Data Wrangling and Data Analysis

Time Series and Demand Forecasting

Hakim Qahtan

Department of Information and Computing Sciences

Utrecht University

Utrecht University

1

Topics for Today

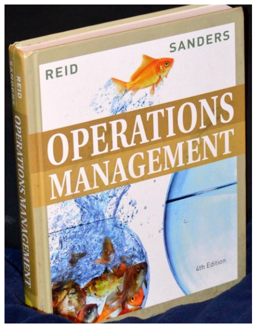
- Time Series Analysis
- Demand Forecasting

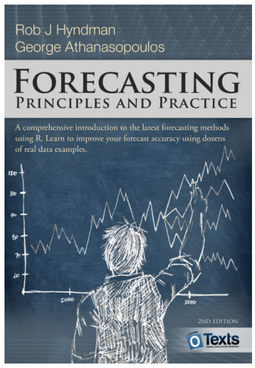
Utrecht University

2

Reading Material

- Operations Management (4-th Edition)
  - Reid Sanders
- A good replacement would be:
- Forecasting: Principles and Practice (2nd ed)
  - Rob J Hyndman and George Athanasopoulos





Utrecht University

3

Time Series Analysis

Utrecht University

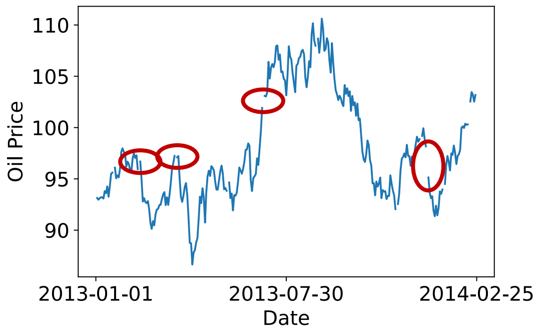
4

Time Series

- A **Time Series** is a set of observations measured at specified, usually equal, time intervals
- Adjacent observations are dependent

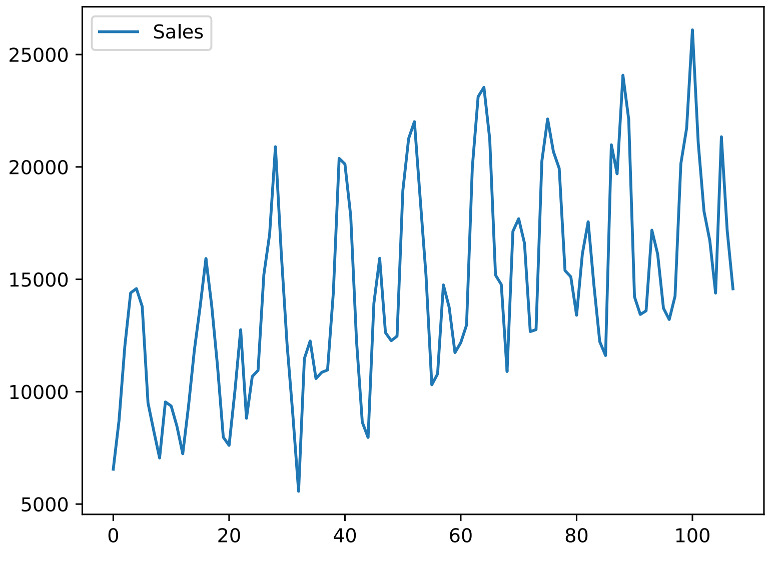
Oil Price in Ecuador 2013

Date	Oil Price
2013-01-01	nan
2013-01-02	93.14
2013-01-03	92.97
2013-01-04	93.12
2013-01-07	93.2
2013-01-08	93.21
2013-01-09	93.08
2013-01-10	93.81
2013-01-11	93.6
2013-01-14	94.27



Utrecht University

5



Car Sales Dataset

Utrecht University

6

Time Series Examples

- Sales data
- Gross national product
- Share prices
- Euro-to-Dollar Exchange rate
- Unemployment rates
- Population
- Interest rates
- Weather readings: temperature, humidity and wind speed
- …

Time Series Analysis

- Adjacent observations in a time series are dependent
- Time series analysis attempts to identify the factors that exert influence on the values in the series
  - Concerned with techniques for the analysis of dependence between the observations

Time Series Analysis

- Areas of application
  - Forecasting
    - Industry and government must forecast future activity to make decisions and plans to meet projected changes
  - Determining the transfer function of a system
    - Determining the effect of any given series of inputs on the output of a system
  - Using indicator input variables in transfer function
    - Assess the effects of unusual intervention events on the behavior of a time series
  - Examining the interrelationships among several related time series variables

Utrecht University

9

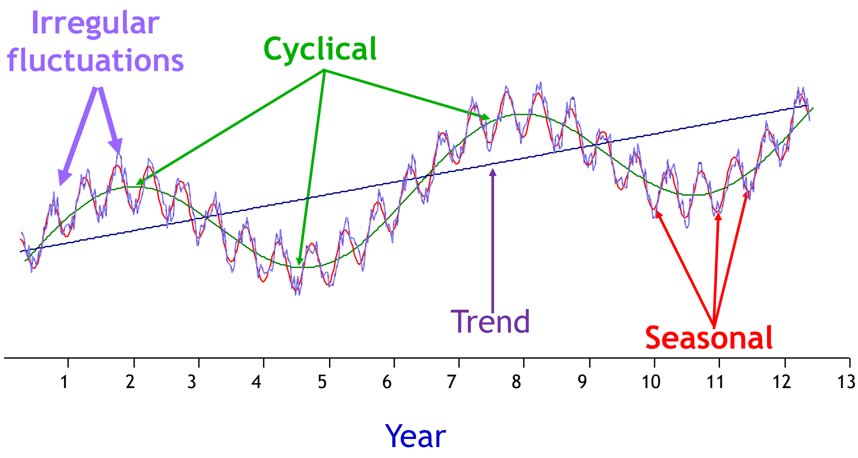
Time Series Components

- Can be broken into these four components:
  - Trend
  - Seasonal variation
  - Cyclical variation
  - Irregular variation

Utrecht University

10

Time Series Components



Utrecht University

11

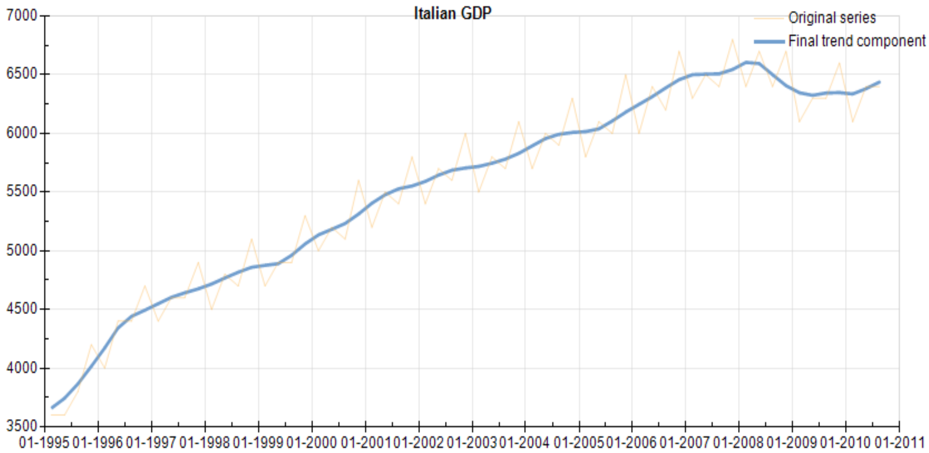
Time Series Components – Trend

- This is the long-term growth or decline of the series
  - In economic terms, long term may mean >10 years
  - Describes the history of the time series
  - Uses past trends to make prediction about the future
- An analysis of the trend of the observations is needed to acquire an understanding of the progress of events leading to prevailing conditions

Utrecht University

12

Time Series Components – Trend



Utrecht University

13

Remarks: Time Series Components – Trend

- Trend estimates are often reliable; however, in some instances the usefulness of estimates is reduced by:
  - high degree of irregularity in original or seasonally adjusted series or
  - abrupt change in the time series characteristics of the original data

Utrecht University

14

Remarks: Time Series Components – Trend

Utrecht University

15

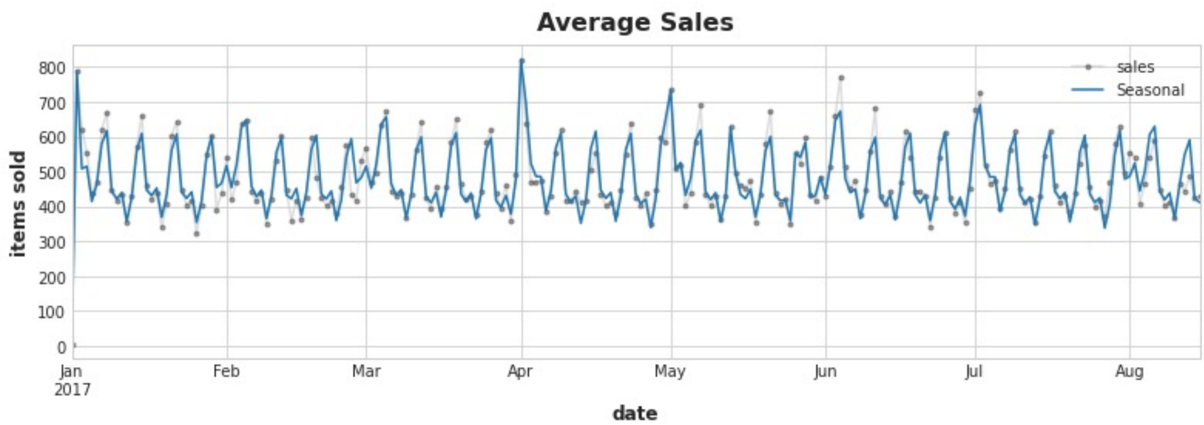
Time Series Components – Seasonal Variation (Seasonality)

- Seasonal variation of a time series is a pattern of change that **recurs** regularly over time.
- Seasonal variations are usually due to the differences between seasons and to festive occasions such as Easter and Christmas.
  - Usually changes occur within a year
- Examples include:
  - Air conditioner sales in Summer
  - Heater sales in Winter
  - Flu cases in Winter
  - Airline tickets for flights during school vacations

Utrecht University

16

Time Series Components – Seasonal Variation (Seasonality)



Seasonality in a time series can be identified by regularly spaced peaks and troughs (Kaggle)

Utrecht University

17

Time Series Components – Cyclical Variation

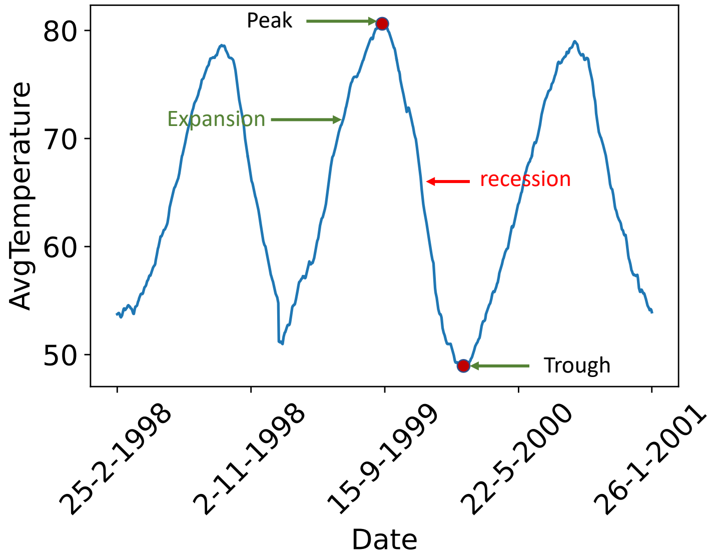
- Cyclical variations have recurring patterns but with a longer and more erratic time scale compared to Seasonal variations (e.g. 2- 10 years)
- The name is quite misleading
  - these cycles can be far from regular
  - it is usually impossible to predict the length of expansion/contraction periods
- There are no guarantees of a regularly returning pattern.
- Examples include:
  - Floods
  - Wars
  - Changes in interest rates
  - Economic depressions or recessions

Utrecht University

18

Time Series Components – Cyclical Variation

Average temperature in Algeria (10-1997 to 2-1999) smoothed by moving average



Utrecht University

19

Time Series Components – Irregular Variation

- An irregular (or random) variation in a time series occurs over varying (usually short) periods.
- It follows no pattern and is, by nature, unpredictable.
- It usually occurs randomly and may be linked to events that also occur randomly.
- Irregular variation cannot be explained mathematically.

Utrecht University

20

Time Series Components – Irregular Variation

- If the variation cannot be accounted for by the trend, season or cyclical variation, then it is usually attributed to irregular variation.

Example include:

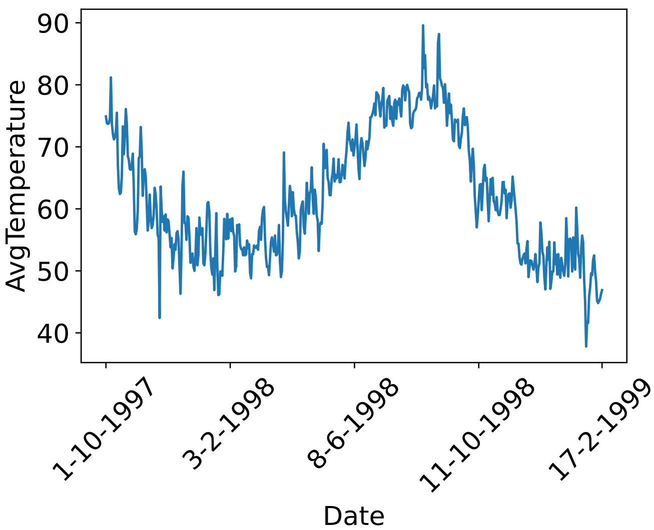
- Sudden changes in interest rates
- Collapse of companies
- Natural disasters
- Sudden shifts in government policy
- Dramatic changes to the stock market
- Effect of Middle East unrest on petrol prices

Utrecht University

21

Time Series Components – Cyclical Variation

Average temperature in Algeria (10-1997 to 2-1999)



Utrecht University

22

Time Series Decomposition

- Additive decomposition $y_t = S_t + T_t + R_t = \hat{S}_t + \hat{T}_t + \hat{R}_t$
  - y_t: time series values
  - S_t: seasonal component
  - T_t: trend component
  - R_t: the reminder
  - Good for situation when the variation in the seasonal fluctuations is almost stable
- Multiplicative decomposition $y_t = S_t \times T_t \times R_t = \hat{S}_t \times \hat{T}_t \times \hat{R}_t$
  - Common with economic time series

Utrecht University

23

Time Series Decomposition – Detrending

- Estimate a smoothed trend $\hat{T}_t$ (details will follow)
- Remove the smoothed $\hat{T}_t$ from y_t to get the$\hat{S}_t$ and $\hat{R}_t$
- In additive decomposition $y_t - \hat{T}_t = \hat{S}_t + \hat{R}_t$
- In multiplicative decomposition $\frac{y_t}{\hat{T}_t} = \hat{S}_t \times \hat{R}_t$

Utrecht University

24

Time Series Decomposition – Seasonal Component

- Compute a seasonal index for each season over the past years
  - Per month $\{\hat{S}^{(1)}, \ldots, \hat{S}^{(12)}\}$ or quarter $\{\hat{S}^{(1)}, \ldots, \hat{S}^{(4)}\}$, …
- If necessary, adjust the indices so that:
  - For additive decomposition $\hat{S}^{(1)} + \cdots + \hat{S}^{(m)} = 0$, m = number of seasons
  - For multiplicative decomposition $\hat{S}^{(1)} + \cdots + \hat{S}^{(m)} = m$

Utrecht University

25

Time Series Decomposition – Reminder Component

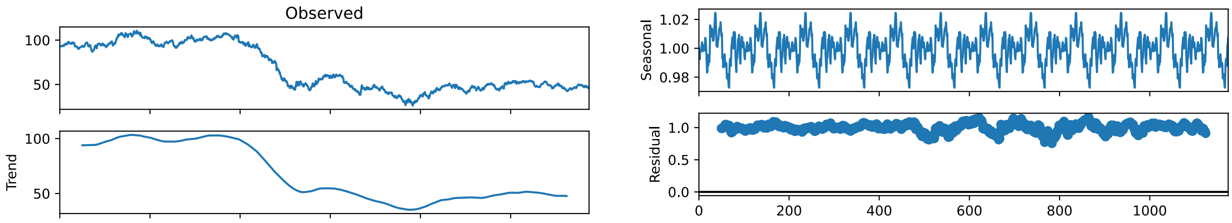
- For additive decomposition $\hat{R}_t = y_t - \hat{T}_t - \hat{S}_t$
- For multiplicative decomposition $\hat{R}_t = \frac{y_t}{\hat{T}_t \hat{S}_t}$

Utrecht University

26

```
from pandas import Series
from matplotlib import pyplot as plt
from statsmodels.tsa.seasonal import seasonal_decompose
%matplotlib inline
series = list(df_oil.dcoilwtico.dropna())
result = seasonal_decompose(series, model='multiplicative', period=100)
result.plot()
plt.savefig("oil_price.pdf", format="pdf", bbox_inches="tight")
plt.show()
```

You can use
'additive'



Utrecht University

27

Time Series Components – Trend

Utrecht University

28

Measuring the Trend

- An essential aim in time series analysis is using the past information to establish plan for the next period.
  - Measuring the trend depicts the general direction of the trend line over time
- The trend can be affected by:
  - Population changes
  - Productivity improvement
  - Technological advancements
  - Global crisis
  - Market changes
  - …

Utrecht University

29

Why Examine the Trend?

- If the current trend is expected to continue, it can be used for future planning:
  - Capacity planning for increased population
  - Utility loads
  - Market progress
  - Required resource for new students
  - Expected workload
    - Emergency calls
    - Taxi demand
    - Occupied beds in a hospital

Utrecht University

30

Depicting the Trend

- Common methods include:
  - Semi-average
  - Moving average
  - Least-square
  - Exponential smoothing

Depicting the Trend – Semi-Average

- Attempts to fit a straight line to describe the trend:
  - Divide the data into 2 equal time ranges
  - Calculate the average of the observations in each of the 2-time ranges
  - Draw a straight line between the 2 points
  - Extend the line slightly past the end of the original observation to make predictions for future years

Depicting the Trend – Moving-Average

- Based on the premise that if values in a time series are averaged over a sufficient period, the effect of short-term variations will be reduced

- The degree of smoothing can be controlled by selecting the number of cases to be included in an average.

Utrecht University

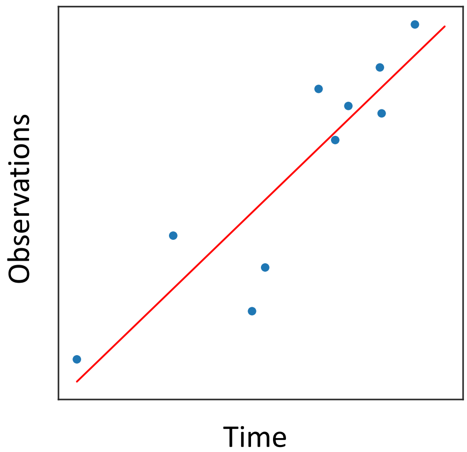
33

Depicting the Trend – Least-Square Linear Regression

- A more sophisticated way for fitting a straight line to a time series is using the method of least squares regression.

Sound familiar?

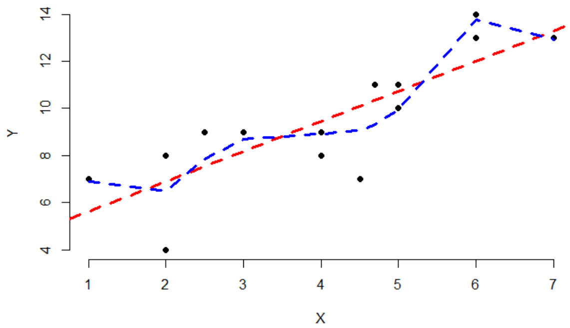
- Observations are the dependant variables (*y*)
- Time is the independent variable (*x*).



Utrecht University

34

Depicting the Trend – Least-Square Linear Regression



Given a set of points (x_i, y_i) such as the points in the scatterplot, find the best fitting line

$$f(x_i) = a + bx_i$$

such that:

$$SSE = \sum_i \left(y_i - f(x_i)\right)^2$$

$$= \sum_i (y_i - a - bx_i)^2$$

is minimized

Utrecht University

35

Depicting the Trend – Least-Square Linear Regression

- The above optimization problem can be solved by:
  1. Taking the partial derivatives of SSE with respect to a and b
  2. Setting $\frac{\partial SSE}{\partial a}$ and $\frac{\partial SSE}{\partial b}$ to 0
  3. Solving the system of linear equations

     Since: $SSE = \sum_i (y_i - a - bx_i)^2$

     Then $\frac{\partial SSE}{\partial a} = -2\sum_i (y_i - a - bx_i) = 0$

     And $\frac{\partial SSE}{\partial b} = -2\sum_i x_i(y_i - a - bx_i) = 0$

Utrecht University

36

Depicting the Trend – Least-Square Linear Regression

- The equations can be summarized by the normal equation:

$$\begin{pmatrix} N & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{pmatrix}$$

Utrecht University

37

Depicting the Trend – Least-Square Linear Regression

- Practically:
  - Determine the number of samples (n)
  - Allocate midpoint in time and replace the time points by their corresponding x values by increasing and decreasing one unit from the midpoint accordingly.
  - The dependent variable is "y"
  - Compute Σx_i^2 and $\Sigma x_i y_i$
    - Σx_i should be 0.
  - Find $y = a + bx$ where $b = \frac{\sum x_i y_i}{\Sigma x_i^2}$ and $a = \frac{\sum y_i}{n}$ (refer to the previous slide and keep in mind that Σx_i = 0)

Utrecht University

38

Example

- Consider the following dataset

Year	2003	2004	2005	2006	2007	2008	2009	2010	2011
x	-4	-3	-2	-1	0	1	2	3	4
y	13	15	17	18	19	20	20	21	22

$$n = 9 \qquad \sum_i x_i = 0 \qquad \sum_i x_i^2 = 60$$

$$\sum_i y_i = 165 \qquad \sum_i x_i y_i = 62$$

By definition, $\sum_i x_i = 0$

Utrecht University

39

Example

- Consider the following dataset

Year	2003	2004	2005	2006	2007	2008	2009	2010	2011
x	-4	-3	-2	-1	0	1	2	3	4
y	13	15	17	18	19	20	20	21	22

$$a = \frac{\sum y_i}{n} = \frac{165}{9} = 18.3$$

$$b = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{62}{60} = 1.03$$

Exercise: Predict the sales for year 2014.

Utrecht University

40

Depicting the Trend – Exponential Smoothing

- History is used to flatten out short term fluctuations

$$S_x = \alpha y + (1 - \alpha) S_{x-1}$$

- S_x = the smoothed value for observation x
- y = the actual value of observation at time x
- S_{x-1} = the smoothed value previously calculated for observation at time $(x - 1)$
- α = the **smoothing constant** where $0 \leq \alpha \leq 1$

Utrecht University

41

Depicting the Trend – Exponential Smoothing

- α = the **smoothing constant** where $0 \leq \alpha \leq 1$
  - α is small $\Longrightarrow$ more weight for the past measurements
  - α is large $\Longrightarrow$ more weight for the present trend
- This approach needs a starting point
  - we choose the first smoothed value (S_1) to be the first observation (y_1)
- The smoothed value of each observation is a function of the smoothed value of the observation immediately before it
- Suffers from propagation error

Utrecht University

42

Seasonal Variation

Seasonal Variations

- Periodic movements in the time series
- It is important to consider seasonal variations for future planning
- A seasonally adjusted series involves estimating and removing the cyclical and seasonal effects from the original data
- For example:
  - employment and unemployment are often seasonally adjusted so that the actual change in employment and unemployment levels can be seen, without the impact of periods of peak employment such as Christmas/New Year when a large number of casual workers are temporarily employed

Seasonal Variations – Example

- Adverse publicity in December about ice-cream
- It would be incorrect simply to compare the sales of ice-cream in June with those in December to determine the effect of the adverse publicity. Sales rate in June is higher in any case, because it is warmer

- Useful comparisons of sales could only be made after removing the seasonal variation so the true impact of the publicity would be more realistic

Utrecht University

45

Compute the Seasonal Index

- To remove the seasonal effect before finding the trend in the data
- Simple average method
  - Take the average for each period (period mean) over at least three years
  - Express that as an index by comparing it to the average of all periods over the same period of time
- Note: indices can be based on periods such as months or weeks

Utrecht University

46

Compute the Seasonal Index – Example

- Consider the data

Year	Quarter			
	1	**2**	**3**	**4**
1994	43	64	63	41
1995	46	64	67	43
1996	51	69	75	39
1997	55	73	79	48
Quarterly total	**195**	**270**	**284**	**171**
Quarterly mean	**48.75**	**67.5**	**71**	**42.75**

Yearly average 1994 = (43 + 64 + 63 + 41)/4 = 211/4 = 52.75

Yearly average 1995 = (46 + 64 + 67 + 43)/4 = 220 /4 = 55.0

Yearly average 1996 = (51 + 69 + 75 + 39)/4 = 234/4 = 58.5

Yearly average 1997 = (55 + 73 + 79 + 48)/4 = 255 / 4 =63.75

Utrecht University

47

Compute the Seasonal Index – Example

- Compute the yearly average of the values and divide the quarterly reading over the yearly average

Year	Quarter			
	1	**2**	**3**	**4**
1994	43/52.75	64/52.75	63/52.75	41/52.75
1995	46/55	64/55	67/55	43/55
1996	51/58.5	69/58.5	75/58.5	39/58.5
1997	55/63.75	73/63.75	79/63.75	48/63.75

Yearly average 1994 = (43 + 64 + 63 + 41)/4 = 211/4 = 52.75

Yearly average 1995 = (46 + 64 + 67 + 43)/4 = 220 /4 = 55.0

Yearly average 1996 = (51 + 69 + 75 + 39)/4 = 234/4 = 58.5

Yearly average 1997 = (55 + 73 + 79 + 48)/4 = 255 / 4 =63.75

Utrecht University

48

Compute the Seasonal Index – Example

- Compute the index as the quarterly average over the years

Year	Quarter			
	1	2	3	4
1994	0.82	1.21	1.19	0.78
1995	0.84	1.16	1.22	0.78
1996	0.87	1.18	1.28	0.67
1997	0.86	1.15	1.24	0.75

Divide the Quarterly value over the yearly average

Year	Seasonal index (Quarterly)			
	1	2	3	4
1994	82	121	119	78
1995	84	116	122	78
1996	87	118	128	67
1997	86	115	124	75
Seasonal Index (Over the years)	**84.8**	**117.5**	**123.3**	**74.5**

Multiply by 100

Utrecht University

49

Compute the Seasonal Index – Example

- Remove the seasonal effect from the data (multiplicative model)

Year	Quarter	Actual Value	Seasonal Index	Adjusted Values
1996	1	51	84.8	60
	2	69	117.5	59
	3	75	123.3	61
	4	39	74.5	52
1997	1	55	84.8	65
	2	73	117.4	62
	3	79	123.3	64
	4	48	74.5	64

$$Seasonally\ Adjusted\ data = \frac{Actual\ Values}{Seasonal\ Index} \times 100$$

Utrecht University

50

Compute the Seasonal Index – Example

- Alternate approach to calculate Seasonal Index number for each quarter
  - Take the quarterly mean over the years
  - Divide each mean value by the mean of means multiplied by 100

Quarter	Quarterly mean	Seasonal Index
1	48.75	48.75/57.5*100 = 84.78
2	67.5	67.5/57.5*100 = 117.39
3	71	71/57.5*100 = 123.478
4	42.75	42.75/57.5*100 = 74.348
Means Total	**230**	**400**
Mean of Means	**57.5**	**100**

Utrecht University

51

Compute the Seasonal Index – Example

- Remove the seasonal effect from the data

Year	**Quarter**	Actual Value	Seasonal Index	Adjusted Values
1996	1	51	84.8	60
	2	69	117.4	59
	3	75	123.5	61
	4	39	74.3	52
1997	1	55	84.8	65
	2	73	117.4	62
	3	79	123.5	64
	4	48	74.3	65

$$Seasonally\ Adjusted\ data\ = \frac{Actual\ Values}{Seasonal\ Index} \times 100$$

Utrecht University

52

Time Series Decomposition – Widely used Approach

- Additive decomposition $y_t = S_t + T_t + R_t = \hat{S}_t + \hat{T}_t + \hat{R}_t$
- Multiplicative decomposition $y_t = S_t \times T_t \times R_t = \hat{S}_t \times \hat{T}_t \times \hat{R}_t$
- Detrended data
  - In additive decomposition $y_t - \hat{T}_t = \hat{S}_t + \hat{R}_t$
  - In multiplicative decomposition $\frac{y_t}{\hat{T}_t} = \hat{S}_t \times \hat{R}_t$
- Detrending and removing the seasonal effect
  - For additive decomposition $\hat{R}_t = y_t - \hat{T}_t - \hat{S}_t$
  - For multiplicative decomposition $\hat{R}_t = \frac{y_t}{\hat{T}_t \hat{S}_t}$

Utrecht University

53

STL Decomposition

- Seasonal and Trend decomposition using Loess
  - Handle any type of seasonality
  - Seasonal component is allowed to change over time
  - Smoothness of the trend can be controlled by the user
  - Can be robust to outliers

```
from statsmodels.tsa.seasonal import STL
df_oil_new = df_oil.dropna()
dcoilwtico = list(df_oil_new.dcoilwtico)
oil_data = pd.Series(
dcoilwtico, index=df_oil_new.date, name="OIL")
stl = STL(oil_data, period = 100)
res = stl.fit()
fig = res.plot()
```

Utrecht University

54

Demand Forecasting

Decisions that Require Forecasting

- What products to produce?
- How many people to hire?
- How many units to purchase?
- How many units to produce?
- How many items to order?
- And so on……

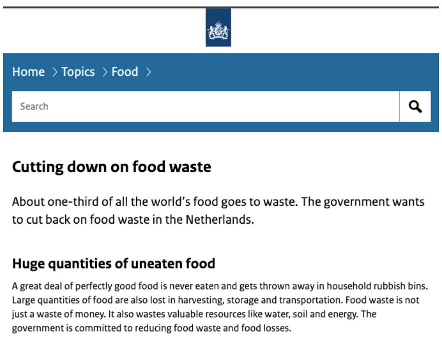
Common Characteristics of Forecasting

- Forecasts are rarely perfect
- Forecasts are more accurate for aggregated data than for individual items
- Forecast are more accurate for shorter than longer time periods

Utrecht University

57

Why Forecasting is Important?



58

Forecasting Techniques

- Naïve Forecasting
- Simple Mean
- Moving Average
- Weighted Moving Average
- Exponential Smoothing

Utrecht University

59

Forecasting – Example

- Determine forecast for periods 11
  - Naïve forecast
  - Simple average
  - 3- and 5-period moving average
  - 3-period weighted moving average with weights 0.5, 0.3, and 0.2
  - Exponential smoothing with alpha=0.2 and 0.5

Period	Orders
1	122
2	91
3	100
4	77
5	115
6	58
7	75
8	128
9	111
10	88

Utrecht University

60

Naïve Forecasting

- Next period's forecast = previous period's actual

$$\hat{y}_{t+1} = y_t$$

$\hat{y}_t$ represents the predicted value at time t

y represents the actual value at time t

Period	Orders	Naïve Forecast
1	122	
2	91	122
3	100	91
4	77	100
5	115	77
6	58	115
7	75	58
8	128	75
9	111	128
10	88	111
11		88

Utrecht University

61

Simple Average

- Next period's forecast = average of previously overserved data

$$\hat{y}_{t+1} = \frac{y_1 + y_2 + \cdots + y_t}{t}$$

Period	Orders	Simple Average
1	122	
2	91	122
3	100	107
4	77	104
5	115	98
6	58	101
7	75	94
8	128	91
9	111	96
10	88	97
11		97

Utrecht University

62

Moving Average

- Next period's forecast = simple average of the last k periods

$$\hat{y}_{t+1} = \frac{y_{t-k+1} + y_{t-k+2} + \cdots + y_t}{k}$$

- Also called **Rolling Window**
- A smaller k makes the forecast more **responsive**
- A larger k makes the forecast more **stable**

Period	Orders	Moving Average (k = 3)	Moving Average (k = 5)
1	122		
2	91		
3	100		
4	77	104	
5	115	89	
6	58	97	101
7	75	83	88
8	128	83	85
9	111	87	91
10	88	105	97
11		109	92

Utrecht University

63

Weighted Moving Average

- Next period's forecast = weighted average of the last k periods

$$\hat{y}_{t+1} = c_1 y_{t-k+1} + \cdots + c_k y_t$$

With

$$c_1 + c_2 + \cdots + c_k = 1$$

We take $c_1 = 0.2$, $c_2 = 0.3$ and $c_3 = 0.5$

Period	Orders	Weighted Moving Average (k = 3)
1	122	
2	91	
3	100	
4	77	102
5	115	87
6	58	101
7	75	79
8	128	78
9	111	98
10	88	109
11		103

Utrecht University

64

Exponential Smoothing

- Next period's forecast = weighted average of the previous reading and the history

$$\hat{y}_{t+1} = \alpha\, y_t + (1-\alpha)\hat{y}_t$$

$$\hat{y}_3 = 0.2 * 91 + 0.8 * 122 = 116$$

- A smaller α makes the forecast more **stable**
- A larger α makes the forecast more **responsive**

Period	Orders	Exponential Smoothing($\alpha = 0.2$)	Exponential Smoothing($\alpha = 0.5$)
1	122		
2	91		
3	100	116	107
4	77	113	104
5	115	106	91
6	58	108	103
7	75	98	81
8	128	93	78
9	111	100	103
10	88	102	107
11		99	98

Utrecht University

65

Forecast Accuracy

- Tests of forecast accuracy are based on the difference between the forecast of the variables' values at time *t* and the actual value at the same time point *t*
- The closer the two to each other $\Longrightarrow$ the smaller the forecast error, i.e. better forecast

Utrecht University

66

Forecast Accuracy – Mean Squared Error (MSE)

- The MSE statistic is defined as:

$$MSE = \frac{\sum_{t=T_1}^{T}(y_t - \hat{y}_t)^2}{T - T_1 + 1}$$

- T is the total number of samples in the time series
- T_1 the index of the first value to be forecast
- $\hat{y}_t$ is the predicted value at time t
- y_t is the actual value at time t
- Another popular measure: Root Mean Squared Error (RMSE) = $\sqrt{MSE}$

Utrecht University

67

Forecast Accuracy – More Measures

- The Mean Absolute Error (MAE) :

$$MAE = \frac{\sum_{t=T_1}^{T}|(y_t - \hat{y}_t)|}{T - T_1 + 1}$$

  - It is also known as Mean Absolute Deviation (MAD)

- Tracking Signal (TS)

$$TS = \frac{\sum_{t=T_1}^{T}(y_t - \hat{y}_t)}{MAE}$$

Utrecht University

68

- Summarize what you learned today in 2-minutes

Utrecht University

69