Zeitreihenanalyse

WS 2005/2006

Gunnar Lischeid
(M. Hauhs, H. Lange)

http://www.bayceer.uni-bayreuth.de/mod

- Module M103, *M409, M509*
- Vorl.-Nr. 282121

- ergänzend dazu: Blockpraktikum Zeitreihenanalyse
  (Februar 2006); *Module M409, M509* (Vorl.-Nr. 28392)

Gliederung

1. Definition und Eigenschaften von Zeitreihen
2. Tests und Trenderkennung für Zeitreihen
3. Fourriertransformation, Powerspektrum
4. Wavelets
5. Zeitreihenmodellierung der ARMA-Klasse (kurzes Gedächtnis)
6. Modellierung von Zeitreihen mit langem Gedächtnis
7. Komplexität und Information von Zeitreihen
8. Wiederkehranalyse
9. Singuläre System-Analyse
10. Skalierung, (Multi-)Fraktale

Literatur zum Thema

Schönwiese, C. (2000): Praktische Statistik für Meteorologen und Geowissenschaftler. Bornträger.

Schlittgen, R. (2001): Angewandte Zeitreihenanalyse. Oldenbourg.

Hipel, K.W., McLeod, A.I. (1994): Time Series Modelling of Water Resources and Environmental Systems. Elsevier.

Tong, H. (1990): Non-linear Time Series. Oxford Science Publ.

Definition: Zeitreihe

Definition: Eine **Zeitreihe** ist eine Menge von Werten, die in einer festgelegten (und bekannten!) **Reihenfolge** vorliegen:

$$X = \left\{ x_i \middle| i = 1, \ldots, n \right\}$$

- Die Zuordnung der Werteposition zum Referenzzeitpunkt ist eine **monotone** Funktion

$$i \rightarrow t_i : i > j \Rightarrow t_i > t_j$$

- Ist der zeitliche Abstand zweier Messungen konstant:

$$t_i - t_{i-1} = \Delta t \neq f(i)$$

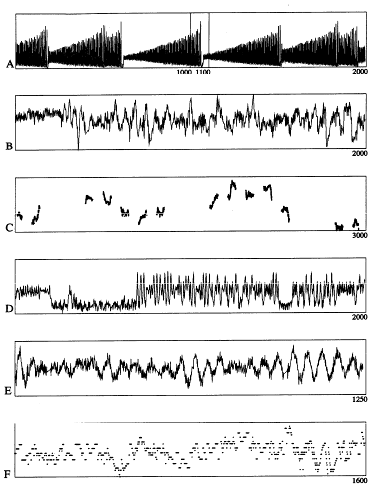
heißt die Zeitreihe **äquidistant**. Es gilt dann $t_i = t_0 + i\Delta t$

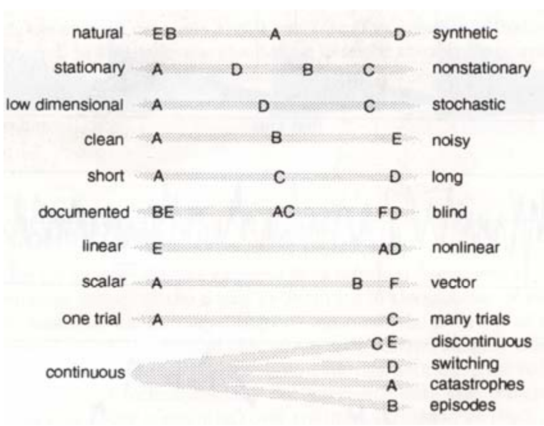
- Fehlt ein i in dieser Liste, hat die Zeitreihe eine **Lücke**.

Anwendungsbeispiele

- Börsenkurse
- Wahlverhalten
- Bevölkerungsentwicklung
- Klimatologie
- Hochwassergefährdungsabschätzung
- ...

Zeitreihen: Beispiele





Zugänge

- aus der Physik:
  - Suche nach der Dynamik des erzeugenden Systems (z.B. Geophysik, Meteorologie)
  - der typische Zugang in den **Geo**wissenschaften
- aus der Mathematik:
  - als Beispiel für geordnete (oder partiell geordnete) Mengen
- aus den Ingenieurwissenschaften (z.B. Hydrologie):
  - als Ausdruck des empirischen Wissens (Abflüsse)
- aus der Modellbildung:
  - als wichtiges Beispiel zur Demonstration der heutigen technischen Möglichkeiten
  - gibt es einen typischen Zugang für Ökosysteme ?

Ziele

- Frühwarnsystem (Trendanalyse)
- Prozessanalyse (Kausalzusammenhänge)
- Klassifizierung (von Verhalten)
- Vorhersage (Abfluss, Börsenkurs)
- Überprüfung/Optimierung von Modellen

Zeitreihenanalyse in der Geoökologie

- direkteste Verbindung zur experimentellen Beschreibung von Systemen (Datenerhebung)
- kommt (i.d.R.) ohne Annahme von Prozessen aus
- kommt mit gar keinem bis wenigen Parametern aus
- konkrete empirische Beschreibung des zeitlich variablen (dynamischen) Verhaltens
- Vorhersage oft erfolgreicher als bei Prozessmodellen
- Klassifikation von Modellen nach ihrer Erklärungsleistung
- sensibler Test von Modellen ("mehr als r^2")

Wichtige Begriffe

- **Univariate** Zeitreihe: Eine (reellwertige) Variable, an einem Ort gemessen
- **Multivariate** Zeitreihe: mehrere Variablen am selben Ort
- **Mehrdimensionale** Zeitreihe: eine Variable an verschiedenen Orten zu jeweils gleichen Zeitpunkten
- **Äquidistante** Zeitreihe
- **Lückenfreiheit**
- **Homogenität**: pdf ändert sich nicht mit der Zeit

-> Generelles Problem: viele Eigenschaften beziehen sich auf / sind nur definiert für unendlich lange Zeitreihen

Charakterisierung der Verteilung I

für eine Datenreihe $x(t_i), i = 1,...,n$:

- Mittelwert: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x(t_i)$
- Varianz $\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x(t_i) - \bar{x}\right)^2$

(Standardabweichung: σ)

- Variationskoeffizient: $cv = \frac{\sigma}{\bar{x}}$

Faustregel :
Zur Berechnung des ***q***-ten Moments benötig man mind. ***2q*** Datenpunkte

- ***q***-tes zentrales Moment: $M_q = \frac{1}{n}\sum_{i=1}^{n}\left(x(t_i) - \bar{x}\right)^q$

Charakterisierung der Verteilung II

Häufigkeitsverteilung: Histogramme

- Klassenbreite (Binbreite): $\Delta x = \frac{(x_{\max} - x_{\min})}{n_{\text{bins}}}$

  Faustregel: 95% der Klassen (Bins) sollten je mind. 5 Datenpunkte enthalten

Häufigkeitsverteilungen $\rightarrow$ Wahrscheinlichkeitsverteilungen (**pdf´s**)

- Median: 50% der Werte sind kleiner
- Modus/Modalwert: Position des Maximums der pdf
- x%-Quantil: x % der Werte sind kleiner

Korrelation

(Produkt-Moment-Korrelation, Pearson ***r***)

Varianz: $$\sigma^2(x) = \frac{\sum_{i=1}^{n}(x-\bar{x})^2}{n} = \frac{\sum_{i=1}^{n}(x-\bar{x})\cdot(x-\bar{x})}{n}$$

Kovarianz: $$\operatorname{cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{n}$$

Korrelation: $$r = \frac{\operatorname{cov}(x,y)}{\sqrt{\sigma^2(x)}\cdot\sqrt{\sigma^2(y)}} = \frac{\operatorname{cov}(x,y)}{\sigma_x\cdot\sigma_y}$$

maximal mögliche Kovarianz

***r*²** = Bestimmheitsmaß = erklärte Varianz

Andere Korrelationsmaße

Spearman-Rangkorrelationskoeffizient:

(D_i: Rangplatzdifferenzen)

$$r_R = 1 - \frac{6\cdot\sum D_i^2}{n^3-n}$$

Kendall's τ:

(*Ko*: Konkordanzen = gleichsinnige Änderungen $x_1 \to x_2$ und $y_1 \to y_2$,

Di: Diskordanzen)

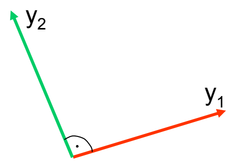
$$r_K = \frac{2\cdot(Ko-Di)}{n\cdot(n-1)}$$

Korrelation

= Maß für die Gemeinsamkeiten zweier Funktionen

-> mathematisch bestimmt durch die Multiplikation der jeweils entsprechenden (normierten) Funktionswerte

geometrisches Analog: Skalarprodukt



=> ***orthogonale*** Funktionen sind nicht korreliert

Faltung (Convolution)

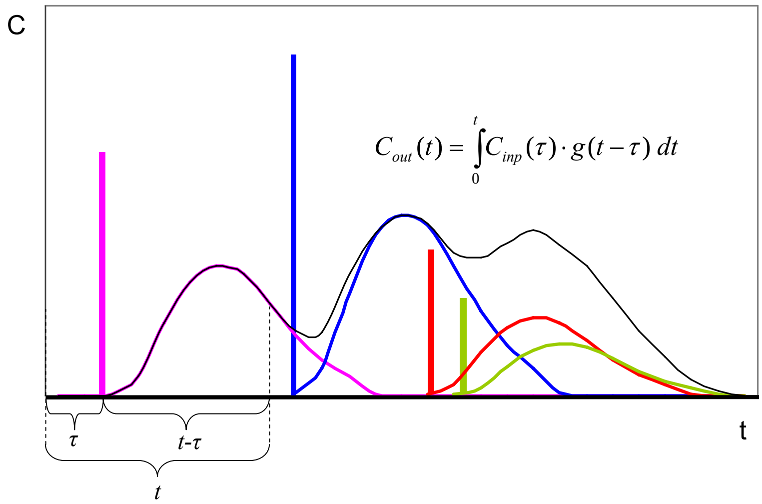
- Verallgemeinerung der Korrelation
- liefert für zwei Funktionen ***f*** und ***g*** eine dritte Funktion, die die "Überlappung" zwischen ***f*** und einer *gespiegelten, verschobenen* Version von ***g*** angibt.
- für kontinuierliche Funktionen:

$$(f * g)(t) = \int f(\tau) \cdot g(t - \tau)\, d\tau$$

- für diskrete Werte:

$$(f * g)(n) = \sum_{k=-\infty}^{\infty} f(k) \cdot g(n - k)$$

Beispiel: Tracer im Vorfluter



Faltung

Anwendungsbeispiele:

- Beschreibung der System***antwort*** auf eine ***Anregung*** durch Faltung der ***Anregungsfunktion*** mit der ***Impulsantwort*** des Übertragungsglieds.
- Charakterisierung der Gewichtung einer Funktion mit einer anderen. Der Funktionswert der Gewichtsfunktion an einer Stelle ***t*** gibt an, wie stark der um τ zurückliegende Wert der gewichteten Funktion in den Wert der Ergebnisfunktion eingeht.

Visualisierung der Faltung

Java-Applets:

- http://www.jhu.edu/~signals/convolve/index.html
- http://www.getsoft.net/fouriertrans/animationen/animation04.html

Zeitliche Korrelation

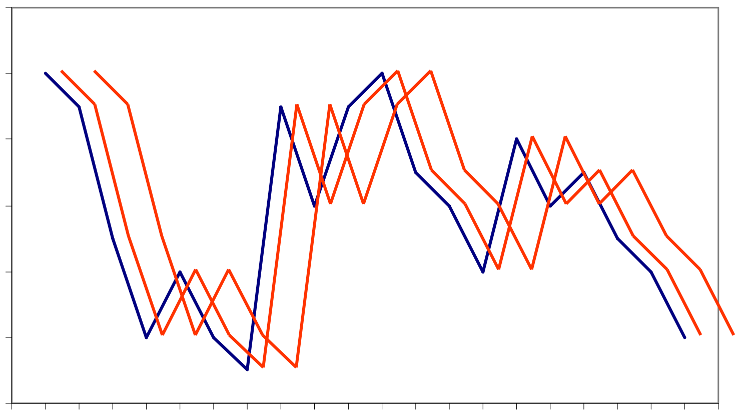
Korrelation: $$r = \frac{1}{\sigma_x \cdot \sigma_y} \cdot \frac{\sum_{i=1}^{n} (x - \bar{x}) \cdot (y - \bar{y})}{n}$$

zeitliche Verschiebung
= time **lag**

Autokorrelation: $$\rho_{ac,k} = \frac{1}{\sigma_x \cdot \sigma_x} \cdot \frac{\sum_{i=1}^{n} (x(t_i) - \overline{x(t)}) \cdot ([x(t_i + k)] - \overline{x(t)})}{n}$$

Kreuzkorrelation: $$\rho_{cc,k} = \frac{1}{\sigma_x \cdot \sigma_x} \cdot \frac{\sum_{i=1}^{n} (x(t_i) - \overline{x(t)}) \cdot ([y(t_i + k)] - \overline{y(t)})}{n}$$

Autokorrelation



Praktische Hinweise

Faustregeln:

- mindestens 30 Datenpunkte
- nur Lags
  k < n/4 (Puristen) bzw.
  k < n/2 (Pragmatiker) vertrauen
- Daten müssen „im Prinzip“ äquidistant vorliegen;
  Lücken sind ein echtes Problem!

Test für (lineare) Unkorreliertheit

Berechnung der Autokorrelation. Für unkorrelierte Daten gilt:

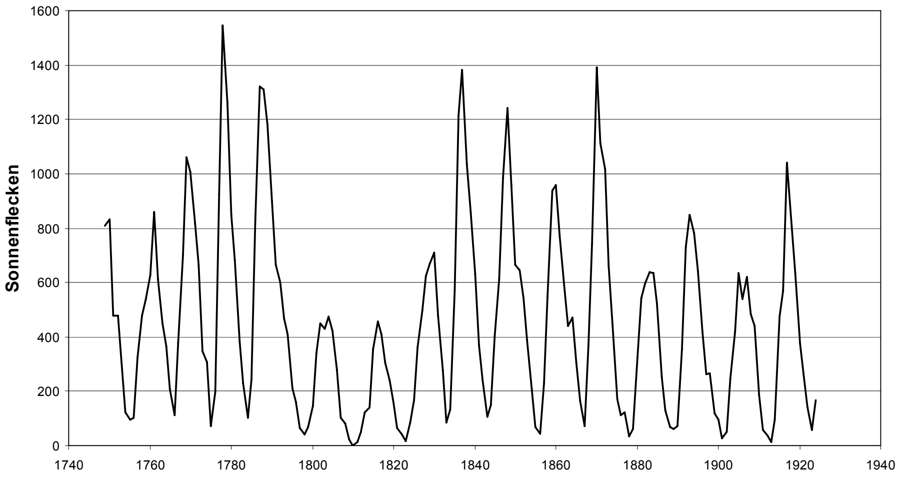
$\rho_k \sim N(0,\frac{1}{n})$ (normalverteilt)

Daher sind die 95%-Signifikanzlinien

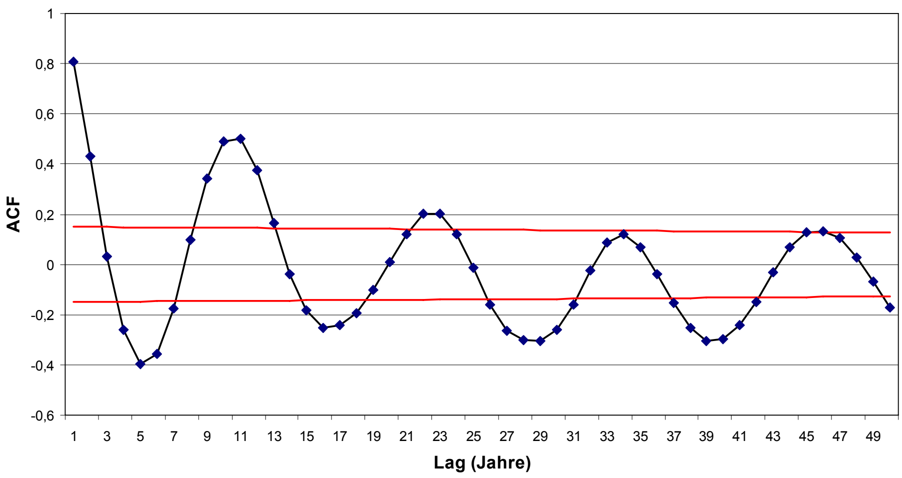
$SL_{0,95} = \pm\frac{1.96}{\sqrt{n}}$ (evtl. Verbesserungen für kleine n)

Liegen weniger als 5% der Werte außerhalb des Intervalls, liegen keine signifikanten Korrelationen vor $\Rightarrow$ Autokorrelationslänge

Beispiel: Sonnenfleckenrelativzahlen



Autokorrelation der Sonnenflecken

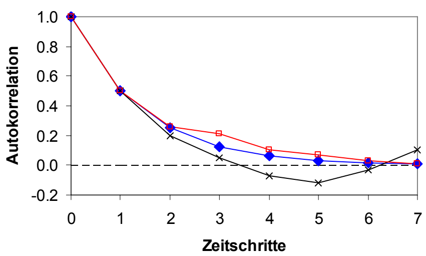


Partielle Autokorrelation

Bsp.: Autokorrelation für ***1*** Zeitschritt = $\rho_{1\cdot\Delta t}$

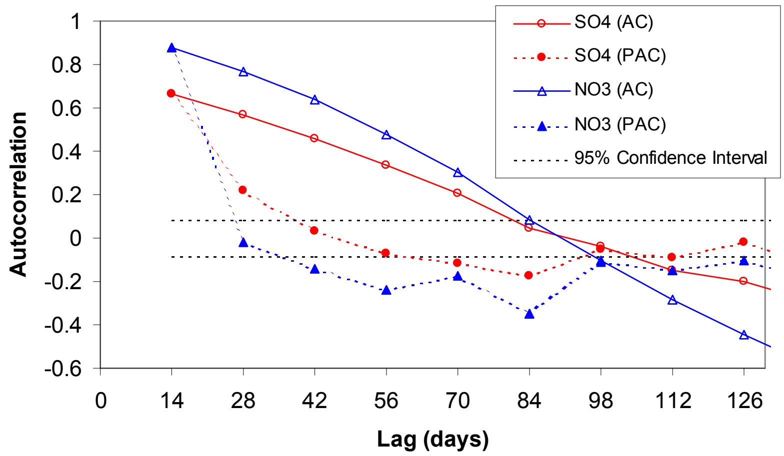
=> Autokorrelation für ***2*** Zeitschritte = $\rho_{1\cdot\Delta t} \cdot \rho_{1\cdot\Delta t} = \rho_{1\cdot\Delta t}^{2}$

=> Autokorrelation für ***n*** Zeitschritte = $\rho_{1\cdot\Delta t}^{n}$



=> Die beobachtete Autokorrelation setzt sich aus den Beiträgen der einzelnen Zeitdifferenzen Δ***t*** zusammen (***Partielle Autokorrelationen***)

Bsp.: Autokorrelation



Wann ist eine Zeitreihe eine Zeitreihe?

Gibt es signifikante Autokorrelationen, ist die zeitliche Reihenfolge wichtig. Die einzelnen Werte sind dann *nicht* unabhängig.

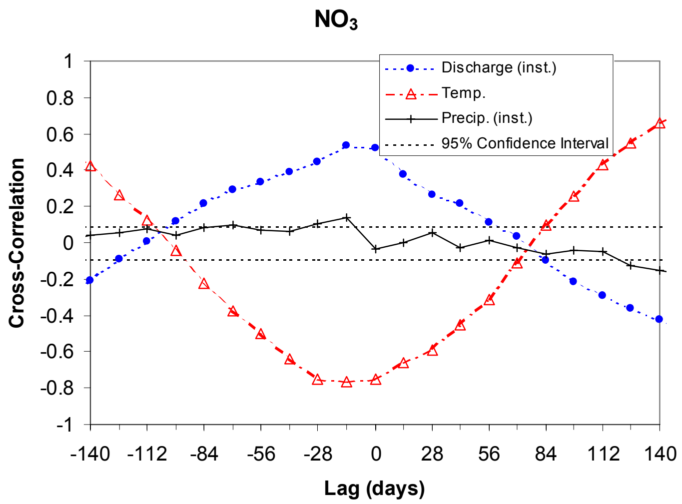
Unabhängigkeit erreicht man durch

- Aggregation
- Wahl einer gröberen Messauflösung

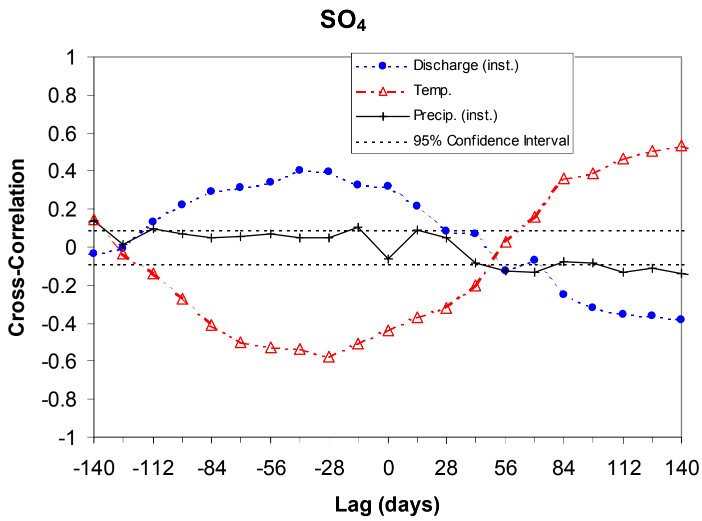
Falls unabhängig:
Zeitreihen als *Realisationen eines stochastischen Prozesses*

I.a. liegen Mischtypen vor (z.B. additives Rauschen)

Bsp.: Kreuzkorrelation



Bsp.: Kreuzkorrelation



Eigenschaften der Kreuzkorrelation

$-1 \le \rho_{uv}(k) \le 1$

$\rho_{uv}(k)$ ist *nicht* symmetrisch: $\rho_{uv}(-k) \ne \rho_{uv}(k), -\rho_{uv}(k)$

95% Signifikanz: $\left|\rho_{uv}(k)\right| > 1.96/(n-|k|)^{1/2}$

Lag-Beschränkung: $-n/4 \le k \le n/4$