bias on Estimates from Panel Surveys. *Journal of the American Statistical Association*, 70, pp. 23-30.

Binder, D.A., and J.P. Dick (1990). A method for the analysis of seasonal ARIMA models. *Survey Methodology*, 16, pp. 239-253.

Brakel, J.A. van den (2005). Small Area Estimators for the Dutch Labour Force Survey using Structural Time Series Models. Research paper, BPA nr: TMO-R&D-2005-05-02-JBRL, Statistics Netherlands, Heerlen.

Brakel, J.A. van den, and S. Krieg (2008a). Estimation of the Monthly Unemployment Rate through Structural Time Series Modelling in Rotating Panel Design. Discussion paper 08003, Statistics Netherlands.

Brakel, J.A. van den, and S. Krieg (2008b). Tijdreeksmodellen voor het schatten van maandelijkse cijfers over werkgelegenheid. CBS-nota, BPA nr.: DMH-2008-07-03-JBRL, Centraal Bureau voor de Statistiek, Heerlen.

Brakel, J.A. van den, and S. Krieg (2009a). Structural time series modelling of the Monthly Unemployment in a Rotating Panel Design. Discussion paper 09031, Statistics Netherlands.

Brakel, J.A. van den, and S. Krieg (2009b). Estimation of the Monthly Unemployment Rate through Structural Time Series Modelling in Rotating Panel Design. *Survey Methodology*, in press.

Brakel, J.A. van den, S. Krieg en M. Souren (2009a). Consistentie van modelschattingen voor werkloosheidscijfers. CBS-nota, BPA nr.: DMH-2009-11-20-JBRL, Centraal Bureau voor de Statistiek, Heerlen.

Brakel, J.A. van den, S. Krieg en M. Souren (2009b). Maandcijfers versus voortschrijdende driemaandsgemiddelden over de beroepsbevolking. CBS-nota, BPA nr.: DMH-2009-11-17-JBRL, Centraal Bureau voor de Statistiek, Heerlen.

CBS (2008). Methoden en definities Enquête Beroepsbevolking 2007. Extern CBS rapport, Centraal Bureau voor de Statistiek, Heerlen. http://www.cbs.nl/NR/rdonlyres/0C9ADB8B-955D-4E53-90FB-6FE49D314229/0/EBBMethodenendefinities2007.pdf

Cuppen, M. en G.H. Martinus (2001). Weegmodel voor de Enquête Beroepsbevolking. Interne CBS nota, bpa nr.: 1948-01-TMO. Centraal Bureau voor de Statistiek, Heerlen.

Huang, E. T. and Fuller, W. A. (1978). Nonnegative regression estimation for survey data, in Proceedings of the Social Statistics Session, American Statistical Association, 300-303.

Doornik, J.A. (2007). *Object-oriented matrix programming using Ox 6th edition*. London: Timberlake Consultants Press.

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

Findley, D.F., Monsell, B.C., Bell, W.R., Otto, M.C. & Chen, B.C. (1998). New capabilities and methods of the X-12-ARIMA Seasonal Adjustment Program. *Journal of Business and Economic Statistics*, 16, 127-176 (with Discussion).

Harvey, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

Hilbink, K., C.A.M. van Berkel and J.A. van den Brakel, (2000). Methodology of the Dutch Labour Force Survey, 1987-1999. Research paper, BPA nr. 2297-00-RSM, Statistics Netherlands, Heerlen.

Huang, E.T. en W.A. Fuller (1978). Nonnegative Regression Estimation for Survey Data, *Proceedings of the Section on Social Statistics*, American Statistical Association 1978, pp. 300-303.

Janssen, B. en M. Souren (2009). Plausibiliteit maandcijfers over de beroepsbevolking. Interne CBS nota, BPA nr.: SAH-2009-07-13-MSUN. Centraal Bureau voor de Statistiek, Heerlen.

Koopman, S.J. (1997). Exact initial Kalman filtering and smoothing for non-stationary time series models. *Journal of the American Statistical Association*, 92, pp.1630-1638.

Koopman, S.J., N. Shephard, and J.A. Doornik (2008). *SsfPack 3.0: Statistical algorithms for models in state space form*. London: Timberlake Consultants Press.

Krieg, S., en J.A. van den Brakel (2008a). Het schatten van maandelijkse werkloosheidscijfers via de gegeneraliseerde regressieschatter in een roterend panelontwerp. CBS-nota, BPA nr.: DMH-2008-11-24-SKRG, Centraal Bureau voor de Statistiek, Heerlen.

Krieg, S., en J.A. van den Brakel (2008b). Consistentie van modelschattingen. CBS-nota, BPA nr.: DMH-2008-12-11-SKRG, Centraal Bureau voor de Statistiek, Heerlen.

Lemaître, G. and J. Dufour (1987). An Integrated Method for Weighting Persons and Families. *Survey Methodology*, 13, 199-207.

Nieuwenbroek, N. and H.J. Boonstra (2002). Bascula 4.0 Reference Manual, BPA nr 279-02-TMO, Statistics Netherlands, Heerlen.

Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business & Economic Statistics*, 9, pp. 163-175.

Pfeffermann, D., M. Feder, and D. Signorelli (1998). Estimation of autocorrelations of survey errors with application to trend estimation in small areas. *Journal of Business & Economic Statistics*, 16, pp. 339-348.

Särndal, C-E., B. Swensson and J. Wretman (1992). *Model Assisted Survey Sampling.* New York: Springer Verlag.

Tiller, R.B. (1992). Time series modelling of sample survey data from the U.S. current population survey, *Journal of Official Statistics*, 8, 149-166.

Tiller, R.B. (2006). Model-based labor force estimates for sub-national areas with large survey errors. Technical report, Bureau of Labor Statistics, Washington, http://www.bls.gov/ore/abstract/st/st060010.htm

Bijlage A: Schattingsmethodiek driemaandsgemiddelden

A.1: Weegmodel voor gegeneraliseerde regressieschatter

De gegeneraliseerde regressieschatter behoort tot een klasse van schatters die kan worden geschreven als de som over de gewogen waarnemingen in de steekproef. Aan de steekproefelementen worden gewichten toegekend, zodanig dat de som over de gewogen waarnemingen in de steekproef een bij benadering zuivere schatter is voor het onbekende populatie totaal. De gewichten die aan de steekproefelementen worden toegekend, zijn in eerste instantie gebaseerd op het steekproefontwerp dat gebruikt is om een kanssteekproef uit de doelpopulatie te trekken. In dat geval wordt de Horvitz-Thompsonschatter verkregen. Bij de EBB worden de insluitgewichten berekend op basis van het steekproefontwerp, waarbij tevens rekening wordt gehouden met regionale verschillen in responskansen. Dit gebeurt door het insluitgewicht te baseren op de netto respons in verschillende regio's.

De nauwkeurigheid van deze schatter wordt doorgaans verbeterd door gebruik te maken van beschikbare hulpinformatie over de doelpopulatie. Hierbij gaat het om sociaaldemografische kenmerken waarvoor de populatietotalen exact bekend zijn. De gewichten worden zodanig aangepast dat de geschatte totalen van de gebruikte hulpvariabelen precies overeenkomen met de hiervoor bekende populatietotalen waarnaar gewogen wordt. Dit kan op diverse manieren gebeuren. Bij de gegeneraliseerde regressieschatter worden de gewichten afgeleid via een regressiemodel dat het verband specificeert tussen de doelvariabele en de hulpvariabelen waarnaar gewogen wordt, zie Särndal et al. (1992), hoofdstuk 6.

Indien het regressiemodel goed past bij de data, dan zal de gegeneraliseerde regressieschatter die op dit model is gebaseerd, de variantie van de Horvitz-Thompsonschatter reduceren. In deze situatie zal het gebruik van hulpinformatie ook leiden tot het gedeeltelijk corrigeren van de vertekening ten gevolge van selectieve non-respons. Via de gegeneraliseerde regressieschatter wordt slechts één keer een set van gewichten berekend waarmee vervolgens alle doelvariabelen van het steekproefonderzoek kunnen worden geschat. Dit komt omdat de gewichten alleen afhangen van het steekproefontwerp en de hulpinformatie in het weegmodel. Dit zorgt tevens voor consistentie tussen de verschillende publicatietabellen van het steekproefonderzoek. Sinds 2008 wordt het volgende model voor de weging van de driemaandsgemiddelden toegepast:

Geslacht2 × *Afkomst20* + *LeeftijdGeslacht3* × *GroteGemeente180* + *Geslacht2* × *Leeftijd43* + *CWIinschrijvingsduur5* + *CWIregio13* + *Inkomen6* + *Looncategorieregio33* + *TypeHuishouden3* + *Arbeidspositie(t-1)5* + *Opleiding(t-1)5* + *LeeftijdGeslachtCAPICATI22*.

In elke weegterm geeft het getal aan in hoeveel klassen een bepaalde categorie verdeeld is. Voor alle weegtermen geldt dat de populatietotalen exact bekend zijn, met uitzondering van de weegtermen *Arbeidspositie(t-1)5* en *Opleiding(t-1)5.* Deze twee variabelen staan voor een indeling van arbeidspositie en opleiding naar vijf categorieën waarvan de populatietotalen zijn geschat op basis van de waarnemingen uit de vorige driemaandsperiode. De laatste term van het weegmodel, *LeeftijdGeslachtCAPICATI22,* is opgenomen om het aandeel CAPI- en CATI-data over de tijd constant te houden in een verhouding (3:7). De weegtermen *Arbeidspositie(t-1)5* en *Opleiding(t-1)5* zijn opgenomen om te corrigeren voor de vertekening

die optreedt tussen de opeenvolgende peilingen van het roterende panel. Omdat deze termen slechts gedeeltelijk corrigeren voor deze vertekening, worden de definitieve schattingen gecorrigeerd via de procedure beschreven in Bijlage A.2.

Bij de weging wordt de methode van Lemaître and Dufour (1987) voor het consistent wegen naar persoons- en huishoudkenmerken toegepast, waardoor alle personen binnen één huishouden hetzelfde gewicht krijgen. *Verder* wordt het algoritme van Huang and Fuller (1978) toegepast om negatieve gewichten te voorkomen. De weging wordt uitgevoerd met het software pakket Bascula, Nieuwenbroek en Boonstra (2002).

Voor meer informatie over de opzet en de methodologie van de EBB gedurende de periode 1987-1999 wordt verwezen naar Hilbink e.a. (2000). Voor een uitgebreide beschrijving van de opzet en de weging ten tijde van het panelontwerp wordt verwezen naar CBS (2008) en Cuppen en Martinus (2001).

A.2: Correctie voor RGB

Zoals aangegeven in paragraaf 2 en 3, zijn de schattingen gebaseerd op de gegeneraliseerde regressieschatter vertekend. Voor de berekening van driemaandsgemiddeldes wordt voor deze vertekening gecorrigeerd door *correcties* op tabelniveau toe te passen. Deze correctie is gebaseerd op een ratio met in de teller een schatting voor de doelvariabele op basis van de eerste peiling over de afgelopen 12 kwartalen en in de noemer een schatting over dezelfde periode op basis van alle vijf de peilingen. Per kwartaal wordt een nieuwe correctiefactor berekend. De correctiefactor wordt berekend voor werkloze en werkzame beroepsbevolking voor een indeling in 10 klassen (geslacht maal leeftijd). Schattingen voor heel Nederland en voor grotere deelpopulaties worden door optellen van de betreffende klassen berekend. Schattingen voor werkloosheidspercentages en totale beroepsbevolking volgen uit de schattingen voor werkloze en werkzame beroepsbevolking. Deze correctie veronderstelt dat de vertekening over een periode van drie jaar constant blijft.

Naast deze correctie wordt door *Nationale* Rekeningen ook nog een correctie op de werkzame beroepsbevolking uitgevoerd. Beide correcties worden in de gewichten verwerkt door de steekproef nogmaals te herwegen naar een tabel met ramingen voor de gecorrigeerde werkloze en werkzame beroepsbevolking.

Bijlage B: Oorzaken voor rotation group bias

Zoals aangegeven in paragraaf 2, treedt er vertekening (rotation group bias) op in de uitkomsten van de verschillende herbenaderingen van het panel. Deze vertekening wordt veroorzaakt door:

1. Selectieve nonrespons tussen de opeenvolgende peilingen door paneluitval.
2. Systematische verschillen tussen de populaties die worden bereikt met de CAPI-mode in de eerste peiling en de CATI-mode in de vervolgpeilingen. Verwacht wordt dat effect gering is omdat tijdens het eerste interview gevraagd wordt naar het

telefoonnummer van het huishouden, zodat geheime nummers ook bij het CBS bekend zijn.

3. Mode-effecten. Dit zijn systematische verschillen in de antwoorden van een respondent doordat de vragenlijst in de vervolgpeilingen CATI in plaats van CAPI wordt afgenomen. Een belangrijk element van het mode-effect is dat bij CAPI de interviewer en de respondent daadwerkelijk met elkaar in contact komen terwijl dit bij CATI niet het geval is. Hierdoor kan een verschil ontstaan in het privacygevoel dat de respondent heeft bij het beantwoorden van de vragen waardoor de uitkomsten beïnvloed kunnen worden. Een ander belangrijk aspect van het mode-effect is dat de snelheid bij telefonische interviews hoger ligt dan bij face-to-face interviews. Hierdoor heeft de respondent bij CATI minder denk- en herinneringstijd waardoor eerder meetfouten worden gemaakt.
4. Toename van het aandeel respondenten dat proxi respondeert onder de CATI mode. Een mogelijke oorzaak is dat via proxi waarneming meer systematische meetfouten worden gemaakt bij het vaststellen of een persoon werkzaam of werkloos is volgens de definities die het CBS hanteert.
5. Vragenlijsteffecten. In de eerste peiling wordt een uitgebreide vragenlijst gebruikt terwijl bij de telefonische herbenaderingen een verkorte vragenlijst wordt gebruikt die uitsluitend gericht is op het vaststellen van veranderingen in de arbeidspositie van de respondent. Verschillen tussen de vragenlijsten kunnen resulteren in systematische effecten op de antwoordpatronen.
6. Paneleffecten. Dit zijn systematische veranderingen in het gedrag van de respondenten in het panel. Het stellen van vragen aan respondenten zonder baan over hun activiteiten om een baan te vinden zou kunnen leiden tot een toename van de zoekactiviteiten van deze respondenten. Hierdoor gaat de arbeidspositie van de respondenten in de vervolgpeilingen afwijken van de rest van de bevolking. Een tweede mogelijkheid is dat respondenten hun antwoordpatroon aanpassen omdat ze leren wat de snelste routing door de vragenlijst is.

Bijlage C: Weegmodel maandcijfers

In deze bijlage wordt beschreven hoe voor iedere afzonderlijke peiling een reeks van gegeneraliseerde regressieschattingen voor een doelvariabele op maandbasis kan worden geconstrueerd. Omdat voor een weging per maand per peiling minder data beschikbaar zijn dan bij de weging voor het voortschrijdende driemaandsgemiddelde wordt een sterk vereenvoudigde versie van het weegmodel uit Bijlage A toegepast:

Afkomst3 + Geslacht2 + Leeftijd21 + TypeHuishouden3 +
Regio44 + Leeftijdgeslacht7 + CWIinschrijvingsduur5 + Looncategorie3.

In elke weegterm geeft het getal aan in hoeveel klassen een bepaalde categorie verdeeld is.

Ook in deze wegingen wordt de methode van Lemaître and Dufour (1987) voor het consistent wegen van personen en huishoudens toegepast, waardoor alle personen binnen één huishouden hetzelfde gewicht krijgen. Het algoritme van Huang and Fuller (1978) wordt toegepast om negatieve gewichten te voorkomen.

Bijlage D: Specificatie tijdreeksmodel

In deze bijlage wordt een technische specificatie gegeven van het tijdreeksmodel beschreven in subparagraaf 3.1. Laat θ_t de waarde voor de onbekende populatieparameter zijn in maand *t*. Verder is Y_t^{t-j} de gegeneraliseerde regressieschatting voor deze populatieparameter voor maand *t* op basis van het panel dat op tijdstip *t-j* het panel instroomde. Omdat een steekproef vijf keer wordt herbenaderd met een interval van drie maanden wordt iedere maand een vector van vijf gegeneraliseerde regressieschattingen waargenomen: $\mathbf{Y}_t = (Y_t^t \; Y_t^{t-3} \; Y_t^{t-6} \; Y_t^{t-9} \; Y_t^{t-12})^T$. Deze vector wordt gemodelleerd met:

$$\mathbf{Y}_t = \mathbf{1}_5 \theta_t + \boldsymbol{\lambda}_t + \mathbf{e}_t . \qquad (1)$$

Hierbij is $\mathbf{1}_5$ een vijfdimensionale vector met ieder element gelijk aan 1, $\boldsymbol{\lambda}_t = (\lambda_t^0 \; \lambda_t^3 \; \lambda_t^6 \; \lambda_t^9 \; \lambda_t^{12})^T$ een vector met tijdsafhankelijke componenten die rekening houden met de vertekening in de trend, en $\mathbf{e}_t = (e_t^t \; e_t^{t-3} \; e_t^{t-6} \; e_t^{t-9} \; e_t^{t-12})^T$ een vector met de steekproeffouten van de gegeneraliseerde regressieschattingen van de vijf peilingen.

D.1 Tijdreeksmodel voor de onbekende populatieparameter

Het structurele tijdreeksmodel voor de onbekende populatieparameter in (1) wordt gegeven door:

$$\theta_t = L_t + S_t + \varepsilon_t , \qquad (2)$$

met L_t een stochastisch trendmodel, S_t een stochastisch seizoensmodel, en ε_t de onverklaarde variatie in de reeks van θ_t. De trend wordt gemodelleerd met het zogenaamde smooth-trend model:

$$\begin{aligned} L_t &= L_{t-1} + R_{t-1}, \\ R_t &= R_{t-1} + \eta_{R,t}, \qquad (3) \\ E(\eta_{R,t}) = 0, \quad & Cov(\eta_{R,t}, \eta_{R,t'}) = \begin{cases} \sigma_R^2 & \text{als} \quad t = t' \\ 0 & \text{als} \quad t \neq t'. \end{cases} \end{aligned}$$

De parameters L_t en R_t worden de trend en de hellingsparameter genoemd. Het seizoenspatroon wordt gemodelleerd met het trigoniometrische model

$$S_t = \sum_{l=1}^{6} S_{l,t} , \qquad (4)$$

waarbij

$$
\begin{aligned}
&S_{l,t} = S_{l,t-1}\cos(h_l) + S^*_{l,t-1}\sin(h_l) + \omega_{l,t} \\
&S^*_{l,t} = S^*_{l,t-1}\cos(h_l) - S_{l,t-1}\sin(h_l) + \omega^*_{l,t}, \quad l = 1,...,6, \\
&h_l = \frac{\pi l}{6}, \quad l = 1,...,6, \\
&E(\omega_{l,t}) = E(\omega^*_{l,t}) = 0, \\
&Cov(\omega_{l,t}, \omega_{l',t'}) = Cov(\omega^*_{l,t}, \omega^*_{l',t'}) = \begin{cases} \sigma^2_\omega & \text{als} \quad l = l' \quad \text{en} \quad t = t' \\ 0 & \text{als} \quad l \neq l' \quad \text{of} \quad t \neq t' \end{cases}, \\
&Cov(\omega_{l,t}, \omega^*_{l,t}) = 0 \text{ voor alle } l \text{ en } t.
\end{aligned}
\tag{5}
$$

De onverklaarde variatie ε_t wordt gemodelleerd als witte ruis:

$$
E(\varepsilon_t) = 0, \quad Cov(\varepsilon_t, \varepsilon_{t'}) = \begin{cases} \sigma^2_\varepsilon & \text{als} \quad t = t' \\ 0 & \text{als} \quad t \neq t'. \end{cases}
\tag{6}
$$

D.2 Tijdreeksmodel voor de rotation group bias

Met de vector $\boldsymbol{\lambda}_t = (\lambda^0_t \ \lambda^3_t \ \lambda^6_t \ \lambda^9_t \ \lambda^{12}_t)^T$ wordt de rotation group bias ten gevolge van het roterende panelontwerp gemodelleerd. Zoals beschreven in subparagraaf 3.1.2 wordt verondersteld dat de gegeneraliseerde regressieschattingen uit de eerste peiling niet vertekend is, dat wil zeggen $\lambda^0_t = 0$. De overige componenten meten het systematische verschil tussen de trends in de reeksen van de gegeneraliseerde regressieschattingen uit de tweede, derde, vierde en vijfde peiling ten opzichte van de eerste peiling. In eerste instantie wordt uitgegaan van een random walk model

$$
\begin{aligned}
&\lambda^j_t = \lambda^j_{t-1} + \eta_{\lambda,j,t}, \qquad j = 3, 6, 9, 12, \\
&E(\eta_{\lambda,j,t}) = 0, \quad Cov(\eta_{\lambda,j,t}, \eta_{\lambda,j',t'}) = \begin{cases} \sigma^2_\lambda & \text{als} \quad t = t' \quad \text{en} \quad j = j' \\ 0 & \text{als} \quad t \neq t' \quad \text{of} \quad j \neq j'. \end{cases}
\end{aligned}
\tag{7}
$$

D.3 Tijdreeksmodel voor de steekproeffout

De variantie en autocorrelatiestructuur in de steekproeffouten worden geschat uit de steekproefdata. De varianties van de steekproeffouten worden als prior informatie aan het tijdreeksmodel meegegeven via het volgende algemene model voor steekproeffouten, dat is voorgesteld door Binder and Dick (1990):

$$
e^{t-j}_t = k^{t-j}_t \tilde{e}^{t-j}_t,
\tag{8}
$$

met k^{t-j}_t de standaardfout van de gegeneraliseerde regressieschattingen van de desbetreffende peiling. De autocorrelaties worden geschat via de procedure van Pfeffermann et al. (1998). Zoals aangegeven in subparagraaf 3.1.3 is de steekproeffout in de eerste peiling niet gecorreleerd met steekproeffouten waargenomen in het verleden. De autocorrelatie voor de overige peilingen kan worden beschreven met een AR(1)-model. Hieruit volgt dat voor de eerste peiling $\tilde{e}^t_t$ wordt gemodelleerd als witte ruis met $E(\tilde{e}^t_t) = 0$ en $Var(\tilde{e}^t_t) = 1$. Doordat de variantie van $\tilde{e}^t_t$ gelijk is aan één, is de variantie van de steekproeffout e^t_t gelijk aan de variantie van de gegeneraliseerde regressieschatter; $Var(e^t_t) = (k^t_t)^2$. Voor de overige peilingen geldt het volgende AR(1)-model:

$$\tilde{e}_t^{t-j} = \rho \tilde{e}_{t-3}^{t-j} + \nu_t^{t-j},$$

$$E(\nu_t^{t-j}) = 0, \quad Cov(\nu_t^{t-j}, \nu_{t'}^{t'-j}) = \begin{cases} \sigma_\nu^2 & \text{als} \quad t = t' \\ 0 & \text{als} \quad t \neq t'. \end{cases}$$

Omdat voor $\tilde{e}_t^{t-j}$ een AR(1)-proces wordt aangenomen, geldt dat $Var(\tilde{e}_t^{t-j}) = \sigma_\nu^2 / (1-\rho^2)$. De variantie van de steekproeffout is gelijk aan de variantie van de gegeneraliseerde regressieschatter indien $\sigma_\nu^2 = (1-\rho^2)$. Zoals gezegd wordt de autocorrelatie coëfficiënt ρ geschat uit de steekproefdata via de procedure van Pfeffermann et al. (1998).

Bijlage E: Toestandsruimtemodellen en het Kalmanfilter

Om schattingen te maken voor maandelijkse cijfers over de beroepsbevolking wordt het structurele tijdreeksmodel uit subparagraaf 3.1 geschreven in de zogenaamde toestandsruimtevorm. Vervolgens kan het Kalmanfilter worden gebruikt om dit model te schatten. Zie Harvey (1989) of Durbin en Koopman (2001) voor een introductie in het schatten van toestandsruimtemodellen met behulp van het Kalmanfilter. In Van den Brakel and Krieg (2009a) is een uitdrukking van het tijdreeksmodel in toestandsruimtevorm te vinden.

Via het Kalmanfilter wordt voor iedere maand een optimale schatting gemaakt voor de doelvariabele en de modelparameters op basis van de informatie die beschikbaar is tot en met deze periode. Dit zijn de zogenaamde gefilterde schattingen. Het Kalmanfilter is een recursief algoritme dat start aan het begin van de tijdreeks en eindigt bij de waarneming van de laatste periode. Vervolgens kunnen de gefilterde schattingen worden verbeterd met de informatie die beschikbaar is gekomen na de periode waarop de gefilterde schatting betrekking heeft. Dit is een recursief algoritme dat start bij de laatst waargenomen periode en eindigt bij het begin van de reeks. Dit proces wordt smoothen genoemd. De gefilterde schattingen voor maand *t* zijn de optimale schattingen gebaseerd op de waarnemingen tot en met periode *t*. De gesmoothte schattingen zijn de optimale schattingen voor maand *t*, gebaseerd op alle informatie uit de beschikbare tijdreeks. In dat geval wordt bij de schatting voor de doelvariabele van maand *t* ook gebruik gemaakt van de informatie die is verkregen in de periode na maand *t*. De gesmoothte schattingen zijn gebaseerd op de fixed interval smoother. Zie Harvey (1989) of Durbin en Koopman (2001) voor technische details.

Voor het schatten en publiceren van maandcijfers zijn vooral de gefilterde schattingen relevant. Dit zijn immers de schattingen die zijn gebaseerd op de informatie die op het moment van publiceren beschikbaar is. Het publiceren van gesmoothte schattingen impliceert dat de publicaties regelmatig moeten worden herzien.

Het tijdreeksmodel voor de maandcijfers wordt geanalyseerd met software die ontwikkeld is in OxMetrics en subroutines van Ssfpack 3.0, zie Doornik (2007) en Koopman et al. (2008). Alle toestandsvariabelen van het toestandsruimtemodel zijn niet-stationair, met uitzondering van de steekproeffouten. De niet-stationaire toestandsvariabelen worden geïnstalleerd met een diffuse prior. Dat wil zeggen dat deze toestandsvariabelen aan het begin van de reeks een startwaarde krijgen die gelijk is aan nul met een diagonale covariantiematrix met zeer grote waarden. De steekproeffouten zijn stationair. Daarom worden de startwaarden voor de steekproeffouten gelijk genomen aan nul met een covariantiematrix die is afgeleid uit het

AR(1)-model. Verder wordt gebruik gemaakt van een exacte diffuse loglikelihoodfunctie via de procedure voorgesteld door Koopman (1997).