UNIVERSITÄT KOBLENZ · LANDAU

Universität Koblenz-Landau Fachbereich Informatik

Klaus G. Troitzsch

Introduction to Time Series Analysis

SICSS Summer School 2001

Author's address: Professor Dr. Klaus G. Troitzsch,Institut für Wirtschafts- und Verwaltungsinformatik,Rheinau 1, D–56075 Koblenz,E-mail: kgt@uni-koblenz.de, WWW: http://www.uni-koblenz.de/~kgt, phone: +49–261–287–2643, fax: +49–261–287–2642.

Summary

The Concept of a Time Series

A time series is defined as a sequence of observations (measurements) ordered by time $\{x_t\}, t \in T$. We restrict ourselves to equidistant time series, i.e. the parameter set is a finite set of equidistant points of time: $T = \{1, 2, 3, \cdots, N\}$.

Time Series Analysis: An Overview

We distinguish two classes of time series analysis approaches:

- one class which represents a time series with a kinetic model (component analysis, classical analysis):

$$x_t = f(t) \tag{1}$$

  the measurements or observations are seen as a function of time;

- one class which represents a time series with a dynamical model ("ARIMA model", "Box-Jenkins procedure"):

$$x_t = f(x_{t-1}, x_{t-2}, x_{t-3}, \cdots) \tag{2}$$

  the measurements or observations are not seen as a function of time, but as a function of their own past (and, perhaps of the past of other measured or observed variables).

The Classical Procedure: Component Model

The classical procedure decomposes the time series function $x_t = f(t)$ into up to four components:

- the trend: a long-term monotonic change of the average level of the time series,
- the trade cycle: a long wave in the time series,
- the seasonal component: a yearly variation in the time series,
- the residual component which represents all the influences on the time series which are not explained by the other three components.

Components

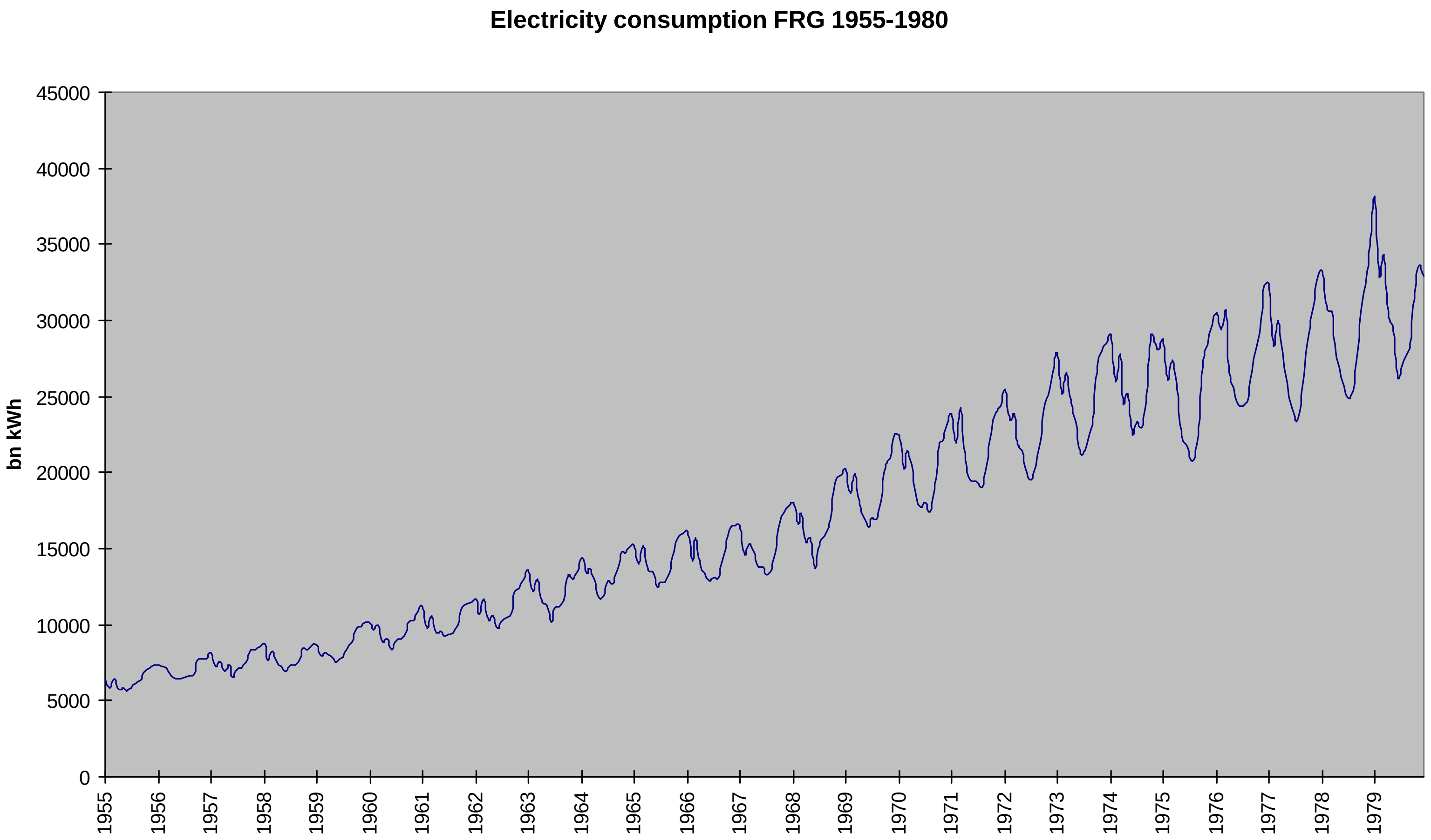
The first two components are often aggregated into the *smooth* component, component two and three are often aggregated into the *cyclic* component.

The simplest case assumes that the four components add up to the time series:

$$x_t = m(t) + k(t) + s(t) + u(t) \tag{3}$$

m is a monotonic function, k is a periodic function with period $\gg 1$ year, s is a periodic function with period $= 1$ year, while u is a random function (stochastic process).

An example: electric power production in Germany



Taking logarithms

In many cases we can observe that the amplitude of $s(t)$ and/or the variance of $u(t)$ increase with t (or with $m(t)$). Hence it is a good idea to model the time series as follows:

$$\log x_t = m(t) + k(t) + s(t) + u(t) \tag{4}$$

which is the same as

$$x_t = \exp[m(t)] \exp[k(t)] \exp[s(t)] \exp[u(t)] \tag{5}$$

(multiplicative model).

Regression vs. smoothing

In both cases one will estimate the parameters of the functions m, k, and s with regression methods (making some assumptions about the period of the trade cycle component). The residual component $u(t)$ is the regression residual (so-called global component model).

Instead one could try to eliminate the residual component by some smoothing procedure such as moving averages (so-called local component model).

Smoothing

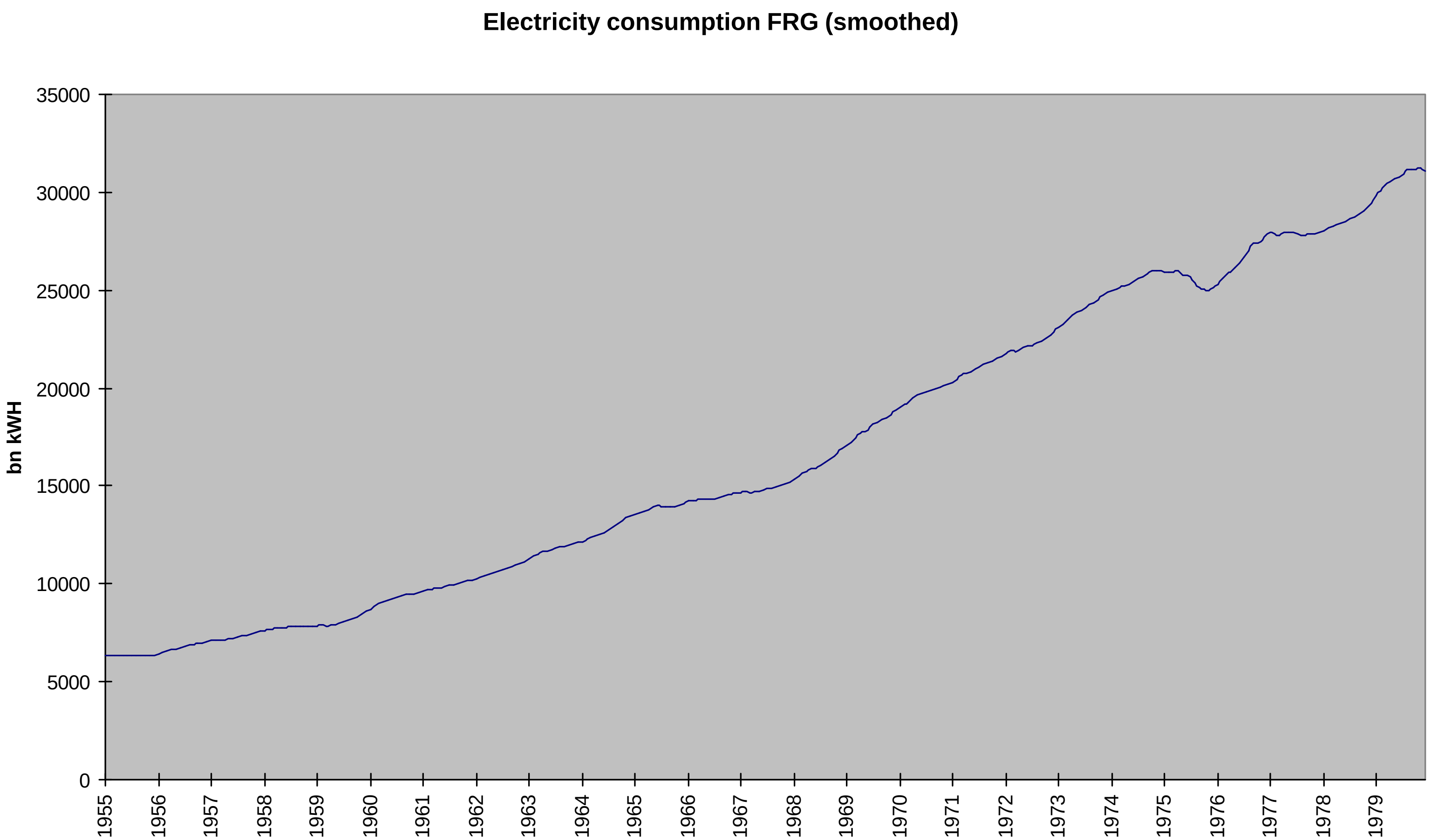
Smoothing by moving averages:

In a monthly time series with a seasonal component (x_t) we calculate means (y_t) of twelve consecutive months each:

$$y_t = \frac{1}{12} \sum_{i=-11}^{0} x_{t+i} \qquad t = 12 \ldots N \tag{6}$$

In this case, no values can be calculated for the first 11 months. Alternatively, one could average from $i = t$ bis $t + 11$, then there are no values for the last 11 months; another approach is averaging from $i = t - 5$ to $t + 6$, then values would be missing both at the beginning and at the end of the time series..

An example of smoothing



Calculating the trend component

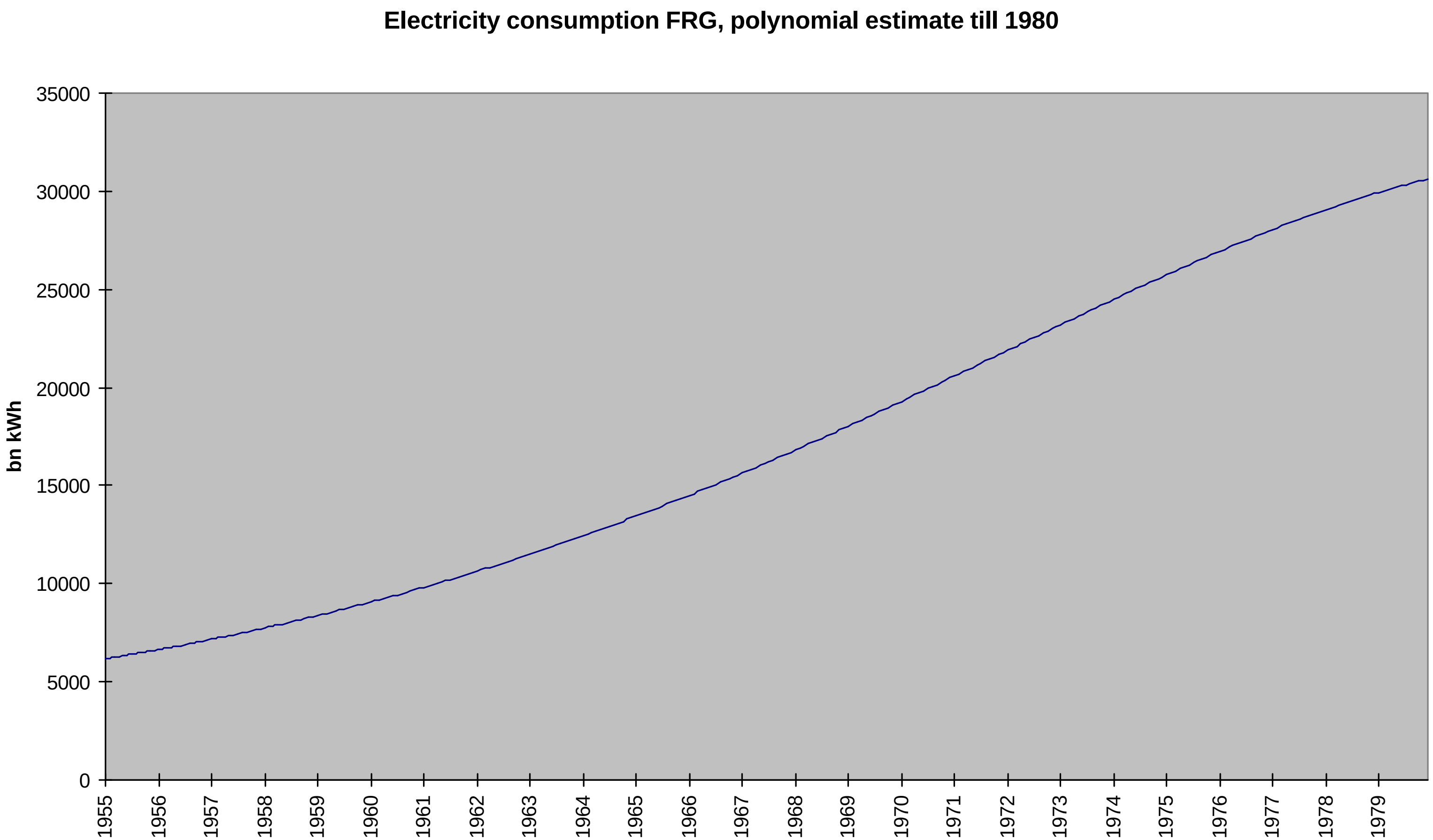
The trend component (or the „smooth“ component as a whole) is mostly estimated by polynomial regression ($\sum_t u_t^2 = \mathrm{min!}$):

$$x_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \ldots + u_t \quad (7)$$

The result for the electric power production is:

$$\ln x_t = 8.719645 + 0.073684t + 0.000971538t^2 - 0.000053744t^3 + u_t \quad (8)$$

Polynomial regression



Estimating the trend component: prediction

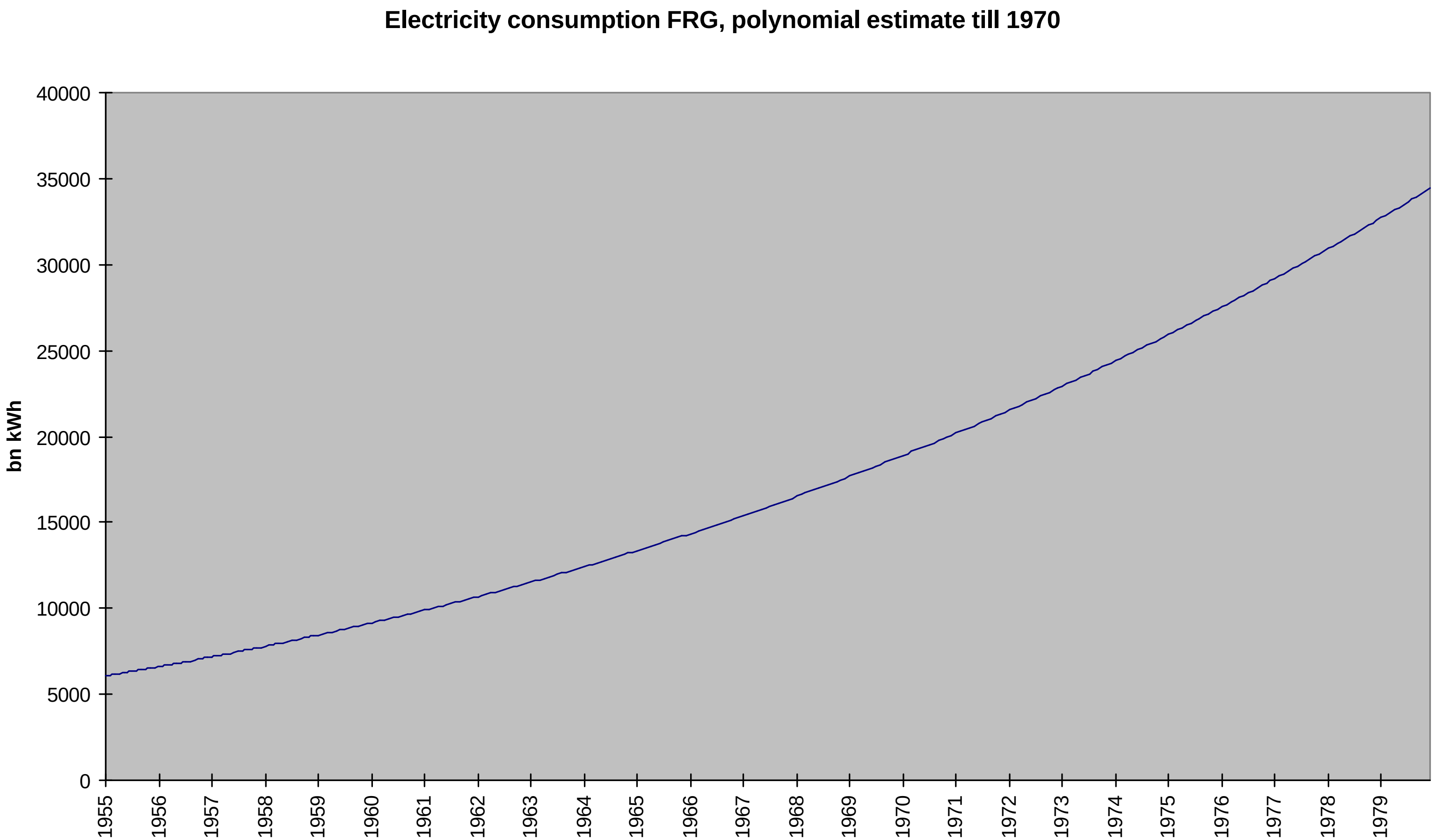
If we use only the first 15 (instead of 25) years for parameter estimation, i.e. if we use only the knowledge available at the end of 1974, we have:

$$\ln x_t = 8.702018 + 0.085957t - 0.000695387t^2 + 0.00000191542t^3 + u_t \quad (9)$$

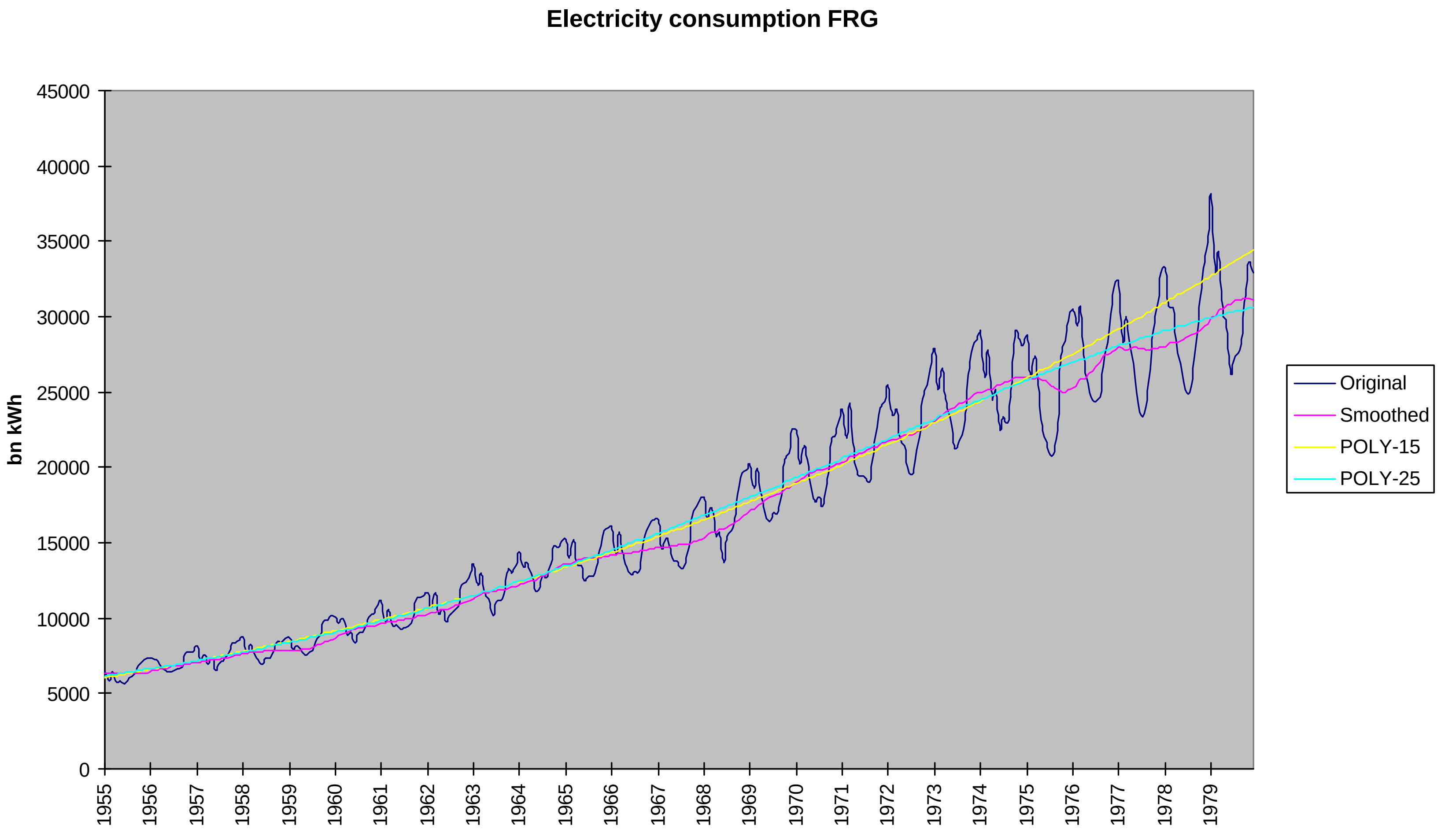
Compare parameters from full knowledge:

$$\ln x_t = 8.719645 + 0.073684t + 0.000971538t^2 - 0.000053744t^3 + u_t \quad (10)$$

An example: prediction after the first 15 years



Comparison: full time series against prediction from the first 15 years



Stochastic processes

A stochastic process is a series $\{X_t\}_{t \in T}$ of random variables X_t. Here, t — the time parameter — is an element of the index set T which we will identify with the set of (positive) integers.

A random variable X is a mapping $X : \Omega \to R$ which attributes real numbers $X(\omega)$ to the outcomes ω of a random process. Thus, a result ω of a random process corresponds to the time series $\{X_t\}_{t \in T}$.

Stochastic processes (2)

$X_t(\omega)$ has different meanings, depending on whether we conceive of ω and t as fixed or variable:

t fixed, ω fixed:	$X_t(\omega)$ is a real number, i.e. the t-th value of the ω-th realisation of the stochastic process (of the ω-th time series).
t variable, ω fixed:	$X_t(\omega)$ is a series of real numbers, i.e. a series of all the values of the ω-th realisation of the stochastic process (it is the ω-th time series).
t fixed, ω variable:	$X_t(\omega)$ is a real-valued random variable, i.e. $X_t(\omega)$ represents all possible realisations of the t-th random variable.
t variable, ω variable:	$X_t(\omega)$ is the complete stochastic process, $X_t(\omega)$ represents an ensemble of time series.

Stochastic processes (3)

ω	1	2	3	4	5	t
a	$X_1(a)$	$X_2(a)$	$X_3(a)$	$X_4(a)$	$X_5(a)$	$\cdots$
b	$X_1(b)$	$X_2(b)$	$X_3(b)$	$X_4(b)$	$X_5(b)$	$\cdots$
c	$X_1(c)$	$X_2(c)$	$X_3(c)$	$X_4(c)$	$X_5(c)$	$\cdots$
d	$X_1(d)$	$X_2(d)$	$X_3(d)$	$X_4(d)$	$X_5(d)$	$\cdots$

Rows a, b, c, d are different realisations of the process, i.e. different time series.

Columns 1, 2, 3, 4 are different random variables of the process.

White-noise process

A white-noise process $\{\varepsilon_t\}_{t \in T}$ is a series of identically and independently distributed random variables ε_t. The distribution function in any column of the table is the same, it is independent of any other column.

Example: Gain/loss of a roulette player (per game) who, in every game, risks € 1,00 on rouge. For every game t the probability function and the distribution function are

Note that “when the zero appears, all even-money bets, such as Red, Black, Odd, or Even, are ‘imprisoned’. On the next spin of the wheel, if the zero appears again the house collects half of each imprisoned bet; if not, it collects all losing bets and returns the original bets to any winners. The bank’s percentage on red, black, odd, or even is $1\frac{13}{37}$ percent, and on all other types of bets it is $2\frac{26}{37}$ percent.” [Encyclopedia Britannica]

White-noise process

Example: Gain/loss of a roulette player (per game) who, in every game, risks € 1,00 on rouge. For every game t (in the zero case: *before* the next spin of the wheel) the probability function and the distribution function are

$$f(\varepsilon) = \begin{cases} \frac{18}{37} & \varepsilon = -1 \\ \frac{1}{37} & \varepsilon = 0 \\ \frac{18}{37} & \varepsilon = 1 \\ 0 & else \end{cases} \quad (11) \qquad F(\varepsilon) = \begin{cases} 0 & \varepsilon < -1 \\ \frac{18}{37} & -1 \leq \varepsilon < 0 \\ \frac{19}{37} & 0 \leq \varepsilon < 1 \\ 1 & 1 \leq \varepsilon \end{cases} \quad (12)$$

All ε_t are identically and independently distributed (iid).

Random-walk process

If $\{\varepsilon_t\}_{t \in T}$ ia a white-noise process then

$$X_t = \begin{cases} \varepsilon_1 & \text{für } t = 1 \\ X_{t-1} + \varepsilon_t & \text{for } t > 1 \end{cases} \tag{13}$$

is a random-walk process.

Random-walk process (2)

Example: The cumulated gain/loss of the same roulette player, after each game. The probability functions $f_t(x)$ and the distribution functions $F_t(x)$ now read:

$$f_1(x) = \begin{cases} \frac{18}{37} & x = -1 \\ \frac{1}{37} & x = 0 \\ \frac{18}{37} & x = 1 \\ 0 & else \end{cases} \quad (14) \qquad F_1(x) = \begin{cases} 0 & x < -1 \\ \frac{18}{37} & -1 \leq x < 0 \\ \frac{19}{37} & 0 \leq x < 1 \\ 1 & 1 \leq x \end{cases} \quad (16)$$

$$f_2(x) = \begin{cases} \frac{324}{1369} & x = -2 \\ \frac{36}{1369} & x = -1 \\ \frac{649}{1369} & x = 0 \\ \frac{36}{1369} & x = 1 \\ \frac{324}{1369} & x = 2 \\ 0 & else \end{cases} \quad (15) \qquad F_2(x) = \begin{cases} 0 & x < -2 \\ \frac{324}{1369} & -2 \leq x < -1 \\ \frac{360}{1369} & -1 \leq x < 0 \\ \frac{1009}{1369} & 0 \leq x < 1 \\ \frac{1045}{1369} & 1 \leq x < 2 \\ 1 & 2 \leq x \end{cases} \quad (17)$$

X_1 are X_2 not identically distributed, and the same holds for any other pair X_s and X_t.

Random-walk process (3)

Besides, they are not independently distributed (divide all numbers by 1369):

	x_2	-2	-1	0	1	2	total
x_1	-1	324	18	324	0	0	666
	0	0	18	1	18	0	37
	1	0	0	324	18	324	666
	total	324	36	649	36	324	1369

$$f_{1,2}(0,1) = \frac{18}{1369} = 0.0131482834186... \quad (18)$$

$$f_1(0)f_2(1) = \frac{36}{1369} \cdot \frac{37}{1369} = 0.000710718022625... \quad (19)$$

Cosinoid process

Let A and B be two independent, normally distributed random variables with mean μ_a and μ_b and variances σ_a^2 and σ_b^2. Then

$$X_t = A \cos 2\pi\lambda t + B \sin 2\pi\lambda t \tag{20}$$

is a cosinoid with respect to the frequency λ. Remember that each realisation of the cosinoid process is a sine wave with wavelength $2\pi\lambda$ but with different phases which depend on A and B.

Gauss process (normal process)

For any finite selection of time points, the random variables have a joint normal distribution.

How to describe stochastic processes

For any finite selection of n time points, the n corresponding random variables of the stochastic process have a well-defined joint distribution which can be described by the n-dimensional distribution function

$$F_{t_1,t_2,\cdots,t_n}(x_1, x_2, \cdots, x_n) = \mathsf{Pr}(X_{t_1} \leq x_1, X_{t_2} \leq x_2, \cdots, X_{t_n} \leq x_n) \tag{21}$$

(see the example of the white-noise and random-walk processes).

If we can write down the joint distribution function for every possible selection of random variables of a stochastic process in a manner that for each random variable the marginal distribution in each selection is the same, then the stochastic process is unambiguously defined (Kolmogorov's consistency theorem).

How to describe stochastic processes (2)

Practically, we will rather be interested in the moments of first and second order (means, variances, covariances) instead of on the distribution functions. So far these moments exist they are called:

mean function:

$$\mu(t) = E\langle X_t \rangle \tag{22}$$

variance function:

$$\sigma^2(t) = E\langle [X_t - \mu(t)]^2 \rangle \tag{23}$$

How to describe stochastic processes (3)

covariance function:

$$\gamma(s,t) = E\langle [X_s - \mu(s)][X_t - \mu(t)] \rangle \tag{24}$$

correlation function:

$$\rho(s,t) = \frac{\gamma(s,t)}{\sigma(s)\sigma(t)} \tag{25}$$

How to describe stochastic processes (4)

From the definitions follows:

$$\sigma^2(t) = \gamma(t,t) \tag{26}$$

— so if a covariance function exists, so does the variance function — and

$$\rho(t,t) = 1 \tag{27}$$

For finite selections of the stochastic process, the values of $\gamma(s,t)$ are collected in the covariance matrix

$$\Sigma = (\gamma(s,t)) \tag{28}$$

Obviously, Σ is a symmetrical matrix.

How to describe stochastic processes (5)

Examples:

In a white-noise process — where all random variables ε_t are identically and independently distributed — we have:

$$\begin{aligned} \mu(t) &= \mu \quad \text{für alle } t && (29) \\ \sigma^2(t) &= \sigma^2 \quad \text{for all } t && (30) \\ \gamma(s,t) &= \begin{cases} \sigma^2 & \text{für } s = t \\ 0 & \text{for } s \neq t \end{cases} && (31) \end{aligned}$$

Describing a random-walk process

In a random-walk process with

$$X_t = \sum_{s=1}^{t} \varepsilon_s \qquad (32)$$

we have a mean function

$$\mu_X(t) = E\langle \sum_{s=1}^{t} \varepsilon_s \rangle = \sum_{s=1}^{t} E\langle \varepsilon_s \rangle = \mu_\varepsilon t \qquad (33)$$

Describing a random-walk process (2)

… and a variance function

$$\begin{aligned}
\sigma_X^2(t) &= E\langle(\sum_{s=1}^{t} \varepsilon_s - \mu_\varepsilon t)^2\rangle && (34)\\
&= E\langle[\sum_{s=1}^{t} (\varepsilon_s - \mu_\varepsilon)]^2\rangle && (35)\\
&= E\langle[\sum_{s=1}^{t} (\varepsilon_s - \mu_\varepsilon)][\sum_{r=1}^{t} (\varepsilon_r - \mu_\varepsilon)]\rangle && (36)\\
&= E\langle \sum_{s=1}^{t} \sum_{r=1}^{t} (\varepsilon_s - \mu_\varepsilon)(\varepsilon_r - \mu_\varepsilon)\rangle && (37)\\
&= \sum_{s=1}^{t} \sum_{r=1}^{t} E\langle(\varepsilon_s - \mu_\varepsilon)(\varepsilon_r - \mu_\varepsilon)\rangle && (38)\\
&= \sum_{s=1}^{t} E\langle(\varepsilon_s - \mu_\varepsilon)^2\rangle = t\sigma_\varepsilon^2 && (39)
\end{aligned}$$

Describing a random-walk process (3)

Only for $\mu_\varepsilon = 0$ the mean function $\mu_X(t)$ of the random-walk process is constant, in all other cases (random-walk process with drift) it is linear in t.

The covariance function of the random-walk process is $(s \leq t)$:

$$\gamma_X(s,t) = \text{Cov}[\sum_{i=1}^{s} \varepsilon_i, \sum_{j=1}^{s} \varepsilon_j] \quad (40)$$

$$= \sum_{i=1}^{s} \sum_{j=1}^{s} \text{Cov}[\varepsilon_i, \varepsilon_j] \quad (41)$$

$$= s\sigma_\varepsilon^2 \quad (42)$$

Because of $\gamma_X(s,t) = \gamma_X(t,s)$ we have

$$\gamma_X(s,t) = \min(s,t)\sigma_\varepsilon^2 \quad (43)$$

Stationary processes

From one realisation of a stochastic process mean function and variance function can only be estimated if we make certain assumptions about the process behind a time series. Note that we can never check whether these assumptions are met.

We will assume that empirical time series are realisations of stationary processes and test of the time series which we will analyse can be a realisation of a stationary process. If this is not the case, then we will try to transform (filter) the time series in a manner that at least the filtered time series is stationary.

Stationary processes (2)

We call a stochastic process $\{X_t\}_{t \in T}$

- stationary with respect to the mean if $\mu(t) = \mu$ for all $t \in T$,
- stationary with respect to the variance if $\sigma^2(t) = \sigma^2$ for all $t \in T$,
- stationary with respect to the covariance if $\gamma(s,t) = \gamma(s+r, t+r)$ for all $r, s, t \in T$,
- weakly stationary if it is both stationary with respect to the mean and to the covariance.

Stationary processes (3)

In processes which are stationary with respect to their covariance we write the covariance and correlation functions

$$\gamma(s,t) = \gamma(s+r, t+r) = \gamma(s-t) = \gamma(\tau) = \gamma(-\tau) \tag{44}$$

and

$$\rho(s,t) = \rho(s+r, t+r) = \rho(s-t) = \rho(\tau) = \rho(-\tau) \tag{45}$$

Linear filters and stochastic processes

A filter transforms a time series (or a stochastic process as a whole) into another time series (or stochastic process):

We call the input time series x_t, the output time series y_t and the filter a_u:

$$y_t = \sum_u a_u x_{t-u} \tag{46}$$

Instead of applying the filter just to one time series x_t (fixed ω!) we can also apply it to the process as a whole $X_t(\omega)$ (ω variable!) and allow filters of infinite length:

$$Y_t = \sum_{u=-\infty}^{\infty} a_u X_{t-u} \tag{47}$$

Filters and moment functions

Mean function:

$$\mu_Y(t) = E\langle Y_t \rangle = \sum_u a_u E\langle X_{t-u} \rangle = \sum_u a_u \mu(t-u) \tag{48}$$

If $\{X_t\}_{t \in T}$ is stationary with respect to the mean, so is $\{Y_t\}_{t \in T}$.

Filters and moment functions (2)

Covariance function:

$$\gamma_Y(s,t) = \text{Cov}[Y_s, Y_t] = \text{Cov}[\sum_v a_v X_{s-v}, \sum_u a_u X_{t-u}] \quad (49)$$

$$= \sum_v \sum_u a_u a_v \text{Cov}[X_{s-v}, X_{t-u}] \quad (50)$$

$$= \sum_v \sum_u a_u a_v \gamma_X(t-u, s-v) \quad (51)$$

If (X_t) is stationary with respect to covariance, so is (Y_t), since then we have

$$\gamma_X(t-u, s-v) = \gamma_X(t-s+v-u) \quad (52)$$

such that $\gamma_Y(s,t)$, too, depends only on $(t-s)$.

If we apply filters with finite sums, stationarity with respect to mean and covariance (weak stationarity) is kept.

The backshift operator and the difference operator

In the following we will often use the backshift operator:

$$\begin{aligned} B(x_t) &= x_{t-1} && (53)\\ B^2(x_t) &= x_{t-2} && (54)\\ &\dots && (55)\\ B^n(x_t) &= x_{t-n} && (56) \end{aligned}$$

and the difference operator

$$\begin{aligned} \Delta(x_t) &= x_t - x_{t-1} = (1-B)(x_t) && (57)\\ \Delta^2(x_t) &= x_t - 2x_{t-1} + x_{t-2} = (1-B)^2(x_t) && (58) \end{aligned}$$

Invertible filters

If it is possible to transform the stochastic process $\{X_t\}_{t\in T}$ into the process $\{Y_t\}_{t\in T}$ by filtering it with the filter $A(B)$, it might be interesting to find out whether $\{Y_t\}_{t\in T}$ can be re-transformed into $\{X_t\}_{t\in T}$ — by applying the filter $A^{-1}(B)$.

We restrict ourselves to “causal” filters whose output depends only on the present and the past of its input:

$$C(B) = \sum_{u=0}^{\infty} c_u B^u \tag{59}$$

Invertible filters (2)

Thus we are looking for a filter $A^{-1}(B)$ which is the inverse of the filter $A(B)$: with

$$y_t = A(B)x_t = \sum_u a_u x_{t-u} \tag{60}$$

the following holds:

$$x_t = A^{-1}(B)y_t = \sum_u c_u y_{t-u} \tag{61}$$

or in short:

$$y_t = A(B)x_t \leftrightarrow x_t = A^{-1}(B)y_t \tag{62}$$

Invertible filters (3)

The filter $A(B)$ is not only

$$A(B) = a_0 B^0 + a_1 B^1 + a_2 B^2 + ... + a_p B^p \tag{63}$$

but it can also be written as follows:

$$A(B) = \alpha_0 (1 - \alpha_1 B)(1 - \alpha_2 B)...(1 - \alpha_p B) \tag{64}$$

Invertible filters (4)

Expansion yields:

$$
\begin{aligned}
a_0 &= \alpha_0 \\
a_1 &= -\alpha_0 \sum_{i=1}^{p} \alpha_i \\
a_2 &= \alpha_0 \sum_{i=1}^{p} \sum_{j=i+1}^{p} \alpha_i \alpha_j \\
&\cdots \\
a_p &= (-1)^p \prod_{i=0}^{p} |\alpha_i| \qquad (65) \\
& \qquad (66)
\end{aligned}
$$

such that the filter $A(B)$ can be replaced by the subsequent application of the filters $(1 - \alpha_i B)$ with the final multiplication by $a_0 = \alpha_0$.

Invertible filters (5)

So we can restrict ourselves to the problem of the inversion of the filter $A(B) = (1 - \alpha B)$. Its inverse is

$$A^{-1}(B) = 1 + \alpha^1 B^1 + \alpha^2 B^2 + \cdots = \sum_{u=0}^{\infty} \alpha^u B^u \qquad (67)$$

Invertible filters (6)

Proof: Let $y_t = x_t - \alpha x_{t-1}$. The following assumption holds:

$$\begin{aligned} x_t &= y_t + \alpha y_{t-1} + \alpha^2 y_{t-2} + \alpha^3 y_{t-3} + \cdots & (68) \\ \alpha x_{t-1} &= \quad + \alpha y_{t-1} + \alpha^2 y_{t-2} + \alpha^3 y_{t-3} + \cdots & (69) \\ x_t - \alpha x_{t-1} &= y_t & (70) \end{aligned}$$

Replacing the operator B with the (complex) number z we have the z-transform $A(z)$ (the characteristic polynomial) of the filter $A(B)$, for which the following holds:

$$A^{-1}(z) = \frac{1}{A(z)} = \frac{1}{\alpha z} = 1 + \alpha z + (\alpha z)^2 + ... = \sum_{u=0}^{\infty} \alpha^u z^u \qquad (71)$$

so with the help of the z-transformation we can calculate with operator polynomials in the same way as with ordinary polynomials.

Invertible filters (7)

The inversion of the filter $A(B)$ is only possible if the series $\sum_{u=0}^{\infty} \alpha^u B^u y_t$ converges, i.e. if $|\alpha| < 1$. The zero of the z-transform $1 - \alpha z$ then is $z_0 = 1/\alpha$ with $|z_0| > 1$. With this result we can write the filter

$$A(B) = a_0 B^0 + a_1 B^1 + a_2 B^2 + \cdots + a_p B^p \tag{72}$$

as

$$A(B) = a_0 (1 - \frac{1}{z_1} B^1)(1 - \frac{1}{z_2} B^2)(1 - \frac{1}{z_3} B^3) \cdots (1 - \frac{1}{z_p} B^p) \tag{73}$$

where the z_i are the zeroes of the z-transformed polynomial

$$A(z) = a_0 z^0 + a_1 z^1 + a_2 z^2 + \ldots + a_p z^p \tag{74}$$

Invertible filters (8)

. . . and we realise that an inverse filter $A^{-1}(B)$ for

$$A(B) = a_0 B^0 - a_1 B^1 - a_2 B^2 - \ldots - a_p B^p \tag{75}$$

exists iff for all zeroes $z_1, z_2, \ldots, z_p$ of the characteristic polynomial

$$A(z) = a_0 z^0 - a_1 z^1 - a_2 z^2 - ... - a_p z^p \tag{76}$$

$|z_i| > 1$ holds, i.e. if all zeroes of the z-transform of the filter lie outside the unit cycle of the complex plane.

Transformations stabilising the mean — polynomials

Stochastic processes whose mean function is a polynomial of d-th order in t may be made stationary with respect to the mean by applying the difference filter $\Delta = 1 - B$ d times.

Let $E\langle X_t\rangle = P_d(t)$ be a polynomial of order d with coefficients $a_0, a_1, a_2, \ldots, a_d$. Then

$$E\langle\Delta^d X_t\rangle = E\langle\binom{d}{0}X_t - \binom{d}{1}X_{t-1} + \binom{d}{2}X_{t-2} + - + \cdots\rangle \quad (77)$$

$$= P_d(t) - \binom{d}{1}P_d(t-1) + \binom{d}{2}P_d(t-2) - + \cdots \quad (78)$$

Transformations stabilising the mean — polynomials (2)

Expanding and simplifying yields

$$E\langle\Delta^d X_t\rangle = d!a_d \tag{79}$$

which is the same as the d-th derivative of the polynomial wrt t.

Transformations stabilising the mean — sine waves

Stochastic processes whose mean function is a sine wave of frequency λ can be made stationary with respect to the mean by filtering it with $(1 - 2\cos 2\pi\lambda B + B^2)$.

Let $E\langle X_t \rangle = a\cos 2\pi\lambda t + b\sin 2\pi\lambda t$ then

$$\begin{aligned}
E\langle X_t - (2\cos 2\pi\lambda) X_{t-1} + X_{t-2} \rangle = \\
= \quad & a\cos 2\pi\lambda t + b\sin 2\pi\lambda t \\
& -(2\cos 2\pi\lambda)[a\cos 2\pi\lambda(t-1) + b\sin 2\pi\lambda(t-1)] \\
& +a\cos 2\pi\lambda(t-2) + b\sin 2\pi\lambda(t-2) \qquad (80) \\
= \quad & a\cos 2\pi\lambda t + b\sin 2\pi\lambda t \\
& -(2\cos 2\pi\lambda)[a(\cos 2\pi\lambda t\cos 2\pi\lambda + \sin 2\pi\lambda t\sin 2\pi\lambda) \\
& \ + b(\sin 2\pi\lambda t\cos 2\pi\lambda - \cos 2\pi\lambda t\sin 2\pi\lambda)] \\
& +a(\cos 2\pi\lambda t\cos 4\pi\lambda + \sin 2\pi\lambda t\sin 4\pi\lambda) \\
& +b(\sin 2\pi\lambda t\cos 4\pi\lambda - \cos 2\pi\lambda t\sin 4\pi\lambda) \qquad (81)
\end{aligned}$$

Transformations stabilising the mean — sine waves (2)

$$\begin{aligned} = \; & a\cos 2\pi\lambda t + b\sin 2\pi\lambda t \\ & -2\cos^2 2\pi\lambda(a\cos 2\pi\lambda t + b\sin 2\pi\lambda t) \\ & -2a\sin 2\pi\lambda\cos 2\pi\lambda\sin 2\pi\lambda t \\ & +2b\sin 2\pi\lambda\cos 2\pi\lambda\cos 2\pi\lambda t \\ & +\cos 4\pi\lambda(a\cos 2\pi\lambda t + b\sin 2\pi\lambda t) \\ & +a\sin 4\pi\lambda\sin 2\pi\lambda t \\ & -b\sin 4\pi\lambda\cos 2\pi\lambda t \end{aligned} \tag{82}$$

Transformations stabilising the mean

Because of $\sin 4\pi\lambda = 2\sin 2\pi\lambda \cos 2\pi\lambda$ lines 3, 4, 6 and 7 of the last equation can be cancelled, and because of $\cos 4\pi\lambda = \cos^2 2\pi\lambda - \sin^2 2\pi\lambda$ we are left with

$$\begin{aligned} & E\langle X_t - (2\cos 2\pi\lambda)X_{t-1} + X_{t-2}\rangle = \\ & = (a\cos 2\pi\lambda t + b\sin 2\pi\lambda t)[1 - 2\cos^2 2\pi\lambda + \cos^2 2\pi\lambda - \sin^2 2\pi\lambda] \\ & = 0 \end{aligned} \tag{83}$$

With the means of description treated so far we cannot draw conclusions from a time series to the mean function, so we cannot know how often we have to apply the difference filter and/or which frequency filter has to be applied.

Transformations stabilising the variance

Without proof we give the Box-Cox transformation for stochastic processes $\{X_t\}_{t \in T}$ for which

$$E\langle X_t\rangle = \mu(t) \quad (84)$$

$$\text{Var}[X_t] = [c(\mu(t))^{1-\theta}]^2 \quad (85)$$

holds, i.e. for which the standard deviation is proportional to some power of the mean. For these cases Box and Cox postulated the transformation

$$Y_t = \begin{cases} (X_t^\theta - 1)/\theta & \theta \neq 0 \\ \ln X_t & \theta = 0 \end{cases} \quad (86)$$

Y_t is stationary with respect to the variance. Using $Y_t = \ln X_t$ for $\theta = 0$ follows from $\lim_{\theta \to 0} \frac{x^\theta - 1}{\theta} = \ln x$.

ARIMA model

The ARIMA model assumes that each value in a time series is a linear function of its own past and of the present and the past of a random process:

$$\begin{aligned} x_t &= \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \alpha_3 x_{t-3} + \cdots + \alpha_p x_{t-p} \\ &\quad + \varepsilon_t - \beta_1 \varepsilon_{t-1} - \beta_2 \varepsilon_{t-2} - \cdots - \beta_q \varepsilon_{t-q} \end{aligned} \tag{87}$$

This is an example of an ARMA(p, q) process. The abbreviation “AR” means “autoregressive” und points to the fact that x_t is represented by a regression equation which contains its own *(auto)* past. The abbreviation “MA” means “moving average” and points to the fact that the residual of the AR part of the regression equation is a generalised (weighted) moving average of a random process.

Trend

The “I” in “ARIMA” represents a trend-like component of the time series. To discuss it in more detail we have to introduce the concept of differencing and the related operator Δ:

$$y_t = \Delta x_t = x_t - x_{t-1} \tag{88}$$

which serves — possibly applied to the time series several times — to make the time series “stationary”, i.e. to eliminate a trend if there is any. Applying the differencing operator several times means:

$$z_t = \Delta^2 x_t = y_t - y_{t-1} = (x_t - x_{t-1}) - (x_{t-1} - x_{t-2}) \tag{89}$$

Trend: a simple example with numbers

That a trend can be eliminated by using the differencing operator can be seen from a simple non-stochastic time series:

t	1		2		3		4		5		6		7
x_t	1		8		27		64		125		216		343
Δx_t		7		19		37		61		91		127	
$\Delta^2 x_t$			12		18		24		30		36		
$\Delta^3 x_t$				6		6		6		6			

ARIMA and Backshift Operator

To make clearer how the differencing operator works, we introduce another operator B (Backshift Operator) which has the following meaning:

$$B^n x_t = x_{t-n} \tag{90}$$

It is obvious that:

$$\Delta = 1 - B \tag{91}$$

$$\Delta^2 = 1 - 2B + B^2 \tag{92}$$

$$\Delta^n = (1 - B)^n \tag{93}$$

ARIMA and Backshift Operator

Hence we can rewrite the ARMA(p, q) process:

$$\begin{aligned} x_t - \alpha_1 B^1 x_t - \alpha_2 B^2 x_t - \alpha_3 B^3 x_t - \cdots - \alpha_p B^p x_t = \\ = \varepsilon_t - \beta_1 B^1 \varepsilon_t - \beta_2 B^2 \varepsilon_t - \cdots - \beta_q B^q \varepsilon_t \end{aligned} \quad (94)$$

Both sides of the equation are polynomials in Bx_t and $B\varepsilon_t$ respectively with the coefficients $\langle 1, -\alpha_1, -\alpha_2, \cdots, -\alpha_p \rangle$ and $\langle 1, -\beta_1, -\beta_2, \cdots, -\beta_q \rangle$. We will write these polynomials in the following manner:

$$\alpha(B)x_t = x_t - \alpha_1 B^1 x_t - \alpha_2 B^2 x_t - \alpha_3 B^3 x_t - \cdots - \alpha_p B^p x_t \quad (95)$$

$$\beta(B)\varepsilon_t = \varepsilon_t - \beta_1 B^1 \varepsilon_t - \beta_2 B^2 \varepsilon_t - \beta_3 B^3 \varepsilon_t - \cdots - \beta_q B^q \varepsilon_t \quad (96)$$

such that the ARMA(p, q) process can be simplified:

$$\alpha(B)x_t = \beta(B)\varepsilon_t \quad (97)$$

ARIMA and ARMA as filtered white-noise processes

So we can write the ARMA(p, q) process

$$\alpha(B)x_t = \beta(B)\varepsilon_t \tag{98}$$

as the result of filtering a white-noise process:

$$x_t = \alpha^{-1}(B)\beta(B)\varepsilon_t \tag{99}$$

ARMA and ARIMA

An ARIMA(p, d, q) process is a process which becomes an ARMA(p, q) process by d times differencing. So if $\{X_t\}_{t \in T}$ is such an ARIMA(p, d, q) process, then

$$W_t = \Delta^d X_t \tag{100}$$

is an ARMA(p, q) process, and we have

$$\alpha(B)(1 - B)^d x_t = \beta(B)\varepsilon_t \tag{101}$$

Motivation of the ARIMA model (1)

To motivate the ARIMA model we consider a system whose state can be given by two variables X and Y at any time. Each of these variables is assumed to depend on the immediate past of the other variable and on random variables:

$$x_t = \gamma_1 y_{t-1} + \varepsilon_t \quad (102)$$

$$y_t = \gamma_2 x_{t-1} + \varepsilon'_t \quad (103)$$

where ε and ε' are two processes which do not depend on each other.

Motivation of the ARIMA model (2)

Because of

$$y_{t-1} = \gamma_2 x_{t-2} + \varepsilon'_{t-1} \tag{104}$$

we have

$$x_t = \gamma_1 \gamma_2 x_{t-2} + \gamma_1 \varepsilon'_{t-1} + \varepsilon_t \tag{105}$$

With σ^2 and σ'^2 as the variances of ε and ε', we can write:

$$\gamma'_1 = \gamma_1 \frac{\sigma^2}{\sigma'^2} \tag{106}$$

and because of the independence of ε und ε' we have:

$$x_t = \gamma_1 \gamma_2 x_{t-2} + \gamma'_1 \varepsilon_{t-1} + \varepsilon_t \tag{107}$$

which yields an ARMA(2,1) process (or an ARIMA(2,0,1) process) with:

$$\alpha_1 = 0 \qquad \alpha_2 = \gamma_1 \gamma_2 \qquad \beta_1 = -\gamma'_1 \tag{108}$$

Estimating ARIMA parameters

To begin with, we have to estimate the parameters p, d and q of an ARIMA model (its „order“).

With an appropriate estimate of p, d und q, $\alpha_1, \ldots, \alpha_p$ and $\beta_1, \ldots, \beta_q$ have to be calculated. This is done by maximising a nonlinear function in $p + q$ variables — and this is far from trivial (and beyond the scope of this introduction!).

Most often, one would calculate several different ARIMA models (for different combinations of p, d and q) and keep the solution which best estimates the parameters with the smallest number of parameters.

Estimating the order of differencing

In many important cases, applying the difference filter to a process stationary with respect to the mean increases its variance as compared to the variance of the original process:

$$\sigma^2_{\Delta X} > \sigma^2_X \quad X_t \text{ stationary with respect to the mean} \tag{109}$$

On the other hand, if a process is not stationary with respect to the mean, then differencing will decrease the variance of the differenced process:

$$\sum(x-c)^2 > \sum(x-\mu)^2 \quad c \neq \mu \tag{110}$$

So to find out how often to difference, we need only to apply the difference filter until the variance starts to increase.

Seasonal differencing

Differencing alone might not be sufficient to make the process under consideration stationary wrt the mean: There might be seasonal influences.

A seasonal influence (in a monthly time series) may be removed by the filter $\Delta_{12} = (1 - B^{12})$. This filter, too, may be applied several times. The method of variate differences is then — as it were — two-dimensional. Example (electric power consumption 1955–1979 after taking logarithms):

	0	1	2	3	4
0	1.000	0.011	0.028	0.091	0.324
1	0.019	0.008	0.024	0.078	0.264
2	0.038	0.025	0.072	0.228	0.763
3	0.126	0.085	0.244	0.774	2.584
4	0.452	0.299	0.866	2.757	9.203

Here, one ordinary differencing and one seasonal differencing would be optimal.

Other necessary transformations

It might also be necessary to remove a sine wave — a periodogram will help to find out which frequencies should be filtered out.

An additional Box-Cox transformation might also be necessary — if necessary it will be the first transformation applied.

Periodogram

Can a harmonic wave explain part of the variance of our time series — and to what extent?

$$\text{Harmonic wave:} \quad m(t) = \beta_1 \cos 2\pi\lambda t + \beta_2 \sin 2\pi\lambda t \tag{111}$$

Estimate β_1 and β_2 with

$$Q = \sum_t \left[(x_t - \bar{x}) - \hat{m}(t)\right]^2 = \text{min!} \tag{112}$$

$\Rightarrow (\lambda = k/N, 0 < \lambda < 0.5)$

$$\hat{\beta}_1 = \frac{2}{N} \sum_t (x_t - \bar{x}) \cos 2\pi\lambda t = 2C(\lambda) \tag{113}$$

$$\hat{\beta}_2 = \frac{2}{N} \sum_t (x_t - \bar{x}) \sin 2\pi\lambda t = 2S(\lambda) \tag{114}$$

Periodogram (2)

$\widehat{\beta}_1$ and $\widehat{\beta}_2$ are twice the covariances of x_t with $\cos 2\pi\lambda t$ and $\sin 2\pi\lambda t$, respectively.

Q is the sum of squared errors of the estimation (the sum of squared residuals):

$$Q = \underbrace{\sum_{t=1}^{N} (x_t - \bar{x})^2}_{\sim\text{total variance}} - \underbrace{2N\left[C(\lambda)^2 + S(\lambda)^2\right]}_{\sim\text{explained variance}} \qquad (115)$$

Periodogram and cumulated periodogram

The periodogram $I(\lambda)$ is that part of the total variance which is explained by the harmonic wave of frequency λ.

$$
\begin{aligned}
I(\lambda) &= 2N\left[C(\lambda)^2 + S(\lambda)^2\right] && (116)\\
&= 2N\left[\frac{1}{N}\sum_t (x_t - \bar{x})\cos 2\pi\lambda t\right]^2 + 2N\left[\frac{1}{N}\sum_t (x_t - \bar{x})\sin 2\pi\lambda t\right]^2 && (117)
\end{aligned}
$$

The cumulated periodogram is (for $\lambda_k = k/N, k = 1, \ldots, N/2$)

$$
S_r = \frac{\sum_{k=1}^{r} I(\lambda_k)}{\sum_{k=1}^{N} I(\lambda_k)} \qquad (118)
$$

Estimating the order of AR and MA

In the early times of ARIMA analysis of time series, the autocorrelation function (ACF) and the partial autocorrelation function (PACF) were used to find an estimate for p and q.

$$\text{ACF:} \qquad \rho_X(\tau) = \frac{\text{Cov}[X_t, X_{t-\tau}]}{\text{Var}[X_t]} \tag{119}$$

$$\text{PACF:} \qquad \pi_X(\tau) = \text{Corr}[X_t, X_{t-\tau}].X_{t-1} \ldots X_{t-\tau+1} \tag{120}$$

A more modern version of this “detecting patterns where there are no patterns” is the technique of vector correlations, which can be done automatically but is much more complicated — and more reliable! — than the original technique.