Tidsserieanalys 7,5 hp Föreläsning 1

Lärare: Lotta Hallberg

E-post-adress: ann-charlotte.hallberg@liu.se

Josef Wilzén

E-post-adress: Josef.wilzen@liu.se

Kursens innehåll

- Projekt i tidsserieanalys, 1,5hp
- Skriftlig tentamen, 6hp

Innehåll:

- Index
- Säsongsrensning (komponentuppdelning)
- Tidsserieregression
- Exponentiell utjämning
- ARIMA/SARIMA modellering
- Några mer avancerade metoder

Kursens upplägg

- LISAM, Teams
- Föreläsningar, Lektioner, Datorlaborationer
- Litteratur:
  - Index, kompendium skrivet av Anders Nordgaard
  - Introduction to TIME SERIES ANALYSIS AND FORECASTING Montgomery, Wiley
  - Plus en bok
- Datorlaborationer
  - Träning på olika moment
  - Två timmar bokad tid för varje laboration
  - Minitab och R
- Examination. Projekt och skriftlig salstentamen
  - Viktigt att labbarna är genomförda för att klara dessa moment
  - Hjälpmedel på tentan: Räknedosa. Två handskrivna eller digitalt skrivna A4 blad med egna anteckningar (båda sidor).
  - För godkänt krävs minst 12 av 20 poäng och för väl godkänt krävs minst 16 av 20 poäng.

Tidsserieanalys
Kapitel 1

- Data som observeras över tid
  - Kan ha olika tidsenhet: år, kvartal, månad, vecka, dag, timme, sekund
- Tidserien ska vara:
  - Samplad med regelbundna intervall
- Vi vill...
  - Förstå tidserien
  - Göra prognoser (forecast) för framtiden

Vilka särdrag har tidsseriedata? Varför behövs nya metoder?

- Observationerna är inte oberoende, eftersom det finns ett beroende i tiden
- Det är samma tidsavstånd mellan alla observationer.
- Inga observationer saknas.

- Observationerna ger ett mönster över tiden
  - en trend: fallande eller stigande värden med tiden
  - en periodisk variation över en tidsperiod av bestämd längd (säsongseffekter eller liknande)
  - konjunktursvängningar
  - förändringar pga regeländring, förändrad definition av y-variabeln osv

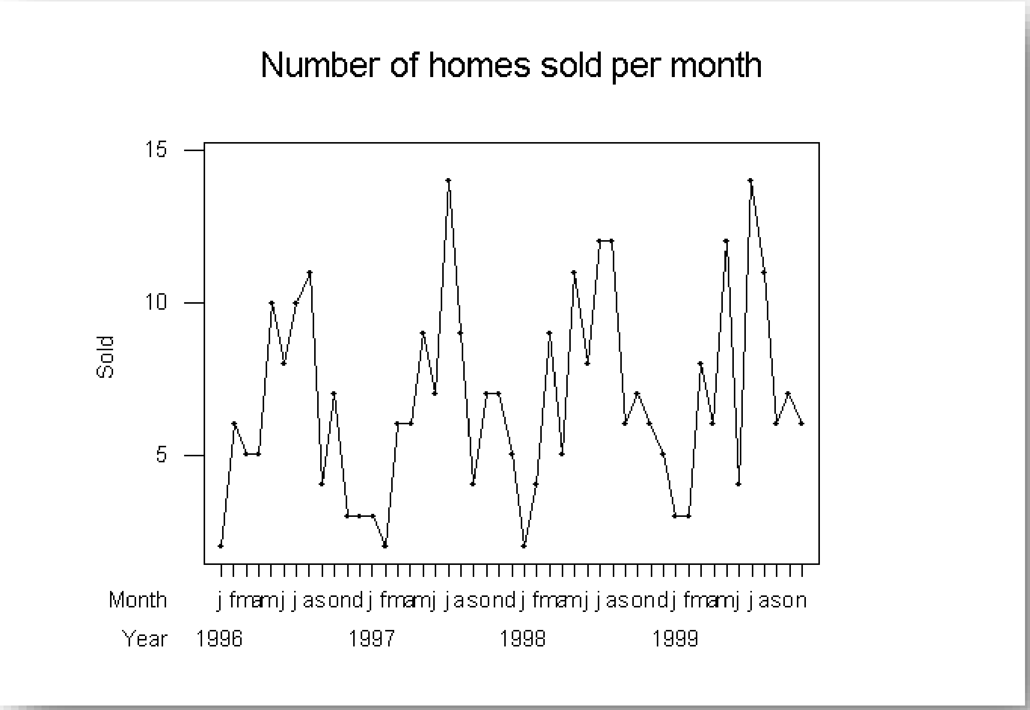
Exempel på tidsseriedata

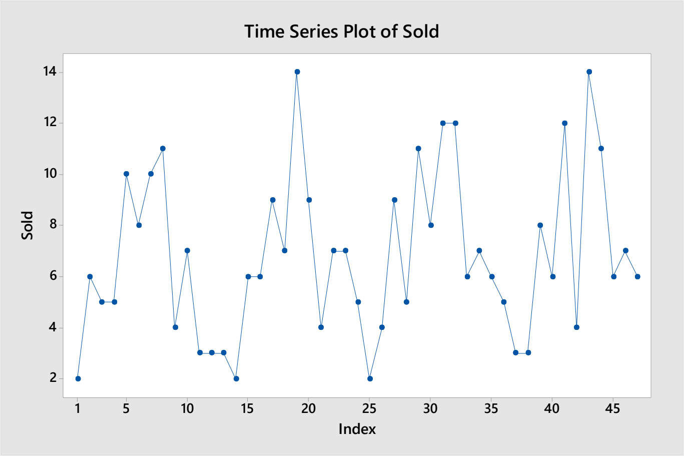
- Olika typer av ekonomiska data:
  - Arbetslöshetssiffror
  - Försäljningsvärden i ett företag
  - Konsumentprisindex och andra index
  - Export- och importmängder

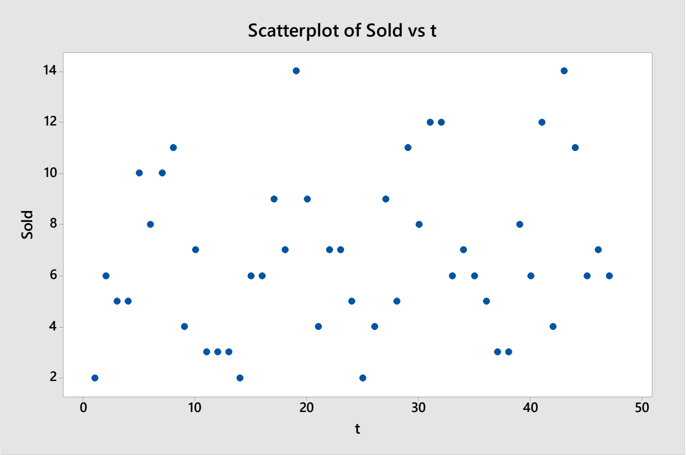
- Miljömätdata:
  - Fosforhalt i havsvattenbassänger
  - Ozonhalt i luftrummet över en storstad

- Medicinska data:
  - Antal fall av viss sjukdom (influensa, påssjuka ...)

Introduktion till tidsserier

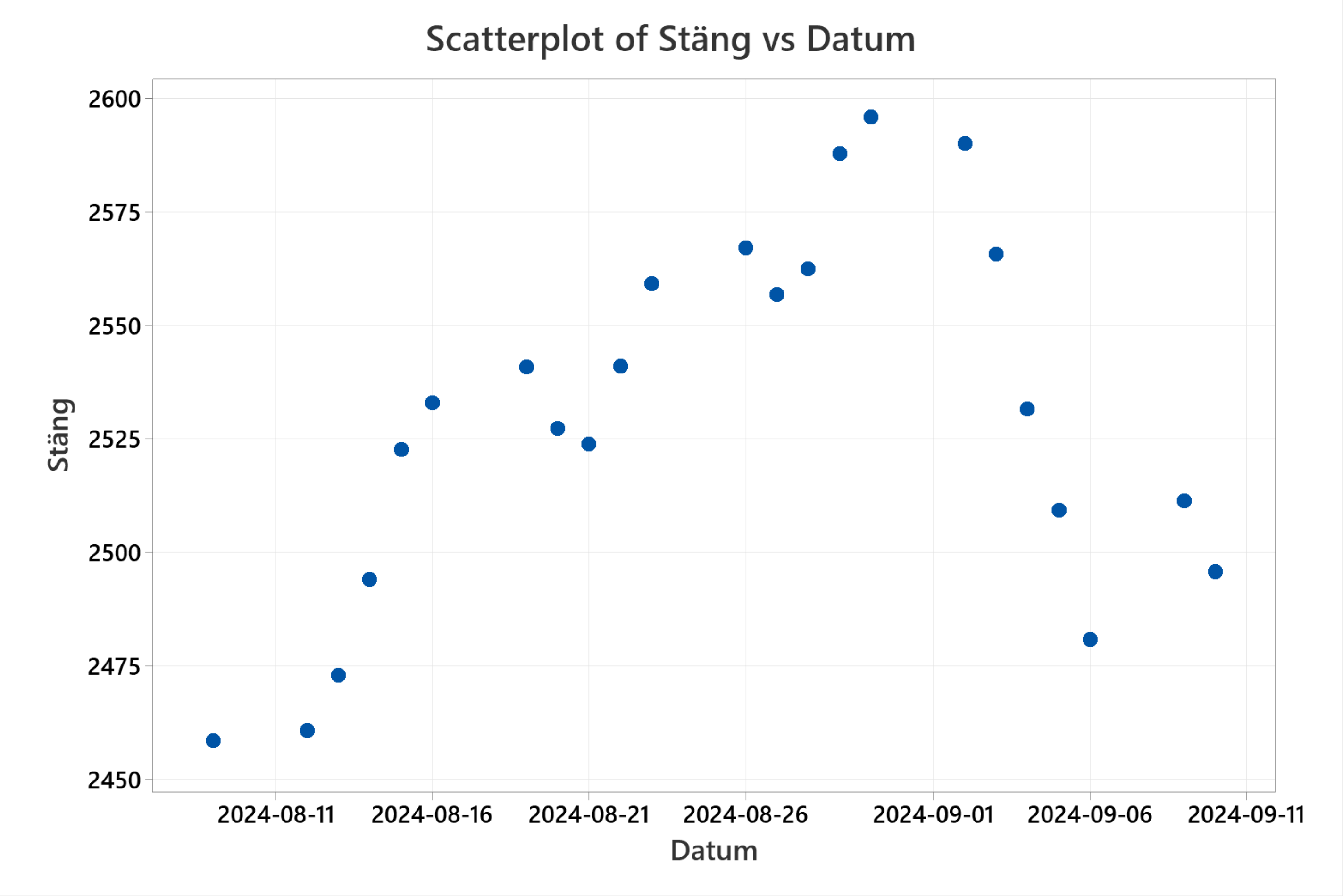


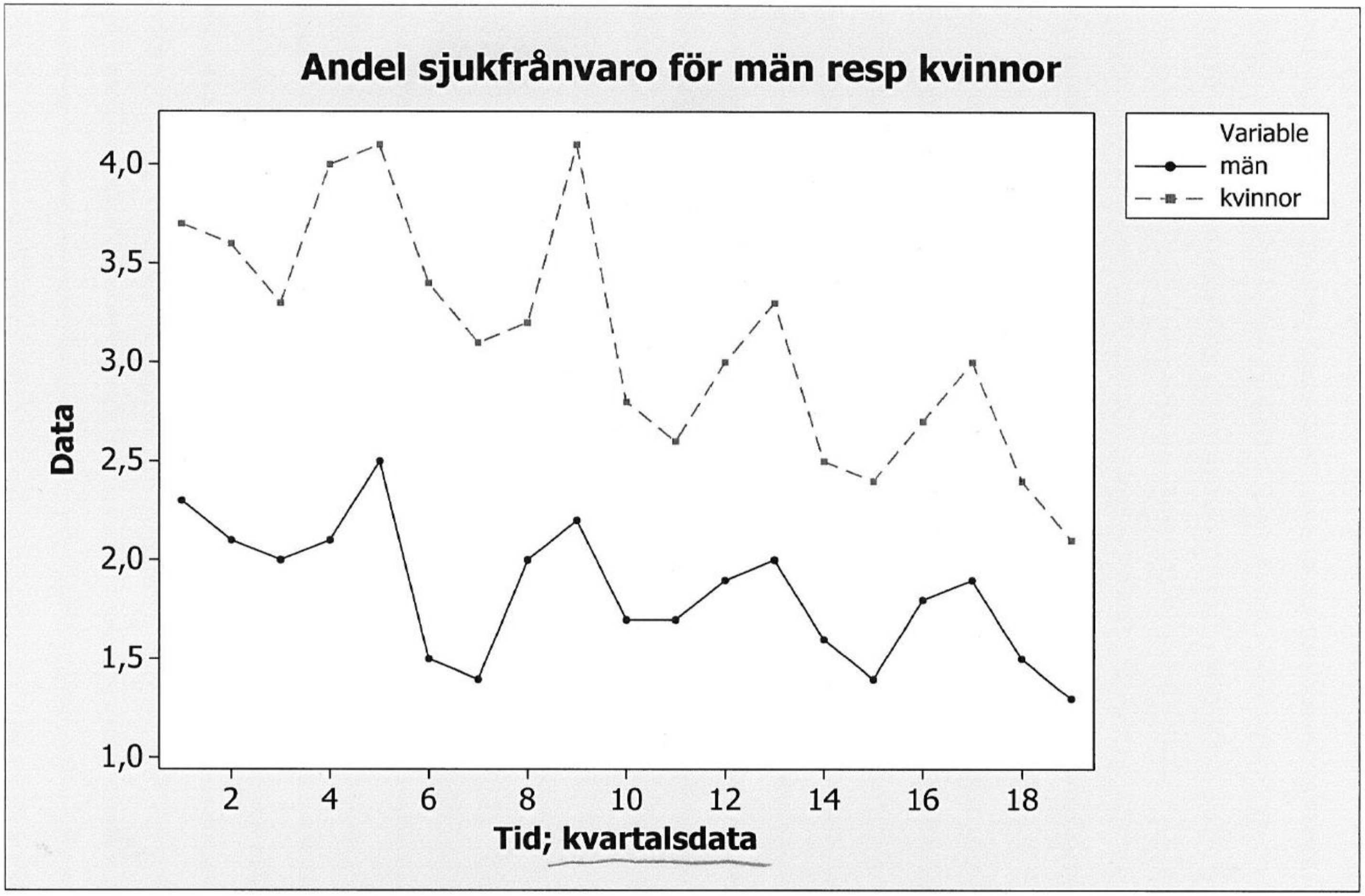




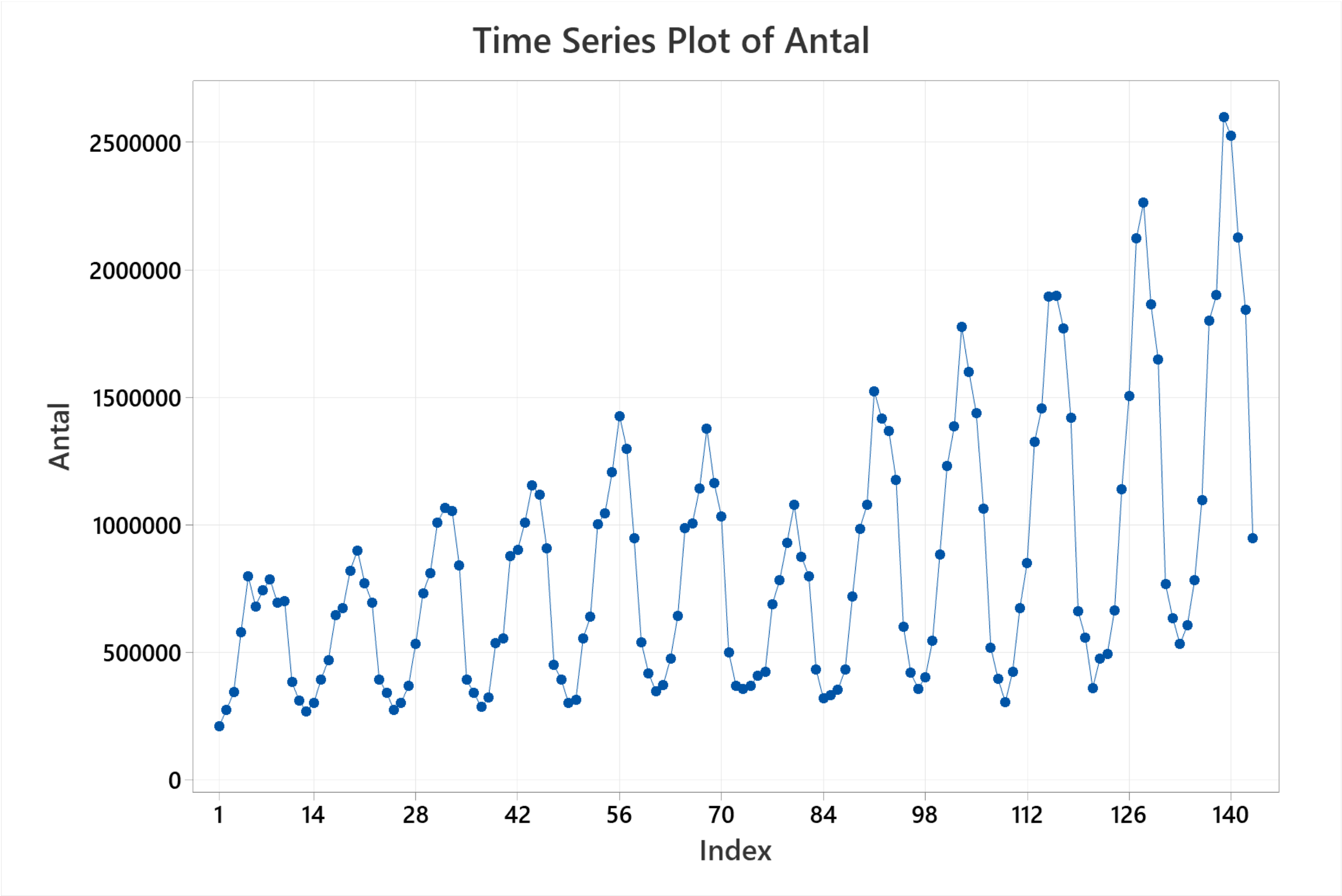
Tidsserieanalys

OMX30 24-08-09 till 24-09-10

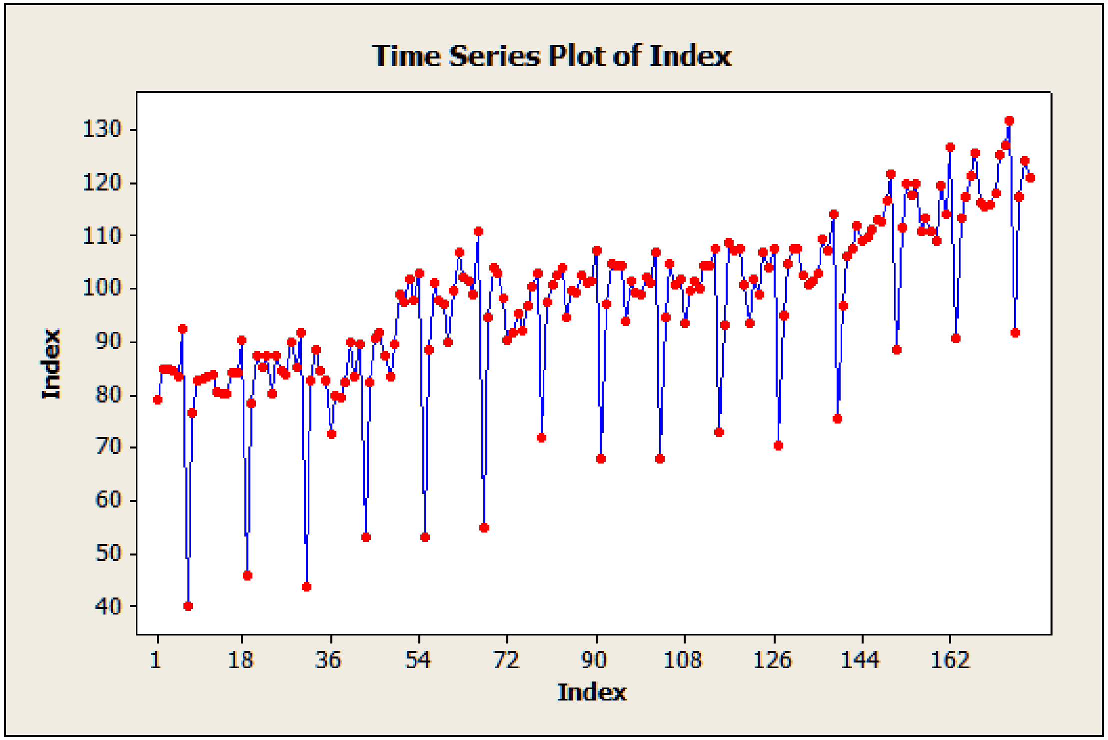



Andel sjukfrånvaro för män resp kvinnor
Data
4,0
3,5
3,0
2,5
2,0
1,5
1,0
2
4
6
8
10
12
14
16
18
Tid; kvartalsdata
Variable
män
kvinnor

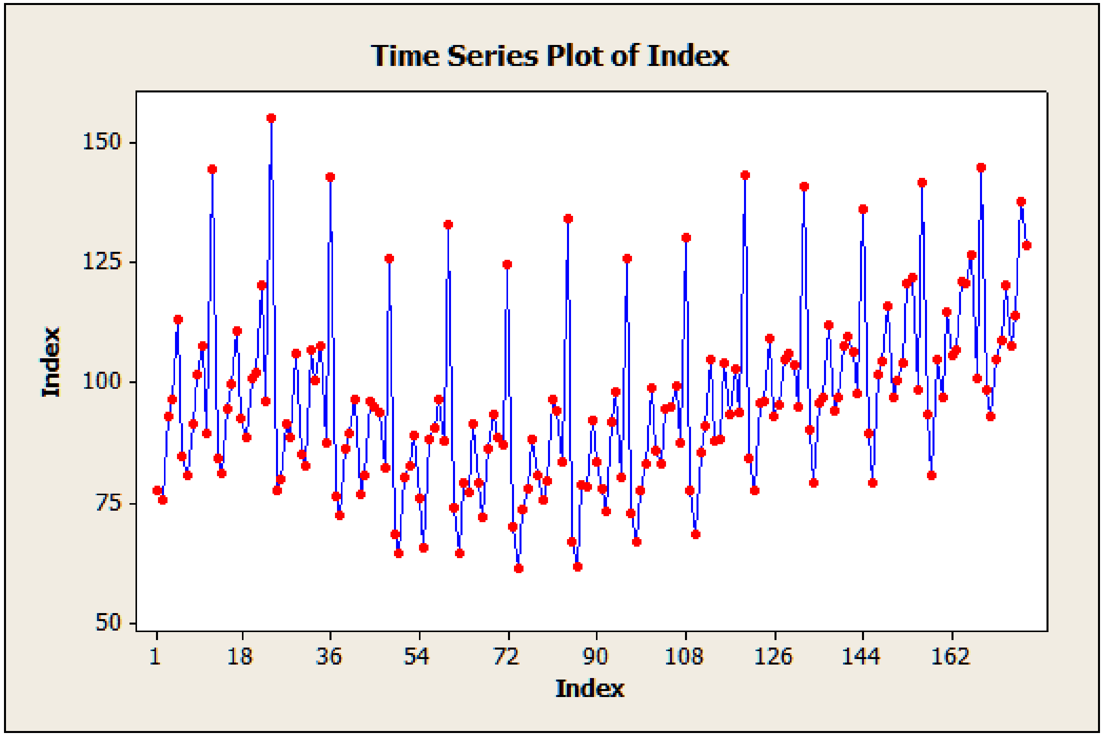
Antal turister till Turkiet Månadsdata



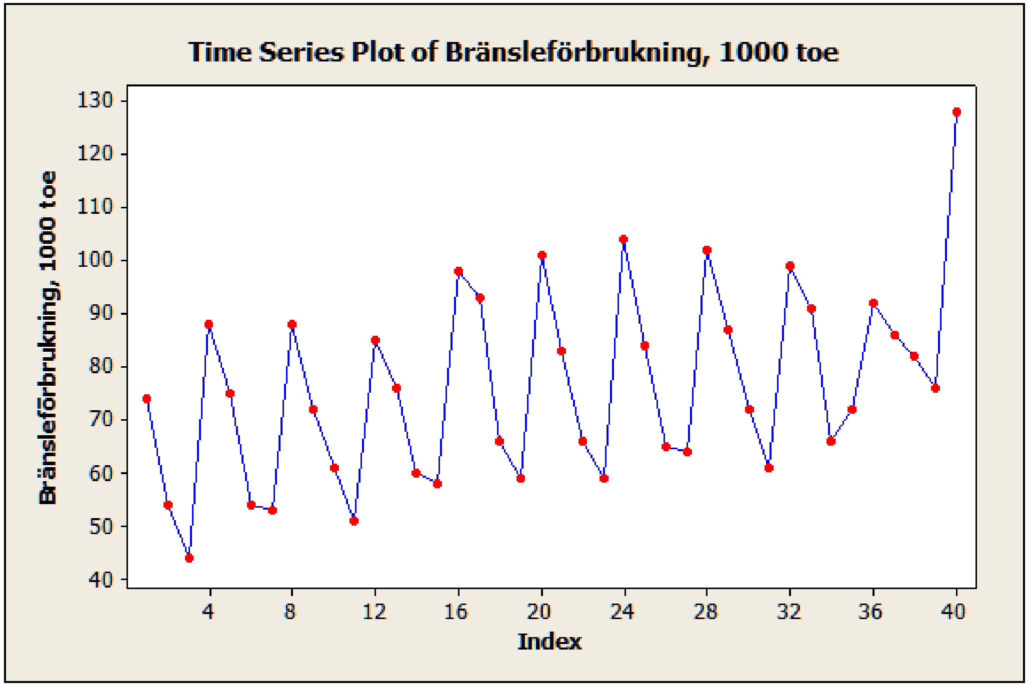
Månadsdata över pappersproduktion jan 1990 till okt 2004



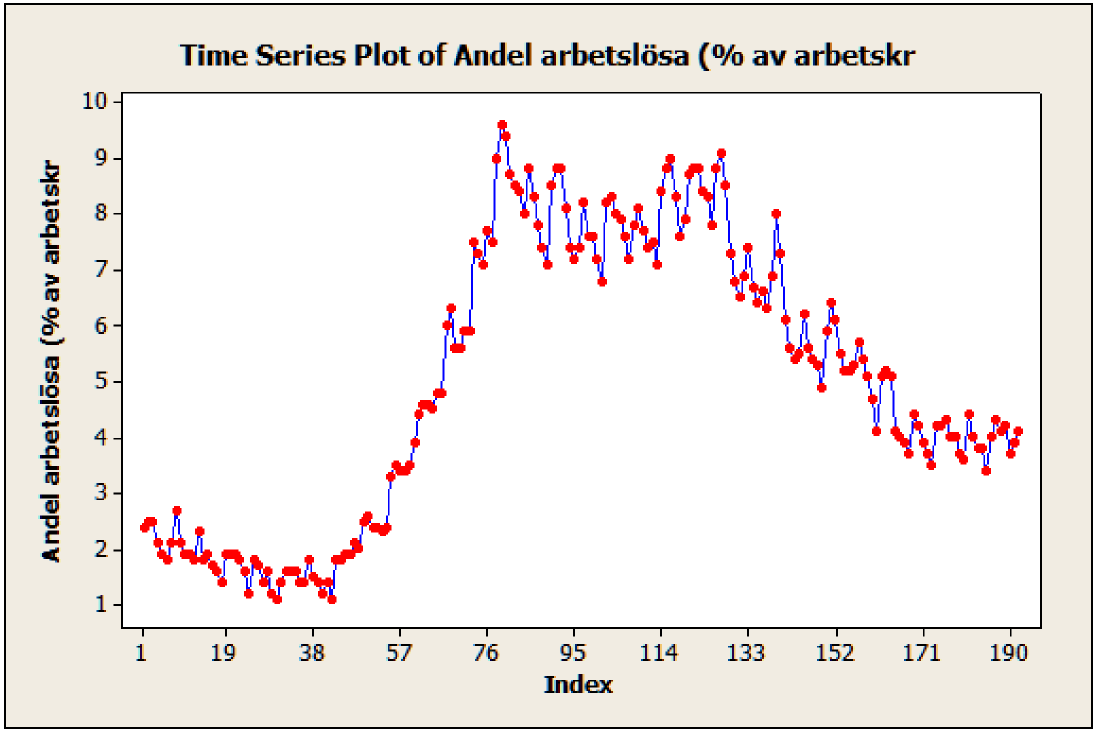
Försäljning kläder och skor jan 1990 till sep 2004



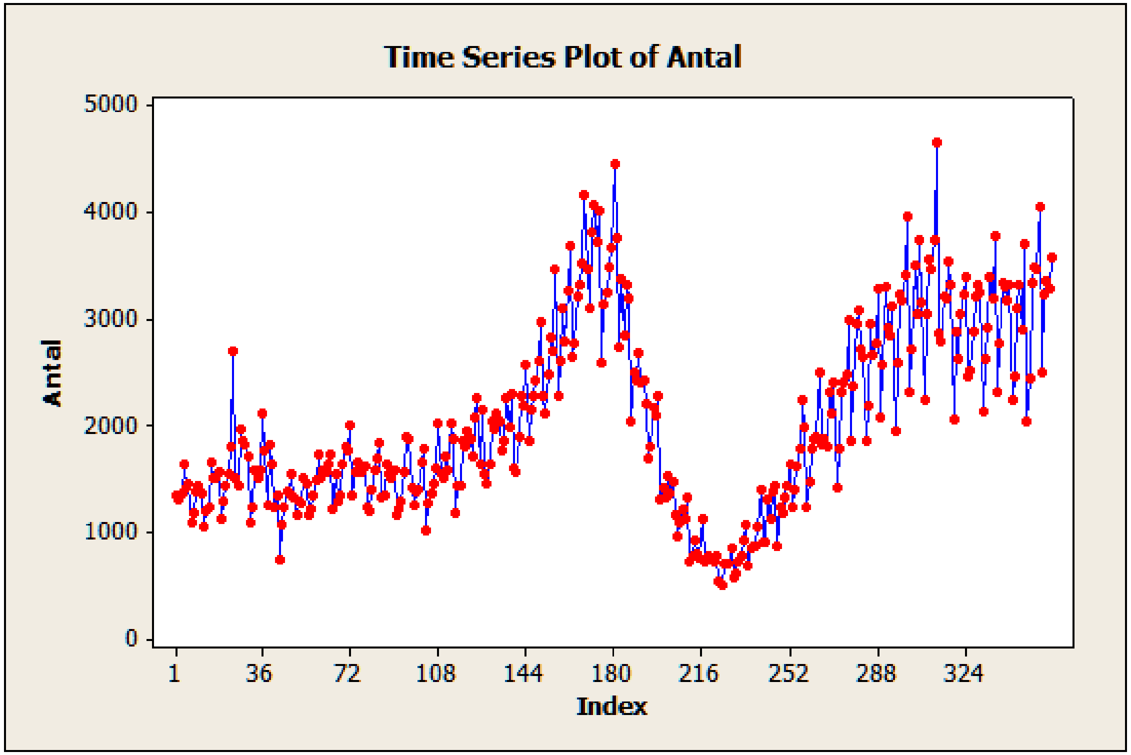
Bränsleförbrukning, naturgas kvartal 1, 1992 till kvartal 4, 2001



Andel arbetslösa
jan 1987 till dec 2002



Antal lastbilsregistreringar jan 1975 till nov 2004



Tidsserieanalys

- En tidsserie har diskreta tidpunkter
  - Vid kontinuerlig tid så har vi en stokastisk process
- Inga saknade värden
  - Imputera om något värde saknas
- Man mäter samma sak genom hela serien
  - Ex: arbetslöshetsdata. Är definitionen på arbetslöshet samma genom hela serien
- Vilken upplösning har serien?
  - År, månad, kvartal, vecka, dag, timme osv

Statiska tidsseriemodeller som kan beskriva tidsserien och även producera prognoser

1. Anpassad modell för tidsserieregression kan användas för att göra prognoser för framtida tidpunkter samt tolka mönster.

2. Klassisk modell för komponentuppdelning kan göra prognoser med hjälp av trend- och säsongskomponenter, men i viss mån även beträffande konjunktur samt även tolka vissa mönster.

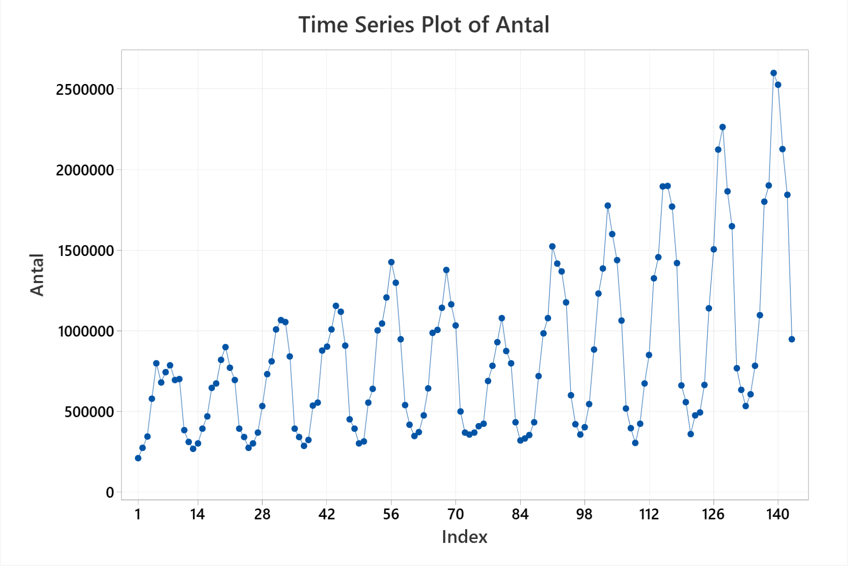
Dynamiska tidsseriemodeller som framförallt syftar till att göra prognoser

3. Exponentiella utjämningsmodeller

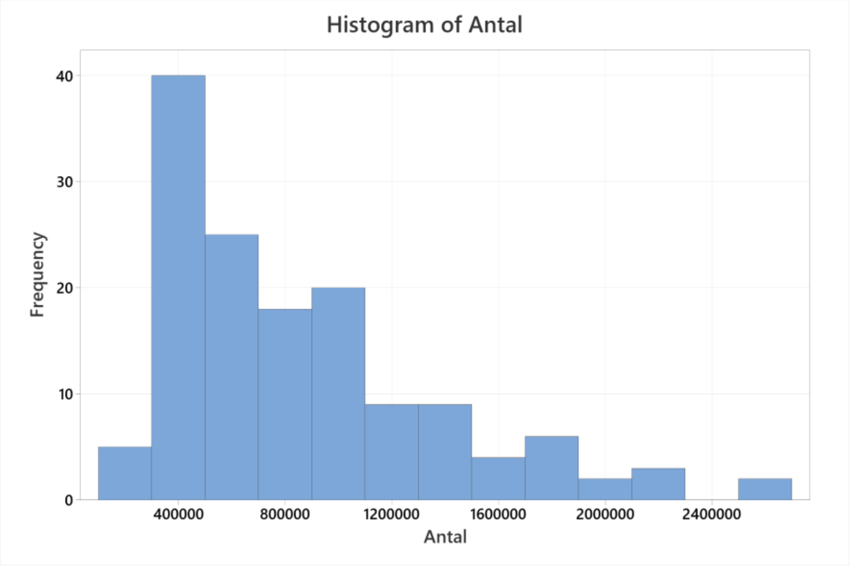
4. Autoregressiv modellering. SARIMA-modeller

Grafiska metoder Kap 2,2

Antal turister

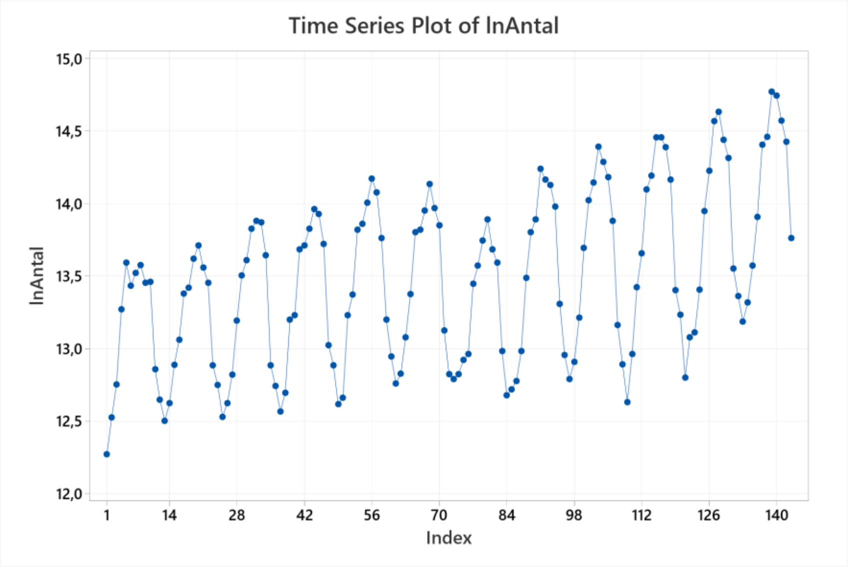


Antal turister

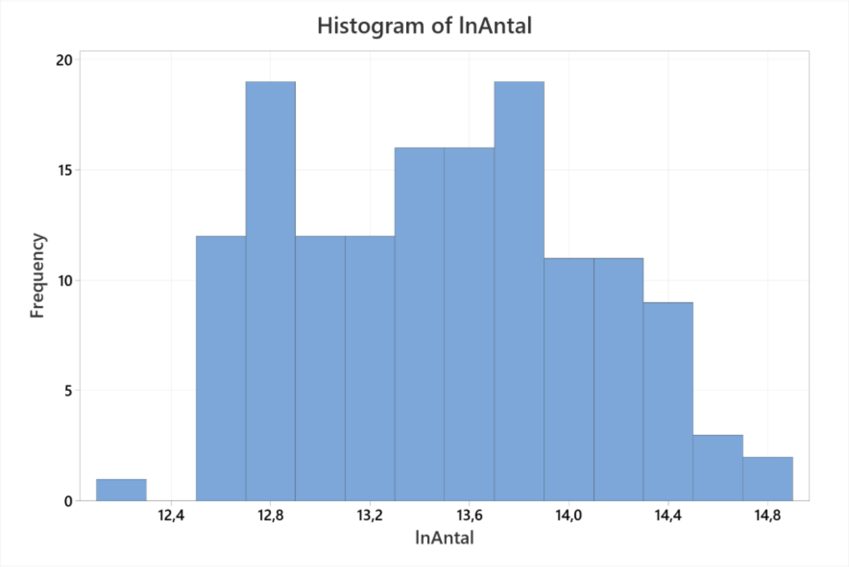


Transformera kap 2,4,1

ln Antal



ln Antal



Prognosprocess
Forecasting Kap 1,3

- Identifiera problemet
- Samla in data
- Analysera tidsserien
- Hitta en lämplig modell
- Validera modellen
- Gör prognos
- Utvärdera prognos

Kapitel 2,1
Några grundläggande begrepp

- Låt y_t , $t = 1,2, \dots, T$ vara tidsserien
- Vi har en observation av slumpvariabeln y_t vid varje tidpunkt t.
- Låt τ beteckna antalet tidpunkter fram i tiden som vi vill göra prognos för (eng. Lead time)
- Låt $\hat{y}_{t+\tau}$(t) beteckna prognos för y för tidpunkt $t+\tau$ då vi står vid tidpunkt t när vi gör prognosen
- Då gäller även $\hat{y}_t(t-\tau)$ är prognos för y_t gjord vid tidpunkt $t-\tau$
- Prognosfel $e_t(\tau) = y_t - \hat{y}_t(t-\tau)$
- Om $\tau = 1$ så är $e_t(1) = e_t = y_t - \hat{y}_t(t-1) = y_t - \hat{y}_t$
- Vi vill ha små prognosfel.

Kapital 2,6
Valideringsmått

- Vi ska titta på tre mått MSD, MAD och MAPE
- Alla tre är mått på hur bra anpassningen är och kan användas för att jämföra olika modeller. Den modell som har lägst MAPE, MAD och/eller MSD har bäst anpassning.
- Oftast visar alla 3 måtten åt samma håll. Men i vissa fall kan man vara tvungen att välja ett av dem.
- Vid val mellan t ex additiv modell och multiplikativ modell kan det hända att något av måtten är högre för den ena modellen medan ett annat mått är lägre. Det gäller alltså att tolka måtten med visst förnuft.

MSD kan också jämföras med *MSE* i den multipla regressionsmodellen:

$$MSD = \frac{1}{n}\sum_{t=1}^{n}\left(y_t - \hat{y}_t\right)^2$$ *Mean Square Deviation*

$$MSE = \frac{1}{n-k-1}\sum_{t=1}^{n}\left(y_t - \hat{y}_t\right)^2$$ *Mean Square Error*

$$MAD = \frac{1}{n}\sum_{t=1}^{n}\left|y_t - \hat{y}_t\right|$$ *Mean Absolute Deviation*

MAD är mindre känslig för avvikande värden och blir mer användbar när vi har något enstaka värde som uppträder konstigt.

Ytterligare en fördel med *MAD* är att dess värde är i samma skala som y_t observationerna själva, vilket gör det lättare att tolka.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|$$

Mean Absolute Percentage Error

Måttet använder också absoluta avvikelser, men mäter dem relativt nivån hos y. Vi får alltså relativa (procentuella) avvikelser.

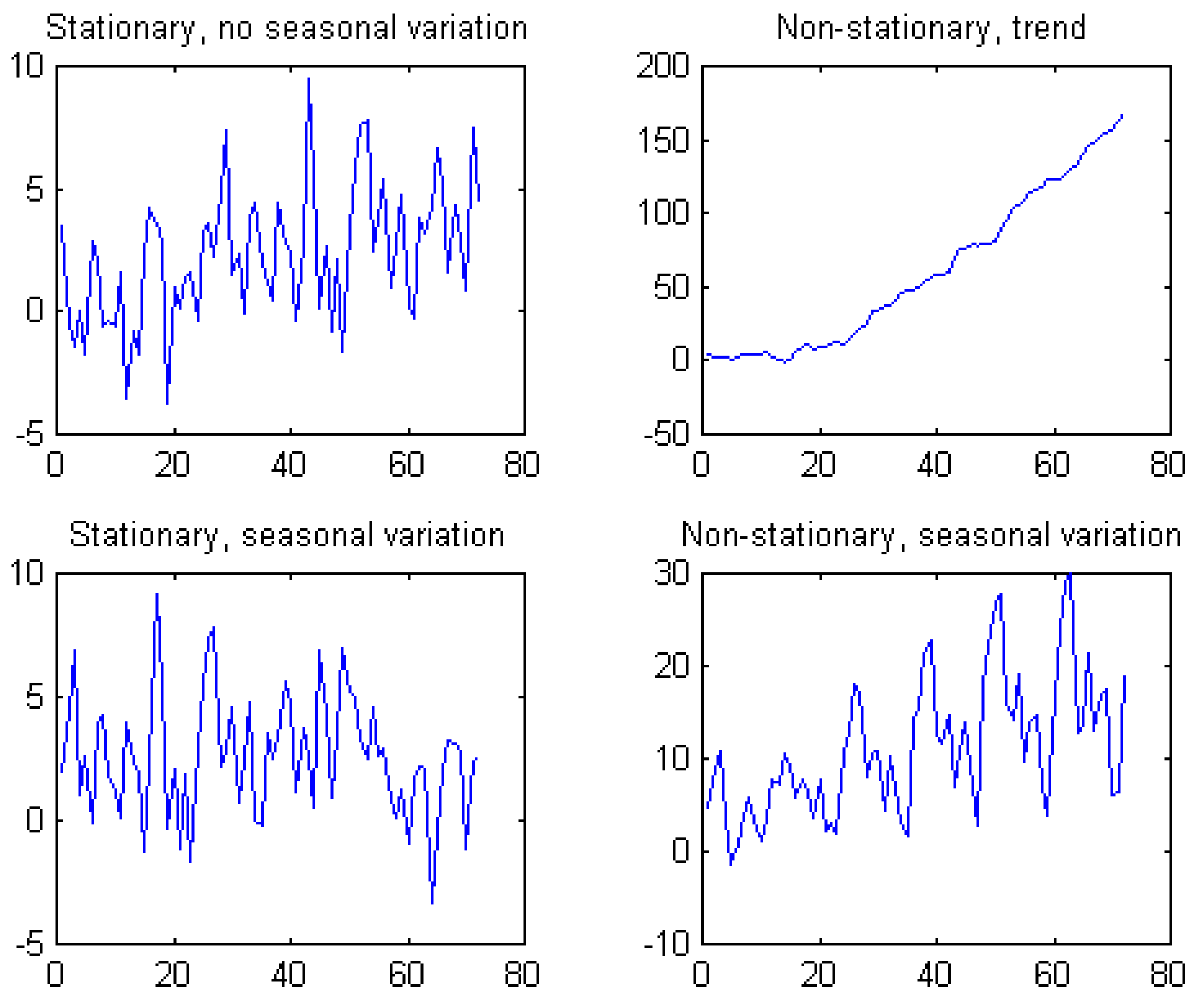
Måttet är praktiskt för multiplikativa modeller där den oregelbundna (slump) komponenten (IR_t) är ganska betydande, eftersom avvikelserna då blir stora när vi har stora värden på y.

Stationäritet Kap 2,3

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.

Exempel på stationära och icke-stationära tidsserier:



Hur kan man avgöra om en tidsserie är stationär?

1. Studera grafen över tidsserien (trender och dyligt syns i regel tydligt)

2. Beräkna och studera den skattade *autokorrelationsfunktionen*, SAC = <u>s</u>ample <u>a</u>uto<u>c</u>orrelation-function

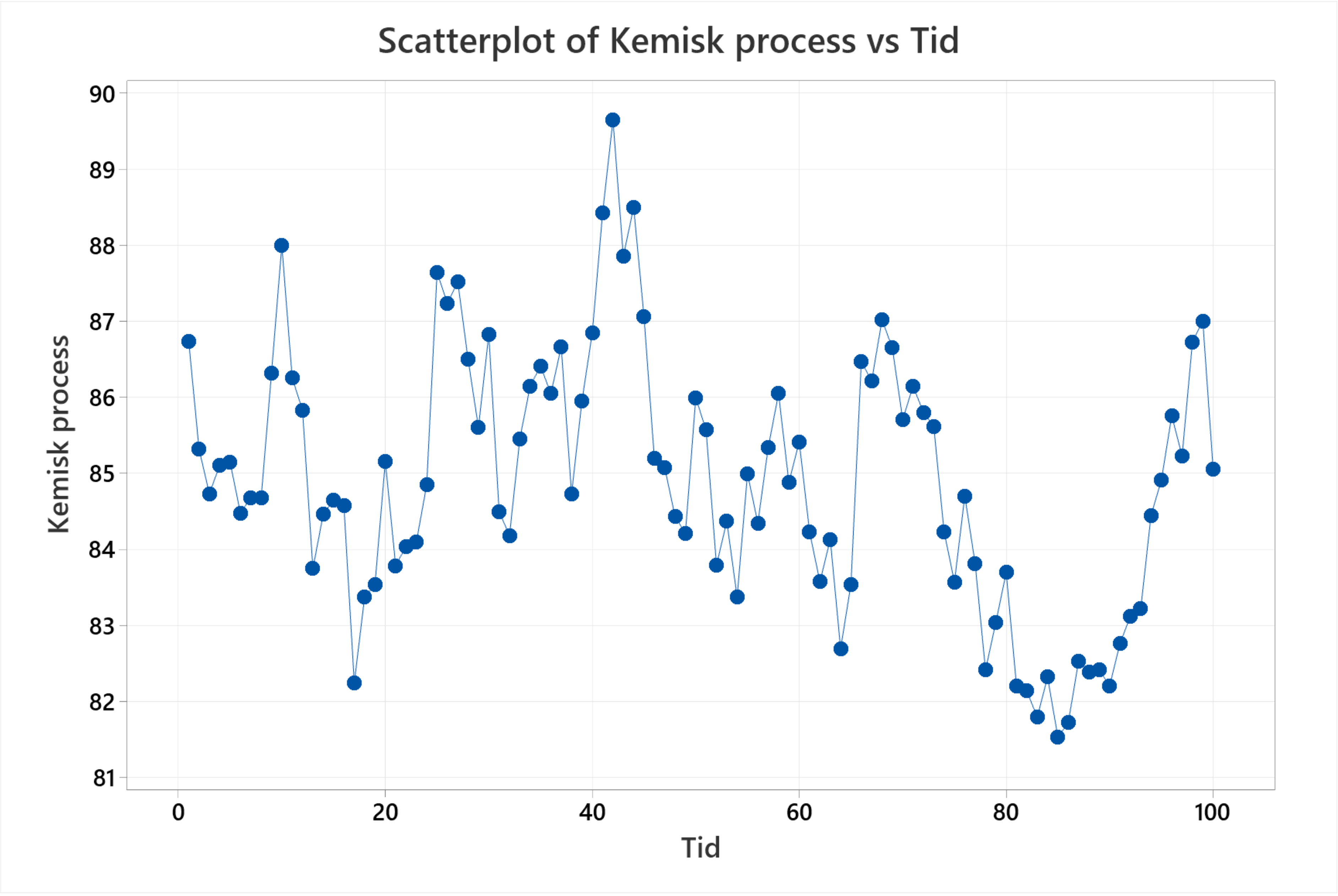
   $r_k = Korr(y_t , y_{t+k})$ för $k = 1,2,3, \ldots$

   Denna ska avta relativt snabbt mot 0 för att serien ska vara stationär.

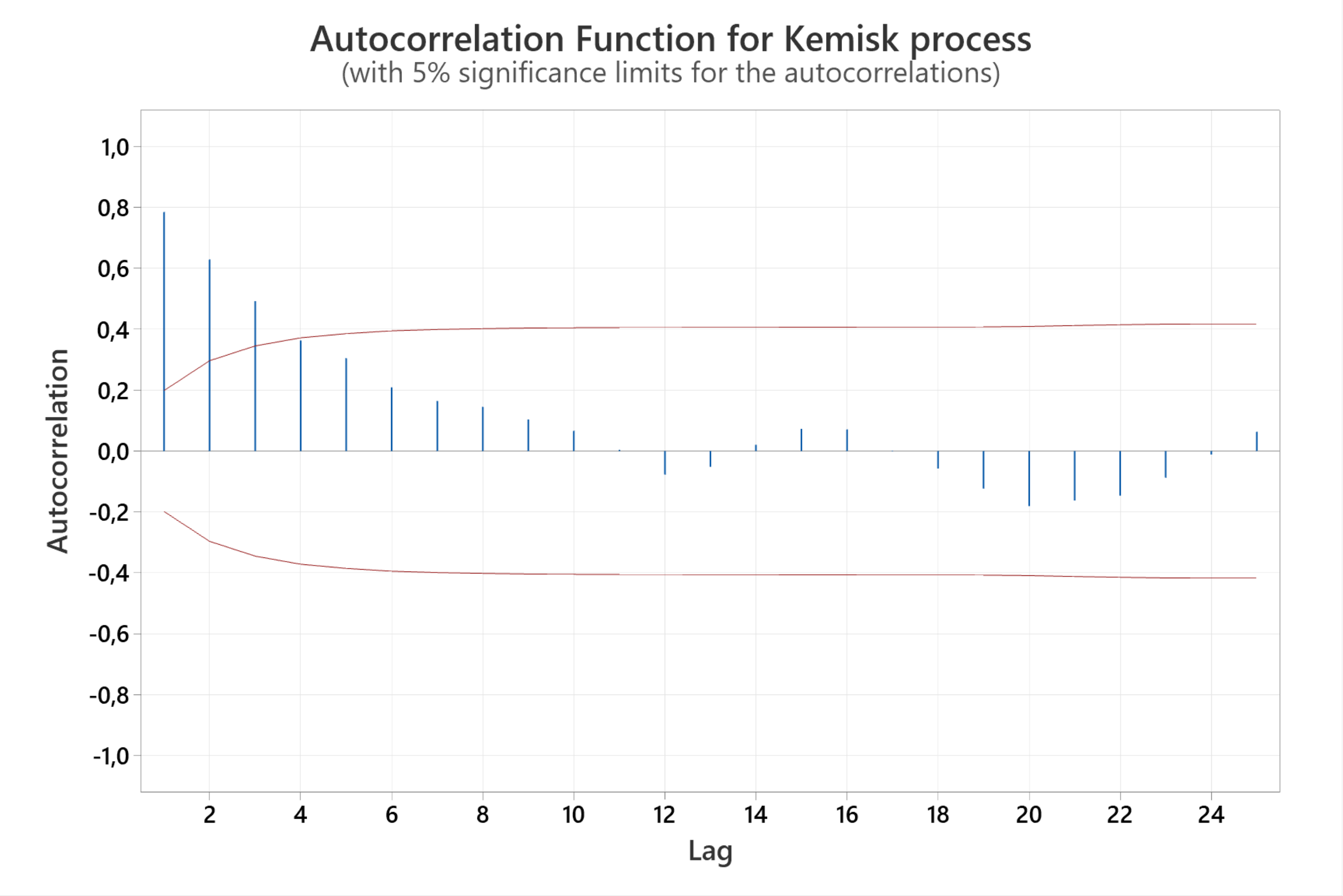
SAC

SAC är den skattade autokorrelationen. Sample AutoCorrelation

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


Scatterplot of Kemisk process vs Tid
Kemisk process
90
89
88
87
86
85
84
83
82
81
0
20
40
60
80
100
Tid

SAC Se sid 40 I boken



Differentiering kap 2,4 sid 50→

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

Om <u>trend</u> ses i tidsserien:

Differentiera originalserien med 1 tidsförskjutning

$z_t = y_t - y_{t-1}$ = Diff1

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

$w_t = y_t - y_{t-1}$ och sedan

$z_t = w_t - w_{t-1}$ = Diff1,1

Om kraftig säsongsvariation finns så kan vi differentiera för säsong:

$$z_t = y_t - y_{t-L} = \text{Diff}L$$

där L är säsongslängden

Månadsdata: $z_t = y_t - y_{t-12} = \text{Diff12}$

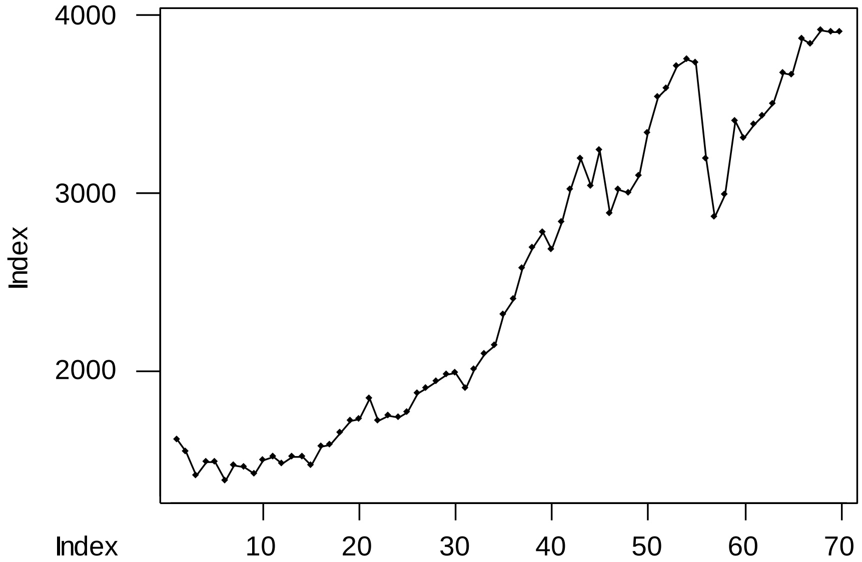
Kvartalsdata: $z_t = y_t - y_{t-4} = \text{Diff4}$

Vi kanske måste differentiera både för trend och säsong:

$$w_t = y_t - y_{t-1} = \text{Diff1}$$

$$z_t = w_t - w_{t-L} = \text{Diff1, L}$$

DiffL, 1 ger samma resultat



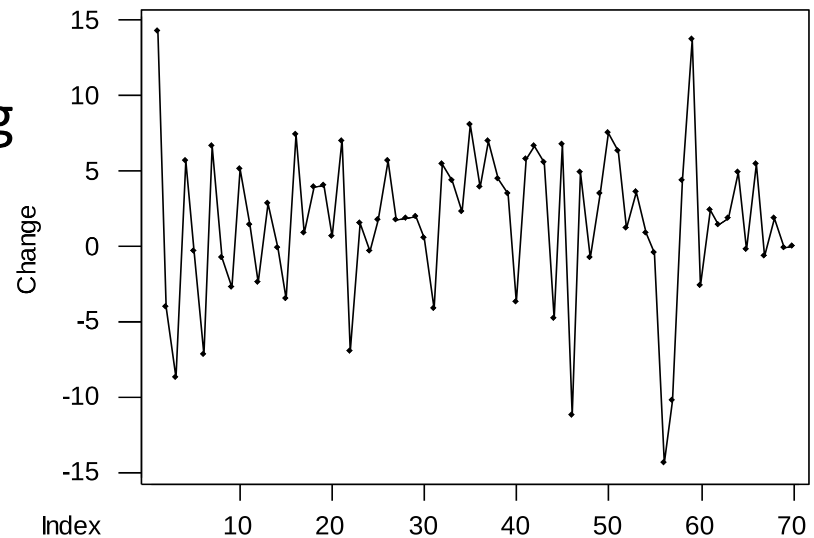
Ursprungliga tidsserien:

inte stationär

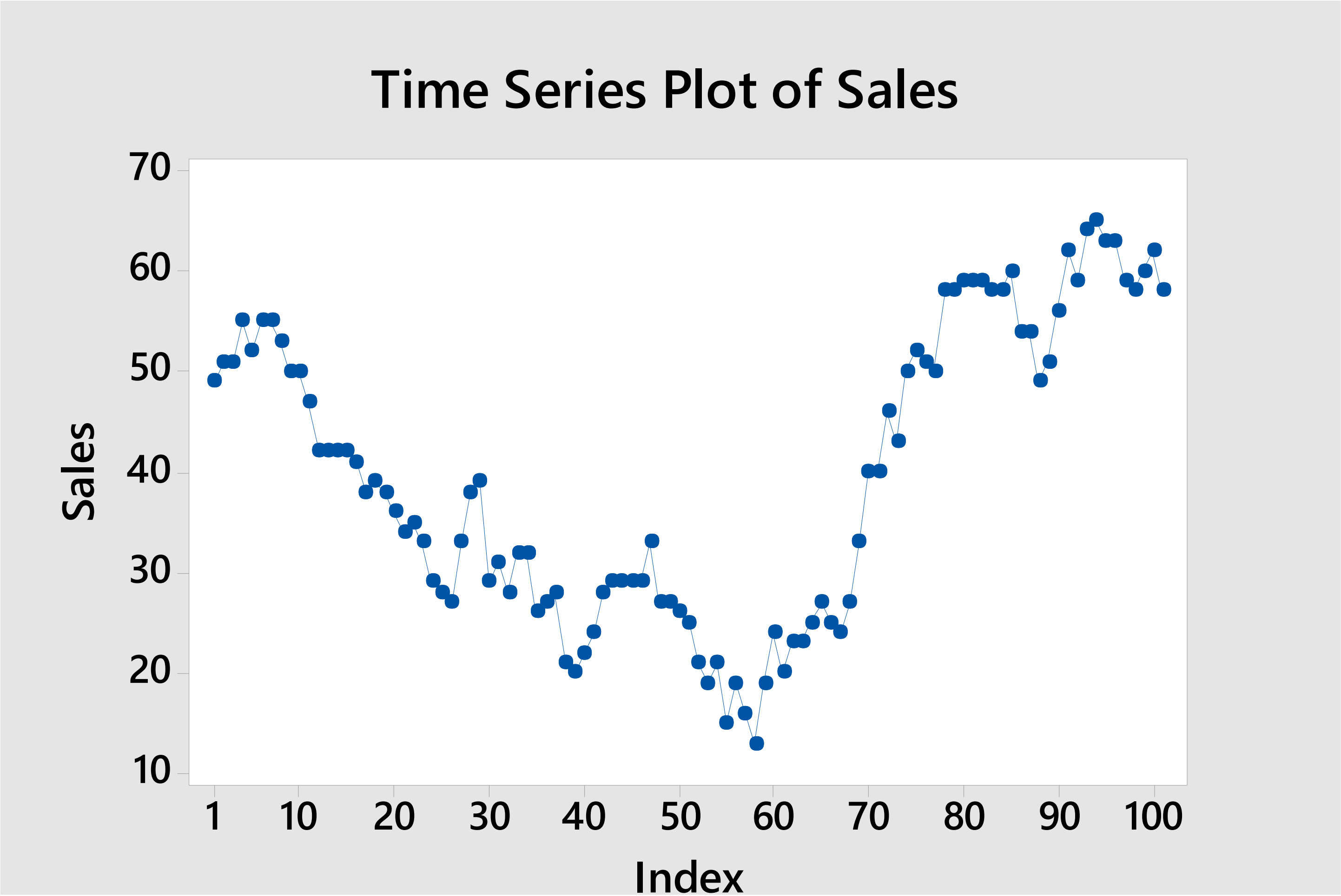
Månadsdata

Differentierad serie (förändring från månad till månad):

stationär

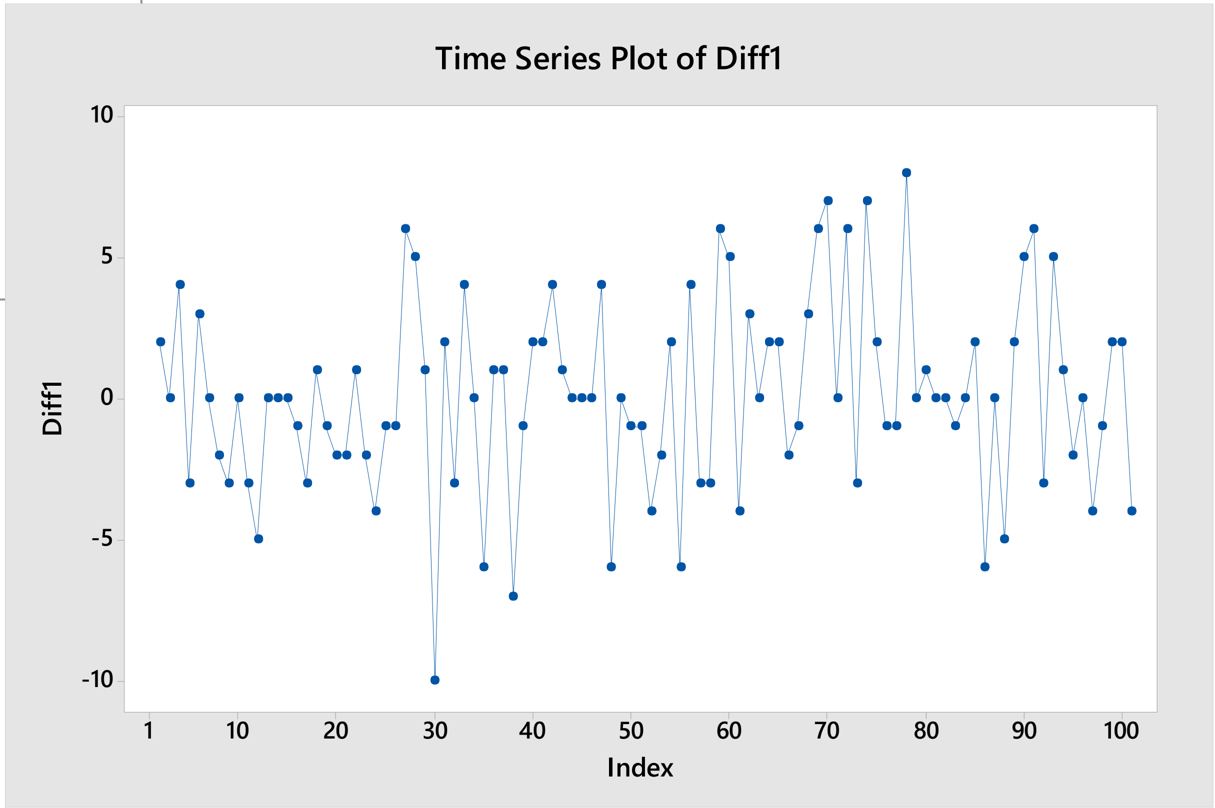


Exempel: Försäljning av tandkräm



t	y_t	y_{t-1}
1	49	*
2	51	49
3	51	51
4	55	51
5	52	55
6	55	52
osv		55

Diff1
49-*=*
51-49=2
51-51=0
55-51=4
52-55=-3
55-52=3
osv



B = bakåtskiftoperator

$By_t = y_{t-1}$ B kan bara operera från vänster

Difference $(1 - B)y_t = y_t - y_{t-1}$

$B^2 y_t = y_{t-2}$ $B^2 = BB$

$B^{12} y_t = y_{t-12}$

$(1 - B^{12})y_t = y_t - y_{t-12}$

$(1 - B)^2 y_t = (1 - \mathrm{B})(y_t - y_{t-1})$

Detta var genomgång av:

Kapitel 1 och mycket ur kapitel 2.

(ej variogram 2,3,3) 2,4 Endast log transformation i 2,4,1

(ej 2,6,3)