

PROMEMORIOR FRÅN P / STM

NR 3

EN PRESENTATION AV BOX - JENKINS METOD
FÖR ANALYS OCH PROGNOSE AV TIDSSERIER
AV ÅKE HOLMÉN

INLEDNING

TILL

Promemorior från P/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1978-1986. – Nr 1-24.

Efterföljare:

Promemorior från U/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1986. – Nr 25-28.

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

Promemorior från P/STM 1979:3. En presentation av Box-Jenkins metod för analys och prognoser av tidsserier / Åke Holmén.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

PROMEMORIOR FRÅN P / STM

NR 3

EN PRESENTATION AV BOX - JENKINS METOD
FÖR ANALYS OCH PROGNOSE AV TIDSSERIER
AV ÅKE HOLMÉN

INNEHÅLL

0	Sammanfattning
1	Inledning
2	Box-Jenkins ansats för tidsserieanalys; filosofi och struktur
2.1	Allmänt
2.2	Box-Jenkins ansatsens struktur och innehåll
3	Några basbegrepp
3.1	Definitioner
3.1.1	AR och MA-modeller
3.1.2	Backward skiftoperator
3.1.3	Stationaritet
3.2	ARMA-modeller
3.3	ARIMA-modeller
4	Univariata modeller
4.1	Inledning
4.2	Identifiering av d och D
4.3	" - p och q
4.4	" - P och Q
4.5	Test av den identifierade modellen
5	Transferfunktioner
5.1	Allmänt
5.2	Beskrivning av transferfunktioner
6	Intervention theory
6.1	Allmänt
6.2	Kortfattad beskrivning av metodiken
Bilaga 1	Teoretiska utseendet på auto- och partiella autokorrelogramen för givna parameterspecifika- tioner
Bilaga 2	Ett exempel Identifiering av en Box-Jenkins modell för industriproduktionsindex

0 Sammanfattning

Box-Jenkins analys är benämning på en speciell analysstrategi vars huvudsyfte är att ge korta prognoser av tidsserier. Metodiken kan även användas för deskription och analys av tidsserier. Man kan exempelvis testa hypoteser rörande trend-förskjutningar som följd av externa händelser (strejker, lag-ändringar, energikriser m m).

Box-Jenkins metoden kräver mycket beräkningar. I praktiken måste man därför ha tillgång till datorer och speciella datorprogram för att kunna genomföra analysen. Analysen kan utföras på 4 nivåer, varav de två "lägsta" behandlas i denna PM, nämligen

i Univariata modeller; där analysen enbart baserar sig på en given series egna historiska förlopp

ii Transferfunktions modeller; där en modell byggs upp av förklarande serier och det unika beteendet i den beroende serien

På de högre nivåerna byggs multivariata modeller upp med ömsesidiga beroenden. Vidare presenteras något av ideerna bakom Box-Jenkins ansatsen.

En Box-Jenkins modell består av autoregressiva (AR) termer och moving-average (MA) termer. Dessa begrepp definieras tillsammans med andra viktiga begrepp. Den mest generella univariata Box-Jenkins modellen är den s k ARIMA (p,d,q) (P,D,Q)-modellen. För att bygga upp denna modell krävs varken att den analyserade serien är stationär eller säsongoberoende. Dessa problem löses inom ramen för modellstrukturen.

Problemet att identifiera antalet parametrar i de univariata samt transferfunktionsmodellerna beskrivs i grova drag.

Ett exempel på en univariat Box-Jenkins analys på SCB-data presenteras och huvuddragen i identifieringsprocessen kan där följas.

1 Inledning

Framställningen syftar till att presentera Box-Jenkins metod för prognoser på ett så enkelt sätt som möjligt, varför den teoretiska framställningen är extremt förenklad. För definitioner, bevis samt hela det logiska sammanhanget hänvisas till Box-Jenkins lärobok; Time Series Analysis: Forecasting and Control, Holden Day 1976.

SCB har införskaffat Jenkins' program för Box-Jenkins analys. Om intresse finns kan STM anordna kurser som behandlar både teori och praktiskt handhavande. Kurserna syftar till att ge tillräcklig förståelse och kunskaper så att deltagarna själva skall kunna göra Box-Jenkinsanalyser på egna serier.

2 Box-Jenkins ansats för tidsserieanalys; filosofi och struktur

2.1 Allmänt

Box-Jenkins ansats kräver ej, som många andra metoder, oberoende observationer. Tvärt om utnyttjar metoden ett givet beroende för prognoserna. Man väljer en modell från en mycket vid klass av matematiska uttryck. Denna modell uttrycker relationerna mellan en viss observation och de tidigare observationerna.

Sedan skattas modellens parametrar, givet det antagna sambandet. Det huvudsakliga hjälpmedlet som används är autokorrelationer mellan observationer på olika tidsavstånd. Man får prognoser genom att använda den skattade modellen till framskrivningar av det samband som modellen beskriver. Alla prognoser presenteras med ett konfidensintervall av valfri nivå.

Box-Jenkins ansatsen kan ses som en generell metodik för tidsserieanalys. Andra tidsserieanalytiska tekniker blir då specialfall, om dessa metoders speciella förutsättningar är uppfyllda. Detta visas t ex för linjär regression. Då Box-Jenkins använder sig av ett komprimerat skrivsätt för sina modeller, kan det nedan presenterade beteckningssystemet även användas för andra analysmetoder, även om dessa ej är härledda ur Box-Jenkins ansatsen.

Som många andra statistiska metoder av senare datum kräver Box-Jenkins metoden att man har tillgång till stora datorer, samt specialskrivna program för att kunna genomföra analysen. Box-Jenkins använder sig av många iterativa beräkningsprocesser vid estimering av parametrarna i modellen. Modellen byggs stegvis upp med hjälp av autokorrelationer och residualserier. Man använder sig också av flera beslutsalgoritmer vilka leder fram till den "bästa" modellen. Det finns inga formella begränsningar i valet av antalet parametrar i modellen, men metodiken innehåller en bakomliggande filosofi som kan sammanfattas i två satser

i sparsamhet¹ med parametrar (för att den uppbyggda modellen skall kunna tolkas)

ii samarbete med användarna av resultatet (för att rimligheten av den framtagna modellen skall kunna avgöras)

Vikten av att följa dessa grundregler förstås lättare om man jämför hur en Box-Jenkins analytiker arbetar med det traditionella angreppssättet som t ex en ekonometriker använder sig av. Ekonometriken skapar en modell utifrån ekonomisk teori och egen erfarenhet. Sedan används data för att estimerar parametrarna i den givna modellen.

¹Parsimony, på engelska

Detta arbetssätt kan beskrivas i följande flödesschema.

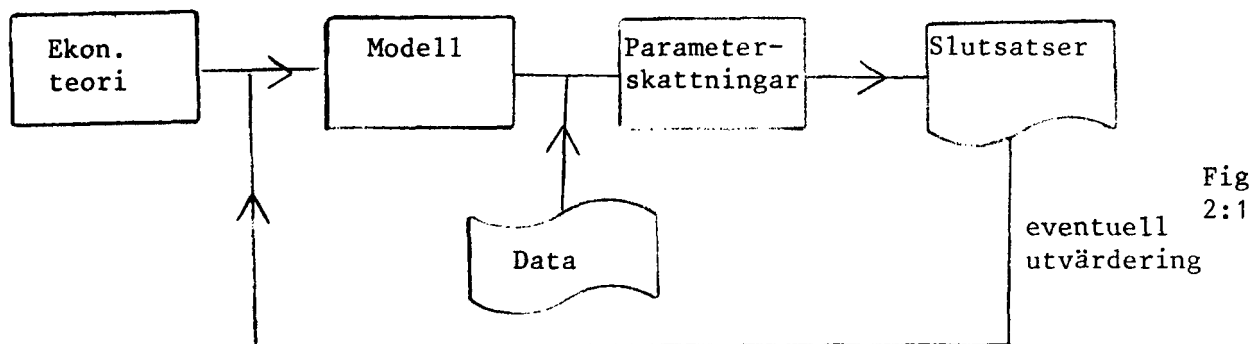


Fig
2:1

En Box-Jenkins analytiker angriper problemet enligt följande

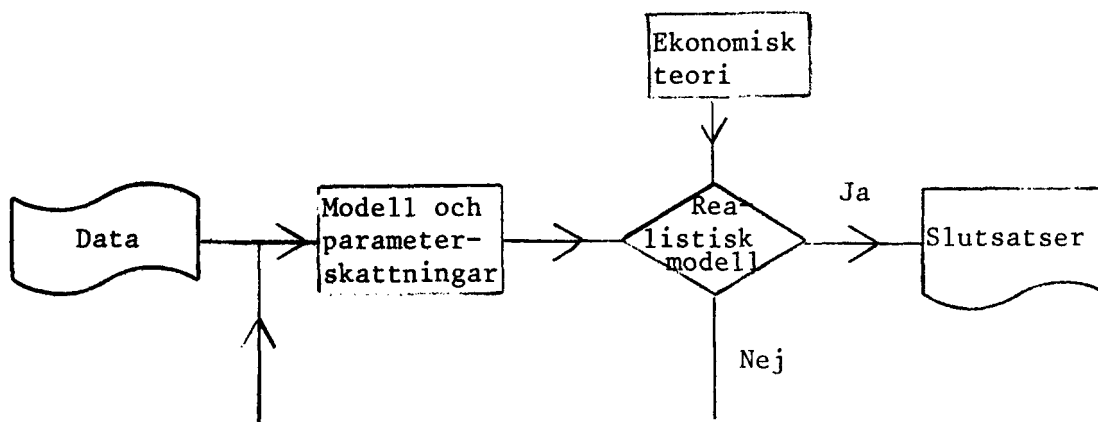


Fig
2:2

Problemet med detta senare angreppssätt är att avgöra om modellen är realistisk inte bara ur matematisk synvinkel utan även ur innehållsmässig, i detta fall ekonomisk - teoretisk. För att man skall kunna avgöra detta, måste således en modell med så få parametrar som möjligt hittas, så att modellens beteende verbalt kan förklaras för en användare/ekonom. Denne får sedan avgöra om modellen har ett realistiskt beteende eller inte. Om modellen ej överensstämmer med ekonomisk teori får man välja en annan modell. Det är möjligt att hitta andra matematiska modeller som kan anpassas till samma data-material, utan att den restserie som uppstår får nämnvärt större varians.

2.2 Box-Jenkins ansatsens struktur och innehåll

Tillvägagångssättet för en Box-Jenkinsanalys kan sammanfattas i nedanstående flödesschema.

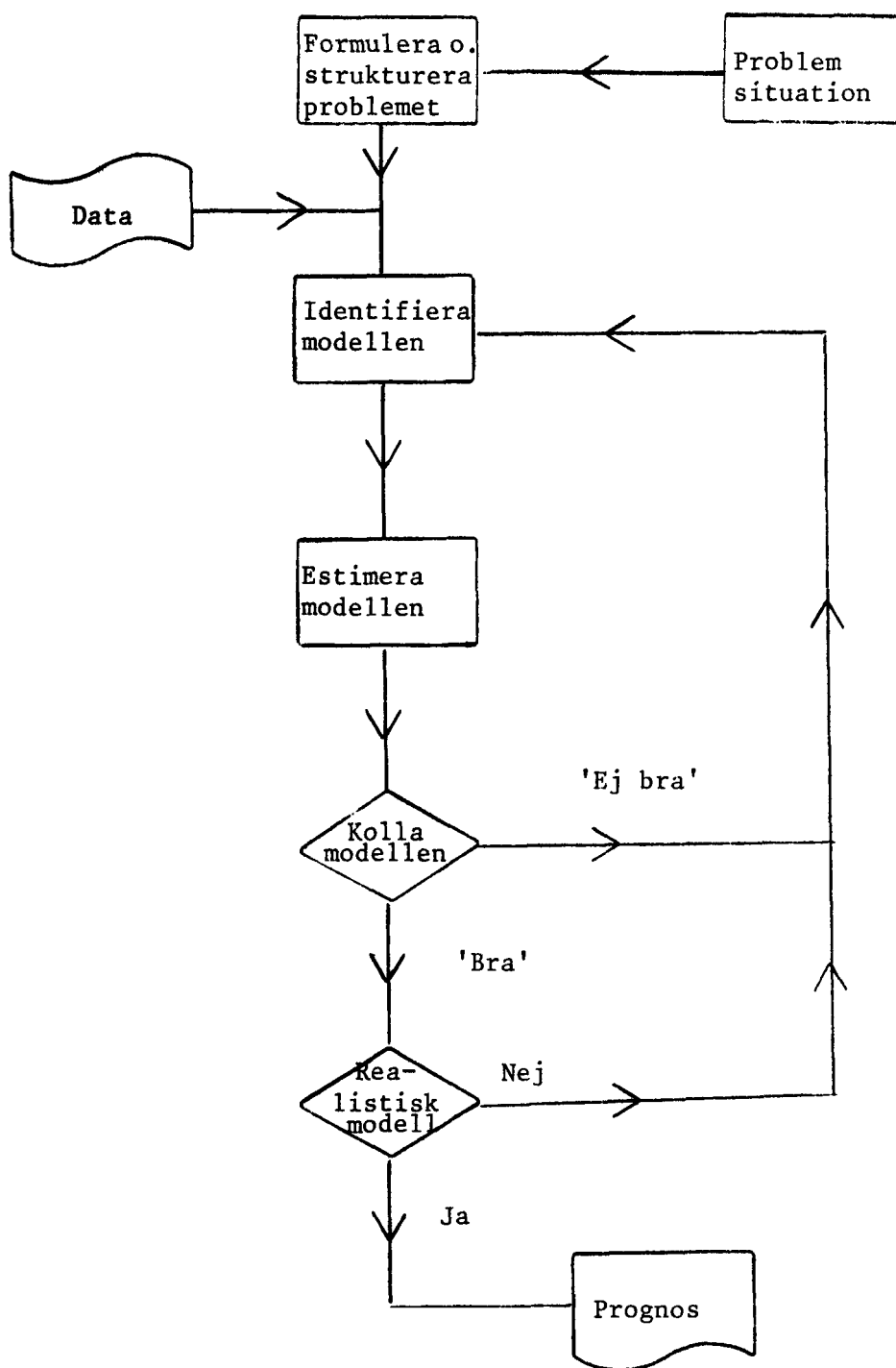


Fig
2:3

Det moment som kräver störst arbetsinsats är identifikationen. Modellen behöver ej identifieras varje gång en prognos skall göras. Någon gång, kanske varje år, bör parameterestimationen kontrolleras.

Metodiken har 4 nivåer.

a) Univariata modeller

På denna nivå har man bara hänsyn till en series egna historiska beteende.

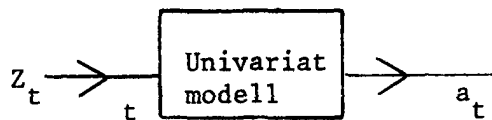


Fig
2:4

Z_t är den observerade tidsserien.

När modellen är tillräckligt bra identifierad och estimerad skall residualserien (a_t) vara ett s k vitt brus (dvs oberoende med fördelningen $N(0, \sigma)$). Den univariata modellen används sedan till att göra prognoser för Z_{t+1} , Z_{t+2} , osv

b) Transfer funktionsmodeller

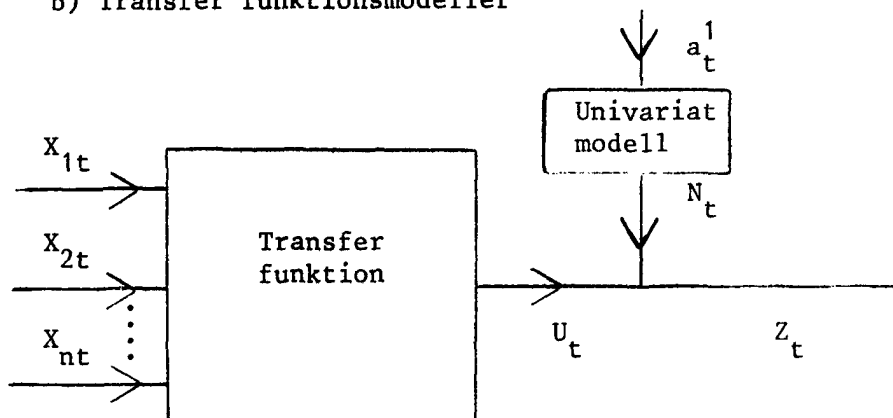


Fig
2:5

Den observerade serien Z_t förklaras dels av de andra seriernas ($X_1, X_2, \dots, X_n$) aggregerade påverkan (U_t) och dels av Z_t -seriens unika historiska beteende (N_t). Vidare finns en slumpkomponent a_t^1

eller

$$Z_t = N_t + U_t + a_t^1$$

På denna nivå tillåts bara kausalt beroende i en riktning.

Inget ömsesidigt kausalt beroende mellan Z_t och X_t antas möjligt
dvs

$$X_t \longrightarrow Z_t \text{ men } Z_t \nrightarrow X_t$$

c) Multivariata stokastiska modeller

I princip som b) men här tillåts feedback

dvs $Z_t \longrightarrow X_{it}$ för åtminstone ett i .

d) Multiple input/Multiple output modeller

Som c) men output består av flera Z-variabler.

Nivåerna c) och d) kommer ej att behandlas vidare i denna PM.

3 Några basbegrepp

3.1 Definitioner

3.1.1 AR och MA modeller

Box-Jenkins modellerna är uppbyggda av

i: autoregressiva modeller (AR)

ii: moving-average modeller (MA)

i: AR modeller

En autoregressiv modell av ordningen 1 definieras

$$Z_t = \phi_1 Z_{t-1} + a_t \quad (3:1)$$

där

a_t är en slumpmässig, oberoende serie s k white noise

Z_t är det observerade värdet på tidsserien vid tidpunkt t

ϕ är en parameter koefficient

En generell autoregressiv modell AR(p) har följande utseende

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \dots + \phi_p Z_{t-p} + a_t \quad (3:2)$$

där p anger antalet perioder bakåt i tiden vi har autoregressivt beroende.

ii: MA modeller

En moving-average modell av ordningen 1 definieras

$$Z_t = a_t - \theta_1 a_{t-1} \quad (3:3)$$

och den generella moving-average modellen MA(q) får följande utseende

$$Z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \dots - \theta_q a_{t-q} \quad (3:4)$$

där q anger antalet perioder bakåt i tiden vi har MA-beroenden.

3.1.2 Backward shiftoperator

För att komprimera modellskrivandet införs en linjär operator B, vilken kallas backwardshiftoperator, definierad som

$$BZ_t = Z_{t-1} \text{ och } B^j Z_t = Z_{t-j}$$

Med hjälp av B kan 3:2 skrivas

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \dots + \phi_p Z_{t-p} + a_t$$

$$Z_t - \phi_1 Z_{t-1} - \phi_2 Z_{t-2} - \dots - \phi_p Z_{t-p} = a_t$$

$$Z_t - \phi_1 BZ_t - \phi_2 B^2 Z_t - \dots - \phi_p B^p Z_t = a_t$$

$$Z_t (1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p) = a_t$$

$$Z_t \phi(B) = a_t \quad (3:5)$$

Motsvarande för 3:4 blir

$$Z_t = \theta(B) a_t \quad (3:6)$$

Det kan visas att varje MA-modell kan skrivas som en AR-modell och tvärt om.

Ex 3:1 Med hjälp av backward shiftoperatorn kan en AR (1) modell skrivas

$$Z_t (1-B\phi) = a_t$$

eller

$$Z_t = a_t / (1-B\phi)$$

Detta kan med känd teknik utvecklas till

$$Z_t = a_t (1 + \phi B + \phi^2 B^2 + \phi^3 B^3 + \dots)$$

Vi får således en oändlig MA-modell där

$$\begin{aligned}\theta_1 &= -\phi \\ \theta_2 &= -\phi^2 \\ &\text{osv}\end{aligned}$$

Det är denna möjlighet som Box-Jenkins metoden använder för att uppfylla sparsamhetskravet avseende antalet parametrar.

Ex 3:2 Säg att vi har identifierat följande MA(3) modell

$$Z_t = a_t (1 + 0.8B + 0.6B^2 + 0.5B^3)$$

Om vi väljer $\phi = 0.8$ i ovanstående AR(1) modellen i exempel 3:1

$$Z_t (1 - 0.8B) = a_t$$

$$Z_t = a_t / (1 - 0.8B) = a_t (1 + 0.8B + 0.64B^2 + 0.51B^3 + \dots)$$

Vi ser att AR (1) modellen med parametern 0.8 beskriver ungefär samma matematiska samband som MA(3) modellen, dock kommer AR(1) modellen ha en större resvarians. Fördelen är dock att bara en parameter behöver estimeras i stället för tre som i MA (3) modellen. Vidare är AR(1)modellen enklare att förstå.

3.1.3 Stationaritet

För att en process skall vara stationär krävs

- i: processen har ett jämviktsläge
- ii: processens egenskaper påverkas ej av en förskjutning av tidsaxeln dvs alla moment förblir lika stora som de var innan tidsaxeln försköts.

Med en svagt stationär process av ordningen K menas att

- i: processen har ett jämviktsläge
- ii: processens K första moment uppfyller kravet att vara oförändrade vid en förskjutning av tidsaxeln.

K brukar väljas till 2 dvs variansen och ändras ej.
Således gäller att en säsongsb beroende tidsserie ej är stationär.

3.2 ARMA-modeller

För att kunna utnyttja valet mellan antalet AR och MA termer till fullo användes modeller där både AR och MA termer tillåts, sk ARMA-modeller.

$$Z_t \phi(B) = \theta(B) a_t \quad (3:7)$$

3:7 är en ARMA-modell av ordningen (p,q) dvs den innehåller p st AR-termer och q st MA termer. Fullt utskrivet blir 3:7 lika med

$$Z_t - \phi_1 Z_{t-1} - \dots - \phi_p Z_{t-p} = a_t - \theta_1 a_{t-1} - \dots - \theta_q a_{t-q} \quad (3:8)$$

Således gäller att en AR(1) modell är det samma som en ARMA (1,0)-modell

Exempel 3:3. Ett specialfall på en ARMA (1,0)-modell fås om $\phi = 1$ dvs

$$Z_t - Z_{t-1} = a_t$$

om $P(a_t = 1) = p$ $P(a_t = -1) = q$ och $P(a_t = 0) = 1-p-q$ fås definitionen för simple random walk.

3.3 ARIMA-modeller

Programsystemet kräver stationaritet för att p och q skall kunna identifieras, men då stationaritet sällan är för handen införs ARIMA-modellen; (ARIMA; autoregressiv integrated moving average).

Ett av kraven för stationaritet var att variansen skulle vara oförändrad vid en förskjutning av tidsaxeln. För att klara av detta krav transformeras de observerade värdena vid behov.

$Z_t^\lambda = Z_t^1$ där Z_t är de observerade värdena

$\lambda = 1$ motsvarar : ingen transformation

$\lambda = 0$ motsvarar: logaritmering

$\lambda = \frac{1}{2}$ motsvarar : rotutdragning

I princip kan λ väljas godtyckligt inom intervallet $[-1,1]$

Den del av variationen som kommer från en säsongcykel löses ej av en transformering, detta problem återkommer vi till.

Det andra kravet på stationaritet, nämligen att serien skall ha ett jämviktsläge, löses genom att serien differensieras en eller flera gånger

$\nabla Z_t = Z_t - Z_{t-1}$ differensiering av första graden

Det gäller således att $\nabla = (1-B)$

$\nabla^2 = (1-B)^2 = \nabla^2 Z_t = Z_t - 2Z_{t-1} + Z_{t-2}$ 2:a graden

En ARIMA-modell skrivs

$$\phi(B)\nabla^d Z_t = \theta(B)a_t \quad (3:9)$$

3:9 är en ARIMA-modell av ordningen (p,d,q) där

p är antalet AR-termer

d " - nödvändiga differensieringar för stationaritet

q " - MA-termer

Z_t är de transformerade värdena.

I exempel 3:3 hade vi en ARMA(1,0)-modell med parametervärdet $\phi=1$. Denna modell kan också ses som en ARIMA(0,1,0)-modell.

För att få fullständig generalitet måste även säsongen införas i ARIMA-modellen. Den kan då skrivas

$$\phi(B) \phi^S(B) \nabla^D \nabla^d Z_t = \theta(B) \theta^S(B) a_t \dots \quad (3:10)$$

Ovanstående är en ARIMA-modell av ordningen $(p,d,q)(P,D,Q)$ med säsongperioden S .

p,d,q har samma betydelse som tidigare

P,D,Q är motsvarande termer för att bestämma säsongens beteende.

4 Univariata modeller

4.1 Inledning

Det första som skall göras, när man bygger en Box-Jenkins modell är att avgöra om en transformering behövs. I en del program finns det hjälpmedel för detta beslut, i en del inte. Den i praktiken vanligaste transformeringen är logaritmering.

För att underlätta det praktiska arbetet med att identifiera en bra modell är det inbakat en del beslutsregler i beräkningarna, men de flesta besluten skall fattas av modellbyggaren själv. Till sin hjälp har han diagram över dels originalserien efter ev transformering, dels över restserien a_t . Framför allt används autokorrelationsfunktionen och dess beteende som beslutsunderlag när man väljer modelltyp och avgör hur många parametrar som behövs. Även vissa andra uppgifter beräknas av programmen för att underlätta arbetet. I Bil 2 beskrivs processen med att bygga en univariat modell för ett praktikfall. Mycket av identifieringsproblematiken är gemensam för alla nivåer av Box-Jenkinsanalysen.

4.2 Identifiering av d och D

För att serien skall kunna göras stationär måste ordningen på d och D bestämmas, men allra först måste man avgöra om transformering är nödvändig. Fortsättningsvis förutsätts att nödvändig transformation är avklarad på ett tillfredsställande sätt.

Den programvara SCB har behandlar värdet på transformeringen (λ) som en parameter. Vilket innebär att det optimala värdet på λ estimeras tillsammans med övriga parametrar.

Av autokorrelationsfunktionens (a.c.f) utseende avgörs om differensering med avseende på säsong (D) och icke säsong (d) behövs. Efter differensering en gång studeras a.c.f på nytt varefter ytterligare differensering görs vid behov.

A.c.f. teoretiska utseende då differensering med avseende enbart på d behöver göras framgår av figur 4:1

A.c.f.

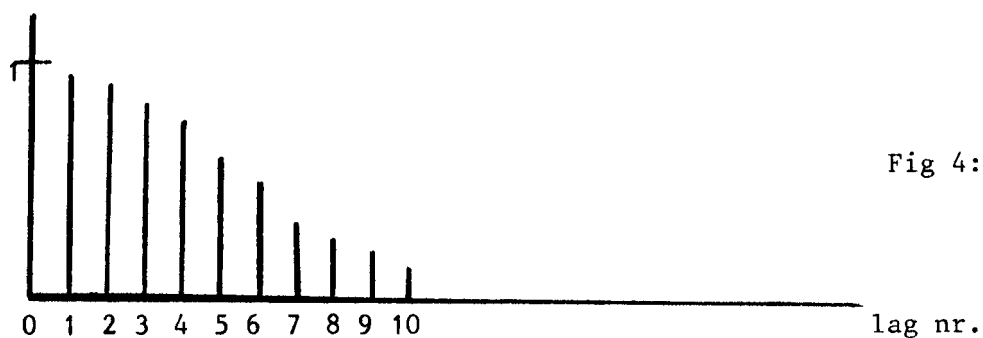
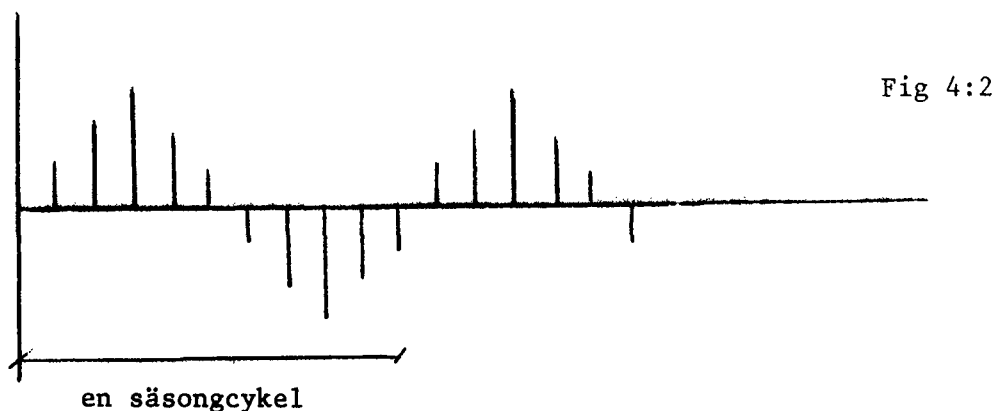


Fig 4:1

Motsvarande a.c.f då differensering enbart med avseende på D behöver göras.

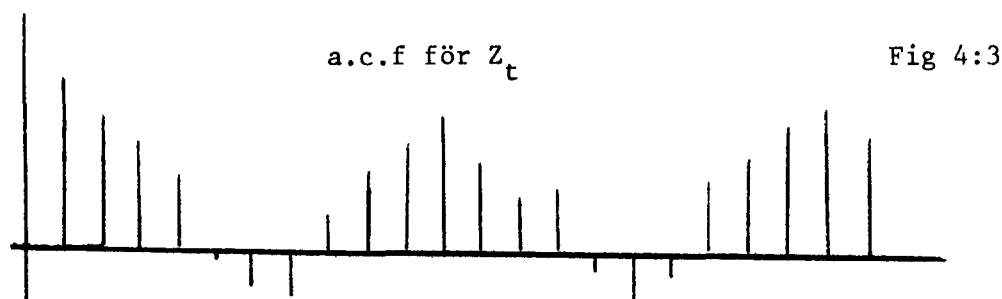


Då differenserings både med avseende på d och D behövs blir det självfallet en kombination av Fig 4:1 och 4:2 som är aktuell. Autokorrelationsfunktionen kan i detta fall i princip se ut hur som helst, då negativa och positiva korrelationseffekter dels kan förstärka och dels ta ut varandra.

Figurerna 4:3 - 4:6 belyser ett faktiskt exempel.

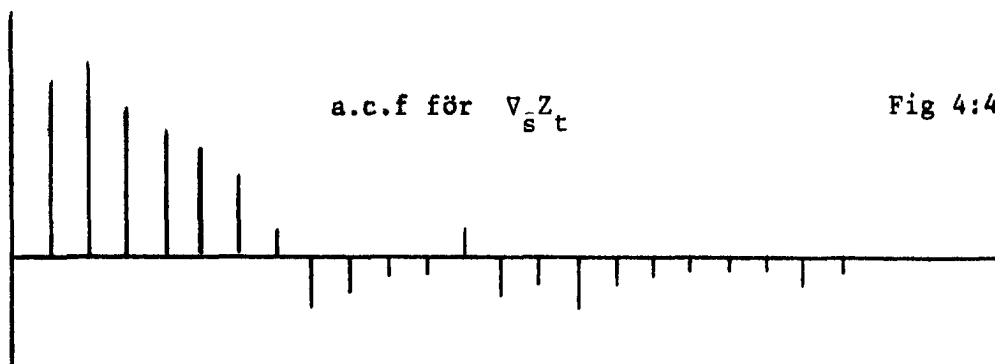
Exempel 4.1

A.c.f för originalserien Z_t har ett utseende i enlighet med fig 4:3



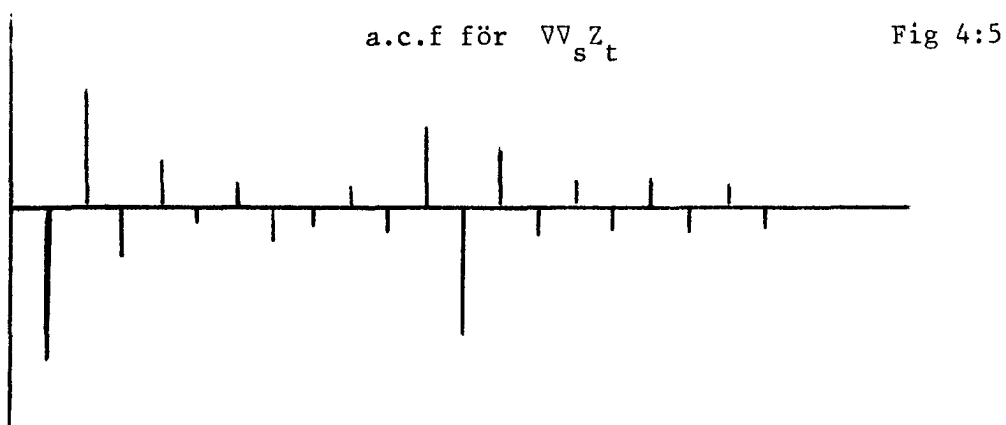
Att en säsongdifferensering behövs framgår av det vågkammade mönstret, vidare ses att de negativa värdena är dämpade varför en differensering även med avseende på icke säsong är trolig. Det bör påpekas att i praktisk tillämpning kommer slumpfaktorn även att påverka a.c.f, varför a.c.f sällan kommer att helt likna den teoretiska.

Differensering med avseende på säsongen ger a.c.f i enlighet med figur 4:4, som således gäller för serien $\nabla_s Z_t$.



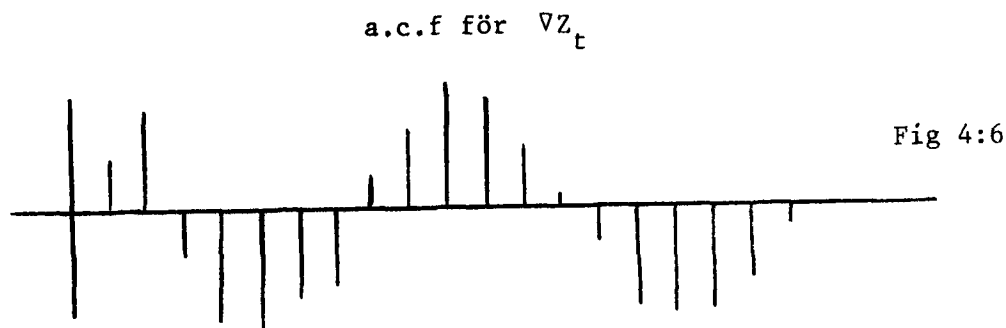
Som synes i figur 4:4 har det vågkammade mönstret försvunnit varför någon ytterligare differensering för att uppfylla stationaritetsskravet avseende säsongen ej är nödvändig. Således är D bestämt till 1.

Dock kan inte serien anses stationär med avseende på icke säsong (se fig 4.1). En differensering motsvarande $d = 1$ görs därför och ger ett utseende på a.c.f i enlighet med figur 4:5, som således gäller för serien $\nabla \nabla_s Z_t$



Inget av de i figurerna 4:1 och 4:2 visade mönsterna återrinns numera i serien, den kan således anses stationär. De återstående markerade stolparna (spikes) i a.c.f används för att identifiera p, q, P, Q . D och d är i och med detta bestämda till $D=1$ och $d=1$. I praktiken är det sällsynt med differenseringar av högre ordning än 1.

Hade vi istället börjat med att differensera med avseende på d och a.c.f för ∇Z_t fått följande utseende.



Vågmönstret framgår klart och de visar att en säsongdifferensering är nödvändig (se fig 4:2). Görs denna kommer man också fram till figur 4:5.

4.3 Identifiering av p och q

Till sin hjälp har modellbyggaren autokorrelationsfunktionen a.c.f) och den partiella autokorrelationsfunktionen (p.a.c.f) för den stationära serien $\nabla^d \nabla_s^D Z_t$.

För definitioner på a.c.f och p.a.c.f. hänvisas t ex till Box-Jenkins bok.

Som tidigare nämnts stör den slumpmässiga variationen a.c.f och p.a.c.f. Därför presenteras diagrammen med konfidensintervall, och de presenterade korrelationerna är således genomsnittsvärden för givna tidsavstånd (lags). Inom detta konfidensintervall kan en korrelationskoefficient ej anses vara signifikant skild ifrån noll. Stora krav ställs här på Box-Jenkins-analytikern. Går man uteslutande efter dessa intervall är det lätt att missa "den bästa" modellen. I identifieringsarbetet måste även restserien (a_t -serien) studeras, samt hjälpinformation av statistiskt-analytiskt slag producerade i programmet. Att studera residualseriens egenskaper är avgörande för ett lyckat resultat med Box-Jenkins ansatsen.

Det finns inbyggda spärrar i metoden som förhindrar att en överidentifiering av modellen sker. En överidentifiering innebär att några parametrar förklarar samma sak, vilket kan resultera i instabila parameterskattningar.

Vid identifiering av p och q i en säsongberoende serie är bara de s första lagen av a.c.f och p.a.c.f av intresse, där s är säsongscykeln. I bilaga 1 och 2 presenteras de teoretiska mönstren för a.c.f och p.a.c.f för låga värden på p och q . Även här går man stegvis fram, av mönstret kan t ex $p = 2$ identifieras. När sedan a_t serien studeras ser man kanske att även en MA-term bör vara med. En ny a_t -serie rensad från 2 AR och 1 MA-term beräknas härafter och studeras. När man har identifierat tillräckligt skall a_t -serien vara en white noise-serie, helst med liten varians.

Den a.c.f som visas i figur 4:5 kan härröra från någon av följande modeller om man bortser från säsong

$$(p,q) = (1,0);(0,1);(2,0);(0,2)$$

Något eller några alternativ kan uteslutas vid studium av p.a.c.f.

4.4 Identifiering av P,Q

Metoden är analog med den i 4.3 beskrivna med den skillnaden att man bara studerar

$S, 2S, 3S, 4S$ osv

I tidigare exempel figur 4:5 kan säsongdelen identifieras med hjälp av bilaga 1 och 2 till $(P,Q) = (1,0);(0,1)$

4.5 Test av den identifierade modellen

När p,q,P,D och Q är bestämda estimeras parametrarna av datorprogrammet. Antalet parametrar är $(p+q+P+Q)$ st.

Den ursprungliga serien Z_t rensas från den preliminära ARIMA-modellen, och restserien a_t erhålles.

Några punkter som skall beaktas innan analysen fortsätter är

- är restvariansen acceptabelt låg? Denna beräknas och anges.
- innehåller modellen för många parametrar? Korrelationen mellan parametrarna beräknas för att visa om de förklarar samma

sak. Hög korrelation mellan 2 parametrar innebär oftast att en kan tas bort ur modellen.

- plotta och studera a_t -serien? Man letar efter extrema värden. Dessa skall studeras särskilt för att man skall kunna avgöra om dessa beror på slumpen eller speciella störningar i serien. I det senare fallet bör ytterligare åtgärder övervägas för att korrigera eller utesluta de tveksamma observationerna, se kap 6.

- studera a.c.f och p.a.c.f för a_t -serien. Beslut om a_t -serien kan anses slumpmässig eller ej fattas på basis av dessa figurer.

- är parameterskattningarna signifikant skilda från 0?
Variansen för parametrarna studeras.

- kan modellens beteende anses rimligt? Modellen får ej beskriva ett tillfälligt matematiskt samband som strider mot t ex ekonomisk teori.

När alla punkter gåttts igenom och nöjaktigt svar erhållits anses den "bästa" ARIMA-modellen vara identifierad. Med hjälp av ovanstående analysmoment kan det dock också hända att man kan ifrågasätta om en Box-Jenkins-modell överhuvudtaget är lämplig.

5 Transfer functions

5.1 Allmänt

Här presenteras en mycket förenklad och populariserad framställning av vad Box-Jenkins menar med transfer functions. Syftet är att ungefär beskriva huvuddragen i identifieringsarbetet. För djupare insikt hänvisas till Box-Jenkins bok.

5.2 Beskrivning av transfer functions

Figur 2:5 visar hur en transfer functions modell är uppbyggd.

$$Y_t = U_t + N_t + a_t \quad 5:1$$

där Y_t är den beroende serien.

U_t är den variation i Y_t som förklaras av de oberoende serierna $(x_{1t}, x_{2t}, \dots, x_{mt})$

N_t är den del av variationen som enbart beror på Y_t -seriens unika historiska förlopp.

Exempel

$$\begin{aligned} \text{Antag att } U_t &= a + bx_t \\ \therefore Y_t &= a + bx_t + N_t \end{aligned}$$

I fallet då U_t förklarar all variation i Y_t sånär som på en slumpfaktor gäller

$$Y_t = a + bx_t + \varepsilon_t \quad \text{ty } \varepsilon_t = N_t + a_t$$

Således gäller: Om alla förutsättningar i en vanlig regressionsanalys är uppfyllda, blir den bästa regressionslinjen estimerad även ur en Box-Jenkinsanalys.

Detta kan även formuleras:

Om ej alla förutsättningar i en regressionsanalys är uppfyllda beskriver en ideal Box-Jenkins modell det historiska förloppet hos en serie bättre

Fördelarna med en Box-Jenkinsansats är således

1 alla samband mellan x-serien och y-serien inom den vida klassen av ARIMA-modeller kan estimeras

2 den beroende serien kan även innehålla en variation som inte förklaras av de oberoende serierna, men som kan förklaras av den beroende serien själv.

För överskådlighetens skull behandlas här bara fallet med en oberoende serie. Vidare förutsätts serierna vara säsongsoberoende. En generalisering till flera serier och säsongssamband är självfallet möjlig.

U_t kan skrivas

$$U_t = V_0 X_t + V_1 X_{t-1} + \dots + V_n X_{t-n} = V(B)X_t \quad 5:2$$

Vidare gäller att N_t -serien kan skrivas

$$\phi(B)\nabla^d N_t = \Theta(B)a_t \quad 5:3$$

5:1 kan då skrivas med hjälp av 5:2 och 5:3

$$\nabla^d Y_t = V(B)\nabla^d X_t + \frac{\Theta(B)}{\phi(B)} \cdot a_t \quad 5:4$$

I princip kan olika antal differenser med avseende på x och y -serierna tillåtas, eller

$$\nabla^{dy} Y_t = V(B) \nabla^{dx} X_t + \frac{\Theta(B)}{\phi(B)} \cdot a_t \quad 5:5$$

Som generell funktion av $V(B)$ används

$$V(B) = \frac{\omega(B)}{\delta(B)} B^b \quad 5:6$$

där

$$\delta(B) = 1 - \delta_1 B - \delta_2 B^2 - \dots - \delta_T B^T$$

$$\omega(B) = \omega_0 - \omega_1 B - \omega_2 B^2 - \dots - \omega_S B^S$$

b är fördröjningen mellan orsak och verkan (delay)

5:6 är en transferfunktion av ordningen (T,S,b)

5:5 och 5:6 ger

$$\nabla^d y_t = \frac{\omega(B)}{\delta(B)} \cdot B^b \nabla^d x_t + \frac{\Theta(B)}{\phi(B)} \cdot a_t \quad 5:7$$

Således skall följande parametrar identifieras; dx , dy , p , q , T , S , b . Då denna identifieringsprocess är svår att formalisera beskrivs förfaringssättet verbalt och bara i grova drag.

Steg 1:

Identifiera den univariata modellen (U.V.M) för X-serien.

Det är viktigt att detta görs noggrant så att restserien α_t blir white noise. Vi får

$$\nabla^x x_t = \frac{\Theta_x(B)}{\phi_x(B)} \cdot \alpha_t \quad 5:8$$

Steg 2:

Identifiera U.V.M för Y-serien

$$\nabla^y y_t = \frac{\Theta_y(B)}{\phi_y(B)} \cdot a_t \quad 5:9$$

Denna identifiering behöver ej göras så att a_t blir white noise, men det fortsatta arbetet underlättas av att denna identifiering görs så bra som möjligt.

Steg 3:

Rensa Y-serien från den i 5:8 presenterade modellen, dvs den variation i Y-serien som förklaras av X-serien tas bort. Detta kallas att göra Y-serien prewhitened.

Den restserie som erhålles kallas β_t

Steg 4:

Korskorrelationen mellan α_t och β_t -serierna beräknas och dessa kan sedan användas till att göra preliminära estimat till para-

metrarna T_y korskorrelationerna är proportionella mot vikterna V_i i 5:2
dvs

$$r_{\alpha\beta}(K) = C \cdot V_k$$

Från dessa korrelationsdiagram identifieras dock först parameterantalet (T , S , b).

Nu är således U_t preliminärt estimerad.

Steg 5:

N_t estimeras preliminärt genom att p och q väljes till samma storlek som de har i modell 5:8.

Givet N_t , de preliminära estimaten enligt steg 4 samt ordningen på q och p kan även estimat på dessa parametrar beräknas.

Nu är således hela modellen 5.6 preliminärt estimerad.

Steg 6:

Modellen kontrolleras genom karaktäristika som beräknas av datorprogrammen. På basis av dessa kan modellens parameterantal reduceras eller eventuellt utvidgas.

Metoden under steg 1-5 ger lätt en överrepresenterad modell genom att ett givet historiskt beteende både förklaras i U_t och N_t -serierna. I så fall skall modellen reduceras. Vidare kan en icke identifierad men förklaringsbar variation hittas, och då kanske modellen måste utvidgas. Med hjälp av alla dessa beräknade karaktäristika förbättras modellen successivt tills man blir nöjd, och slutprodukten blir att modellen definitivt spikas.

Metodiken är helt analog då flera X-serier används och då säsongsb beroende existerar.

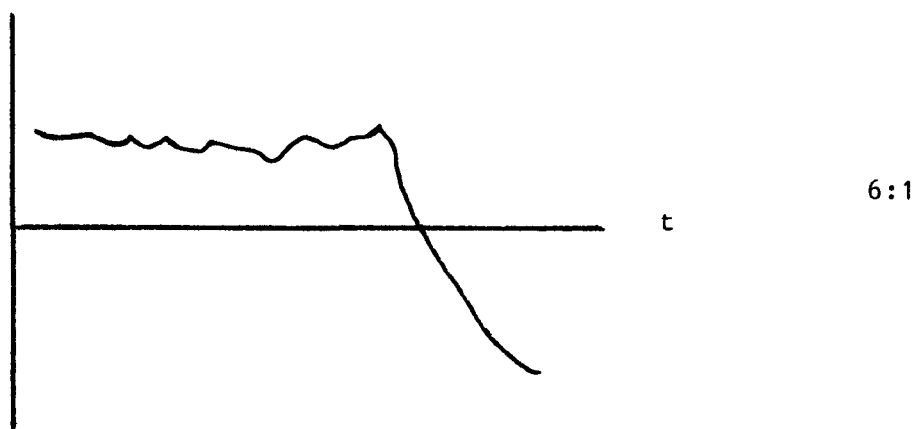
6 Intervention theory

6.1 Allmänt

Denna aspekt av Box-Jenkins ansatsen är relativt ny och av speciellt intresse för SCB och användare av SCB-data.

Metoden som kan ses som ett specialfall av transferfunction analysen syftar till att komma till rätta med speciella problem som kan finnas i tidsseriedata som t ex strejker, förändringar på grund av politiska beslut, definitionsändringar, ändring i datainsamlingsmetoderna och andra störningar i tidsserien av tillfällig eller permanent natur.

Vidare kan man analysera serier som av en eller annan anledning har en markerad ändring i lutningen som i fig 6:1



Genom att studera residualserien med avseende på extrema värden och förändringar i variationen avgörs om en interventionsanalys är nödvändig.

6.2 Kortfattad beskrivning av metodiken

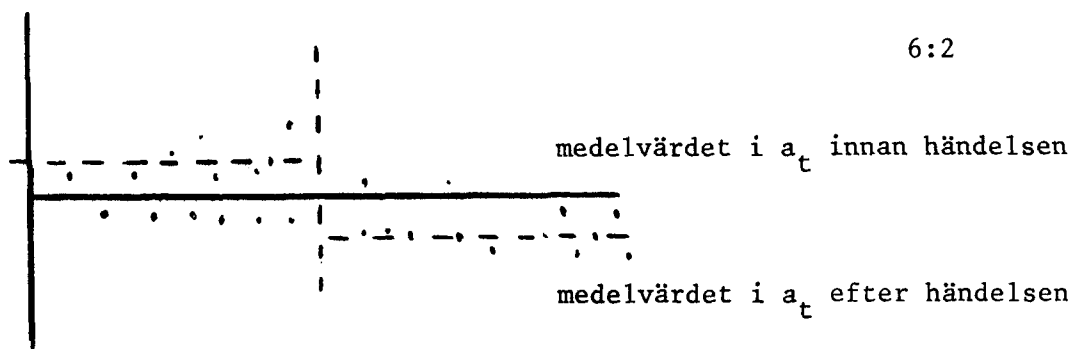
Identifikationsprocessen är densamma som i fallet med transferfunktion, utom när det gäller att bestämma ordningen på de aktuella parametrarna. Skillnaden beror på att man vid interventionsanalysen intruducerar så kallade dummyvariabler. Det första steget i identifikationsprocessen måste baseras på mer eller mindre korrekta gissningar. Sedan testas gissningarna och

modellen ändras med samma teknik som beskrivs i kap. 5. Självfallet gäller att ju korrektare gissning man gör i det första steget ju mindre arbete behöver läggas ned i det fortsatta arbetet.

Exempel 6:2

Säg att fr o m observation 40 har t ex en skatteändring. Från a_t -serien ser vi att det blivit en kraftig påverkan

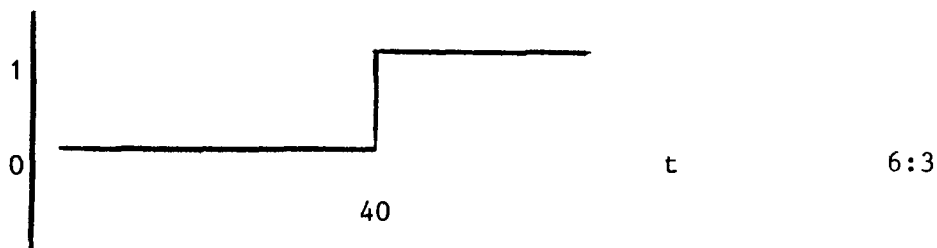
a_t -serien kan se ut



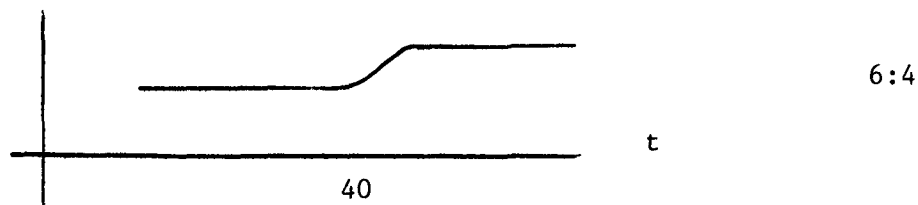
Vi definierar en hjälpvariabel

$$\xi_t = \begin{cases} 0 & \text{för alla } t < 40 \\ 1 & \text{" - } t > 40 \end{cases}$$

eller



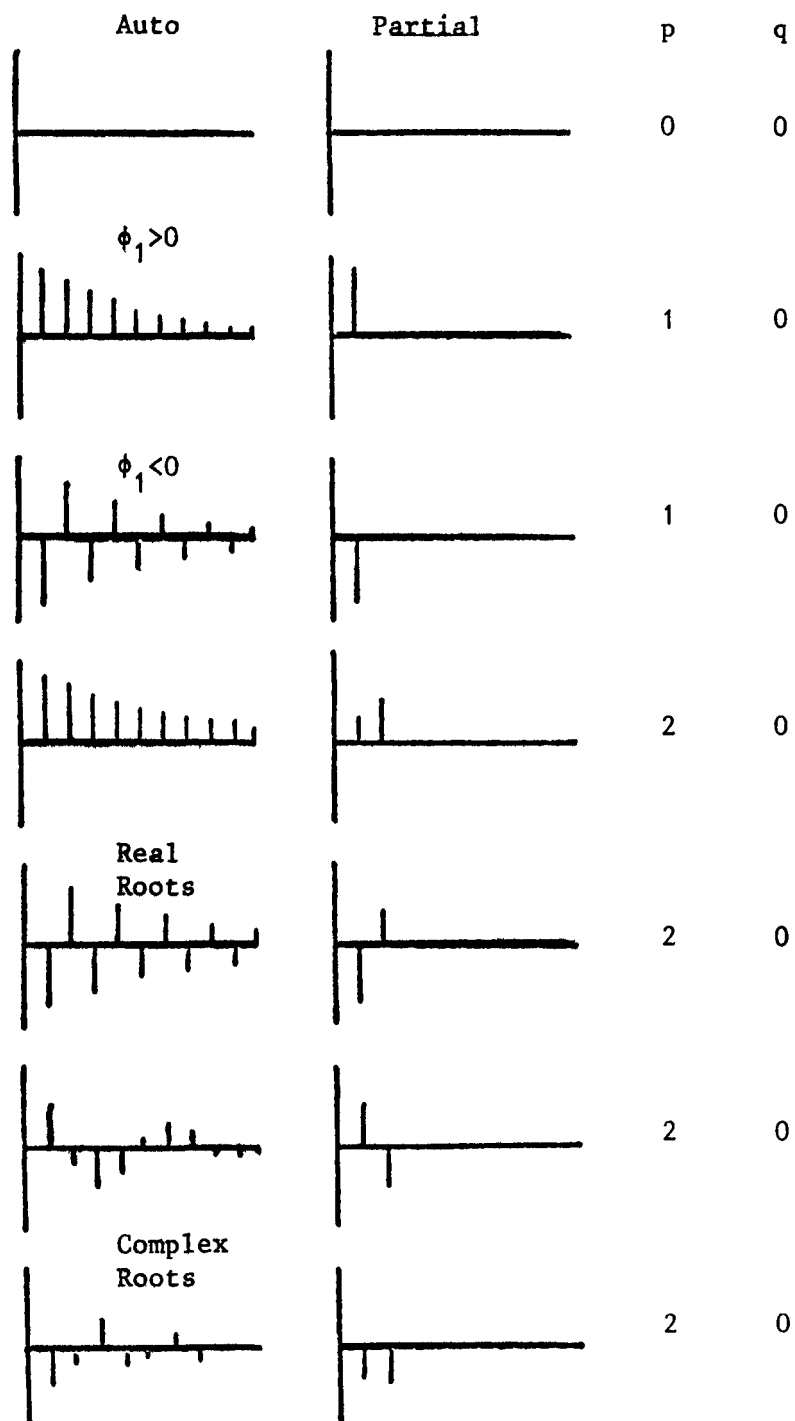
Vi 'tror' att effekten av ξ_t blir

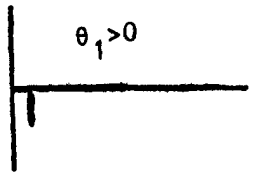
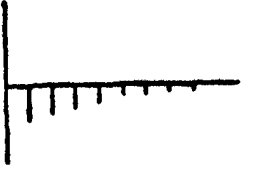
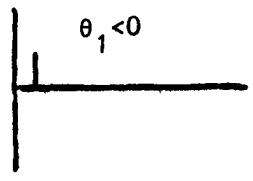
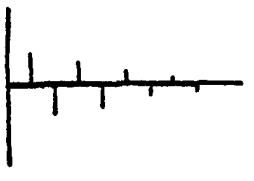
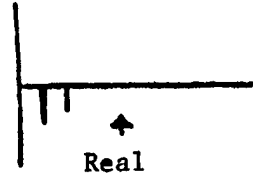
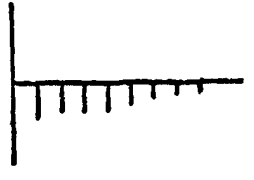
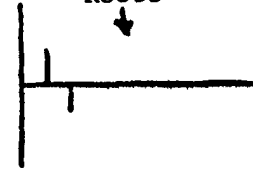
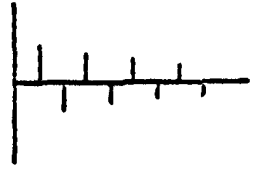
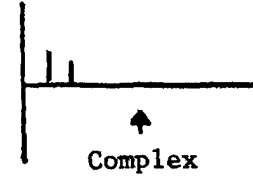
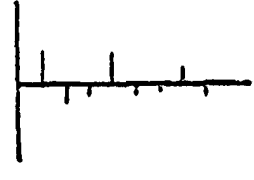
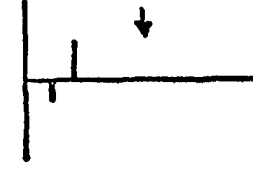
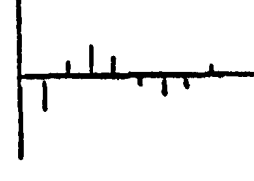
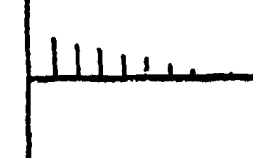
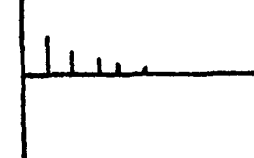


Figur 6:4 kan beskrivas av $\omega_0/1-\delta$. Som första försök används således en transfer funktion med $T=0$ $S=1$ $b=0$.

Vi gör sedan en preliminär estimation av δ och N_t . Utifrån korskorrelationer och data från N_t -serien beslutar vi om och hur vi skall gå vidare i modellbyggandet.

Teoretiska utseendet på auto- och partiella autokorrelo-
grammen för givna parameterspecifikationer.



Auto	Partial	p	q
$\theta_1 > 0$ 		0	1
$\theta_1 < 0$ 		0	1
 ↑ Real Roots ↓		0	2
		0	2
 ↑ Complex Roots ↓		0	2
		0	2
		1	1

Ett exempel som visar tillvägagångssättet vid byggandet av en univariat modell.

Industrivolyminindex 68-78. Kvartalsdata för SNI 2 + 3

Program USID (i figur B:1-B:5); Preliminär modellidentifikation.

USID används för att göra serien stationär dvs bestämma

- preliminärt värde på λ
- ordningen på d
- ordningen på D

Kommentarer till figurerna

B:2 Vi observerar stora värdena i lag 4, 8, 12, 16 varför vi börjar med att differensera m a p säsong en gång dvs $D=1$.

Vad som då händer i a.c.f syns i figur B:3.

B:3 Av de stora värdena i lag 1, 2, 3 beslutar vi att differensera även m a p icke säsong. De kraftiga värdena i multiplar av säsongperioden har försvunnit varför D är bestämt till $D=1$. Vi ser att en säsongdifferensering reducerade variansen från 259.4 till 31.1.

B:4 Slutsatsen av figuren är att ingen mer differensering är nödvändig. Således är även d bestämd till $d=1$. Variansen är nu nere i 13.0. Vi har signifikanta värden på a.c.f. i lag 3 och 4. För att kunna identifiera p och q på bästa sätt måste vi även studera p.a.c.f. som visas i figur B:5.

Identifiering av p och q

I både figur B:4 och B:5 har vi ett signifikant värde i lag 3 vidare har vi ett högt värde i lag 1. Genom att studera bilaga 1 ser vi att detta kan härröra både från

en AR(3) och en MA(3) modell. Vi bör därför prova båda.

Identifiering av P och Q

Figur B:4 har ett signifikant värde i lag 4 medan figur B:5 har värden i 4, 8, 12, 16 som beskriver en exponentiellt avtagande funktion. Från bilaga 1 vet vi att detta motsvaras av det teoretiska mönstret från en MA(1)-process dvs $P=0$ och $Q=1$.

Sammanfattning av informationen från USID

- $\lambda = 1.0$ dvs ingen transformation (i detta sammanhang berörs ej λ beroende på att beräkningen skall ändras i nästa version av programvaran).

- två modeller verkar troligast nämligen

ARIMA (310)(011) och ARIMA (013)(011)

Program USES (Figur B:6-B:12); Estimation och Modellkontroll

B:6 Konvergens är ej uppnådd. Från andra sidor i outputen vet vi dock att vi är tillräckligt nära dvs 3:e decimalen är korrekt.

AR(3)-processen faktoriseras och uppgifter fås om period m m. Alla parameterskattningar redovisas med standardavvikelsen.

B:7 Korrelationen mellan parametrarna beräknas. Om denna är hög (>0.75), betyder det att parametrarna förklarar samma sak och således indikerar att modellen kan reduceras.

B:8 Residualserien skall studeras noggrant m a p

- olika varians i olika delar av serien
- outliers

I detta fall har vi bara en outlier vid $t=37$ på 2.7 S.E.

B:9 Histogrammet skall i princip visa en normalfördelning.

B:10-B:11 A.c.f och p.a.c.f visar att inga signifikanta autokorrelationer finns kvar (dvs inget beroende mellan observationerna).

B:12 All riktig information samlas på en sida vilken är den enda som i princip behöver sparas.

Sammanfattning av USES

- Modellen (310)(011) har bäst data
- ingen korrelation mellan parametrarna
- inget beroende i restserien, vidare kan den med lite god vilja anses normalfördelad
- residualserien innehåller bara en outlier och denna föranleder inga ytterligare åtgärder
- den första autoregressiva parametern är ej signifikant skild ifrån 0.
- vi har nu reducerat den ej förklarade variansen till 9.1 (från 259.4 dvs 96.5 % av variansen)
- χ^2 -värdet säger ej emot att modellen passar till datamaterialet.

Vi anser modellen identifierad och använder den för prognossyfte.

Program USFO; Prognoser (B:13-B:15)

B:13 Vi gör en prognos för fyra kvartal framåt, och använder bara de kända värdena t o m t = 44

Prognosen för kv 45 blir 127.7

46 136.5

Dessa siffror kan jämföras med de sanna som var 130 resp 140. Vi har således ett prognosfel på 1.73 resp 2.43 %.

B:15 Här kan vi se hur prognosen för Kv 46 ändras i och med att det kända värdet för Kv 45 används. Prognosen för Kv 46 blir nu 138.5 med ett prognosfel på 1.1 %.

Genom att på detta sätt studera hur prognoserna ändras kan man studera hur modellen fungerar, t ex vid tillfällen då man vet att en trendändring skett, kan man studera hur lång tid det tar innan modellen kommer på 'rätt nivå'.

Självfallet kan dessa studier hur modellen fungerar utvidgas, genom att man simulerar olika händelser och utvecklingar. Genom att simulera kan man således få en mycket god bild av modellens egenskaper vid olika realistiska händelseförlopp.

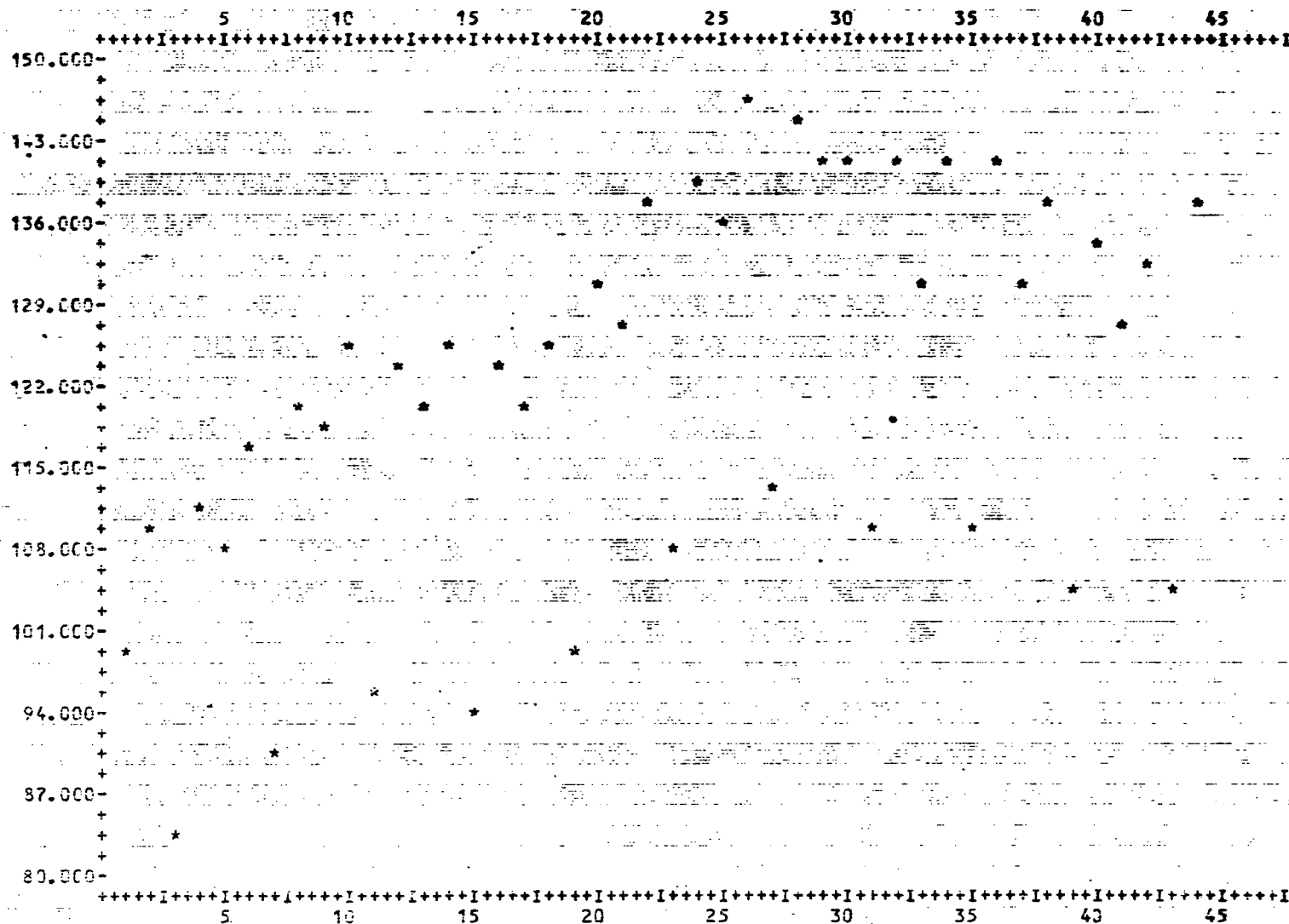
ANALYSIS FOR TRANSFORMATION LAMBDA = 1.00 AND DIFFERENCING D = 0 DS = 0 AND S = 4

GRAPH OF INDEX KVARTAL SNI 2

Figur B1

TRANSFORMATION PARAMETER LAMBDA = 1.00
NON SEASONAL DIFFERENCING D = 0
SEASONAL DIFFERENCING DS = 0
SEASONAL PERIOD S = 4

VALUE



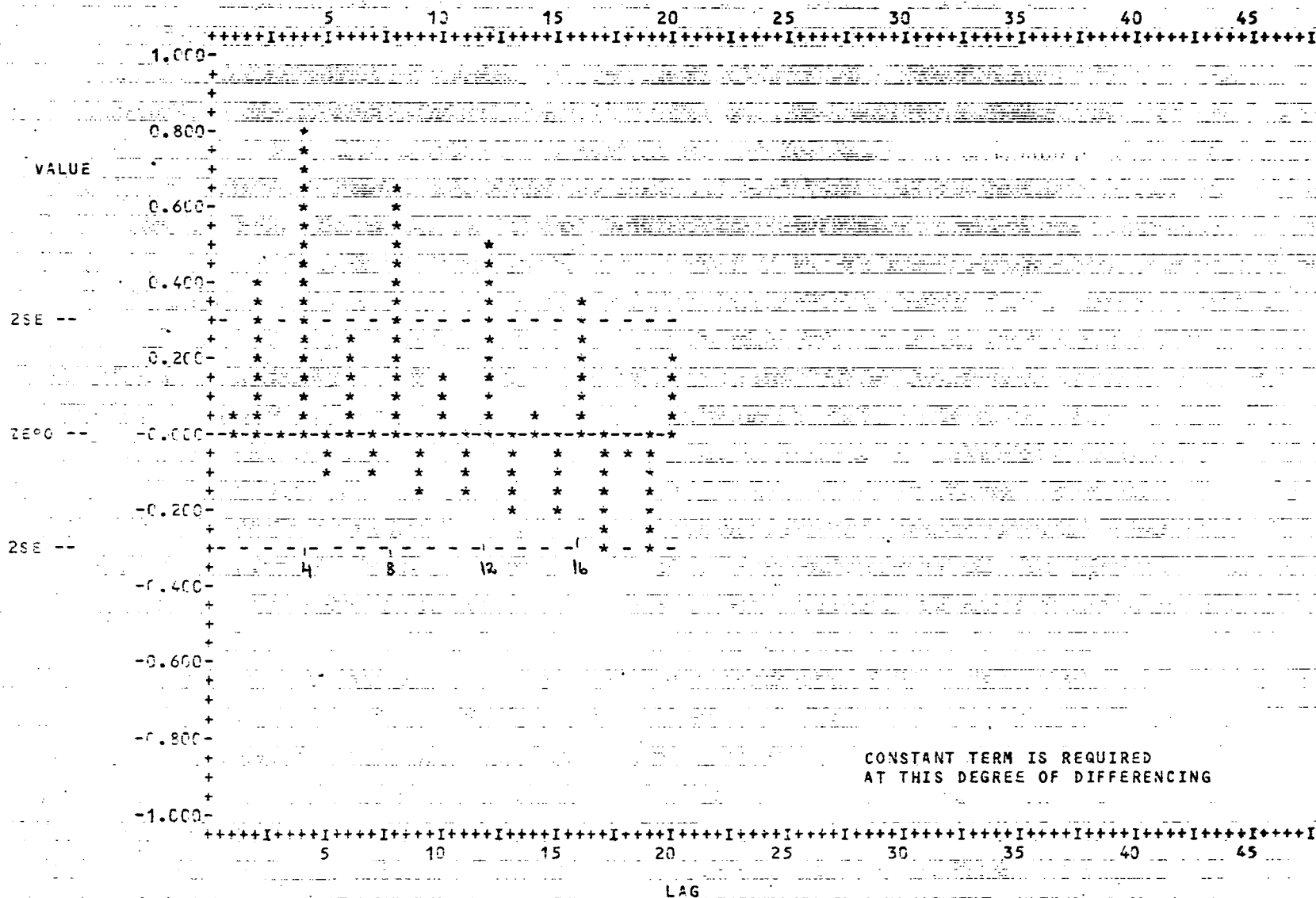
DATING PARAMETERS INCONSISTENT - NORMAL NUMBERING FACILITY USED

VARIANTS 259.4
MODEL 220 2.4

GRAPH OF AUTOCORRELATION FUNCTION OF INDEX KVARTAL SNI 2

Figur B:2

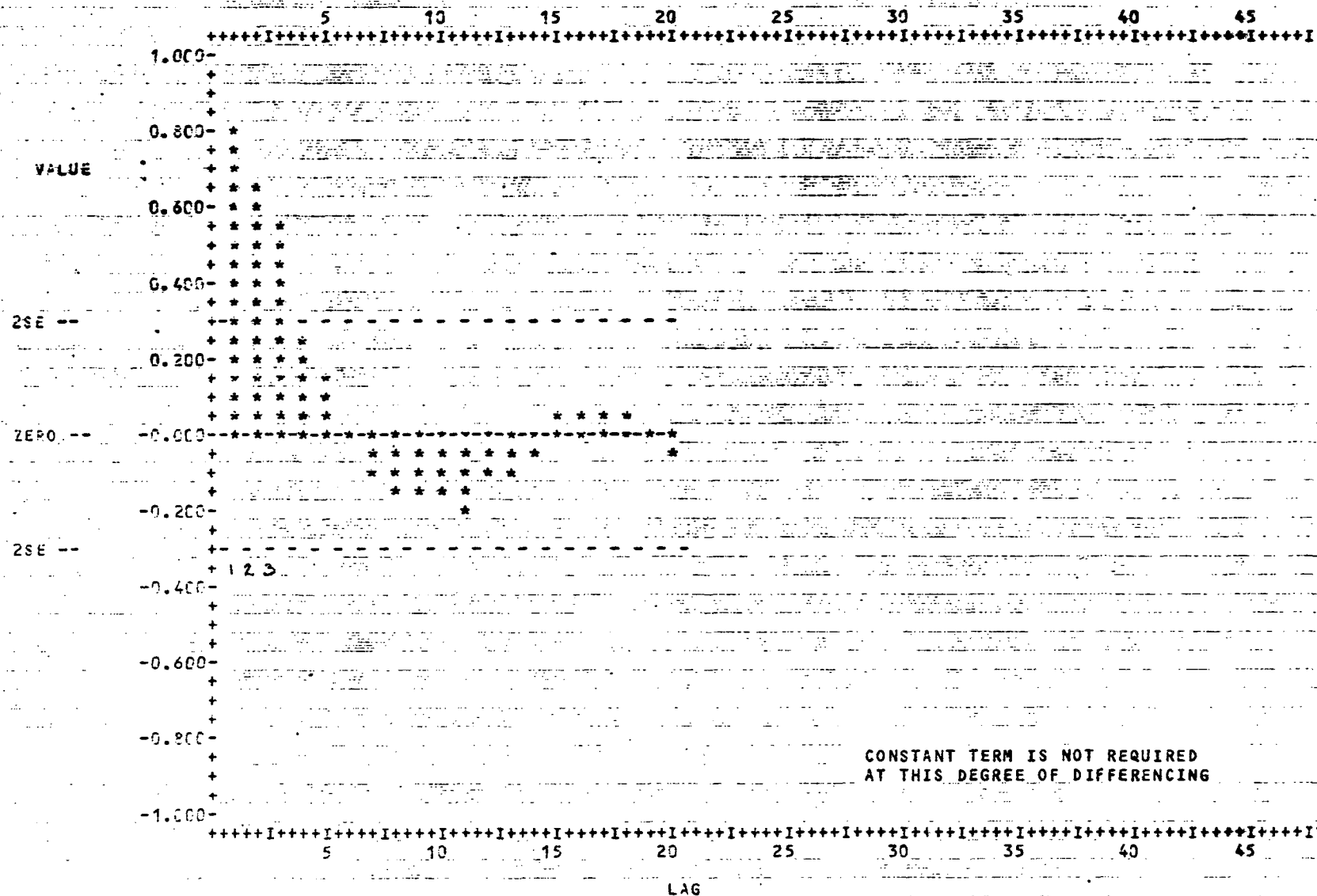
TRANSFORMATION PARAMETER LAMBDA = 1.00
 NON SEASONAL DIFFERENCING D = 0
 SEASONAL DIFFERENCING DS = 0
 SEASONAL PERIOD S = 4



GRAPH OF AUTOCORRELATION FUNCTION OF INDEX KVARTAL SNI 2

Fig B.3

TRANSFORMATION PARAMETER LAMBDA = 1.00
 NON SEASONAL DIFFERENCING D = 0
 SEASONAL DIFFERENCING DS = 1
 SEASONAL PERIOD S = 4

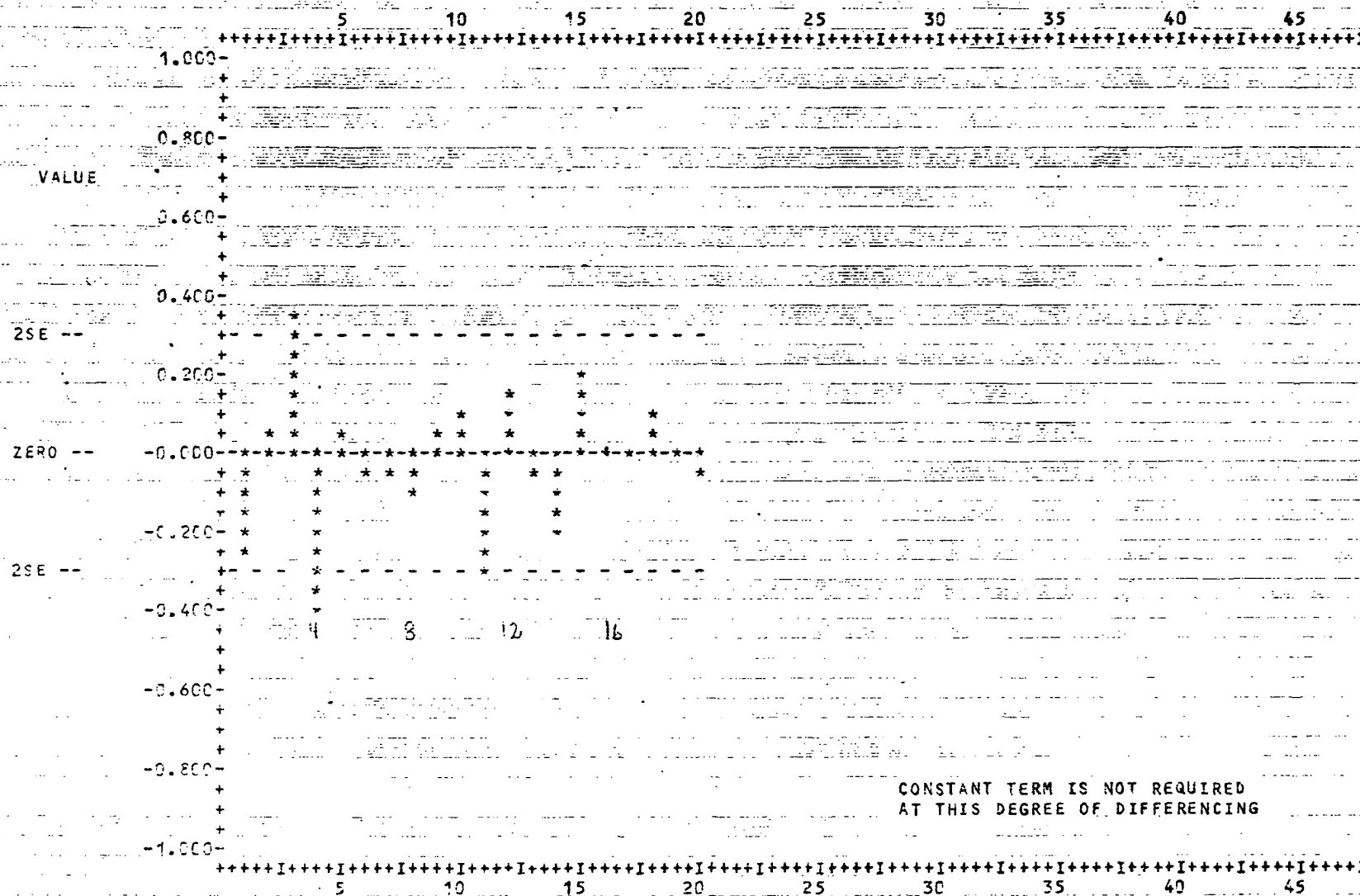


VARIANS 31.1
 MEAN VALUE 4.5

GRAPH OF AUTOCORRELATION FUNCTION OF INDEX KVARTAL SNI 2

Figur B 4

TRANSFORMATION PARAMETER LAMBDA = 1.00
 NON SEASONAL DIFFERENCING D = 1
 SEASONAL DIFFERENCING DS = 1
 SEASONAL PERIOD S = 4

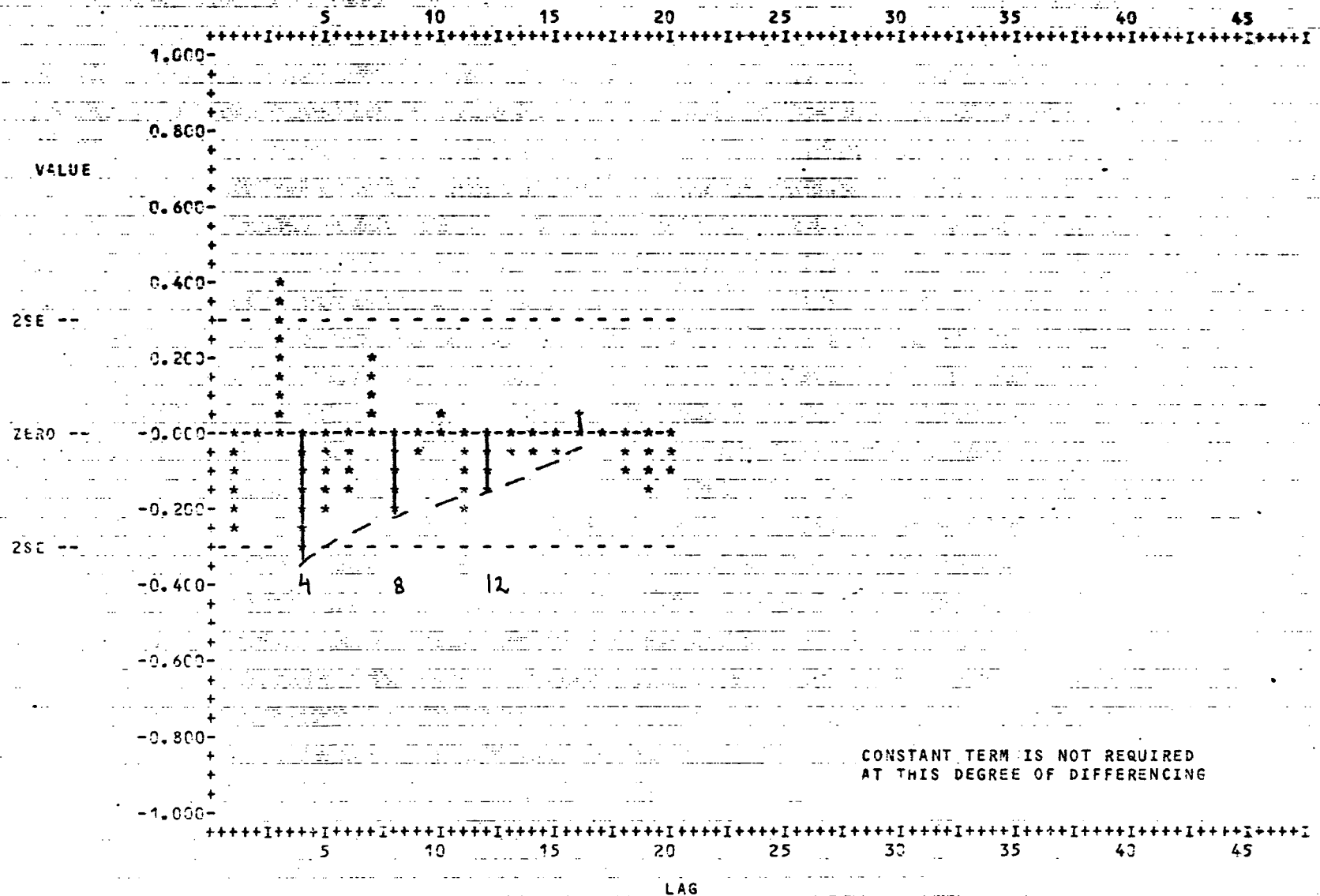


VARIANS 13.0
 MODELVAR 0.1

GRAPH OF PARTIAL AUTOCORRELATION FUNCTION OF INDEX KVARTAL SNI 2

TRANSFORMATION PARAMETER LAMBDA = 1.00
 NON SEASONAL DIFFERENCING D = 1
 SEASONAL DIFFERENCING DS = 1
 SEASONAL PERIOD S = 4

Figur 85



FINAL PARAMETER ESTIMATES

ITERATION - 10 SUM OF SQUARES = 0.374706464D+03

	PHI	THETA	SPHI	STHETA
VALUE 1	-0.1761			0.8417
S.E.	0.1478			0.0722
VALUE 2	0.2801			
S.E.	0.1443			
VALUE 3	0.2991			
S.E.	0.1484			

VARIANCE OF BASE SERIES	=	254.431
VARIANCE OF DIFFERENCED SERIES	=	15.949
ESTIMATED RESIDUAL VARIANCE	=	9.139
ESTIMATED RESIDUAL S.D.	=	3.023

MODEL IN FACTORED FORM

LAMBDA	=	1.00
NON-SEASONAL DIFFERENCING	=	1
SEASONAL DIFFERENCING	=	1
SEASONAL PERIOD	=	4

$$\text{PHI}(B) * (W(T) - C) = \text{STHETA}(B) * A(T)$$

C = 0 (NO CONSTANT REQUIRED)

$$\text{PHI}(B) = (1 - 0.74B) (1 + 0.92B + 0.40B^2)$$

$$\text{STHETA}(B) = (1 - 0.84B)$$

PERIOD, FREQUENCY AND DAMPING FACTORS OF COMPLEX OPERATORS

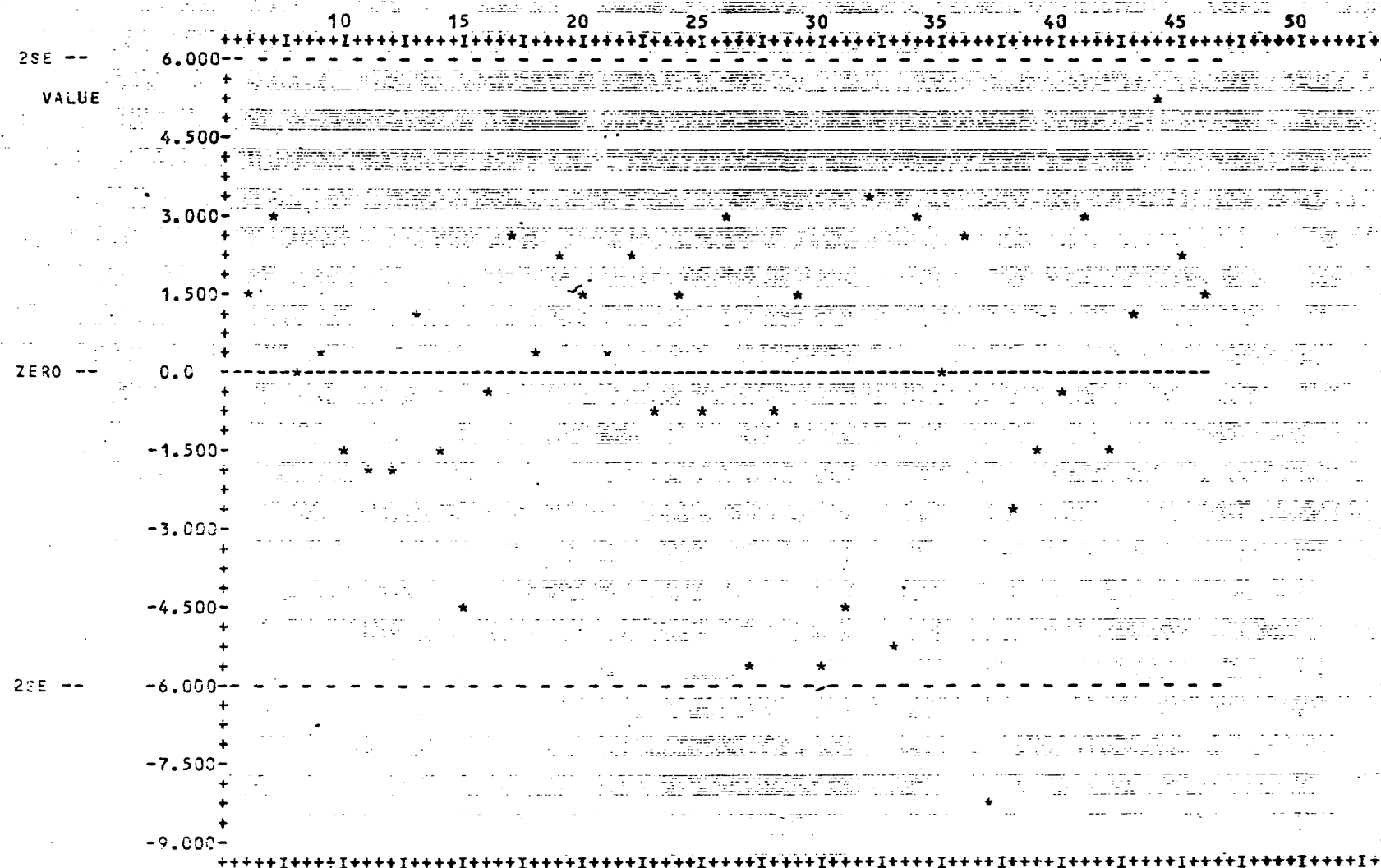
OPERATOR	DAMPING FACTOR	FREQUENCY	PERIOD
$(1 + 0.92B + 0.40B^2)$	0.63	0.379	2.639

Figur 8.7

CORRELATION MATRIX OF PARAMETER ESTIMATES-

	PHI 1	PHI 2	PHI 3	STHETA 1
PHI 1	1.00			
PHI 2	0.12	1.00		
PHI 3	-0.19	0.10	1.00	
STHETA 1	0.18	0.18	-0.17	1.00

LAMBDA = 1.00
 NON-SEASONAL DIFFERENCING = 1
 SEASONAL DIFFERENCING = 1
 SEASONAL PERIOD = 4

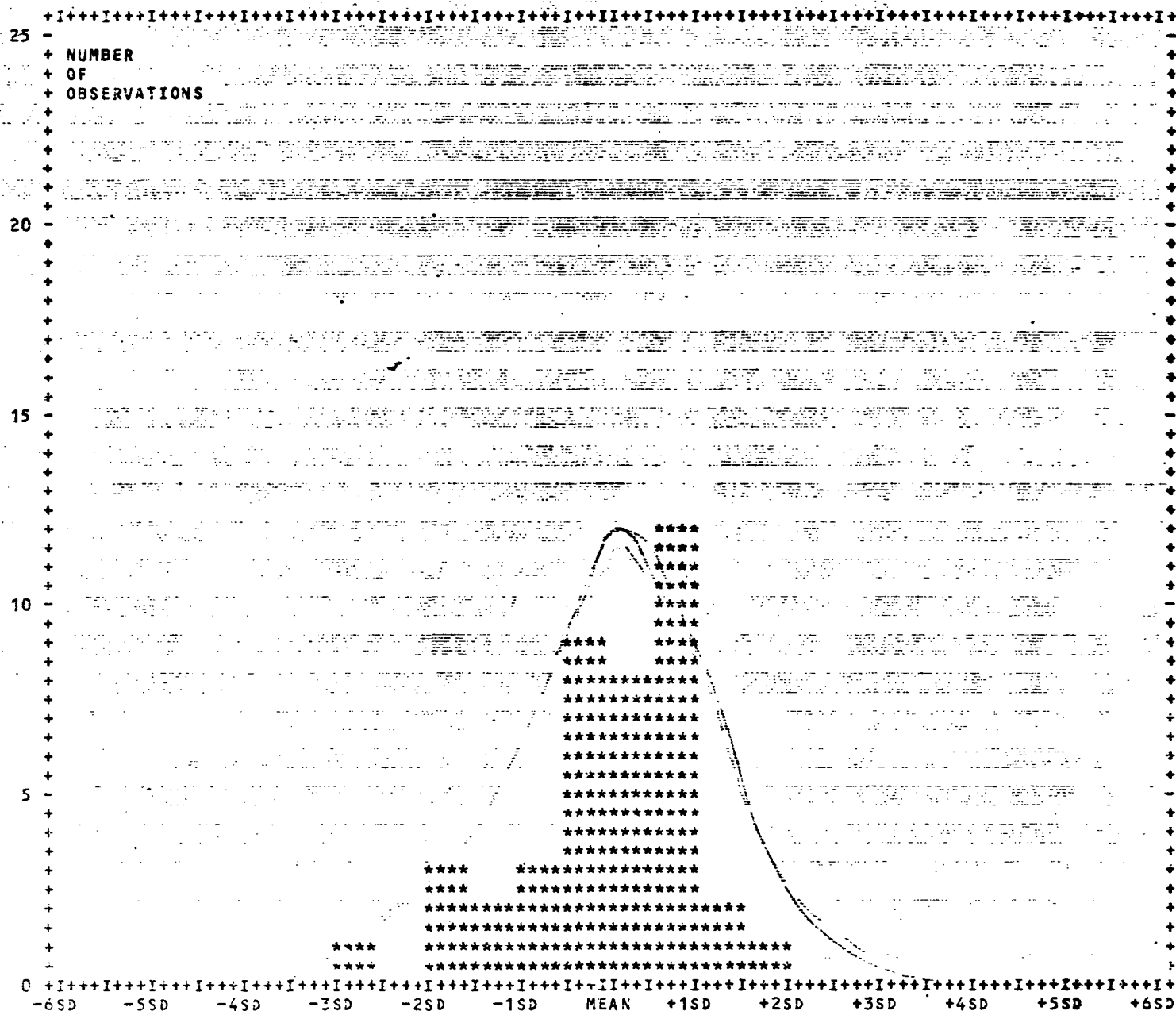


50

STANDARD DEVIATION = 3.023

HISTOGRAM OF RESIDUAL SERIES FOR INDEX KVARTAL SNI 2

Figur B9



MEAN OF RESIDUALS

-0.13475883

1 STANDARD DEVIATION

3.082



45

LAG

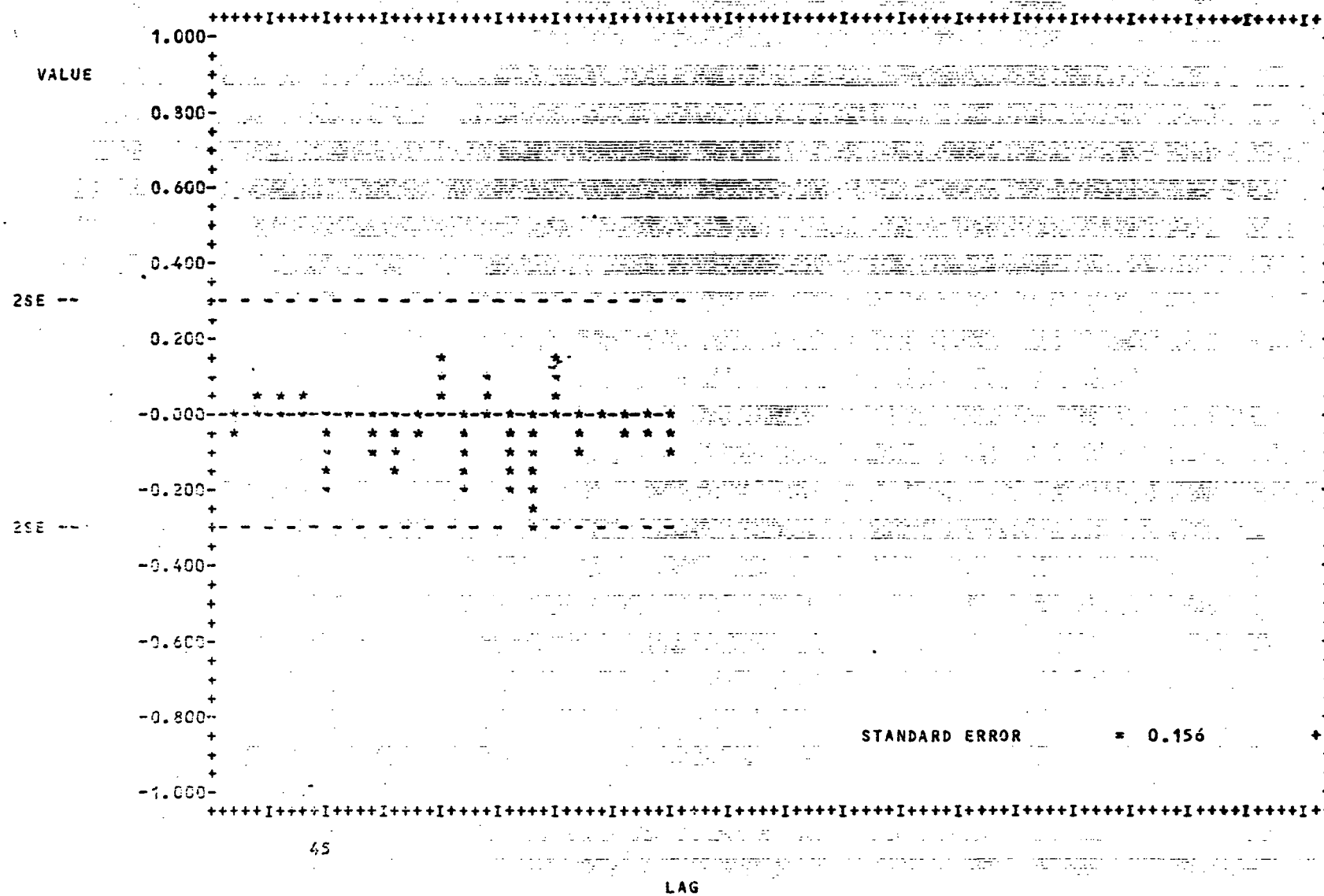
AUTOCORRELATION FUNCTION OF RESIDUAL SERIES OF INDEX KVARTAL SNI 2

```
LAG      0
VALUE    1.000
```

LAG	1	2	3	4
VALUE	-0.029	0.264	0.052	0.038

Figur B.11

PARTIAL AUTOCORRELATION FUNCTION OF RESIDUAL SERIES OF INDEX KVARTAL SNI 2



USES SUMMARY SHEET - INDEX KVARTAL SNI 2

LENGTH OF SERIES SUBSET N = 46 LAMEDA = 1.00 D = 1 DS = 1 S = 4

VARIANCE OF BASE SERIES = 254.431
VARIANCE OF DIFFERENCED BASE SERIES = 15.949

ESTIMATED MODEL

ESTIMATED RESIDUAL VARIANCE OF A(T) = 9.139
ESTIMATED RESIDUAL S.D. OF A(T) = 3.023

A. UNFACTORED OPERATORS

PHI (B) = (1 + 0.176B - 0.280B² - 0.299B³)
S.E. 0.148 0.144 0.148
STHETA(B) = (1 - 0.342B)
S.E. 0.072

B. FACTORED OPERATORS

PHI(B) * (W(T) - C) = STHETA(B) * A(T)

C = 0 (NO CONSTANT REQUIRED)

PHI (B) = (1 - 0.74B) (1 + 0.92B + 0.40B²)
STHETA(B) = (1 - 0.84B)

PERIOD, FREQUENCY AND DAMPING FACTORS OF COMPLEX OPERATORS

OPERATOR	DAMPING FACTOR	FREQUENCY	PERIOD
(1 + 0.92B + 0.40B ²)	0.63	0.379	2.639

COMMENTS

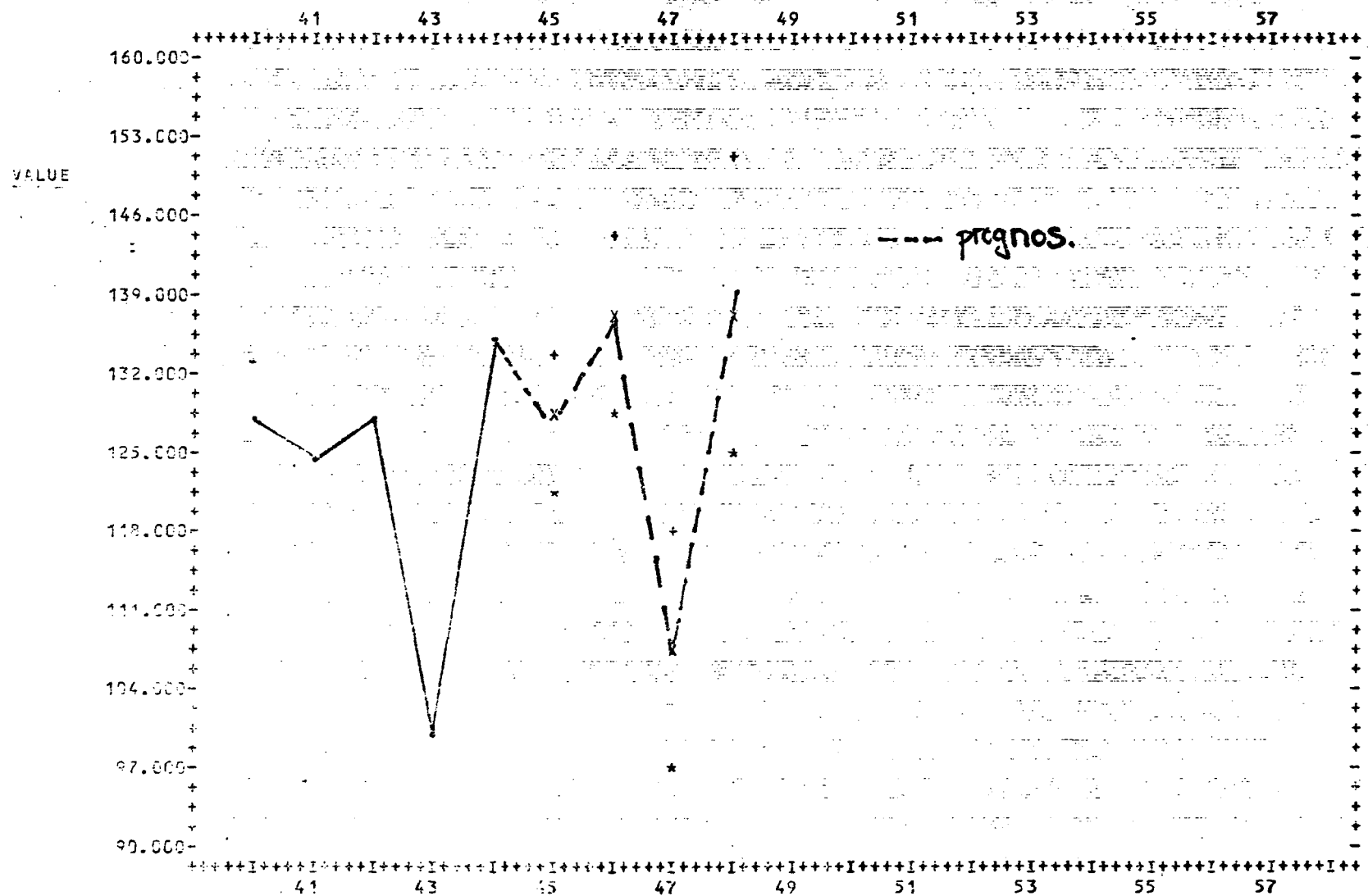
1. CONVERGENCE NOT OBTAINED IN 10 ITERATIONS
2. PARAMETERS WHICH ARE NOT SIGNIFICANT
PHI 1,2
3. NO PARAMETER CORRELATIONS GREATER THAN 0.75
4. NUMBER OF SIGNIFICANT RESIDUALS = 1 OCCURRING AT T = 37
5. MEAN OF RESIDUALS NOT SIGNIFICANTLY DIFFERENT FROM ZERO
6. NO SIGNIFICANT AUTOCORRELATIONS
7. CHI SQUARE VALUE = 10.7 ON 16 DEGREES OF FREEDOM
8. NO SIGNIFICANT PARTIAL AUTOCORRELATIONS

FORECAST ORIGIN T = 44

B13

FORECAST FOR POINT T	LOWER LIMIT 95%	FORECAST	UPPER LIMIT 95%	ACTUAL VALUE	FORECAST ERROR	PER CENT FORECAST ERROR	PSI WEIGHT	S.D. OF FORECAST ERROR OF TRANSFORMED VARIABLE (LAMBDA = 1.00)
45	0.121820+03	0.1277460+03	0.1336720+03	0.1300000+03	0.2253610+01	1.73	1.00	0.3023030+01
46	0.1289240+03	0.1366020+03	0.1442790+03	0.1403000+03	0.3398340+01	2.43	0.82	0.3917160+01
47	0.9682390+02	0.1070300+03	0.1172370+03				1.13	0.5207370+01
48	0.1250720+03	0.1379660+03	0.1508610+03				1.33	0.6578900+01

B.N



DATING PARAMETERS INCONSISTENT - NORMAL NUMBERING FACILITY USED

PREVIOUS VALUES -
 FORECASTS X
 UPPER LIMITS +
 LOWER LIMITS *

FORECAST ORIGIN T = 45

8.15

FORECAST FOR POINT T	LOWER LIMIT 95%	FORECAST	UPPER LIMIT 95%	ACTUAL VALUE	FORECAST ERROR	PER CENT FORECAST ERROR	PSI WEIGHT	S.D. OF FORECAST ERROR OF TRANSFORMED VARIABLE (LAMBDA = 1.00)
46	0.132533D+03	0.138459D+03	0.144384D+03	0.140000D+03	0.154136D+01	1.10	1.00	0.302303D+01
47	0.101910D+03	0.109588D+03	0.117266D+03				0.82	0.391716D+01
48	0.130757D+03	0.140963D+03	0.151170D+03				1.13	0.520737D+01
49	0.123179D+03	0.136073D+03	0.148968D+03				1.33	0.657890D+01