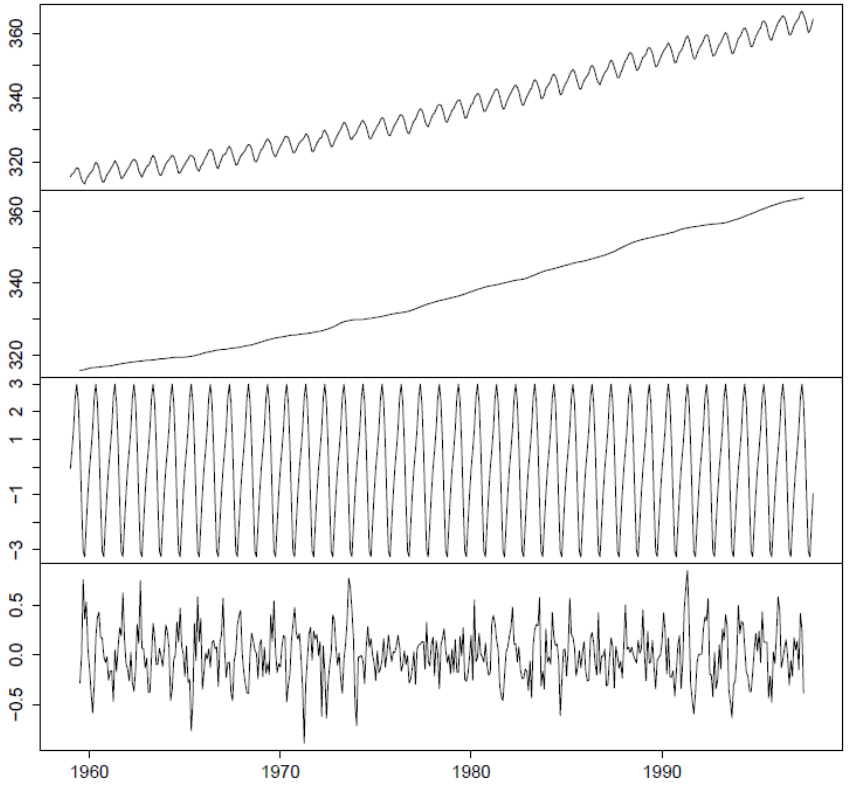
ALGORITMI ANALIZE MASIVNIH PODATKOV

DOMEN MONGUS

P05 – Analiza časovnih vrst

Motivacija - V 4 Velocity

- Časovna vrsta:
  - Beli šum, naključni sprehod,…
- Razlogi za variabilnost vrednosti
  - Dekompozicija signala:
  - Trend:
  - Sezonski efekti:
  - Neregularne fluktuacije



Avtokorelacija in linearna regresija

- Pearsonov korelacijski koeficient

$$r = r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)}\sqrt{(\sum y_i^2 - n\bar{y}^2)}}.$$

- Avtokorelacija je korelacija med signalom $X = \{x_t\}$ in njegovo zakasnjeno kopijo $X_h = \{x_{t+h}\}$

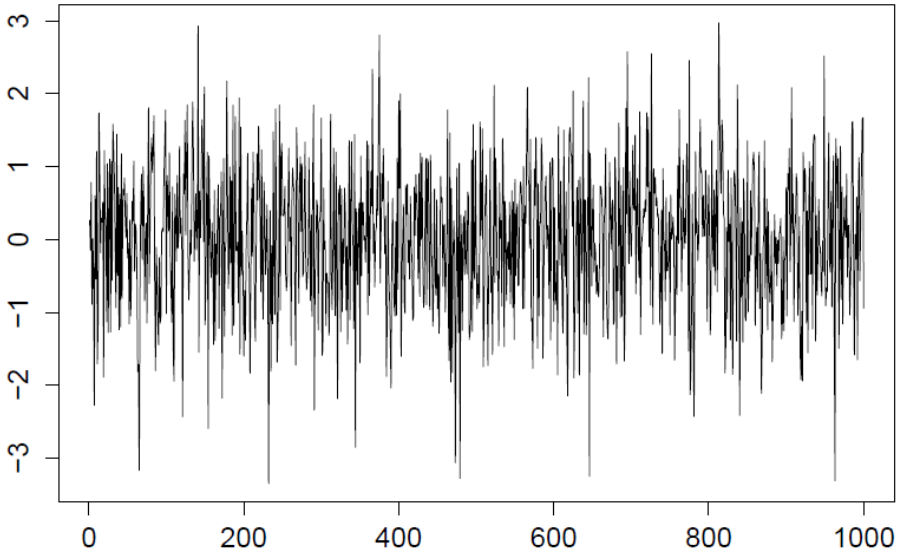
- Generalizirana regresijska enačba $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$
  - Metoda najmanjših kvadratov $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

Vsebina

- Tradicionalni pristopi, ki napovedovanju vrednosti v časovnih vrstah:
  - Premikajoče povprečje
  - Avtoregresijski model
  - ARIMA
- Načrtovanje napovedovalnih modelov

Beli šum

- Definicija
  - Povprečje = *0*
  - Varianca = *konstanta*
  - Je nekoreliran
- V signalu ne ugotovimo vzorca
  - Množica statističnih testov
- Zaključni kriterij vsake časovne analize.



Nelinearna regresija z metodo najmanjših kvadratov

- Matrična predstavitev metode najmanjših kvadratov v linearnem sistemu:

$$y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_k x_{i,k} + \varepsilon_i$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,k} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,k} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

- Polinomska regresija:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_m x_i^m + \varepsilon_i \ (i = 1, 2, \ldots, n)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^m \\ 1 & x_2 & x_2^2 & \ldots & x_2^m \\ 1 & x_3 & x_3^2 & \ldots & x_3^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^m \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

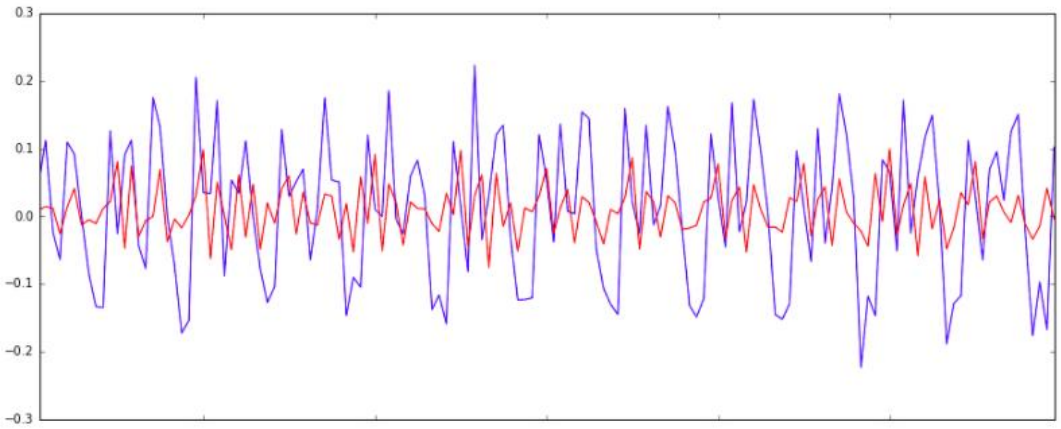
Kako z več spremenljivkami?

Avtoregresijski model

- Notacija *AR(p)*
  - *p* – število preteklih vrednosti, ki jih uporabimo za napoved
- Definicija: $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + w_t$
  - Rešljivo s tradicionalno linearno regresijo
- Zahteva stacionarne vrednosti!
  - Povprečje = 0
  - Standardna deviacija = konstanta

Premikajoče povprečje

- Ang. moving average
- Notacija: MA(q)
  - q – določa dolžino modela
- Definicija: $X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$
  - θ_q - koeficienti (povezani z napakami), ki jih računamo
  - ε_t, ε_{t-1},…, ε_{t-q} - napake prejšnjih napovedih (stacionarni)
  - μ - povprečje zadnjih *q* vrednosti



Reševanje MA modelov

- Napake v napovedovanju so nedoločljive, zato potrebujemo iterativni pristop.

- Množica možnih pristopov:
  - Metoda Yule–Walker
  - Metoda največje verjetnosti (Maximum Likelihood)
    - Newton–Raphsonov in Scoring Algorithmi
  - Iterativni Gauss – Newtonov algoritem

Avtoregresijsko premikajoče povprečje

- Ang. autoregresive moving average
- *Notacija: ARMA(p,q)*
  - *p* – število avtoregresijskih členov
  - q – ševilo členov belega šuma (napak)

- Definicija:

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}$$

Integrirana ARMA

- Ang. Avtoregresive Integrated Moving Average
- *Notacija: ARIMA(p,d,q)*
  - *p* in *q* prevzeta iz ARMA
  - *d* – red diferenciacije
- Diferenciacija:
  - Prvi red: $y'_t = y_t - y_{t-1}$
  - Drugi red: $$\begin{aligned} y^*_t &= y'_t - y'_{t-1} \\ &= (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) \\ &= y_t - 2y_{t-1} + y_{t-2} \end{aligned}$$

Višje običajno ne gremo…

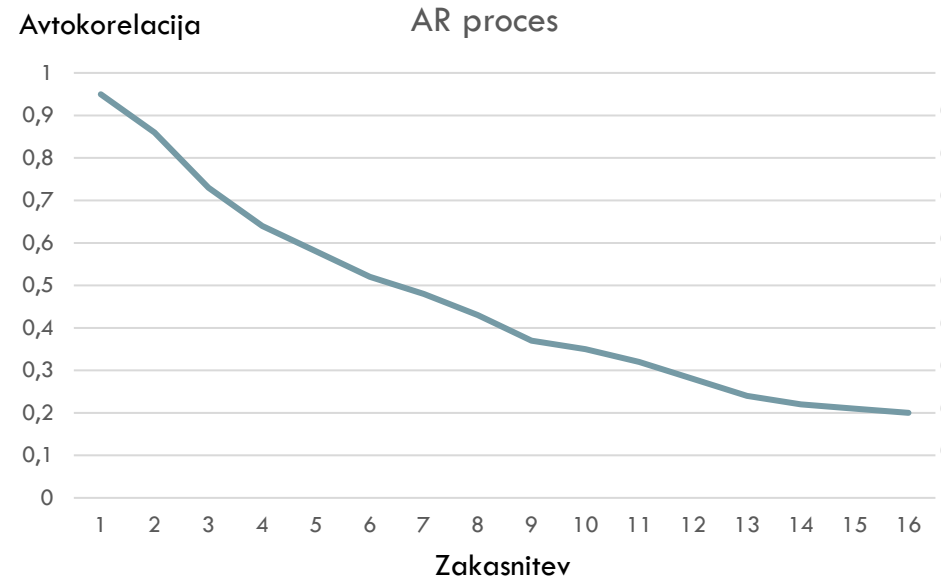
Načrtovanje modelov

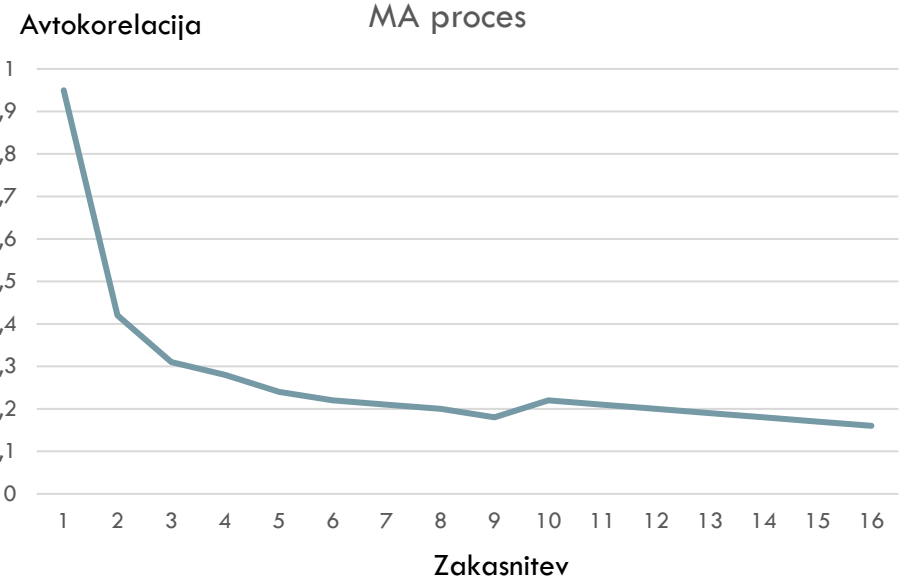
- Metoda Box-Jenkins
  - Identifikacija modela
  - Izvedba modela
  - Diagnostika

- Pred uporabo B-J izvedi:
  - Ali so podatki beli šum?
  - Ali je časovna vrsta stacionarna?
    - Če ni, izvedi diferenciacijo

Izračun korelograma

- Procese modeliramo glede ne vrste, ki jih razberemo na osnovi korelograma:
  - To je graf avtokorelacije glede na zakasnitev
  - Lahko izberemo optimalno zakasnitev?





Delna avtokorelacija

- Korelacija je približek regresije z navadnimi najmanjšimi kvadrati:
  - x: {22,17,16,14,13,10,1215,21,19,18,16,19,20,24}
  - x_1 : {17,16,14,13,10,1215,21,19,18,16,19,20,24, 21}
  - Korelacijski faktor: 0.68132
  - Regresijski koeficient (AR(1) model): 0.69608
    - Vsaka vrednost „ima 69% vpliva na naslednjo vrednost“
- PRIMER AR(3) modela:
  - $X = k_1 x_1 + k_2 x_2 + k_3$
  - Zanima nas koliko točno vpliva $k_2 x_2$ npr. brez $k_1 x_1$
    - *Izkaže se, da je vpliv n-tega zamika enak koeficientu n-tega člena regresije n-zamikov.*

Metoda Box-Jenkins

- Identifikacija modela in optimalne zakasnitve
  - Analiza avtokorelacije za določitev najprimernejše zakasnitve (optimal lag).
  - Upoštevanje delne avtokorelacije
    - Izločanje ekstremnih dogodkov

MODEL	Avtokorelacija	Delna avtokorelacija
AR(p)	Pada počasi proti 0	Takoj upade blizu 0
MA(q)	Takoj upade blizu 0	Pada počasi proti 0
ARMA(p,q)	Pada počasi proti 0	Pada počasi proti 0

Metoda B-J

- Diagnostika:
  - Običajno izračun korena povprečne kvadratne napake (ang. root mean square error)

$$\mathrm{RMSE} = \sqrt{\sum \frac{(y_{pred} - y_{ref})^2}{N}}$$

  - Izris grafa napake in preveri ali so napake res beli šum?