KM 44 (1993)
pg. 45-57

EVALUATIE VAN MODELSELECTIE MET VOORSPELCRITERIA

Edwin van GAMEREN
Debora E.G. MOOLENAAR
Richard PAAP*

SAMENVATTING

In dit artikel onderzoeken we aan de hand van simulaties de prestaties van een aantal voorspelcriteria op het gebied van modelselectie van univariate autoregressieve (AR) tijdreeksmodellen en modellen met verdeelde vertragingen (DL-modellen). Het blijkt dat modelselectie aan de hand van 1-staps voorspellingen betere resultaten geeft dan wanneer men voorspellingen voor meer perioden tegelijk evalueert. Bovendien blijkt er weinig verschil te bestaan tussen de voorspelcriteria onderling. Wanneer modelselectie het doel is, blijkt dat de keuze beter kan worden gebaseerd op selectiecriteria binnen de steekproef, zoals het Schwarz Criterium, dan op voorspelcriteria toegepast op extra waarnemingen.

---

*De auteurs zijn studenten Econometrie aan de Erasmus Universiteit Rotterdam. Dit artikel is een aangepaste versie van het onderzoeksverslag geschreven in het kader van het vak Toegepaste Econometrie. De auteurs willen alle medestudenten en docenten bedanken voor hun commentaar en in het bijzonder hun begeleider dr. P.H.B.F. Franses. Ook gaat onze dank uit naar een anonieme referee voor zijn/haar waardevolle suggesties.

Correspondentie adressen:

Edwin van Gameren
Vakgroep Macro-Economie
Erasmus Universiteit Rotterdam
Postbus 1738
3000 DR ROTTERDAM

Debora Moolenaar
OCFEB
Erasmus Universiteit Rotterdam
Postbus 1738
3000 DR ROTTERDAM

Richard Paap
Vakgroep Econometrie
Erasmus Universiteit Rotterdam
Postbus 1738
3000 DR ROTTERDAM

1. INLEIDING

In de voorspelliteratuur worden veel criteria gebruikt om voorspellingen te evalueren, zoals kan worden gezien in enkele recente jaargangen van de *Journal of Forecasting*, de *International Journal of Forecasting* en de *Journal of Business & Economic Statistics*. Zelden wordt verklaard waarom een bepaald criterium gebruikt wordt. In dit artikel beschrijven we de resultaten van een onderzoek naar de prestaties van voorspelcriteria op het gebied van selectie van autoregressieve (AR-)modellen en modellen met verdeelde vertragingen (*distributed lag*, DL-modellen). Het doel van het onderzoek is een keuze uit het grote aantal criteria te verantwoorden en meer inzicht te krijgen in de mogelijkheden en beperkingen van modelselectie met voorspelcriteria.

De selectiecriteria die in dit artikel zijn opgenomen, geven een redelijk beeld van de criteria die in recente literatuur vermeld worden. In paragraaf 2 worden de AR-modelselectiecriteria behandeld. Deze zijn niet expliciet ontwikkeld om voorspellingen te evalueren, maar om het juiste model te kiezen op basis van waarnemingen binnen de steekproef. Een beschrijving van de voorspelcriteria, die we zullen onderzoeken, is te vinden in paragraaf 3. In paragraaf 4 worden de prestaties van de verschillende criteria op het gebied van modelselectie van AR-modellen aan de hand van simulaties onderzocht. Paragraaf 5 bespreekt de overeenkomstige simulatieresultaten voor DL-modellen. De belangrijkste conclusies, die we op grond van deze simulaties trekken, worden in paragraaf 6 kort samengevat.

2. AR-MODELSELECTIECRITERIA BINNEN DE STEEKPROEF

In de literatuur bestaan een aantal modelselectiecriteria speciaal bestemd voor het bepalen van de orde van een AR-proces. In deze paragraaf bespreken we de door ons in de simulaties toegepaste AR-modelselectiecriteria. We zullen volstaan met het tonen van de formules. Volledige afleidingen kunnen worden gevonden in de betreffende artikelen.

Een veel gebruikt selectiecriterium is het Informatie Criterium van Akaike (1974):

$$AIC(k) = T \log \hat{\sigma}^2(k) + 2k \tag{2.1}$$

waarbij T staat voor het aantal waarnemingen, k voor de orde van het AR-model en $\hat{\sigma}^2(k)$ voor de maximale aannemelijkheidsschatting van de variantie van de storingen van een geschat AR(k)-model. Deze formule herbergt een balans tussen de mate van fit en het aantal parameters.

Het volgende criterium, afgeleid in Schwarz (1978), straft meer voor het opnemen van extra parameters en staat bekend als het Schwarz Criterium:

$$SC(k) = T \log \hat{\sigma}^2(k) + k \log T \tag{2.2}$$

Hannan & Quinn (1979) stellen een strafterm voor het aantal parameters voor, waarbij de straf voor het opnemen van een extra parameter kleiner is dan bij het Schwarz Criterium maar groter dan bij het Akaike Informatie Criterium:

$$HQ(k) = T \log\hat{\sigma}^2(k) + 2kc\log(\log T) \qquad \text{met } c>1 \tag{2.3}$$

Het is overigens onbekend welke waarde van c het beste is. Aangeraden wordt voor c een waarde te nemen die dicht bij 1 ligt.

Naast bovenstaande AR-modelselectiecriteria hebben we ook gebruikt gemaakt van modelkeuze op grond van de voorspelkwaliteit van AR-modellen van verschillende orde. Akaike (1969) stelt de *Final Prediction Error* voor, een criterium dat gebaseerd is op de asymptotische kwadratische fout van een 1-staps voorspelling gemaakt met een AR(k)-model.

$$FPE(k) = \hat{\sigma}^2(k)\left(\frac{T+k}{T-k}\right) \tag{2.4}$$

Modelselectie aan de hand van de vier besproken criteria vindt in alle gevallen plaats door het minimaliseren van de betreffende formule naar k. De waarde k, waarvoor het criterium de kleinste waarde aanneemt, geeft de orde van het AR-model aan dat de data binnen de steekproef het best beschrijft. Vaak blijken deze specifieke AR-modelselectiecriteria ook goed te presteren bij de modelkeuze van andere dan AR-modellen. Wij hebben daarom besloten deze criteria ook toe te passen op DL-modellen.

In de volgende paragraaf zullen we de door ons onderzochte voorspelcriteria bespreken. Deze criteria worden in tegenstelling tot de in deze paragraaf besproken selectiecriteria toegepast op waarnemingen buiten de steekproef, die gebruikt wordt om het model te schatten. De voorspelcriteria zijn dan ook niet speciaal ontwikkeld voor AR-modellen, maar kunnen ook voor andere modellen gebruikt worden.

3. VOORSPELCRITERIA

In deze paragraaf zullen de voorspelcriteria behandeld worden, waarvan we de prestaties op het gebied van modelselectie van AR- en DL-modellen aan de hand van simulaties willen onderzoeken. De vermelde voorspelcriteria geven een redelijk beeld van de criteria die in de recente literatuur worden gebruikt. Voor een uitgebreider overzicht wordt verwezen naar het oorspronkelijke onderzoeksverslag, dat te verkrijgen is bij de auteurs.

Modelselectie op basis van voorspelcriteria vindt plaats door dat model te kiezen waarvoor het criterium de kleinste waarde aanneemt. Het model dat voorspellingen genereert waarvoor het voorspelcriterium optimaal is, voorspelt volgens het betreffende criterium het beste ten opzichte van de overige modellen. In tabel 3.1 staan de door ons onderzochte criteria. Voor de vermelde criteria geldt dat N het aantal voorspellingen representeert, F_t de voorspelling voor periode t, en A_t de eigenlijke waarde voor periode t.

Het meest eenvoudige voorspelcriterium is natuurlijk de gemiddelde fout (ME). Dit criterium wordt echter weinig gebruikt om verklaarbare redenen. De gemiddelde fout zegt namelijk niets over de grootte van de fouten, aangezien positieve en negatieve fouten tegen elkaar wegvallen. Om dit laatste probleem te omzeilen kan de gemiddelde absolute fout (MAE) gebruikt worden. Beide criteria

Tabel 3.1. Overzicht van voorspelcriteria

NAAM	CRITERIUM		REFERENTIE
Mean Error	$ME = \frac{1}{N}\sum_{t=1}^{N}(A_t - F_t)$	(3.1)	Wheelwright & Makridakis (1985)
Mean Absolute Error	$MAE = \frac{1}{N}\sum_{t=1}^{N}\lvert A_t - F_t\rvert$	(3.2)	Wheelwright & Makridakis (1985)
Mean Squared Error	$MSE = \frac{1}{N}\sum_{t=1}^{N}(A_t - F_t)^2$	(3.3)	Ash, Smith & Heravi (1990)
Mean Squared Percentage Error	$MSPE = \frac{1}{N}\sum_{t=1}^{N}\left[\frac{A_t - F_t}{A_t}\right]^2$	(3.4)	Wolfe & Flores (1990)
Mean Percentage Error	$MPE = \frac{1}{N}\sum_{t=1}^{N}\frac{A_t - F_t}{A_t}$	(3.5)	Shkurti & Winefordner (1989)
Mean Absolute Percentage Error	$MAPE = \frac{1}{N}\sum_{t=1}^{N}\left\lvert\frac{A_t - F_t}{A_t}\right\rvert$	(3.6)	Shkurti & Winefordner (1989)
Adjusted Mean Absolute Percentage Error	$MAPE^* = \frac{1}{N}\sum_{t=1}^{N}\left\lvert\frac{A_t - F_t}{\frac{1}{2}(A_t + F_t)}\right\rvert$	(3.7)	Armstrong (1985)
Theils Ongelijkheids-Coëfficiënt	$U2 = \frac{RMSE}{\sqrt{\frac{1}{N}\sum_{t=1}^{N}A_t^2}}$	(3.8)	Theil (1966)
Accuracy Ratio	$AcRa = \frac{1}{N}\sum_{t=1}^{N}\max\left(\frac{A_t}{F_t}, \frac{F_t}{A_t}\right)$	(3.9)	Armstrong (1985)

zijn echter sterk afhankelijk van hetgeen dat voorspeld wordt en op welke manier. Wel kunnen deze criteria aangewend worden als indicatie van over- of onderschatting.

Twee veelvuldig aangetroffen criteria zijn de gemiddelde kwadratische fout (MSE) en de wortel hieruit (RMSE). Vanzelfsprekend levert een selectie op basis van de MSE en de RMSE dezelfde resultaten op. Door het kwadraat te nemen wordt opnieuw het probleem van tegen elkaar wegvallende fouten vermeden, maar het is nog steeds de fout zelf waar de meeste aandacht naar uit gaat, zodat deze criteria dezelfde nadelen hebben als de MAE.

Een manier om dit op te lossen is gebruik te maken van de gemiddelde kwadratische procentuele fout (MSPE) of de wortel hieruit (RMSPE). Net als hiervoor geeft een selectie op basis van de MSPE en de RMPSE natuurlijk dezelfde resultaten. Het voordeel van deze criteria is dat ze niet gebaseerd zijn op de voorspelfout maar op de voorspelfout ten opzichte van de werkelijke waarneming, met andere woorden de relatieve voorspelfout. Indien men echter kleine waarden wil voorspellen kan deze deling instabiel worden.

Twee andere criteria die ook van relatieve fouten gebruik maken, zijn de gemiddelde procentuele fout (MPE) en de gemiddelde absolute procentuele fout (MAPE). Aan de MPE kleeft hetzelfde bezwaar als aan de ME: fouten kunnen tegen elkaar wegvallen, zodat een vertekend beeld ontstaat van het gemiddelde. Dit wordt natuurlijk vermeden met de MAPE. De MAPE wordt dan ook veelvuldig vermeld in de literatuur.

Zoals gezegd is het nadeel van de vier laatstgenoemde criteria dat zij niet geschikt zijn als de waarnemingen van de te voorspellen reeks zeer kleine waarden aannemen. Armstrong (1985) stelt daarom een aangepaste versie voor van de MAPE voor (MAPE'). Hieraan is echter weer een ander nadeel verbonden. Als de reeks of de voorspellingen ook negatieve waarden aannemen, dan kan het voorkomen dat $A_t \approx -F_t$, zodat er gedeeld moet worden door een klein getal.

Naast de veelgebruikte RMSE komt men in de literatuur ook wel Theils Ongelijkheidscoëfficiënt (U2) tegen. In de praktijk levert selectie op basis van de (R)MSE en de Theils Ongelijkheidscoëfficiënt echter geen verschillende resultaten op. De U2 kan dan ook beschouwd worden als een geschaalde RMSE. Verder vermelden we nog een enigszins ongewoon voorspelcriterium, namelijk de *Accuracy Ratio* (AcRa). Het nadeel van dit criterium is dat zowel de voorspellingen als de ware waarnemingen hetzelfde teken moeten hebben.

4. SIMULATIERESULTATEN VAN AR-MODELLEN

In deze paragraaf zullen de resultaten uiteengezet worden van simulatie-experimenten met AR-modellen die we uitgevoerd hebben om de bruikbaarheid van voorspelcriteria bij het selecteren van een model te onderzoeken.

De opzet van de simulaties is als volgt. Met behulp van een AR(3)-model werden $200+N$ waarnemingen gegenereerd. De keuze is op een AR(3)-model gevallen omdat dit een relatief eenvoudig univariaat model is, dat een cyclus kan bevatten. De eerste honderd waarnemingen dienen als opstartperiode om de cyclus in het AR(3)-model op gang te helpen. De volgende honderd waarnemingen

hebben we gebruikt om verschillende AR-modellen terug te schatten, te weten een AR(1), AR(2), AR(3), AR(4) en een AR(5)-model, zodat we naaste het ware model twee grotere en twee kleinere modellen hebben. Zo zijn we in staat om te onderzoeken of er symmetrie bestaat tussen onder- en overspecificatie. Op basis van de teruggeschatte modellen hebben we een voorspellingspad gecreëerd voor de *N* perioden, op twee verschillende manieren. Allereerst voorspellen we 1 tot en met *N* perioden vooruit, zonder gebruik te maken van de waarnemingen buiten de steekproef. Deze voorspellingen worden aangeduid als h-staps voorspellingen. Daarnaast voorspellen we ook *N* keer één periode vooruit, waarbij we steeds de meest recente waarnemingen (buiten de steekproef) gebruiken. Deze voorspellingen zullen we 1-staps voorspellingen noemen. Dit hebben we 1000 maal herhaald voor de in tabel 4.1 vermelde data genererende processen (DGP's). De vijf DGP's zijn zo gekozen, dat ze een cyclus van 4 à 5 perioden hebben. Verder hebben ze een reële en twee complexe wortels, die binnen de eenheidscirkel liggen. We hebben in elk DGP een constante opgenomen, om mogelijke problemen met de relatieve criteria te voorkomen (zie paragraaf 3). Tot slot krijgt elk DGP in iedere periode een standaardnormaal verdeelde storing, $\varepsilon_t \sim N(0,1)$. Door middel van andere simulaties, die hier niet vermeld staan, is gebleken dat de resultaten in deze paragraaf vrijwel onafhankelijk zijn van de grootte van de variantie van de storingen. De verschillen bleven beperkt tot enkele tiende procenten en kwamen vooral tot uiting bij de relatieve voorspelcriteria. De weergegeven prestaties van ieder criterium zijn een gemiddelde over deze vijf DGP's.

Tabel 4.1. Gebruikte DGP's: $y_t = 10 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-3} + \varepsilon_t$, met $\varepsilon_t \sim N(0,1)$.

β_1	β_2	β_3	reële wortel	complexe wortels	modulus	cyclus
0.900	-0.500	0.300	0.761	0.070 ± 0.624 i	0.630	4.3
2.160	-1.640	0.430	0.671	0.745 ± 0.294 i	0.801	5.3
1.090	-0.585	0.445	0.970	0.060 ± 0.678 i	0.681	4.2
1.300	-1.000	0.500	0.823	0.238 ± 0.742 i	0.779	5.0
1.300	-1.080	0.620	0.875	0.212 ± 0.814 i	0.841	4.8

Voor we ons concentreren op de voorspelcriteria kijken we eerst naar de AR-modelselectiecriteria binnen de steekproef uit paragraaf 2. In tabel 4.2 zien we dat alle criteria in een groot deel van de simulaties het ware AR(3)-model terugvinden, waarbij het Schwarz Criterium het hoogste percentage scoort. Alleen bij het DGP met alle parameters absoluut kleiner dan 1 vindt het criterium van Hannan & Quinn vaker het AR(3)-model. Bij dit DGP is het moeilijk het ware model terug te vinden, het gemiddelde percentage van de simulaties waarin het AR(1)- of AR(2)-model wordt gevonden wordt vooral door dit DGP omhoog getrokken. We zien dat het AR(3)-model vaker teruggevonden wordt naarmate de straf op het opnemen van extra parameter groter wordt: Schwarz scoort beter dan Akaike, en Hannan & Quinn zit tussen beide in. De *Final Prediction Error* geeft bij ieder DGP percentages die overeenkomen met het Akaike Informatie Criterium. Overspecificatie blijkt vaker voor te komen dan onderspecificatie.

Tabel 4.2. AR-modelselectiecriteria binnen de steekproef.

criterium	AR(1)	AR(2)	AR(3)	AR(4)	AR(5)
AIC (2.1)	0.3	1.7	68.3	16.4	13.3
SC (2.2)	2.8	4.7	84.7	5.8	2.0
HQ (2.3)	1.0	3.2	77.4	11.9	6.5
FPE (2.4)	0.3	1.8	68.6	16.2	13.1

[1] Voor de vijf DGP's in tabel 4.1 hebben we onderzocht welk model door een criterium wordt geselecteerd. Het percentage dat een criterium het betreffende AR(k)-model verkiest, wordt aan de hand van 1000 simulaties berekend, waarna over de vijf DGP's wordt gemiddeld.
[2] De constante c in het Hannan & Quinn criterium (2.3) is gelijk aan 1.0001 genomen.

Na de AR-modelselectiecriteria binnen de steekproef zullen de voorspelcriteria uit paragraaf 3 nader onderzocht worden. We zullen eerst kijken naar de resultaten van de h-staps voorspellingen. In tabel 4.3 valt direct op dat het AR(3)-model veel minder vaak wordt teruggevonden dan bij de AR-modelselectiecriteria binnen de steekproef. Het percentage is niet alleen veel lager dan in tabel 4.2, maar het AR(2)-model wordt zelfs vaker aangewezen als beste voorspelmodel dan het AR(3)-model, en bij sommige criteria vinden we ook voor AR(1) en/of AR(5) hogere percentages dan het AR(3)-model krijgt. Het lijkt er op dat naarmate de parameters groter zijn, het AR(3)-model iets vaker wordt teruggevonden, maar de verschillen tussen de DGP's zijn niet erg groot. Het gemiddelde geeft een goed beeld van de resultaten. Verder valt op dat er weinig verschil is tussen de verschillende voorspelcriteria. Alleen de twee criteria waarbij positieve en negatieve voorspelfouten tegen elkaar wegvallen (de ME en de MPE) vallen, zoals verwacht, een beetje uit de toon. Deze doen het nog slechter dan de andere. Een ondergespecificeerd model wordt vaker geselecteerd dan een overgespecificeerd model.

Tabel 4.3. Voorspelcriteria voor tien perioden vooruit voorspellen (h-staps voorspellingen).

criterium	AR(1)	AR(2)	AR(3)	AR(4)	AR(5)
ME	23.4	30.6	17.4	11.2	17.4
MAE	20.1	27.5	20.3	11.8	20.3
MSE, U2	19.6	28.2	20.3	11.8	20.1
MPE	23.3	30.7	17.2	11.4	17.4
MAPE	20.2	27.4	20.3	11.8	20.3
MAPE*	20.1	27.5	20.3	11.9	20.2
MSPE	19.6	28.2	20.5	11.7	20.0
AcRa	20.2	27.4	20.4	11.8	20.2

[1] Voor de vijf DGP's in tabel 4.1 hebben we onderzocht welk model door een voorspelcriterium wordt geselecteerd. Het percentage dat een criterium het betreffende AR(k)-model als beste voorspelmodel verkiest, wordt aan de hand van 1000 simulaties berekend, waarna over de vijf DGP's wordt gemiddeld.
[2] Voor voorspelcriteria zie tabel 3.1.

Tabel 4.4. Voorspelcriteria voor tien perioden vooruit voorspellen (1-staps voorspellingen).

criterium	AR(1)	AR(2)	AR(3)	AR(4)	AR(5)
ME	17.1	19.7	23.6	16.1	23.5
MAE	11.2	17.3	28.7	18.1	24.7
MSE, U2	10.4	16.2	31.5	18.7	23.2
MPE	16.6	19.7	23.8	16.1	23.8
MAPE	11.0	17.5	28.7	18.2	24.6
MAPE*	11.1	17.4	28.8	18.1	24.6
MSPE	10.5	16.2	31.4	18.6	23.3
AcRa	11.1	17.4	28.8	18.1	24.6

[1] Voor de vijf DGP's in tabel 4.1 hebben we onderzocht welk model door een voorspelcriterium wordt geselecteerd. Het percentage dat een criterium het betreffende AR(k)-model als beste voorspelmodel verkiest, wordt aan de hand van 1000 simulaties berekend, waarna over de vijf DGP's wordt gemiddeld.
[2] Voor voorspelcriteria zie tabel 3.1.

Als we 1-staps voorspellingen genereren, dan zijn de resultaten minder dramatisch. In tabel 4.4 zien we dat bij deze wijze van voorspellen het AR(3)-model door alle criteria het meest wordt aangewezen als het beste model. Als we de voorspelcriteria vergelijken met de AR-modelselectiecriteria, dan blijven de resultaten ver achter. Het maakt hierbij nauwelijks uit welk voorspelcriterium gekozen wordt. We zien dat criteria waarbij de (relatieve) fout gekwadrateerd wordt (de MSE en de MSPE) iets vaker het AR(3)-model terugvinden, en dat opnieuw de ME en de MPE het slechtst scoren. Een overgespecificeerd model wordt vaker geselecteerd dan een ondergespecificeerd model.

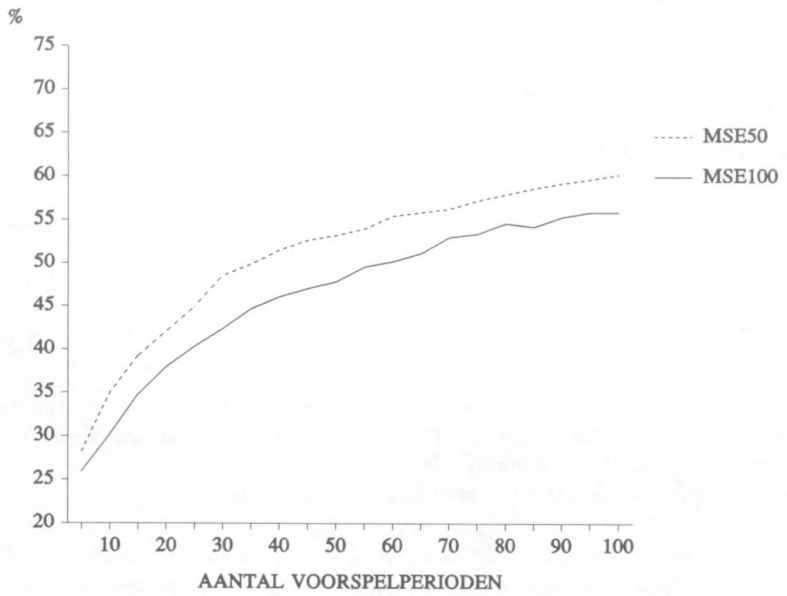
Tabel 4.5. Voorspelcriteria voor honderd perioden vooruit voorspellen (1-staps voorspellingen).

criterium	AR(1)	AR(2)	AR(3)	AR(4)	AR(5)
ME	18.6	5.4	31.8	18.8	25.4
MAE	1.3	2.3	52.9	23.5	20.0
MSE, U2	0.8	1.6	56.0	22.5	19.1
MPE	19.0	5.6	31.5	18.6	25.3
MAPE	1.3	2.3	52.9	23.5	20.0
MAPE*	1.3	2.3	52.7	23.8	19.9
MSPE	0.8	1.6	55.7	22.7	19.2
AcRa	1.3	2.2	52.8	23.8	19.9

[1] Voor de vijf DGP's in tabel 4.1 hebben we onderzocht welk model door een voorspelcriterium wordt geselecteerd. Het percentage dat een criterium het betreffende AR(k)-model als beste voorspelmodel verkiest, wordt aan de hand van 1000 simulaties berekend, waarna over de vijf DGP's wordt gemiddeld.
[2] Voor voorspelcriteria zie tabel 3.1.

Het is niet waarschijnlijk dat modelselectie op basis van het evalueren van slechts tien voorspellingen betere of gelijkwaardige resultaten kan geven dan het evalueren van honderd waarnemingen binnen de steekproef. We hebben daarom de voorspelcriteria ook toegepast op honderd voorspellingen. Het juiste model wordt dan aanzienlijk vaker gekozen dan voorheen. Het blijkt dat de stijging in het percentage dat een AR(3)-model gekozen wordt bij een groter aantal voorspellingen, bijna geheel voortkomt uit een daling van het aantal keren dat een ondergespecificeerd model geselecteerd wordt.

figuur 1.



[1] Weergegeven zijn de gemiddelde percentages van de vijf in tabel 4.1 genoemde DGP's waarmee de MSE het AR(3)-model selecteert, gebaseerd op 1000 simulaties per DGP, bij een oplopend aantal voorspellingen. Bij "MSE100" is het aantal waarnemingen binnen de steekproef gelijk aan 100, terwijl dit bij "MSE50" beperkt is tot 50.

Om een overzicht te krijgen over het effect van de voorspelhorizon op de selectieprestaties van de voorspelcriteria, hebben we prestaties van de MSE op grond van 1-staps voorspellingen in een grafiek geplaatst. Achtereenvolgens werden voorspellingen voor 5, 10, 15, ..., 100 perioden geëvalueerd. In figuur 1 wordt aangegeven in hoeveel procent van de simulaties de MSE het AR(3)-model selecteert. We zien dat dit percentage groter wordt naarmate er meer voorspellingen geëvalueerd worden, maar ook dat we niet snel boven de 50% komen. Uit figuur 1 blijkt dat wanneer men een

grotere steekproef gebruikt om het AR-model te schatten, er ook een groter aantal voorspellingen geëvalueerd moet worden om een gelijke nauwkeurigheid van modelselectie met de MSE te krijgen.

Op grond van de resultaten van de MSE uit tabel 4.5, zou het idee kunnen ontstaan dat een MSE met een strafterm op extra parameters betere resultaten geeft. Wij hebben daarom besloten om de AR-modelselectiecriteria uit paragraaf 2 ook buiten de steekproef toe te passen, hoewel we niet weten of dit theoretisch te onderbouwen is. Wanneer we tien 1-staps voorspellingen evalueren, dan kiezen de modelselectiecriteria wel vaker het juiste model dan de voorspelcriteria, maar een AR(1)-model wordt beduidend vaker verkozen boven het ware AR(3)-model. Als we echter honderd 1-staps voorspellingen evalueren dan blijkt dat de AR-modelselectiecriteria toegepast buiten de steekproef zelfs beter presteren dan binnen de steekproef (vergelijk tabel 4.6 met tabel 4.2). Het AR(3) model wordt door alle modelselectiecriteria in $\pm 89\%$ van de gevallen geprefereerd, waarbij de overige modellen een percentage variërend van 0.1% tot 7.1% krijgen toegewezen. Opmerkelijk is dat de percentages waarmee de criteria het juiste model selecteren binnen de steekproef verschillend zijn, terwijl deze verschillen buiten de steekproef nagenoeg zijn verdwenen.

Tabel 4.6. Modelselectiecriteria voor AR-model (buiten de steekproef, honderd 1-staps voorspellingen).

criterium	AR(1)	AR(2)	AR(3)	AR(4)	AR(5)
AIC (2.1)	2.3	2.4	88.4	5.6	1.3
SC (2.2)	7.1	3.4	88.2	1.2	0.1
HQ (2.3)	3.8	2.9	89.9	2.9	0.5
FPE (2.4)	2.3	2.4	88.4	5.6	1.3

[1] Voor de vijf DGP's in tabel 4.1 hebben we onderzocht welk model door een criterium wordt geselecteerd. Het percentage dat een criterium het betreffende AR(k)-model verkiest, wordt aan de hand van 1000 simulaties berekend, waarna over de vijf DGP's wordt gemiddeld.

[2] De constante c in het Hannan & Quinn criterium (2.3) is gelijk aan 1.0001 genomen.

Om er achter te komen hoe de resultaten van de AR-modelselectiecriteria binnen en buiten de steekproef met elkaar samenhangen, hebben we onderzocht in hoeveel procent van de simulaties binnen (buiten) de steekproef het juiste model gekozen wordt, gegeven het feit dat buiten (binnen) de steekproef het juiste model gekozen is. Deze conditionele kansen liggen niet meer dan één procent boven de onconditionele kansen uit tabel 4.2 en tabel 4.6. De kans dat zowel binnen als buiten de steekproef het AR(3)-model gekozen wordt is nauwelijks groter dan het produkt van de afzonderlijke kansen. Dit betekent ook dat de kans klein is dat in beide gevallen een fout model wordt geselecteerd (kleiner dan 5%). Er bestaat daarentegen wel een redelijke kans dat in één van de twee selecties een verkeerde modelkeuze gemaakt wordt.

We hebben dus gezien dat de AR-modelselectiecriteria toegepast op waarnemingen binnen de steekproef beter presteren dan alle voorspelcriteria, zelfs wanneer voor de modelselectie door de twee soorten criteria evenveel waarnemingen worden gebruikt. Bovendien presteren de AR-modelselectiecriteria, wanneer ze worden toegepast op waarnemingen buiten de steekproef beter dan binnen de steekproef.

5. SIMULATIERESULTATEN VAN DL-MODELLEN

In deze paragraaf bespreken we de resultaten van overeenkomstige simulatieproeven met DL-modellen. De simulaties hebben dezelfde opzet als in paragraaf 4. De waarnemingen worden nu gegenereerd met een DL(2)-model en een DL(0) tot en met DL(4) worden teruggeschat, waarbij DL(k) aangeeft dat y_t verklaart wordt door $x_t, \ldots, x_{t-k}$. De DL variabele is getrokken uit een standaard normale verdeling. Tabel 5.1 vermeldt de parameters van de door ons onderzochte DGP's.

Tabel 5.1. Gebruikte DGP's: $y_t = 10 + \beta_1 x_t + \beta_2 x_{t-1} + \beta_3 x_{t-2} + \varepsilon_t$, met $\varepsilon_t \sim N(0,1)$ en $x_t \sim N(0,1)$

parameters	DGP 1	DGP 2	DGP 3	DGP 4	DGP 5
β_1	0.700	0.750	0.800	0.850	0.900
β_2	0.490	0.563	0.640	0.723	0.810
β_3	0.343	0.422	0.512	0.614	0.729

Toegepast op de DL-modellen presteren het Schwarz Criterium en het Hannan & Quinn criterium binnen de steekproef beter in vergelijking met de AR-modellen, maar het verschil is niet groter dan vier procent. De andere criteria zitten op een zelfde niveau als bij de AR-modellen. De voorspelcriteria selecteren, bij tien 1-staps voorspellingen, vaker het juiste model dan bij de AR-modellen. Het verschil is echter niet groter dan twee procent. Alleen de ME en de MPE, die al relatief slecht presteerden, doen het niet beter. Als er honderd voorspellingen worden geëvalueerd valt direct op dat de ME en de MPE het bij DL-modellen ongeveer acht procent minder goed presteren dan bij de AR-modellen. Voor de voorspelcriteria, die verhoudingsgewijs goed presteerden, varieert het van iets beter tot iets slechter (vergelijk tabel 4.5 en tabel 5.2).

Tabel 5.2. Voorspelcriteria voor honderd perioden vooruit voorspellen (1-staps voorspellingen).

criterium	DL(0)	DL(1)	DL(2)	DL(3)	DL(4)
ME	25.2	14.2	23.3	16.3	21.0
MAE	0.2	3.8	51.6	23.9	20.5
MSE, U2	0.1	2.6	55.1	23.2	19.0
MPE	21.7	13.0	23.8	18.2	23.3
MAPE	0.2	3.4	52.1	23.4	20.9
MAPE*	0.2	4.1	50.5	24.4	20.8
MSPE	0.2	2.4	51.7	23.2	22.5
AcRa	0.2	3.6	52.0	24.2	20.0

[1] Voor de vijf DGP's in tabel 5.1 hebben we onderzocht welk model door een voorspelcriterium wordt geselecteerd. Het percentage dat een criterium het betreffende DL(k)-model als beste voorspelmodel verkiest, wordt aan de hand van 1000 simulaties berekend, waarna over de vijf DGP's wordt gemiddeld.

[2] Voor voorspelcriteria zie tabel 3.1.

6. CONCLUSIES

In dit artikel hebben we getracht aan de hand van simulaties inzicht te krijgen in welke criteria de beste prestaties leveren bij de selectie van AR- en DL-modellen. Hierbij hebben we ons beperkt tot de meest toegepaste criteria. Ter vergelijking hebben we ook enkele AR-modelselectiecriteria binnen de steekproef in het onderzoek betrokken.

Uit de simulaties blijkt dat het van wezenlijk belang is op welke wijze de voorspellingen gegenereerd worden. Als de voorspellingen gemaakt worden door vanaf de laatste waarneming binnen de steekproef de volgende waarnemingen te voorspellen, dan wordt door alle onderzochte voorspelcriteria niet het ware model als beste voorspelmodel aangewezen; een ondergespecificeerd model wordt in een groter deel van de simulaties geselecteerd. Worden de voorspellingen echter gegenereerd door een aantal keer één periode vooruit te voorspellen, waarbij de voorgaande waarnemingen dus bekend worden verondersteld, dan wordt het ware model vaker geselecteerd dan elk van de alternatieve modellen afzonderlijk. Als men voorspellingen wil gebruiken om een model te selecteren, dan verdient het dus de voorkeur om gebruik te maken van 1-staps voorspellingen. Daarnaast heeft het aantal gemaakte voorspellingen ook een grote invloed op de prestaties van de voorspelcriteria. Naarmate er meer perioden vooruit voorspeld wordt, neemt het percentage dat het juiste model gekozen wordt toe. Ongeacht de voorspelmethode en het aantal voorspellingen zijn de resultaten van de AR-modelselectiecriteria altijd beter dan de voorspelcriteria.

Het gebruik van voorspelcriteria bij het selecteren van een model is geen verstandige keuze. Beter is het om de AR-modelselectiecriteria op waarnemingen binnen de steekproef toe te passen. Hoewel ontwikkeld voor binnen de steekproef, blijken deze criteria, wanneer ze worden toegepast op waarnemingen buiten de steekproef, betere prestaties te leveren. Nader onderzoek zal moeten uitwijzen of er een theoretische onderbouwing mogelijk is.

ontvangen 1 - 10-1992
geaccepteerd 20- 5 -1993

REFERENTIES

Akaike, H., "Fitting Autoregressive Models for Prediction", *Annals of the Institute of Statistical Mathematics*, vol. 21, 1969, p. 243-247.

Akaike, H., "A New Look at the Statistical Model Identification", *IEEE Transactions on Automatic Control*, AC-19, 1974, p. 716-723.

Armstrong, J. S., "Long-Range Forecasting: From Crystal Ball to Computer", 2e ed., 1985, Wiley, New York.

Ash, J. C. K., Smyth, D. J. & Heravi, S. M., "The accuracy of OECD forecasts of the international economy, demand, output and prices", *International Journal of Forecasting*, vol. 6, 1990, p. 379-392.

Hannan, E. J. & Quinn, B. G., "The Determination of the Order of an Autoregression", *Journal of the Royal Statistical Society B*, vol. 41, 1979, p. 190-195.

Schwarz, G., "Estimating the Dimension of a Model", *Annals of Statistics*, vol. 6, 1978, p. 461-464.

Shkurti, W. J. & Winefordner, D., "The politics of state revenue forecasting in Ohio 1980-1987, a case study and research implications", *International Journal of Forecasting*, vol. 5, 1989, p. 360-371.

Theil, H., "Applied economic forecasting", 1966, North-Holland, Amsterdam.

Wheelwright, S. C. & Makridakis, S., "Forecasting Methods for Management", 4e ed., 1985, Wiley, New York.

Wolfe, C. & Flores, B., "Judgemental adjustment of Earnings Forecast", *Journal of Forecasting*, vol. 9, 1990, p. 389-405.