Predictive Modeling

Time Series Analysis

Mirko Birbaumer

HSLU T&A

Example Time Series : PAN AM

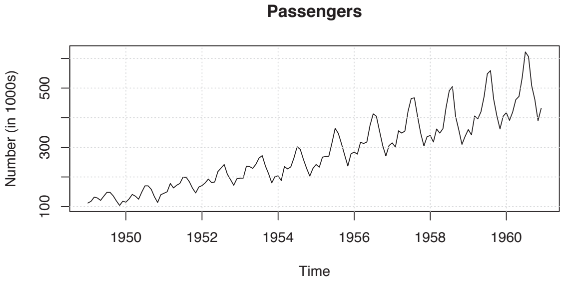
- The following table contains the number of airline passenger bookings (in thousands) per month of the airline *PanAm* (1927-1991) from 1949 to 1960

```
##      Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
## 1949 112 118 132 129 121 135 148 148 136 119 104 118
## 1950 115 126 141 135 125 149 170 170 158 133 114 140
## 1951 145 150 178 163 172 178 199 199 184 162 146 166
## 1952 171 180 193 181 183 218 230 242 209 191 172 194
## 1953 196 196 236 235 229 243 264 272 237 211 180 201
## 1954 204 188 235 227 234 264 302 293 259 229 203 229
## 1955 242 233 267 269 270 315 364 347 312 274 237 278
## 1956 284 277 317 313 318 374 413 405 355 306 271 306
## 1957 315 301 356 348 355 422 465 467 404 347 305 336
## 1958 340 318 362 348 363 435 491 505 404 359 310 337
## 1959 360 342 406 396 420 472 548 559 463 407 362 405
## 1960 417 391 419 461 472 535 622 606 508 461 390 432
```

- Is there a better way of representing this data set?

Example : PAN AM

- Presentation of data as table not convenient $\rightarrow$ **visualize data**



- The following patterns become visible:
  - global *increase* of flight bookings over time $\rightarrow$ **trend**
  - more flight bookings during summer months $\rightarrow$ **seasonality**
  - observations are *not* independent $\rightarrow$ **serial correlation**

Examples Time Series

- `PAN AM` data was originally used to forecast future flight bookings in order to plan aircraft and crew demand

- Many real life measuring and data recording processes result in data sets that are **serially correlated**. Examples of such situations are:
  - *Machine monitoring:* Temperature, pressure, acoustic emissions, vibrations, etc. are measured at various locations in/at/around an operating engine (motor, generator, compressor, etc.)
  - *Stock:* Stock prices and exchange rates are recorded at the end of a trading day
  - *Environmental observations:* Temperature, humidity, pollen concentration, pollution, precipitation are recorded at a specific weather station
  - *Federal statistics:* Population census, income, accidents, ...

- This kind of data is called **time series data**

Time Series Analysis

- There are several goals that one wants to achieve with time series analysis :

1. **Descriptive Analysis**:
   By means of summary statistics and visualizations, the basic properties of a time series are understood

2. **Modeling and Interpretation**:
   By modeling the underlying process that governs the observed time series, a deeper understanding can be gained. In particular, tests and confidence intervalls/bands can be constructed from the model. The sequential dependency of the time series is also often quantified.

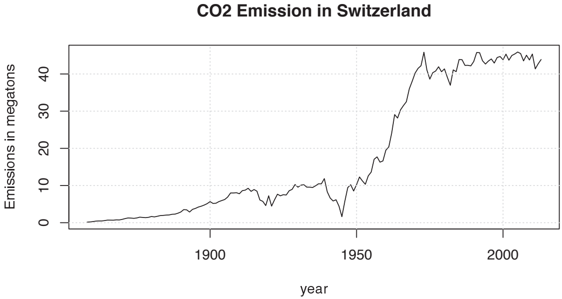
Time Series Analysis

4. **Decomposition**:
   - **Seasonality**, in particular a periodic pattern in the data
   - **Trend**, gradually changing average of the series which is directly correlated with the time axis
4. **Prediction**: By means of the model, future values of the time series can be predicted. Prediction for time series is often alternatively termed **forecasting**
5. **Regression**:
   One often tries to explain a time series (*response*) by several other time series (*predictors*). This idea is wide-spread in industry, where the goal is to replace in a multi-sensor setup a particular (expensive or hard to install) sensor by a model that predicts its values from the other sensor values. This is called *virtual sensoring* or *soft sensoring*.

Example: Kyoto Protocol

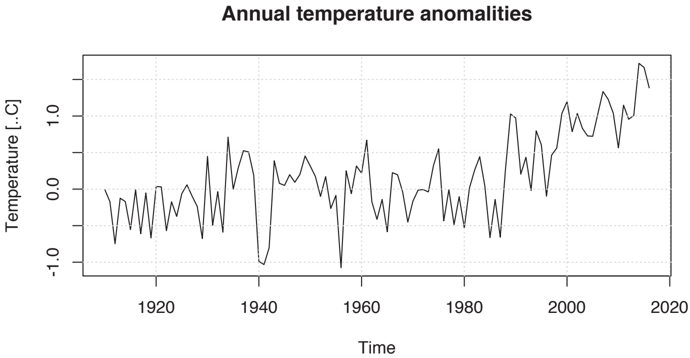
- **Kyoto Protocol**: amendment to the United Nations Framework Convention on Climate Change
- It opened for signature in December 1997 and came into force on February 16, 2005
- The arguments for reducing greenhouse gas emissions rely on a combination of science, economics, and time series analysis
- Decisions made in the next few years will affect the future of the planet

Example: Kyoto Protocol



- Figure shows the yearly emission of the greenhouse gas CO_2 in Switzerland from 1858 to 2013
- No seasonal effect (due to the yearly averaging)
- Peculiar trend is observable: The emissions increased heavily in the post WW2 era between the 1950s and 1970s

- Global warming:



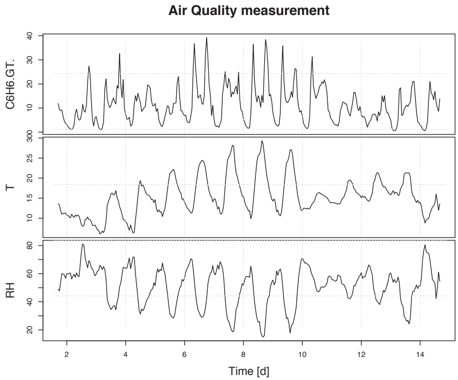
- Annual average temperature anomalities with respect to the average between 1910 and 2000 in Europe are shown
- Upward trend starting in the 1980s
- Previous figure (CO_2): a correlation between CO_2 emission and global warming is not deniable : physical theories support causal relation

Example: Air Quality

- `AirQuality` data set contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device
- The device was located on the field in a significantly polluted area, at road level, within an Italian city
- Data were recorded from March 2004 to February 2005
- All in all, 13 variables are measured by the device

Example: Air Quality

- Data set `AirQuality` :



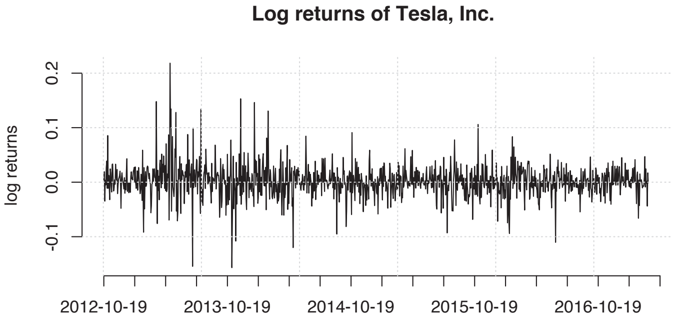
- Concentration of benzene (C_6H_6), the air temperature in C° and the relative humidity (in %) are plotted over a period of 2 weeks

Example: Stock Index of Tesla

- Typical instance of time series are trading indices, and exchange rates in economics.
- Stock indices are often analyzed and subject to prediction attempts
- Trends in stock indices, however, are nearly impossible to forecast
- Here: stock index of Tesla

Example : Stock Index of Tesla

- Daily closing of 1112 consecutive trading days starting at 19.12.2012
- The Tesla stock index : impressive increase between March and June 2013
- Breakdown around February 2016 with subsequent sharp increase
- Presentation of the Model 3 in April 2016



- Instead stock indices, **log-returns** are displayed, i.e. day-to-day changes of the logarithm of the index, are analyzed and modeled
- log-returns : approximation of relative change with respect to previous trading day
- No trend anymore $\quad \rightarrow \quad$ this data is uncorrelated.
- Making predictions for log-returns based on historical data is fruitless endaveour (however, change of variance may be predicted)

Time Series with `R` and `Python`

Please see examples `2.1`, `2.2`, `2.3`, and `2.4` of the `Introduction to Time Series` chapter

Basic transformation, Visualization and Decomposition of Time Series

- Analysis of time series begins with the description, transformation and visualization of data $\rightarrow$ No modeling
- Does not give rise to proper predictions, confidence intervalls etc.
- Important insights and a profound understanding of the data can be achieved by these techniques
- Here:
  - Most important data transformations in the context of time series
  - Toolbox of helpful visualization techniques for exploring time series
  - Decomposition of time series into seasonal, trend and irregular components

Transformation of Time Series Data

- It is desirable or even necessary to transform a time series before the application of models and predictions
- In partiular, many methods assume a:
  - **Gaussian** or at least a **symmetric** distribution of the data
  - **Linear** trend relationship between time and data
  - **Constant variance** across time
- Example: for highly skewed or heteroskedastic data, it is often better not to use the original series

$$\{x_1, x_2, \ldots\}$$

- but a transformed series

$$\{g(x_1), g(x_2), \ldots\}$$

Box-Cox-Transformationen

- A family of transformations, that is well suited for correcting skewness and variance are the *Box-Cox-transformations*:

Box-Cox-Transformations

For a time series $\{x_1, x_2, \dots\}$ with positive values the Box-Cox transformations are defined as

$$g(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(x) & \text{if } \lambda = 0. \end{cases}$$

- **Goal**: choose the parameter λ such that desired properties hold

- See examples 3.1 of the `Introduction to Time Series` chapter

Time-shift Transformation

- Box-Cox family of transforms amounts to a modification of the *values* of the time series
- Sometimes, it is necessary to transform the *time*-axis as well
- Most simple version of time transforms : *shifting*

Time-shift transformation

Let $\{x_1, x_2, \dots\}$ be a time series.

1. The time-shift by a *lag* of $k \in \mathbb{Z}$ is defined by

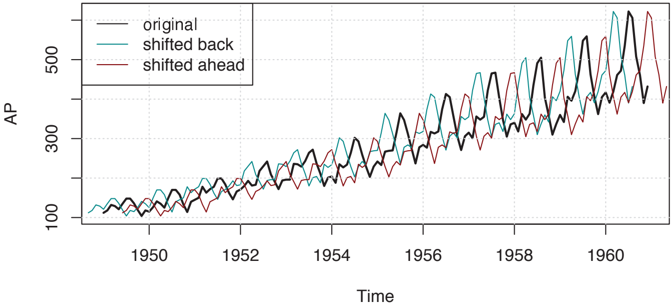
$$g(x_i) = x_{i-k}$$

2. For the particular case where $k = 1$ the time-shift is called *backshift*

$$B(x_i) = x_{i-1}$$

- Applying a time-shift to a time series amounts to go **back** k steps (if $k > 0$) or go **ahead** $-k$ steps (if $k < 0$) in the series

Example Time-shift: `AirPassengers`



- Time-shift for $k = 4$ und $k = -5$
- See example 3.2 of the `Introduction to Time Series` chapter

Log-Returns

- Back-shift operator is applied when *differences* of time series are computed, since

$$x_i - x_{i-1} = x_i - B(x_i)$$

- Differencing is often combined with Box-Cox transformations

- Example: **log-returns** of a (financial) time series are defined as

$$y_i = \log(x_i) - \log(x_{i-1}) = \log\left(\frac{x_i}{x_{i-1}}\right) = \log\left(\frac{x_i - x_{i-1}}{x_{i-1}} + 1\right) \approx \frac{x_i - x_{i-1}}{x_{i-1}}$$

- Last equation: Taylor series expansion of the logarithm:

$$\log(s+1) = s - s^2/2 + \dots$$

Log-Returns

- Log-return time series y_i approximates the relative increase of the time series x_i at each time instance
- This quantity is often studied in financial applications: The original series
$$\{x_1, x_2, \ldots\}$$
may exhibit *seemingly significant patterns*
- but the series $\{y_1, y_2, \ldots\}$ often is rather *random*
- See example `3.3` of the `Introduction to Time Series` chapter for log-return on Tesla stock index
- Tesla log-return looks quite random despite some large fluctuations from time to time
- Analysts try to model the waiting time between these fluctuations

Visualizations of Time Series

- See examples 3.4 to 3.6 of the `Introduction to Time Series` chapter
- Please solve exercises 1

Decomposition of Time Series

- Many time series are dominated by a **trend** and/or **seasonal effects**
- Models in this section are based on these components
- A simple **additive decomposition** model is given by

$$x_k = m_k + s_k + z_k$$

  where

  - k time index
  - x_k observed time series
  - m_k trend
  - s_k seasonal effect
  - z_k error term that is, in general, a sequence of *correlated* random variables with mean zero

Decomposition of Time Series

- `AirPassengers` data : the seasonal effects may increase as the trend increases

- Thus, a **multiplicative** model seems more appropriate:

$$x_k = m_k \cdot s_k + z_k$$

- If the noise is multiplicative as well, i.e., $x_k = m_k \cdot s_k \cdot z_k$, the logarithm of x_k is a linear model again

$$\log(x_k) = \log(m_k) + \log(s_k) + \log(z_k)$$

Moving Average

- **Moving average** : method for estimating the **trend** m_k and the **seasonal effect** s_k by means of the moving average filter:

Moving average filter

Assume that $\{x_1, x_2, \ldots, x_n\}$ is a time series and that $p \in \mathbb{N}$. The *moving average filter* of length p is defined as follows

- If p is odd, then $p = 2l + 1$ and the filtered sequence is defined by

$$g(x_i) = \frac{1}{p}(x_{i-l} + \cdots + x_i + \cdots + x_{i+l})$$

- If p is even, then $p = 2l$ and the filtered sequence is defined by

$$g(x_i) = \frac{1}{p}\left(\frac{1}{2}x_{i-l} + x_{i-l+1} + \cdots + x_i + \cdots + x_{i+l-1} + \frac{1}{2}x_{i+l}\right)$$

The value p is referred to as *window width*.

Moving Average

- **Moving average filter** amounts to replace the i-th value in the time series by the average of the p nearest neighbors of x_i
- If p is **odd**, then the window stretches symmetrically around x_i
- For an **even** p one constructs a window of length $p + 1$ (which is then odd) but counts the end points only by one half.
- Example: If a time series has the frequency 12 (for monthly data), then the trend component of the time series can be estimated by applying the moving average filter with window width $p = 12$
- Since we average at each point exactly over one period, the seasonal effects vanish and the trend component remains. This yields the trend estimator $\widehat{m}_k$

Seasonal Effect

- To estimate the **seasonal additive effect** one computes

$$\widehat{s}_k = x_k - \widehat{m}_k$$

- Now the time series $\widehat{s}_k$ is averaged for each time point in one cycle

- We obtain a single estimate of the effect for each cycle point

Remainder Effect

- We subtract the trend and seasonality estimates and arrive at the estimate for the **remainder term**:

$$\widehat{r}_i = x_i - \widehat{m}_i - \widehat{s}_i$$

- The remainder term should consist of (possibly correlated) random values without structure/periodicity

- See examples 3.7 and 3.8 of the `Introduction to Time Series` chapter

Seasonal Decomposition of Time Series by Loess (STL)

- Although the decomposition method gives promising results for our example data set, it is rarely used in practice, for several reasons:
  - Lacking robustness with respect to outliers in the data.
  - The seasonal component is assumed to be constant over time
- State-of-the-art method for decomposing time series that does not suffer from the above drawbacks is **seasonal decomposition of time series by loess (STL)**
- Please see `3.10` of the `Introduction to Time Series` chapter

STL Decomposition

- STL procedure is iterative : outliers in the estimated remainder terms are detected and their effect mitigated by proper reweighting

- Moving average smoothing is replaced by **loess regression** which gives more flexibility and better results as the moving average.

- Loess regression is a form of **local** (mostly linear) regression which means that the regression line through data $(x_1, y_1), \ldots, (x_n, y_n)$ at a point x is only computed using the observations in **a neighborhood** of x

STL Decomposition

- Seasonal component is not assumed to be constant : the method considers **cycle-subseries**, i.e. the subseries of values at each position of the seasonal cycle
- For example, for a monthly series with frequency 12, the first **cycle-subseries** consists of the January values, the second of the February values etc
- These **sub-cycles** are also smoothed by loess and may change over time