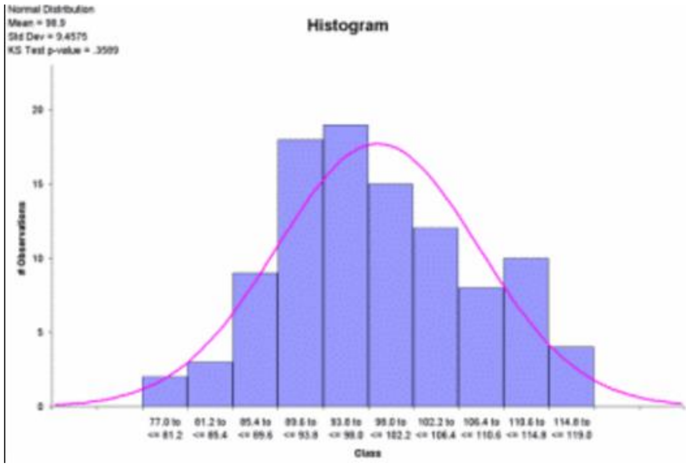
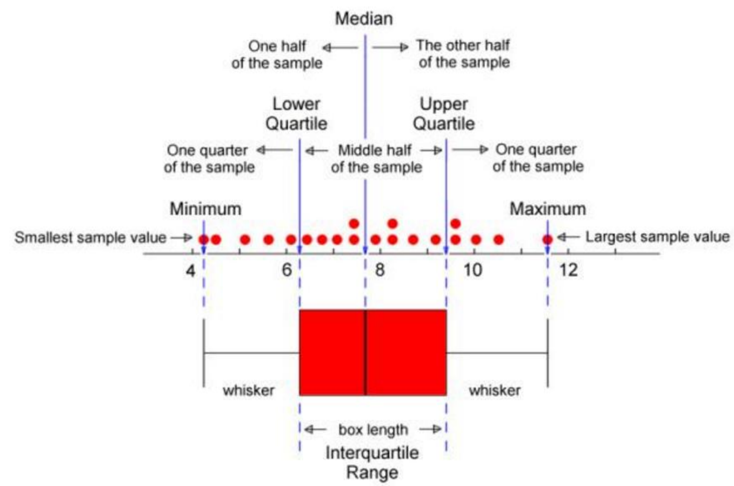
MSci 523 – Forecasting

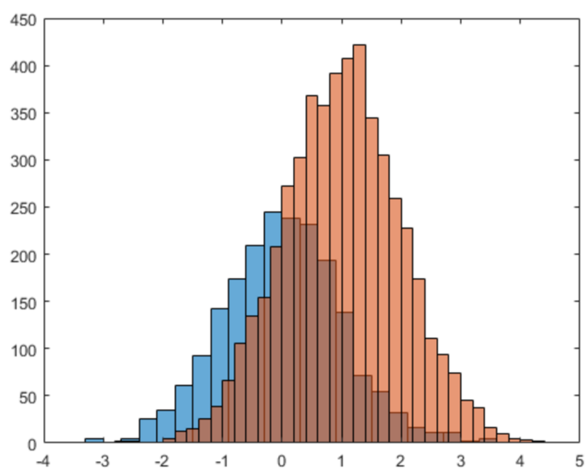
Lecture 2

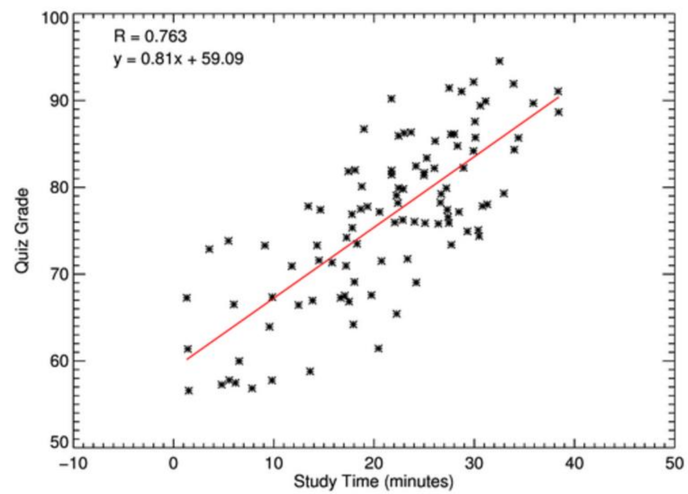
Time Series Analysis:
Data Exploration & Transformation

Traditional Data Exploration



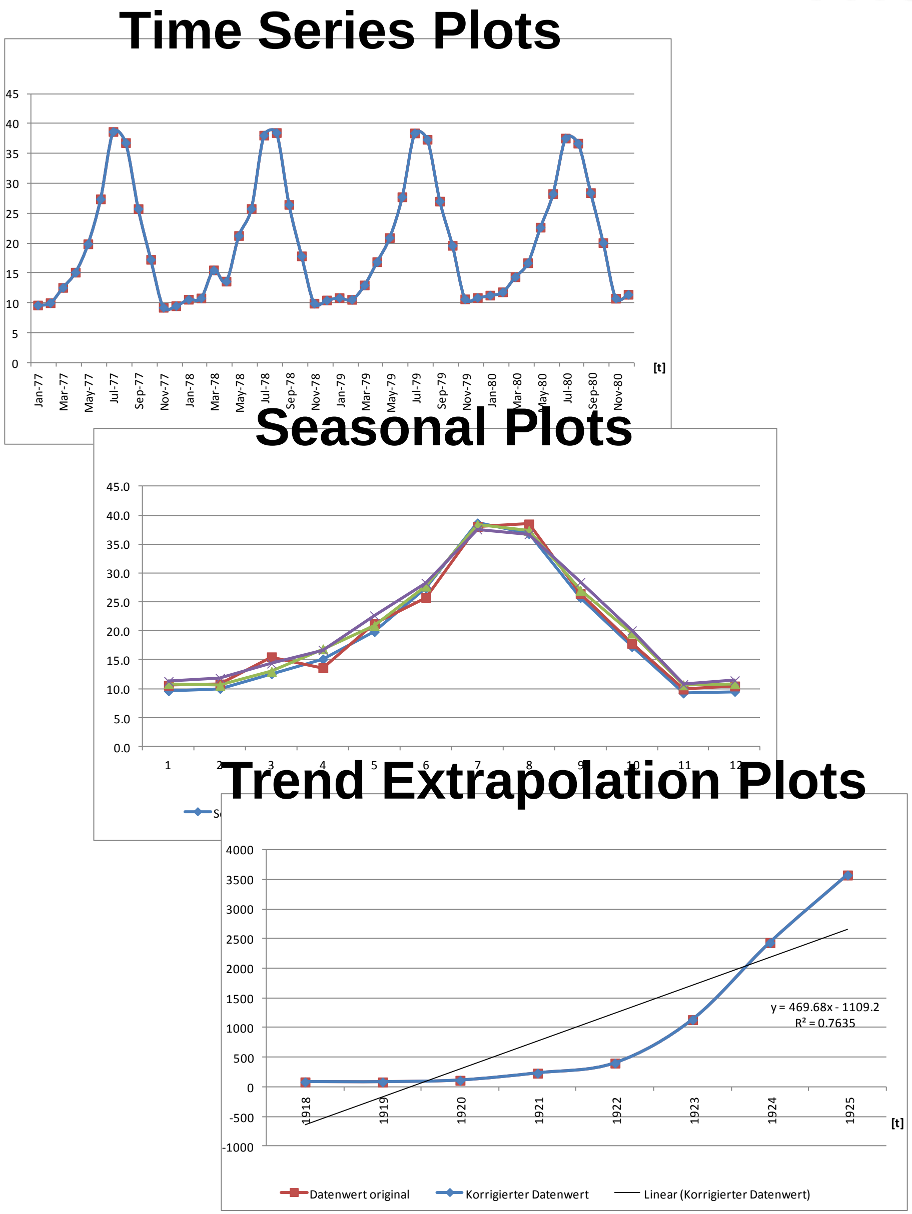


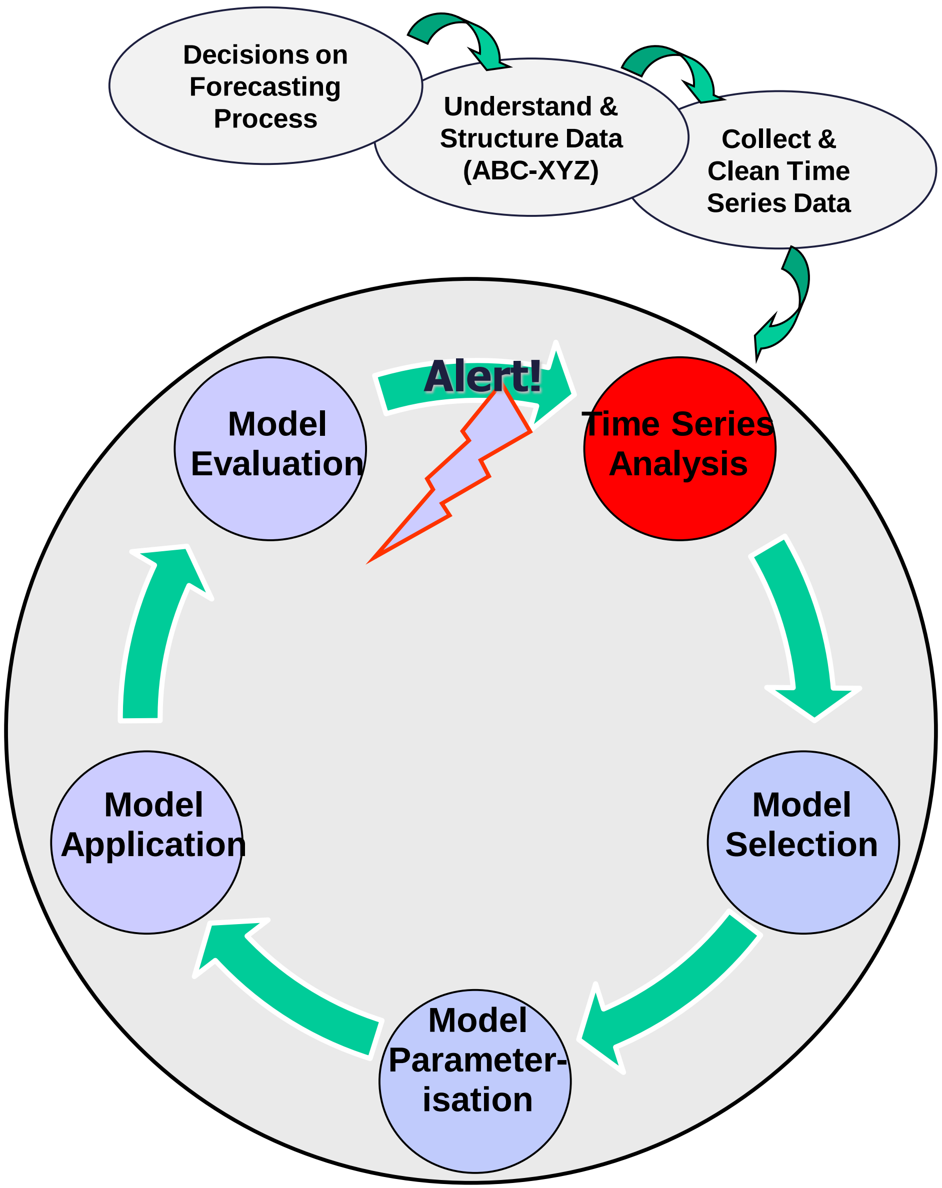




→ Data Exploration for Time Series data?

The Forecasting Process





Agenda

Time Series Analysis

1. Data Exploration
    1. Time Series Patterns
        1. Level, Trend & Season
        2. Outliers & Structural breaks
    2. Graphical Identification of Time Series Patterns
        1. Time Series & Seasonal Plots
        2. Time Series Decomposition
        3. A related „discipline“: Technical Analysis
    3. Statistical Identification of Time Series Patterns
        1. Autocorrelation Analysis (Corelograms)
        2. Spectral Analysis
        3. Statistical Tests
2. Data Transformations
    1. Transforming level of series
    2. Transforming variance of time series

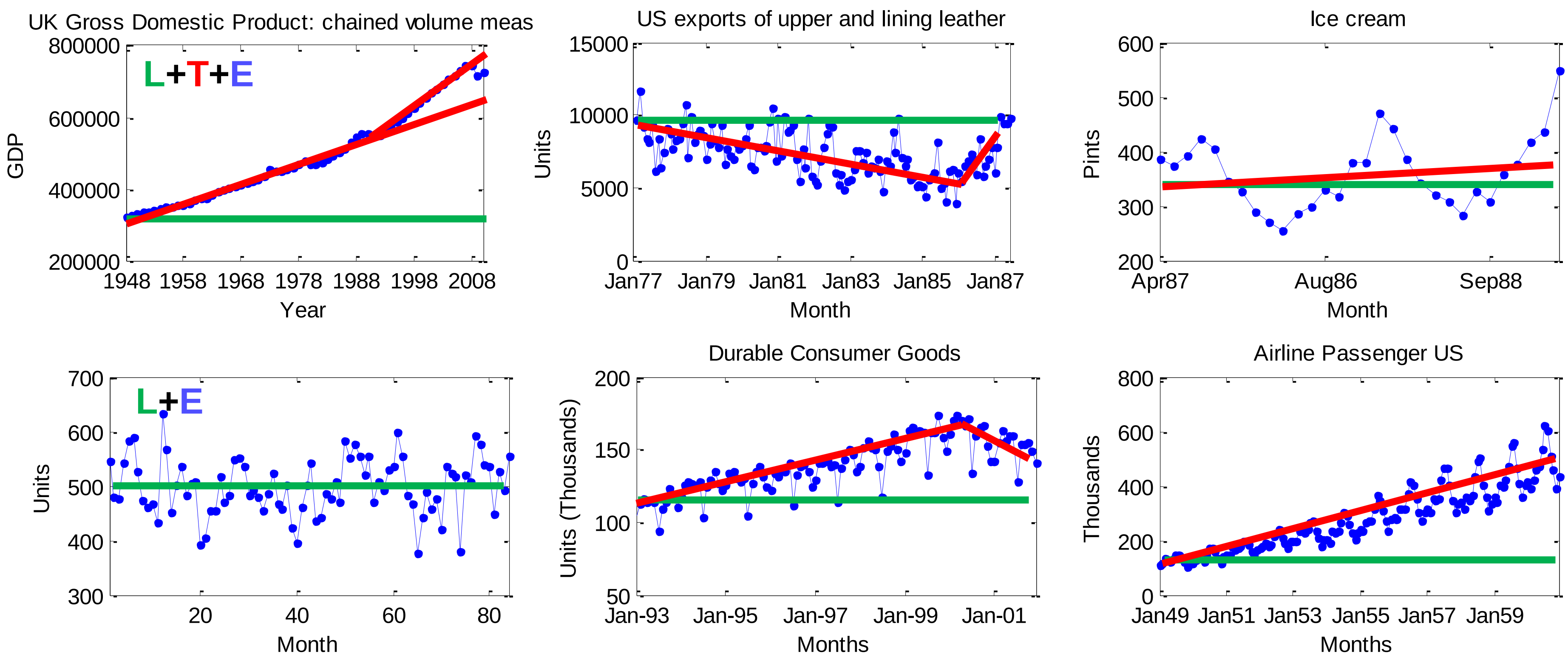
Concept of Time Series

- **An observed measurement is made up of a**
  - one or more <u>systematic / regular components</u> **and a**
  - one or more <u>irregular components</u>
    (stochastic part, noise, variability, randomness …)

Regular Components	Irregular Components
• Level • Trend • Season (one or multiple) • Cycle	• Noise • Outliers (& patches) • Structural breaks (in Level, Trend, Season)

- **Approach**
  - **Unfortunately we cannot observe either of these seperately**
  - **Forecasting methods try to isolate the systematic part from random part**
  - **Forecasts are based on the systematic part**
  - **The random part determines the distribution shape of errors / residuals**

Challenge of Time Series Prediction



Time series can have several different structural forms. These can be disaggregated into 4 main components:

(L)evel – (T)rend – (S)eason – (E)rror

Time series have always Level and Error and they may have Trend and/or Season. Constant time series have only Level and Error.

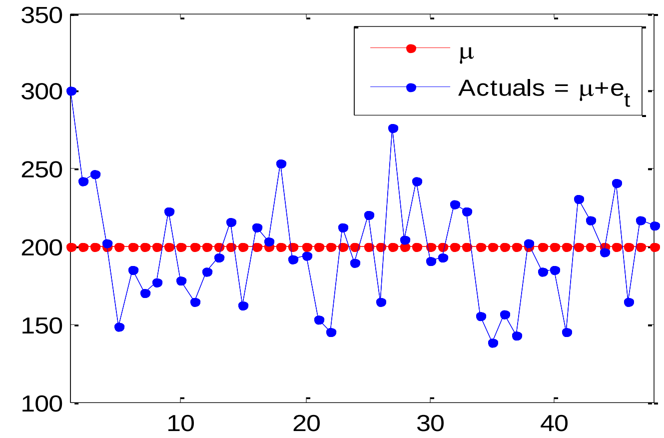
Time Series Components

Level

A stationary time series fluctuates around a constant level. Let $\{A_t\}$ be a stationary time series, then:

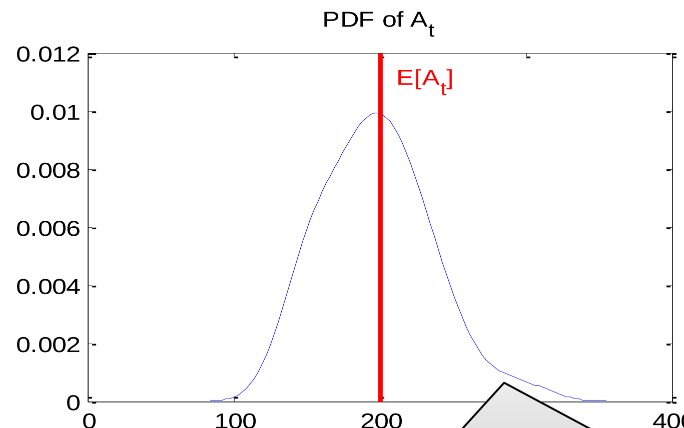
$$A_t = \mu + e_t \quad \text{and} \quad E[A] = \mu$$

- A_t has constant level μ over time.
- Noise e_t is overlaid on top of μ.
- The expected value of A_t is equal to μ.



<u>Strict stationarity</u> requires that the moments of the distribution of the time series do not change over time.

- The expected value E[A] is the 1st moment.
- The standard deviation of A, the noise in this case, is the 2nd moment → homoscedastic.
- Higher moments follow suit.



The deviation around the mean is the noise – unforecastable!

<u>Weak stationarity</u> requires only the 1st (or up to 2nd) moment of the distribution to be constant over time.

Therefore, for stationary time series the forecasting problem is how to identify a good estimation of the expected value, the rest is noise!

Lancaster Centre for Forecasting
www.forecasting-centre.com

Time Series Components

Trend

A long term upwards or downwards movement is a trend. Trend can be of two forms:

- **Deterministic:** $A_t = \mu + \beta t + e_t$

  where β is a constant, μ the level, e_t the random part and t is time.

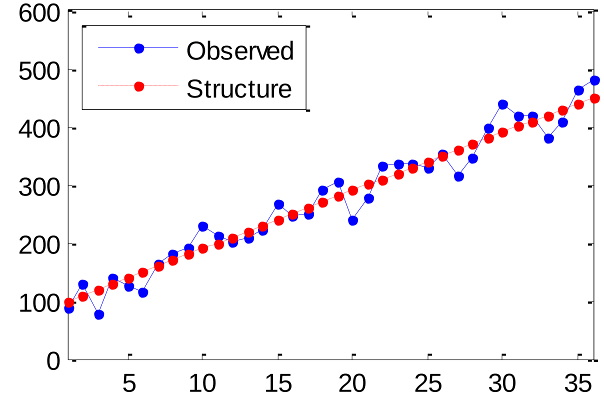
- **Stochastic:** $\phi \nabla A_t = A_t - \phi A_{t-1} = e_t$

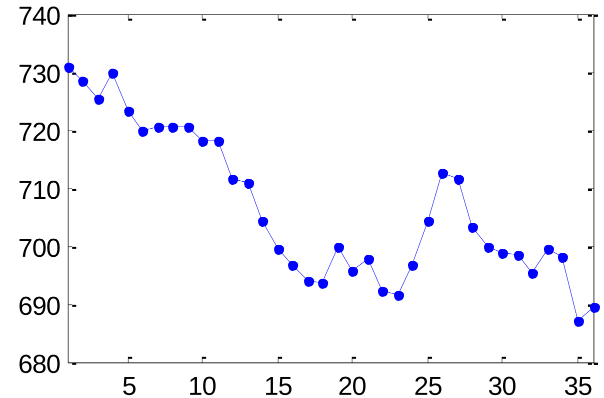
  where φ is a constant and delta is the difference operator.

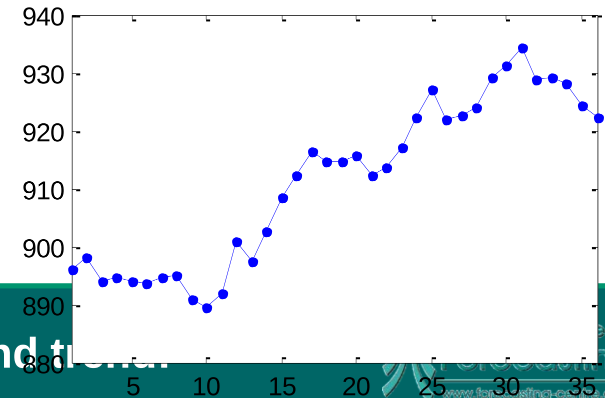
Similar to level, trend is overlaid by randomness e_t.

The forecasting problem for trended time series is:

$$E[A] = \mu + E[\nabla A]$$



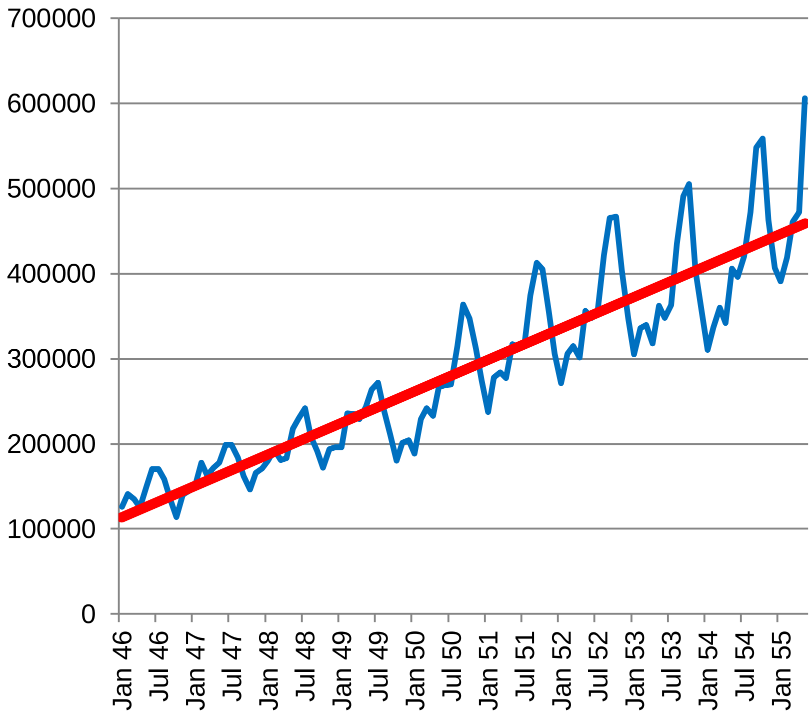




Therefore we can smooth noise separately in level and trend.

Trend Patterns

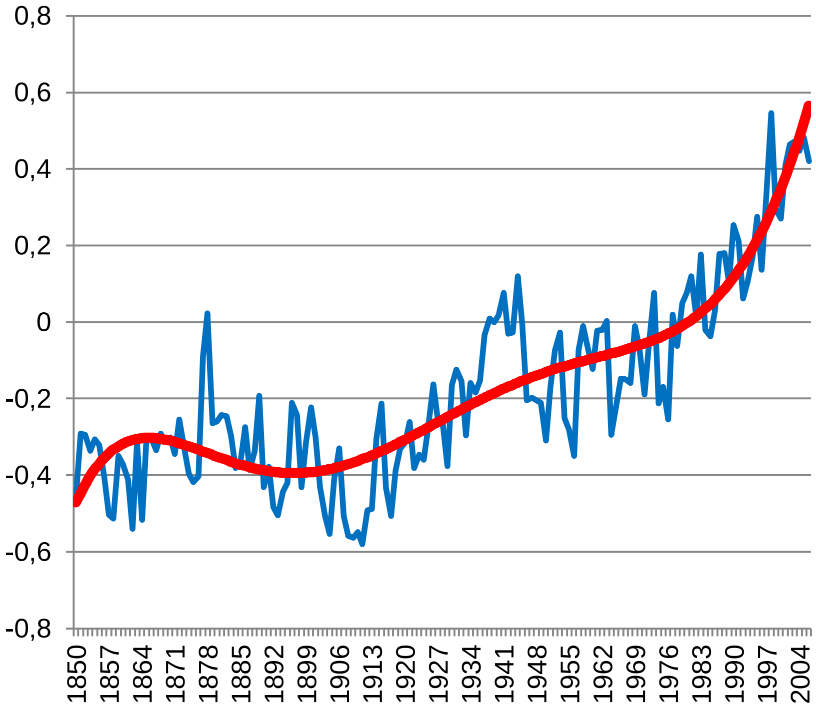
Deterministic (global) Trend



Trend remains constant over time

OR

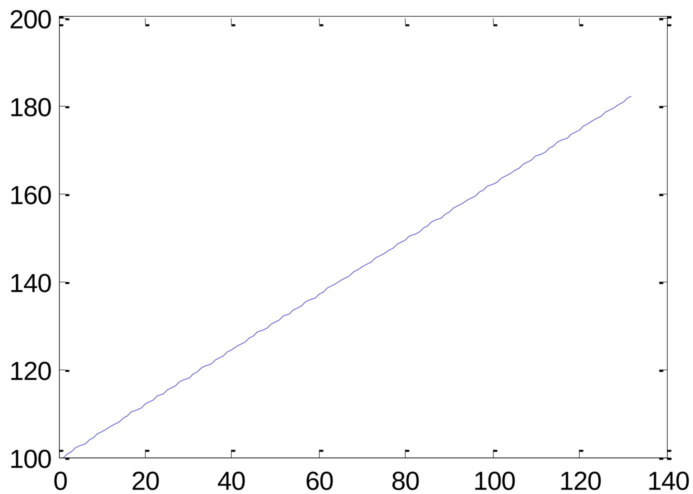
Stochastic (local) Trend



Trend changes over time

Deterministic Trend Patterns

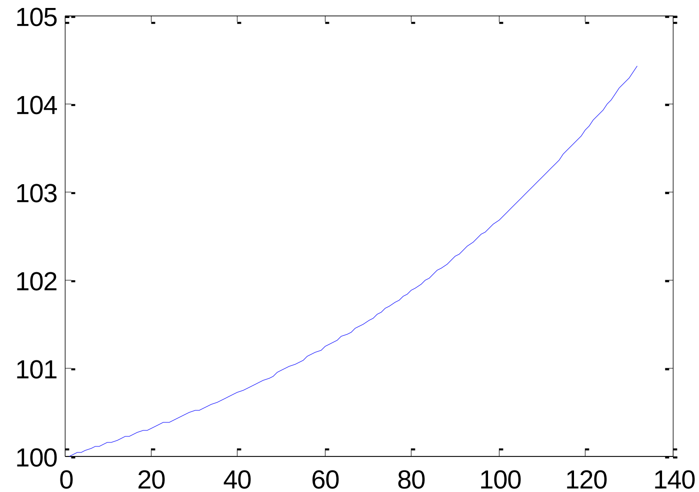
Linear Trend



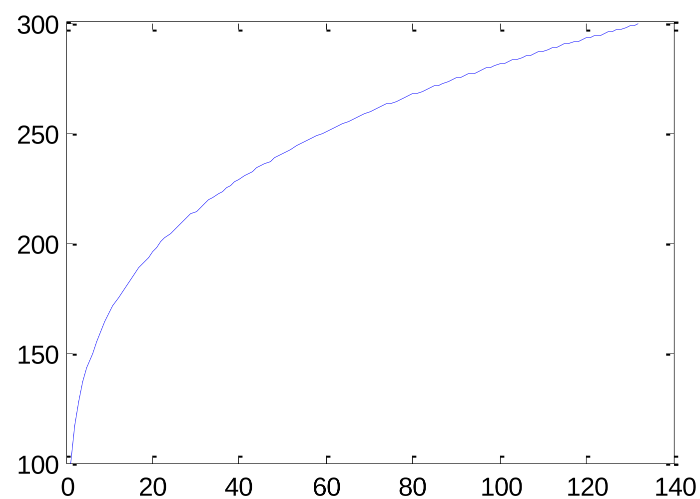
Steady change of the level over time

Nonlinear Trends

Exponential Trend



Damped Trend



Time Series Components

Season

A periodic pattern that can be of two forms:

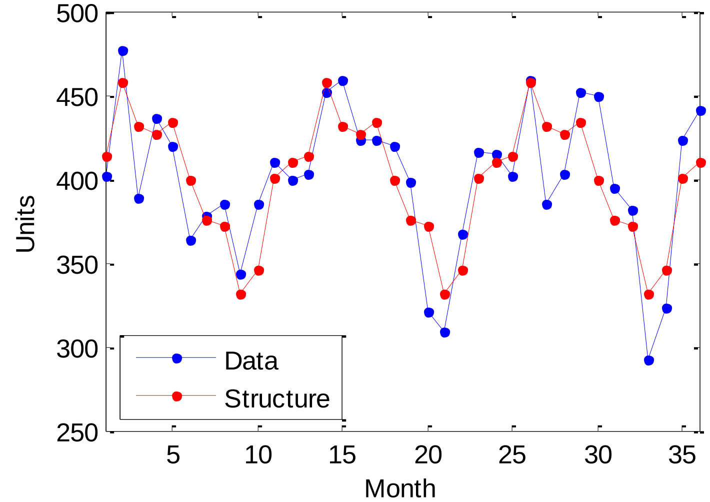
- **Deterministic:** $$A_t = \mu + \sum_{s=1}^{S} m_s d_{st} + e_t$$

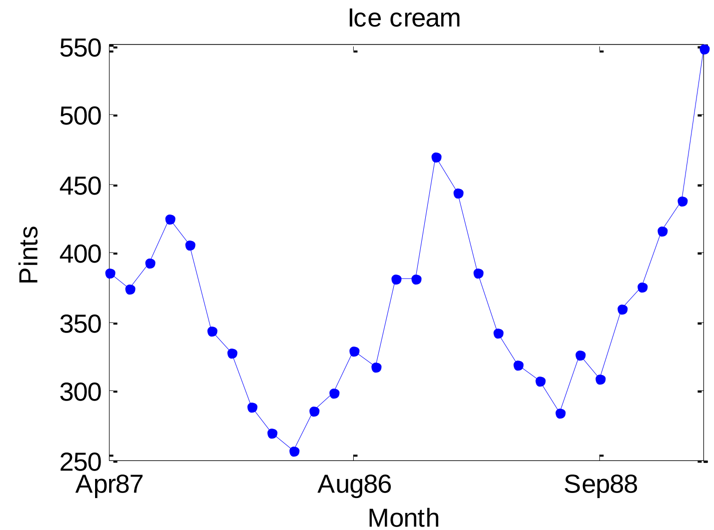
  where *S* is the length of the season, d_{st} are *S* dummy binary variables, *m* is a vector of *S* seasonal indices and μ is the level.

- **Stochastic:** $$\phi \nabla_S A_t = \phi(A_t - A_{t-s}) = \theta \varepsilon_\tau$$

  where φ and θ are constants and and delta is the difference operator, in this case of seasonal differences.

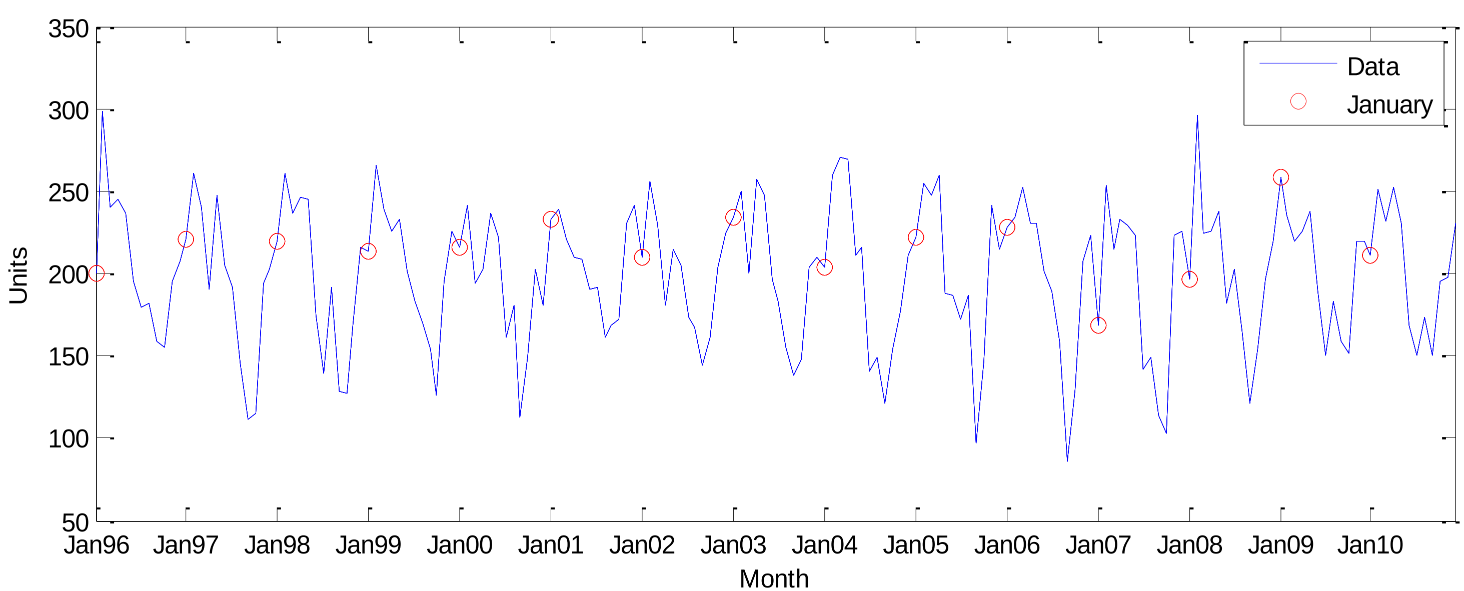
Season is overlaid by randomness e_t.

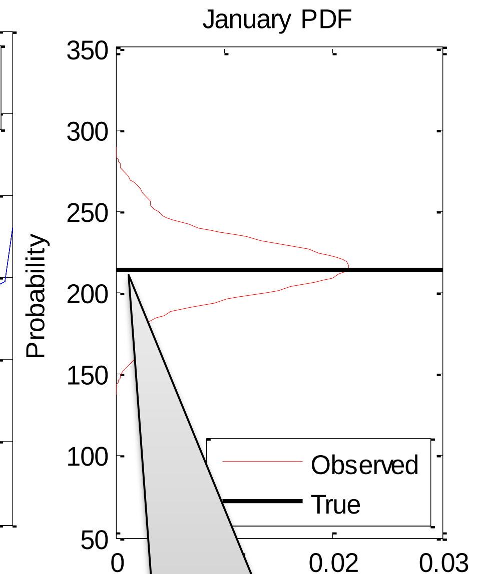




Time Series Components

Season



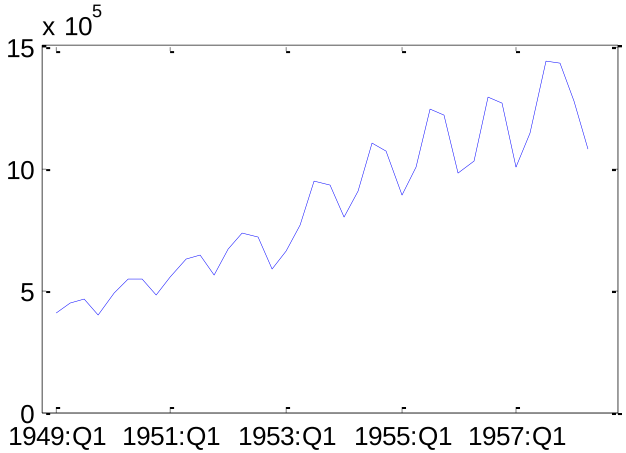


The deviation around the true January is due to noise

Similarly to level or trend, the seasonal component is affected by noise, which needs to be filtered

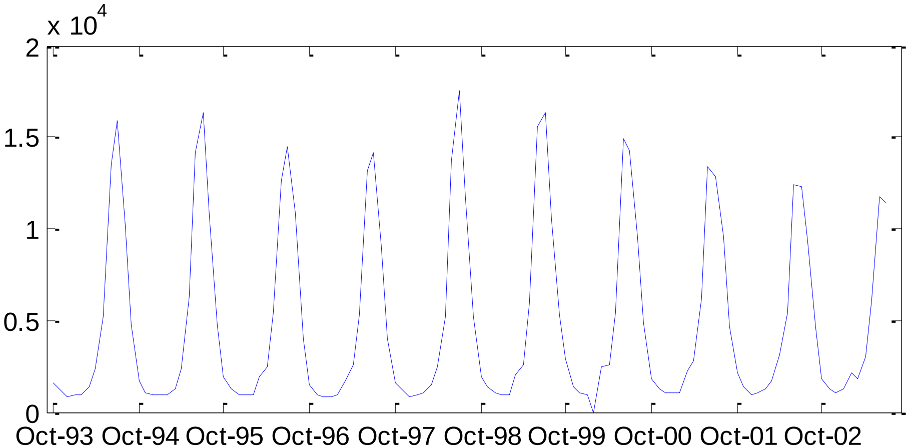
Seasonal Pattern

Quarterly Time Series



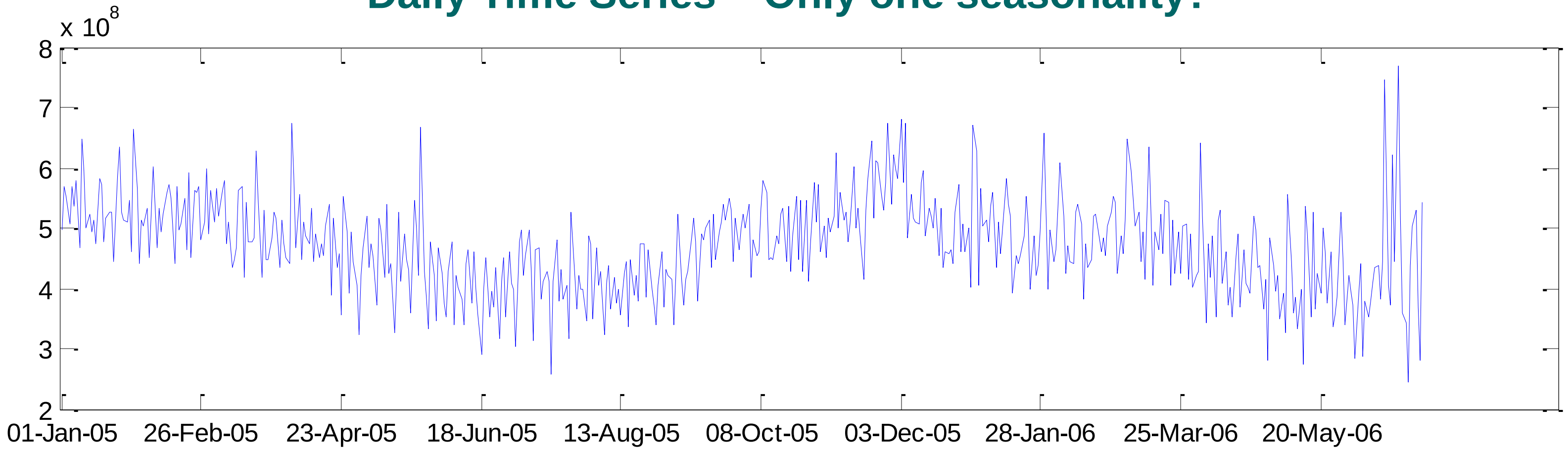
Airline Passenger Data

Monthly Time Series



U.S. Foreign Waterborne Freight- tons (thousands) - Alcan

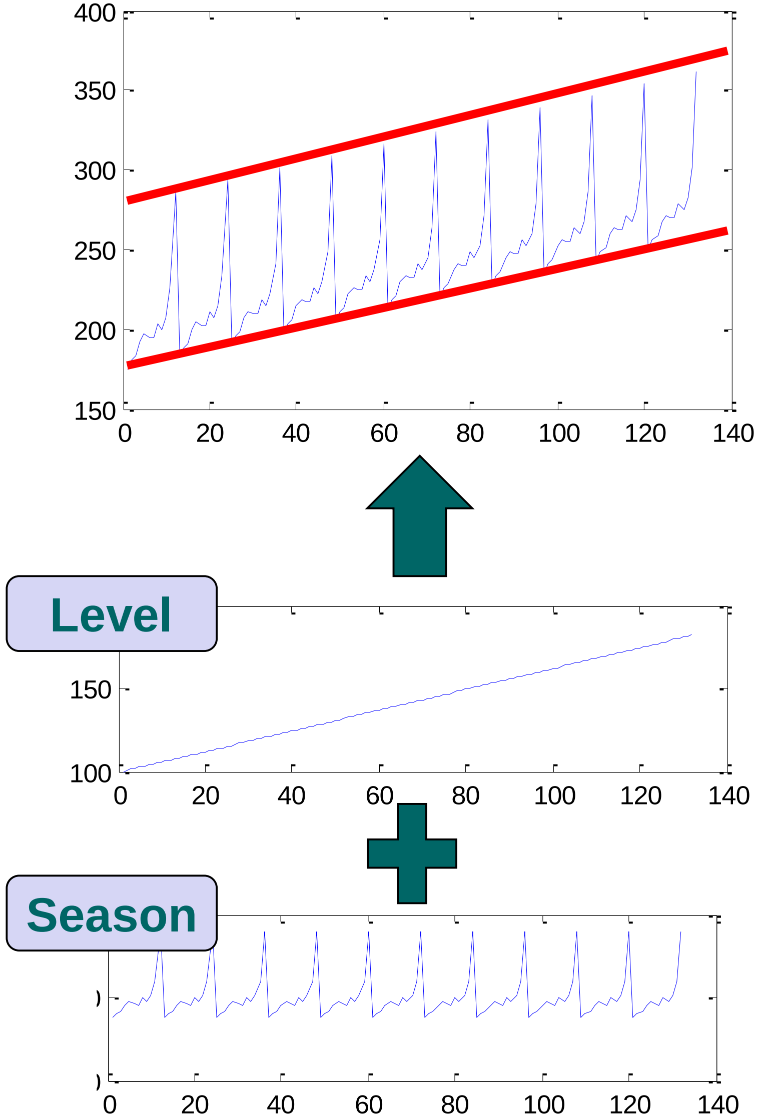
Daily Time Series – Only one seasonality?



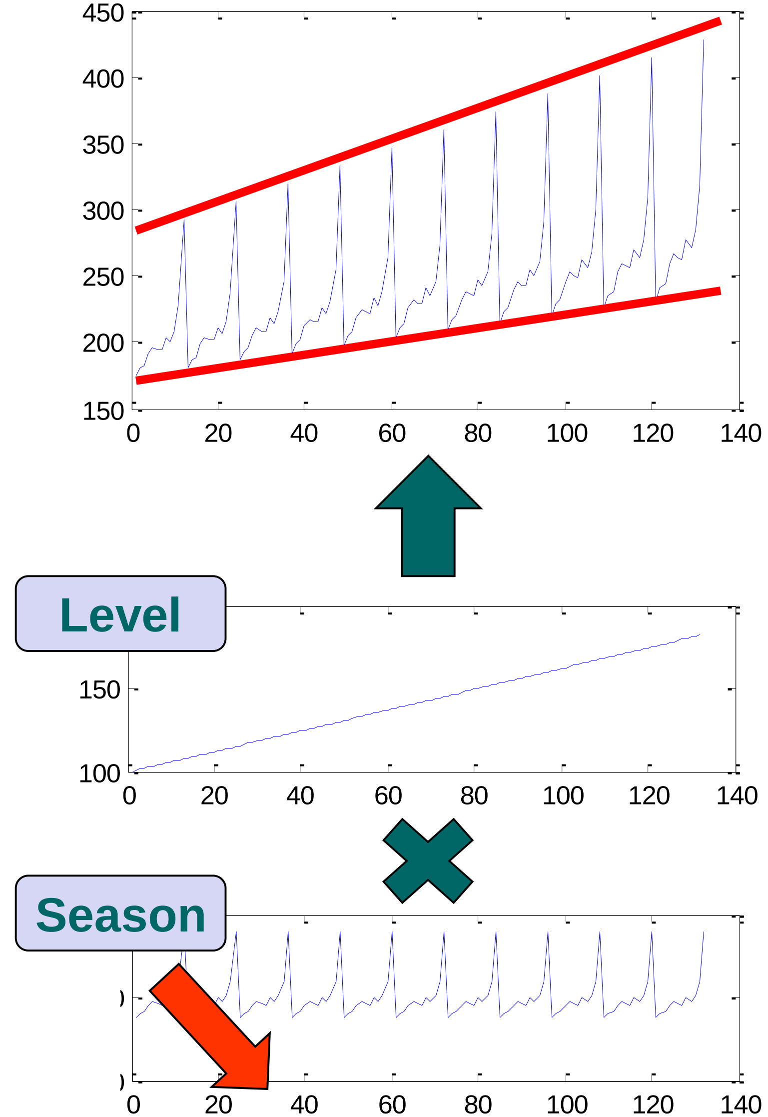
Television viewership - UK

Seasonal Pattern

Additive seasonality



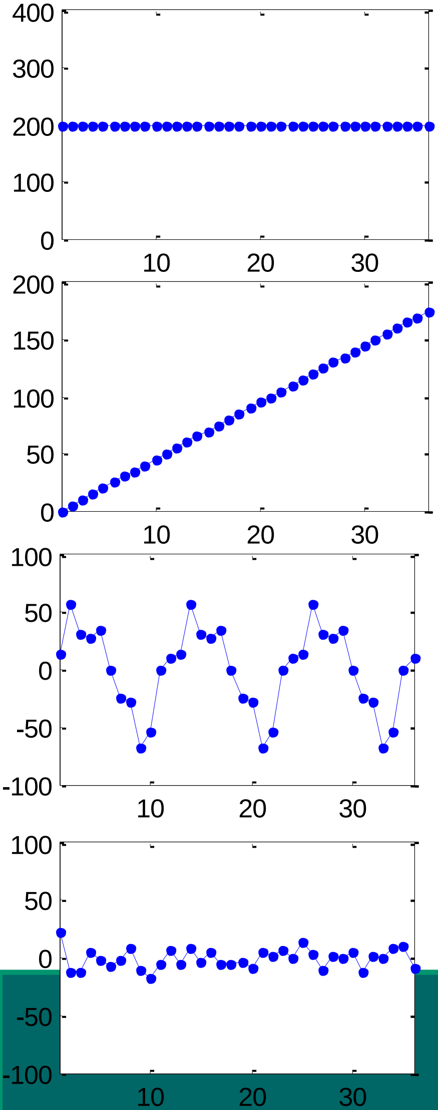
Multiplicative seasonality



Expressed in percentages!

Time Series Components
Additive Time Series

These components can be added together:



Level

+

Trend

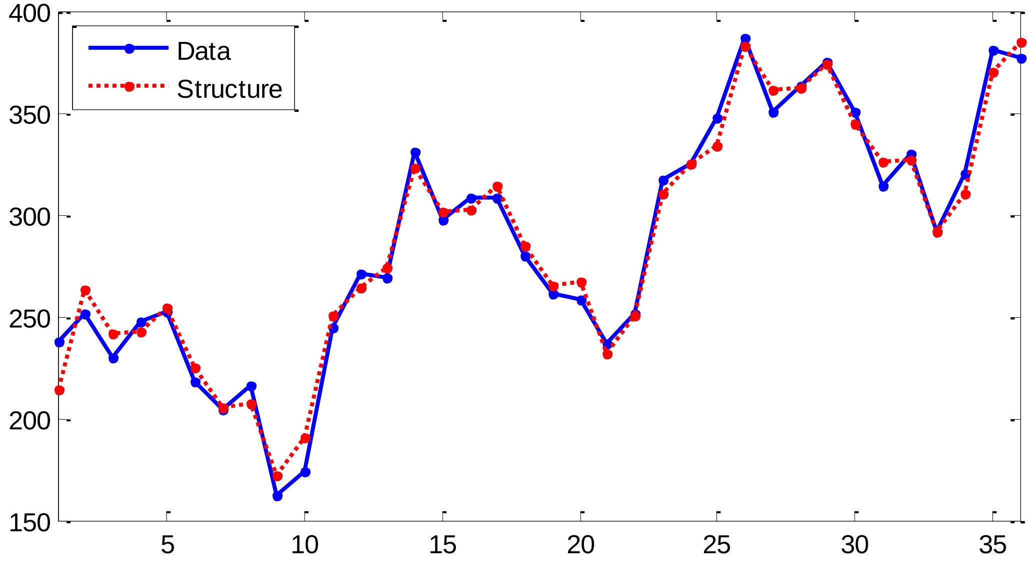
+

Season

+

Error

=



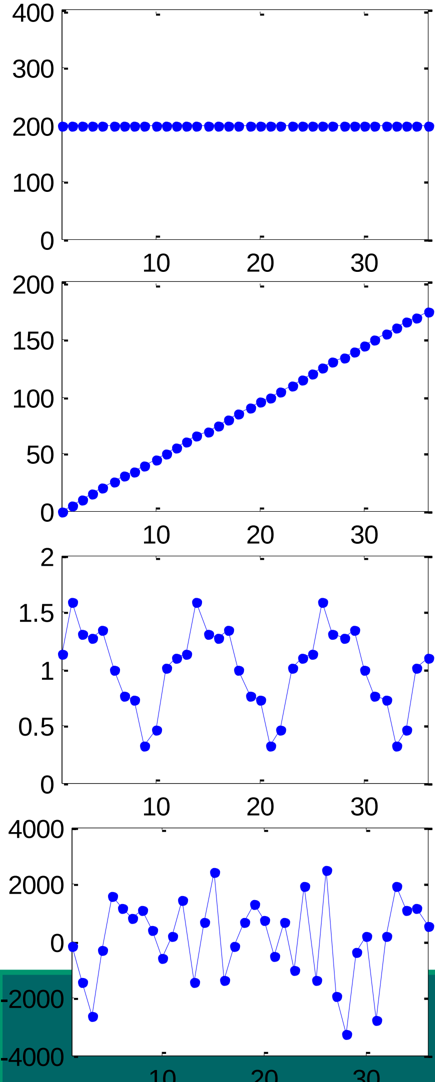
This is the additive form.

Lancaster Centre for Forecasting
www.forecasting-centre.com

Time Series Components

Multiplicative Time Series

... or multiplied together:



Level

×

Trend

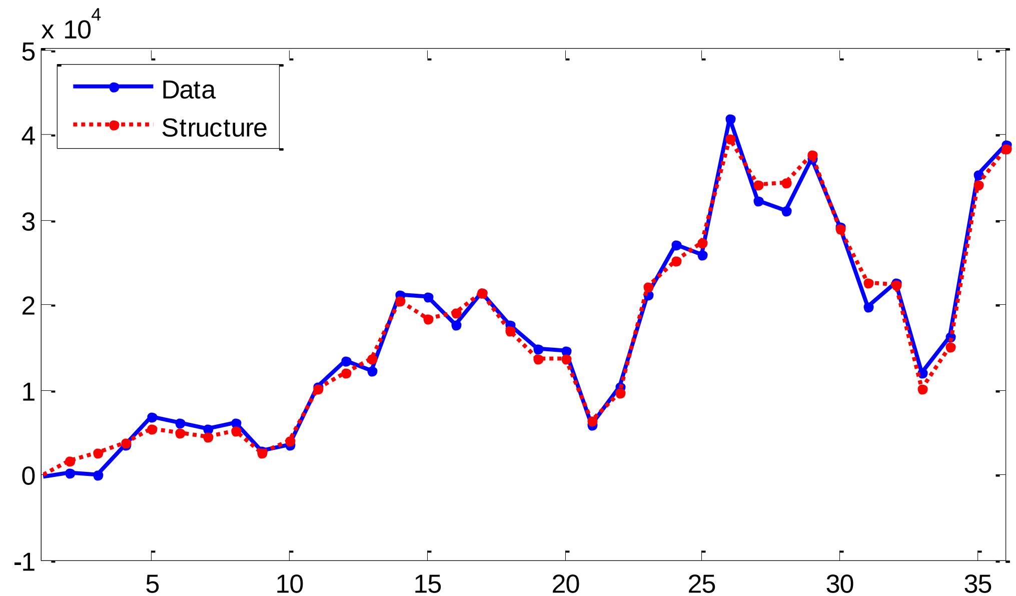
×

Season

+

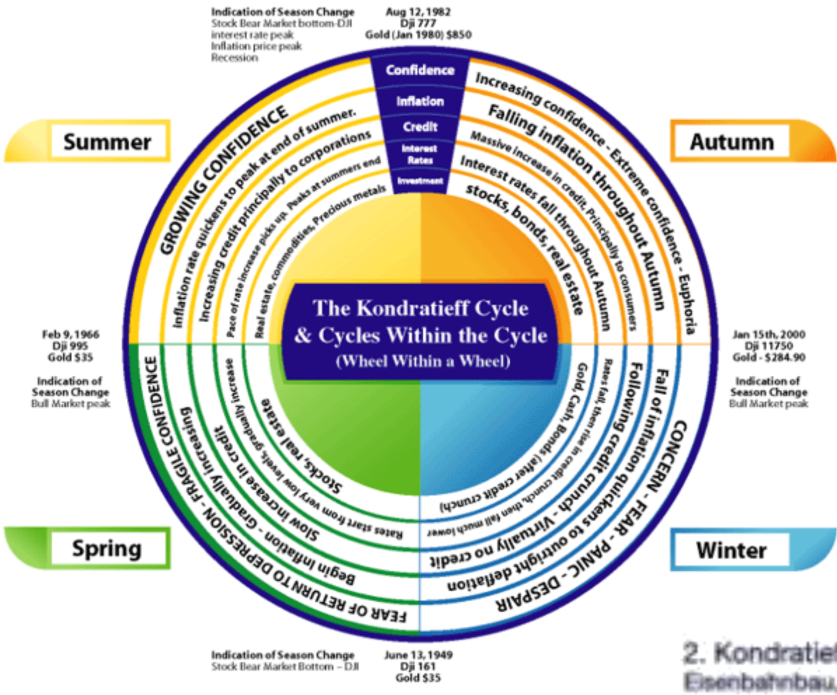
Error

=

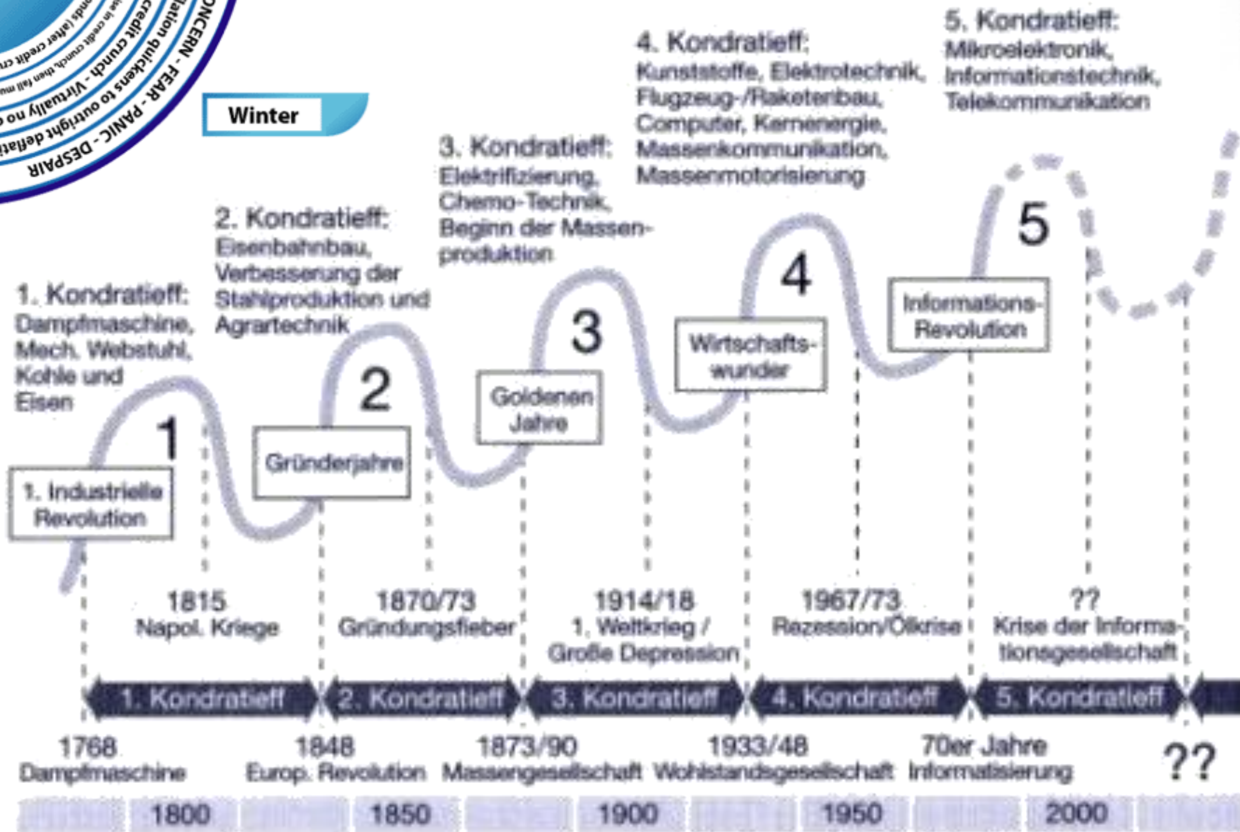


This is the <u>multiplicative</u> form. Mixed forms are common

The Kondratieff Business Wave cycle



Nikolai Dmitrijewitsch Kondratjew (1892–1938)



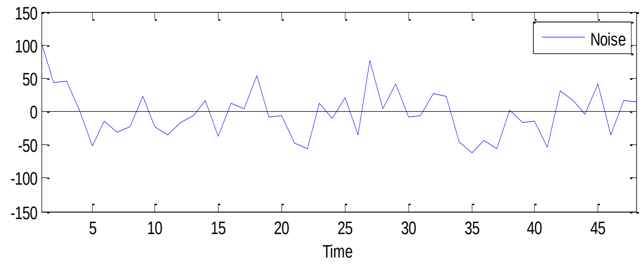
Agenda

Time Series Analysis

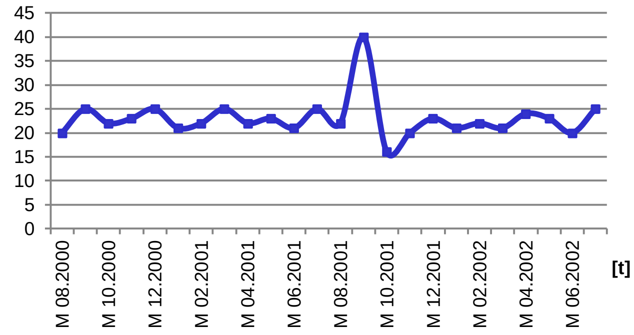
1. Data Exploration
    1. Time Series Patterns
        1. Level, Trend & Season
        2. Outliers & Structural breaks
    2. Graphical Identification of Time Series Patterns
        1. Time Series & Seasonal Plots
        2. Time Series Decomposition
        3. A related „discipline“: Technical Analysis
    3. Statistical Identification of Time Series Patterns
        1. Autocorrelation Analysis (Corelograms)
        2. Spectral Analysis
        3. Statistical Tests
2. Data Transformations
    1. Transforming level of series
    2. Transforming variance of time series

Irregular Components of Time Series

- **Noise**
  - **Constant or non-constant randomness around the regular components**

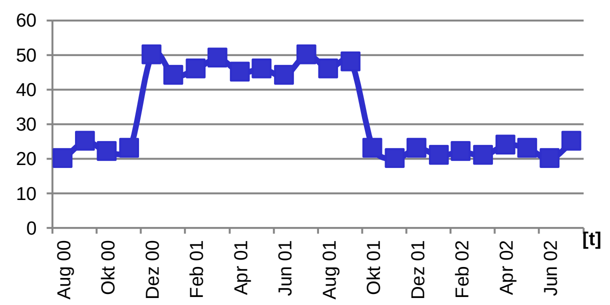


- **Pulses / Outliers**
  - **one time occurrence**
  - **on top of systematic stationary / trended / seasonal patterns**

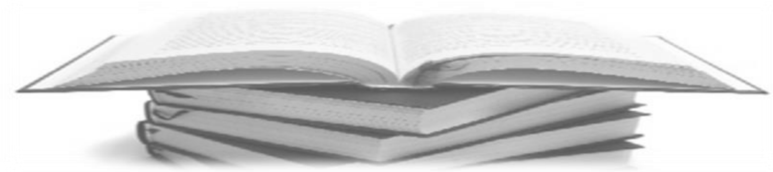


STATIONARY time series with PULSE

- **Structural Breaks**
  - **one time / multiple level shifts**
  - **Local time trends / trend breaks**
  - **Changing seasonal shapes**
  - **on top of systematic stationary / trended / seasonal patterns**



STATIONARY time series with level shift

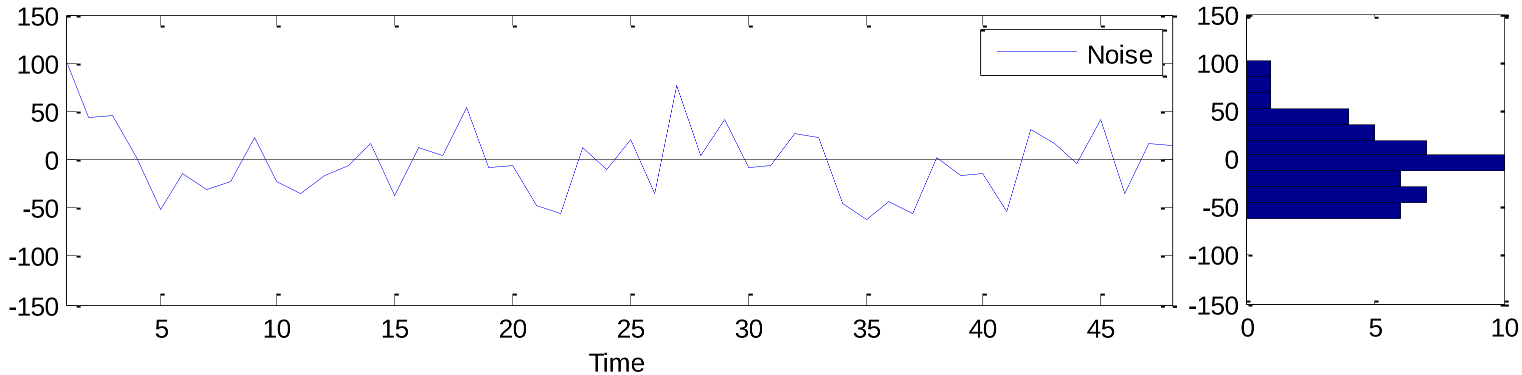


Forecasting Basics

Time Series Structure and Noise

Noise (or random part) is inherently unforecastable since it has no structure, <u>otherwise it should be captured by the forecasting method used</u>.

Noise in practice consists of all the information that is impossible to collect or observe, such as the actions of individuals or the mechanics of ill-understood systems. As such it follows an unknown distribution that has no structure.



It is impossible to predict whether the noise time series will go up or down, as there is no structure. We may be fooled in seeing patterns in the noise, but there are none! Typically once more sample is collected this becomes clear → Noise cannot be forecasted!

Note that no structure means <u>NO DYNAMICS</u>, <u>NO TREND</u> and <u>NO SEASONALITY</u> in the noise, on the long run it just randomly fluctuates around zero.

Outliers

There are two forms of outliers:

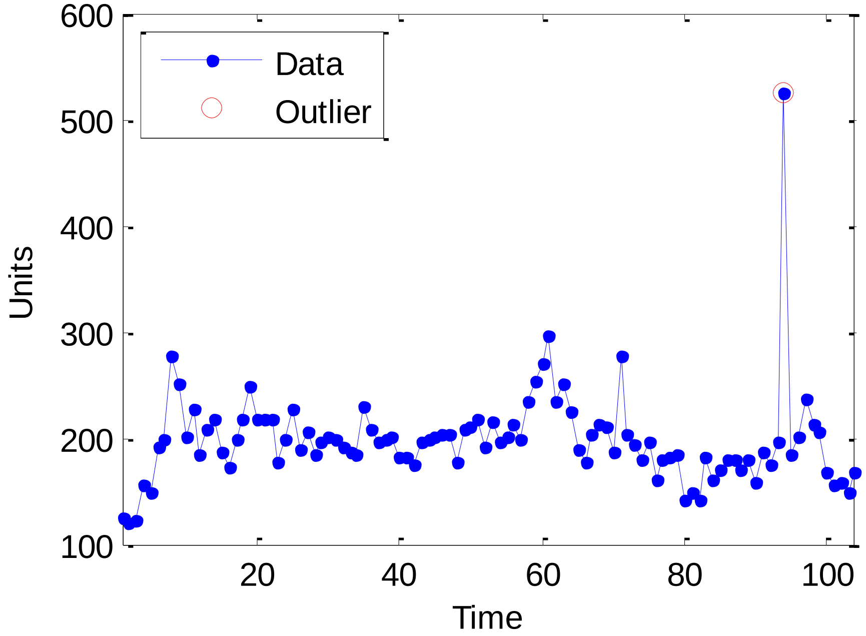
1. **Additive**
2. **Innovation**

Additives outliers can be seen as:

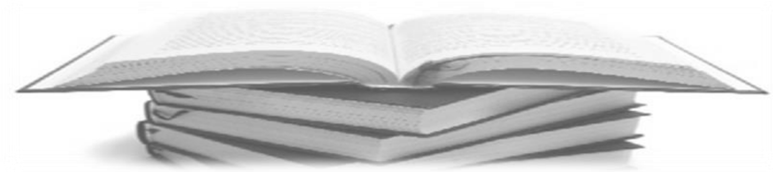
$$A_t = \mu + e_t + dI(t = \tau)$$

At time $t=\tau$, marked by indicator variable I, d is added to the level μ.

Innovation outliers occur in the errors instead.



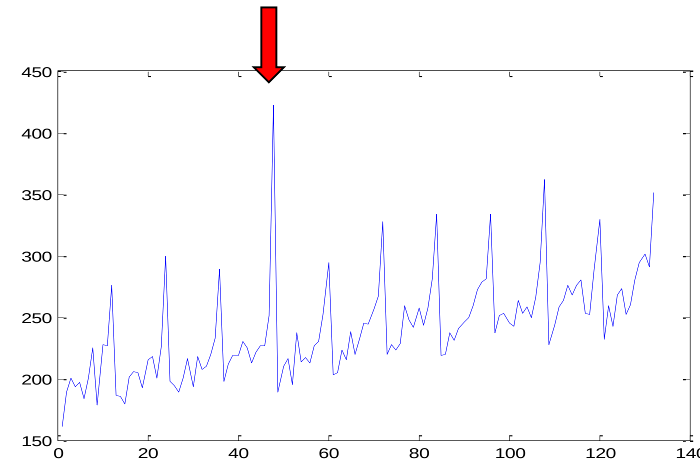
Outliers can cause problems to the estimation of the expected value, i.e. in forecasting future values.



Forecasting Process

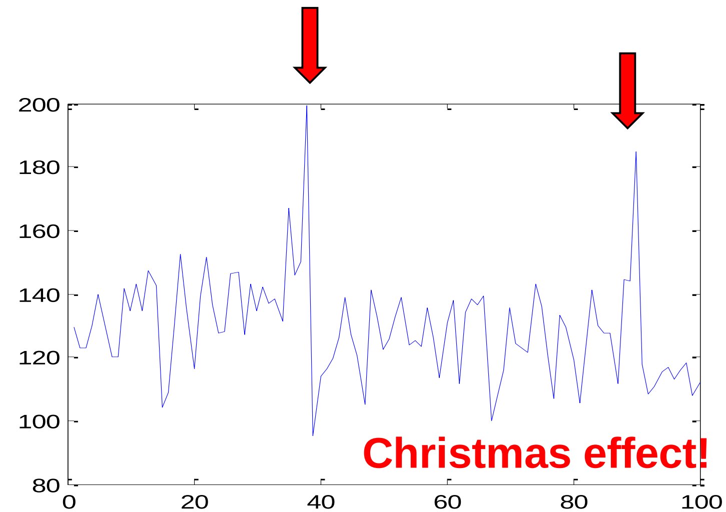
Statistical Modelling: Time Series Analysis

Irregular time series components



Outlier (pulse)

- one time occurrence
- on top of systematic stationary / trended / seasonal structure
- Various causes:
    - Special (unknown) one-off events such as NBA finals, royal wedding, etc.



Seasonal Outliers (pulses) due to seasonally reoccurring events!

- E.g. christmas peak
  → special cases of seasonality

Level shifts

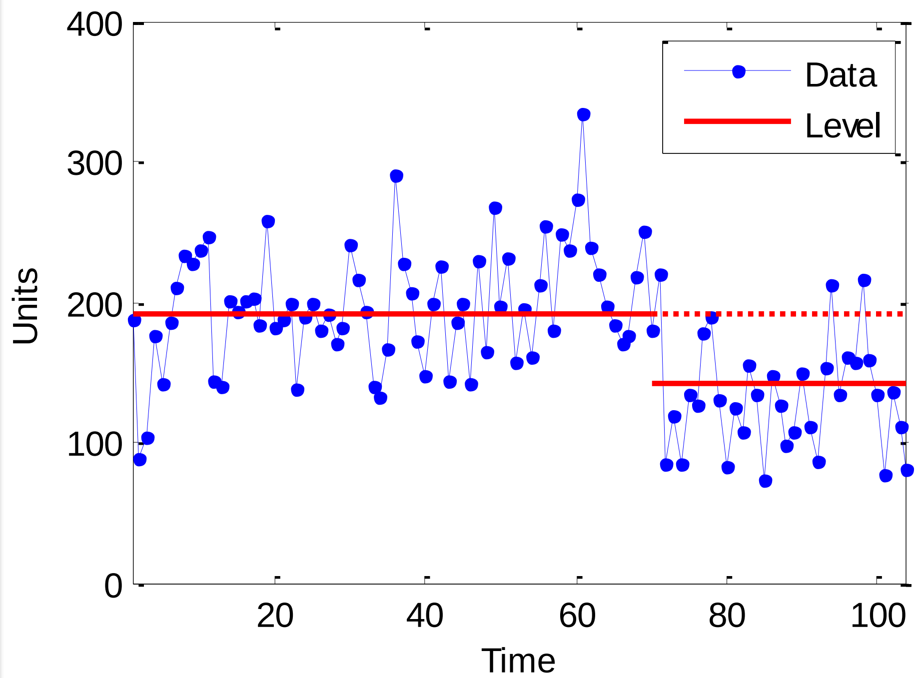
These can be interpreted as persistent outliers:

$$A_t = \mu + e_t + dI(t = \tau)$$

where τ is a vector of time indices, or as a breaks in the structure of the time series:

$$A_t = \begin{cases} \mu_1 + e_t, \forall t \le \tau \\ \mu_2 + e_t, \forall t > \tau \end{cases}$$

where a new level μ_2 is reached after time τ.



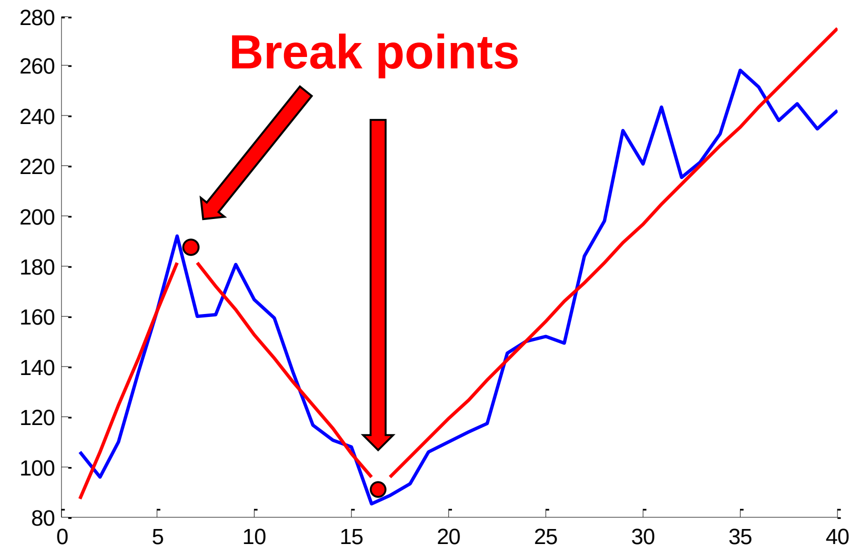
Level shifts can cause problems to the estimation of the expected value, i.e. in forecasting future values.

Irregular Components of Time Series

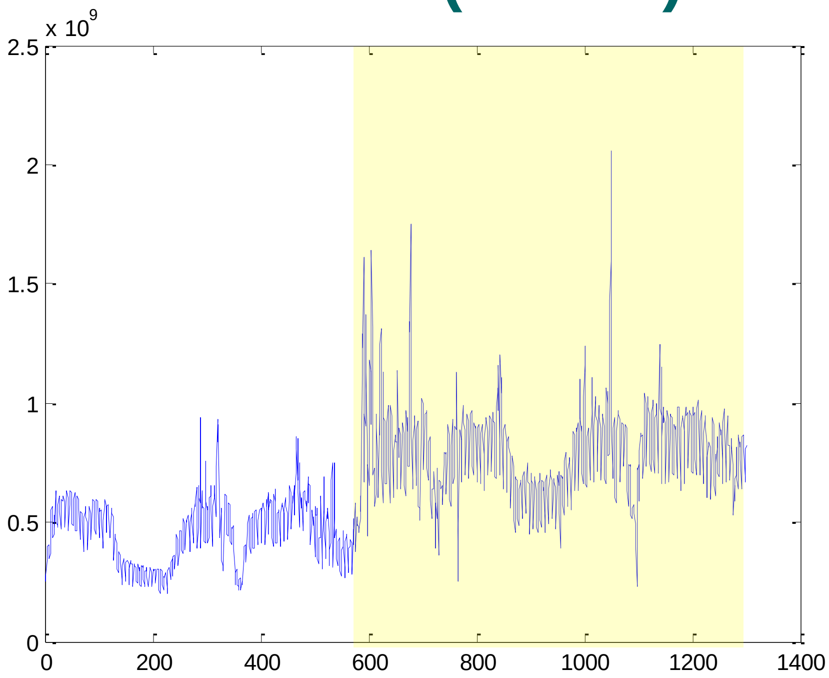
STRUCTURAL BREAKS

- A change in one of the components
  - Level, Trend, Season…

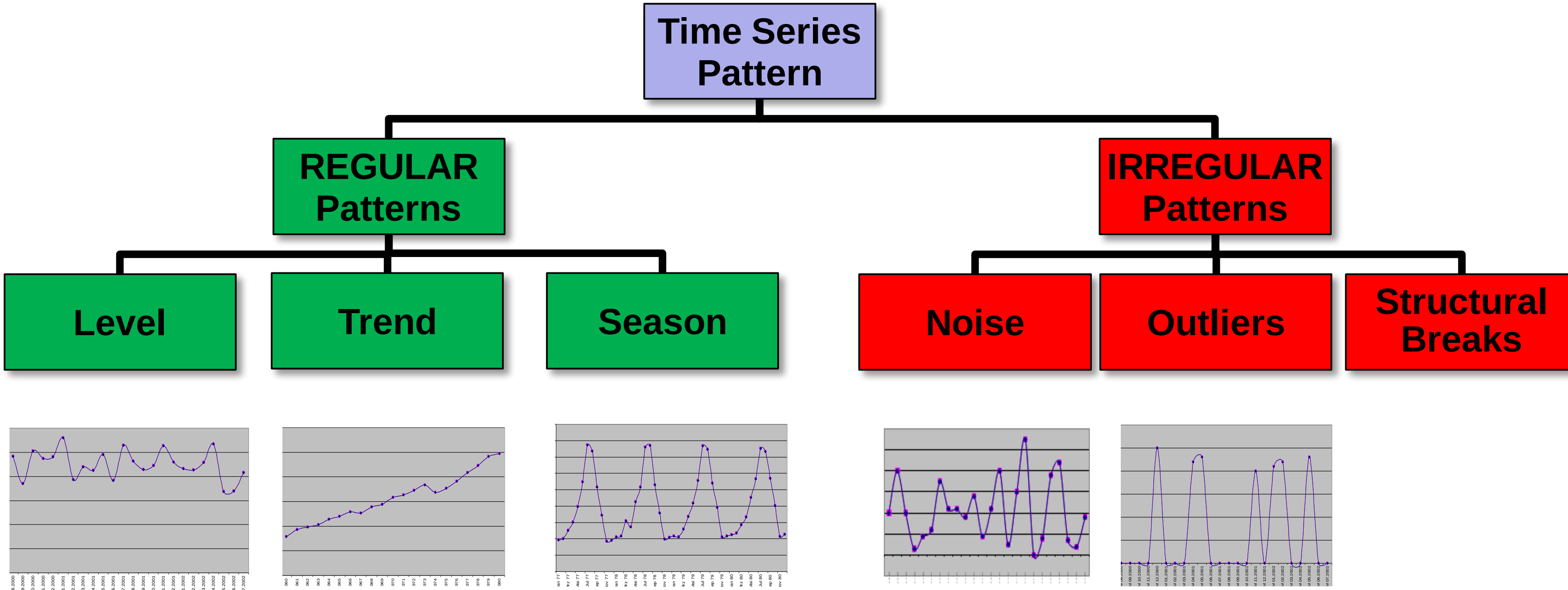
Trend



Season (+Level)



Time Series Patterns



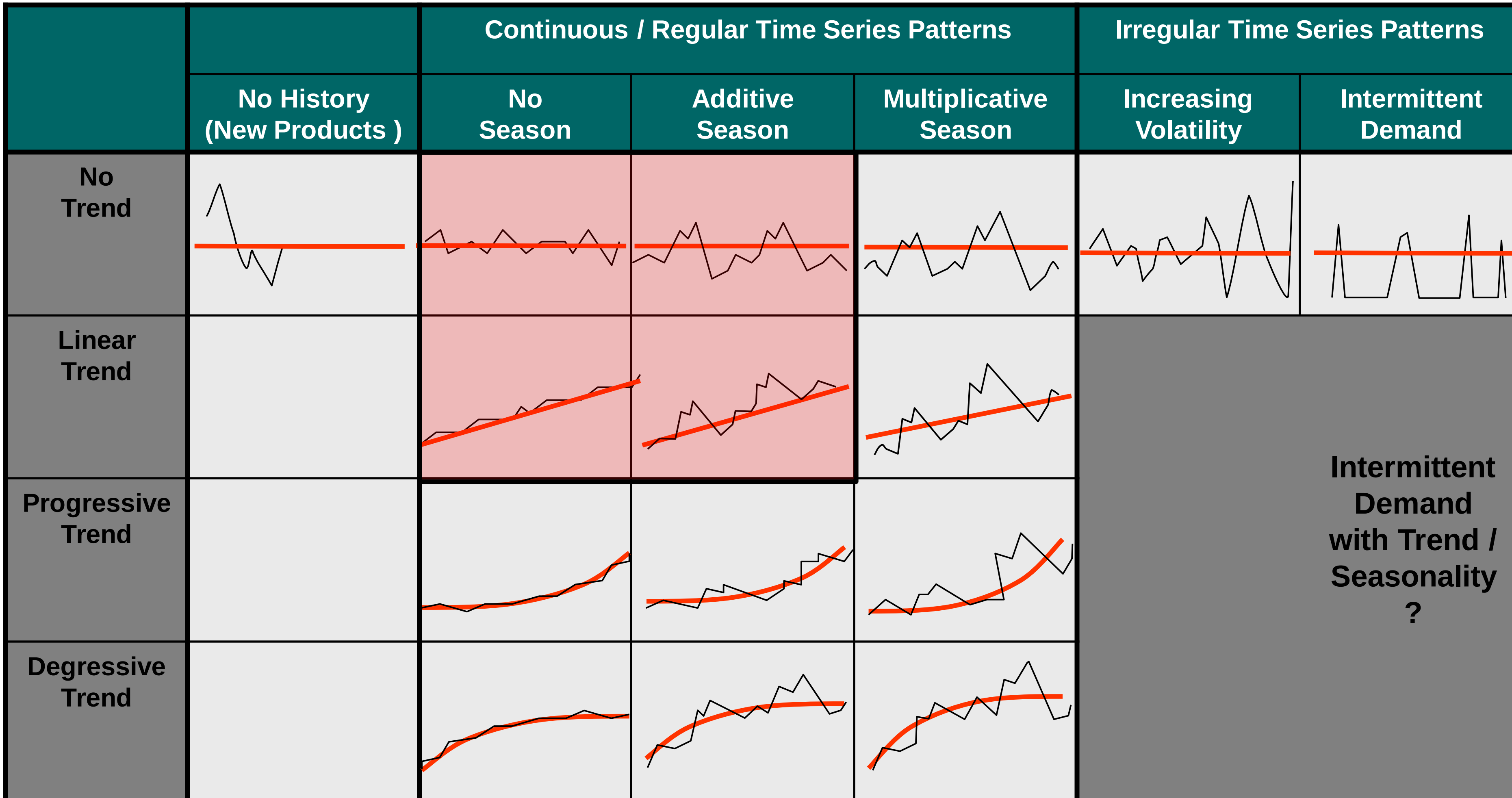
Actual Time Series are a combination of individual Components

$$Y_t = f(L_{t,} S_t, T_t, E_{t\,,} P_{t,} S_t)$$

Time series can have several different structural forms. These can be disaggregated into 4 main components:

(L)evel – (T)rend – (S)eason – (E)rror

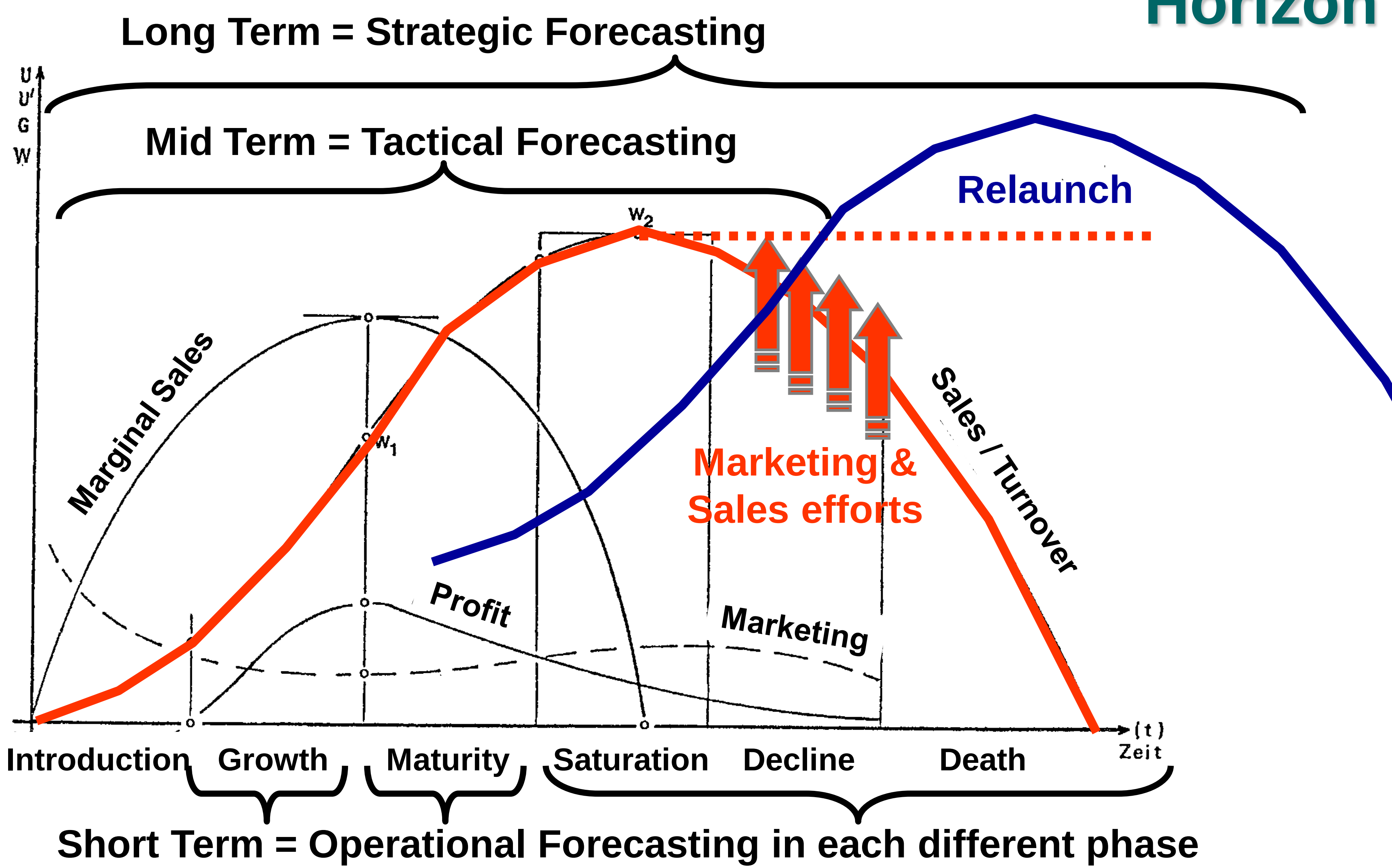
Time Series Patterns



Pegels [1969], Gardner [1985]

- → **12 archetypical systematic time series patterns (4 main ones)**
  - → **Extensions: New Products, Intermittent Series, multiple Seasonality, Structural Breaks / stochastic patterns (level shifts …), Outliers**
    - → **External influences: Calendar effects, controllable policies etc.**

Forecasting Horizon



→ What constitutes the shortest / longest horizon for you? Operational planning?

Agenda

Time Series Analysis

1. Data Exploration
    1. Time Series Patterns
        1. Level, Trend & Season
        2. Outliers & Structural breaks
    2. Graphical Identification of Time Series Patterns
        1. Time Series & Seasonal Plots
        2. Time Series Decomposition
        3. A related „discipline“: Technical Analysis
    3. Statistical Identification of Time Series Patterns
        1. Autocorrelation Analysis (Corelograms)
        2. Spectral Analysis
        3. Statistical Tests
2. Data Transformations
    1. Transforming level of series
    2. Transforming variance of time series

Numbers or graphs?

Jan 2002	Feb 2002	Mrz 2002	Apr 2002	Mai 2002	Jun 2002	Jul 2002	Aug 2002	Sep 2002	Okt 2002	Nov 2002	Dez 2002
447024	165048	282449	255336	271888	199464	274848	166500	235440	333350	339792	316078

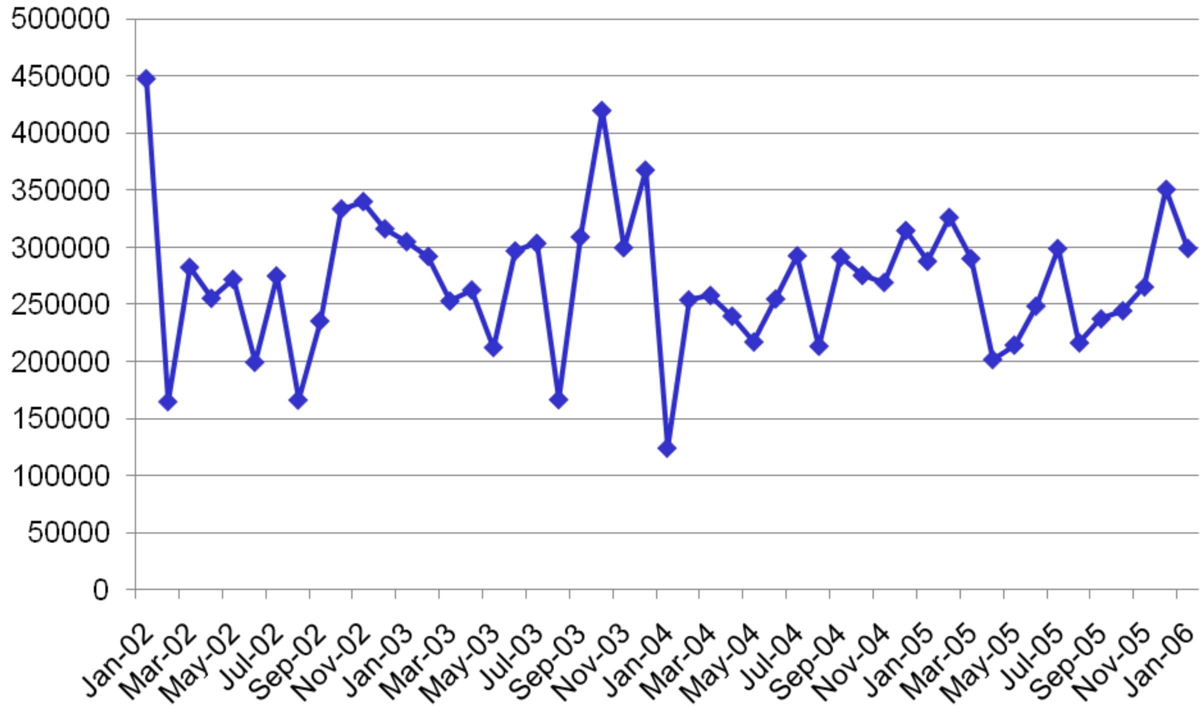
Jan 2002	Feb 2002	Mrz 2002	Apr 2002	Mai 2002	Jun 2002	Jul 2002	Aug 2002	Sep 2002	Okt 2002	Nov 2002	Dez 2002
447024	165048	282449	255336	271888	199464	274848	166500	235440	333350	339792	316078
Jan 2003	Feb 2003	Mrz 2003	Apr 2003	Mai 2003	Jun 2003	Jul 2003	Aug 2003	Sep 2003	Okt 2003	Nov 2003	Dez 2003
304776	291960	252912	262536	212376	296760	303336	166848	308872	419295	299514	367297

Jan 2002	Feb 2002	Mrz 2002	Apr 2002	Mai 2002	Jun 2002	Jul 2002	Aug 2002	Sep 2002	Okt 2002	Nov 2002	Dez 2002
447024	165048	282449	255336	271888	199464	274848	166500	235440	333350	339792	316078
Jan 2003	Feb 2003	Mrz 2003	Apr 2003	Mai 2003	Jun 2003	Jul 2003	Aug 2003	Sep 2003	Okt 2003	Nov 2003	Dez 2003
304776	291960	252912	262536	212376	296760	303336	166848	308872	419295	299514	367297
Jan 2004	Feb 2004	Mrz 2004	Apr 2004	Mai 2004	Jun 2004	Jul 2004	Aug 2004	Sep 2004	Okt 2004	Nov 2004	Dez 2004
124464	253992	257832	239688	217224	254760	292512	213504	291204	275280	269112	314491
Jan 2005	Feb 2005	Mrz 2005	Apr 2005	Mai 2005	Jun 2005	Jul 2005	Aug 2005	Sep 2005	Okt 2005	Nov 2005	Dez 2005
287664	325944	290040	201884	214488	248496	298752	216432	237432	244452	265224	350424

	Jan	Feb	Mrz	Apr	Mai	Jun	Jul	Aug	Sep	Okt	Nov	Dez
2002	447024	165048	282449	255336	271888	199464	274848	166500	235440	333350	339792	316078
2003	304776	291960	252912	262536	212376	296760	303336	166848	308872	419295	299514	367297
2004	124464	253992	257832	239688	217224	254760	292512	213504	291204	275280	269112	314491
2005	287664	325944	290040	201884	214488	248496	298752	216432	237432	244452	265224	350424

Numbers or graphs?

- **What is the pattern? Horizontal (→stationary)**



Numbers or graphs?

- **What is the pattern? (4 years of data)**

Time Series 1

	Jan	Feb	Mrz	Apr	Mai	Jun	Jul	Aug	Sep	Okt	Nov	Dez
2002	48918	38286	45621	40104	33485	24762	25924	18492	37887	51947	44994	35097
2003	43068	40224	39798	33216	38184	22782	25188	17172	31728	39114	38346	41052
2004	31968	29478	39504	32472	29574	28542	23592	21948	35970	33498	34500	18828
2005	29214	32868	28770	26874	26628	18840	22338	20778	30342	33390	35280	30942

Time Series 2

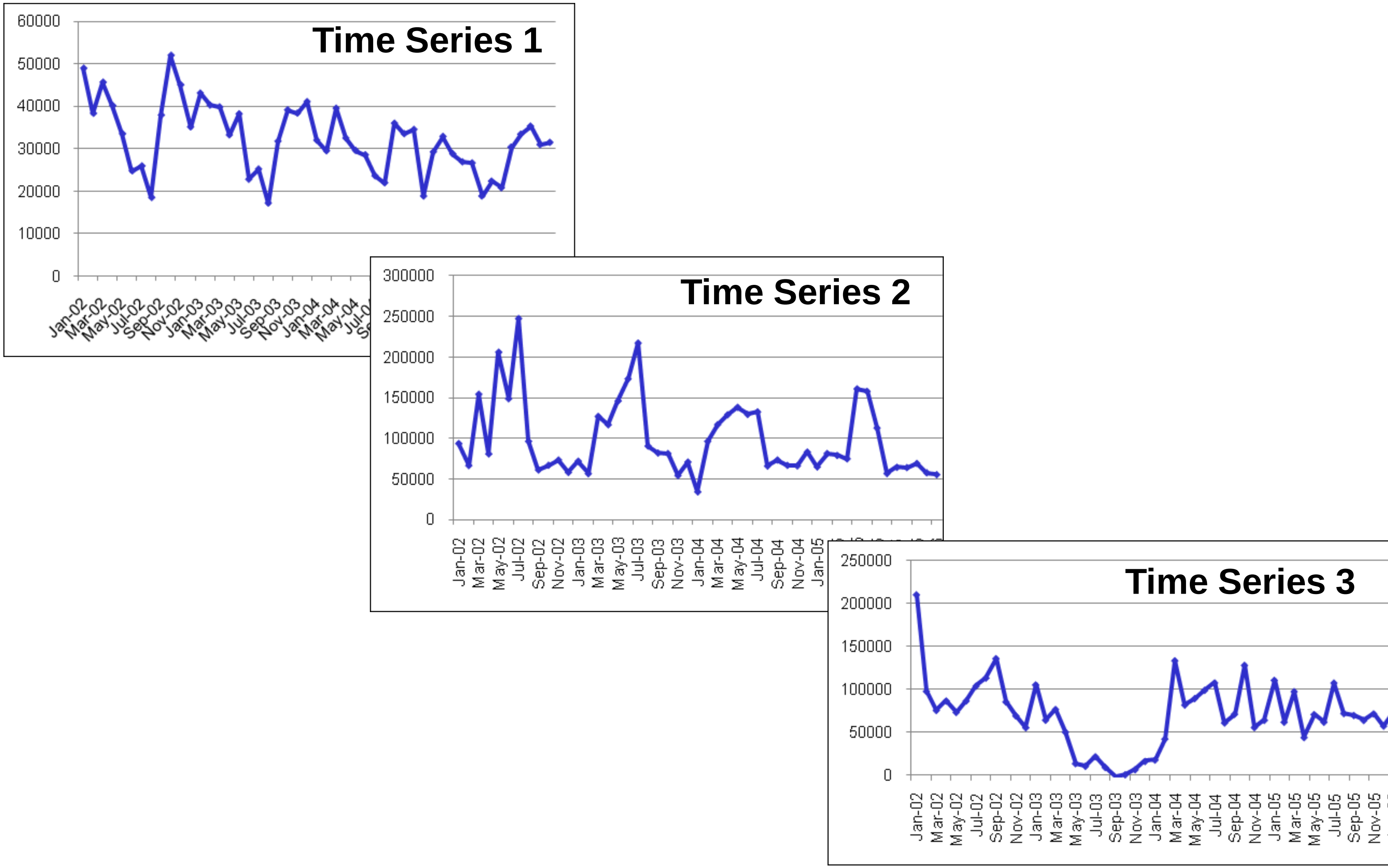
	Jan	Feb	Mrz	Apr	Mai	Jun	Jul	Aug	Sep	Okt	Nov	Dez
2002	93888	66840	154691	81204	206430	149074	247758	96707	61428	67249	73716	58384
2003	72192	56864	127260	117072	146412	173544	217626	90900	82224	81660	54564	71124
2004	34428	96636	117264	129348	138552	129755	132924	66264	73740	67080	66492	83376
2005	65112	81672	79236	75108	161184	158172	113064	57012	64632	64332	69516	57971

Time Series 3

	Jan	Feb	Mrz	Apr	Mai	Jun	Jul	Aug	Sep	Okt	Nov	Dez
2002	210228	97968	75798	86736	73428	86699	103980	113423	136080	85632	69384	55704
2003	105396	64428	76884	50532	13956	10860	21936	9504	-1452	996	6840	16476
2004	18096	42480	133536	82000	89616	99168	107760	61200	71340	128172	55896	64212
2005	110760	62064	97692	44244	71016	62052	107712	72228	69924	64080	71724	57564

Numbers or graphs?

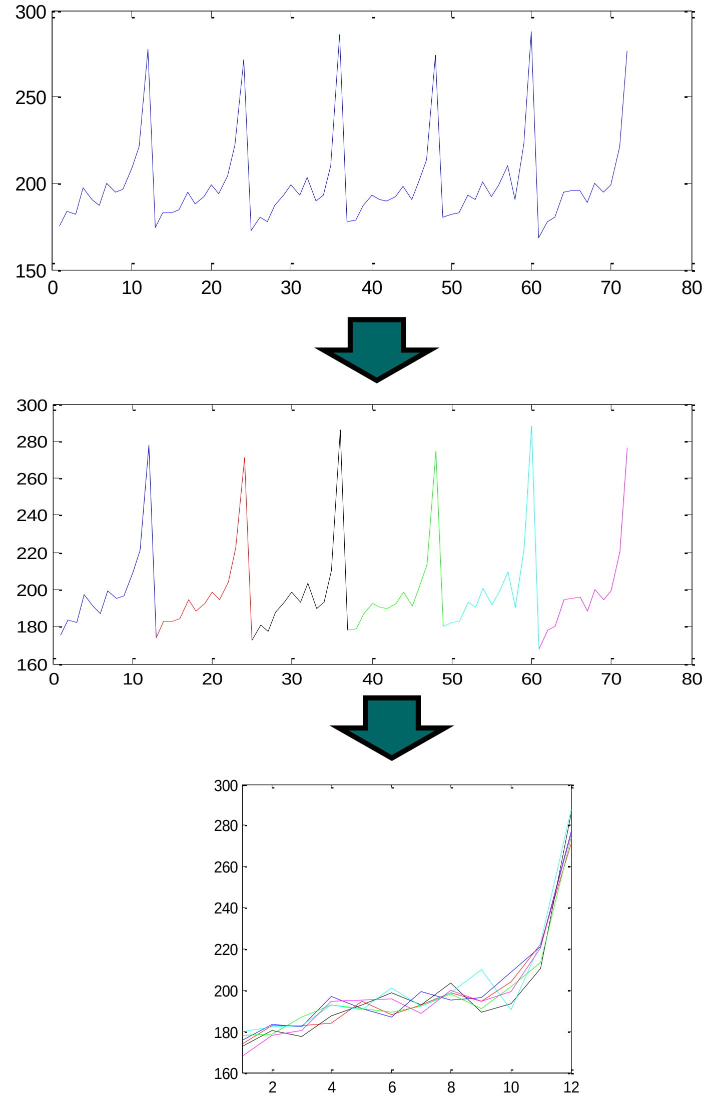
- **What is the pattern? (4 years of data)**



Data Exploration
Seasonal Diagrams

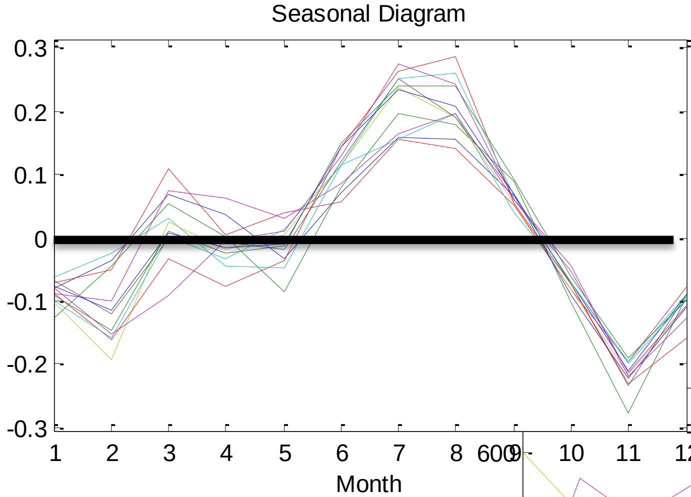
To identify seasonality we can use seasonal plots:

1. **Starting from a time series we identify different seasons (typically years).**
2. **If the time series is trended we need to de-trend it: subtract moving average.**
3. **We plot each season is plotted on top of each other.**
4. **If there is a clear repeating pattern, the time series is seasonal.**

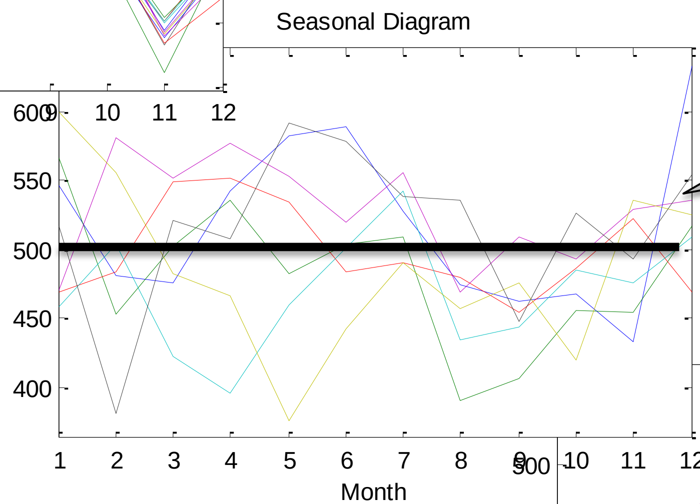


How many years are needed to identify seasonality?

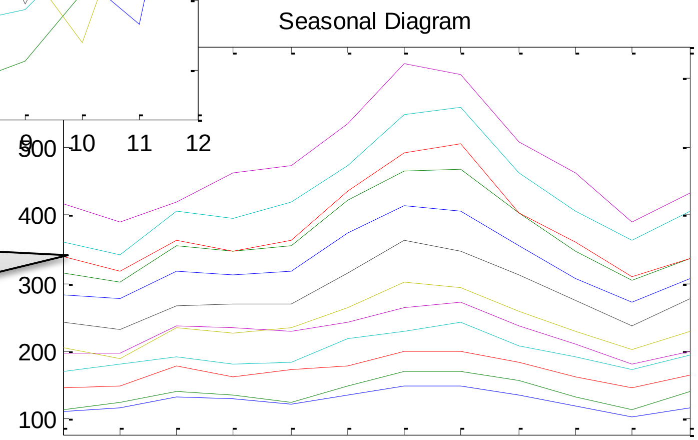
Data Exploration
Seasonal Diagrams



Imagine the mean on the seasonal diagram, do you see consistent deviations from it?

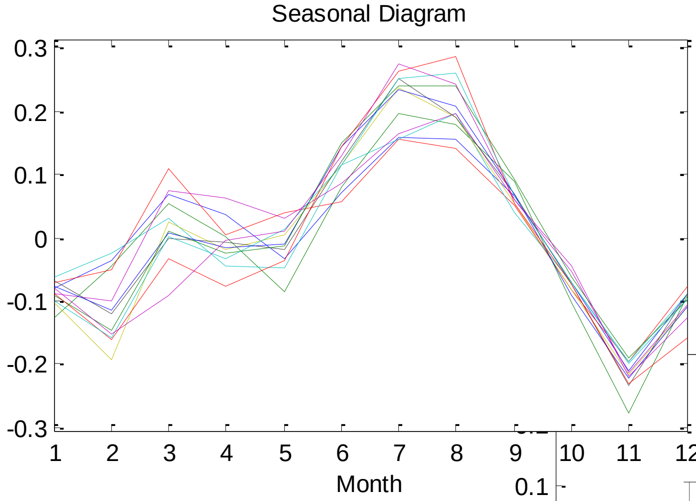


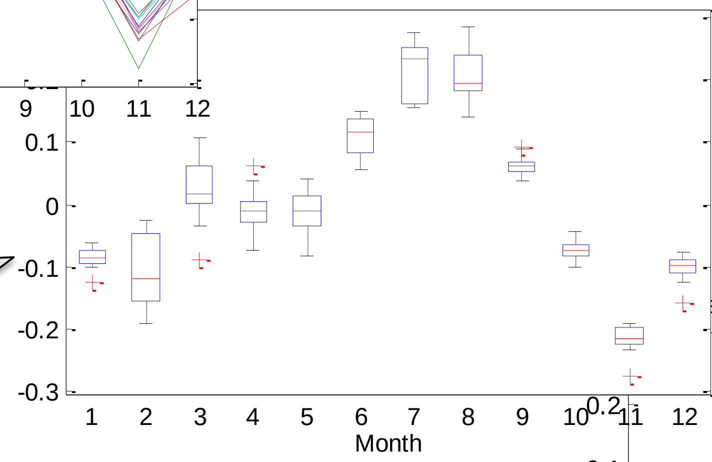
No clear pattern → No seasonality

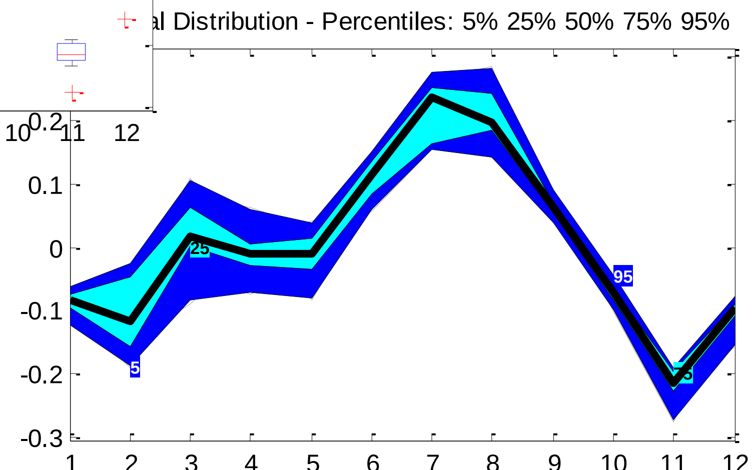


Can also hint the existence of trend

Forecasting Process
Statistical Modelling: Time Series Analysis





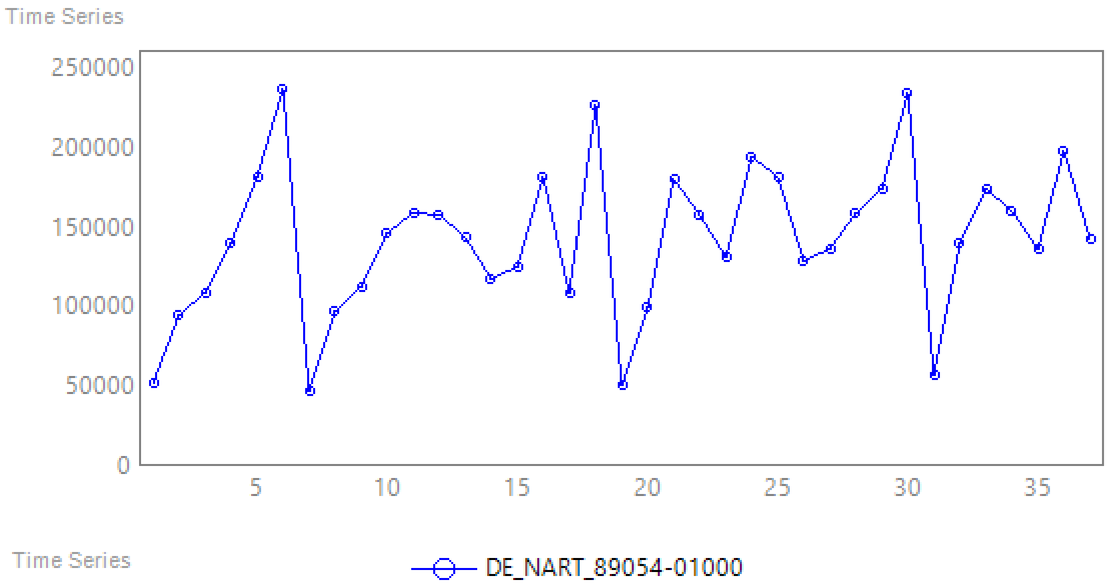


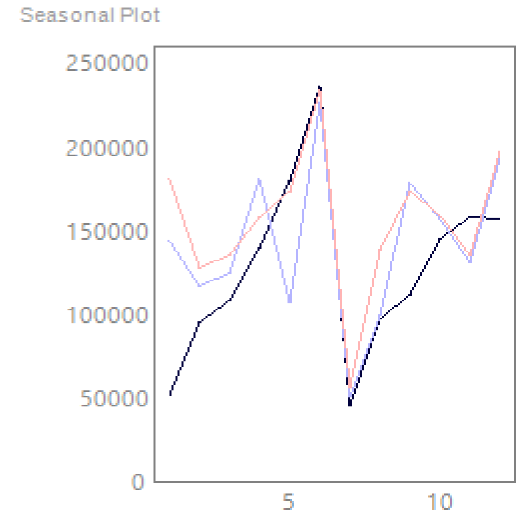
Essentially we check whether the means/median are different

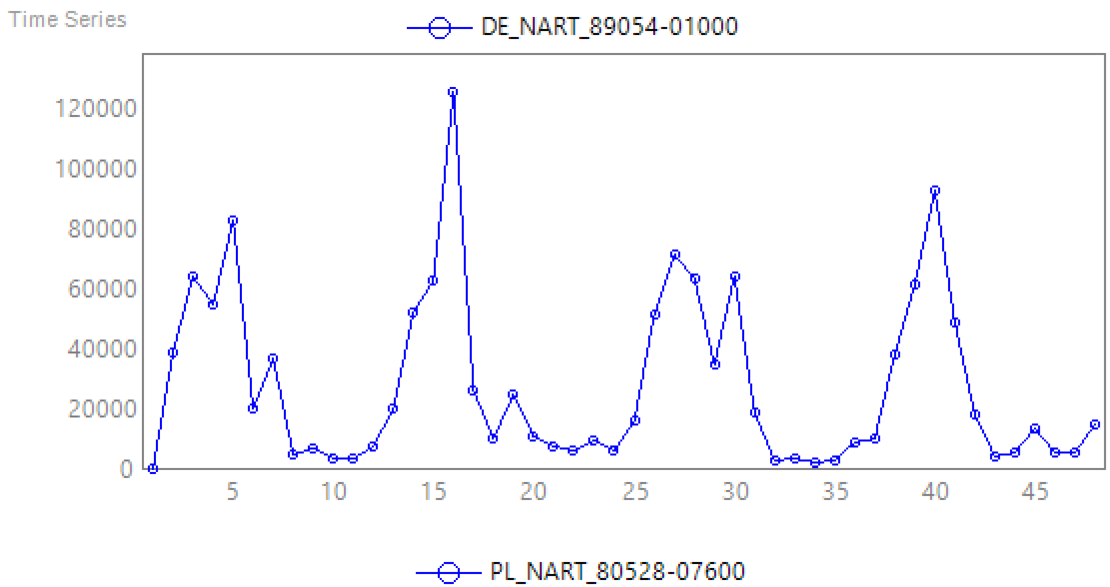
We can also visualise how much is the uncertainty at each month

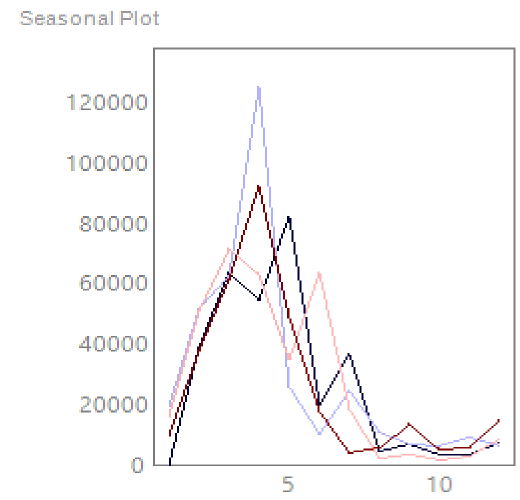
Forecasting Process

Statistical Modelling: Time Series Analysis





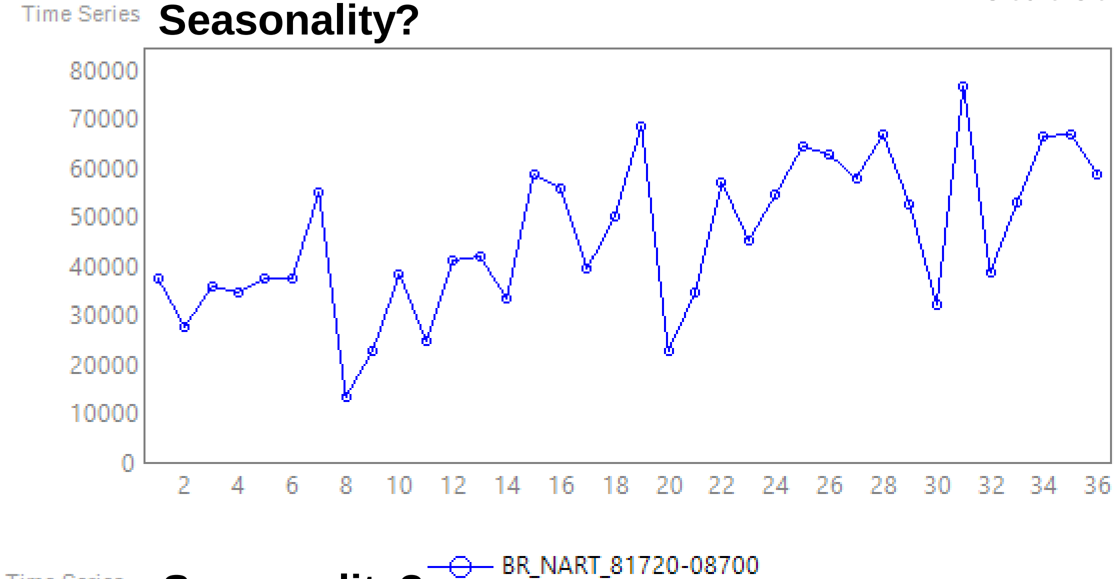


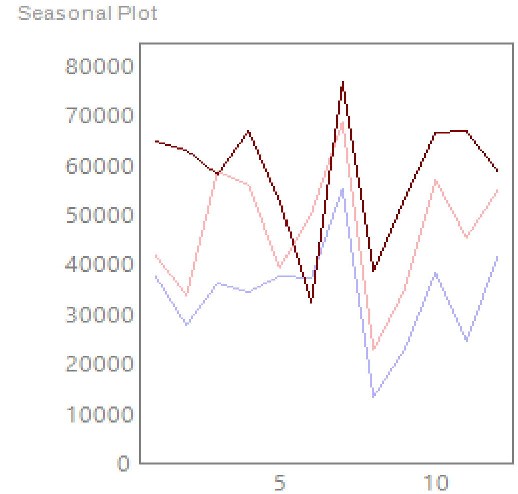


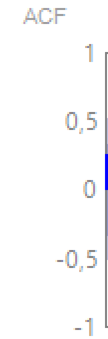
Forecasting Process

Statistical Modelling: Time Series Analysis

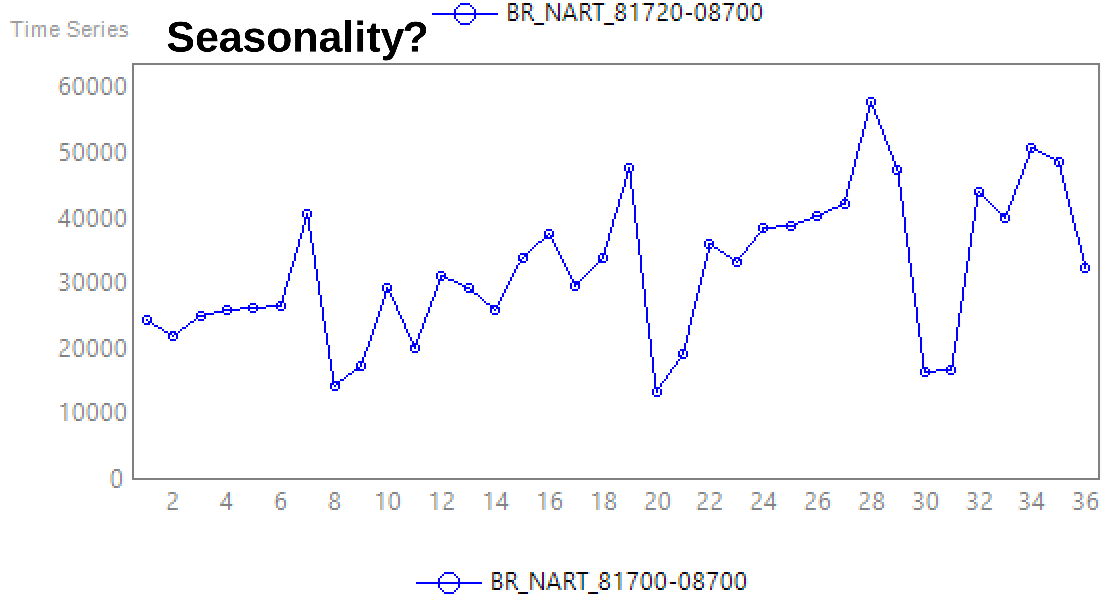
Seasonality?

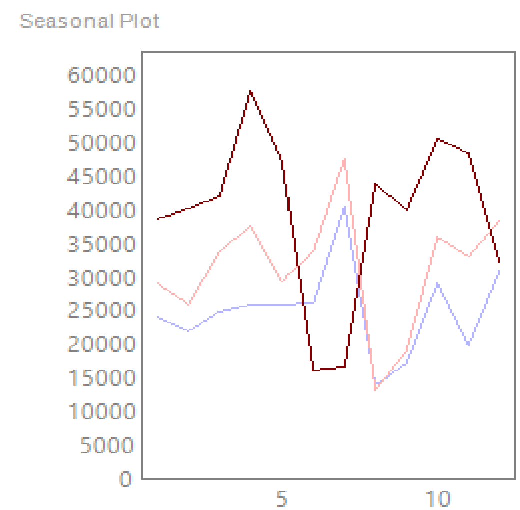


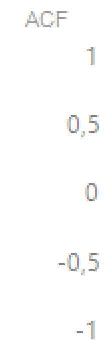




Seasonality?



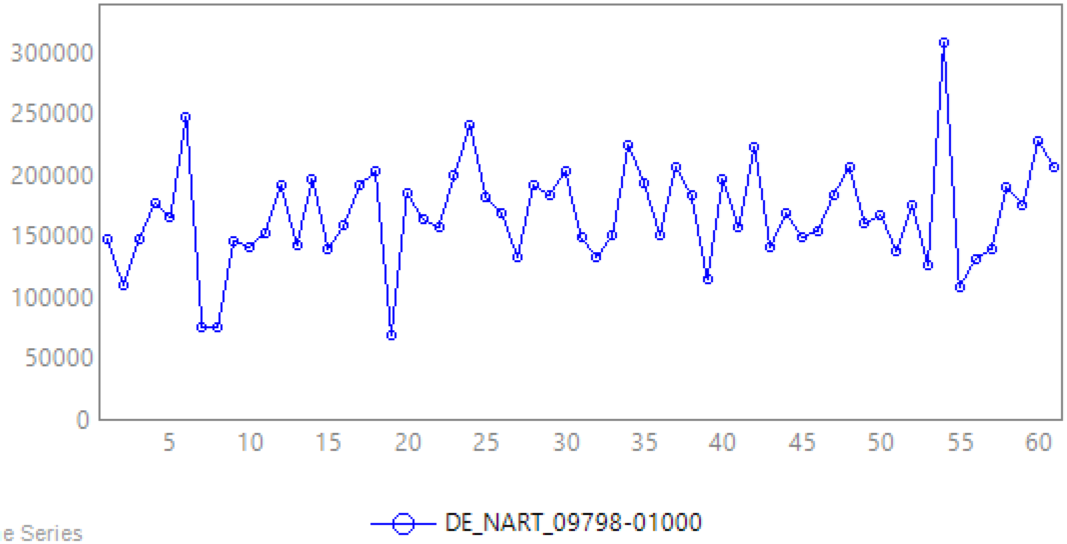




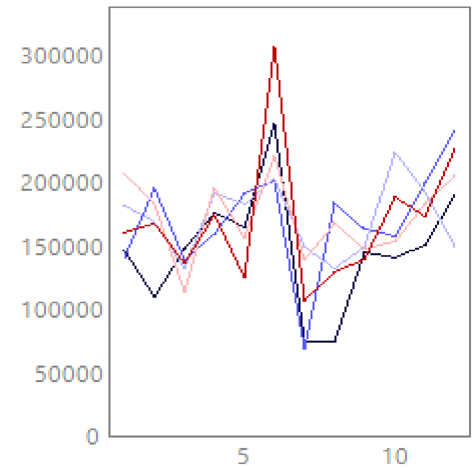
Forecasting Process

Statistical Modelling: Time Series Analysis

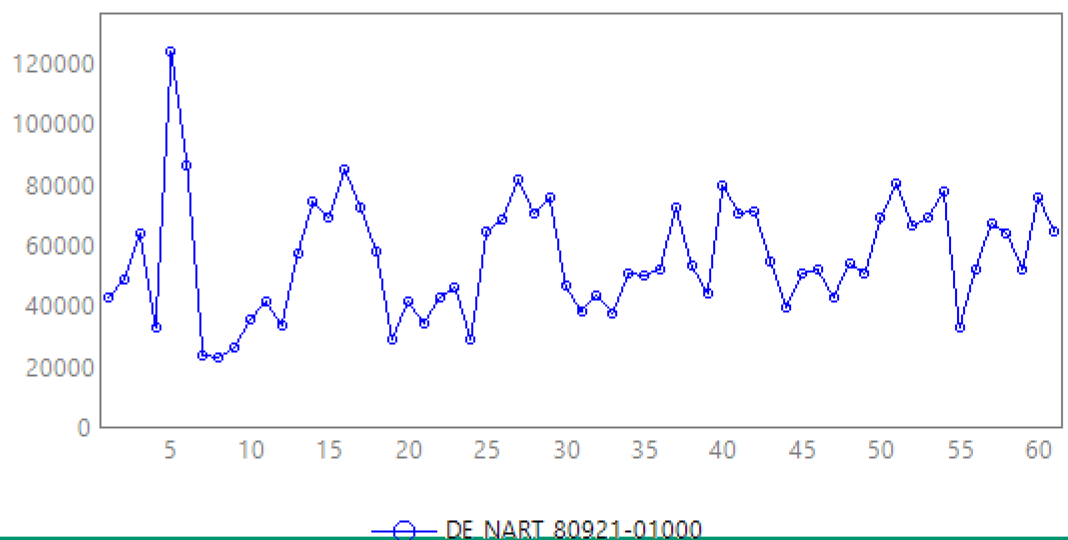
Time Series



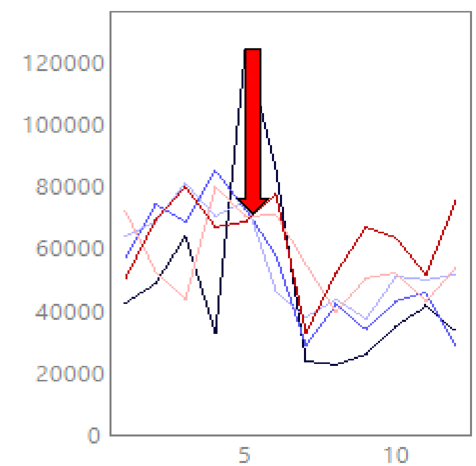
Seasonal Plot



Time Series



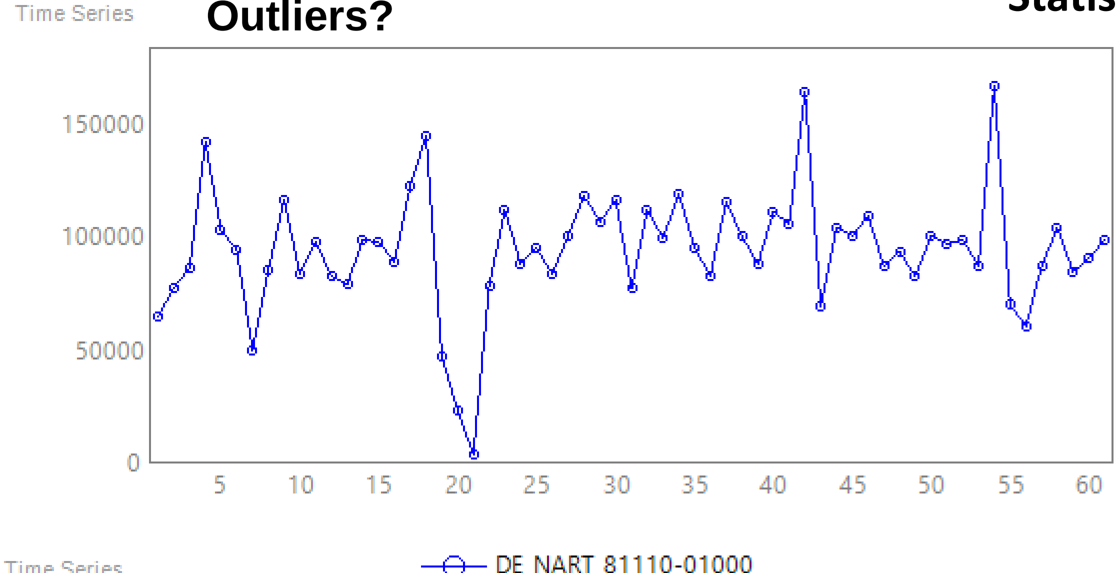
Seasonal Plot

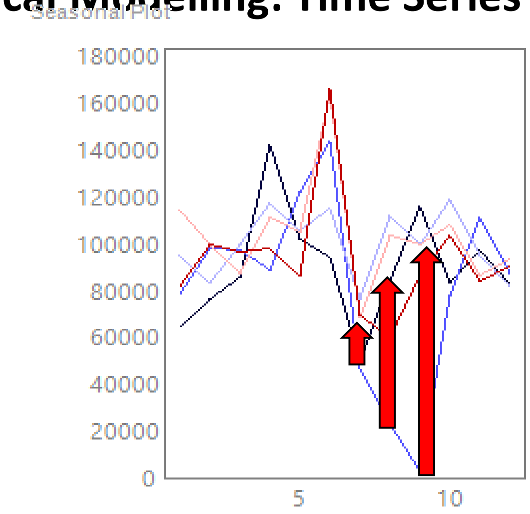


Forecasting Process

Statistical Modelling: Time Series Analysis

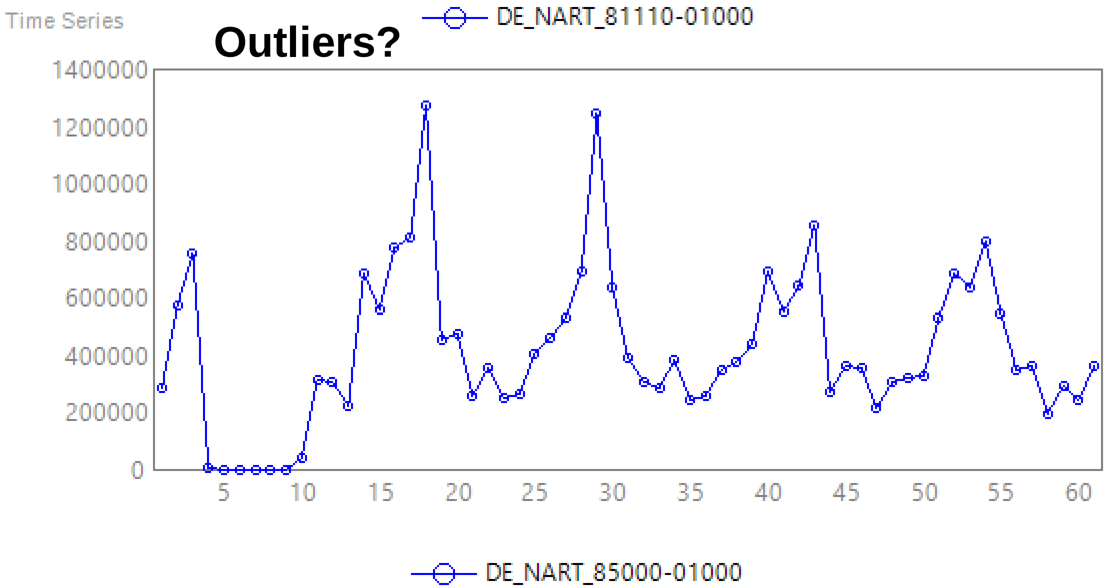
Outliers?

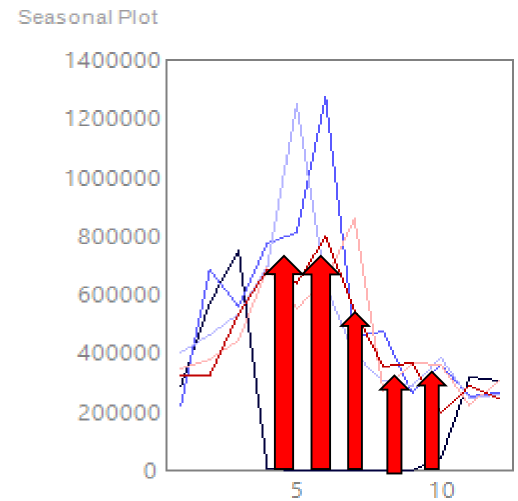




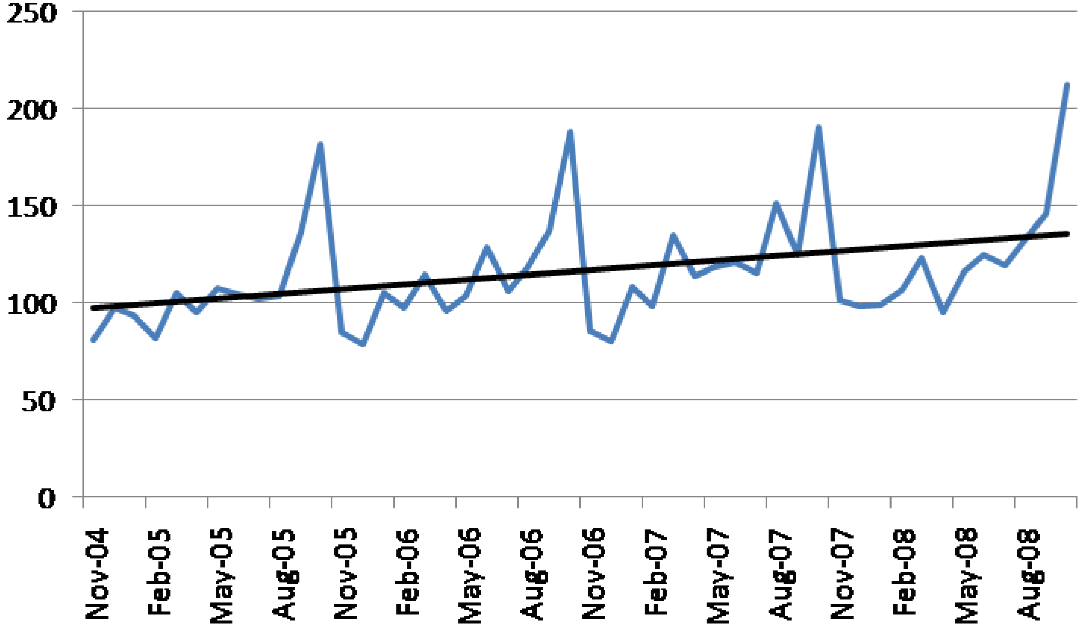
Find „true“ values!

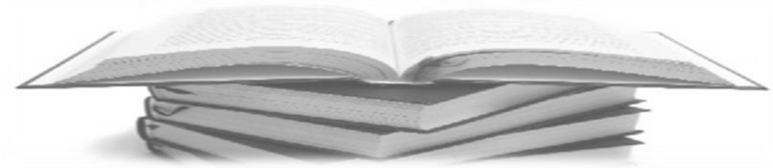
Outliers?





Trend Lines using Regression



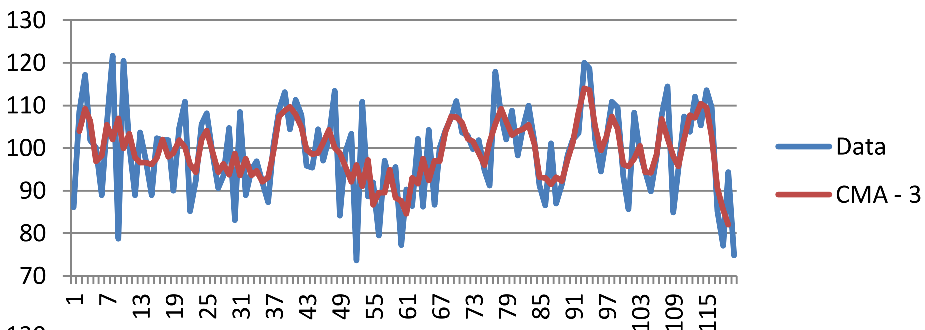


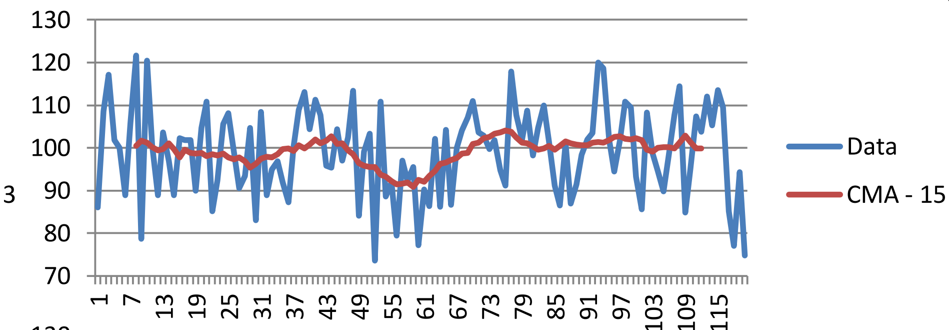
Data Exploration

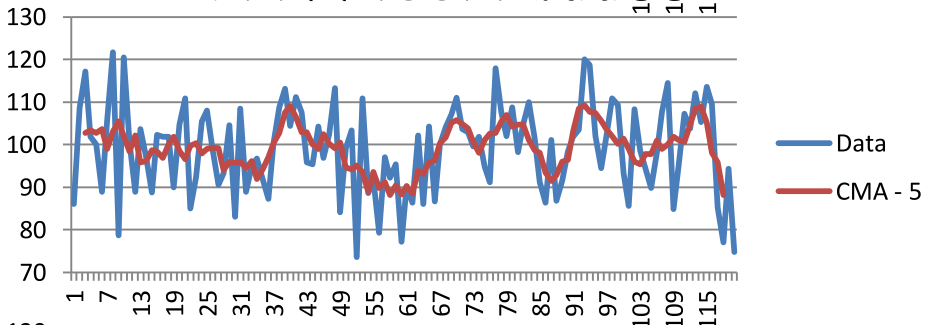
Level & Trend Diagrams

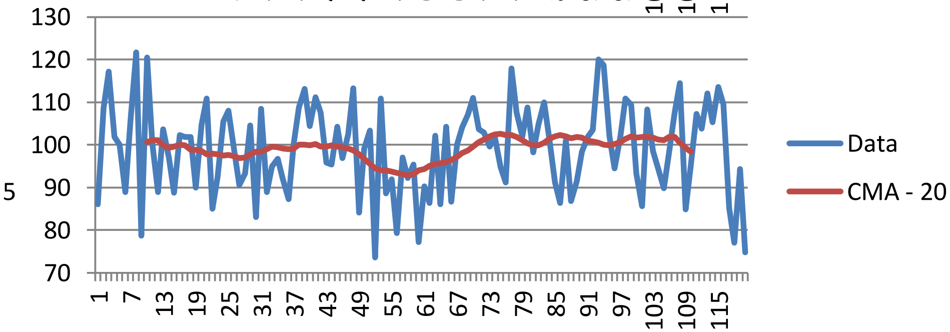
A simple tool to identify trend in the time series is the centre moving average (CMA). CMA is characterised by its order, i.e. how long is the average.

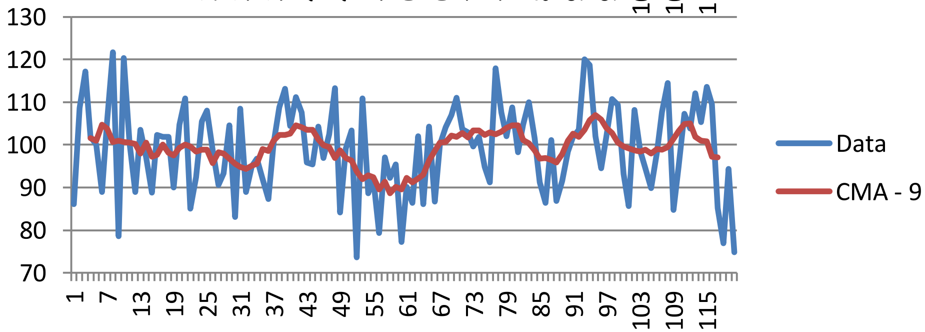
- **It can filter the noise → Longer CMA filters more**
- **It can filter seasonality when the CMA order = seasonal length**

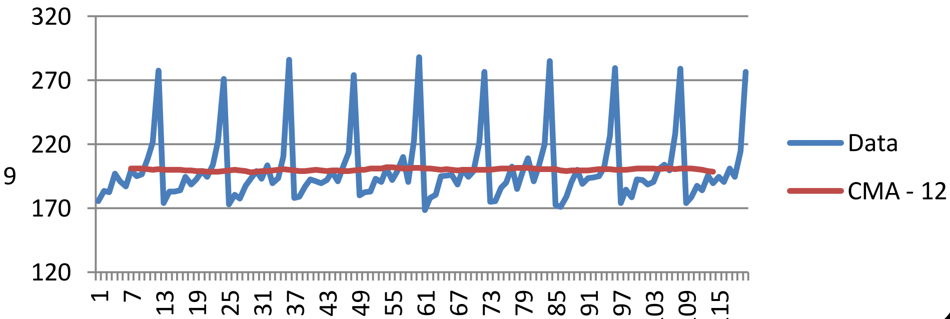












Data Exploration

Level & Trend Diagrams

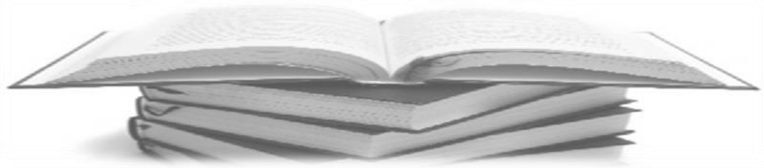
Period t	Time Series	CMA – 3			Time Series	CMA – 4	
1	86			½	86		
2	109	104.0	Middle value	1	109		
3	117	109.3		1	117	105.3 ?	Middle value
4	102	106.3		1	102	104.5	
5	100	97.0		½	100	100.6	
6	89	98.3			89	101.8	
7	106	105.7			106	101.6	
8	122	102.3			122	102.9	
9	79	107.0			79		
10	120				120		

$$\text{CMA} - 3 = \frac{X_1 + X_2 + X_3}{3}$$

$$\text{CMA} - 4 = \frac{X_1/2 + X_2 + X_3 + X_4 + X_5/2}{4}$$

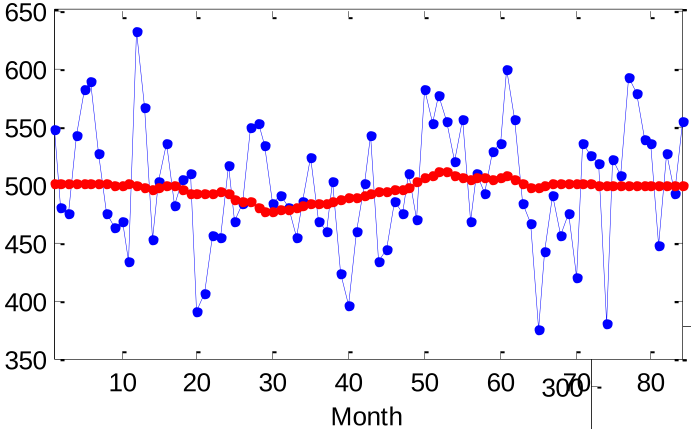
½ + 1 + 1 + 1 + ½ = 4!

It is similar to a moving average, but it is placed in the centre, instead of the end of the series of observations used to form it. Cannot be used to forecast!

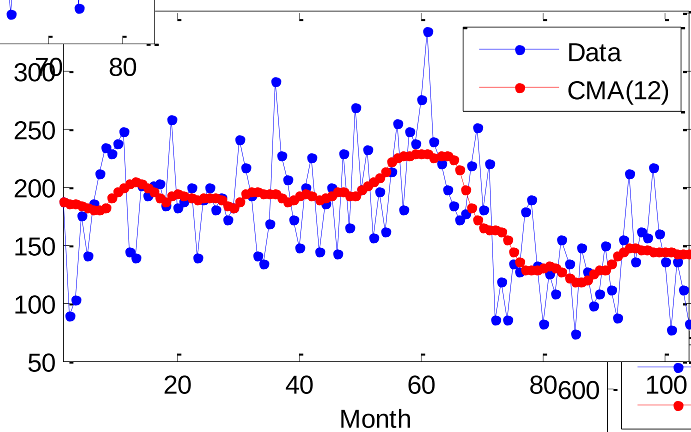


Forecasting Process

Statistical Modelling: Time Series Analysis

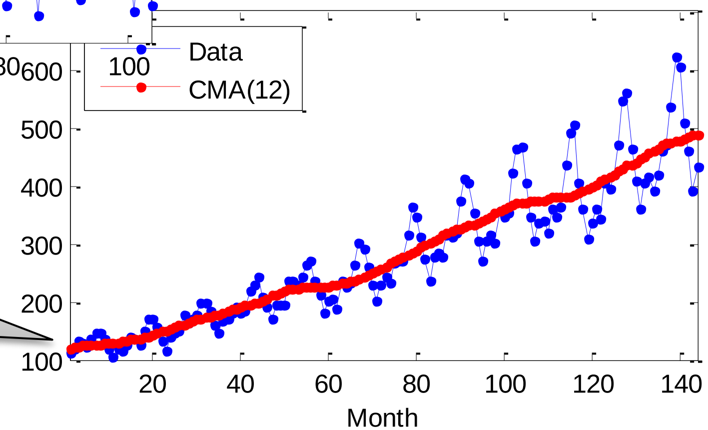


Relatively flat CMA → No trend



Abrupt change in the level → Level shift

Plots of CMA allow a quick visual identification of trended time series



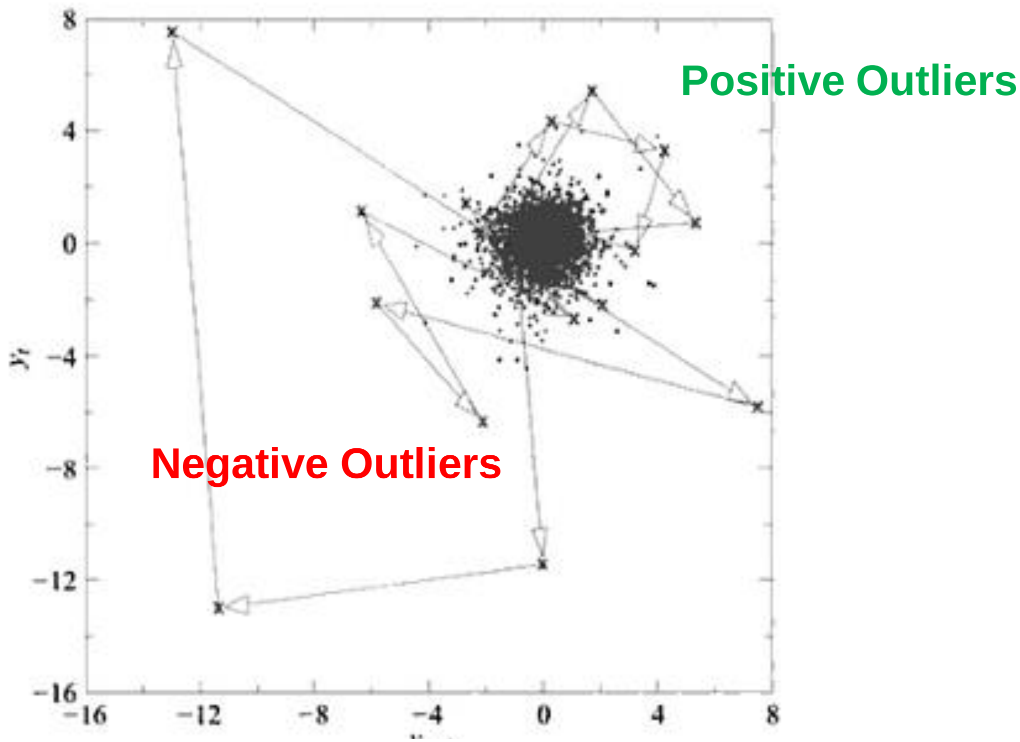
Long term movement of CMA → Trend

Statistical tests can be employed to help automating the process, such as: OLS linear or robust regression tests, Cox-Stuart test, ...

Lancaster Centre for Forecasting
www.forecasting-centre.com

Outlier Plots

Identifying innovation Outliers



Agenda

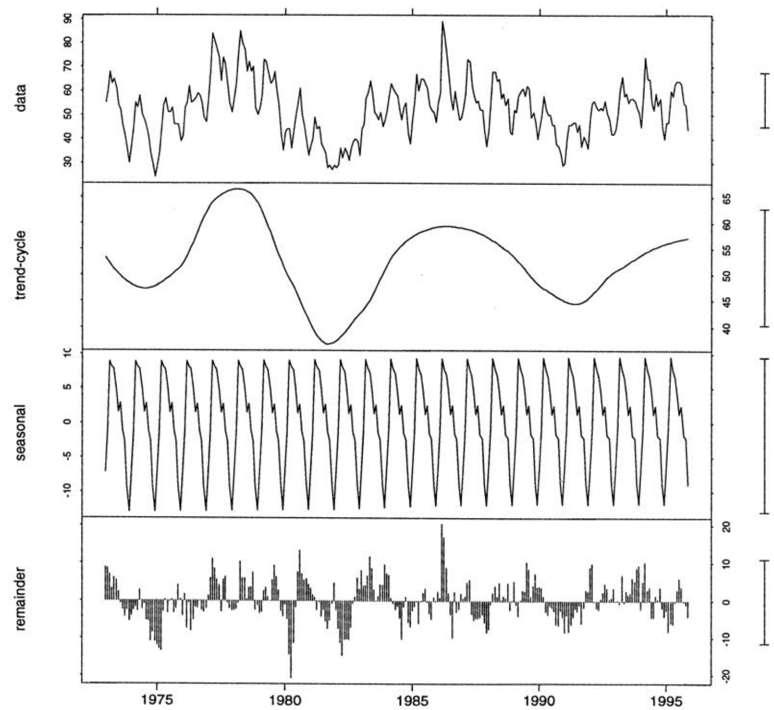
Time Series Analysis

1. Data Exploration
    1. Time Series Patterns
        1. Level, Trend & Season
        2. Outliers & Structural breaks
    2. Graphical Identification of Time Series Patterns
        1. Time Series & Seasonal Plots
        2. Time Series Decomposition
        3. A related „discipline“: Technical Analysis
    3. Statistical Identification of Time Series Patterns
        1. Autocorrelation Analysis (Corelograms)
        2. Spectral Analysis
        3. Statistical Tests
2. Data Transformations
    1. Transforming level of series
    2. Transforming variance of time series

Regular Components of Time Series

A Time Series consists of superimposed components / patterns:

- **Signal**
  - **level ‘L‘**
  - **trend ‘T‘**
  - **seasonality ‘S‘**
- **Noise**
  - **irregular,error 'e'**

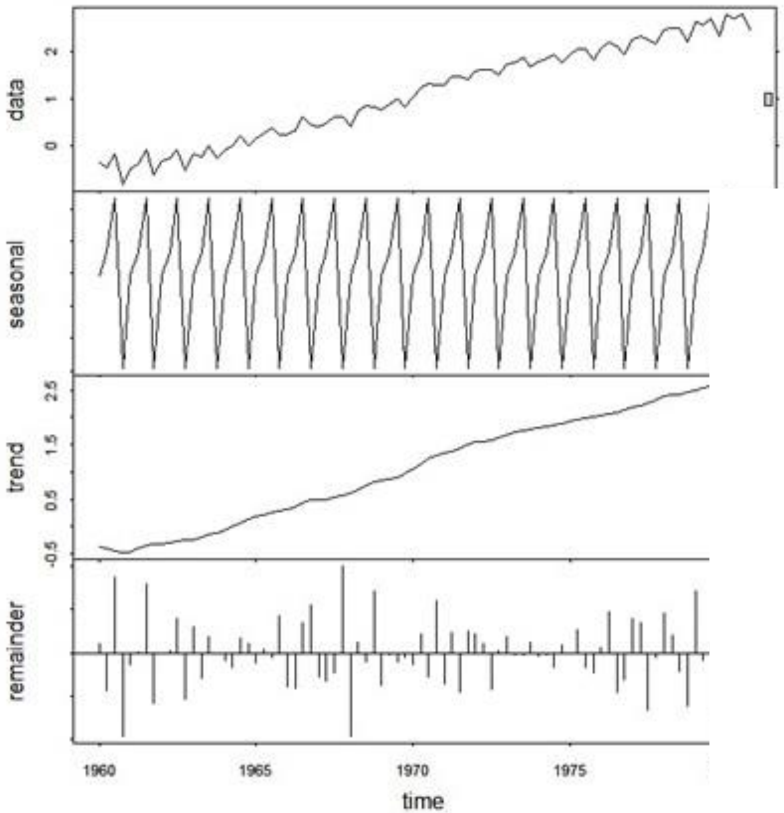


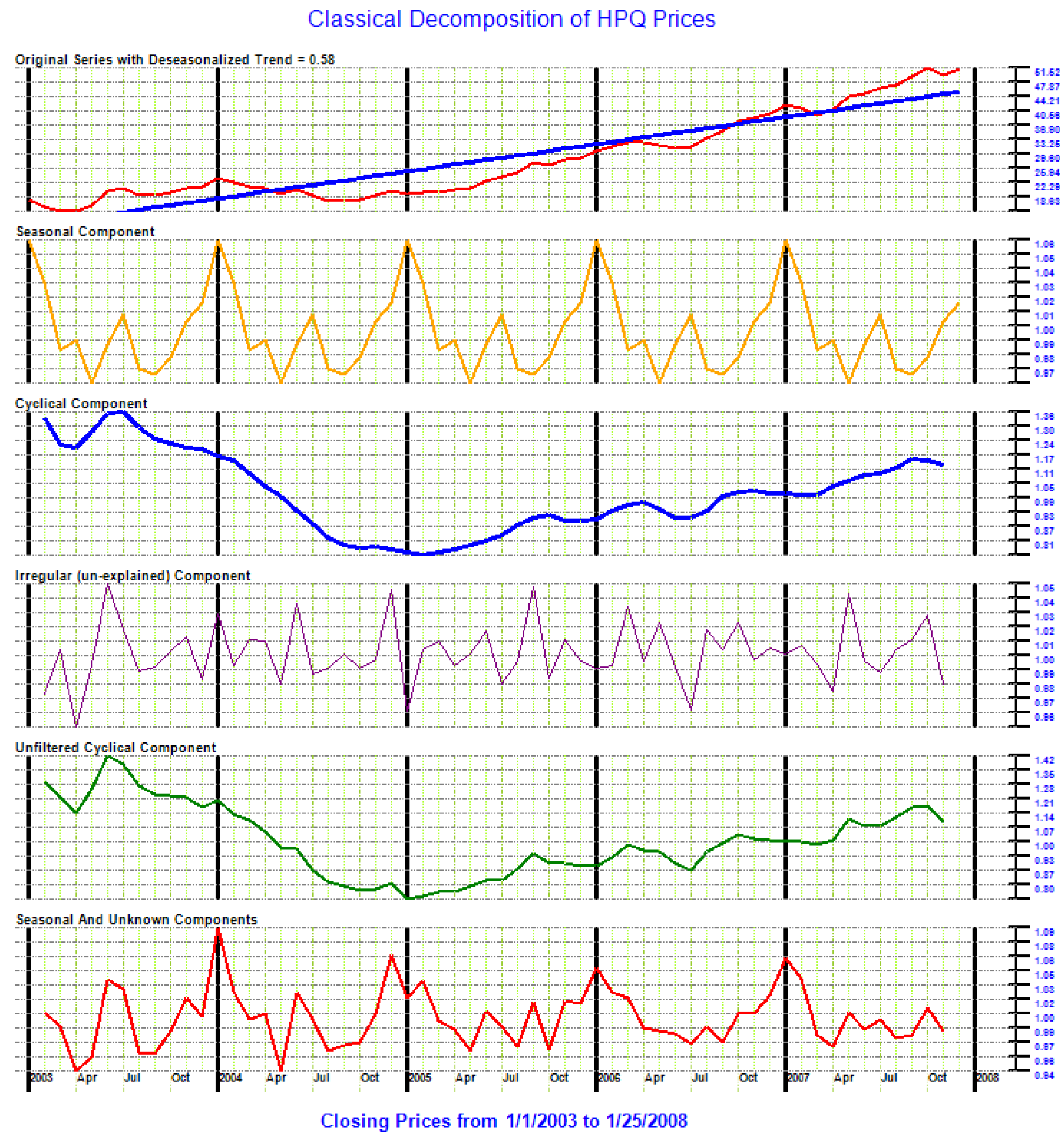
Y = L + S + T + E

Sales = LEVEL + SEASONALITY + TREND + RANDOM ERROR

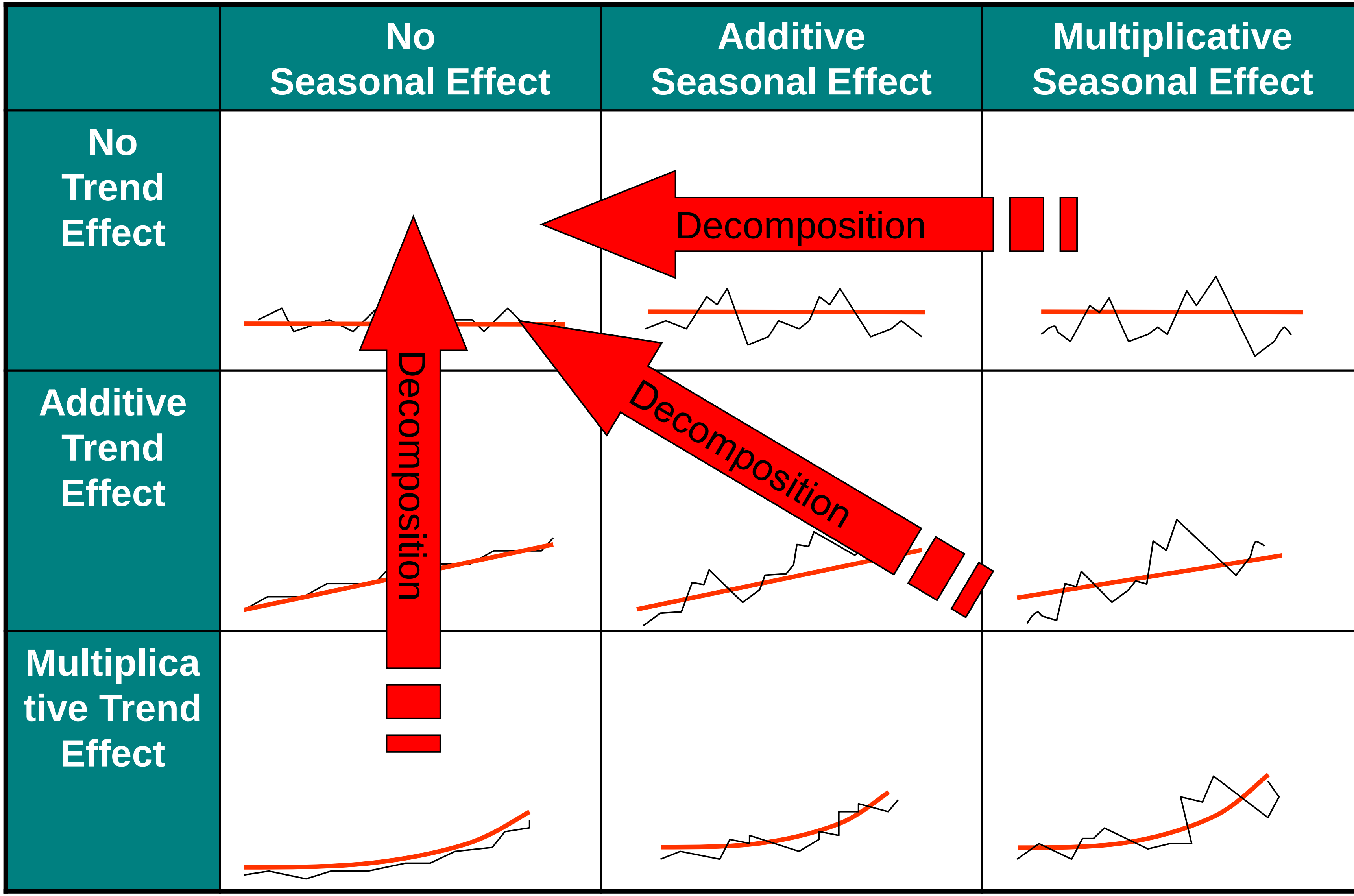
Y = L * S * T * E

Examples of Time Series Decomposition





Decomposition?



Time Series Decomposition

Objective: Isolate individual components to understand structure

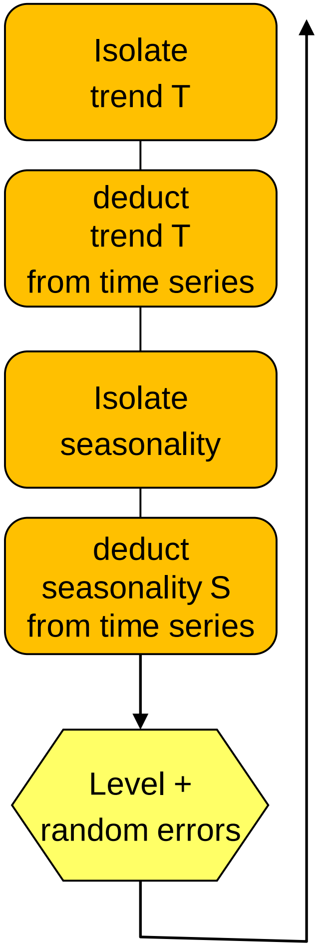
Y= S+T+E |search Y? --> Y=S+T+E|-T

Approach

1. **Disaggregate individual components through successive calculation**
   1. **calculate trend through12-month moving average (or Regression)**
   2. **deduct trended time series from original series -T**
   3. **calculate seasonal indices as mean of all months (Sum divided by number of values each period)**
   4. **deduct seasonal time series from detrended time series –S**
2. **Analysis of 3 time series components in individual graphs**
3. **continuation / extrapolation of base value (optional)**
4. **aggregation of previously disaggregate components (optional)**

WARNING

- **Decomposition is NO TOOL for valid forecasting**
- **Useful for UNDERSTANDING & separate VISUALIZATION of time series**
  - **development of individual components over time → stability**
  - **interaction of components → pattern identification**

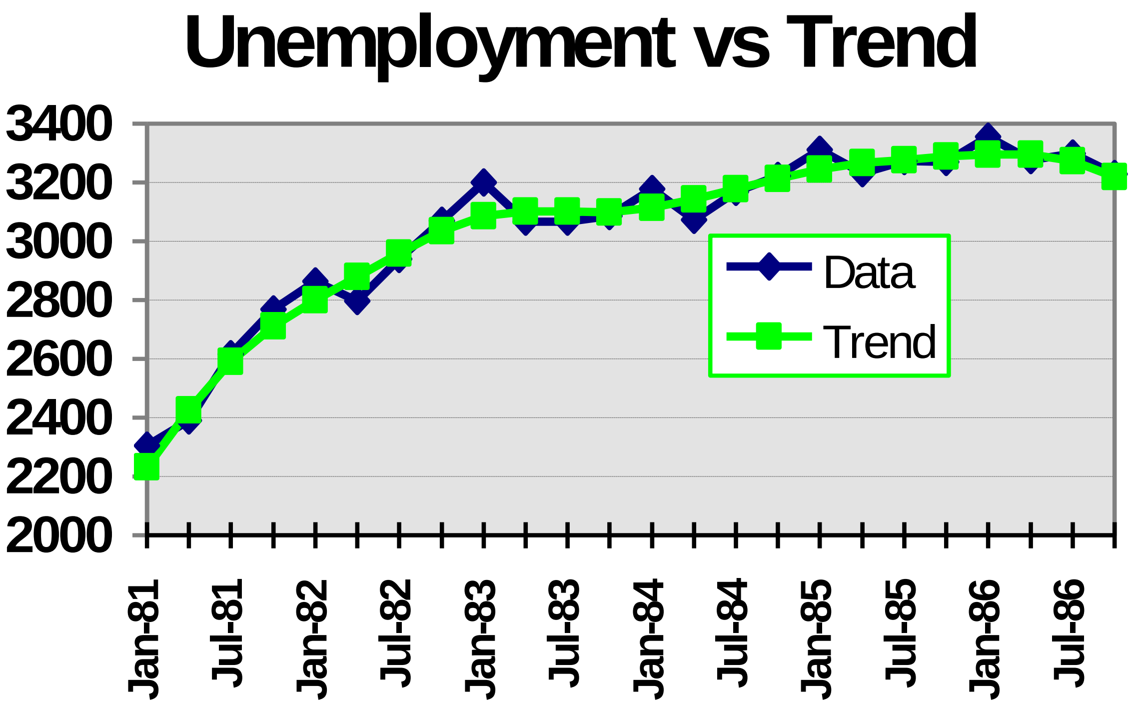


Decomposition of Time Series Patterns

- Calculating a moving average (for quarterly data)
  - average of two consecutive 4 period averages

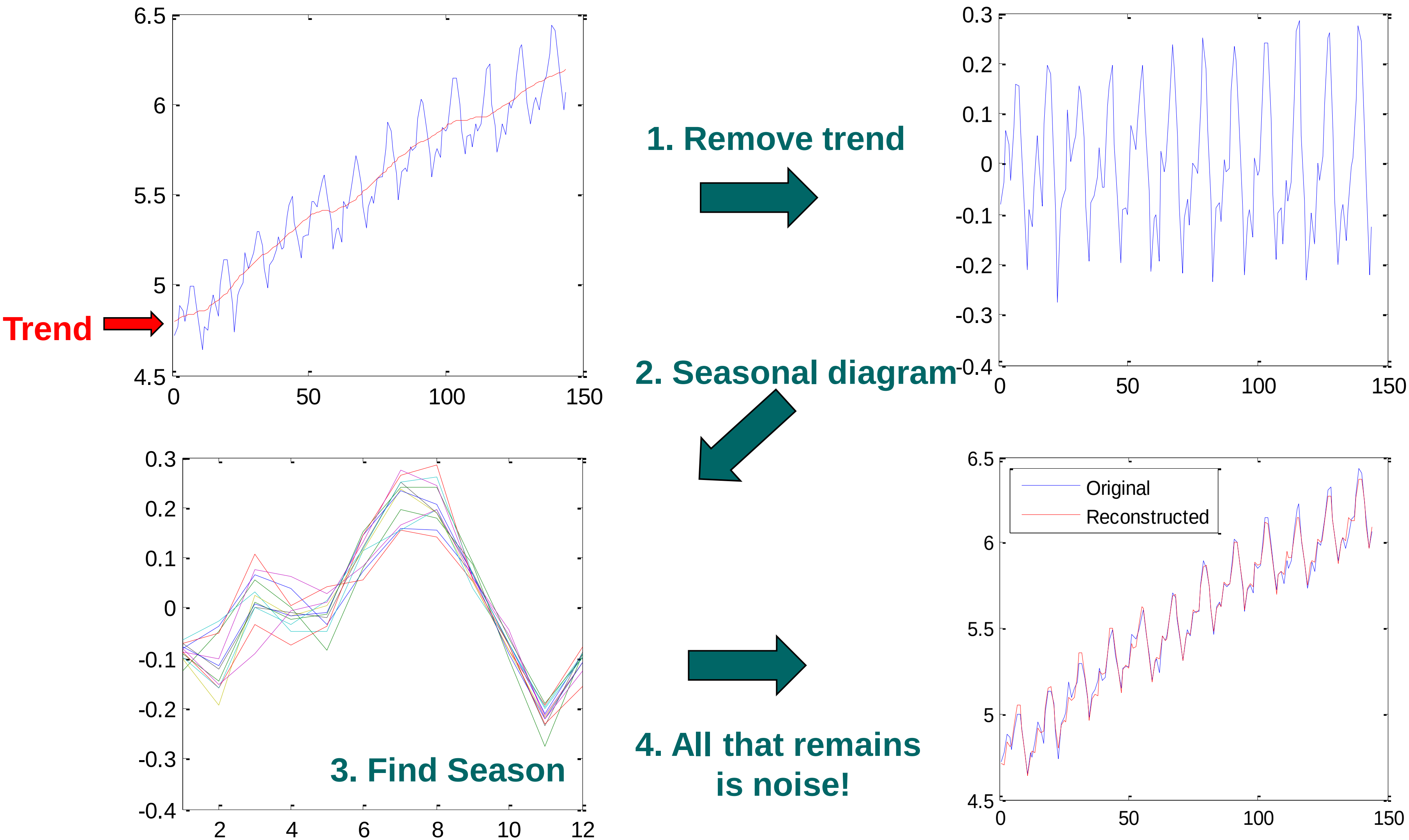
$$MA_t = \frac{1}{8}(Y_{t+2} + 2(Y_{t+1} + Y_t + Y_{t-1}) + Y_{t-2})$$

	Data	Moving Average
Jan-81	2306	2232
Apr-81	2392	2426
Jul-81	2616	2590
Oct-81	2768	2710
Jan-82	2862	2801
Apr-82	2796	2879
Jul-82	2939	2959
Oct-82	3070	3035
Jan-83	3199	3085
Apr-83	3068	3103
Jul-83	3066	3102
Oct-83	3086	3100
Jan-84	3176	3113
Apr-84	3074	3143



Decomposition - An example

Airline Passengers (logarithm)



Agenda

Time Series Analysis

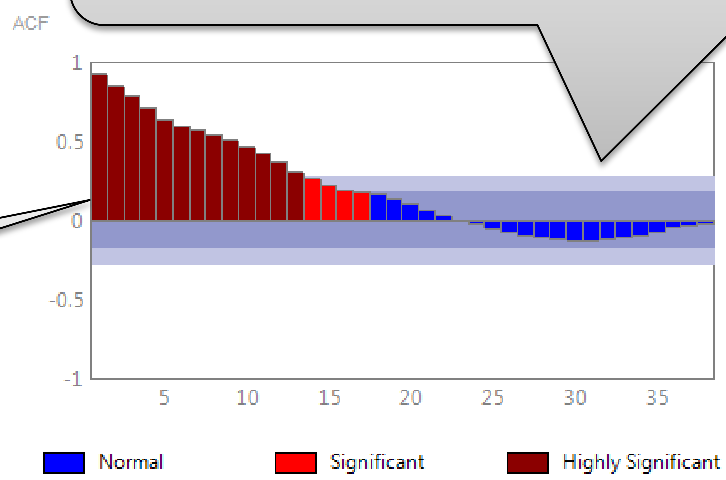
1. Data Exploration
    1. Time Series Patterns
        1. Level, Trend & Season
        2. Outliers & Structural breaks
    2. Graphical Identification of Time Series Patterns
        1. Time Series & Seasonal Plots
        2. Time Series Decomposition
        3. A related „discipline“: Technical Analysis
    3. Statistical Identification of Time Series Patterns
        1. Autocorrelation Analysis (Corelograms)
        2. Spectral Analysis
        3. Statistical Tests
            1. Tests for Stationarity
                1. Durbin Watson Test
                2. Augmented Dickey Fuller Test
            2. Tests for Seasonality
            3. Tests for Trend
2. Data Transformations
    1. Transforming level of series
    2. Transforming variance of time series

Model Selection
ACF/PACF

The Autocorrelation Function (ACF) is a time series statistic that shows how closely correlated is A_t with lagged A_{t-k}, and as such it is bounded between [-1, 1]. A correlation close to 1 is a strong positive correlation, i.e. A_t and A_{t-k} move in the same direction. A correlation of -1 is a strong negative correlation, i.e. A_t and A_{t-k} move in the opposite direction, while a correlation close to zero means that the two lags are unrelated.

$$\hat{\rho}_k = \frac{\sum_{k+1}^{n}(A_t - \bar{A})(A_{t-k} - \bar{A})}{\sum_{1}^{n}(A_t - \bar{A})^2}$$

The confidence intervals the z-score from the normal distribution divided by the sample size: z/n



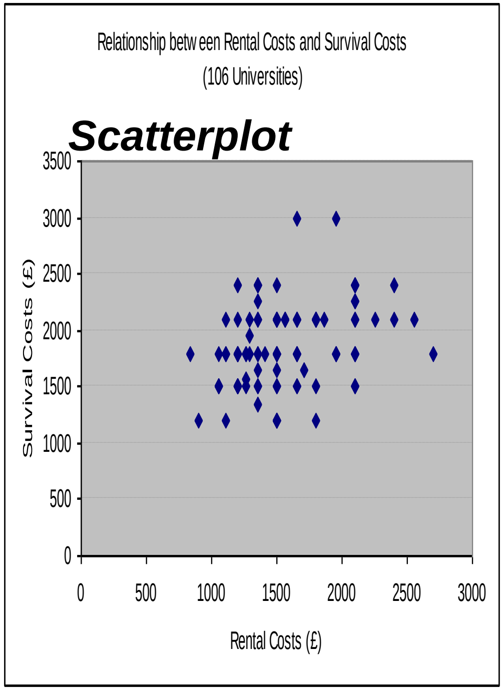
The strength of each autocorrelation is denoted by a bar between [-1, 1]

RECAP: ***Correlation*** **= Measure of Association**

Definition:

measures the <u>strength</u> and <u>direction</u> of a <u>linear</u> relationship between two numerical variables

→Analysing bivariate relationships



e.g.

- **relationship between advertising expenditure and sales?**
- **relationship between price and sales?**
- **relationship between attendance and exam performance?**
- **between university fees and student applications?**
- **traffic density and accident rates?**
- **debt and insurance crime?**
- **physical attributes and risk of birth problems?**
- **share activity in London and New York?**

Q: is there a pattern in the data?

Measuring the strength of a relationship – correlation coefficient

-1 perfect negative relationship

0 no relationship

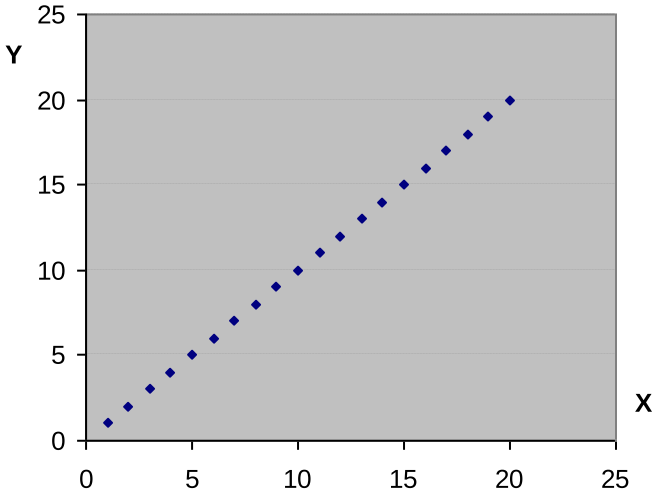
1 perfect positive relationship

correlation coefficient, *r*

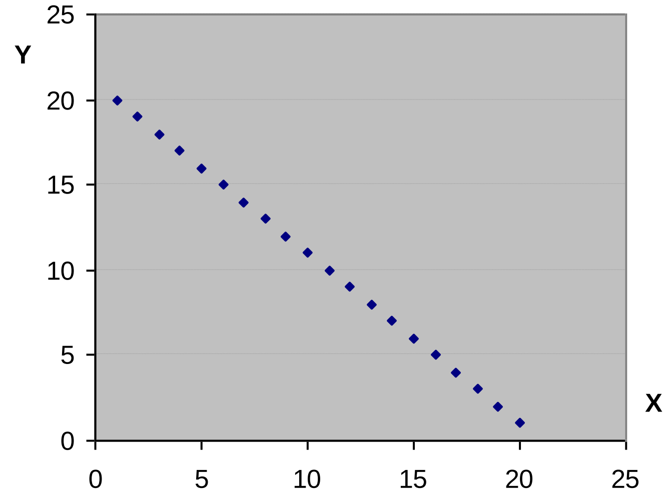
$$= \frac{\sum xy - \left(\sum x\right)\left(\sum y\right)/n}{\sqrt{\sum x^2 - \left(\sum x\right)^2/n}\sqrt{\sum y^2 - \left(\sum y\right)^2/n}}$$

the bigger the absolute value of r, the stronger the relationship

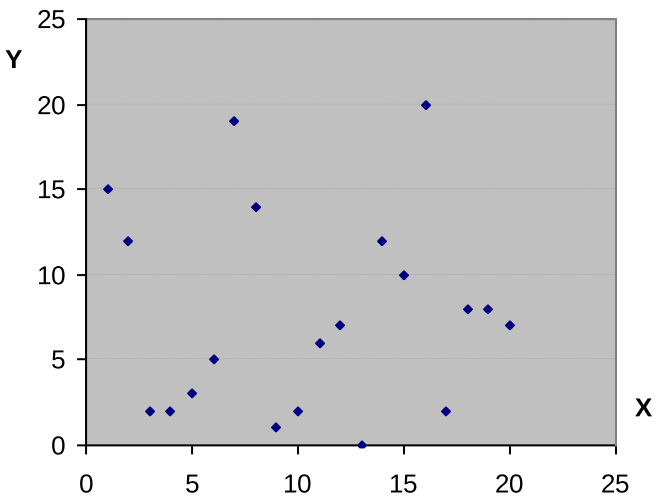
the value of *r* doesn't tell the whole story though……..



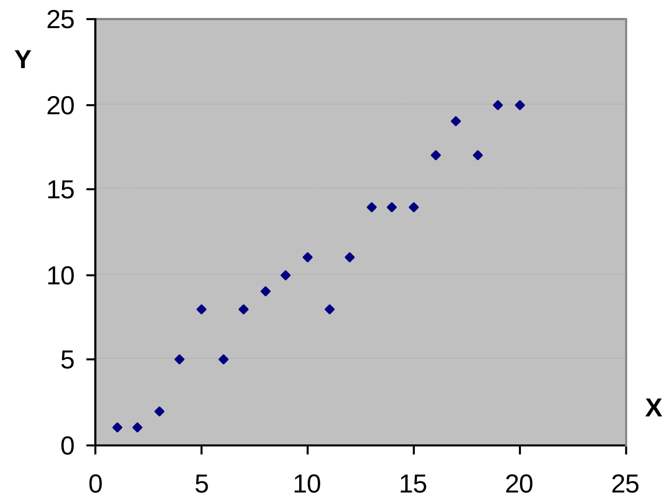
perfect positive relationship, r = 1



perfect negative relationship, r = -1



no relationship, r = 0.004

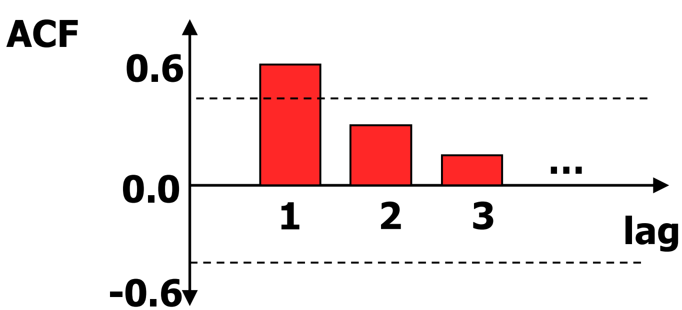


strong positive relationship, r = 0.975

- **E.g. time series Y_t**

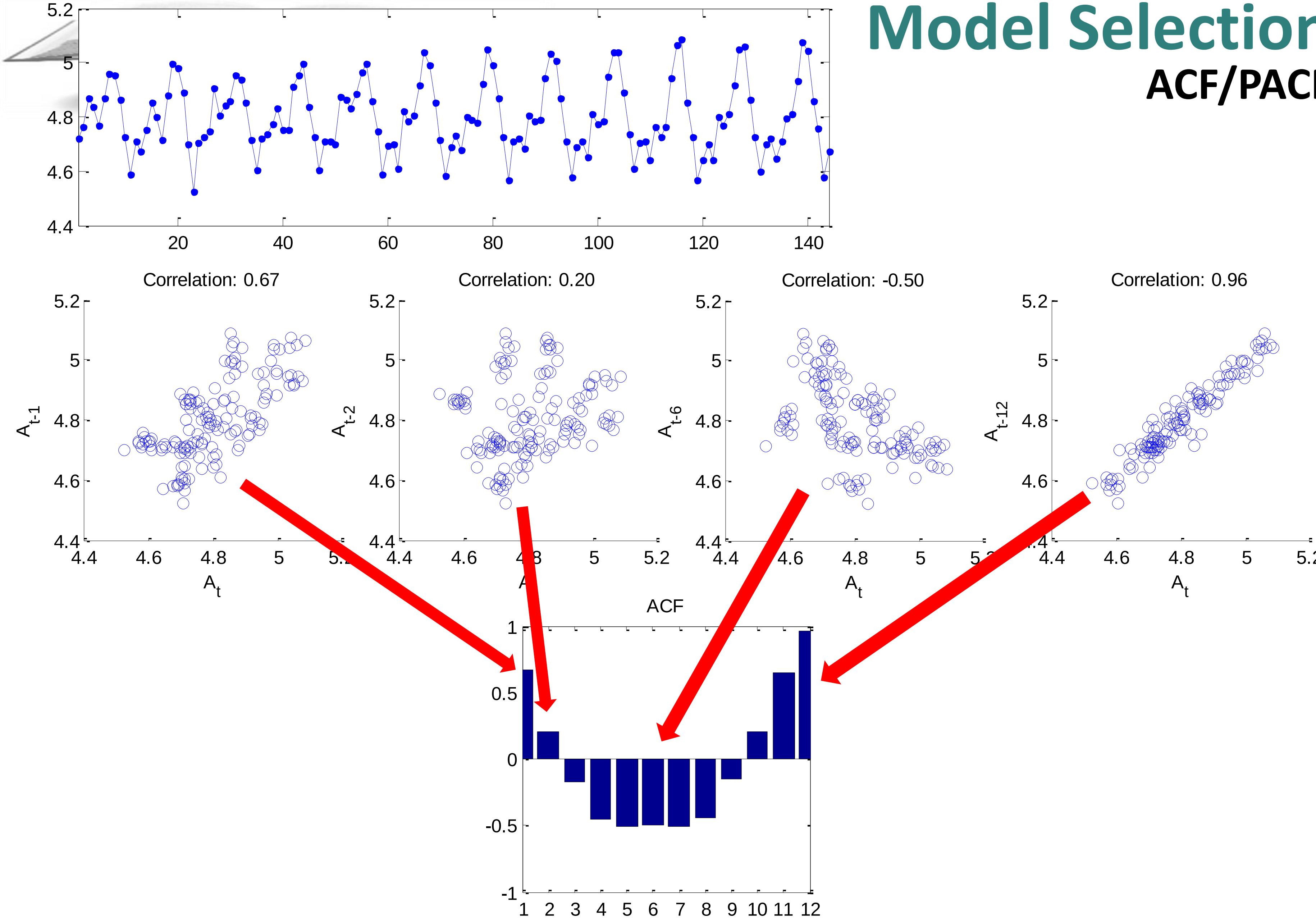
7, 8, 7, 6, 5, 4, 5, 6, 4.

lag 1:	lag 2:	lag 3:
7, 8	7, 7	7, 6
8, 7	8, 6	8, 5
7, 6	7, 5	7, 4
6, 5	6, 4	6, 5
5, 4	5, 5	5, 6
4, 5	4, 6	4, 5
5, 6	5, 4	
6, 4		
r_1=.62	r_2=.32	r_3=.15



→ **Autocorrelations r_t gathered at lags 1, 2, ... make up the autocorrelation function (ACF)**

Model Selection
ACF/PACF



Model Selection

ACF/PACF

The ACF contains in the correlation between A_t and A_{t-k} all the previous lags $A_{t-k+i,\ i<k}$. The Partial Autocorrelation Function (PACF) provides the strength of relationship between lags clean of any effects from intermediate lags. The derivation of the autocorrelation function comes from the iterative fitting of an dynamic regression model.

The model is fitted up to lag k+1 and all its coefficients are kept fixed. Then the coefficient of lag k is estimated, which is also the partial autocorrelation of that lag.

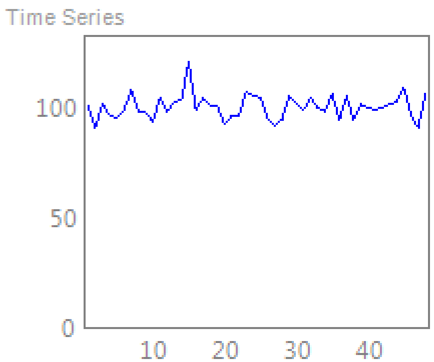
The PACF is bounded between [-1, 1] and can be interpreted in a similar fashion to ACF.

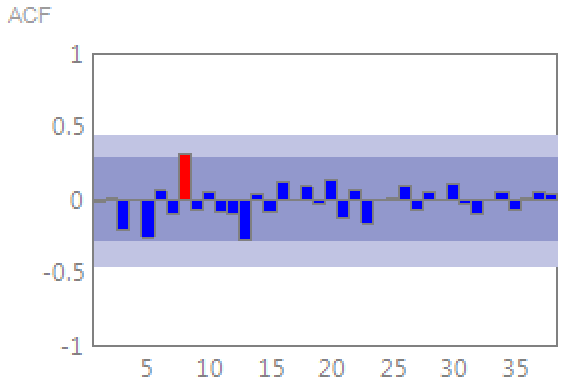
Both ACF/PACF can aid in data exploration. Depending on the components of each time series the ACF/PACF exhibit particular patterns, that allows us to classify the time series.

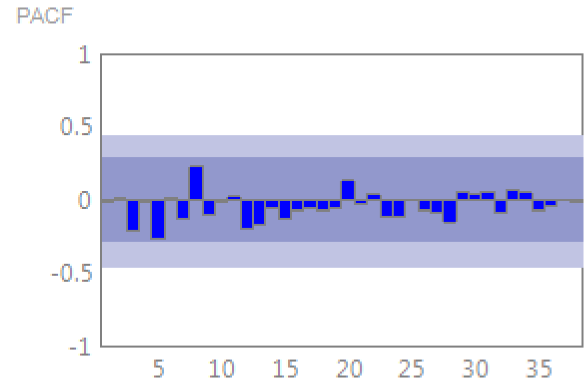
Model Selection

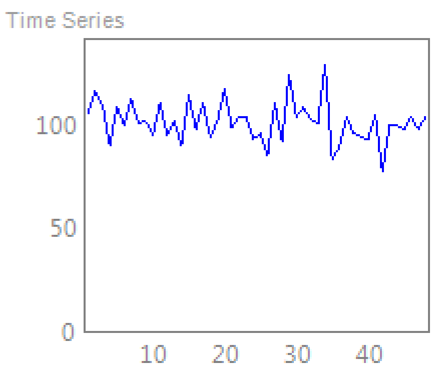
ACF/PACF stationary time series

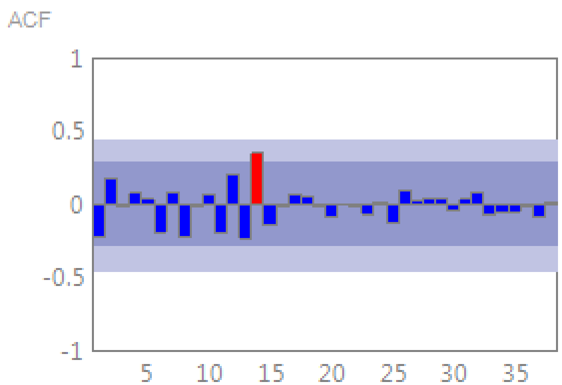
ACF | **PACF**

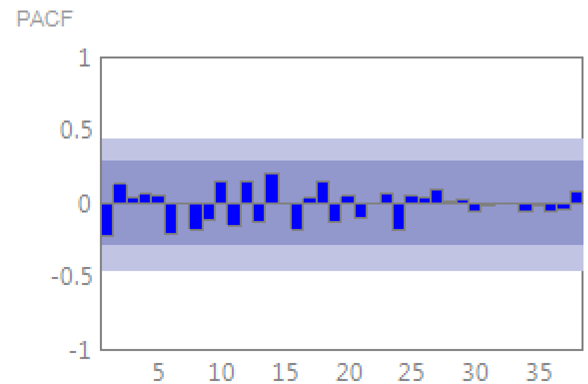


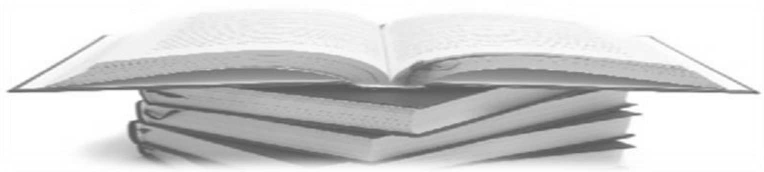








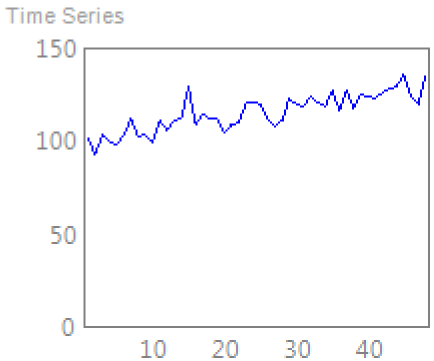


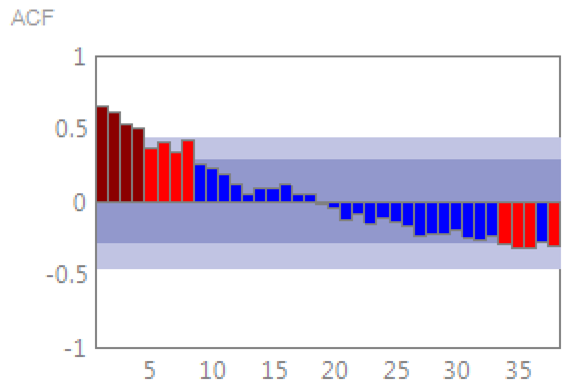


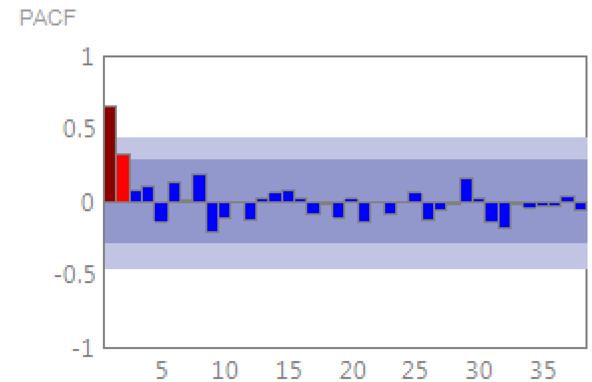
Model Selection

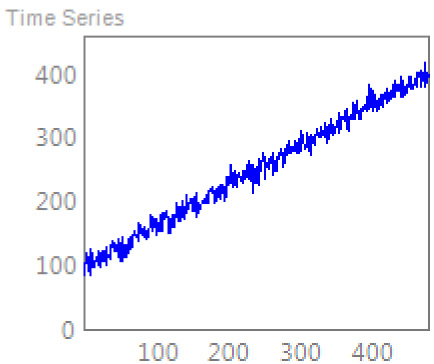
ACF/PACF trended time series

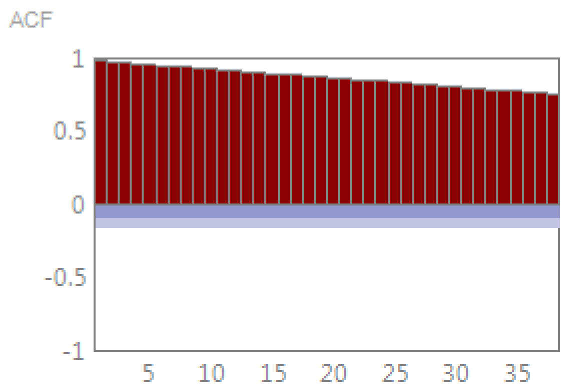
ACF | **PACF**

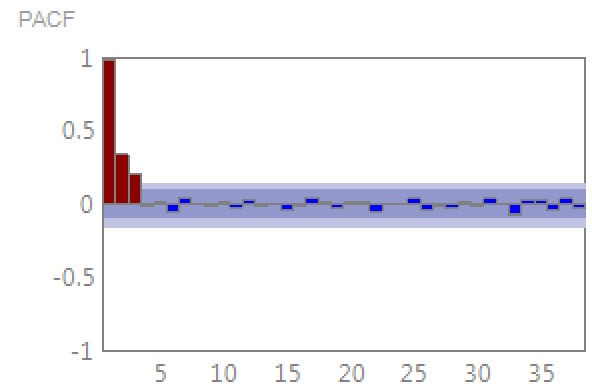






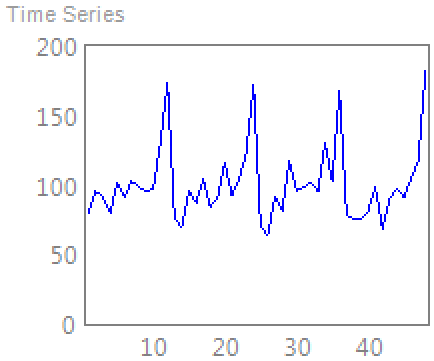


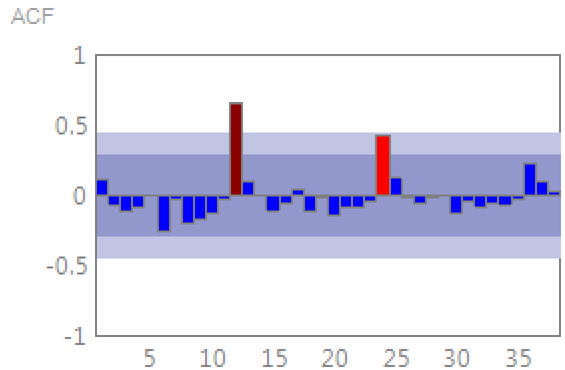


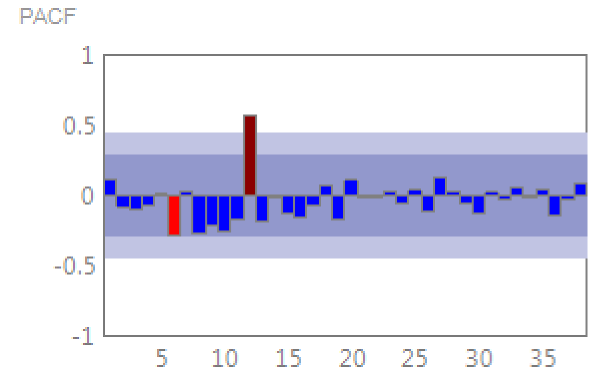


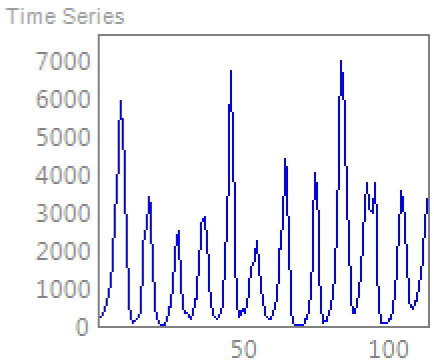
Model Selection
ACF/PACF seasonal time series

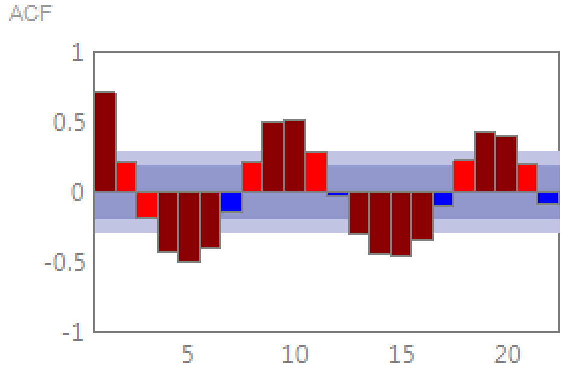
	ACF	PACF

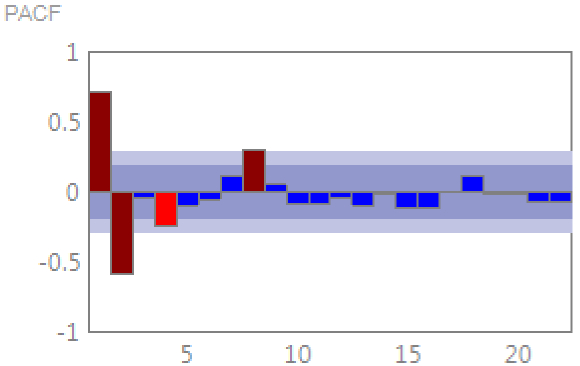












Partial Autocorrelations

- **Partical Autocorrelations are used to measure the degree of association between Y_t and Y_{t-k} when the effects of other time lags $1,2,3,\ldots,k-1$ are removed**
  - **Significant AC between Y_t and Y_{t-1} → significant AC between Y_{t-1} and Y_{t-2} → induces correlation between Y_t and Y_{t-2} ! (1st AC = PAC!)**

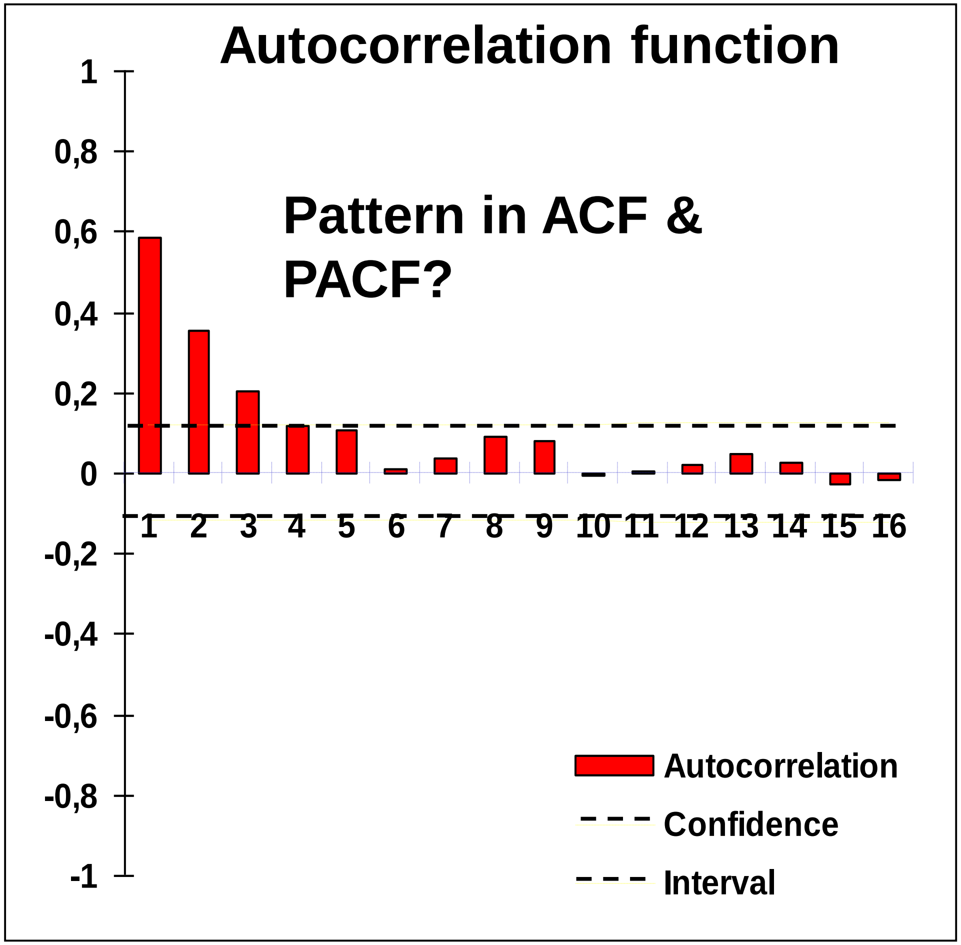
- **When fitting an AR(p) model to the time series, the last coefficient p of Y_{t-p} measures the excess correlation at lag p which is not accounted for by an AR(p-1) model. π_p is called the pth order partial autocorrelation, i.e.**

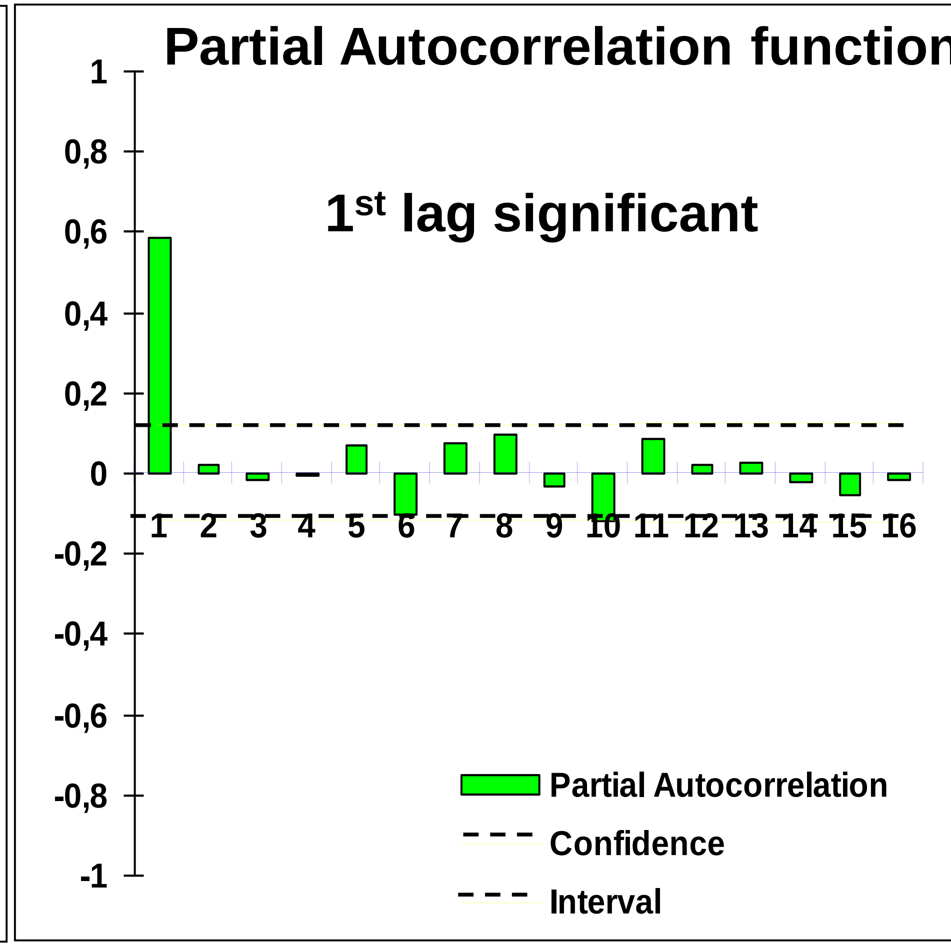
$$\pi_p = \text{corr}\left(Y_t, Y_{t-p} \mid Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p+1}\right)$$

- **Partial Autocorrelation coefficient measures true correlation at Y_{t-p}**

$$Y_t = \varphi_0 + \varphi_{p1} Y_{t-1} + \varphi_{p2} Y_{t-2} + \ldots + \pi_p Y_{t-p} + \nu_t$$

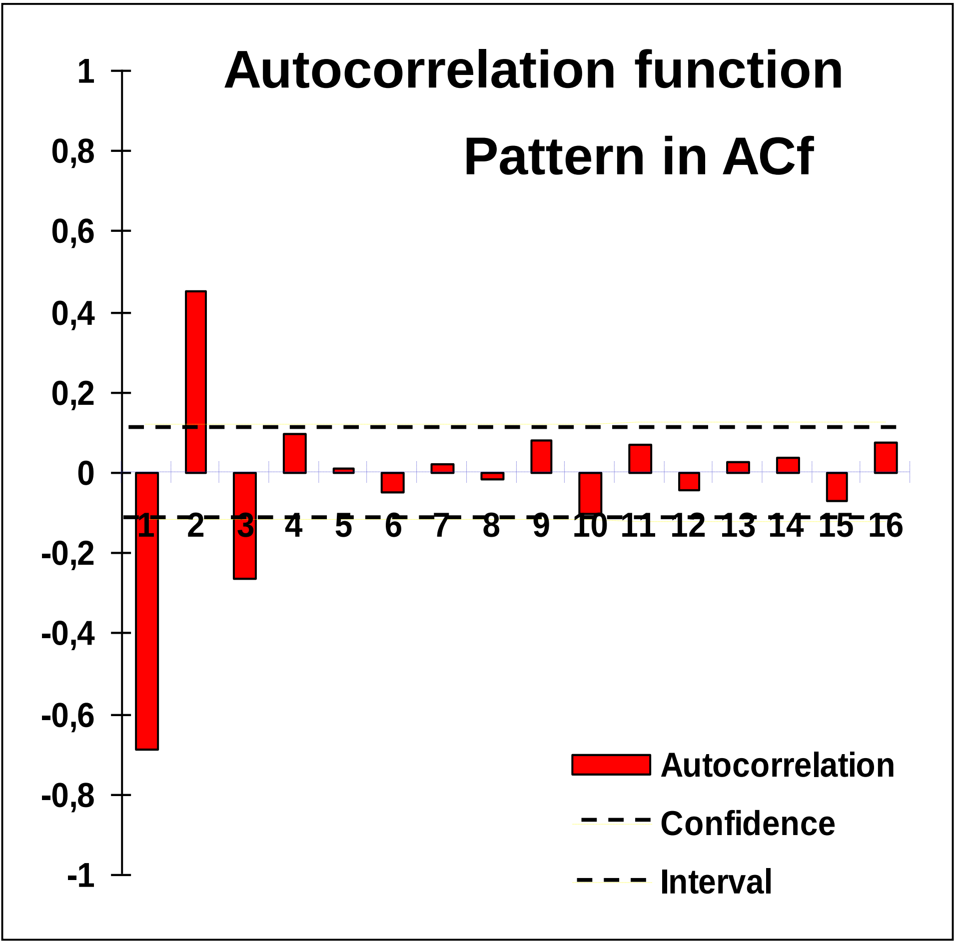
AR-Model patterns

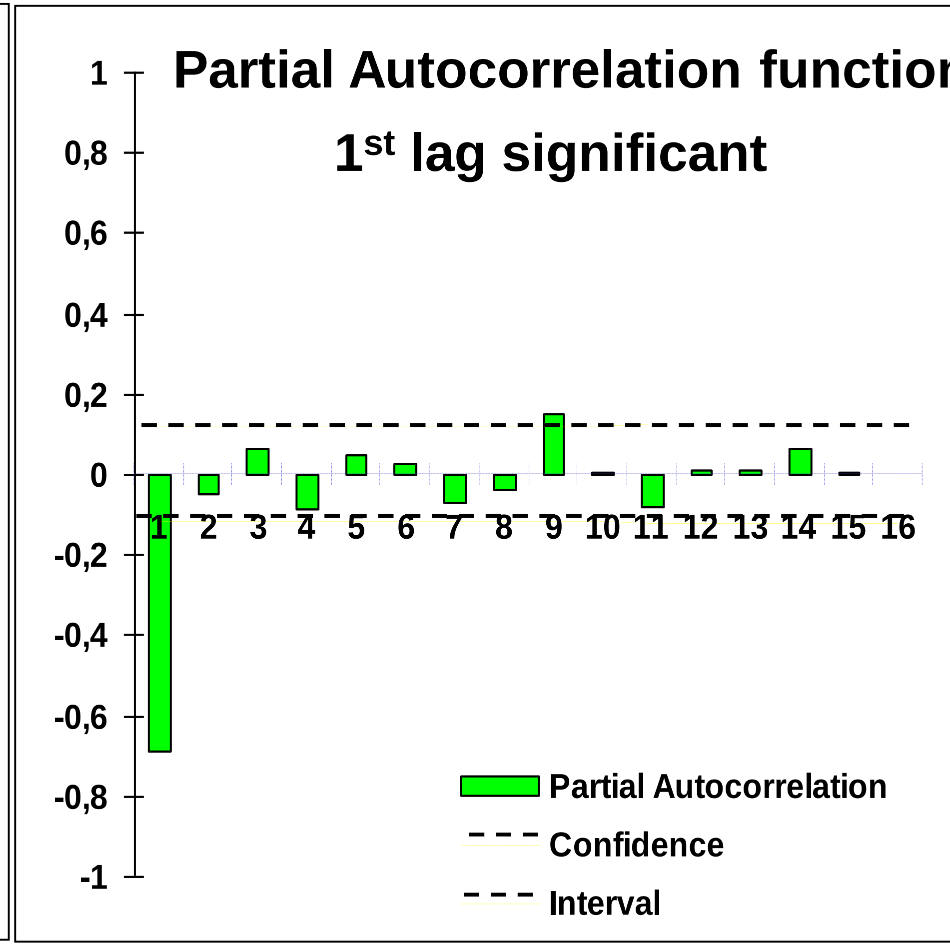




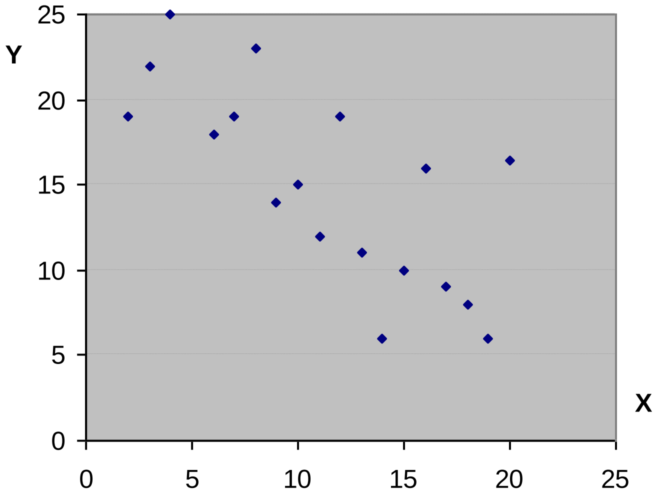
- **AR(1) model:** $Y_t = c + \phi_1 Y_{t-1} + e_t$ **=ARIMA (1,0,0)**

AR-Model patterns

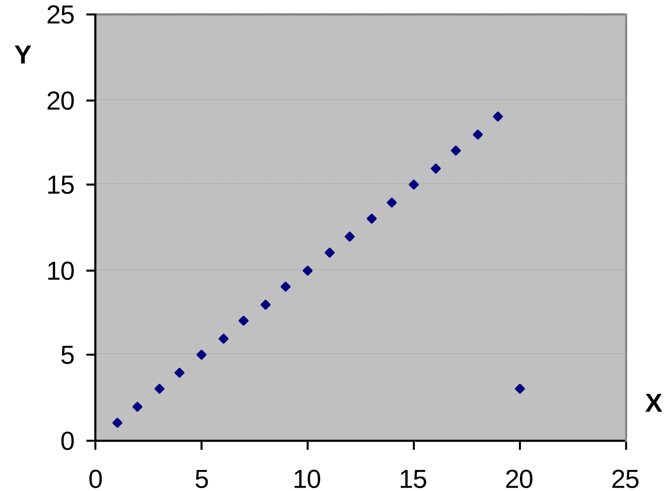




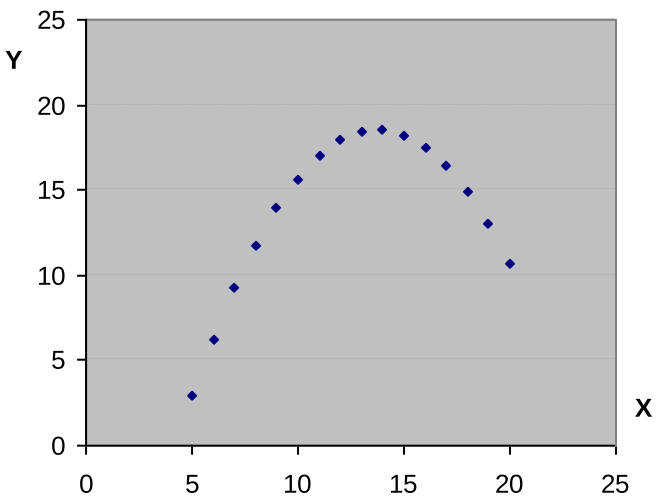
- **AR(1) model:** $Y_t = c + \phi_1 Y_{t-1} + e_t$ **=ARIMA (1,0,0)**



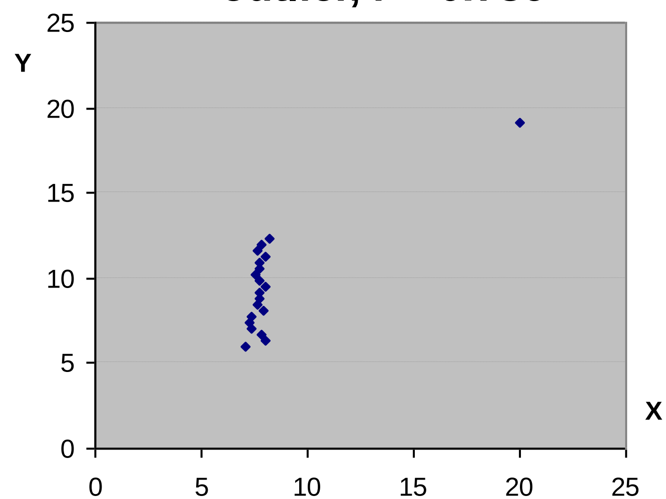
negative relationship, r = -0.786



positive linear relationship with one outlier, r = 0.786

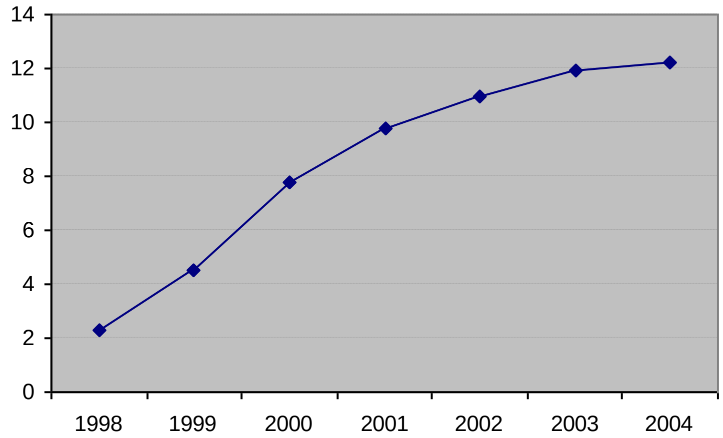


non linear relationship, r = 0.786

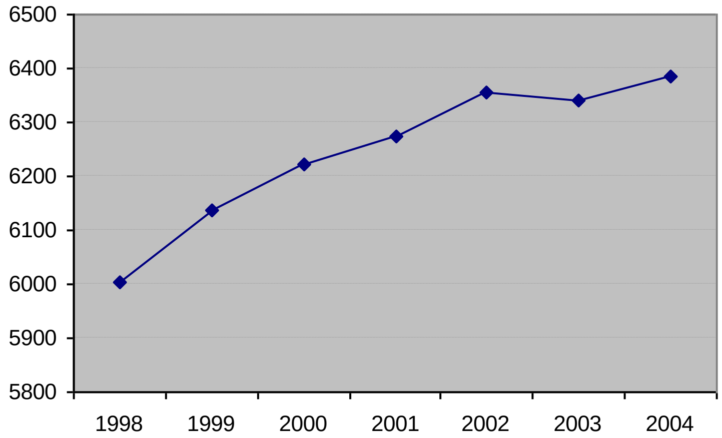


**no relationship,
with one outlier, r = 0.786**

Correlation and Causality



UK internet usage (millions of households)

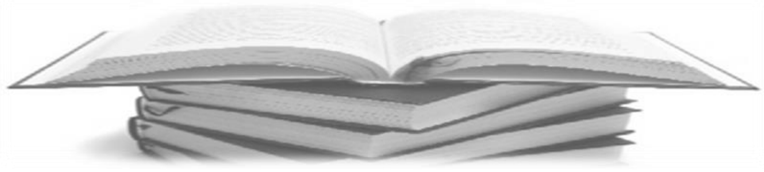


US beer sales (billions of gallons)

Correlation Matrix

	UK Pop	*Austria Pop*	*UK Internet*	*US Beer Sales*
UK Pop	1			
Austria Pop	0.989	1		
UK Internet	0.947	0.893	1	
US Beer Sales	0.937	0.883	0.986	1

the matrix shows strong correlations – but no causality



Model Selection

Pre-filtering and Data Exploration

Manual Data exploration can be time consuming, however if done by experts it allows identifying the structure of the time series; hence producing robust fits of appropriate models.

Furthermore, data exploration easily allows to identify data irregularities, such as outliers and structural breaks.

Once the time series have been classified according to their structure, adequate models can selected and parameterised.

A degree of automation on data exploration can be achieved through statistical testing.

Expert knowledge should be superimposed, when available.

Agenda

Time Series Analysis

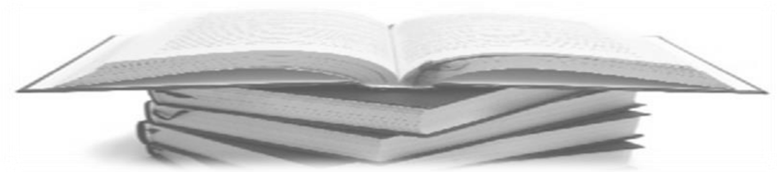
1. Data Exploration
   1. Time Series Patterns
      1. Level, Trend & Season
      2. Outliers & Structural breaks
   2. Graphical Identification of Time Series Patterns
      1. Time Series & Seasonal Plots
      2. Time Series Decomposition
      3. A related „discipline“: Technical Analysis
   3. Statistical Identification of Time Series Patterns
      1. Autocorrelation Analysis (Corelograms)
      2. Spectral Analysis
      3. Statistical Tests
2. Data Transformations
   1. Transforming level of series
   2. Transforming variance of time series

Model Selection
Statistical Tests for Filtering

- **Statistical Test for Regular Time Series Components**
  - **Cycle**
    - Irrelevant to operational forecasting, often stochastic
  - **Trend**
    - Existence of monotonic trend : Kendall’s Test, Spearman's Rho, Cox-Stuart (Trend)
    - Trend-form: Linear Coefficient (Monotonic linear trend) , Noether's Cyclical Trend
  - **Seasonality**
    - Existence of Seasonality: Kruskal Wallis Test, Chi-Squared-Mod-Test, F-Test, ACF-Heuristic
    - Deterministic / stochastic Seasonality pending; Multiple Seasonality pending
  - **Noise**
    - Cox-Stuart (Change in Dispersion)
    - Structure (Any type of non-randomness): Runs (Mean), Runs (Median), Runs (Up-Down)
  - **Other Characteristics**
    - Zero Values → Intermittent Demand
    - Length → New Products

- **Statistical Test for Irregular Time Series Components**
  - Structural Breaks → pending
  - Length / New Products → heuristic

→ Allow categorisation of time series in assortment
→ Ex ante Filter-approach (pre-model building)
→ limit model search space (e.g. non seasonal→ no dummies)



Model Selection
Statistical Tests

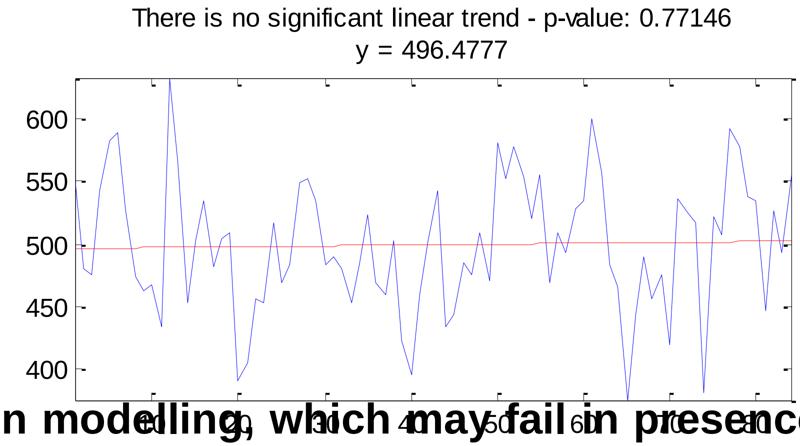
The visual tools of data exploration can be complimented by statistical tests, allowing to automate parts of the exploration.

Tests for trend:

1. **Ordinary least square or robust linear regression → Fit a regression conditional on time to the time series and assess whether the time coefficient is significant, in which case the time series is trended.**

$$\hat{A}_t = b_o + b_1 t$$

***b_0* is a constant and *b_1* is the trend coefficient. We test *b_1* for significance.**



This test is based on static regression modelling, which may fail in presence of outliers, level shifts, structural breaks, nonlinear trends or incomplete seasons.

Tests for trend:

2. **Nonparametric Cox-Stuart test. This test splits the time series in half. Then compares the two halves in pairs and finds the numbers that the first half is higher or lower than the second half. The number of positive, or negative differences, whichever is lowest, is taken as the test statistic that is compared against a binomial distribution with sample/2 degrees of freedom.**

Date	Value	Date	Value	Sign
1997	65.96	2004	77.77	+
1998	68.18	2005	68.49	+
1999	99.35	2006	1.77	-
2000	41.94	2007	104.73	+
2001	46.78	2008	33.70	-
2002	17.29	2009	82.14	+
2003	68.96	2010	28.40	-

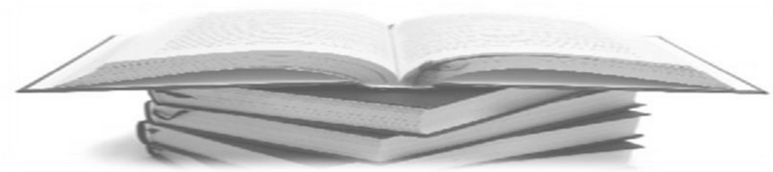
T = 3 (for three "-")

T ~ Binomial(7,0.5)

Number of samples/2

Probability of either direction

This test can mix trends and level shifts. Also, as nonparametric it's power is relatively weak.

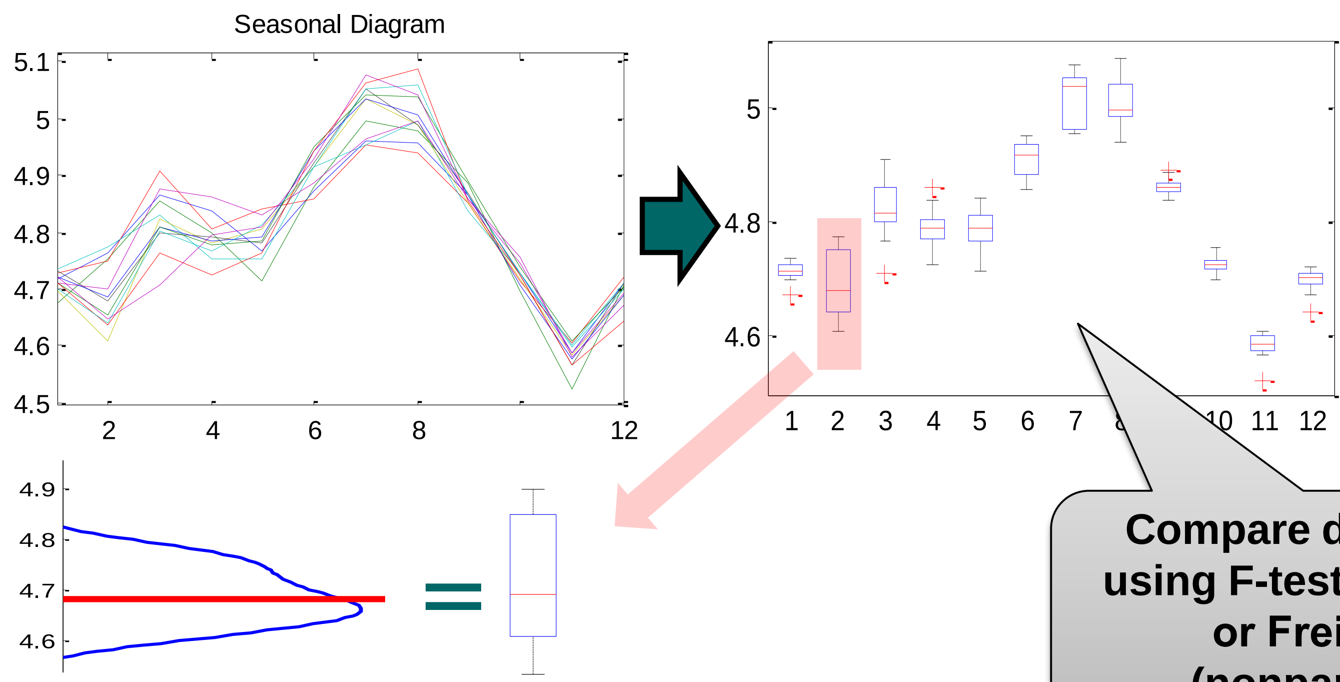


Model Selection

Statistical Tests for Seasonality

Tests for season:

1. **Variance of location (parametric/nonparametric) → From the seasonal diagram from each point in the season (e.g. month) construct the distribution and compare the location of the distributions (means for parametric case) across all points. If at least one is different we cannot reject the possibility of seasonality.**



Compare distributions using F-test (parametric) or Freidman's (nonparametric)

The performance of statistical tests heavily relies on the availability of data.

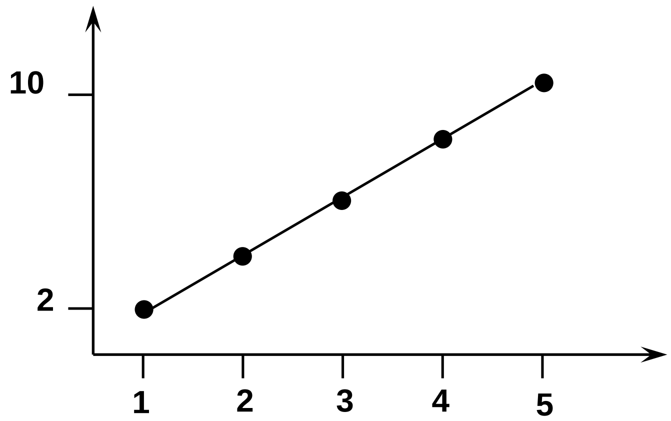
Agenda

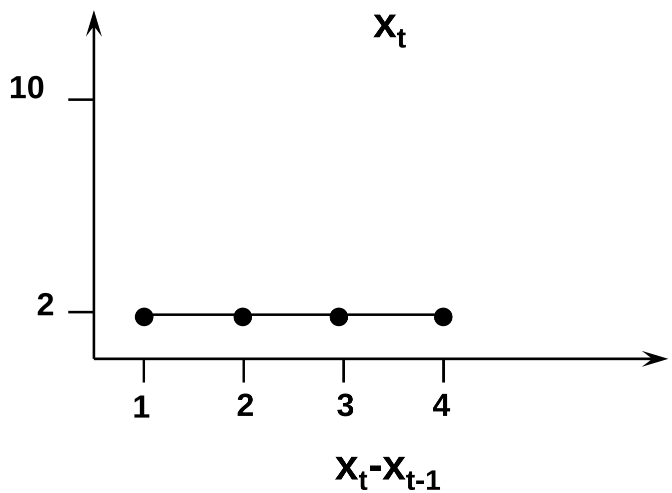
Time Series Analysis

1. Data Exploration
    1. Time Series Patterns
        1. Level, Trend & Season
        2. Outliers & Structural breaks
    2. Graphical Identification of Time Series Patterns
        1. Time Series & Seasonal Plots
        2. Time Series Decomposition
        3. A related „discipline“: Technical Analysis
    3. Statistical Identification of Time Series Patterns
        1. Autocorrelation Analysis (Corelograms)
        2. Spectral Analysis
        3. Statistical Tests
2. Data Transformations
    1. Transforming level of series
    2. Transforming variance of time series

ARIMA-Modells
Differencing for Stationariry

Differencing time series





$$Z_t = Y_t - Y_{t-1}$$

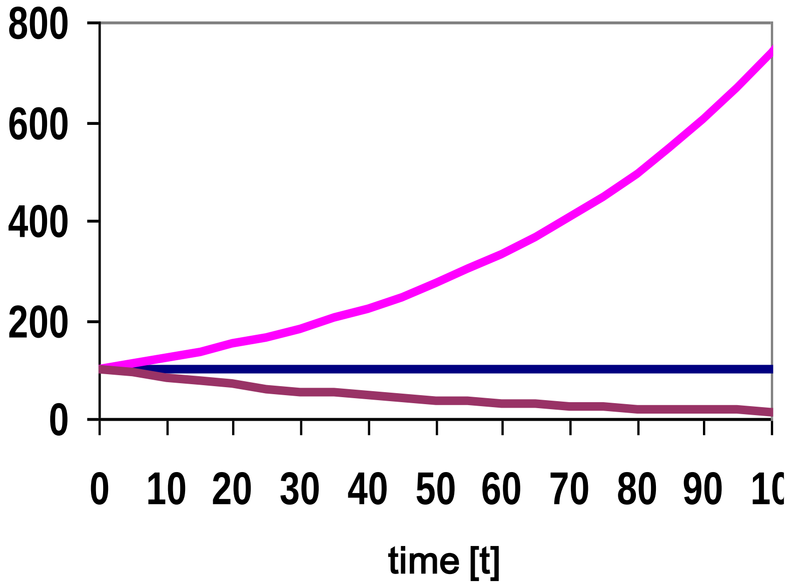
- **E.g. : time series Y_t={2, 4, 6, 8, 10}.**
- **time series exhibits linear trend**
- **1st order differencing between observation Y_t and predecessor Y_{t-1} derives a transformed time series:**

4-2=2
6-4=2
8-6=2
10-8=2

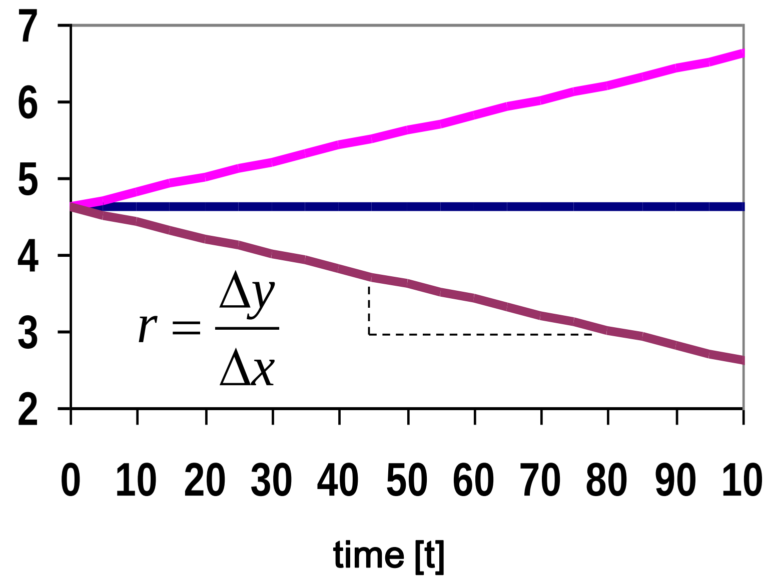
→ **The new time series ΔY_t={2,2,2,2} is stationary through 1st differencing**

- **Logarithms**

Population size N(t)



Logarithm of population size Ln(N(t))



r-values: 0.02, 0.00, -0.002

→ **Transform to constant rate of growth**

Dat Transformations
- Calendar Day Corrections

PROBLEM: Calendar effects in time series data

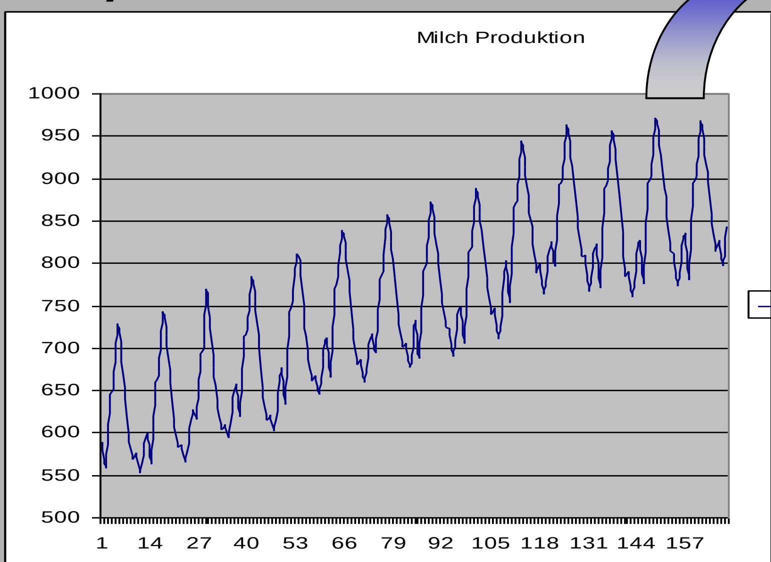
- Analysis of data in different periods then actually seasonal pattern distorts data
- Time series may be analyzed in different granularities e.g. daily / weekly / monthly
- distance between time points of observations is important for asymmetries from changing calendars (different number of working days in each country / month / region → keep calendar)

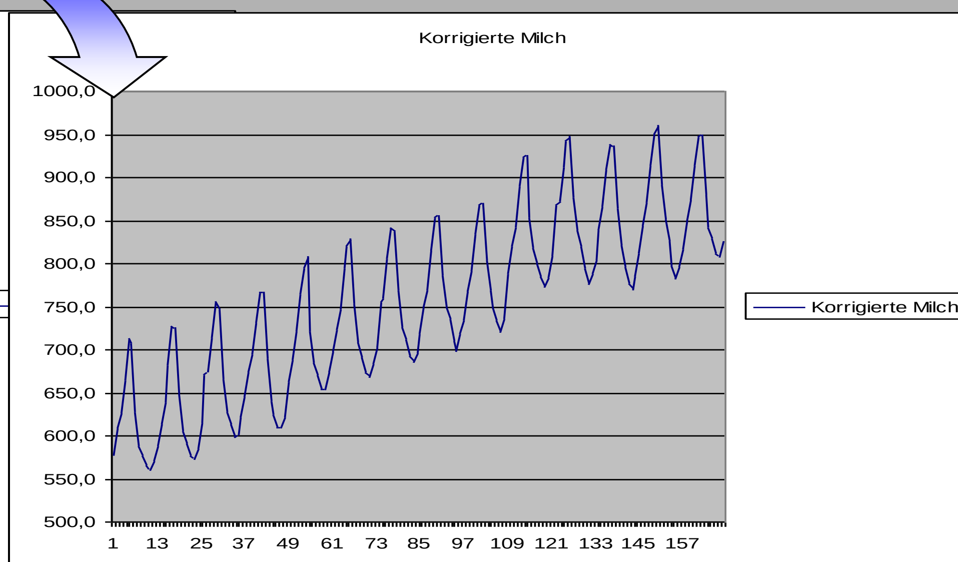
Problems through forecasts on MONTHLY BASIS instead of weekly!

- Correction of monthly asymmetries / weekly asymmetries
  - Length of month is different: January 31; February 28; March 31; ... (31,28,31,30,31,30,31,31,30,31,30,31)
  - Length of week is almost constant → Number of days is most constant
  - Correction of calendar day / working-day / bank holiday- / day-of-the-week asymmetries
    - Single bank holidays are in different week / month of the year
    - Changing number of working days per calendar week / calendar month
  - Different LENGTH of time series through collection interval
    - 3 years = 36 monthly observation = 156 weekly values = 1095 daily values → SELECTION PROBLEM
    - ...

$$\frac{31-28}{30,4}=0,98=9,8\%$$

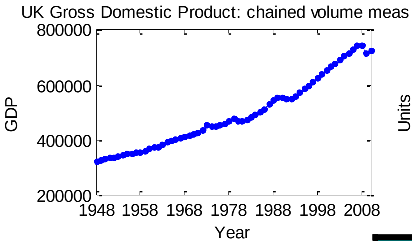
Daily Milk Production Dataset

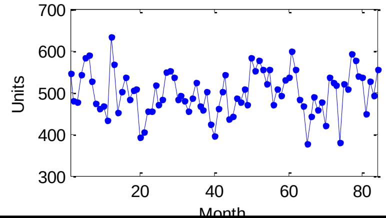


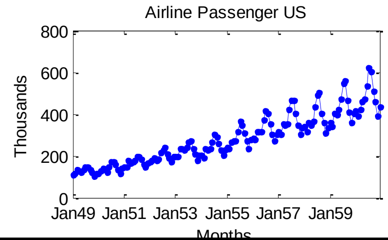


Model Selection

What is Model Selection?







Identify time series pattern … ?

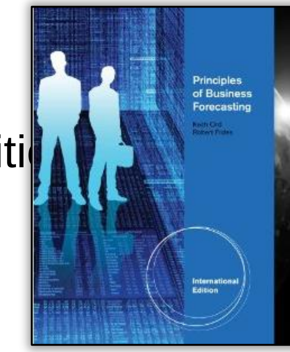
Select „best"/ suitable forecast algorithm … ?

	No Seasonal Effect	Additive Seasonal	Multiplicative Seasonal
No Trend Effect	• naïve methods (-) • averages (-) • moving averages (-) • single exponential smoothing (1st order: α) • linear regression (-)	• Seasonal Exponential Smoothing by Winters – additive (3rd order α,β=0,δ) • seasonal regression (-)	• Seasonal exponential Smoothing by Winters – multiplicative (3rd order α, β=0, δ) • seasonal regression (-)
Additive Trend Effect	• Double Exponential Smoothing by Brown (2nd order: α, β= α) • Linear Exponential Smoothing by Holt (2nd order: α, β) • linear Regression (-)	• Seasonal Exponential Smoothing by Holt-Winters – additive (3rd order α, β, δ)	• Seasonal Exponential Smoothing by Holt-Winters – multiplicative (3rd order: α, β=0, δ)
Multiplicative Trend Effect	• Double Exponential Smoothing by Brown (2nd order: α, β= α) • Linear Exponential Smoothing by Holt (2nd order: α, β) • linear Regression (-)	• Seasonal Exponential Smoothing by Holt-Winters – additive (3rd order α, β, δ)	• Seasonal Exponential Smoothing by Holt-Winters – multiplicative (3rd order: α, β=0, δ)

→Forecasting algorithms were developed for specific Time Series Patterns
→Algorithms are not downwards compatible
→requires (1.) model identification & (2.) matching to algorithms!

Lancaster Centre for Forecasting
www.forecasting-centre.com

Homework!



MUST HAVE (!): Principles of Business Forecasting

1. **<u>Recap Lecture 2 (today):</u>**
   Read Chapter 4.0 - 4.5 without 4.21 (pages 99-111) on Decompositi
2. **<u>Prepare Lecture 3 (next session):</u>**
   Read Chapter 3 (pages 59-91)

NICE TO HAVE

1. N/A

Questions?

Dr. Sven F. Crone
Lancaster University Management School
Centre for Forecasting - Lancaster, LA1 4YX
email   n.kourentzes@lancaster.ac.uk

Copyright