Föreläsning 6
ARIMA-modellering

Kapitel 5

ARIMA

Auto Regressive Integrated Moving Average

Modellen består av tre delar

Auto Regressive AR

Integrated   I (har med differentiering att göra)

Moving Average   MA (modellering av feltermen)

Autoregressiva modeller

Istället för att tvinga in ett antal bestämda komponenter (trend, säsong, cyklisk komponent) kan man låta tidsserien successivt ”förklara sig själv”:

Den enklaste modellen är en AR(1):

$$y_t = \delta + \phi_1 y_{t-1} + a_t$$

- a_t är normalfördelad slumpterm (vitt brus) med väntevärde 0 och konstant varians. Ej observerbar
- Värdet i tidpunkt *t* antas främst bero av värdet vid tidigare tidpunkter och ev. av motsvarande säsongstidpunkt föregående år.
- Parametrarna kan ibland skattas med Minsta-Kvadrat-metoden (som i regression), men vanligare är Maximum-Likelihood-metoden.

AR(1): $y_t = \delta + \phi_1 y_{t-1} + a_t$

Denna modell kan skrivas med bakåtshiftoperatorer *B*

$By_t = y_{t-1}$

$$y_t = \delta + \phi_1 y_{t-1} + a_t$$

$$y_t = \delta + \phi_1 B y_t + a_t$$

$$y_t - \phi_1 B y_t = \delta + a_t$$

$$(1 - \phi_1 B) y_t = \delta + a_t$$

$$\Phi_1(B) y_t = \delta + a_t$$

$$\Phi_1(B) = (1 - \phi_1 B)$$

Enkel och dubbel exponentiell utjämning är specialfall av autoregressiva modeller. Dvs. autoregressiva modeller kan beskriva tidsserier med komplexare samband mellan historiska och framtida observationer.

För att kunna anpassa en autoregressiv modell krävs att tidsserien är *stationär*.

Arbetsgång vid modellering av ARIMA

- Behöver tidsserien transformeras för att y_t ska ha konstant varians och vara normalfördelad?
- Är tidsserien y_t stationär? Differentiera tills stationäritet nås.
- Studera SAC och SPAC för att bestämma lämplig ARIMA-modell
- Anpassa vald modell och utvärdera med residualanalys.
- Studera bl a SAC på reisualerna (som vanligt)
- Kan man se i SAC för residualerna att något mer behöver modelleras?
- Gör prognoser

Stationäritet Kap 2,3 (Från Föreläsning 1)

En tidsserie är stationär om följande gäller:

- Tidsserien ska ha ett konstant väntevärde över tiden (bortsett från säsongssvängningar).
- Tidsserien ska ha en konstant varians över tiden.
- Korrelationen mellan två tidpunkter i serien får endast bero på avståndet mellan dessa punkter, och inte på <u>var</u> i tiden de ligger.

Differentiering kap 2,4 sid 50→ (Från Föreläsning 1)

Om tidsserien inte är stationär så kan tidsserien differentieras tills den blir stationär.

- Den stationära tidsserien betecknas z_t.
- Originaltidsserien betecknas y_t
- Andra beteckningar på en tidsserie kan vara w_t v_t osv

(Från Föreläsning 1 men även med B)

Om trend ses i tidsserien:

Differentiera originalserien med 1 tidsförskjutning

$z_t \;=\; y_t - y_{t-1} = (1 - B)y_t =$ Diff1

Om y_t uppvisar en kvadratisk trend differentierar man två gånger, dvs

$w_t \;=\; y_t - y_{t-1} = (1 - B)y_t$ och sedan

$z_t \;= w_t - w_{t-1} = (1 - B)(1 - B)y_t = \;(1 - B)^2 y_t =$ Diff1,1

(Från Föreläsning 1)

Om kraftig säsongsvariation finns så kan vi differentiera för säsong:

$$z_t = y_t - y_{t-L} = \left(1 - B^L\right)y_t = \mathrm{Diff}L$$

där L är säsongslängden

Månadsdata: $z_t = y_t - y_{t-12} = (1 - B^{12})y_t = \mathrm{Diff12}$

Kvartalsdata: $z_t = y_t - y_{t-4} = (1 - B^4)y_t = \mathrm{Diff4}$

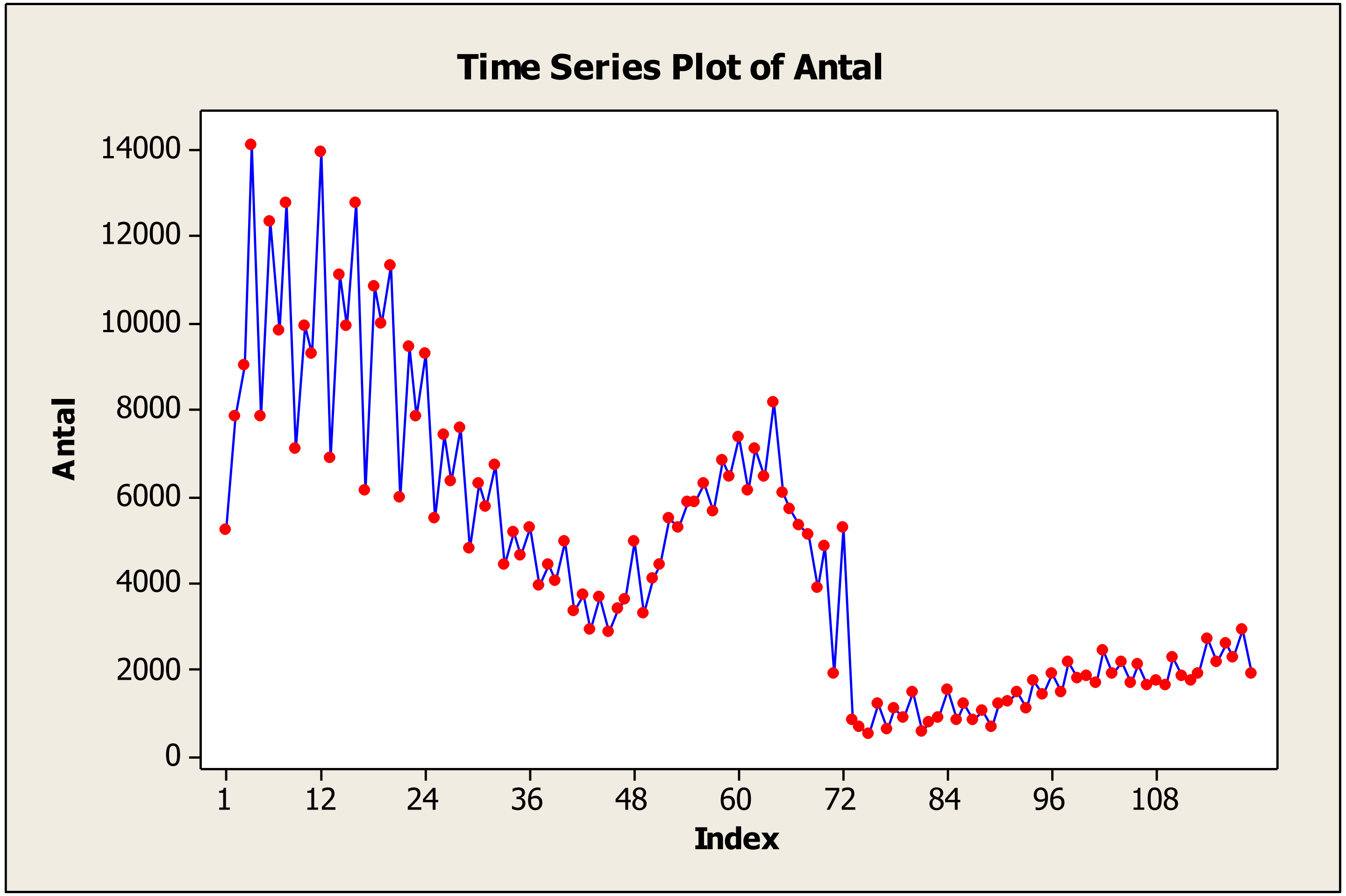
Vi kanske måste differentiera både för trend och säsong:

$$w_t = y_t - y_{t-1} = (1 - \mathrm{B})y_t = \mathrm{Diff1}$$

$$z_t = w_t - w_{t-L} = \left(1 - B^L\right)w_t = \left(1 - B^L\right)(1 - \mathrm{B})y_t = \mathrm{Diff1, L}$$

$\mathrm{DiffL, 1}$ ger samma resultat

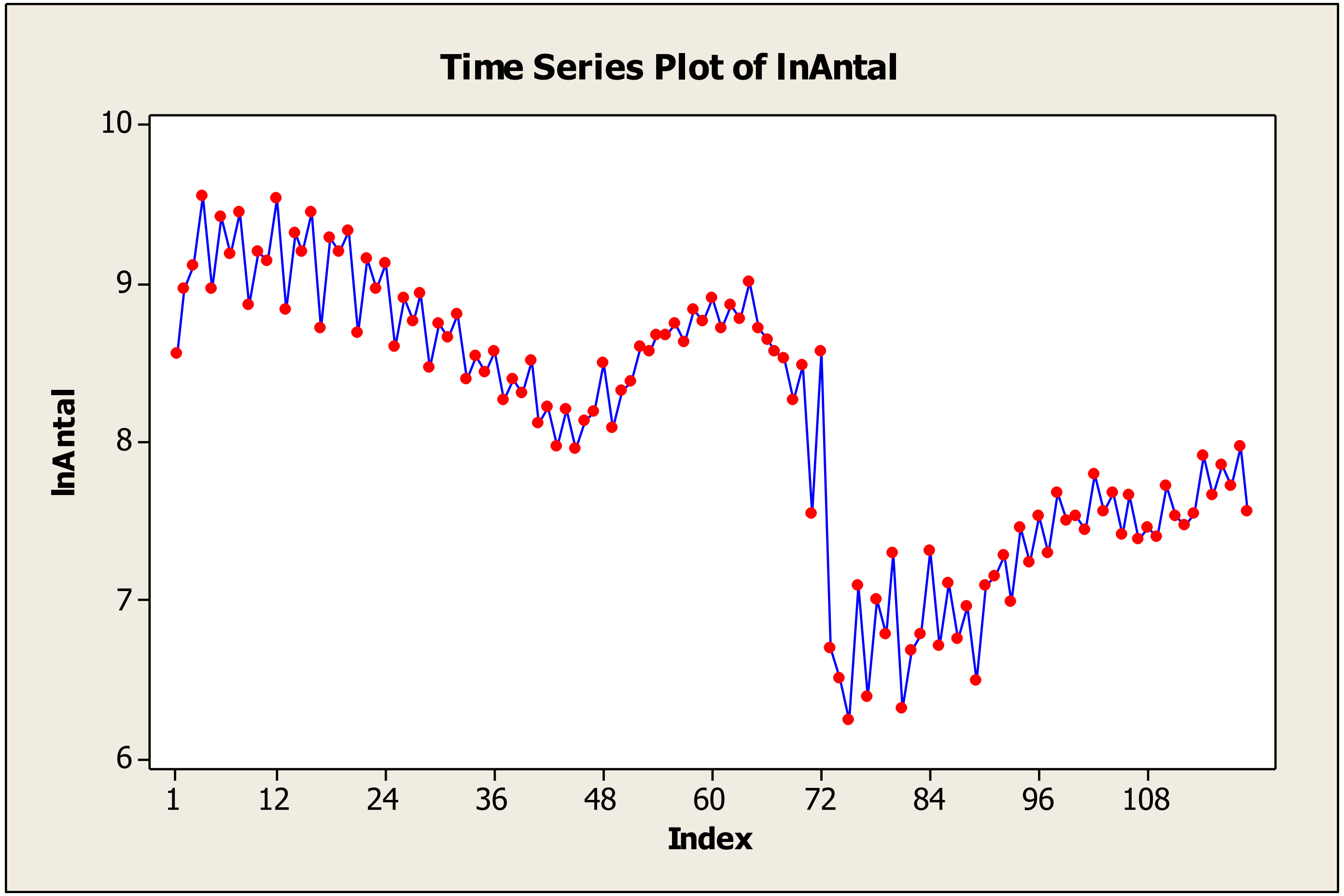
Exempel: Antal nybyggnationer Kvartalsdata kv 1 1975 till kv 3 2004



Först måste vi undersöka om tidsserien har konstant varians.

Om den inte har det så måste vi transformera data så att konstant varians uppnås.

Antal nybyggnationer logaritmeras. lnAntal



SAC och SPAC

SAC är den faktiska autokorrelationen.

SPAC = Sample Partial AutoCorrelation, är en ’renare’ autokorrelation.

Båda är mycket viktiga för att välja rätt modell.

Tidsserien bör vara stationär när man studerar SPAC

SAC och SPAC

SAC: $\widehat{\mathrm{Korr}}[z_t, z_{t+k}] = r_k = \frac{\sum_{t=1}^{n-k}(z_t - \bar{z})(z_{t+k} - \bar{z})}{\sum_{t=1}^{n}(z_t - \bar{z})^2}$

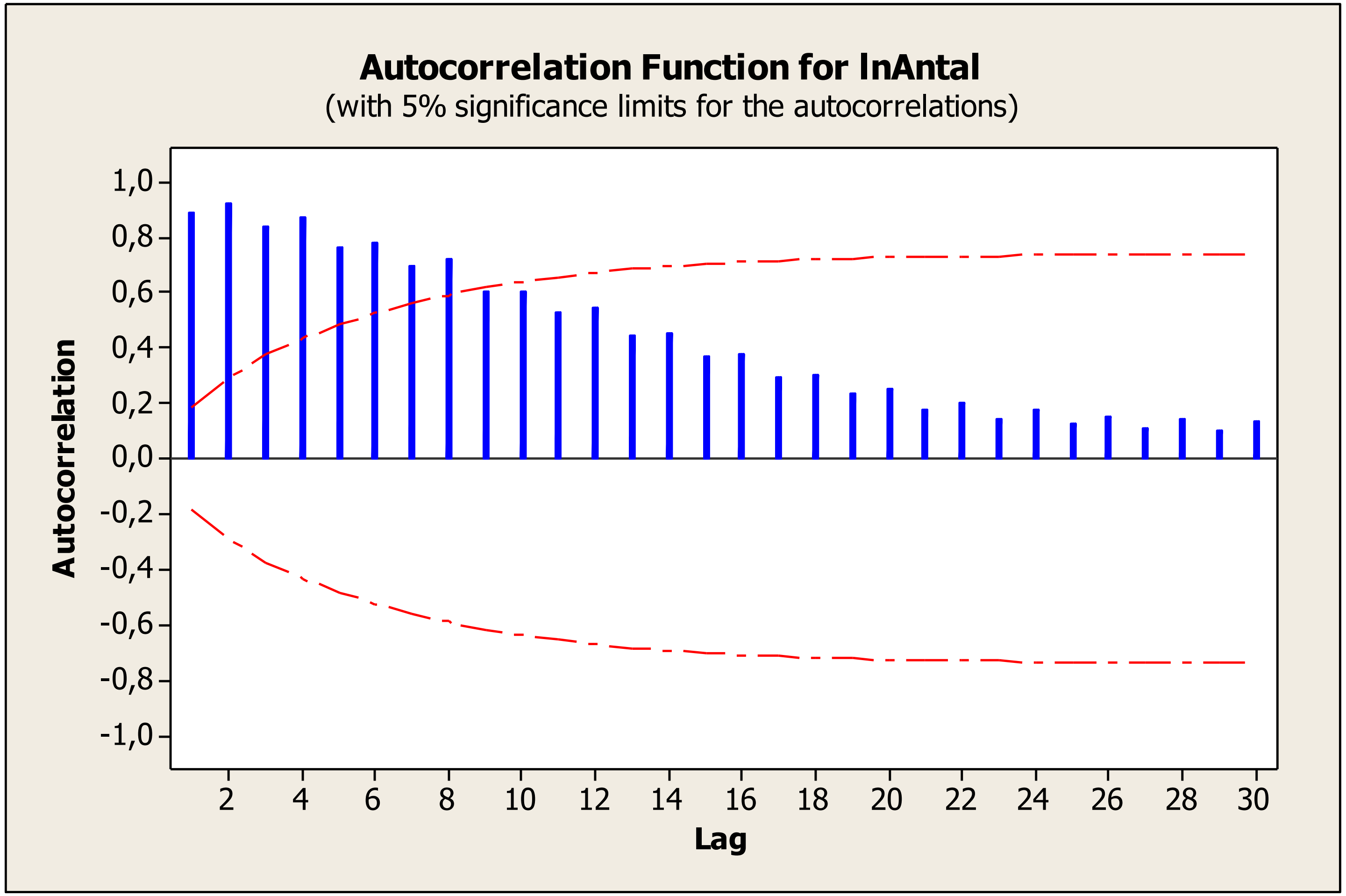
SPAC: $\widehat{\mathrm{Korr}}[z_t, z_{t+k} | z_{t+1}, \ldots, z_{t+k-1}] = r_{kk}$

$$r_{kk} = \begin{cases} r_1 & \text{om } k = 1 \\ \frac{r_k - \sum_{j=1}^{k-1} r_{k-1,j} r_{k-j}}{1 - \sum_{j=1}^{k-1} r_{k-1,j} r_j} & \text{om k} = 2,3, \ldots \end{cases}$$

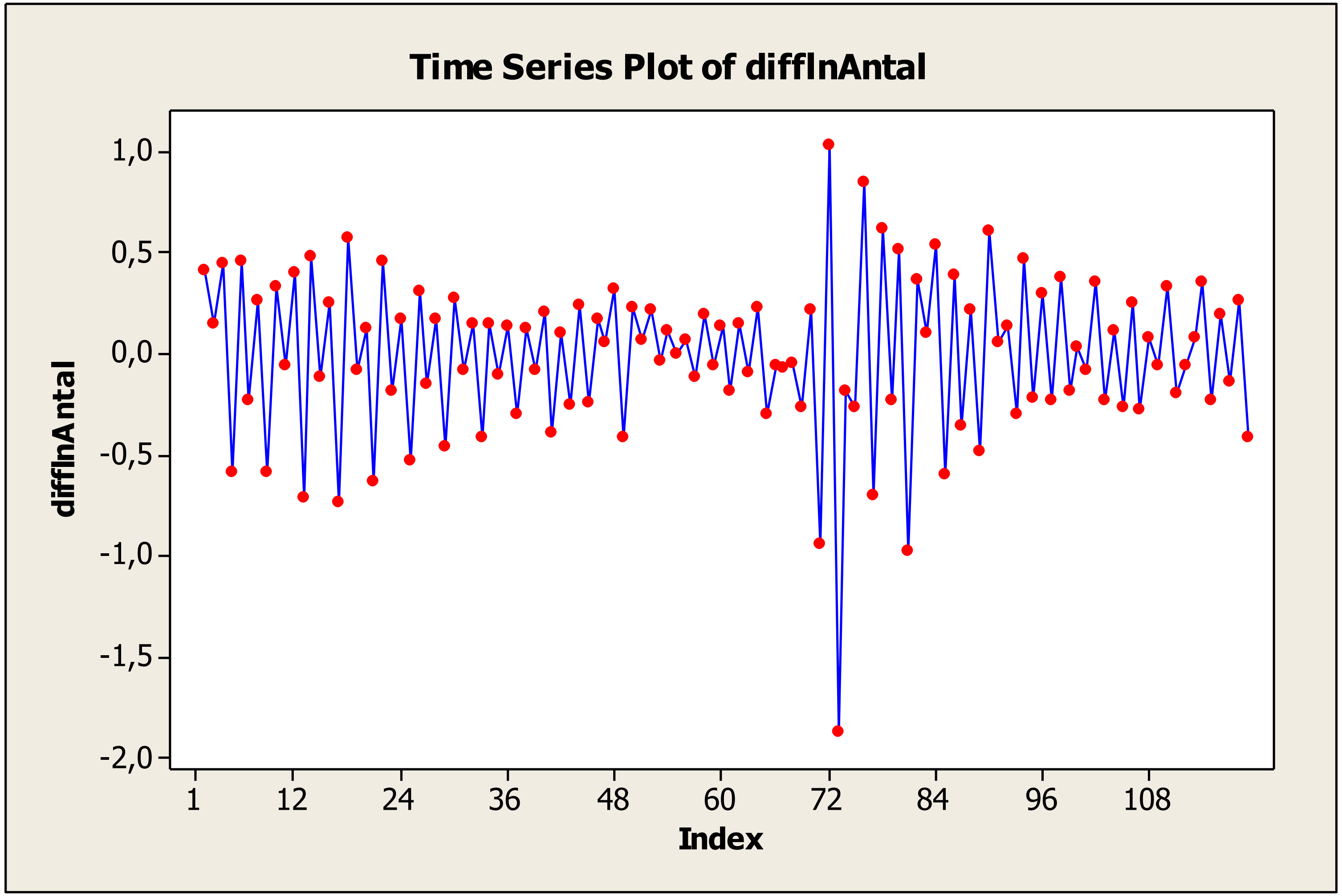
där $r_{kj} = r_{k-1,j} - r_{kk} r_{k-1,k-j} \qquad j = 1,2, \ldots, k-1$

SAC på lnAntal. Ej stationär. Avtar ganska långsamt.

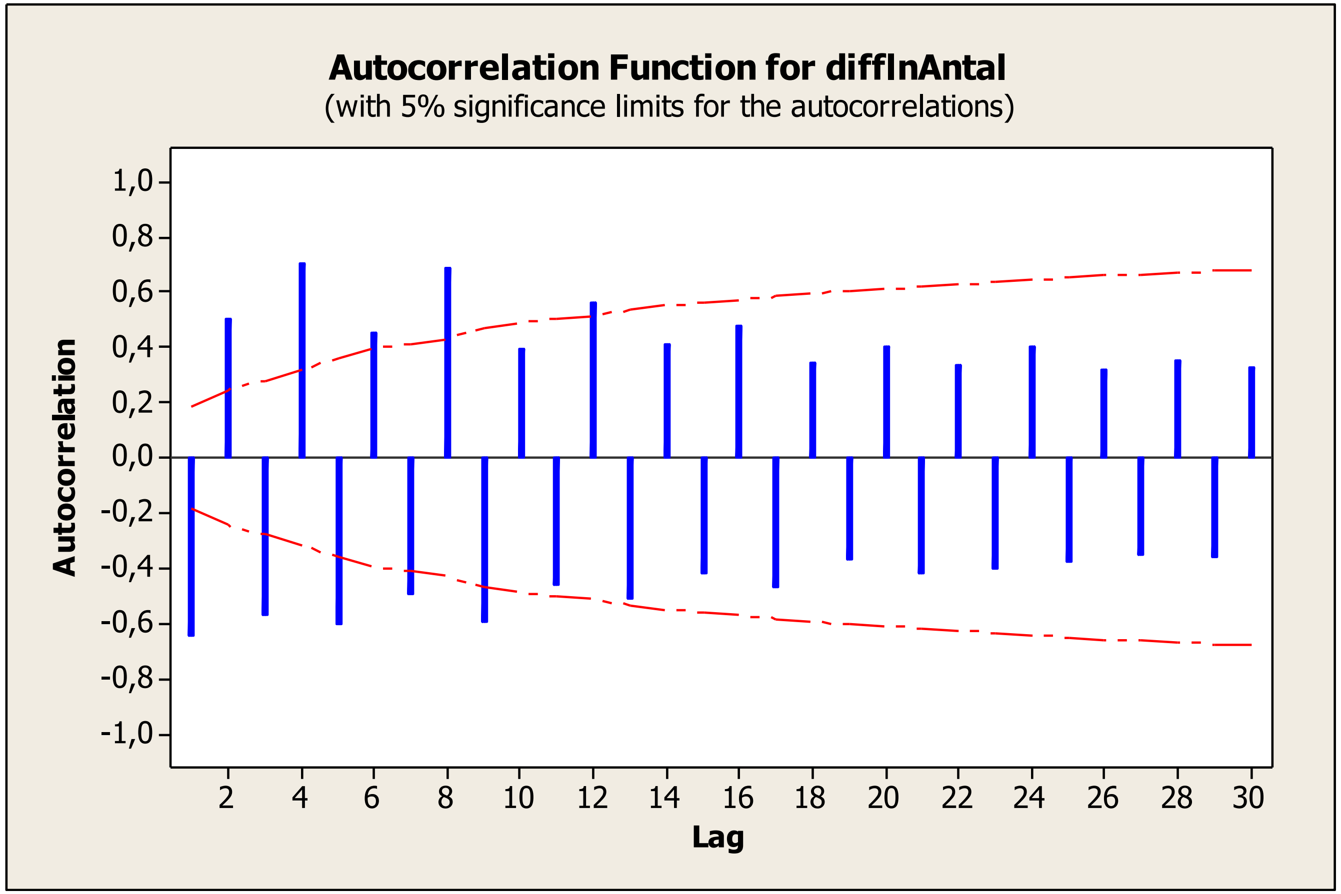
Det är alltså i SAC vi ser om vi har stationäritet. Även i tidsseriegrafen
Vid stationäritet avtar SAC relativt snabbt.



Differentiera lnAntal

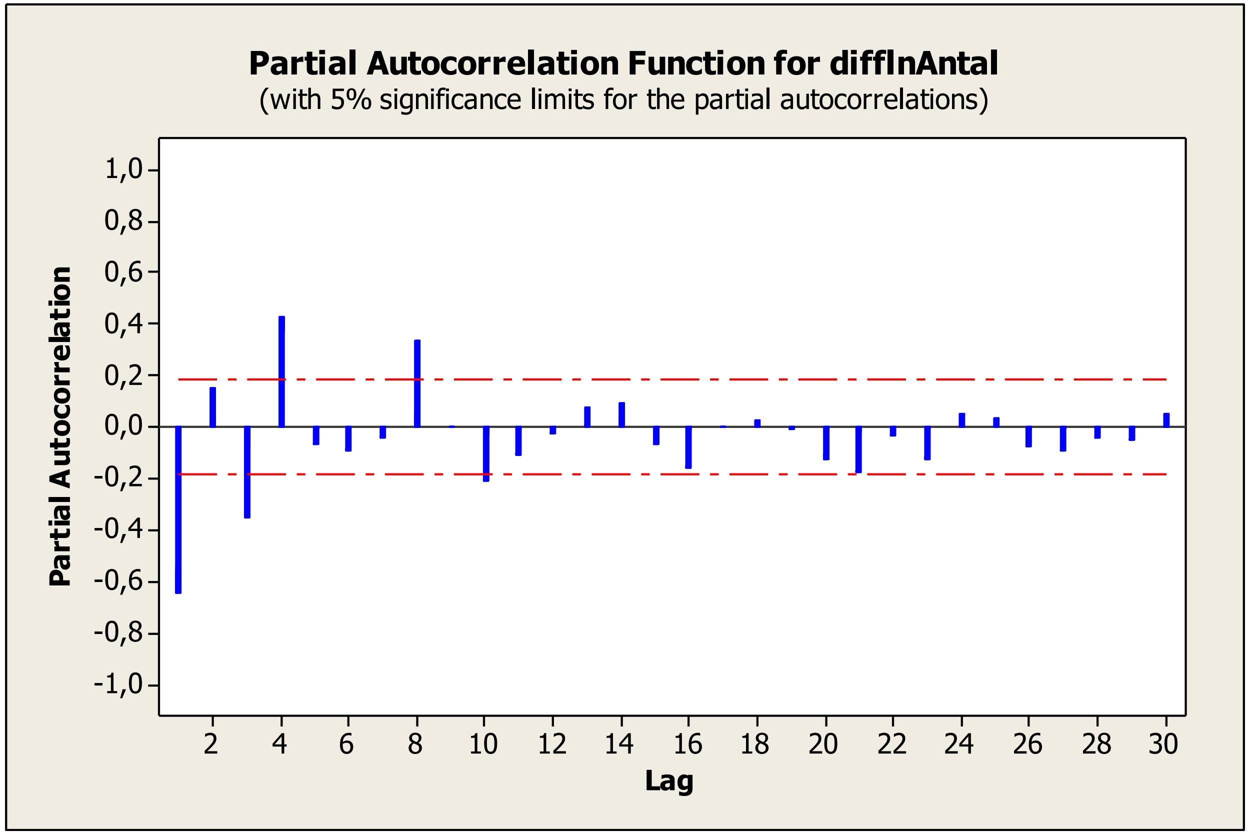


SAC på Diff1=difflnAntal. Avtar ganska långsamt.



Serien tycks ändå inte bli stationär. Studera SPAC

SPAC på diff1=difflnAntal

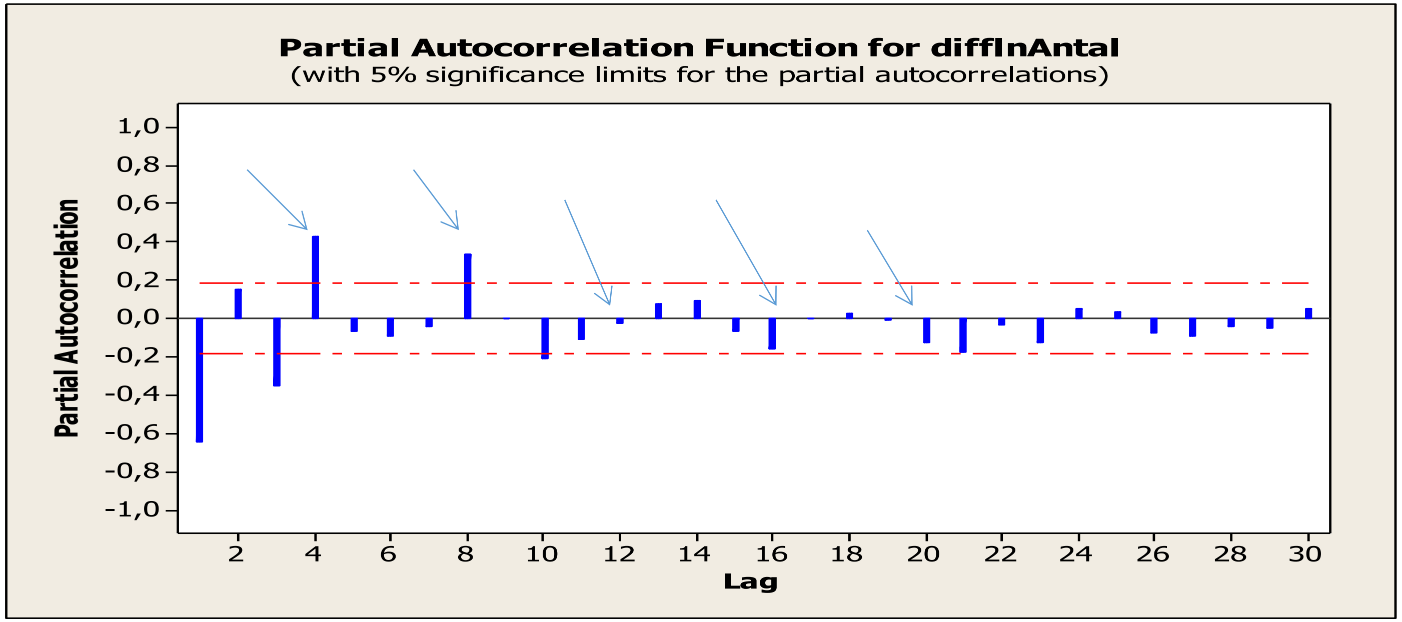


SPAC ger en ’Renare’ autokorrelation.

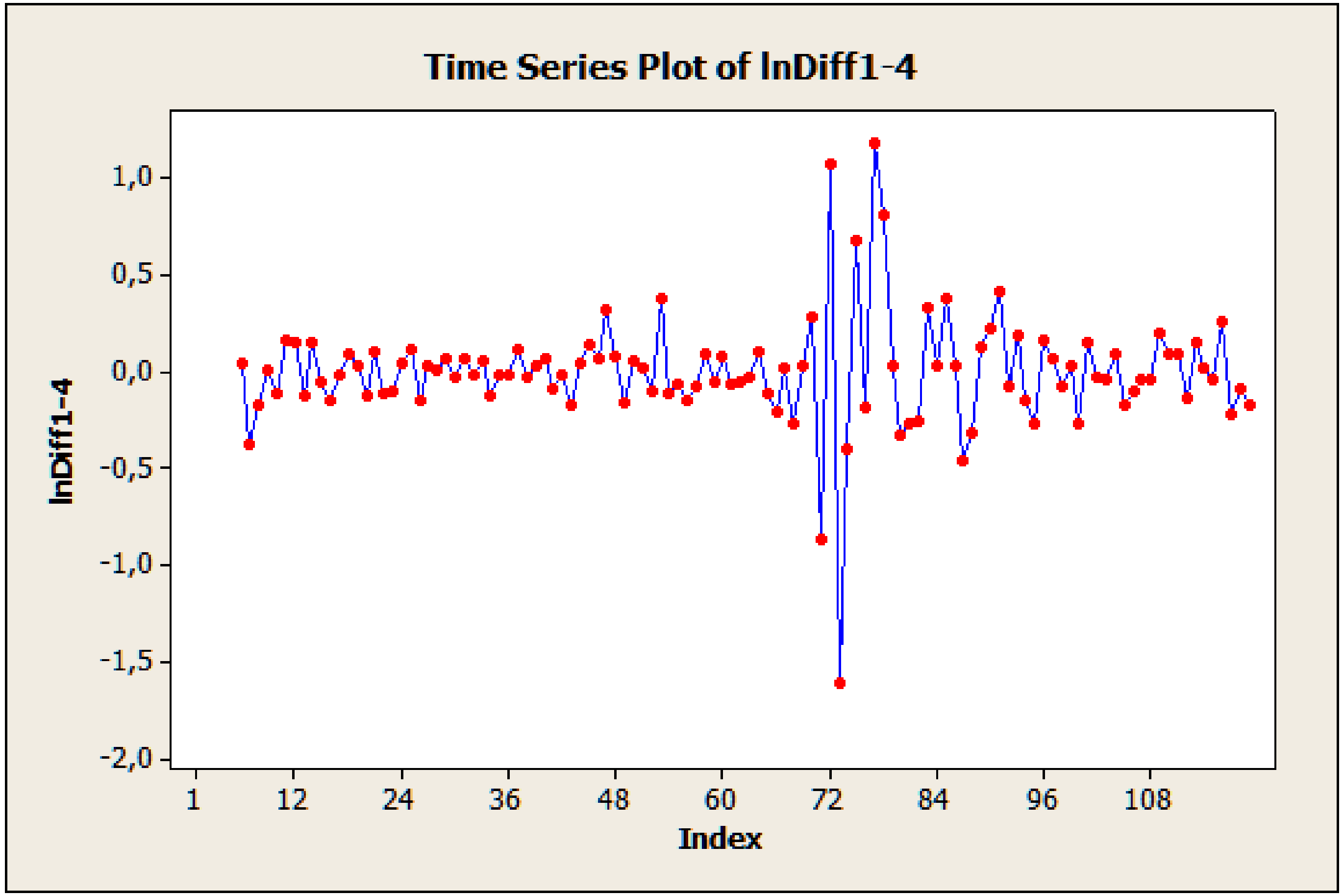
Tolkning SPAC

Vi ser en spik på lag 4 och lag 8. Detta påminner oss om att vi har kvartalsdata.

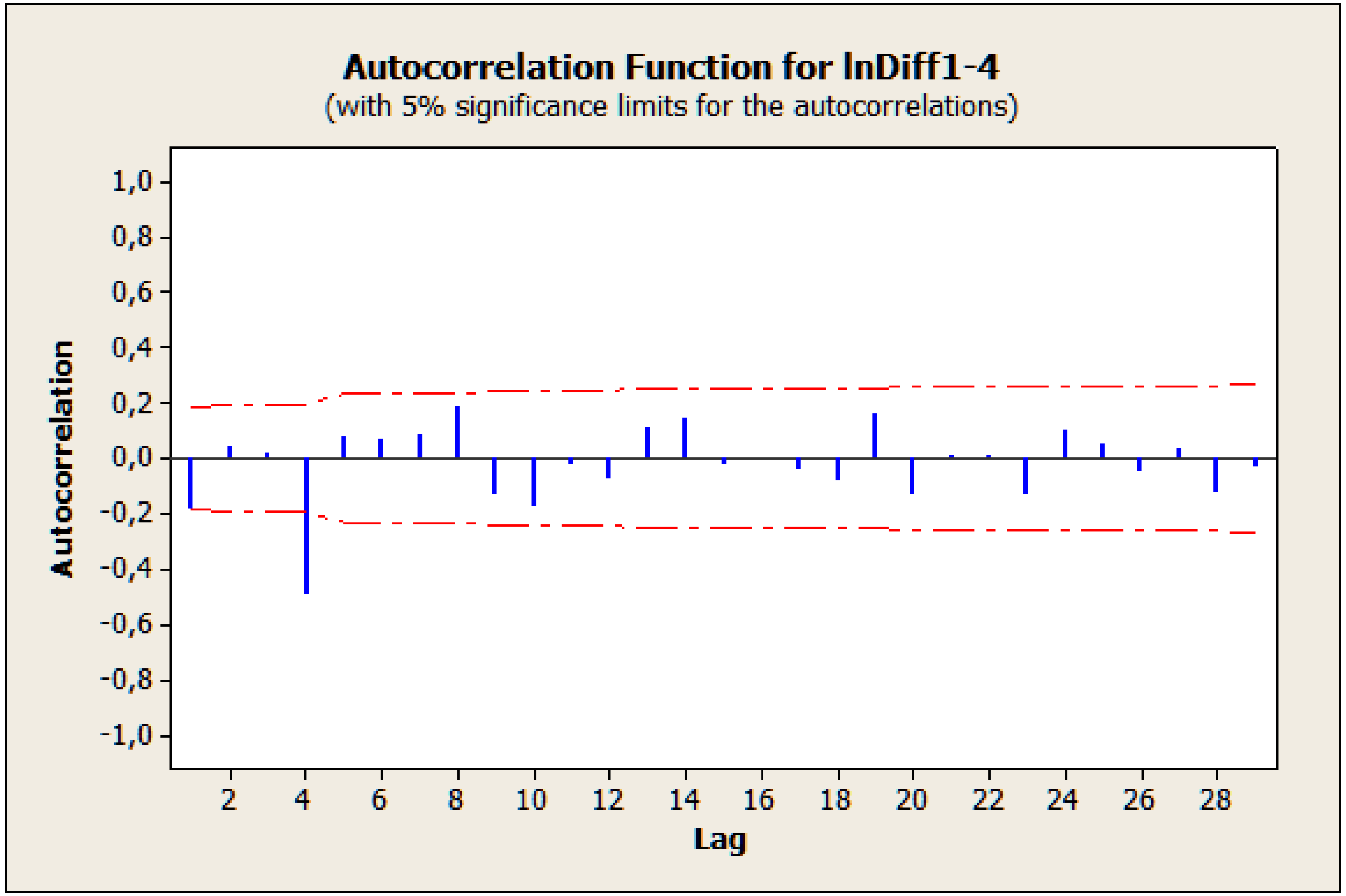
Vi kan nu differentiera, diffad lnAntal för säsong.



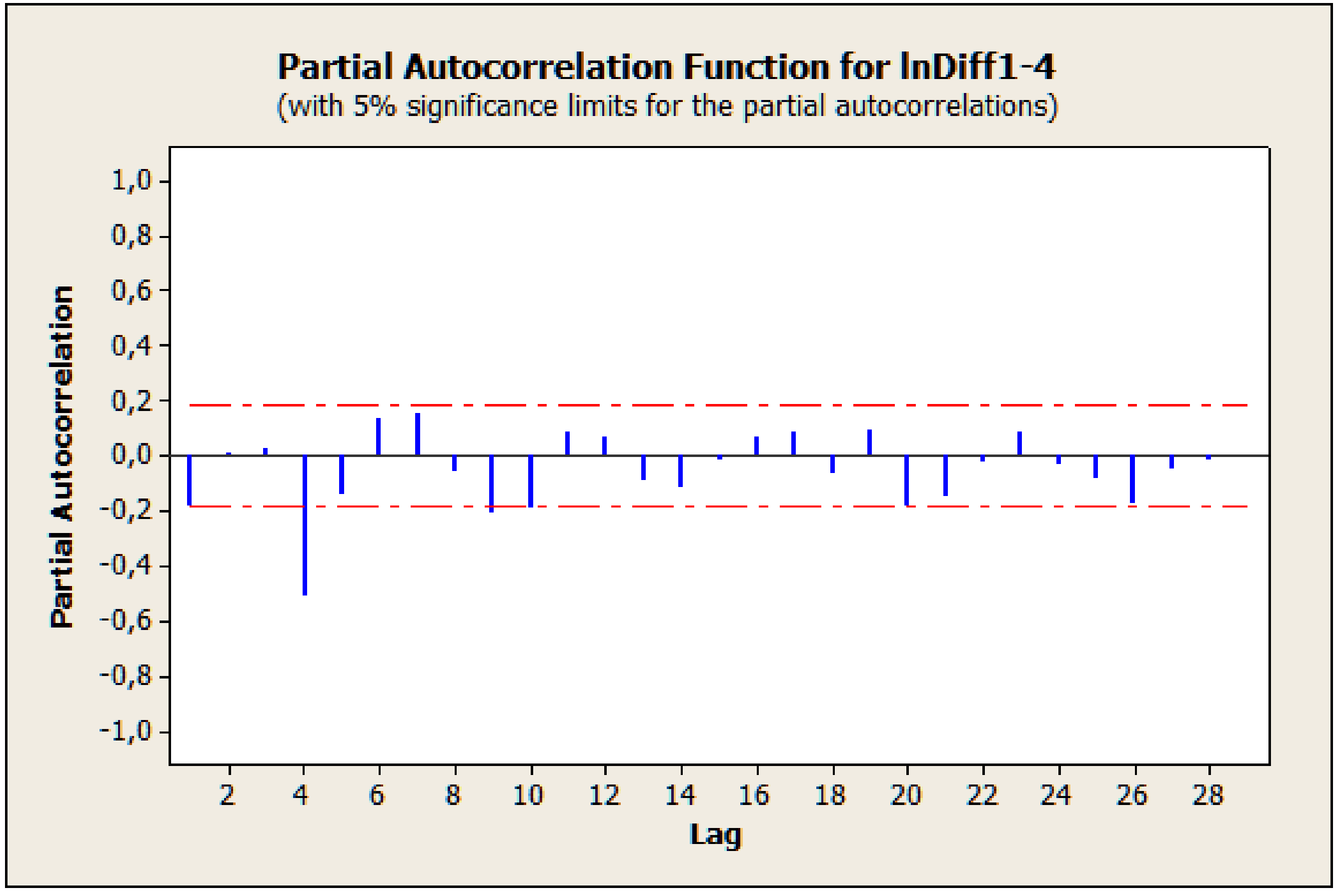
Så nedanstående graf är Diff1,4 på lnAntal



SAC på lnDiff1,4. STATIONÄR!! Yes ☺



SPAC på lnDiff1,4



Modellbestämning

Med hjälp av SAC och SPAC för den stationära tidsserien, ska vi nu gå vidare och bestämma en ARMA-modell.

Det är viktigt att inte differentiera i onödan. Vi vill INTE få bort ALLA spikar i SAC och SPAC.

ARIMA-modell

Auto **R**egressive **I**ntegrated **M**oving **A**verge modell ARIMA(p,d,q)

Låt y_t , $t = 1,2, \ldots , n$ vara tidsserien med n värden.

Låt vidare z_t vara den stationära tidsserien. Dvs y_t har differentierats för säsong och trend om det behövs.

AR modell med ordning p = AR(p):

$$z_t = \delta + \phi_1 z_{t-1} + \phi_2 z_{t-2} + \ldots + \phi_p z_{t-p} + a_t$$

a_t är **vitt brus** (=slumpterm). Alla a_t är oberoende och normalfördelade med väntevärde 0 och konstant varians.

a_t kan inte observeras

AR modell med ordning $p \;=\; AR(p)$:

$$z_t = \delta + \phi_1 z_{t-1} + \phi_2 z_{t-2} + \ldots + \phi_p z_{t-p} + a_t$$

Med bakåtskiftoperatorer *B*

$z_t - \phi_1 z_{t-1} - \phi_2 z_{t-2} - \ldots - \phi_p z_{t-p} = \delta + a_t$

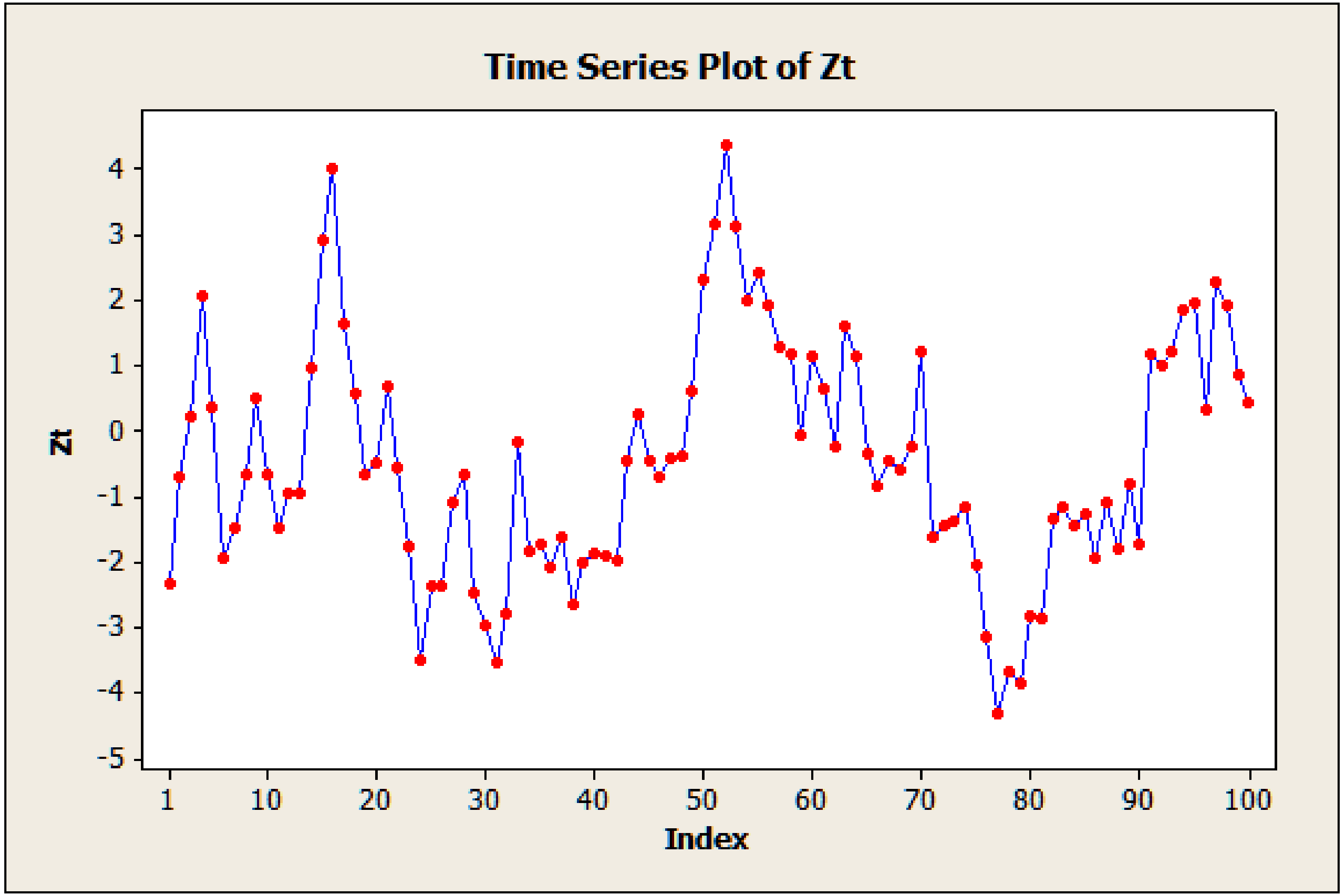
$(z_t - \phi_1 B z_t - \phi_2 B^2 z_t - \ldots - \phi_p B^p z_t) = \delta + a_t$

$(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)\, z_t = \delta + a_t$

$$\Phi_p(B) z_t = \delta + a_t$$

$\Phi_p(B) = (1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)$

Simulerad tidsserie AR(1)



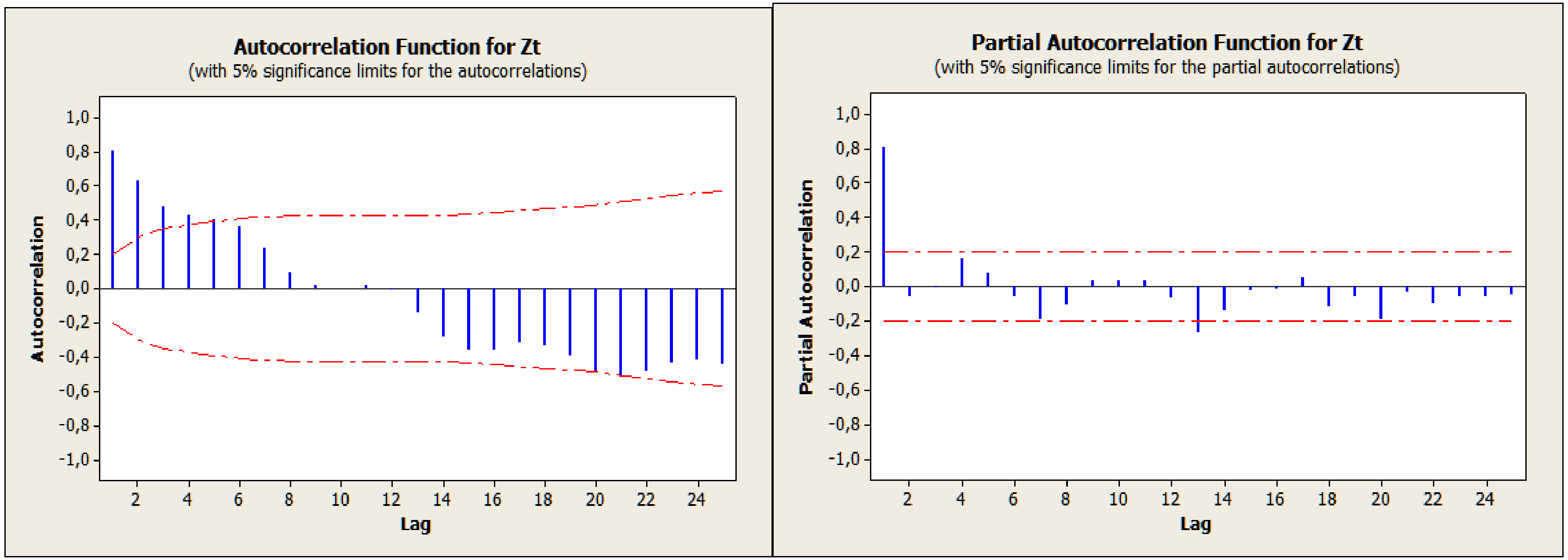
Val av AR-modell

Studera SAC och SPAC för z_t

SAC ska ha ett ganska snabbt avtagande mönster och

SPAC ska ha p spikar om AR(p).

Ex: AR(1)



Simulerad tidsserie

Den tidsserie som har simulerats är

$$z_t = 0{,}8z_{t-1} + a_t$$

där a_t är $N(0{,}1)$ och oberoende

$$t = 1{,}2, \dots , 100$$

Enkelt att simulera då det är en rekursionsformel som man adderar normalfördelade slumpvärden till.

Teori om AR(1)

AR(1): $z_t = \delta + \phi_1 z_{t-1} + a_t$

$$E[z_t] = \frac{\delta}{1 - \phi_1}$$

Teoretisk AutoCorrelation TAC

$$\rho(k) = Korr[z_t, z_{t+k}] = Korr[z_t, z_{t-k}] = \rho(-k)$$

$$\rho(k) = \phi_1^k \qquad k = 0,1,2 \ldots$$

TAC avtar ganska snabbt. När vi studerar SAC så letar vi efter mönster som finns I TAC

Se bl a figure 5,6 sid 341

TPAC $\neq 0$ då $k = 1$ och noll annars Teoretisk Partiell AutoCorrelation

Teori om AR(2)

AR(2): $z_t = \delta + \phi_1 z_{t-1} + \phi_2 z_{t-2} + a_t$

$$E[z_t] = \frac{\delta}{1 - \phi_1 - \phi_2}$$

Teoretisk AutoCorrelation TAC

$\rho(1) = \frac{\phi_1}{1-\phi_2}$ $\rho(2) = \phi_1\rho(1) + \phi_2$ $\rho(3) = \phi_1\rho(2) + \phi_2\rho(1)$ osv

TAC avtar ganska snabbt men kan även svänga.

Se bl a figure 5,7 och 5,8

TPAC $\neq 0$ då $k = 1, 2$ och noll annars

MA-modell

Medan en AR-modell ser ut som en viss typ av regressionsmodell, autoregression,

så är MA-modellen mycket svårbegriplig. Men den är ändå mycket användbar.

Man modellerar här z_t med endast vitt brus.

MA-modell med ordning *q* = $MA(q)$

$$z_t = \delta + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q}$$

MA-modell med ordning q = $MA(q)$

$$z_t = \delta + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q}$$

Med bakåtskiftoperatorer *B*

$$z_t = \delta + (a_t - \theta_1 B a_t - \theta_2 B^2 a_t - \ldots - \theta_q B^q a_t)$$

$$z_t = \delta + (1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q)\, a_t$$

$$z_t = \delta + \Theta_q(B) a_t$$

$$\Theta_q(B) = (1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q)$$

Val av MA-modell

Studera SAC och SPAC

SAC har q spikar och

SPAC avtar ganska snabbt.

Det är det omvända jämfört med en AR-modell.

Teori om MA(q)

$$z_t = \delta + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q}$$

Eftersom $E[a_t] = 0$ så är $E[z_t] = \delta$

Autokorrelationen mellan alla a_t är 0 så $\rho(k) \neq 0$ då $k \leq q$

ARMA-modell

Sätt ihop en AR(p) och en MA(q) så fås en ARMA(p,q)

$$z_t = \delta + \phi_1 z_{t-1} + \phi_2 z_{t-2} + \ldots + \phi_p z_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q}$$

SAC och SPAC är här en sammanblandning för vad som gäller för en AR(p) och en MA(q).

Man måste ofta pröva sig fram för att finna en bra modell.

Anpassa flera olika modeller och avgör vilken som är bäst.

ARMA(p,q)

$$z_t = \delta + \phi_1 z_{t-1} + \phi_2 z_{t-2} + \ldots + \phi_p z_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q}$$

Med bakåtshiftoperatorer

$$z_t - \phi_1 z_{t-1} - \phi_2 z_{t-2} - \ldots - \phi_p z_{t-p} = \delta + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q}$$

$$z_t - \phi_1 B z_t - \phi_2 B^2 z_t - \ldots - \phi_p B^p z_t = \delta + a_t - \theta_1 B a_t - \theta_2 B^2 a_t - \ldots - \theta_q B^q a_t$$

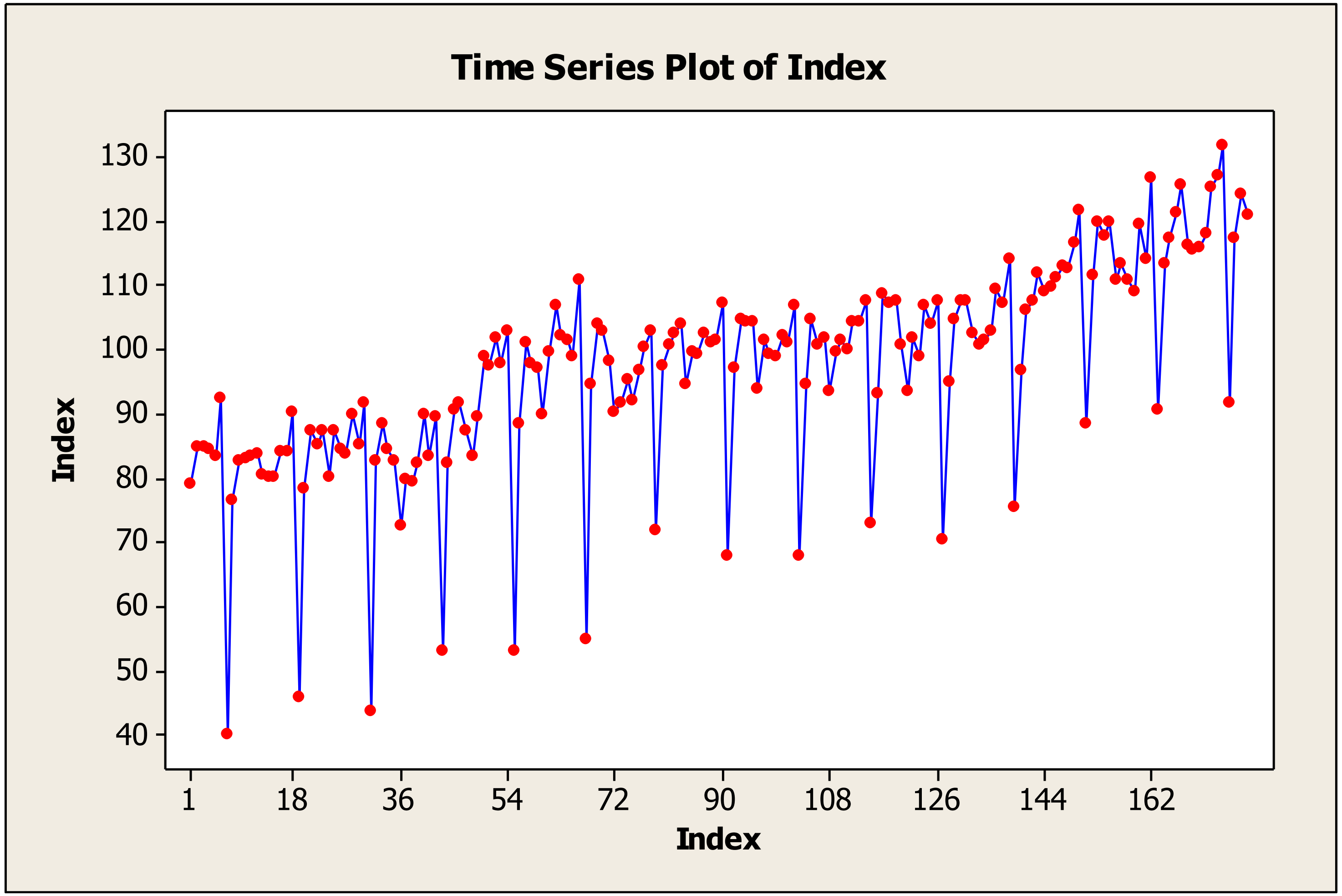
$$(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p) z_t = \delta + (1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q) a_t$$

$$\Phi_p(B) z_t = \delta + \Theta_q(B) a_t$$

$\Phi_p(B) = (1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)$

$\Theta_q(B) = (1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q)$

Ex: Pappersproduktion



Vi ska följa detta ex.

Som ses finns det säsongsvariation och detta kan vi modelleras med en ARIMA-modell. När säsong modelleras så kan den kallas SARIMA

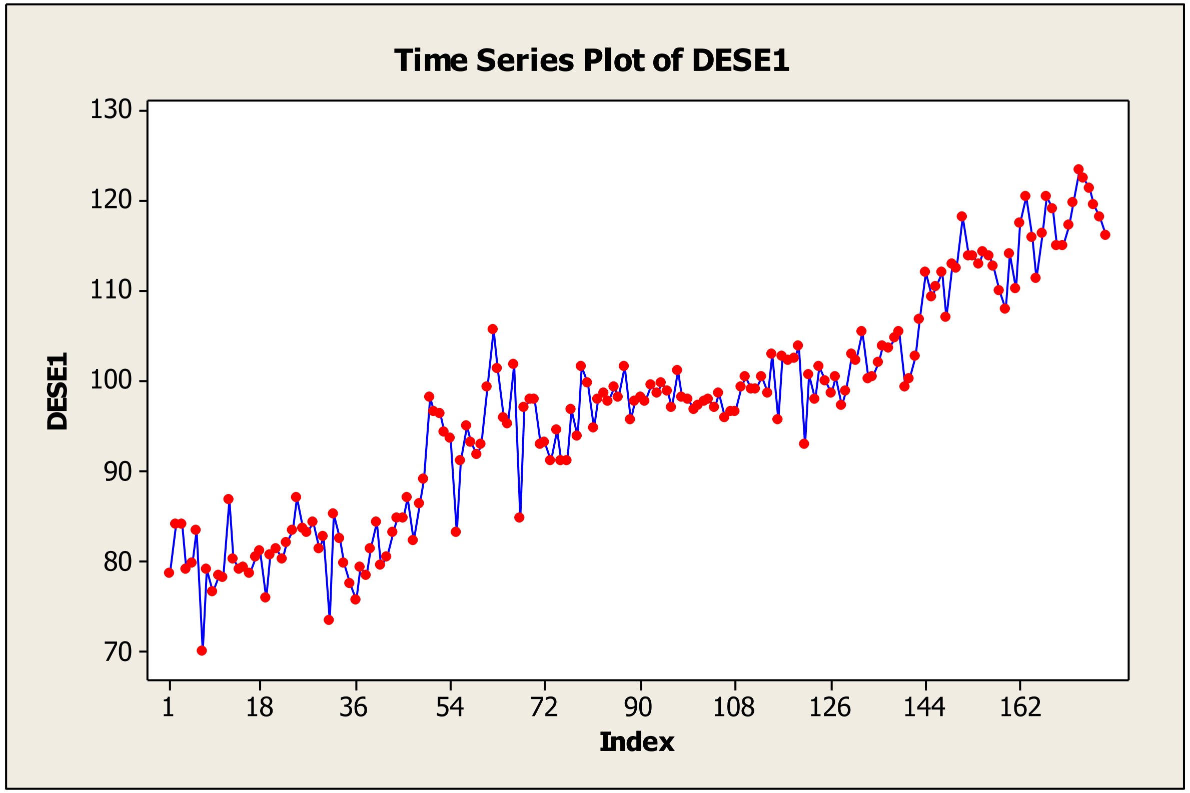
Att hitta en SARIMA-modell väntar vi med en stund.

På något sätt måste vi ta bort den kraftigaste säsongsvariationen. Vi prövar två olika sätt

1. Säsongsrensa tidsserien med klassisk komponentuppdelning
2. Differentiera för säsong

1. Säsongsrensa

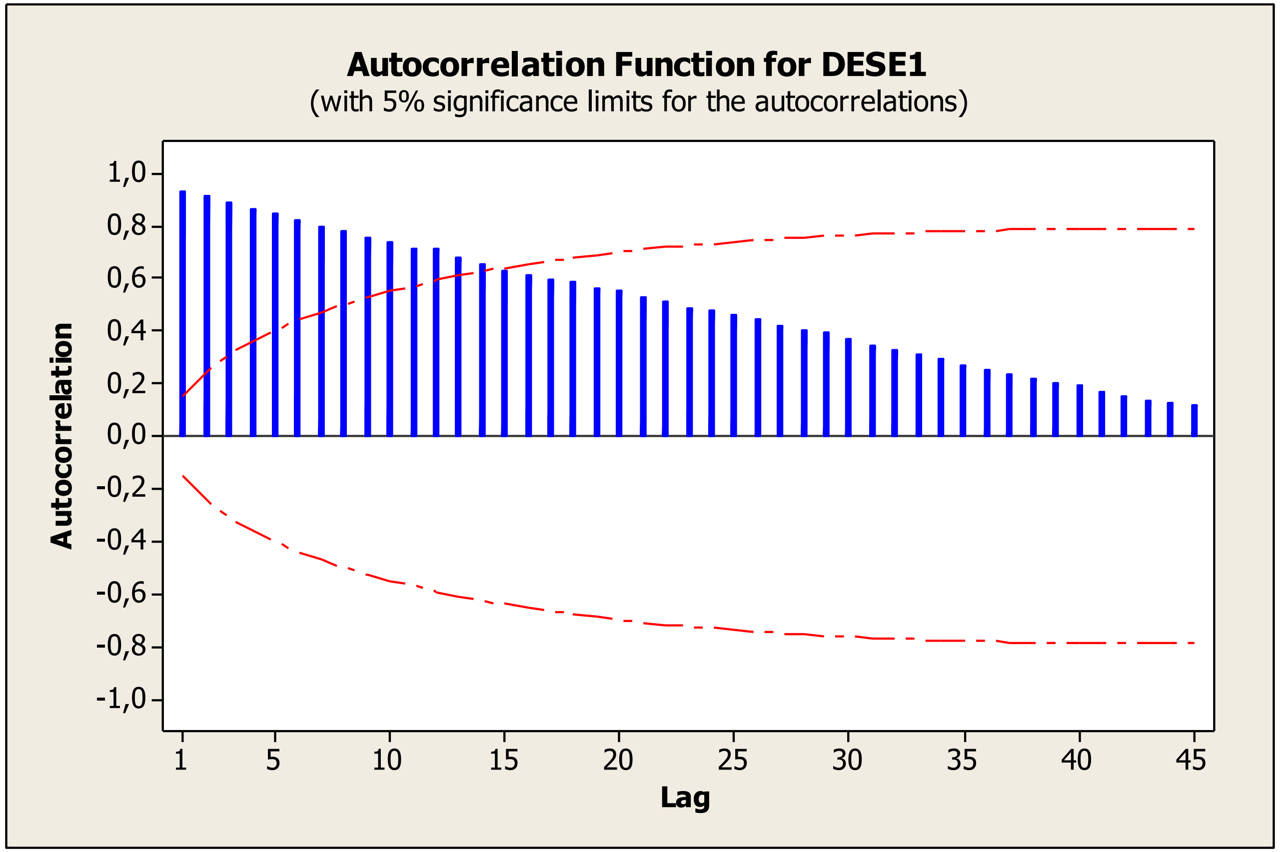
Säsongsrensad serie. Spara säsongskomponenterna.



Seasonal Indices

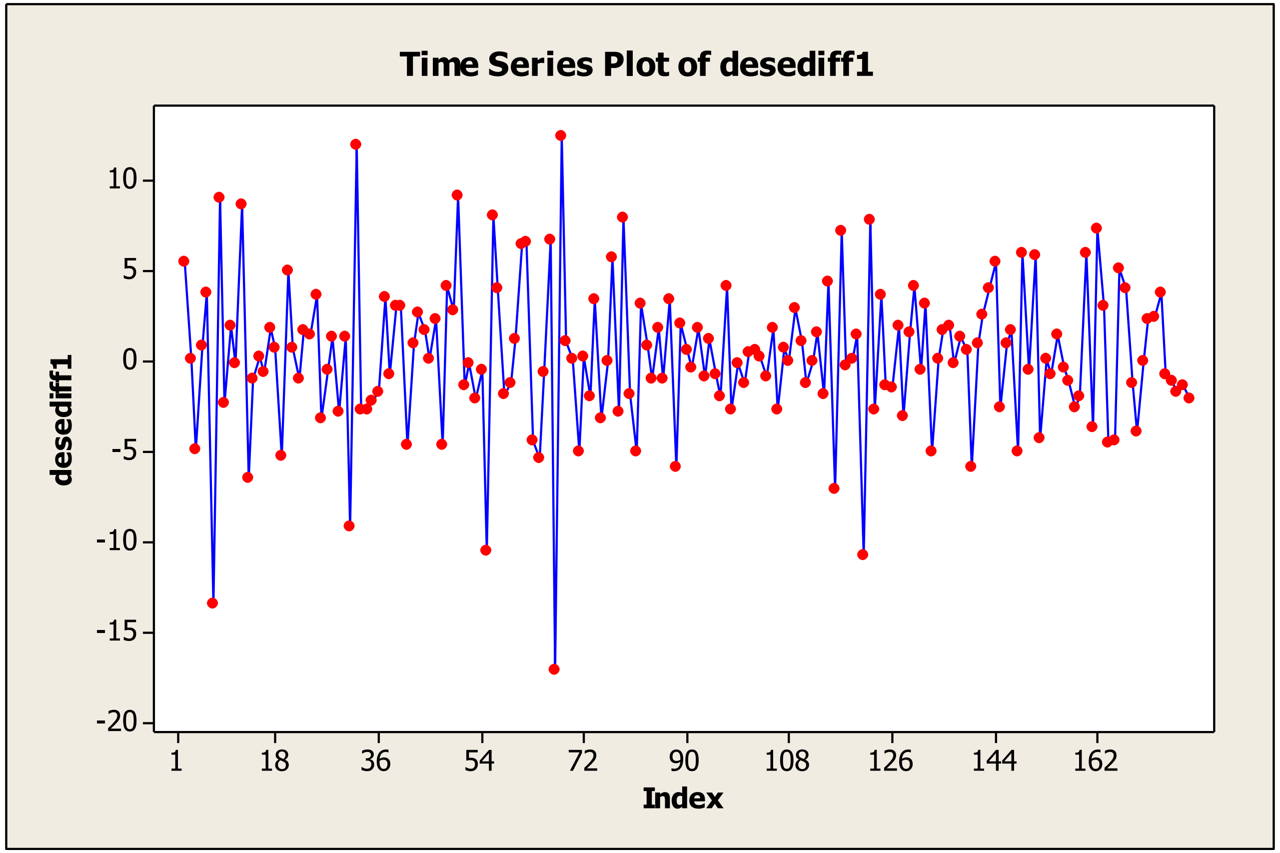
Period	Index
1	0,3795
2	0,8295
3	0,7441
4	5,3920
5	3,6420
6	8,9587
7	-29,9517
8	-2,5601
9	5,8753
10	4,6399
11	5,1753
12	-3,1247

SAC på DESE1

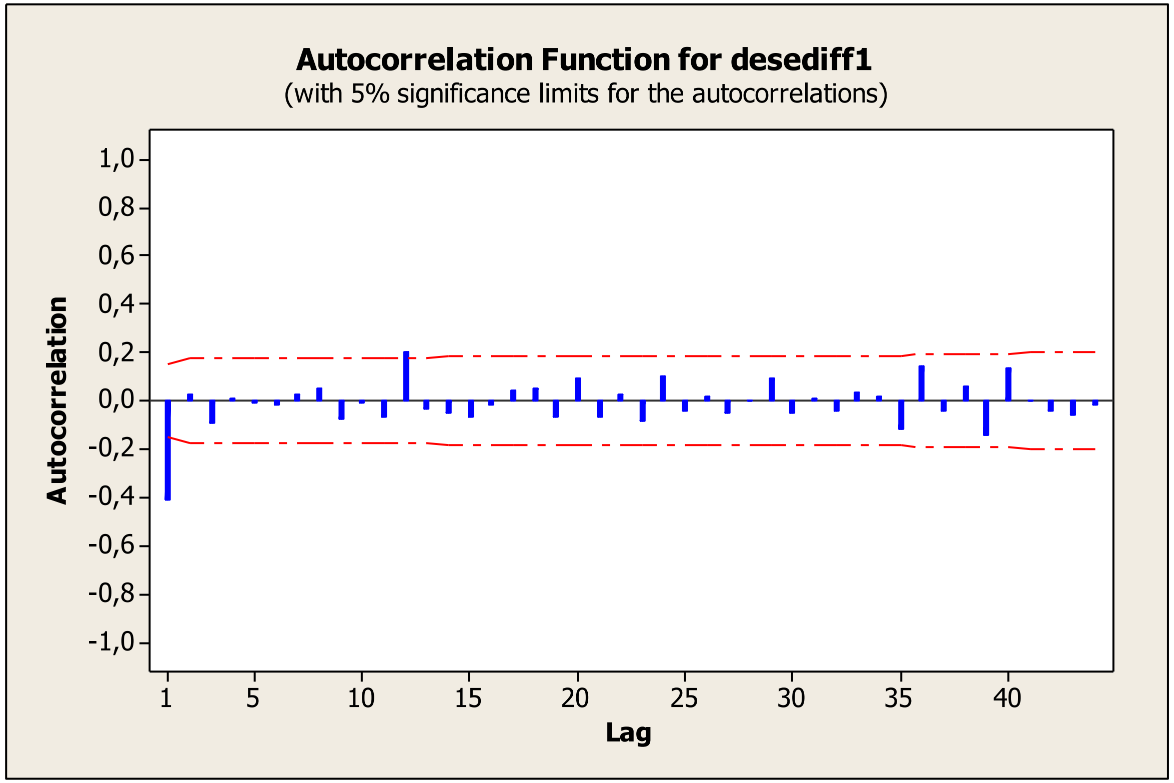


Inte stationär

Diffa för trend. desediff1

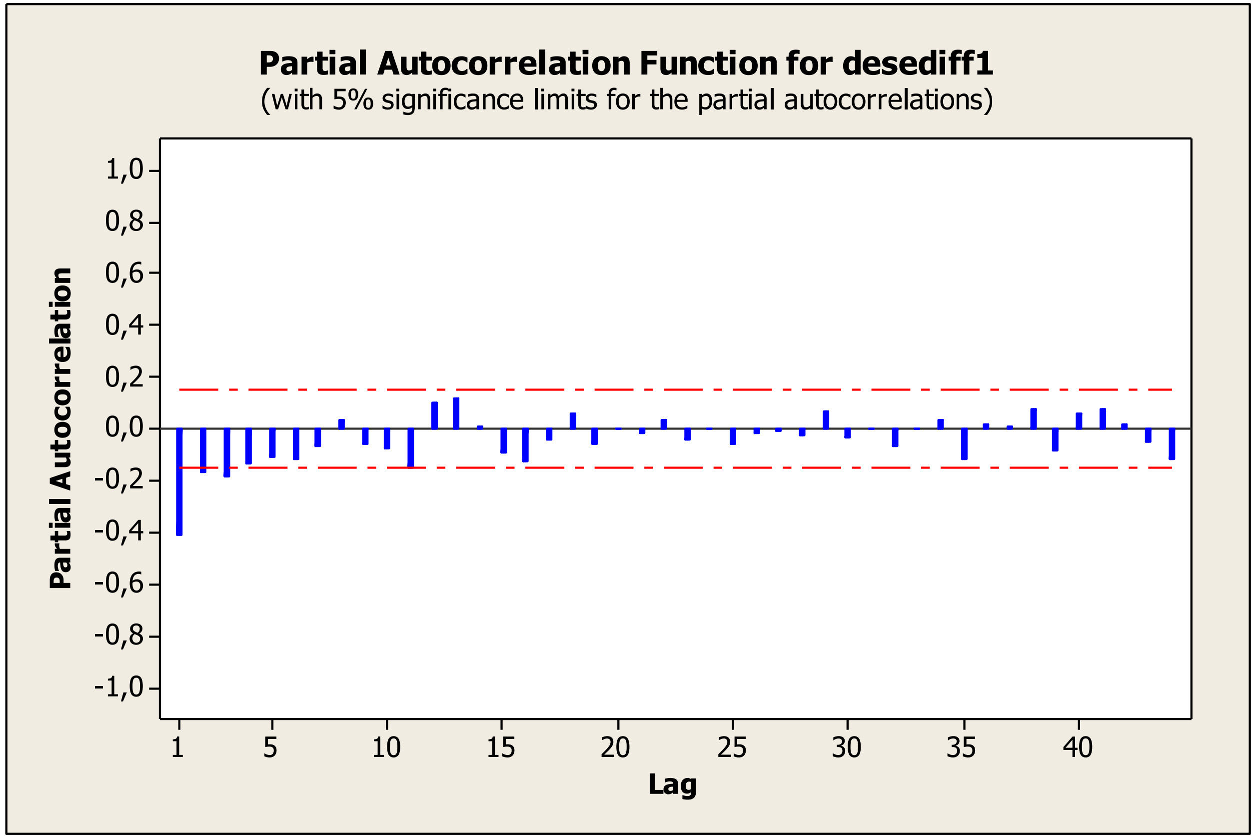


SAC på desediff1

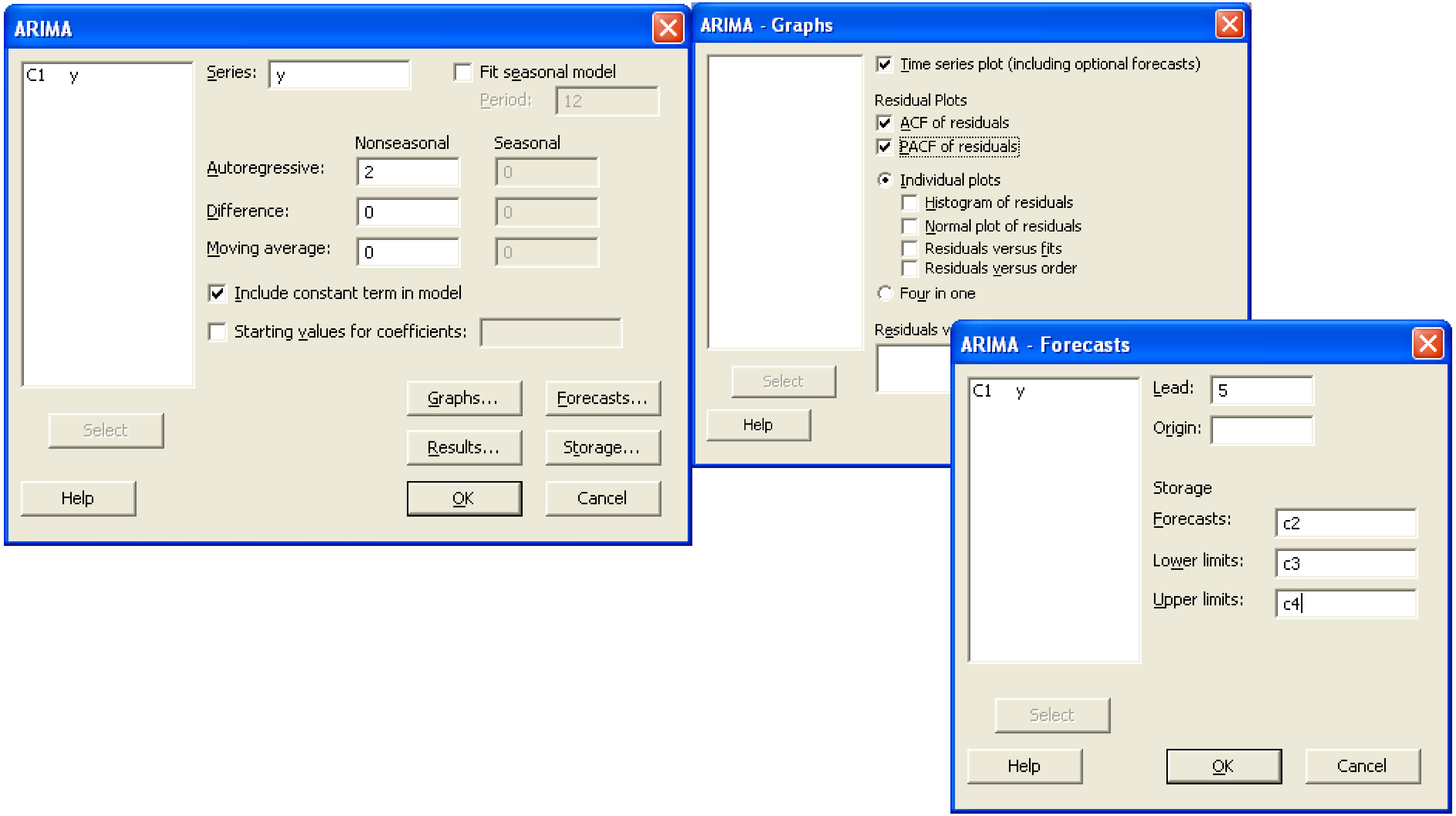


Tidsseriegraf och SAC visar att desediff1 är stationär

Förslag på modell för desediff1. Beräkna SPAC



Minitab


ARIMA
C1 y
Series: y
Fit seasonal model
Period: 12
Nonseasonal
Seasonal
Autoregressive: 2 0
Difference: 0 0
Moving average: 0 0
Include constant term in model
Starting values for coefficients:
Select
Graphs...
Forecasts...
Results...
Storage...
Help
OK
Cancel
ARIMA - Graphs
Time series plot (including optional forecasts)
Residual Plots
ACF of residuals
PACF of residuals
Individual plots
Histogram of residuals
Normal plot of residuals
Residuals versus fits
Residuals versus order
Four in one
Residuals v
Select
Help
ARIMA - Forecasts
C1 y
Lead: 5
Origin:
Storage
Forecasts: c2
Lower limits: c3
Upper limits: c4
Select
Help
OK
Cancel

ARMA(0,1)= MA(1)

SAC har en spik och

SPAC avtar ganska snabbt.

Detta är rätt tydligt

ARMA(1,0)= AR(1)

SAC avtar snabbt och
SPAC har en spik

Känns inte lika troligt men
kan vara värt att pröva.

Vi provar att anpassa både en MA(1) och en AR(1)

ARMA(0,1) = MA(1)

ARIMA Model: desediff1

Final Estimates of Parameters

Type	Coef	SE Coef	T	P
MA 1	0,6522	0,0577	11,31	0,000
Constant	0,21802	0,09463	2,30	0,022
Mean	0,21802	0,09463		

Number of observations: 177

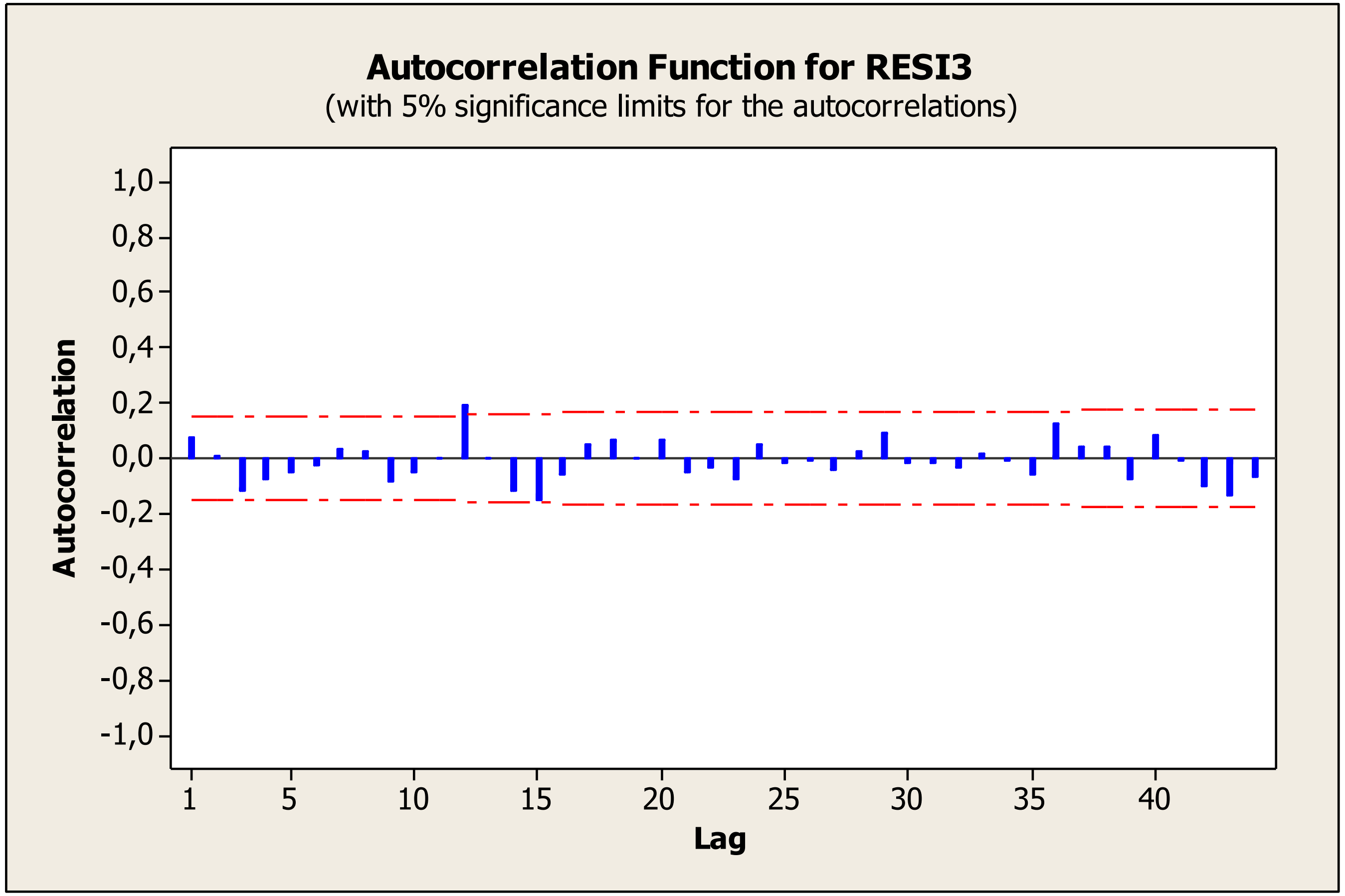
Residuals: SS = 2273,62 (backforecasts excluded)

MS = 12,99 DF = 175

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	13,8	26,7	33,8	48,9
DF	10	22	34	46
P-Value	0,182	0,223	0,476	0,357

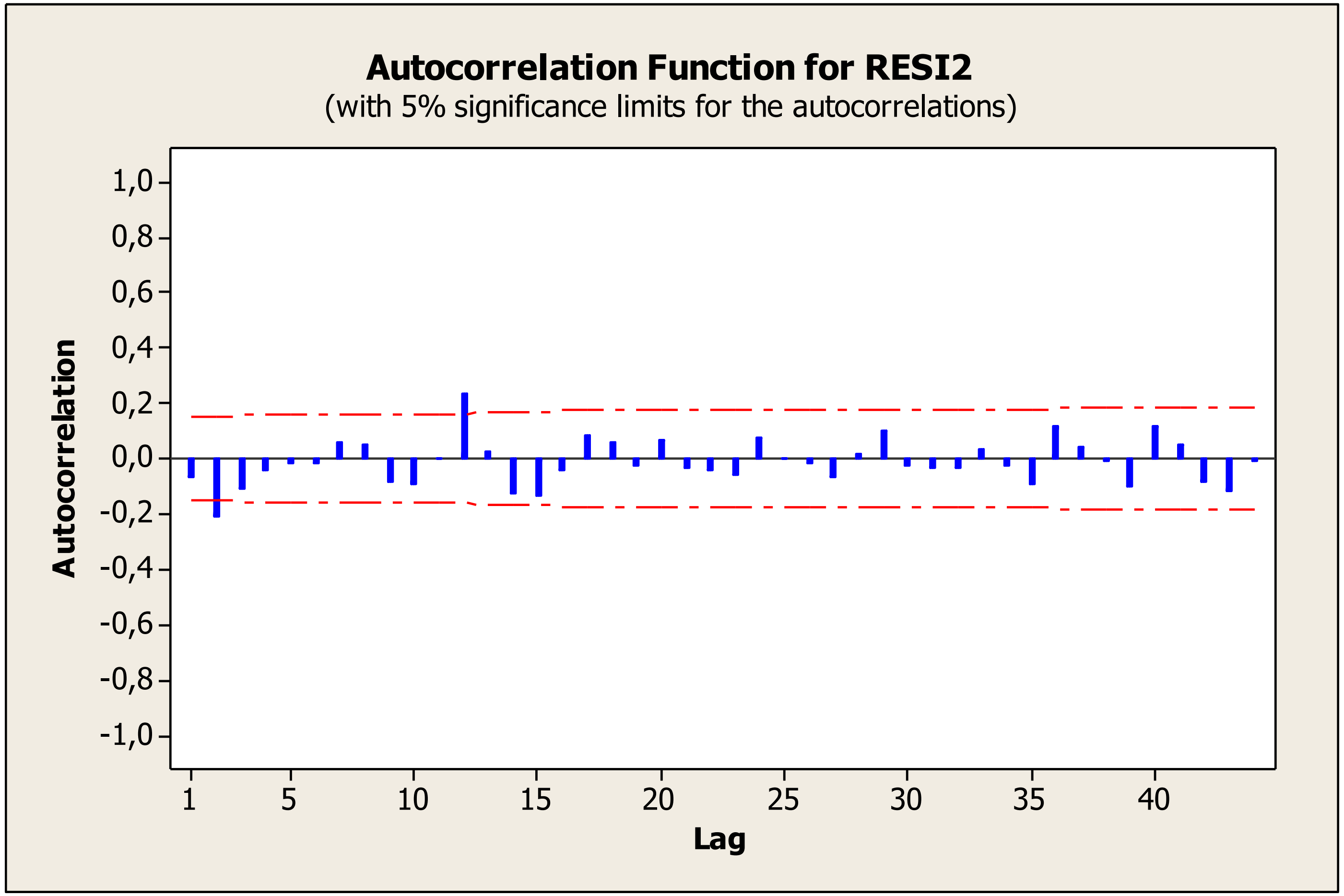
SAC på residualerna för MA(1)



ARMA(1,0) = AR(1)

```
ARIMA Model: desediff1
Final Estimates of Parameters

Type          Coef  SE Coef      T      P
AR   1     -0,4172   0,0687  -6,07  0,000
Constant    0,2938   0,2839   1,04  0,302
Mean        0,2073   0,2003

Number of observations:  177
Residuals:    SS =  2495,89 (backforecasts excluded)
              MS =  14,26  DF = 175

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag             12     24     36     48
Chi-Square    25,9   38,9   48,3   65,4
DF              10     22     34     46
P-Value      0,004  0,015  0,054  0,032
```

SAC på residualerna för AR(1)



Vilken modell är bäst

- T-kvoter
- MSE
- Residualanalys

ARIMA(p,d,q)

d anger hur många differentieringar för trend vi gjort.

I vår modell är d=1.

Vald modell är ARIMA(0,1,1)

Låt Minitab diffa automatiskt.

Så nu skattar vi återigen de skattade parametrarna i den valda modellen

ARIMA Model: DESE1

Final Estimates of Parameters

Type	Coef	SE Coef	T	P
MA 1	0,6522	0,0577	11,31	0,000
Constant	0,21802	0,09463	2,30	0,022

Differencing: 1 regular difference

Number of observations: Original series 178, after differencing 177

Residuals: SS = 2273,62 (backforecasts excluded)

MS = 12,99 DF = 175

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	13,8	26,7	33,8	48,9
DF	10	22	34	46
P-Value	0,182	0,223	0,476	0,357