

## Lab 6: Data Mining. Serii temporale

---

Analiza seriilor temporale are ca scop să modeleze și să explice dependența unor date de momente de timp succesive. Exemple tipice de serii temporale sunt: temperatura înregistrată zilnic, curs de schimb valutar, prețul unor acțiuni, valori ale unor mărimi fiziologice înregistrate de dispozitive de monitorizare (tensiune arterială, puls etc) etc.

Principalele prelucrări care pot fi efectuate asupra unei serii de timp sunt:

- *Vizualizare:*
  - Vizualizarea evoluției în timp a mărimii analizate (**R**: pentru obiecte de tip `ts` se poate folosi `ts.plot` sau `plot.ts`; cele două funcții diferă în principal prin modul de vizualizare a mai multor serii)
- *Pre-procesare*
  - completarea valorilor absente prin interpolare;
- *Analiză: identificare*
  - Tendință (trend)
    - se poate determina din seria inițială prin eliminarea zgomotului folosind o tehnică de *netezire* (smoothing): medie mobilă (Moving Average) sau netezire exponențială. In **R** se poate folosi funcția `filter` pentru a defini filtrul de netezire
    - se poate elimina prin “*diferențiere*”: noua serie se obține din cea anterioară prin calculul diferențelor dintre elementele vecine; pentru diferențe de ordin mai mare decât 1 se aplică repetat strategia; în practică diferențele de ordin 1 permit eliminarea tendinței liniare, iar cele de ordin 2 permit eliminarea tendinței pătratice. **R**: funcția `diff`
  - Caracter sezonier (seasonal)
    - Se poate determina calculând media valorilor corespunzătoare unei perioade. De exemplu pentru date înregistrate lunar pentru care există caracter sezonier la nivel de an (există un tipar de comportament de-a lungul unui an) se calculează mediile pentru toate lunile și se obține o estimare a tiparului corespunzător unui an.
    - Se poate elimina calculând diferențe între valori corespunzătoare unor întârzieri în timp egale cu perioada caracterului sezonier. De exemplu pentru date înregistrate lunar pentru care există caracter sezonier la nivel de an (există un tipar de comportament de-a lungul unui an) se poate folosi întârziere (lag)  $L=12$ .
  - Zgomot (noise, residuals)
    - Se obține prin eliminarea tendinței și a caracterului sezonier
  - Staționaritate (stationary): seria este staționară dacă caracteristicile sale statistice locale nu se schimbă de-a lungul seriei (nu există tendință și nici caracter sezonier). O serie nestaționară poate fi transformată într-una staționară prin diferențiere (calcul diferențe).

**Obs:** cele 3 componente ale unei serii (tendința, componenta sezonieră și zgomotul) se pot determina direct în **R** folosind funcția `decompose`.

- *Predicție*: estimarea valorilor ulterioare din serie pe baza valorii curente și a celor anterioare (folosind un model care descrie dependența valorii curente din serie de valorile anterioare). Principalele etape:
  - se elimină tendința și componenta sezonieră
  - se alege un model care explică variabilitatea din componenta de zgomot
    - ARMA(p,q) – pt serii staționare
    - ARIMA(p,d,q) – pt serii nestaționare

**Obs:** în cazul în care mai multe modele conduc la valori similare ale erorii (suma pătratelor erorilor) se alege modelul mai simplu, de exemplu cel pentru care valoarea AIC (Akaike Information Criterion) este mai mică.

**Obs:** Un proces de predicție este caracterizat prin:

- *Intrare*: datele de intrare sunt valori anterioare din serie
- *Iesire*: rezultatul reprezintă valoarea/valorile următoare din serie
- *Model*: un model de regresie care descrie legătura dintre valoarea curentă a seriei și valorile anterioare (numărul de valori anterioare despre care se consideră că influențează valoarea curentă este denumit întârzierea seriei (*time-lag*))

Considerăm seria  $X_1, X_2, \dots, X_n$  și întârzierea  $T$ . Deci valoarea curentă  $X_i$  depinde de valorile  $X_{i-1}, X_{i-2}, \dots, X_{i-T}$ . Prin urmare secvența de valori din serie poate fi transformată într-un alt set de date în care sunt  $T$  atribute predictor și un atribut prezis:

<i>Atribute predictor</i>	<i>Atribut prezis</i>
$X_1 \ X_2 \ \dots \ X_i \ \dots \ X_T$	$X_{T+1}$
$X_2 \ X_3 \ \dots \ X_{i+1} \ \dots \ X_{T+1}$	$X_{T+2}$
...	
$X_{n-T} \ X_{n-T+1} \ \dots \ X_{n-i} \ \dots \ X_{n-1}$	$X_n$

Folosind acest set de date se poate construi un model de regresie (în aceeași manieră ca pentru date care nu sunt temporale). În acest context pot fi folosite modele de regresie, rețele neuronale sau rețele RBF. Una dintre principalele dificultăți este alegerea adecvată a valorii  $T$ .

**Exercițiul 1.** Analiza în R a unor serii temporale și construirea unor modele de predicție bazate pe netezire (Holt-Winters) și a unor modele de tip ARIMA. Parcurgeți și executați comenzile din fișierul [Ex1.r](#).

**Exercițiul 2.** Parcurgeți și executați comenzile din fișierul [Ex2.r](#).

### Exercițiul 1/Weka.

- Deschideți fișierul [airlines.arff](#) (conținând nr de pasageri ai unei companii aeriene înregistrat lunar în perioada 1949 – 1960)
- Construiți un nou set de date folosind o întârziere  $T=12$ . *Indicație*: utilizați Weka pt eliminarea atributului corespunzător datei și Excel (sau un limbaj de programare) pt construirea noului set de date
- Aplicați un model de regresie pentru noul set de date și analizați rezultatele obținute

### Exercițiul 2/Weka.

(doar pt versiune Weka  $\geq 3.7.3$ )

- a) Instalați pachetul **Time Series Forecasting** utilizând **Weka GUI Chooser** ->**Tools->Package manager** și selectând pentru instalare **timeSeriesForecasting**
- b) Deschideți fisierul **airlines.arff**
- c) Preziceti următoarele 6 valori utilizând unul dintre următoarele modele: (i) linear regression; (ii) multilayer perceptron; (iii) random forests. *Indicație:* selecția modelului se realizează utilizând panelul **Advanced Configuration->Based Learner**

Obs: detalii privind pachetul **TimeSeriesForecasting** pot fi găsite la <http://wiki.pentaho.com/display/DATAMINING/Time+Series+Analysis+and+Forecasting+with+Weka>.

(pentru Weka 3.8.1) se utilizează **Forecast (Basic configuration)** și se specifică doar **Number of time units to forecast** (6 dacă se dorește estimarea următoarelor 6 valori).

### **Temă.**

1. Scrieți funcții R care transformă o serie prin:
  - a. Netezire folosind medie mobilă bidirecțională/centrată (moving average smoothing) – dimensiunea ferestrei mobile este parametru al funcției
  - b. Netezire exponențială (exponential smoothing) – parametrul regulii de netezire este parametru al funcției
  - c. Diferențiere (diferențe între valori vecine) – ordinul diferențierii este parametru al funcției

Obs. valorile de la extremitățile seriei vor fi tratate adecvat (nr de elemente din fereastra mobilă poate fi mai mic la extremități).