Applied Time Series Analysis
FS 2012 – Week 02

Marcel Dettling

Institute for Data Analysis and Process Design

Zurich University of Applied Sciences

marcel.dettling@zhaw.ch

http://stat.ethz.ch/~dettling

ETH Zürich, February 27, 2012

Applied Time Series Analysis
FS 2012 – Week 02

Stochastic Model for Time Series

Def: A *time series process* is a set $\{X_t, t \in \mathrm{T}\}$ of random variables, where T is the set of times. Each of the random variables $X_t, t \in \mathrm{T}$ has a univariate probability distribution F_t.

- If we exclusively consider time series processes with equidistant time intervals, we can enumerate $\{T = 1, 2, 3, ...\}$
- An observed time series is a realization of $X = (X_1, \ldots, X_n)$, and is denoted with small letters as $x = (x_1, \ldots, x_n)$.
- We have a multivariate distribution, but only 1 observation (i.e. 1 realization from this distribution) is available. In order to perform "statistics", we require some additional structure.

Applied Time Series Analysis
FS 2012 – Week 02

Stationarity

For being able to do statistics with time series, we require that the series “doesn’t change its probabilistic character” over time. This is mathematically formulated by **strict stationarity**.

Def: A time series $\{X_t, t \in \mathrm{T}\}$ is strictly stationary, if the joint distribution of the random vector $(X_t, \ldots, X_{t+k})$ is equal to the one of $(X_s, \ldots, X_{s+k})$ for all combinations of t, s and k.

➔

$X_t \sim F$	all X_t are identically distributed
$E[X_t] = \mu$	all X_t have identical expected value
$Var(X_t) = \sigma^2$	all X_t have identical variance
$Cov(X_t, X_{t+h}) = \gamma_h$	the autocov depends only on the lag h

Applied Time Series Analysis
FS 2012 – Week 02

Stationarity

It is impossible to „prove“ the theoretical concept of stationarity from data. We can only search for evidence in favor or against it.

However, with strict stationarity, even finding evidence only is too difficult. We thus resort to the concept of *weak stationarity*.

Def: A time series $\{X_t, t \in \mathrm{T}\}$ is said to be *weakly stationary*, if

$$E[X_t] = \mu$$
$$Cov(X_t, X_{t+h}) = \gamma_h \quad \text{for all lags } h$$

and thus also: $Var(X_t) = \sigma^2$

Note that weak stationarity is sufficient for „practical purposes“.

Applied Time Series Analysis
FS 2012 – Week 02

Testing Stationarity

- In time series analysis, we need to verify whether the series has arisen from a stationary process or not. Be careful: stationarity is a property of the process, and not of the data.
- Treat stationarity as a hypothesis! We may be able to reject it when the data strongly speak against it. However, we can never prove stationarity with data. At best, it is plausible.
- Formal tests for stationarity do exist (→ see scriptum). We discourage their use due to their low power for detecting general non-stationarity, as well as their complexity.

→ **Use the time series plot for deciding on stationarity!**

Applied Time Series Analysis
FS 2012 – Week 02

Evidence for Non-Stationarity

- **Trend**, i.e. non-constant expected value
- **Seasonality**, i.e. deterministic, periodical oscillations
- **Non-constant variance**, i.e. multiplicative error
- **Non-constant dependency structure**

Remark:

Note that some periodical oscillations, as for example in the lynx data, can be stochastic and thus, the underlying process is assumed to be stationary. However, the boundary between the two is fuzzy.

Applied Time Series Analysis
FS 2012 – Week 02

Strategies for Detecting Non-Stationarity

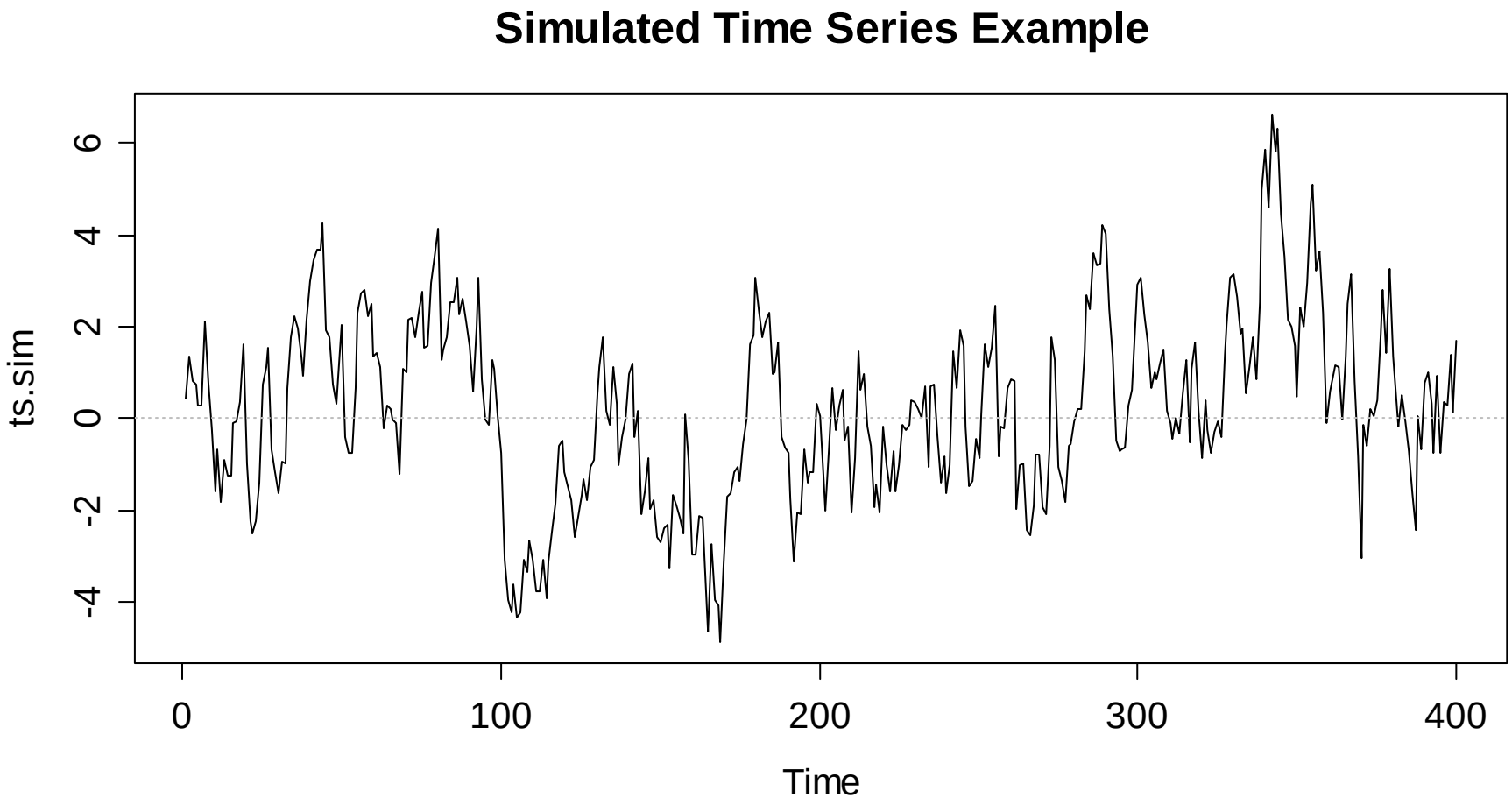
1) **Time series plot**
   - non-constant expected value (trend/seasonal effect)
   - changes in the dependency structure
   - non-constant variance

2) **Correlogram (presented later...)**
   - non-constant expected value (trend/seasonal effect)
   - changes in the dependency structure

A (sometimes) useful trick, especially when working with the correlogram, is to split up the series in two or more parts, and producing plots for each of the pieces separately.

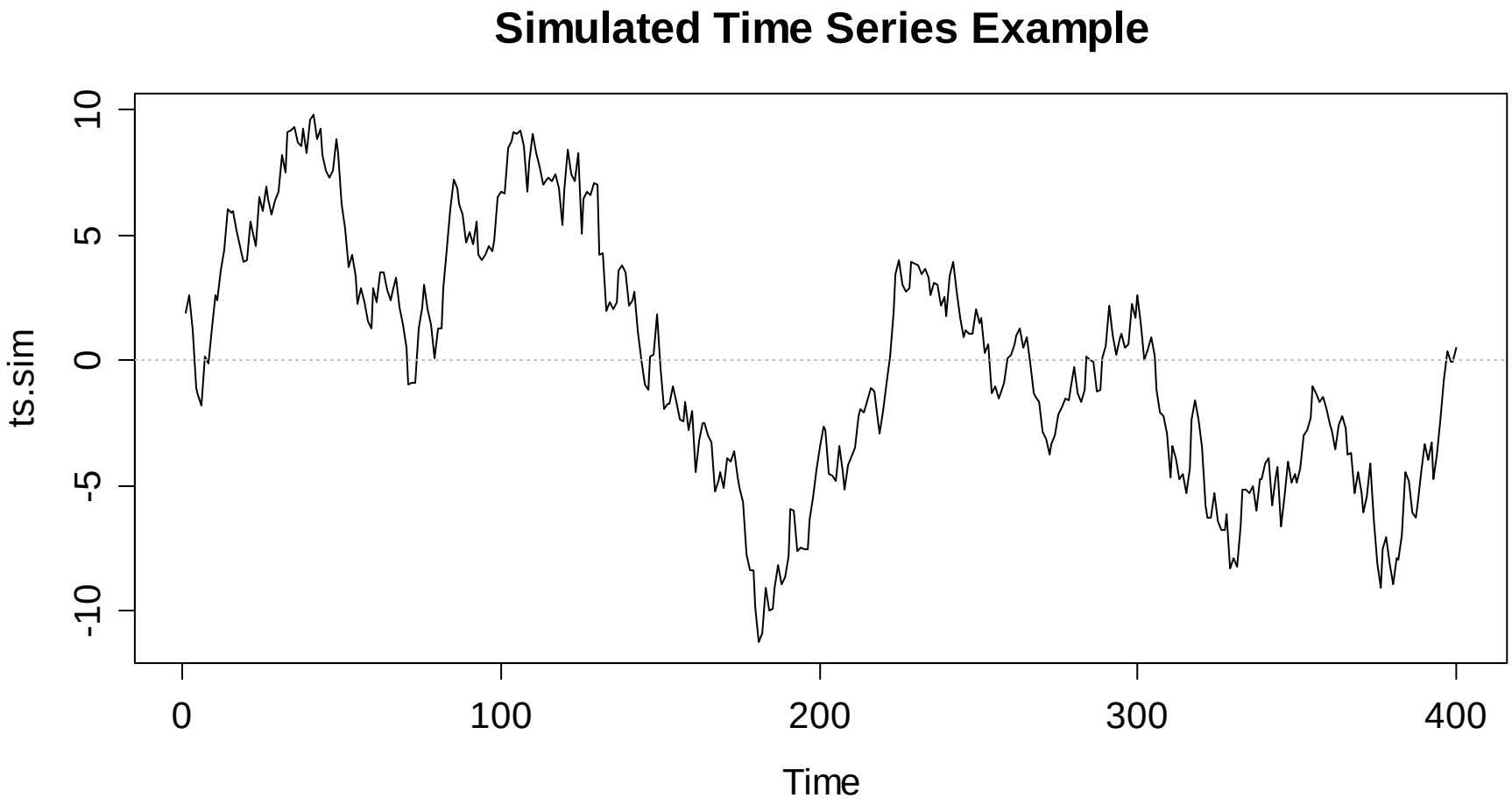
Applied Time Series Analysis
FS 2012 – Week 02

Example: Simulated Time Series 1



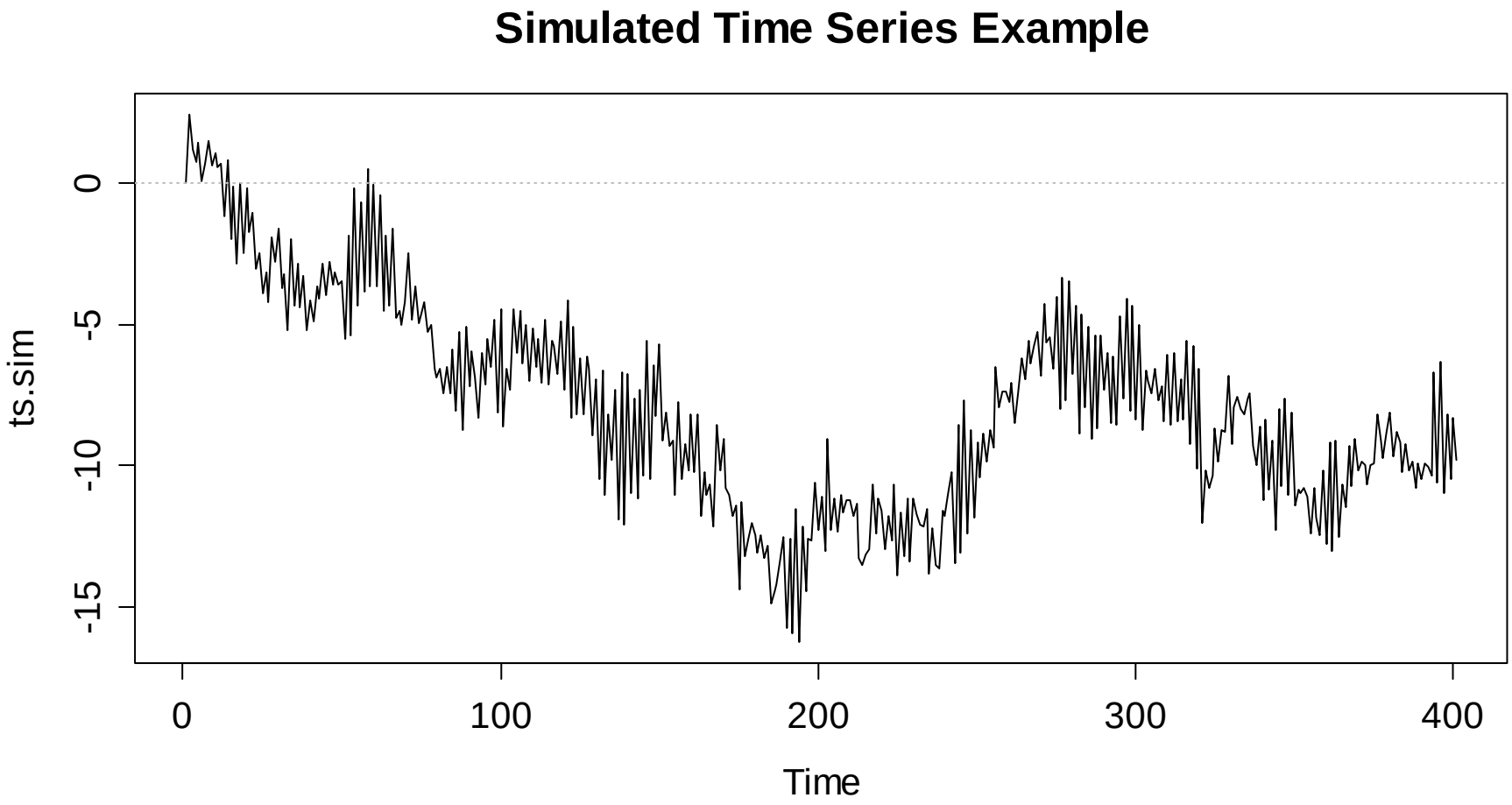
Applied Time Series Analysis
FS 2012 – Week 02

Example: Simulated Time Series 2



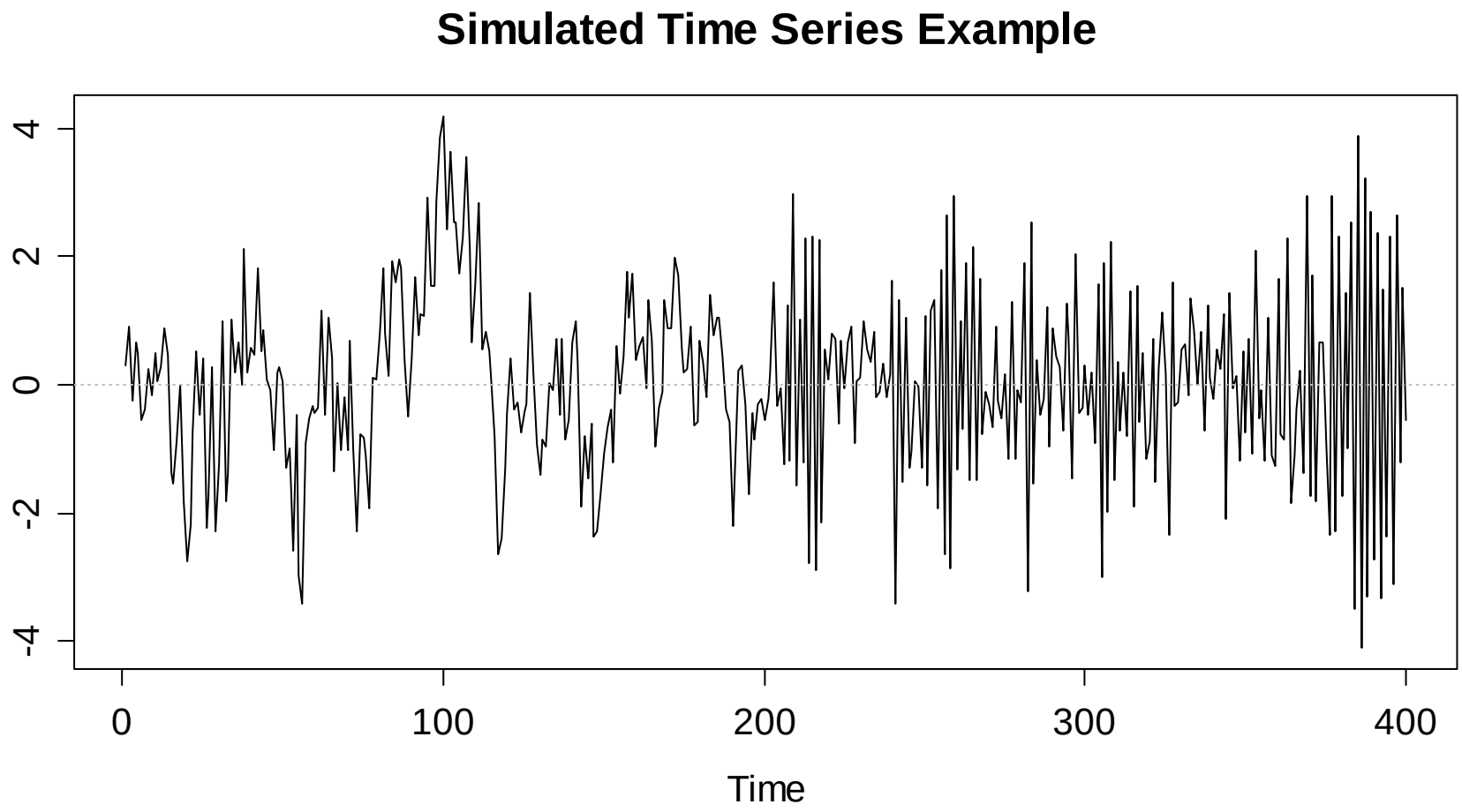
Applied Time Series Analysis
FS 2012 – Week 02

Example: Simulated Time Series 3



Applied Time Series Analysis
FS 2012 – Week 02

Example: Simulated Time Series 4



Applied Time Series Analysis
FS 2012 – Week 02

Time Series in R

- In **R**, there are *objects*, which are organized in a large number of *classes*. These classes e.g. include *vectors*, *data frames*, *model output*, *functions*, and many more. Not surprisingly, there are also *several classes for time series*.
- We focus on **ts**, the basic class for regularly spaced time series in **R**. This class is comparably simple, as it can only represent time series with *fixed interval records*, and *only uses numeric time stamps*, i.e. enumerates the index set.
- For defining a **ts** object, we have to supply the *data*, but also the *starting time* (as argument start), and the *frequency* of measurements as argument frequency.

Applied Time Series Analysis
FS 2012 – Week 02

Time Series in R: Example

Data: number of days per year with traffic holdups in front of the Gotthard road tunnel north entrance in Switzerland.

2004	2005	2006	2007	2008	2009	2010
88	76	112	109	91	98	139

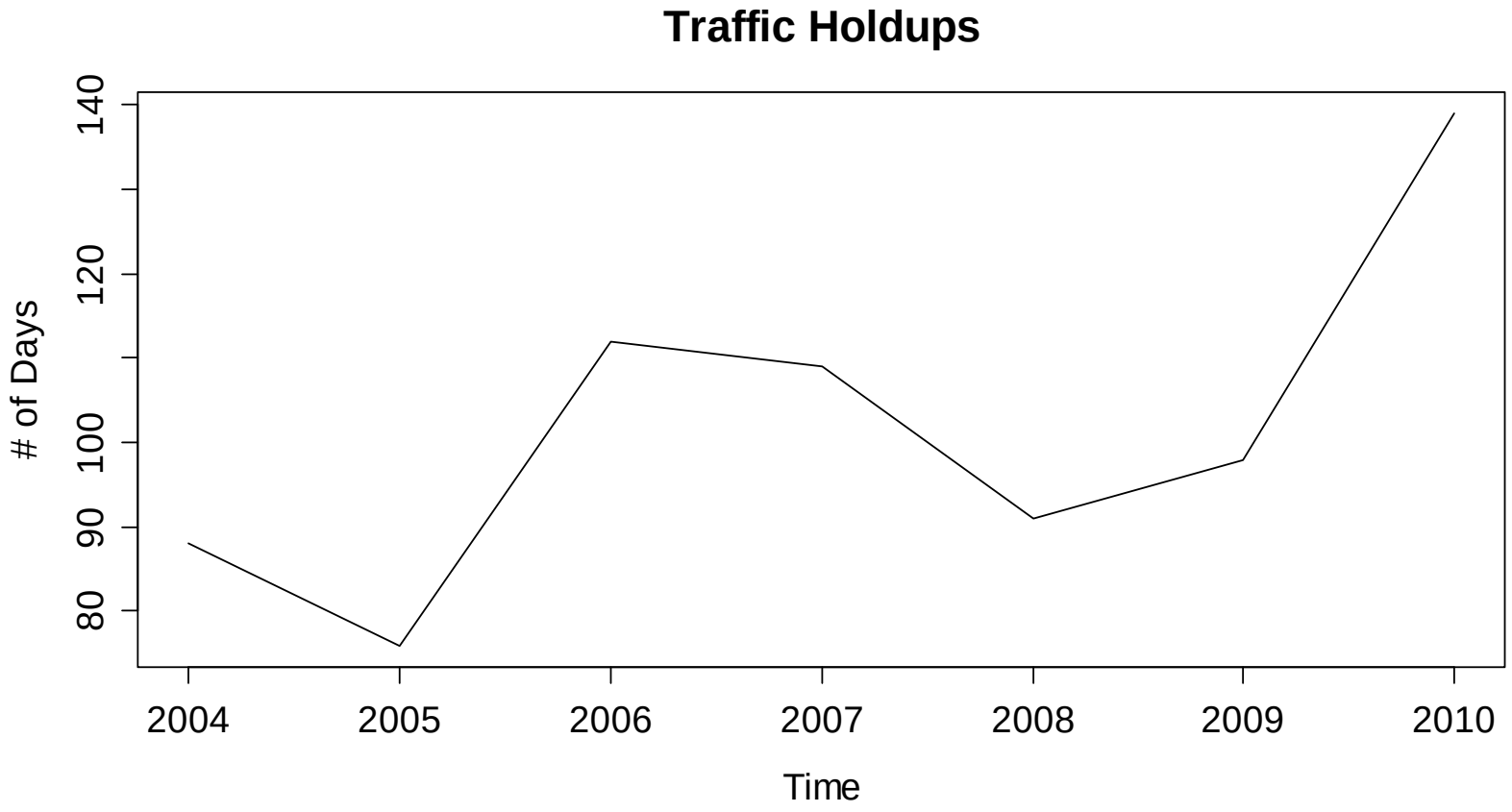
```
> rawdat <- c(88, 76, 112, 109, 91, 98, 139)
> ts.dat <- ts(rawdat, start=2004, freq=1)

> ts.dat
Time Series: Start = 2004
End = 2010; Frequency = 1
[1]  88  76 112 109  91  98 139
```

Applied Time Series Analysis
FS 2012 – Week 02

Time Series in R: Example

```
> plot(ts.dat, ylab="# of Days", main="Traffic Holdups")
```



Applied Time Series Analysis
FS 2012 – Week 02

Addendum: Daily Data and Leap Years

Example from Exercises:

Rainfall Data, 8 years with daily data from 2000-2007. While 2001-2003 and 2005-2007 have 365 days each, years 2000 and 2004 are leap years with 366 days.

- Do never cancel the leap days, and neither introduce missing values for Feb 29 in non-leap years.
- Is this a (deterministically) periodic series? Using the Gregorian calendar, we can say the time unit is 4 years, and the frequency is $366 + (3 \cdot 365) = 1461$.
- Physically, we can say that the frequency equals 365.25.

Applied Time Series Analysis
FS 2012 – Week 02

Further Topics in R

The scriptum discusses some further topics which are of interest when doing time series analysis in R:

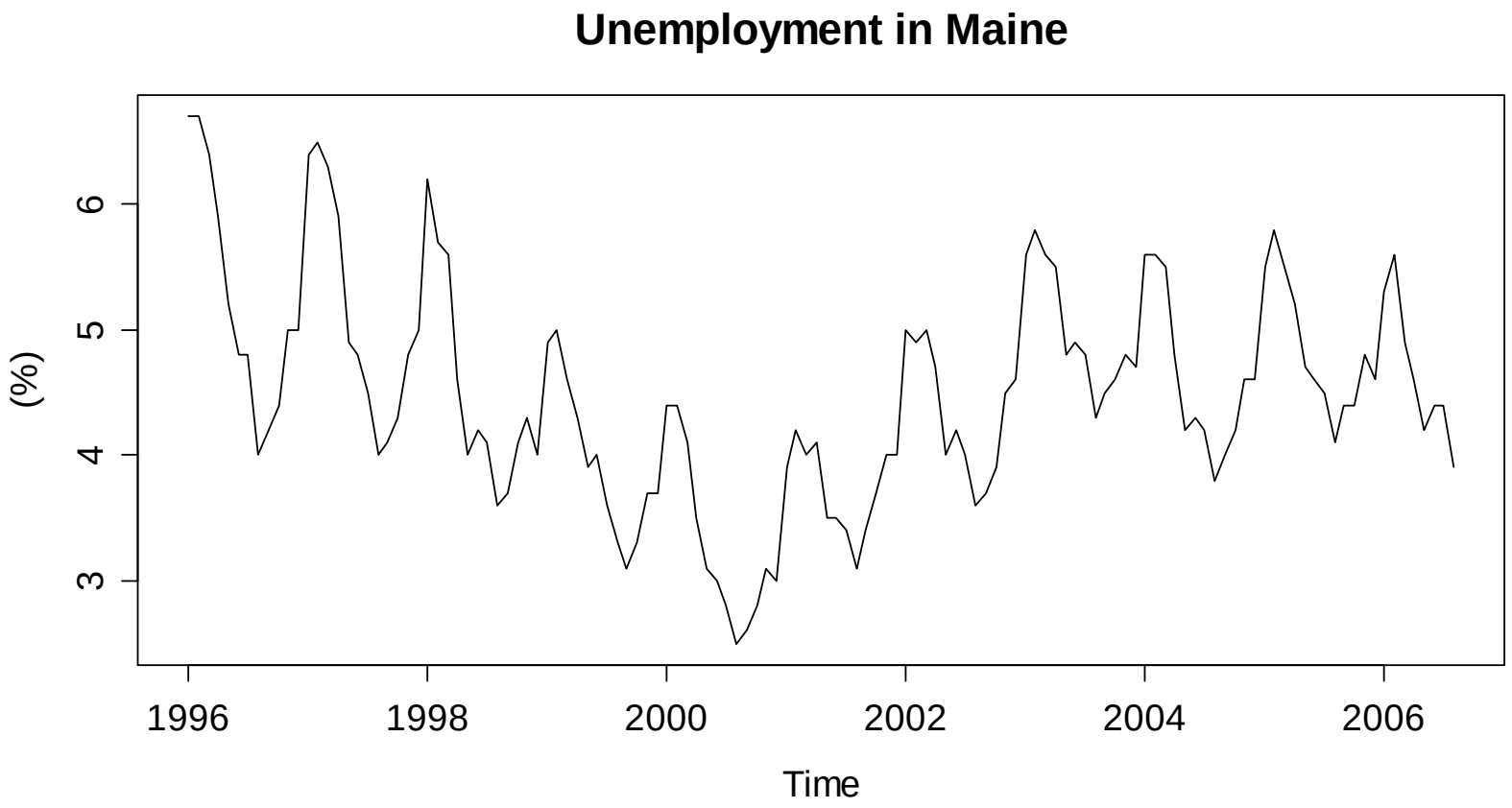
- *Handling of dates and times in R*
- *Reading/Importing data into R*

→ **Please thoroughly read and study these chapters. Examples will be shown/discussed in the exercises.**

Applied Time Series Analysis
FS 2012 – Week 02

Visualization: Time Series Plot

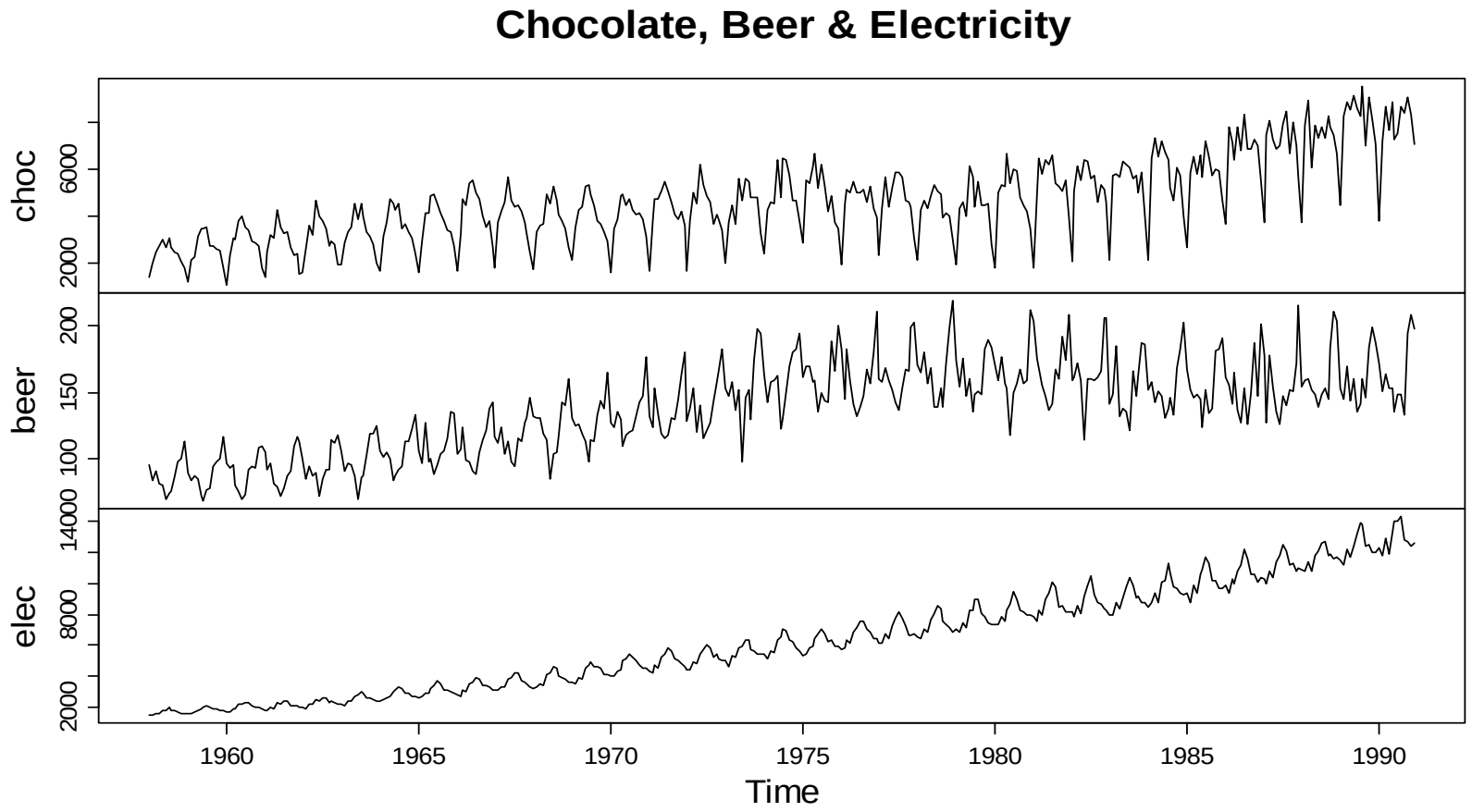
```
> plot(tsd, ylab="(%)", main="Unemployment in Maine")
```



Applied Time Series Analysis
FS 2012 – Week 02

Multiple Time Series Plots

```
> plot(tsd, main="Chocolate, Beer & Electricity")
```

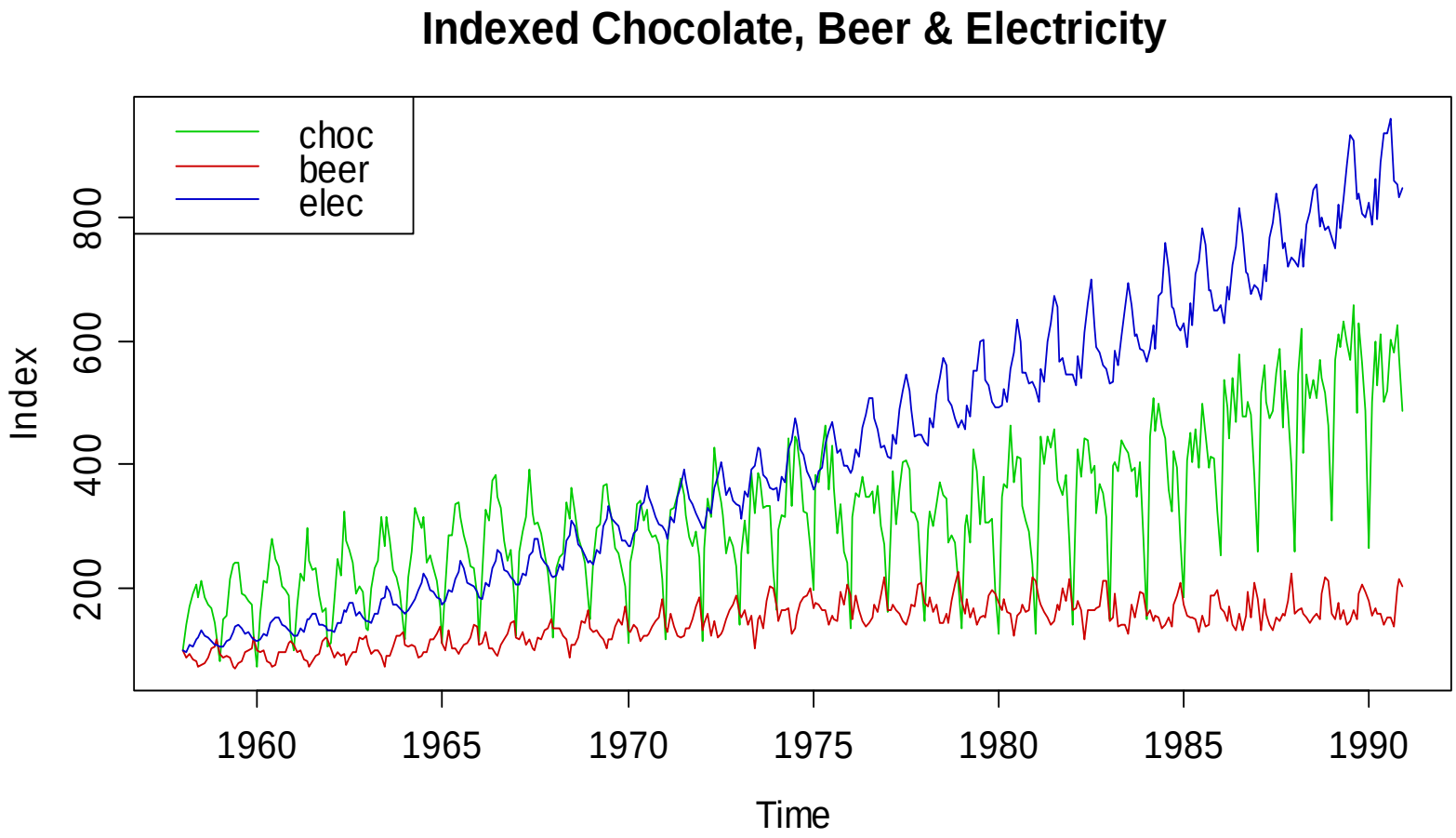


Only One or Multiple Frames?

- Due to different scale/units it is often impossible to directly plot multiple time series in one single frame. Also, multiple frames are convenient for visualizing the series.
- If the relative development of multiple series is of interest, then we can (manually) index the series and (manually) plot them into one single frame.
- This clearly shows the magnitudes for trend and seasonality. However, the original units are lost.
- For details on how indexing is done, see the scriptum.

Applied Time Series Analysis
FS 2012 – Week 02

Multiple Time Series Plots



Applied Time Series Analysis
FS 2012 – Week 02

Descriptive Decomposition

It is convenient to describe non-stationary time series with a simple decomposition model

$$X_t = m_t + s_t + E_t$$

= **trend + seasonal effect + stationary remainder**

The modelling can be done with:

1) ***taking differences with appropriate lag (=differencing)***
2) ***smoothing approaches (= filtering)***
3) ***parametric models (= curve fitting)***

Applied Time Series Analysis
FS 2012 – Week 02

Differencing: Theory

In the absence of a seasonal effect, a piecewise linear trend of a non-stationary time series can by removed by taking differences of first order at lag 1:

$$Y_t = X_t - X_{t-1}$$

The new time series Y_t is then going to be stationary, but has some new, strong and artificial dependencies.

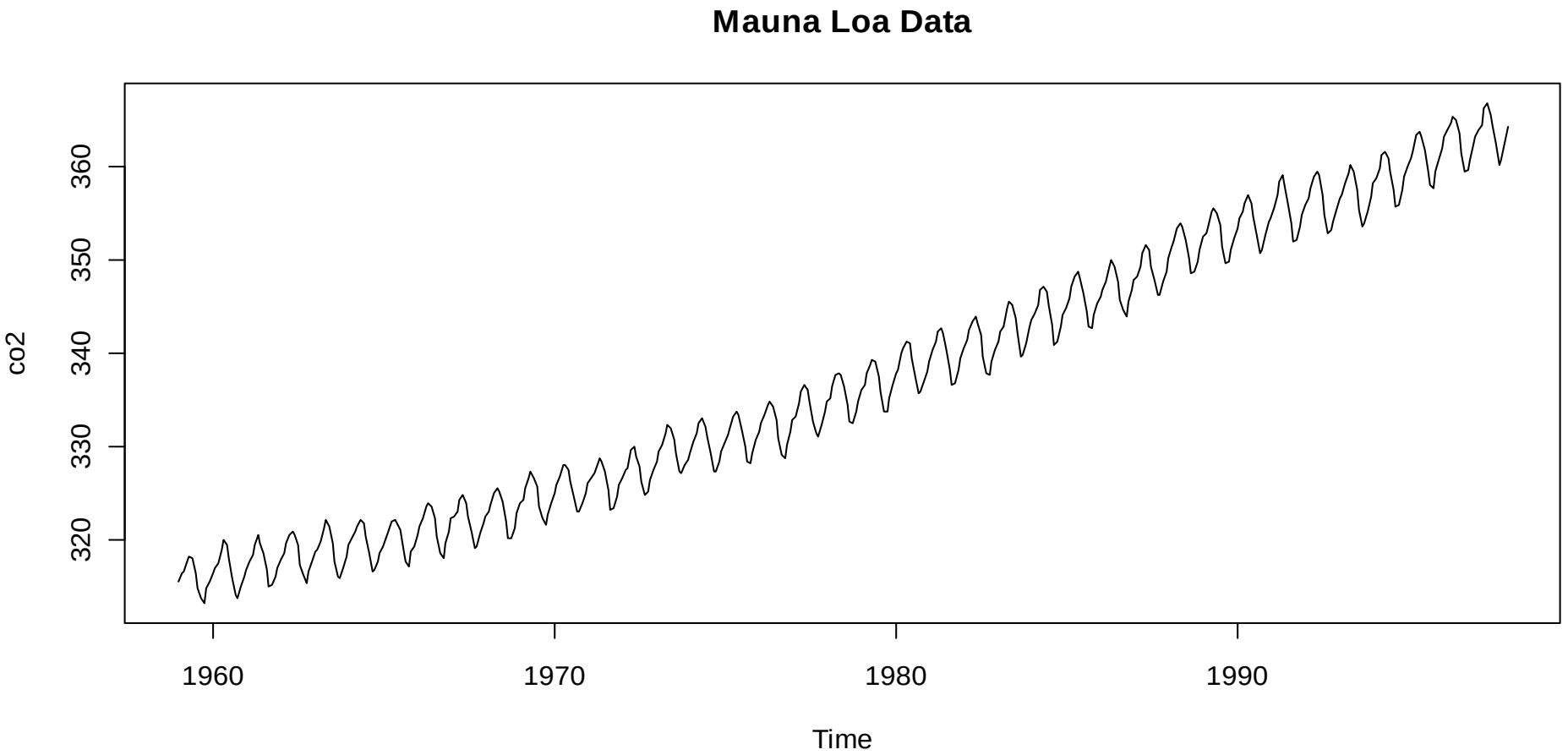
If there is a seasonal effect, we have to take first order differences at the lag p of the period, which removes both trend and season:

$$Y_t = X_t - X_{t-p}$$

Applied Time Series Analysis
FS 2012 – Week 02

Differencing: Example

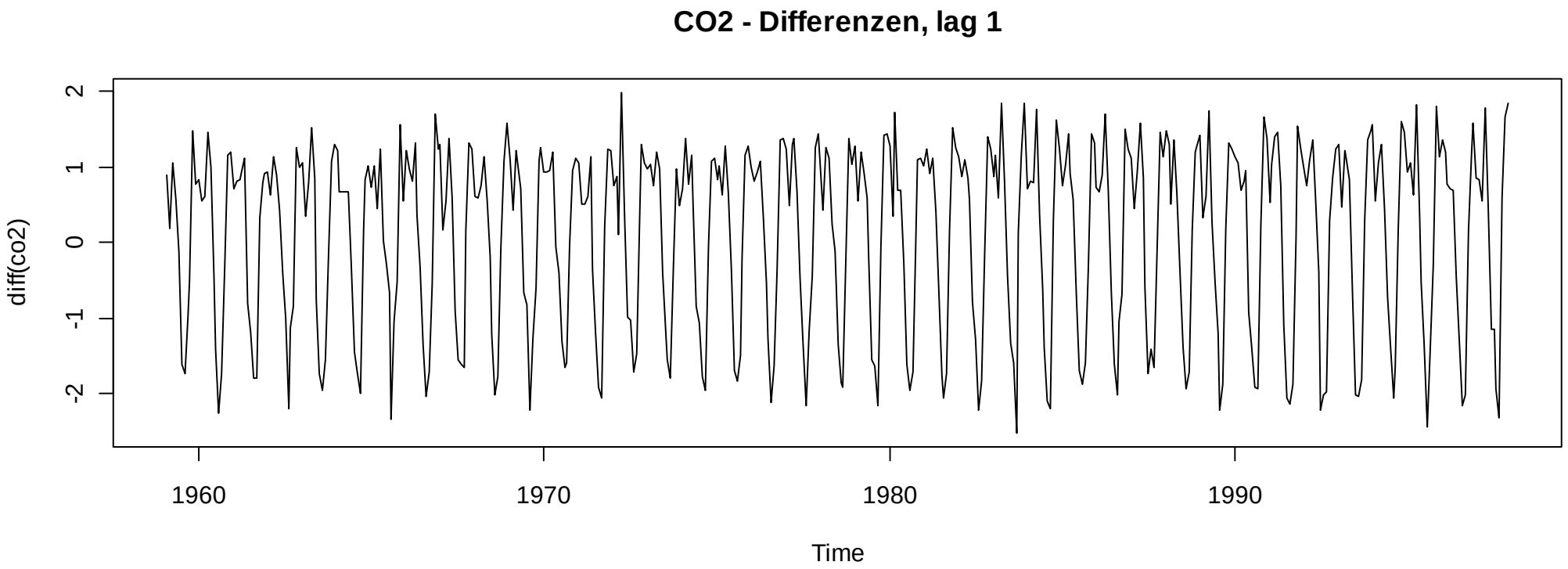
Mauna Loa Data: original series, containing trend and season



Applied Time Series Analysis
FS 2012 – Week 02

Differencing: Example

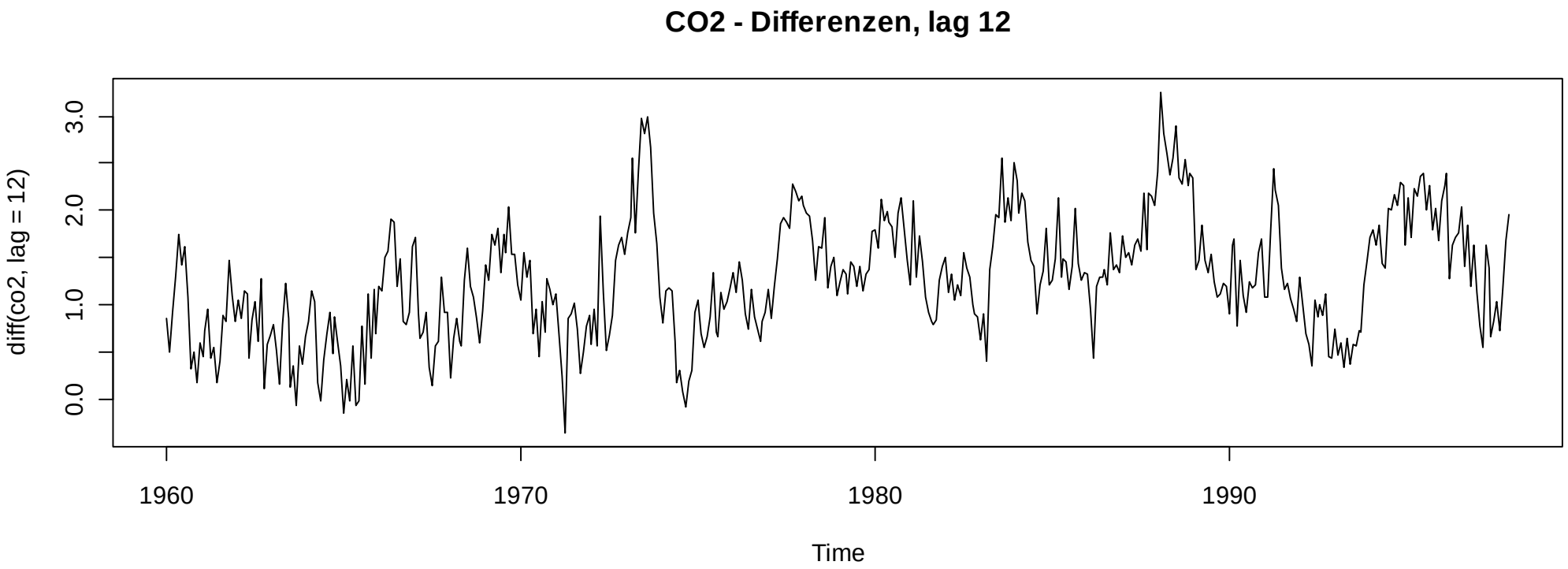
Mauna Loa Data: first order differences with lag 1



Applied Time Series Analysis
FS 2012 – Week 02

Differencing: Example

Mauna Loa Data: first order differences with lag 12



Differencing: Remarks

Some advantages and disadvantages:

+ trend and seasonal effect can be removed
+ procedure is very quick and very simple to implement

- $\hat{m}_t$ and $\hat{s}_t$ are not known, and cannot be visualised
- resulting time series will be shorter than the original
- differencing leads to strong artificial dependencies
- extrapolation of $\hat{m}_t$, $\hat{s}_t$ is not possible

Applied Time Series Analysis
FS 2012 – Week 02

Smoothing, Filtering: Part 1

In the absence of a seasonal effect, the trend of a non-stationary time series can be determined by applying any **additive, linear filter**. We obtain a new time series $\hat{m}_t$, representing the trend:

$$\hat{m}_t = \sum_{i=-p}^{q} a_i X_{t+i}$$

- the window, defined by p and q, can or can‘t be symmetric
- the weights, given by a_i, can or can‘t be uniformly distributed
- other smoothing procedures can be applied, too.

Applied Time Series Analysis
FS 2012 – Week 02

Smoothing, Filtering: Part 2

In the presence a seasonal effect, smoothing approaches are still valid for estimating the trend. We have to make sure that the sum is taken over an entire season, i.e. for monthly data:

$$\hat{m}_t = \frac{1}{12}\left(\frac{1}{2}X_{t-6} + X_{t-5} + \ldots + X_{t+5} + \frac{1}{2}X_{t+6}\right) \quad for \; t = 7,...,n-6$$

An estimate of the seasonal effect s_t at time t can be obtained by:

$$\hat{s}_t = x_t - \hat{m}_t$$

By averaging these estimates of the effects for each month, we obtain a single estimate of the effect for each month.

Applied Time Series Analysis
FS 2012 – Week 02

Smoothing, Filtering: Part 3

- The smoothing approach is based on estimating the trend first, and then the seasonality.
- The generalization to other periods than $p = 12$, i.e. monthly data is straighforward. Just choose a symmetric window and use uniformly distributed coefficients that sum up to 1.
- The sum over all seasonal effects will be close to zero. Usually, it is centered to be exactly there.
- This procedure is implemented in R with function: **`decompose()`**

Applied Time Series Analysis
FS 2012 – Week 02

Smoothing, Filtering: Remarks

Some advantages and disadvantages:

+ trend and seasonal effect can be estimated
+ $\hat{m}_t$ and $\hat{s}_t$ are explicitly known, can be visualised
+ procedure is transparent, and simple to implement

- resulting time series will be shorter than the original
- averaging leads to strong artificial dependencies
- extrapolation of $\hat{m}_t$, $\hat{s}_t$ are not entirely obvious

Applied Time Series Analysis
FS 2012 – Week 02

Smoothing, Filtering: STL-Decomposition

The Seasonal-Trend Decomposition Procedure by Loess

- is an iterative, non-parametric smoothing algorithm
- yields a simultaneous estimation of trend and seasonal effect

→ similar to what was presented above, but more robust!

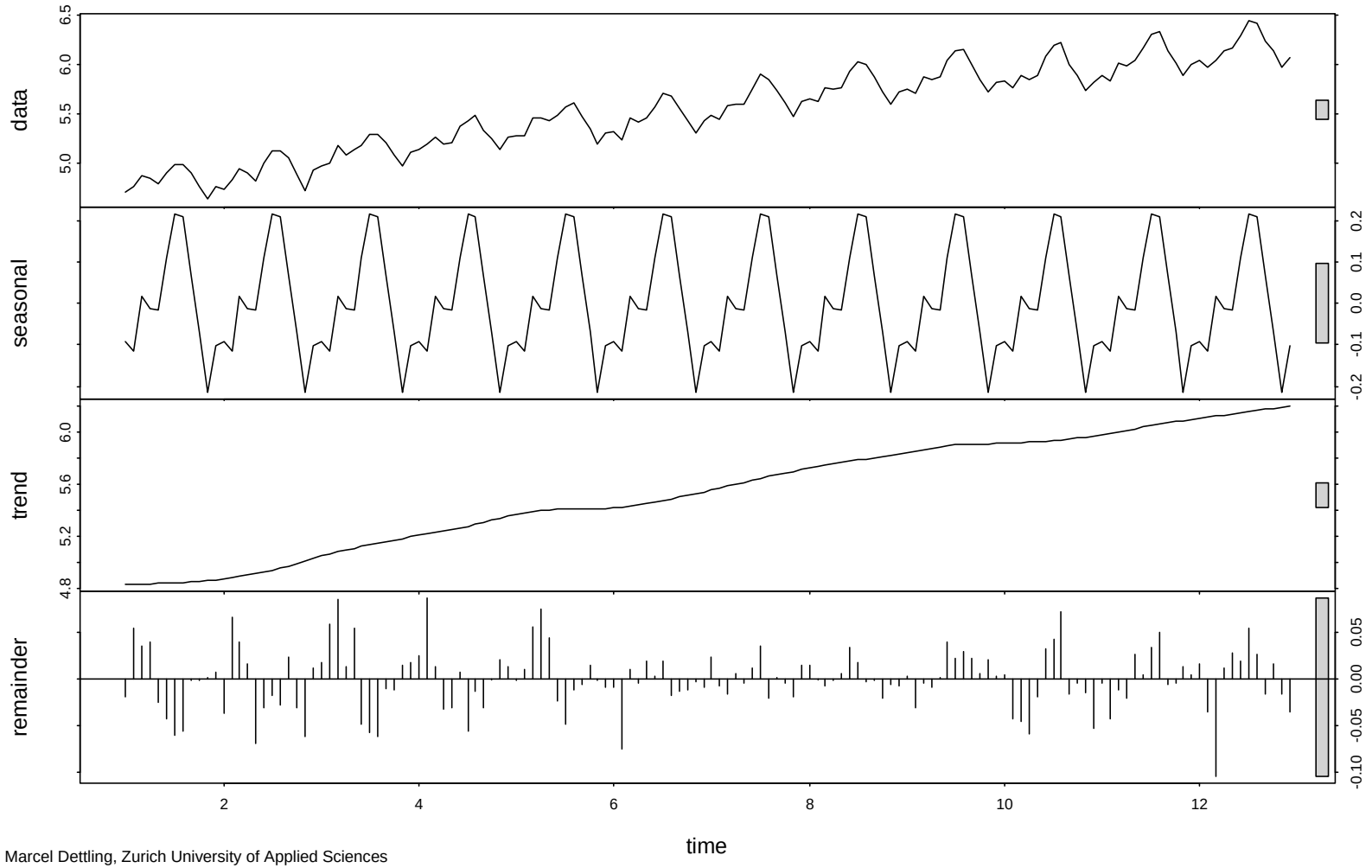
+ very simple to apply
+ very illustrative and quick
+ seasonal effect can be constant or smoothly varying
- model free, extrapolation and forecasting is difficult

→ **Good method for „having a quick look at the data“**

Applied Time Series Analysis
FS 2012 – Week 02

STL-Decomposition: Constant Season

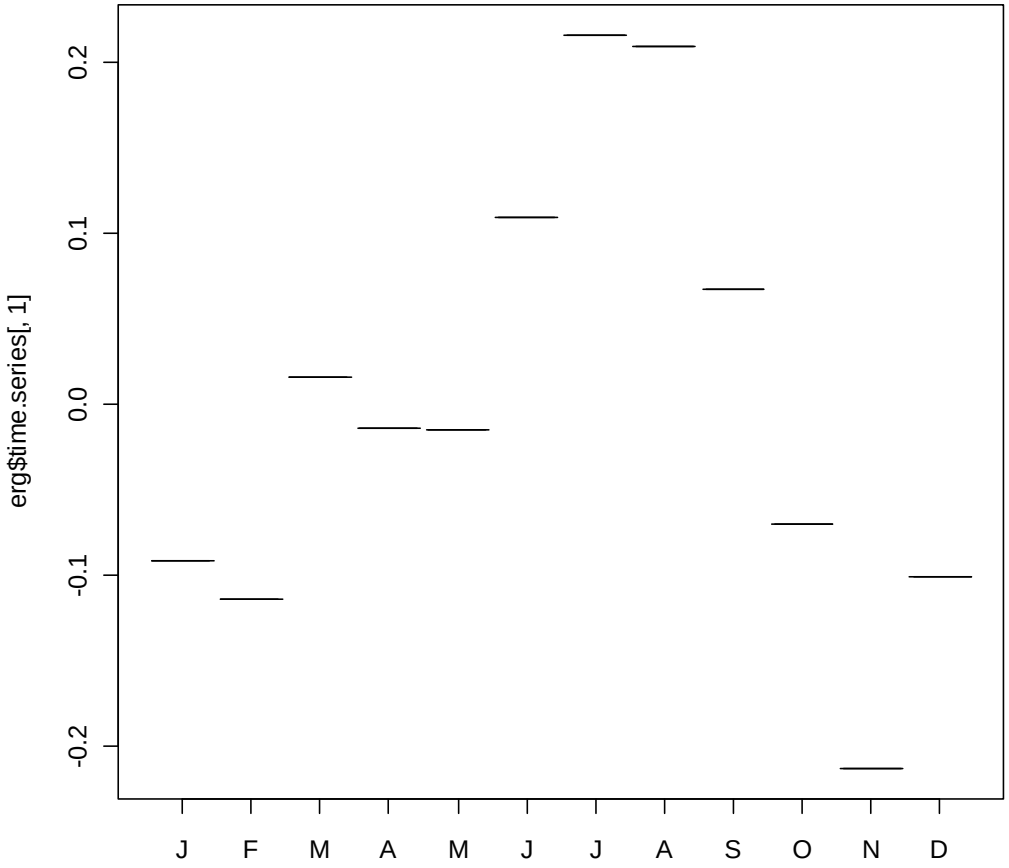
`stl(log(ts(airline,freq=12)),s.window=„periodic“)`



Applied Time Series Analysis
FS 2012 – Week 02

STL-Decomposition: Constant Season

```
stl(log(ts(airline,freq=12)),s.window=„periodic“)
```

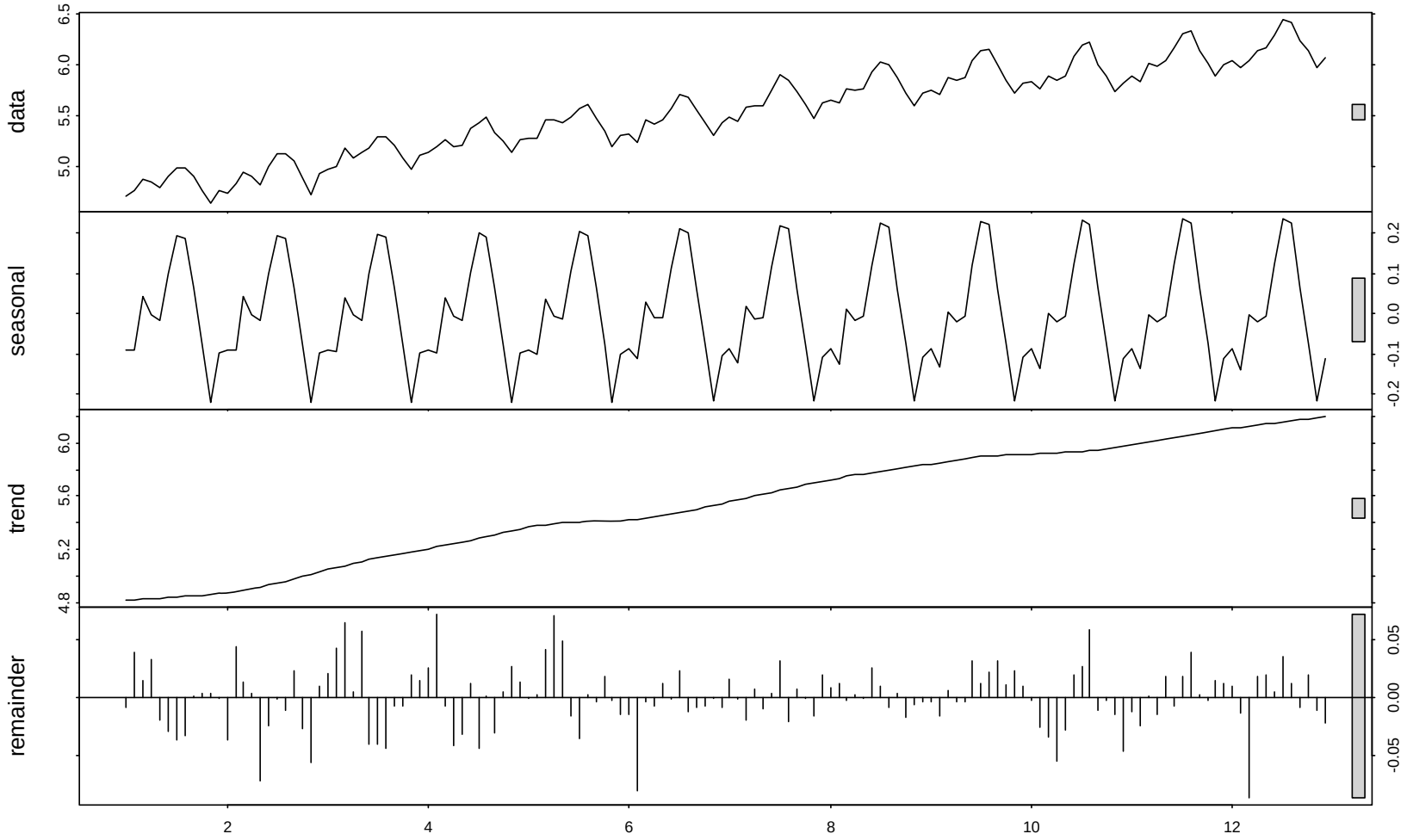


the seasonal effect here is not time varying

Applied Time Series Analysis
FS 2012 – Week 02

STL-Decomposition: Evolving Season

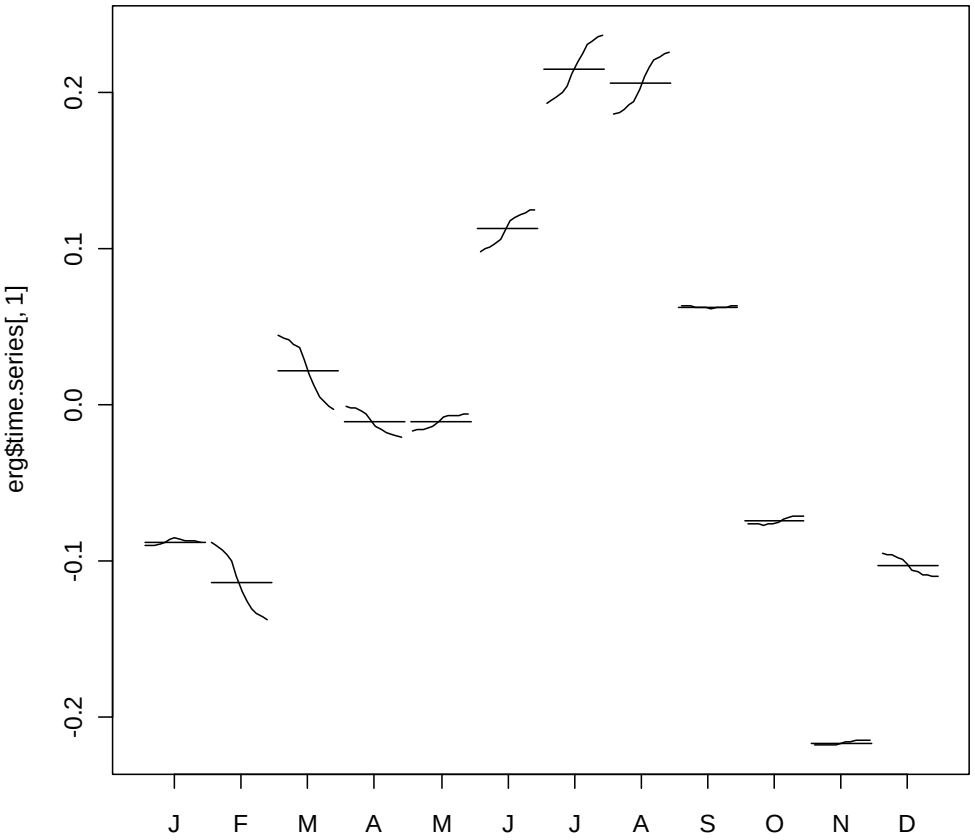
`stl(log(ts(airline,freq=12)),s.window=15)`



Applied Time Series Analysis
FS 2012 – Week 02

STL-Decomposition: Evolving Season

```
stl(log(ts(airline,freq=12)),s.window=15)
```

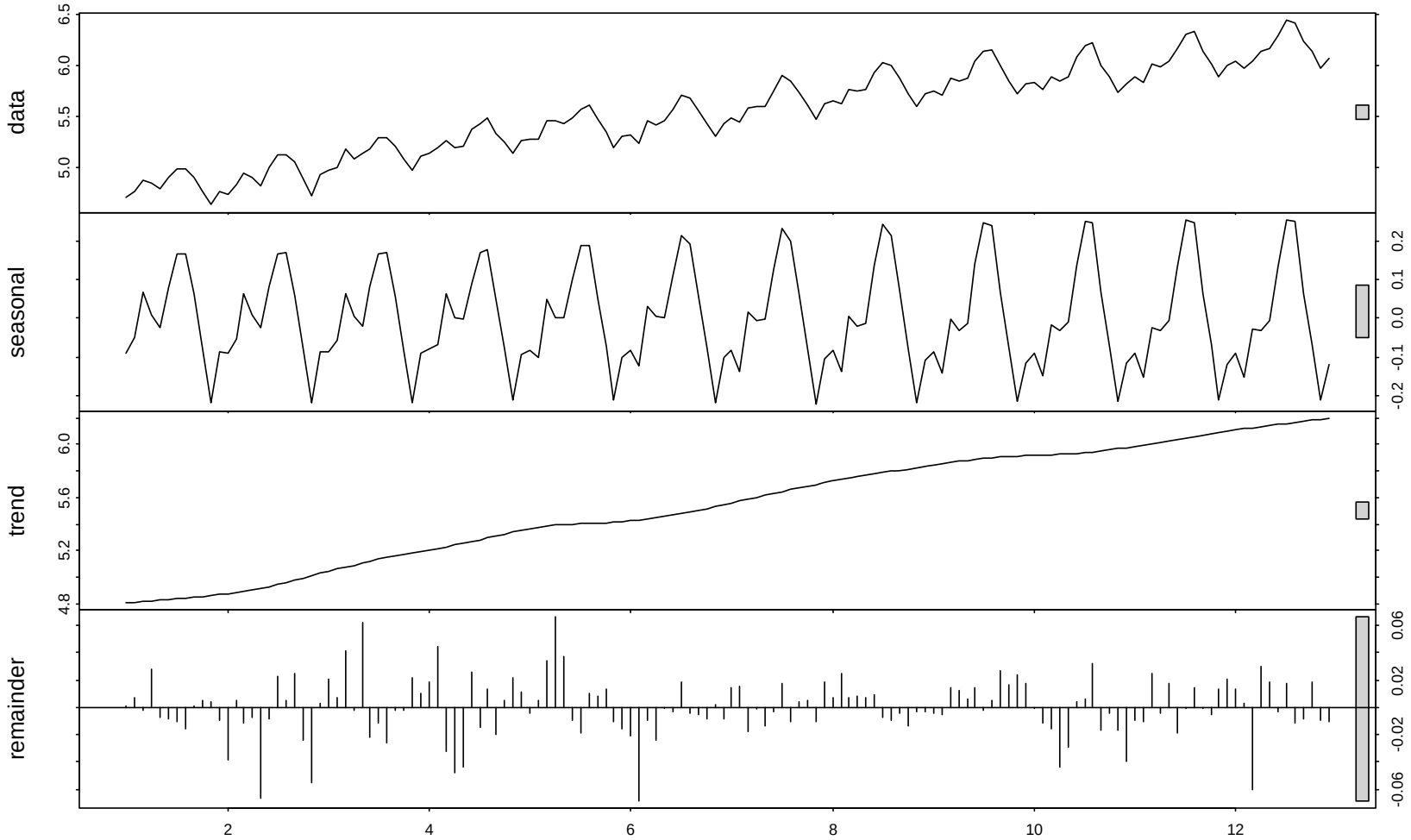


correct amount of smoothing on the time varying seasonal effect

Applied Time Series Analysis
FS 2012 – Week 02

STL-Decomposition: Evolving Season

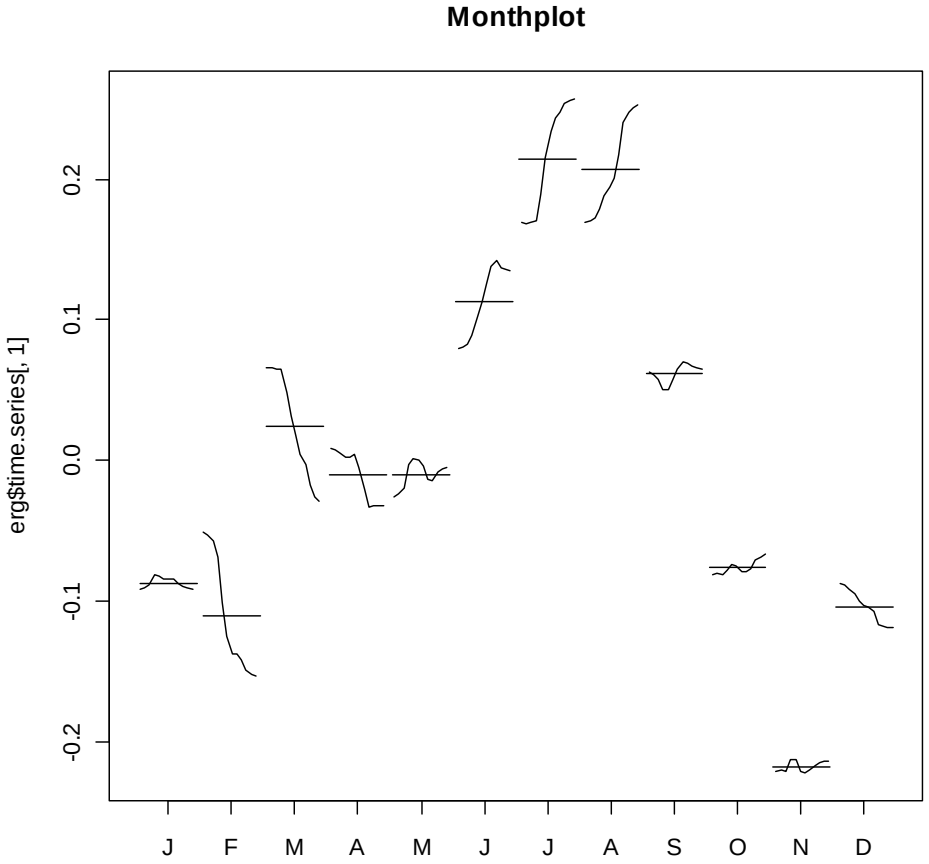
`stl(log(ts(airline,freq=12)),s.window=7)`



Applied Time Series Analysis
FS 2012 – Week 02

STL-Decomposition: Evolving Season

```
stl(log(ts(airline,freq=12)),s.window=7)
```



not enough smoothing on the time varying seasonal effect

Applied Time Series Analysis
FS 2012 – Week 02

Parametric Modelling

When to use?

→ Parametric modelling is often used if we have previous knowledge about the trend following a functional form.

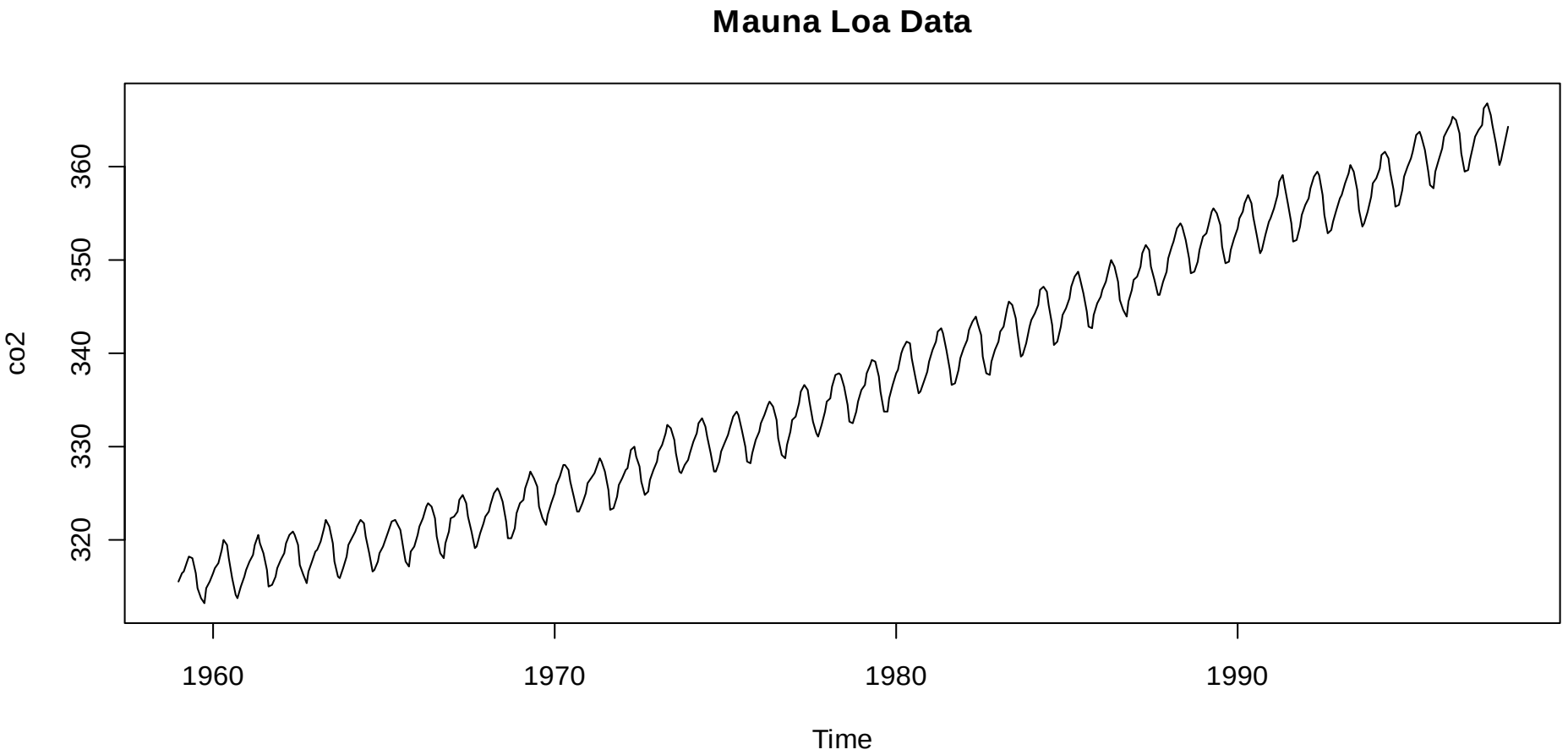
→ If the main goal of the analysis is forecasting, a trend in functional form may allow for easier extrapolation than a trend obtained via smoothing.

→ It can also be useful if we have a specific model in mind and want to infer it. Caution: correlated errors!

Applied Time Series Analysis
FS 2012 – Week 02

Parametric Modeling: Example

Mauna Loa Data: original series, containing trend and season



Applied Time Series Analysis
FS 2012 – Week 02

Parametric Modeling for the Mauna Loa Data

Most often, time series are parametrically decomposed by using regression models. For the trend, polynomial functions are widely used, whereas the seasonal effect is modelled with dummy variables (= a factor).

$$X_t = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot t^2 + \beta_3 \cdot t^3 + \alpha_{i(t)} + E_t$$

where $t \in \{1, 2, ..., 468\}$

$i(t) \in \{1, 2, ..., 12\}$

Remark: choice of the polynomial degree is crucial!

Applied Time Series Analysis
FS 2012 – Week 02

Parametric Modeling: Remarks

Some advantages and disadvantages:

+ trend and seasonal effect can be estimated
+ $\hat{m}_t$ and $\hat{s}_t$ are explicitly known, can be visualised
+ even some inference on trend/season is possible
+ time series keeps the original length

- choice of a/the correct model is necessary/difficult
- residuals are correlated: this is a model violation!
- extrapolation of $\hat{m}_t$, $\hat{s}_t$ are not entirely obvious