KM 31(1989)
pag 43 - 53

DE IDENTIFIKATIE EN EVALUATIE VAN UNIVARIATE ARIMA-MODELLEN ALS GEFORMALISEERD PROCES

Enkele kanttekeningen bij Van der Hoeven en Hundepool

Hans Anker en Erik Oppenhuis*

Samenvatting

Deze notitie is een reaktie op de bespreking van de automatische komponent van het tijdreeksprogramma AUTOBOX door H. van der Hoeven en A.J. Hundepool in Kwantitatieve Methoden 27. In deze notitie laten wij zien dat de door de auteurs gepresenteerde reeksen met behulp van ARIMA niet op één, maar op verschillende wijzen gemodelleerd kunnen worden. Voorts tonen wij aan dat de keuze van een uiteindelijk model uit een verzameling konkurrerende modellen niet alleen gebaseerd is op statistische kriteria, maar ook op inhoudelijke overwegingen rust. Wij konkluderen dat de automatische optie binnen AUTOBOX hooguit een extra model toe kan voegen aan een door de onderzoeker te evalueren verzameling konkurrerende modellen.

* Universiteit van Amsterdam, Politiek Sociaal-Culturele Faculteit, Vakgroep Methoden en Technieken, Oudezijds Achterburgwal 237, 1012 DL Amsterdam, telefoon: 020 - 5252636.

1 Inleiding[1]

Zo langzamerhand lijkt een trend waarneembaar in de ontwikkeling van software waarmee gekompliceerde statistische technieken zonder tussenkomst van de onderzoeker automatisch kunnen worden toegepast. De nieuwste release van het bekende LISREL-programma bevat een optie waarmee geheel automatisch LISREL-modellen gegenereerd kunnen worden en ook op het gebied van tijdreeksanalyse, in het bijzonder zogenaamde ARIMA-modellen, zijn soortgelijke mogelijkheden beschikbaar gekomen in de vorm van het programma AUTOBOX. De automatische komponent van dit programma werd onlangs door Van der Hoeven en Hundepool in Kwantitatieve Methoden besproken. Een uitstekend initiatief, te meer daar de opvattingen over de noodzaak van dergelijke programmatuur flink uiteenlopen. Voorstanders verwachten dat hierdoor gekompliceerde analysetechnieken toegankelijk worden gemaakt voor statistisch minder onderlegde onderzoekers, terwijl tegenstanders vrezen dat een dergelijke mogelijkheid zijn tol zal eisen in de vorm van kwalitatief minder goed onderzoek. In het specifieke geval van ARIMA is de kernvraag daarbij in hoeverre de identifikatie en evaluatie van univariate modellen als een geformaliseerd proces kan worden opgevat waarbij de interpretatie van de onderzoeker geen rol meer speelt. In hun bespreking van AUTOBOX zijn Van der Hoeven en Hundepool echter ten onrechte aan deze vraag voorbijgegaan. Aan de hand van de door hen gepresenteerde reeksen zullen we laten zien dat zowel statistische als inhoudelijke kennis vereist is om met succes van ARIMA gebruik te kunnen maken.

Bij de bepaling van een univariaat ARIMA-model voor een tijdreeks kan een viertal fasen worden onderscheiden:

1) identifikatie

Met behulp van de (partiële) autokorrelatiefunktie worden

---

1. Wij bedanken Cees van der Eijk voor zijn kommentaar op een eerdere versie van dit artikel.

aanwijzingen verkregen over welke ARIMA-filters (Auto Regressive, Integrated of Moving Average) het beste kunnen worden toegepast.

2) schatting

De parameter(s) van de gespecificeerde filter(s) worden geschat, waarna geïnspekteerd wordt of deze signifikant zijn, eventuele AR-parameters binnen de grenzen van stationariteit vallen en mogelijke MA-parameters binnen de grenzen van invertibiliteit liggen. Indien dit niet het geval is, wordt de identifikatie van het model opnieuw ter hand genomen.

3) diagnose

De residuen van het gepostuleerde model dienen ongekorreleerd te zijn (witte ruis). Ter evaluatie van een tentatief model wordt aan de hand van de (partiële) autokorrelatiefunktie van de residuen vastgesteld of deze nog autokorrelatie bevatten. Indien dit het geval is gaat men weer terug naar fase 1 of 2 en specificeert een alternatief model. Fase 1 tot en met 3 kunnen daarom opgevat worden als een iteratief proces, dat beëindigd wordt zodra een (statistisch) bevredigend model gevonden is.

4) meta-diagnose

Vaak doet zich de situatie voor dat meer dan één model voldoet in termen van aanvaardbaarheid van de schatters en ongekorreleerdheid van de residuen. Dit is niet erg verwonderlijk wanneer men zich realiseert dat hogere orde AR-modellen te schrijven zijn als lagere orde MA en andersom. Bij de keuze van een definitief model uit een verzameling konkurrerende modellen let men vooral op de mate waarin de autokorrelatie gemodelleerd is, de verklaarde variantie en zuinigheid. In het model opgenomen seizoensinvloeden behoren inhoudelijk plausibel te zijn.

Aan de hand van een viertal tijdreeksen illustreren Van der Hoeven en Hundepool de automatische optie van het programma AUTOBOX. Zij slaan daarbij echter de fase van de meta-diagnose volledig over. Bovendien aanvaarden zij zonder

kommentaar de door het programma gegenereerde grootheden, wordt het aspekt van zuinigheid door hun nauwelijks in beschouwing genomen en wordt de (non-)plausibiliteit van door hen gemodelleerde seizoensinvloeden evenmin besproken.

2 Evaluatie van konkurrerende ARIMA-modellen.

De keuze van een definitief model uit een verzameling konkurrerende modellen geschiedt aan de hand van een drietal kriteria: verklarende kracht (R^2), ongekorreleerdheid van de residuen (LBQ) en zuinigheid (aantal geschatte parameters).

De R^2 van een univariaat ARIMA-model geeft aan in welke mate de in het model opgenomen filters de in een reeks aanwezige variantie verklaren. De berekening van de R^2 kan echter, net als bij de klassieke regressie, gemakkelijk geïnfleerd raken door het gemiddelde van de reeks, met als gevolg dat de R^2 iets anders aangeeft dan de intuïtieve notie van verklaarde variantie. De oorzaak hiervan ligt in de berekening van de R^2 in termen van totale (TSS) en residuele (RSS) kwadratensommen:

$$R^2 = 1 - RSS/TSS = 1 - RSS/(n*\bar{y}^2 + n^2*st.err_y^2) \qquad 1.$$

In deze TSS zijn zowel het gemiddelde van de reeks als de variantie verdiskonteerd. In het algemeen is men echter niet in het niveau van een reeks geïnteresseerd maar in het verloop ervan. Het is daarom beter formule 2 te gebruiken waarin het niveau niet als te verklaren variantie wordt beschouwd.[2]

$$R^2 = 1 - RSS/(n^2*st.err_y^2) \qquad 2.$$

Behalve het gemiddelde kan ook eventueel in de reeks aanwezige non-stationariteit tot hoge R^2-en leiden. Indien de

---

2. Dit is de formule zoals die ook in AUTOBOX is opgenomen.

onderzoeker geïnteresseerd is in variantie die niet door non-stationariteit wordt veroorzaakt, is het verstandiger om, in navolging van McCleary & Hay (1980), de variantie van enigerlei verschilreeks als uitgangspunt voor de TSS te nemen. Welke verschilreeks daartoe gebruikt moet worden, is afhankelijk van de aanwezigheid van trend, drift en/of seizoenseffekten binnen een reeks. Dit impliceert dat de TSS op verschillende manieren berekend kan worden. Men kan daarom nooit volstaan met de presentatie van 'de' R^2; op z'n minst zal aangegeven moeten worden ten opzichte van welke TSS deze is gedefinieerd.

In zijn algemeenheid kunnen de TSS op de volgende wijzen worden gedefinieerd:

1- De totale TSS (formule 1). Dit type is in analytisch opzicht niet erg interessant en wordt daarom hier verder buiten beschouwing gelaten.
2- De TSS, berekend door het gemiddelde van de reeks uit de berekening weg te laten (formule 2).
3- De TSS, berekend over een verschilreeks, waardoor een trend of drift niet in de berekening wordt opgenomen.
4- De TSS, berekend over een seizoensverschilreeks, waardoor seizoensinvloeden niet in de berekening worden opgenomen.
5- De TSS, berekend over alle in het model opgenomen verschilreeksen, waardoor alle in de reeks aanwezige non-stationariteit uit de berekening wordt weggelaten.

De R^2 is niet het enige kriterium bij de keuze van het uiteindelijke model. Er dient ook gelet te worden op de mate waarin de in de reeks aanwezige autokorrelatie is gemodelleerd. Hiervoor kan gebruik gemaakt worden van de autokorrelatiefunktie (ACF) van de residuen en de daarop gedefinieerde chi-kwadraat verdeelde manteltoets LBQ. De LBQ geeft aan of, en in welke mate, er sprake is van witte ruis. Het aantal vrijheidsgraden is gelijk aan het aantal 'lags'

waarover de ACF berekend wordt minus het aantal geschatte parameters.[3] In tabel 1 zijn van een aantal konkurrerende modellen voor de reeksen van Van der Hoeven en Hundepool de R^2-en, de LBQ en het aantal geschatte parameters gerapporteerd.[4] Met behulp van deze informatie kunnen de modellen op verantwoorde wijze worden geëvalueerd.

Allereerst blijkt uit tabel 1 dat de R^2-en van Van der Hoeven en Hundepool zo nu en dan iets van die van ons (type 2) verschillen. Deze (minieme) verschillen zijn wellicht het gevolg van verschillen in gebruikte komputerprogrammatuur.[5] Opvallender zijn de zeer uiteenlopende waarden van de R^2-en die op basis van de verschillende typen TSS worden verkregen: voor bijvoorbeeld de reeks Werkloze Bouwarbeiders valt, wanneer rekening wordt gehouden met de eerste orde verschilreeks, de verklaarde variantie terug van 99.6 tot 86.4%. Na modellering van alle in de reeks aanwezige non-stationariteit (in dit geval trend en een seizoenseffekt) wordt slechts 18%

---

3. Opname van een extra schatter in een model impliceert het verlies van 1 vrijheidsgraad. Bij 25 vrijheidsgraden en $\alpha = .05$ leidt een LBQ met een waarde groter dan 37.65 tot een verwerping van de nul-hypothese dat de residuen verdeeld zijn als witte ruis. De opname van een extra parameter (met het daaraan verbonden verlies van een vrijheidsgraad) zal een LBQ op moeten leveren die de nulhypothese met een minstens zo grote waarschijnlijkheid weerlegt als het model zonder de extra parameter.

4. Aangezien n kleiner wordt naarmate er meer parameters in het model opgenomen worden, kunnen de RSS en TSS op verschillende aantallen waarnemingen betrekking hebben. Het leek ons daarom beter om de RSS en TSS door het respektievelijke aantal waarnemingen te delen, hetgeen betekent dat zowel in de teller als in de noemer van Mean Sums of Squares gebruik wordt gemaakt.

5. Van der Hoeven en Hundepool gebruiken het programma AUTOBOX, terwijl wij onze berekeningen met behulp van BMDP2T op de het Cyber-mainframe van de Stichting Academisch Rekencentrum Amsterdam (SARA) hebben uitgevoerd. Beide programma's maken gebruik van het zogenaamde 'Pack-algorithme'. BMDP2T werkt met een woordlengte van 60 bits, terwijl AUTOBOX met een kleinere woordlengte werkt. De berekeningen met AUTOBOX zijn daardoor minder nauwkeurig dan die van BMDP2T.

Tabel 1: R^2-en, LBQ en aantal geschatte parameters van konkurrerende modelspecifikaties

Variabele	Model	R^2-en H/H	type 2	type 3	type 4	type 5	LBQ(25)	Aantal parameters AR	MA	TREND/ GEMIDD.[6]
Werkloze Bouwarbeiders	$(3,1,0)(0,1,1)_{12}$ *	.997	.996	.864	.989	.180	37	3	1	-
	$(3,1,0)(0,1,0)_{12}$ AR(1): t=1.60	-	.996	.858	.988	.147	31	3	-	-
	$(0,1,0)(1,0,0)_2(1,0,0)_3(0,1,0)_{12}$	-	.996	.858	.988	.143	29	2	-	-
Verkeers-ongevallen met letsel	$(1,0,0)(0,1,1)_{12}$	-	.844	-	.411	.411	29	1	1	-
	$(1,0,0)(0,1,1)_{12}$ *	.843	.866	-	.494	.494	22	1	1	1
	$(1,0,0)(1,0,1)_{12}$ * AR(12)=.92	.849	.834	-	-	-	22	2	1	1
	$(1,0,1)(0,1,1)_{12}$	-	.852	-	.442	.442	26	1	2	-
	$(1,0,1)(0,1,1)_{12}$ MA(1): t=1.62	-	.866	-	.495	.495	22	1	2	1
Prijzen van Nieuwbouw-Woningen	$(0,1,0)(0,0,1)_2(1,0,0)_4(0,0,1)_{12}$* MA(12): t=-1.72, trend: t=1.93	.997	.997	.339	-	.339	12	1	2	1
	$(1,1,0)(0,0,1)_4$	-	.997	.184	-	.184	26	1	1	-
	$(0,1,0)(0,0,1)_2(1,0,0)_4$	-	.997	.285	-	.285	20	1	1	-
	$(0,1,0)(0,0,1)_2(1,0,0)_4$	-	.997	.312	-	.312	17	1	1	1
Beleidssepots	$(0,1,1)$ *	.645	.634	.266	-	.266	21	-	1	1

* Door Van der Hoeven en Hundepool gepresenteerde modellen

R^2 H/H = Door Van der Hoeven en Hundepool gerapporteerde R^2-en
R^2 type 2 = 1 - RMS(model)/MTSS(0,0,0)
R^2 type 3 = 1 - RMS(model)/MTSS(0,1,0)
R^2 type 4 = 1 - RMS(model)/MTSS(0,1,0)$_{\text{seizoen}}$
R^2 type 5 = 1 - RMS(model)/MTSS(0,1,0)(0,1,0)$_{\text{seizoen}}$
RMS = Residual Mean Square = Residual Sum of Squares/n
MTSS = Mean Total Sum of Squares
n = Het aantal meetmomenten waarover residuen zijn geschat.

verklaard door de AR- en MA-filters. Voor de andere reeksen is het beeld hetzelfde: modellering van non-stationariteit heeft aanzienlijk lagere verklaarde varianties tot gevolg dan wanneer de TSS berekend worden door alleen het gemiddelde

6. Voor een aantal modellen is een trend of gemiddelde geschat. Het verschil tussen de twee $(1,0,0)(0,1,1)_{12}$-modellen voor de verkeersongevallen reeks wordt veroorzaakt doordat in het ene wel, en in het andere geen trend geschat is.

uit de berekening weg te laten.[7] Voorts blijkt uit tabel 1 dat alle R^2-en waarin nog enigerlei non-stationariteit verrekend is, niet of nauwelijks veranderen bij alternatieve modelspecifikaties. De R^2 van het type 5 laat echter duidelijke verschillen zien en kan daarom het beste worden gebruikt bij de evaluatie van konkurrerende modellen. De door Van der Hoeven en Hundepool gepresenteerde R^2-en (type 2) zijn het minst gevoelig en daarom het minst geschikt om gebruikt te worden bij de keuze van een uiteindelijk model.

Bij de reeks Werkloze Bouwarbeiders wordt de LBQ aanzienlijk verbeterd door de MA(12)-parameter buiten het model te laten, maar zien we tevens een aanzienlijke afname in de R^2 (type 5) van 18% naar 14.7%. Verder kan, bij een min of meer gelijkblijvende R^2, in termen van LBQ een beter model bereikt worden door de niet-signifikante AR(1)-parameter uit het model weg te laten; op inhoudelijke gronden valt echter nauwelijks te rechtvaardigen dat de eerste orde autoregressieve parameter niet, en de tweede en derde orde parameter wèl van het model deel uitmaken. Vandaar dat wij ondanks een iets minder goede LBQ en een niet signifikante parameter de voorkeur geven aan het model $(3,1,0)(0,1,0)_{12}$.
Bij de reeks Verkeersongevallen met letsel wordt zowel voor het $(1,0,0)(0,1,1)_{12}$-model als voor het $(1,0,1)(0,1,1)_{12}$-model een signifikante verbetering in de LBQ bereikt wanneer een trendparameter wordt meegeschat. Ook de R^2-en nemen behoorlijk in omvang toe. Op basis van deze gegevens wordt de voorkeur gegeven aan de modellen met een trendparameter. Er blijven dan drie konkurrerende modellen over, alle met een (lage) LBQ-waarde van 22. Op basis van de door ons berekende

---

7. De tabel laat ook op andere wijze zien dat 'de' R^2 niet bestaat. Voor bijvoorbeeld de twee verkeersletsel-modellen zou men, wanneer geen acht op de aan de R^2 ten grondslag liggende typen TSS wordt geslagen, gemakkelijk kunnen konkluderen dat in het eerste geval slechts 49.4% van de variantie is verklaard en in het tweede geval maar liefst 83.4%.

R^2-en van het type 2 konstateren we dat het $(1,0,0)(1,0,1)_{12}$-model minder goed voldoet dan de andere twee modellen.[8] De resterende twee modellen onderscheiden zich niet naar verklarende kracht (zowel type 2 als 5). Omdat in het $(1,0,0)(0,1,1)_{12}$-model minder parameters geschat zijn dan in het $(1,0,1)(0,1,1)_{12}$-model is het eerste model zuiniger en verdient het daarom de voorkeur. Bovendien is de MA(1)-parameter in het laatste model niet signifikant.

Het door Van der Hoeven en Hundepool gepresenteerde model voor de reeks Prijzen van Nieuwbouwwoningen achten wij in een tweetal opzichten ongelukkig. De MA(12)-parameter en ook de trendparameter zijn niet signifikant bij een betrouwbaarheidsinterval van 95 procent. Bezwaarlijker vinden wij echter de geringe interpreteerbaarheid van dit model: een driejaarlijkse trend in de prijsontwikkeling van nieuwbouwhuizen komt ons niet erg plausibel voor (het gaat hier om kwartaalcijfers). De overige drie modellen hebben alle een jaarlijkse seizoensinvloed gemeen. Het schatten van een trend in het model $(0,1,0)(0,0,1)_2(1,0,0)_4$ leidt niet tot een signifikante verbetering in de LBQ maar wel tot een redelijke verhoging van de R^2, en verdient daarom de voorkeur boven het model zonder trendparameter. De keuze tussen het $(0,1,0)(0,0,1)_2(1,0,0)_4$-model met trendparameter en het model $(1,1,0)(0,0,1)_4$ valt, wanneer uitsluitend gelet wordt op de LBQ en de R^2, in het voordeel van de eerste uit. De vraag is echter of de veronderstelling van een half-jaarlijkse seizoensinvloed theoretisch plausibel is. Deze vraag kan het best beantwoord worden door ter zake deskundigen.
Voor de reeks Beleidssepots vonden we één model dat zowel qua fit als qua interpreteerbaarheid voldoet.

---

8. Bovendien duidt de hoogte van de AR(12)-parameter op de aanwezigheid van een intergretatief proces.

3 Konklusie

In deze notitie hebben we laten zien dat voor een aantal reeksen konkurrerende ARIMA-modellen mogelijk zijn. De bepaling van het uiteindelijke univariate model geschiedt op basis van zowel statistische als inhoudelijke overwegingen. Men gebruikt hierbij de volgende statistische kriteria: het aantal parameters, de LBQ en de R^2 die niet beinvloed wordt door enigerlei vorm van non-stationariteit. Indien de verschillen in R^2 en LBQ in tegengestelde richting wijzen is het niet mogelijk om in statistische termen van een beste model te spreken en zal de onderzoeker aan moeten geven welk kriterium het zwaarst weegt. Een dergelijke beslissing wordt bepaald door het doel waarvoor het univariate model ontwikkeld wordt: voor het maken van voorspellingen heeft een hoge R^2 de voorkeur; voor bijvoorbeeld het analyseren van interventies is een lage LBQ wenselijker. Indien de verschillen in R^2 en LBQ wèl in dezelfde richting wijzen is er in statistische termen sprake van een 'beste' model. Dit model hoeft echter geenszins identiek te zijn aan het uiteindelijk geprefereerde ARIMA-model. Zo kan op grond van de interpreteerbaarheid van het model toch besloten worden om een niet-signifikante parameter in het model op te nemen, ook al gaat dit ten koste van de LBQ. Om dezelfde reden kunnen wèl signifikante parameters, ondanks een lagere R^2, uit het model weggelaten worden.

Wij twijfelen er ernstig aan of de hierboven beschreven afwegingen in een komputeralgorithme kunnen worden vastgelegd. Naar onze mening voegt de automatische optie binnen AUTOBOX hooguit een extra model toe aan een verzameling konkurrerende modellen, waaruit, op grond van zowel inhoudelijke als statistische overwegingen, het uiteindelijke univariate ARIMA-model wordt gekozen. Onze konklusie luidt dan ook dat de identifikatie en evaluatie van univariate ARIMA-modellen niet als een volledig geformaliseerd proces kan worden opgevat.

Literatuurverwijzingen

Beer, J.J.A. de and F.J.R. van de Pol (1988)
Time Series Analysis with Intervention Effects: Method and Application in: Sociometric Research Vol. 2 Data Analysis, edited by W.E Saris and I.N. Gallhofer, McMillan Press, Houndhills/Basingstoke/Hampshire/London, 1988

Hoeven, H. van der en A.J. Hundepool (1988)
Autobox, een automatisch Box-Jenkins Tijdreeksanalyse Pakket voor de Micro, Kwantitatieve Methoden 27, 1988, pag.171-194

McCleary, R. and R.A. Hay Jr. (1980)
Applied time series for the social sciences
Sage, Beverly Hills 1980

Ontvangen: 03-08-1988
Geaccepteerd: 10-03-1989