Idősorok

Nagyméretű adathalmazok kezelése

Bartók Ferenc

2014.03.31.

Tartalom

- **Bevezetés**
- Modellezés
- Szegmentálás
- Anomáliák

Idősor

- Megfigyelések egy sorozata
- Tipikusan adott időközönkénti mérések
  - Pl. naponta, óránként, percenként
- Itt már számít a sorrend
  - „Eddigi” adatbázisokra ez nem volt jellemző
- Adatokhoz időbélyeg is tartozik
  - Nem pusztán szekvenciális adatbázis

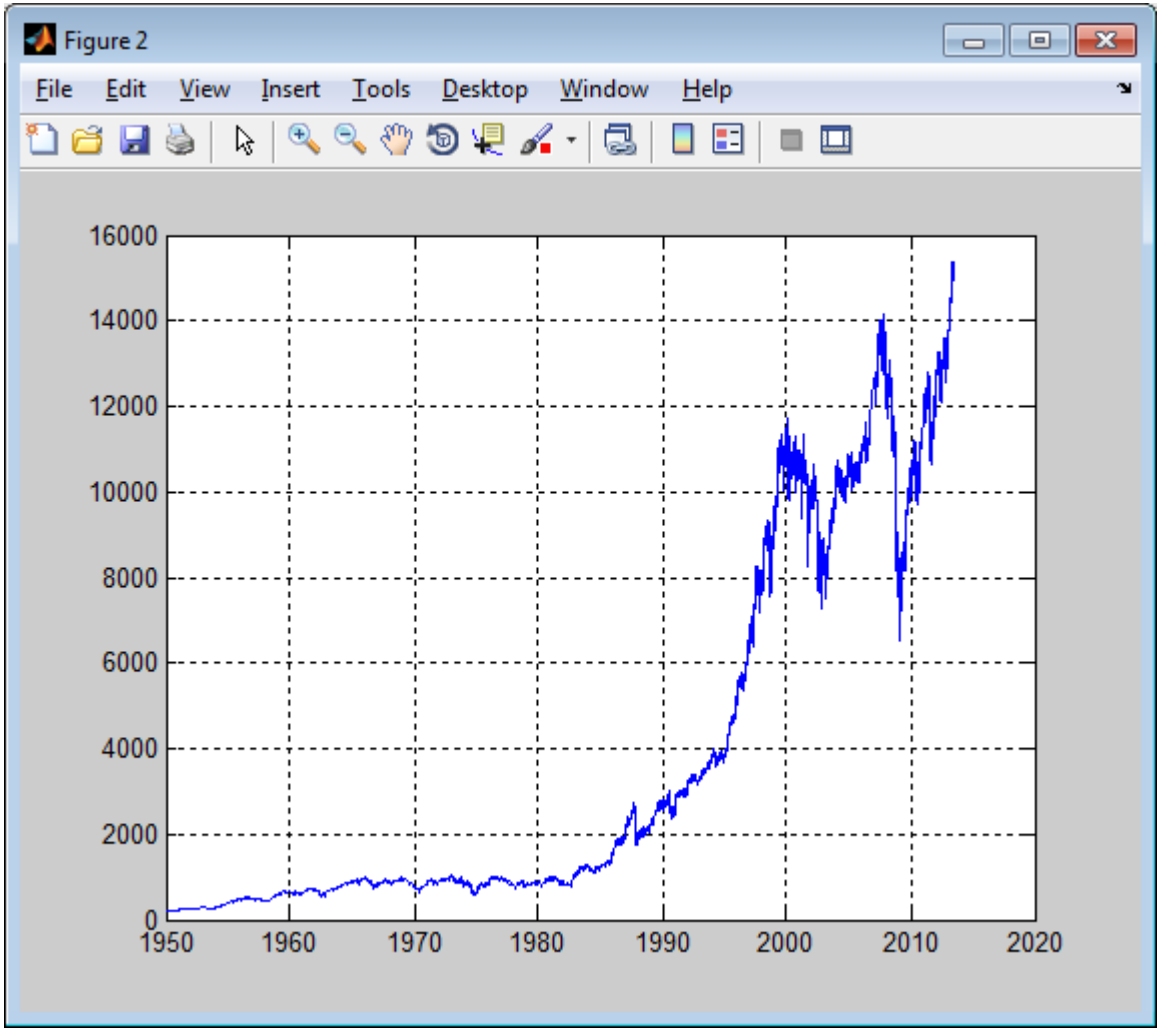
Idősor

- Elemi reprezentáció, t hosszú idősor:

$$x = (x[0], \ldots, x[t-1])$$

- Fontos tulajdonságok
  - Egymást követő megfigyelések erősen korrelálnak egymással
    - Pl. hőmérséklet 10:10-kor és 10:11-kor
  - Különböző hosszúságú idősorok

Példa – Dow Jones idősor



Idősorok típusai

- Attribútum száma szerint
  - Egyváltozós (univariate)
    - Pl. Levegő hőmérséklete vagy tőzsde záró érték
  - Többváltozós (multivariate)
    - Pl. tőzsde záró és nyitó érték, napi kereskedett mennyiség…
    - Pl. levegő hőmérséklete, páratartalma…
    - Pl. az előző két sor együtt

Idősorok típusai

- Stacionaritás szerint (nem definíció)
  - Nem stacionárius
    - Variancia, átlag, más jellemző változik az idő haladása során
    - Lehetnek pl. trendek, ciklusok az idősorban
    - Gyakran jellemző idősorokra!
  - Stacionárius
    - Fentiek nem jellemzőek
    - Pl. konstans variancia az időtől függetlenül
    - Megjegyzés: létezik erős és gyenge stacionaritás fogalom is

Alkalmazások

- Nagyon sok területen alkalmazzák
  - Tőzsde
  - Időjárás
  - Autó forgalom
  - Földrengés
  - Víz, gáz, stb. fogyasztások
  - Népesség
  - …
- Elsősorban előrejelzés céljából

Kérdések

- Hogyan tudunk trendeket felismerni?
- Hogyan tudjuk az idősort reprezentálni?
- Hogyan tudunk anomáliákat detektálni?

- Hogyan tudjuk vizsgálni az idősorokat?

Tartalom

- Bevezetés
- **Modellezés**
- Szegmentálás
- Anomáliák

Modellezés

- Betekintést nyerhetünk azon mechanizmusokba, amelyek generálják az idősort
- Fontos kérdés a modell bonyolultsága (példa)
  - „Prediction is very difficult, especially if it's about the future.” Nils Bohr
    - Figyelmeztetésként szolgál, hogy nem ismert adatokra is validáljuk a modellt, illetve, hogy könnyű olyan modellt találni ami az eddigi adatokra jól illeszkedik, de az előrejelzés nehéz
- Törekedni kell az egyszerű modellre

Egy modell trendelemzéshez

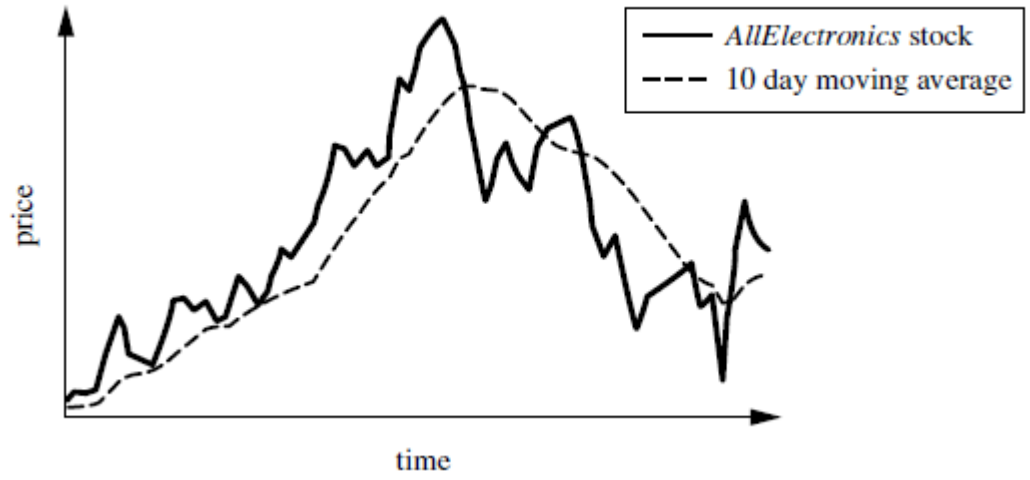
- $Y = F(t)$: idősor (változója)
  - Pl. tőzsde napi záró értékei
- Idősor 4 fő komponense:
- T : trendmozgás
- C : ciklikus mozgás
- S : szezonális mozgás
- I : irreguláris mozgás
- Y dekompozíciója a 4 változóba

$$Y = T \times C \times S \times I$$

(vagy összeg)

Komponensek magyarázata 1.

- Trendmozgás: általános irány egy hosszú időszakon keresztül



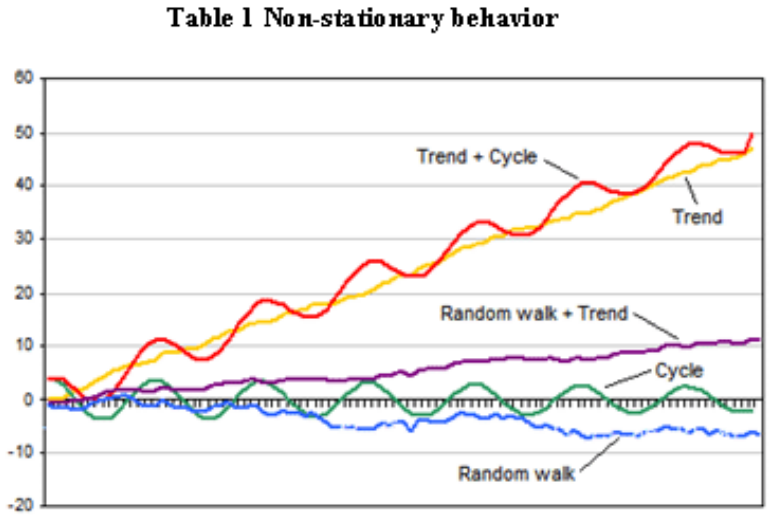
Forrás: Data Mining: Concepts and Techniques

Komponensek magyarázata 2.

- Ciklikus mozgás: ciklusok, hosszú távú változások a trend körül (nem feltétlen periodikus)
  - Pl. tömegközlekedést használó ingázók száma – szabályos csúcspontok és mélypontok

Komponensek magyarázata 2.

- Ciklikus mozgás: ciklusok, hosszú távú változások a trend körül (nem feltétlen periodikus)
  - Pl. tömegközlekedést használó ingázók szám[illegible]tok és mély



Forrás: investopedia.com

Komponensek magyarázata 3.

- Szezonális mozgás: rendszeresen előforduló jelenségek, naptárhoz köthető (periódus max. egy év)
  - Pl. karácsony előtti nagy vásárlások, nőnapi virágeladás
- Irreguláris mozgás: véletlen**szerű** események
  - Pl. árvíz, háború, áramkimaradás, bankrobbantás, tőzsde manipulálása …

Egy másik dekompozíció

$$X_t = m_t + \Delta_t + Y_t$$

- m_t : trend
- Δ_t : szezonalitás, periodikus függvény
- Y_t : stacionárius folyamat

- Általánosabbnak mondható modell

Trendelemzés

- Trend meghatározására néhány módszer:
  - Szabad kézzel (egy egyenes, vagy görbe)
  - Mozgó átlag
  - Regresszió

Mozgó átlag

$$\frac{y_1+y_2+\cdots+y_n}{n}, \frac{y_2+y_3+\cdots+y_{n+1}}{n} \ldots$$

- 3-ad rendű egyszerű és súlyozott mozgó átlag

	3	7	2	0	4	5	9	7	2
Egysz.	-	4							
Súly. (1,4,1)									

Mozgó átlag

$$\frac{y_1+y_2+\cdots+y_n}{n}, \frac{y_2+y_3+\cdots+y_{n+1}}{n} \ldots$$

- 3-ad rendű egyszerű és súlyozott mozgó átlag

	3	7	2	0	4	5	9	7	2
Egysz.	-	4	3						
Súly. (1,4,1)									

Mozgó átlag

$$\frac{y_1+y_2+\cdots+y_n}{n}, \frac{y_2+y_3+\cdots+y_{n+1}}{n} \ldots$$

- 3-ad rendű egyszerű és súlyozott mozgó átlag

	3	7	2	0	4	5	9	7	2
Egysz.	-	4	3	2	3	6	7	6	-
Súly. (1,4,1)	-	5.5							

Mozgó átlag

$$\frac{y_1+y_2+\cdots+y_n}{n}, \frac{y_2+y_3+\cdots+y_{n+1}}{n} \ldots$$

- 3-ad rendű egyszerű és súlyozott mozgó átlag

	3	7	2	0	4	5	9	7	2
Egysz.	-	4	3	2	3	6	7	6	-
Súly. (1,4,1)	-	5.5	2.5						

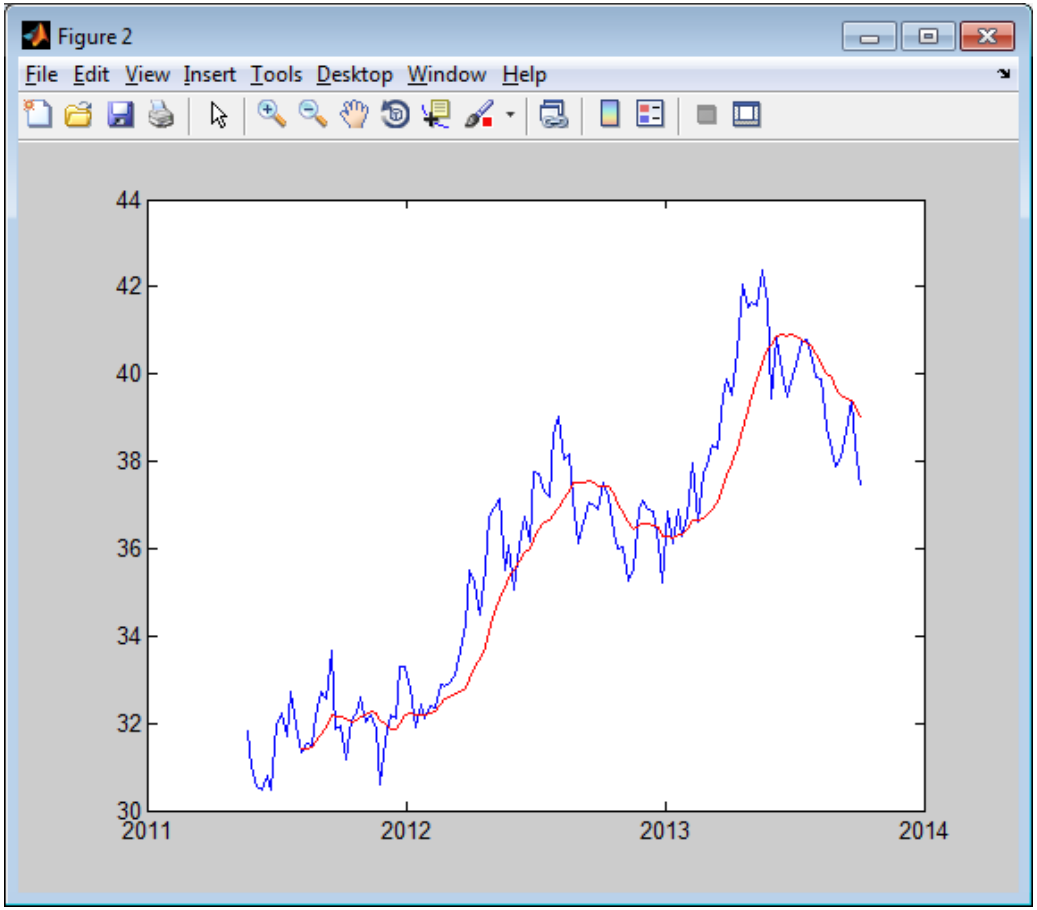
Mozgó átlag

$$\frac{y_1+y_2+\cdots+y_n}{n}, \frac{y_2+y_3+\cdots+y_{n+1}}{n} \ldots$$

- 3-ad rendű egyszerű és súlyozott mozgó átlag

	3	7	2	0	4	5	9	7	2
Egysz.	-	4	3	2	3	6	7	6	-
Súly. (1,4,1)	-	5.5	2.5	1	3.5	5.5	8	6.5	-

Mozgó átlag – Coca-Cola



Mozgó átlag

- A mozgó átlag hajlamos csökkenteni a változások nagyságát
- „Simítja” az idősort
- DE:
- Adatot veszít az idősor elejéről és végéről
- Ciklusokat, vagy egyéb mozgásokat generálhatnak
- Extrém értékek erősen befolyásolhatják
  - Súlyozott mozgó átlag csökkentheti ezt a hatást megfelelő súlyokkal

Regresszió

- Adott:
  - Független (vagy előrejelző) változók
  - Függő (vagy válasz) változó
- Modellezi a független változók és a függő változó közötti kapcsolatot
- Leginkább előrejelzésre szokták használni
  - Trendelemzésre is jó
- Típusok:
  - Lineáris
  - Nem lineáris regresszió

Lineáris regresszió

- Feltételezi a függő és független változók közötti lineáris kapcsolatot
- Az adatok pontfelhőjére próbál egyenest illeszteni
- Amennyiben 1 darab függő változó van, akkor egyszerű lineáris regresszióról beszélünk
- Több függő változó esetén beszélünk többváltozós lineáris regresszióról

Lineáris regresszió

- Egyszerű lineáris regresszió:
  - minták száma: n adat $\{(x_i, y_i), i = 1, \dots, n\}$
  - Független változó: x
  - Függő változó: y

$$y = \alpha + \beta x$$

  - α, β: regressziós együttható

  - Találjuk meg azt az egyenletmegoldást (egyenest), ami a legjobb illesztése az adatpontoknak

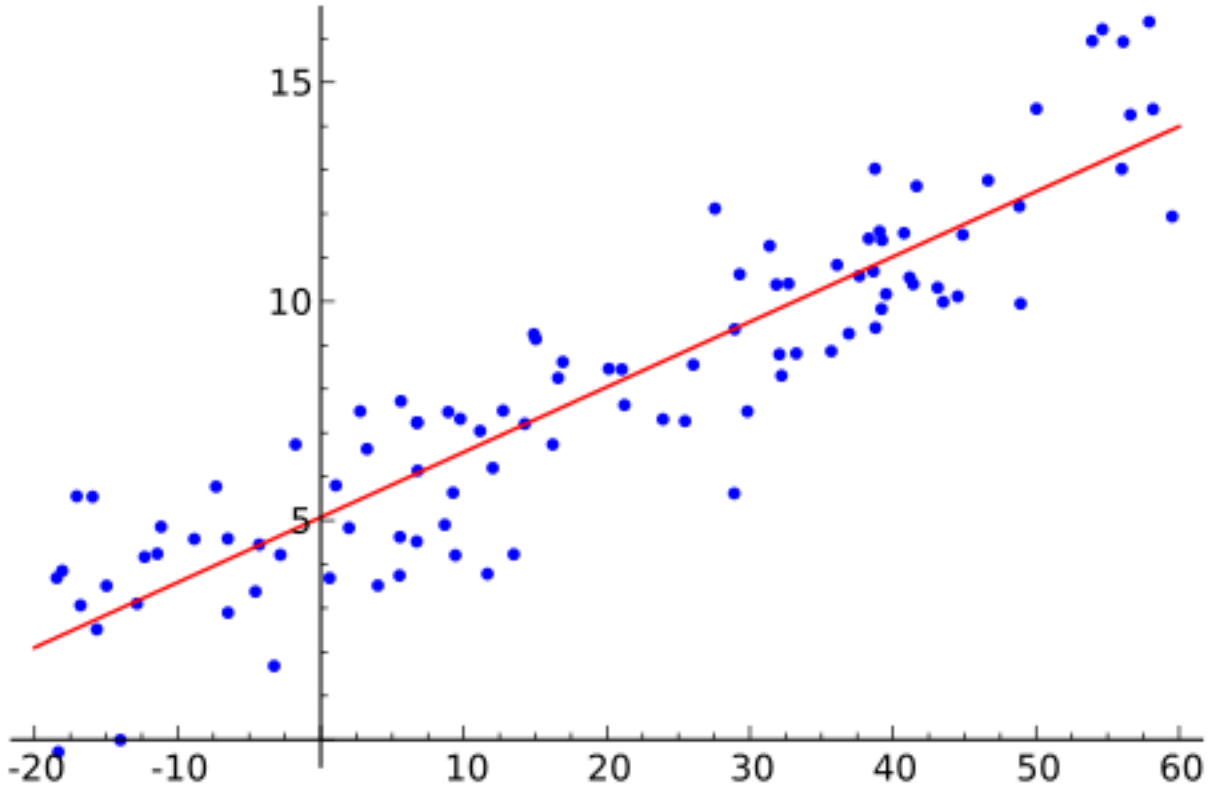
Lineáris regresszió

- Legjobb illesztés megtalálása:
  - Legkisebb négyzetek módszere:
  - Minimalizálja a lineáris regresszió modell négyzetes hibaösszegét
  - Tehát azt az alfa és béta paramétert találja meg, ahol ez a hiba a legkisebb
  - A hiba:

$$\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

  - Ezt a hibát minimalizáljuk (példa)

Lineáris regresszió



Forrás: wikipedia

Lineáris regresszió

- Paraméterek kiszámolása (levezetés nélkül)
  - $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$
  - $\hat{\beta} = \frac{Cov[x,y]}{Var[x]}$
  - $\bar{y}$: függő változó mintaátlaga
  - $\bar{x}$: független változó mintaátlaga

ARIMA modell

- AutoRegressive Integrated Moving Average
- Box-Jenkins módszerhez köthető
- ARIMA(p,d,q)
  - 3 fő rész: AR(p), I(d), MA(q)
  - p, d, q nem negatív egész számok
  - Ha valamelyik 0, akkora az a rész „kiesik”
- Gyakran használják idősorok elemzéséhez és előrejelzéséhez

ARIMA

- Nem stacionárius adatokra is használják
- Ekkor kezdeti lépésként: differenciálás
  - Egyes szintű differenciálás: I(1)

$$diff(i) = y(i) - y(i-1)$$

- Ennek segítségével stacionáriussá (vagy közel stacionáriussá) **alakítható** az idősor

ARIMA


Figure 2
File Edit View Insert Tools Desktop Window Help
x 10^4
1.4
1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
Q1-08
Q2-08
Q3-08
Q4-08
Q1-09
Q2-09
Q3-09

ARIMA

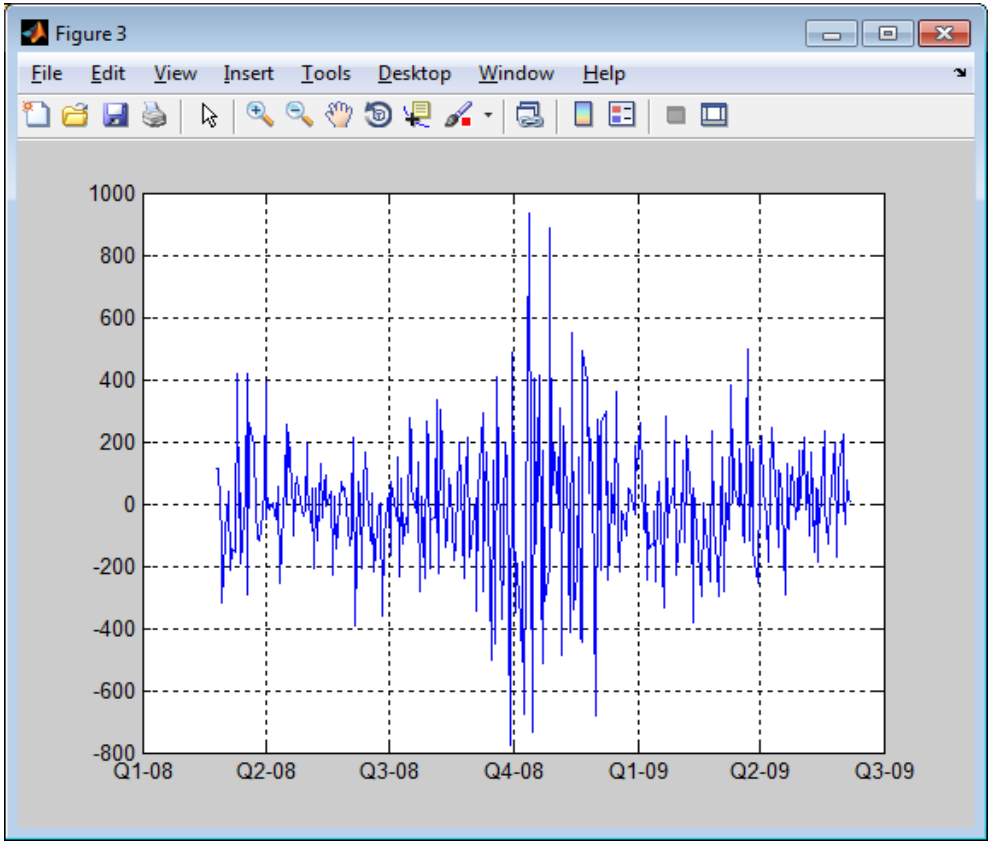


Figure 3
File Edit View Insert Tools Desktop Window Help
1000
800
600
400
200
0
-200
-400
-600
-800
Q1-08
Q2-08
Q3-08
Q4-08
Q1-09
Q2-09
Q3-09

ARIMA

- Differenciálás, különbségképzés „másképpen”
  - Backward shift (B) operátor

$$BX_t = X_{t-1}$$

  - Különbségképző operátor

$$\nabla = 1 - B$$

- Különbségképzés:

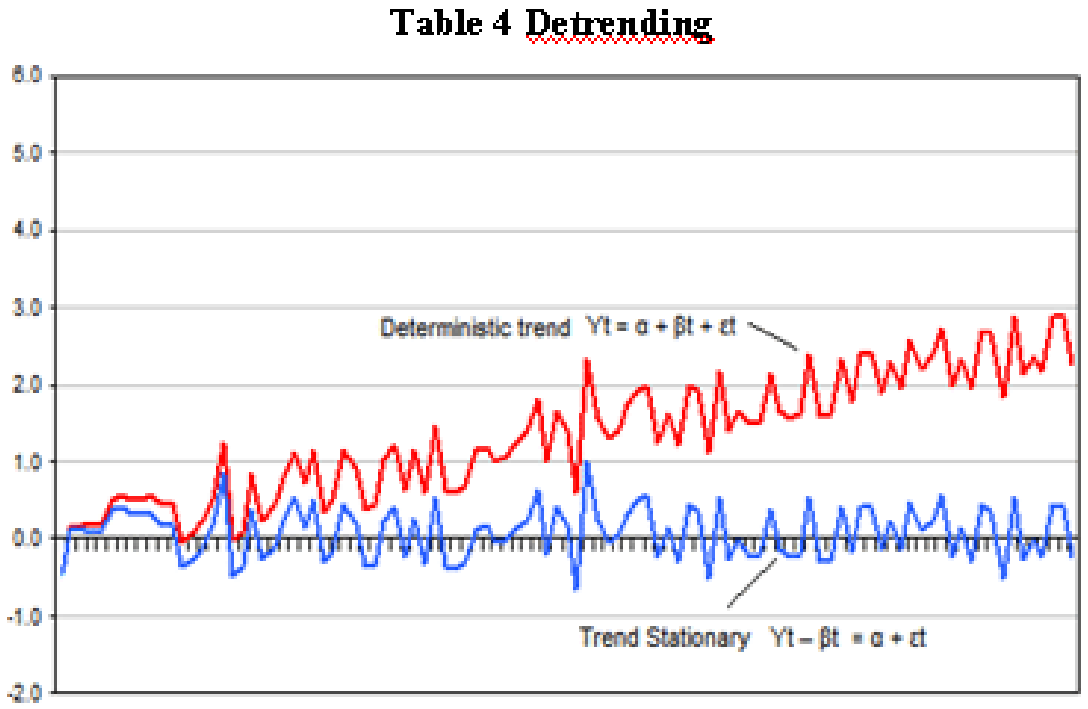
$$\nabla X_t = (1 - B)X_t = X_t - X_{t-1}$$

ARIMA

- Nem stacionárius folyamatokra szintén szokták alkalmazni a „detrending” műveletet is
- Azaz megpróbálják eltávolítani (kivonni) a trendet az idősorból
  - Ehhez illeszteni kell egy trend egyenest
- Itt adatvesztés nem történik

ARIMA

- Detrending



Forrás: investopedia.com

ARIMA

- Stacionaritás vizsgálatára:
  - Dickey-Fuller teszt
  - Augmented Dickey-Fuller teszt
    - (Egy fajta bizonyosságot is megad)

ARIMA

- Dickey-Fuller teszt
  - Azt vizsgálja, hogy „unit root” jelen van e az autoregresszív modellben
  - Egy egyszerű AR(1) modell a következőképp néz ki:

$$y_t = \rho y_{t-1} + u_t$$

  - u_t: hibatag
  - Unit root jelen van, ha $\rho = 1$, ez a nem stacionárius eset
  - Illetve ha $|\rho| < 1$, akkor erősen stacionárius

ARIMA

- Unit root hatása:

Legyen $y_0 = 0$

$$y_t = 1 \times y_{t-1} + u_t$$

- Behelyettesítésekkel:

$$y_t = y_0 + \sum_{j=1}^{t} u_j$$

$$Var(y_t) = \sum_{j=1}^{t} \sigma^2 = \mathrm{t}\sigma^2$$

- Tehát a variancia függ t-től

ARIMA

$$Var(y_1) = \sigma^2$$
$$Var(y_2) = 2\sigma^2$$

- Tehát a Dickey-Fuller teszt a unit root jelenlétét vizsgálja
  - Nullhipotézis: van unit root

ARIMA

- Kitérő:
  - A legtöbb statisztikai előrejelző módszer arra a feltételezésre épít, hogy az idősorok közel stacionáriussá alakíthatóak matematikai transzformációk segítségével (differencing, detrending…)
  - Egy ilyen idősor előrejelzése azon alapul, hogy a statisztikai paraméterek hasonlóak lesz a jövőben is
  - A kapott előrejelzést pedig vissza lehet transzformálni az eredeti idősor előrejelzéséhez

ARIMA – Box-Jenkins

- Lépések

1. Modell azonosítása, választása
2. Paraméterek becslése
3. Modell ellenőrzése

ARIMA – Box-Jenkins

1. Modell azonosítása, választása

- Idősor stacionáriussá tétele (I(d) meghatározása)
- Szezonalitás felismerése, ha szükséges kiiktatása
- „Korreláció”, ACF, PACF kirajzolása, segítségével AR(p), MA(q) komponensek becslése

ARIMA – Box-Jenkins

- ACF = AutoCorrelation Function
  - Pl. lag=1: Pl. Y(t) és Y(t-1)-ek korrelációja
- PACF = Partial ACF
  - A tényleges korreláció és „várható” korreláció különbsége
  - („várható”: korreláció tovább terjedhet)
- Azt méri, hogy az adatok adott időbeli távolságokra („lag”) mennyire korrelálnak egymással
- Értékük: [-1,+1]
  - Minél nagyobb, annál erősebb pozitív korrelációról beszélünk

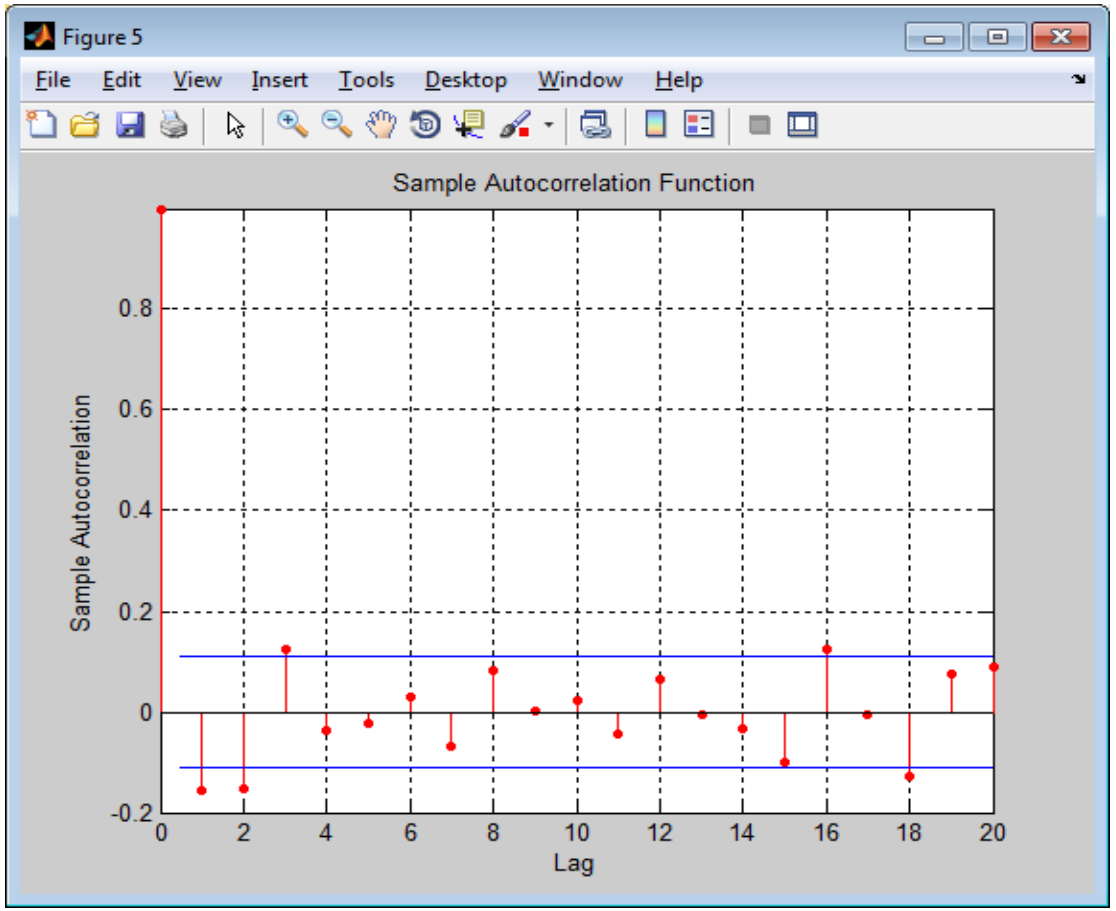
ARIMA – Box-Jenkins

- Megjegyzés:
  - Az empirikus autokorrelációs függvényt a tapasztalati autokovariancia függvénnyel definiáljuk:

$$\hat{\rho}(h) := \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$$

  - Ahol $\hat{\gamma}(h)$ a tapasztalati autokovariancia függvény

ARIMA – Box-Jenkins



ARIMA – Box-Jenkins

2. paraméterek becslése

- Legáltalánosabb módszer:
  - Maximum likelihood becslés
  - Yule-Walker egyenletek

3. Modell ellenőrzése

- Előrejelzéssel összehasonlítás

ARIMA – Box-Jenkins

- Bizonyos kutatók szerint ez a megközelítés problémás a következő miatt:
  - Pl. a közgazdasági, társadalmi valós idősorok sohasem stacionáriusak, akárhány differenciálást is alkalmazunk

Tartalom

- Bevezetés
- Modellezés
- **Szegmentálás**
- Anomáliák

Szegmentálás

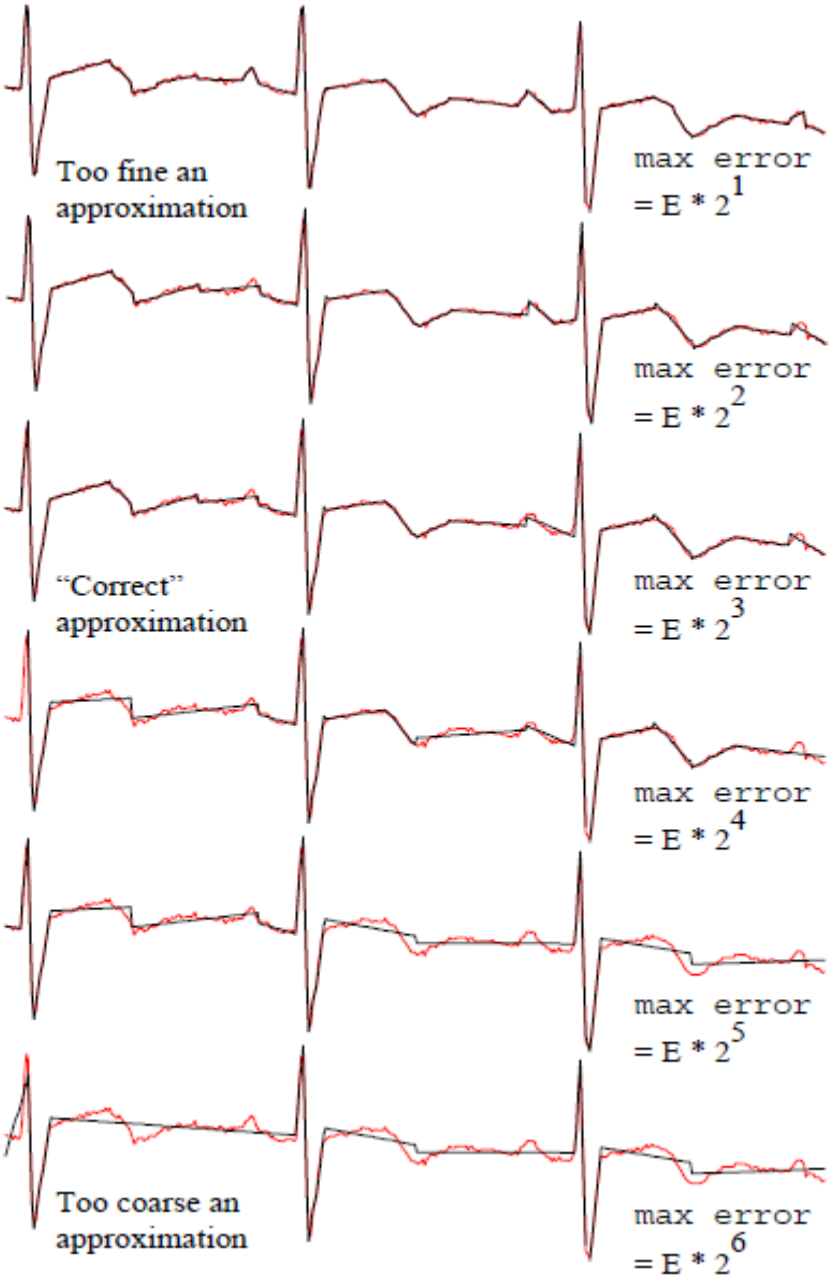
- Az adatok reprezentálásának a módja fontos
  - Pl. Fourier transzformálás
  - Pl. Szimbólummal történő ábrázolás
  - Pl. Piecewise Linear Representation (PLR)
    - n hosszúságú T idősort K darab szakasszal közelítünk (példa)

Szegmentálás

- PLR:
- K sokkal kisebb, mint n
  - Azaz sokkal kevesebb szakasszal közelítünk, helyettesítünk
- Az adatok tárolása, továbbítása, számításai hatékonyabbak
- Támogatja a következőket:
  - Gyors hasonlóság keresés
  - Változási pont (changepoint) detektálás
  - Újszerű klaszterezési és osztályozási algoritmusok
  - …

Szegmentálás

Forrás: Segmenting Time Series: A Survey and Novel Approach



Forrás: Segmenting Time Series: A Survey and Novel Approach

Szegmentálás

- A probléma/feladat:
  - Adott egy T idősor, állítsuk elő a legjobb reprezentációját úgy, hogy
    - csak K darab szegmenst használhatunk fel (kötött szegmensszám)
    - szegmensenként a maximális hiba ne lépje túl a felhasználó által definiált küszöböt (max_error)
    - a szegmensek összesített hibája ne lépje túl a felhasználó által definiált küszöböt (total_max_error)

Szegmentálás

- Típusok:
- Feldolgozás szerint:
  - Valósidejű/online
  - Batch (nem valósidejű, minden megfigyelés rendelkezésre áll)
- Approximáció jellege szerint:
  - Lineáris approximáció
    - Pl. interpoláció, regressziós közelítés…
  - Nem lineáris approximáció

Szegmentálás

- Szegmentálási eljárások:
  - Csúszó-ablak szegmentálás (Sliding Window)
  - Top-down
  - Bottom-Up
  - Sliding-Window and Bottom-Up

Szegmentálás

- Csúszó ablak szegmentálás
  - Amikor az új pont hozzávétele az aktuális szegmenshez az új közelítő egyenes szakasszal meghaladja a max_error korlátot, lezárjuk a szegmenst (az előző pontnál) és újat kezdünk.

Szegmentálás

```
Algorithm Seg_TS = Sliding_Window(T , max_error)
anchor = 1;
while not finished segmenting time series
i = 2;
 while  calculate_error(T[anchor: anchor + i ]) < max_error
   i = i + 1;
 end;
 Seg_TS = concat(Seg_TS, create_segment(T[anchor: anchor + (i-1)]);
 anchor = anchor + i;
end;
```

Forrás: Segmenting Time Series: A Survey and Novel Approach

Szegmentálás

- Csúszó ablak szegmentálás jellemzők
  - Egyszerű
  - Valósidejű
  - Egyszerűen gyorsítható (offline esetben)
  - Sok területen elterjedt (pl. orvosi)
  - DE!: nem ad túl jó eredményeket

Szegmentálás

- Top-down algoritmus
- Nem valósidejű
- Kiindulás:
  - A teljes hosszúságot egyetlen szegmensnek vesszük, ha a hiba túl nagy, akkor kettéosztjuk a szakaszt és rekurzívan mindkét felét újravizsgáljuk

Szegmentálás

```
Algorithm Seg_TS = Top_Down(T , max_error)
best_so_far = inf;
for i = 2 to length(T) - 2  // Find best place to split the time series.
improvement_in_approximation = improvement_splitting_here(T,i);
  if improvement_in_approximation < best_so_far
    breakpoint = i;
    best_so_far = improvement_in_approximation;
  end;
end;
                              // Recursively split the left segment if necessary.
if calculate_error(T[1:breakpoint]) > max_error
  Seg_TS = Top_Down(T[1: breakpoint]);
end;
                              // Recursively split the right segment if necessary.
if calculate_error( T[breakpoint + 1:length(T)] ) > max_error
  Seg_TS = Top_Down(T[breakpoint + 1: length(T)]);
end;
```

Forrás: Segmenting Time Series: A Survey and Novel Approach

Szegmentálás

- Bottom-Up algoritmus
  - N pont esetén N-1 szegmenssel indul
  - Megvizsgáljuk minden lépésben, hogy az összes kis szegmensre az őt előzővel való egyesítés milyen hibanövekedést okoz
  - Megvizsgáljuk, hogy melyik egyesítésnél legkisebb a romlás, és ezután a hibakritérium még teljesül e

Szegmentálás

- Bottom-Up folytatás
  - Ezután ha a feltétel teljesül, akkor végrehajtuk az egyesítést
  - Majd újrakezdjük, és addig csináljuk amíg tudunk egyesíteni
- Tehát 1. lépés után: N-3 db egy hosszú és 1 db kettő hosszú szegmens lesz

Szegmentálás

```
Algorithm Seg_TS = Bottom_Up(T , max_error)
for i = 1 : 2 : length(T)          // Create initial fine approximation.
 Seg_TS = concat(Seg_TS, create_segment(T[i: i + 1 ]));
end;
for i = 1 : length(Seg_TS) - 1     // Find cost of merging each pair of segments.
 merge_cost(i) = calculate_error([merge(Seg_TS(i), Seg_TS(i+1))]);
end;


while min(merge_cost) < max_error                  // While not finished.
 index = min(merge_cost);                          // Find "cheapest" pair to merge.
 Seg_TS(index) = merge(Seg_TS(index), Seg_TS(index+1)); // Merge them.
 delete(Seg_TS(index+1));                                  // Update records.
 merge_cost(index) = calculate_error(merge(Seg_TS(index), Seg_TS(index+1)));
 merge_cost(index-1) = calculate_error(merge(Seg_TS(index-1), Seg_TS(index)));
end;
```

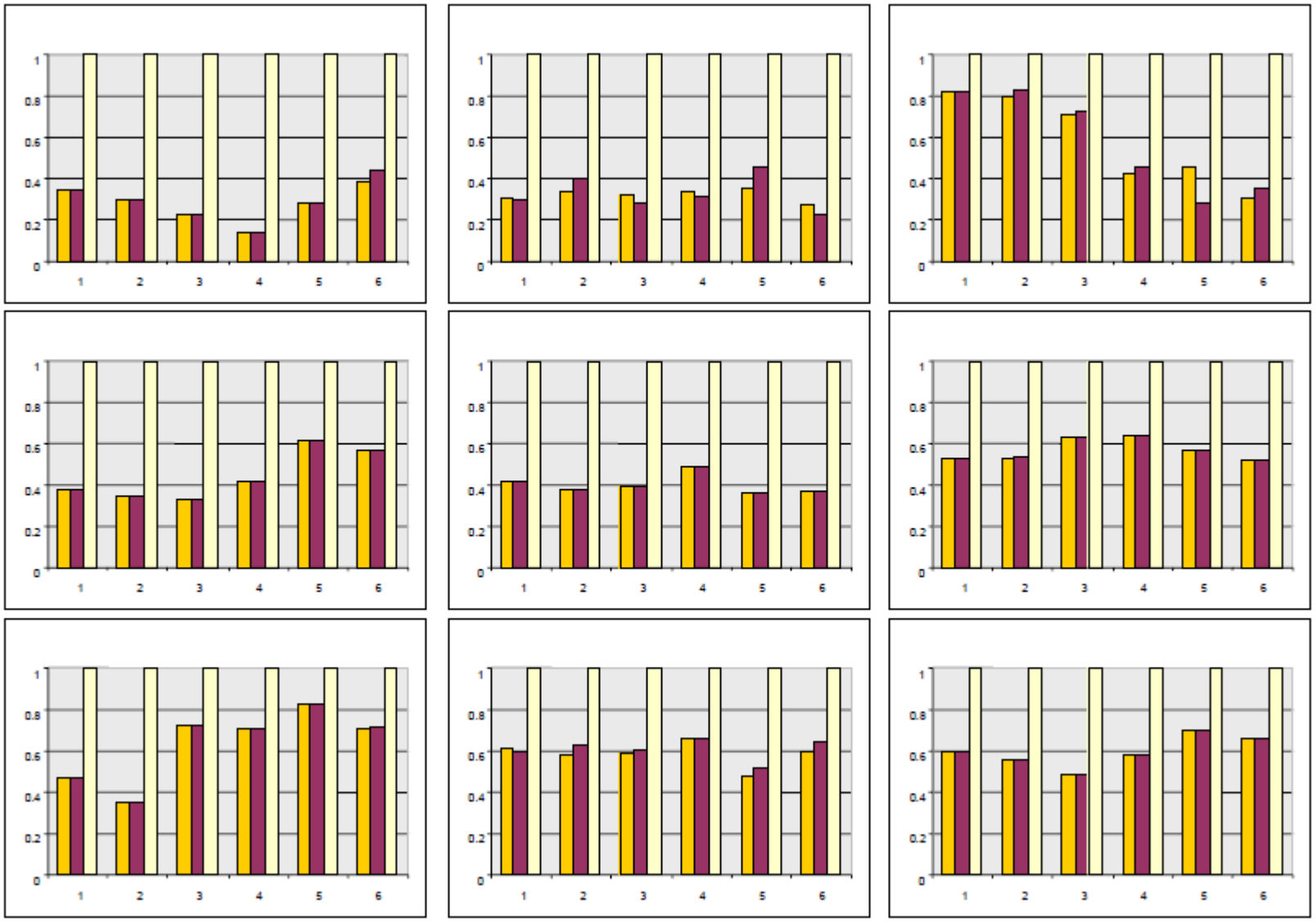
Forrás: Segmenting Time Series: A Survey and Novel Approach

Szegmentálás

- Sliding Window and Bottom-up (SWAB)
  - Előzetes becslés a szegmenshosszra (sL)
  - Olvassunk be a bemeneti bufferből (nyers adat) 5-ször sL adatot a munkabufferbe
  - 2. A munkabuffer adatain B-U szegmentálás
  - 3. A kialakult első szegmenst a kimenetre, pontjait töröljük
  - Ha van még adat a munkabufferben GOTO 2

Szegmentálás

- Ha már nincs adat a munkabufferben, de a bemeneti bufferben még van, akkor olvassunk be és GOTO 2
- Ha elfogyott az adat, akkor vége



narancs=SWAB, lila=BU, sárga=Sliding-W
0 a tökéletes approximáció, 1-6 különböző hibaküszöböknél
Forrás: Segmenting Time Series: A Survey and Novel Approach

Tartalom

- Bevezetés
- Modellezés
- Szegmentálás
- **Anomáliák**

Anomáliák

- Mi az az anomália?
  - Szabad megfogalmazásban: egy olyan adat, amely értéke erősen eltér a várhatótól/elvárttól
- Az anomáliát sokszor kiugró/kilógó értéknek (outlier) is hívják
- Az anomália lehet pl. valamilyen újdonság, kivétel, zaj …

Anomáliák

- Anomália lehet például:
  - Banki csalás – jellemző alkalmazási terület
    - Pl. bankkártyával „hirtelen” másik országban kezdenek el költekezni
  - Természeti katasztrófák
  - Egészségügyi problémák
    - Pl. leáll a szívverés, vagy nagyon felgyorsul stb.
  - Hiányzó, vagy félreírt adat
    - Jellemző (adatbányászatban) az ilyen adatsor
  - …

Anomáliák

- Megjegyzés:
  - Az anomáliákat 2 csoportra lehet osztani az alapján, hogy egy tényleges, valós értékről van szó, vagy pedig valamilyen hiba folytán kaptuk az anomáliát

Anomáliák

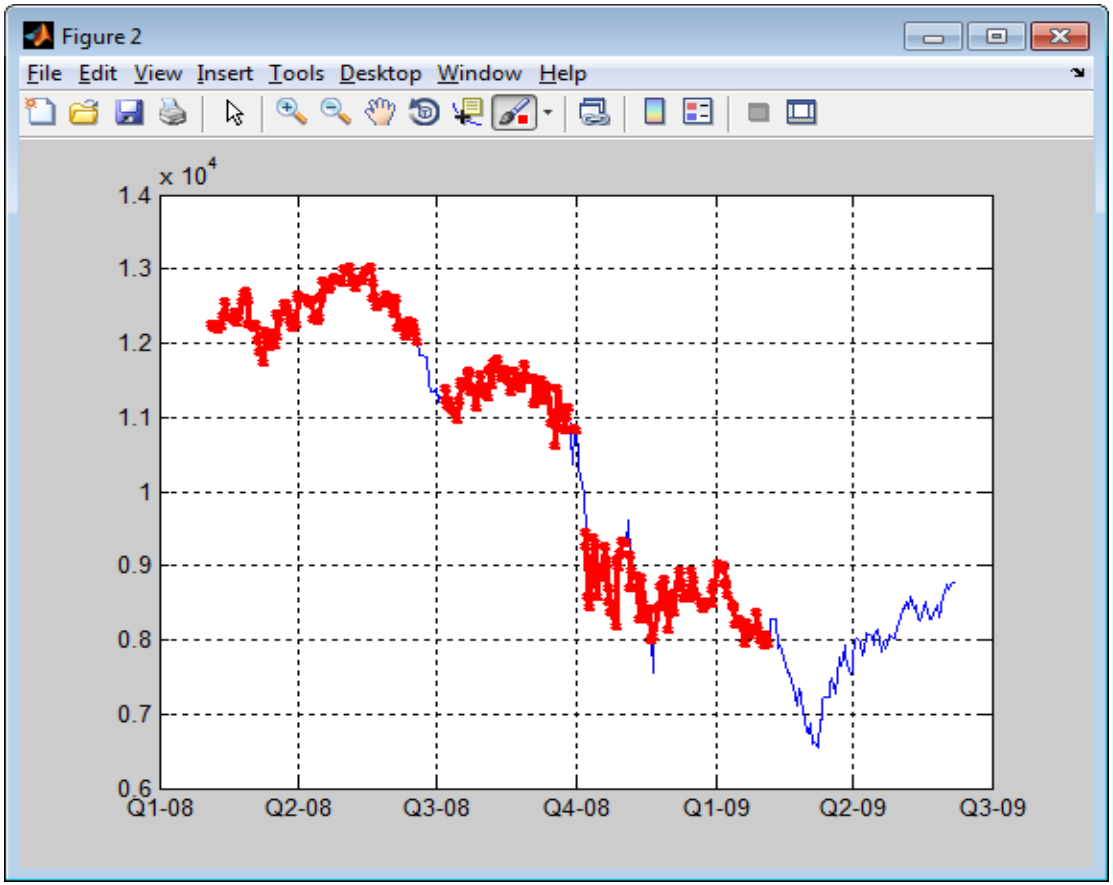
- Egy csoportosítás:
  - Kilógó értékek (outlier)
    - Egy olyan megfigyelés, amely jelentősen eltér ugyanazon minta többi tagjától
  - Változási pontok (changepoint)
    - Strukturális változásnak is hívják
    - Egy olyan változáshoz köthető, ahol megváltoztak a folyamat statisztikai jellemzői

Anomáliák

- Dow Jones

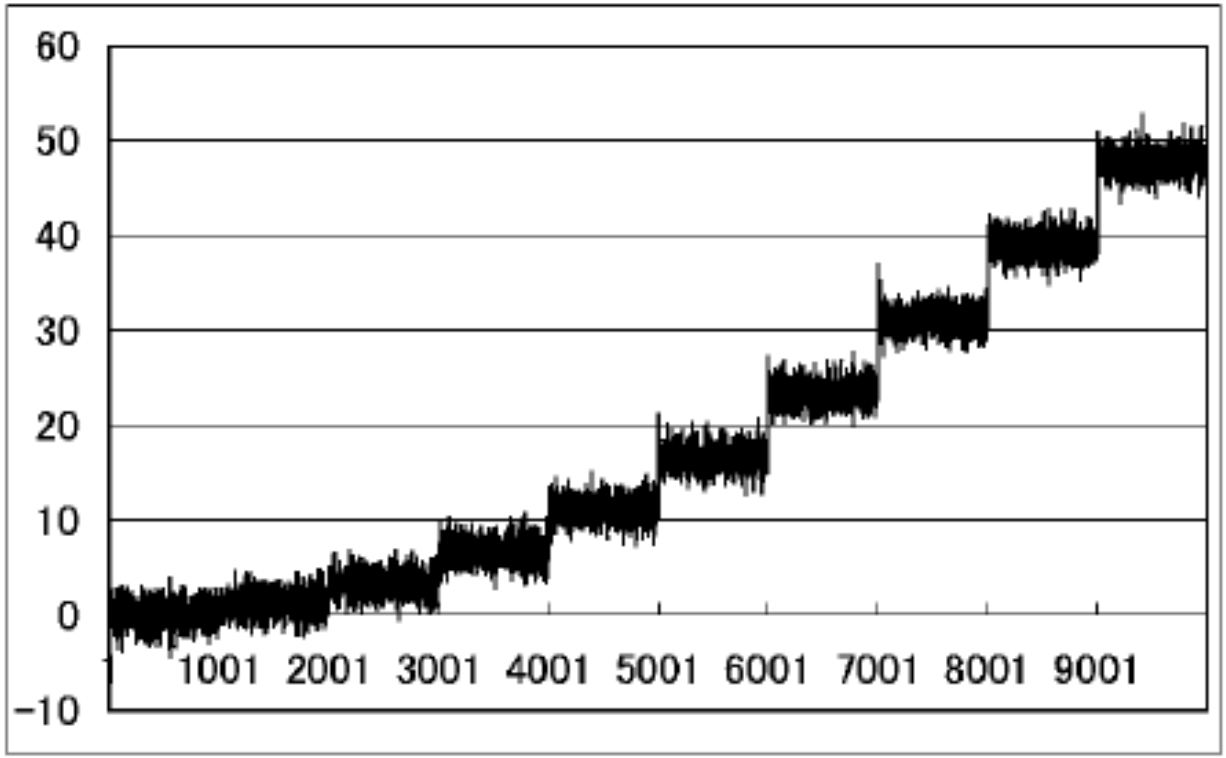
Anomáliák

- Változási pontok (ponthalmazok)



Anomáliák

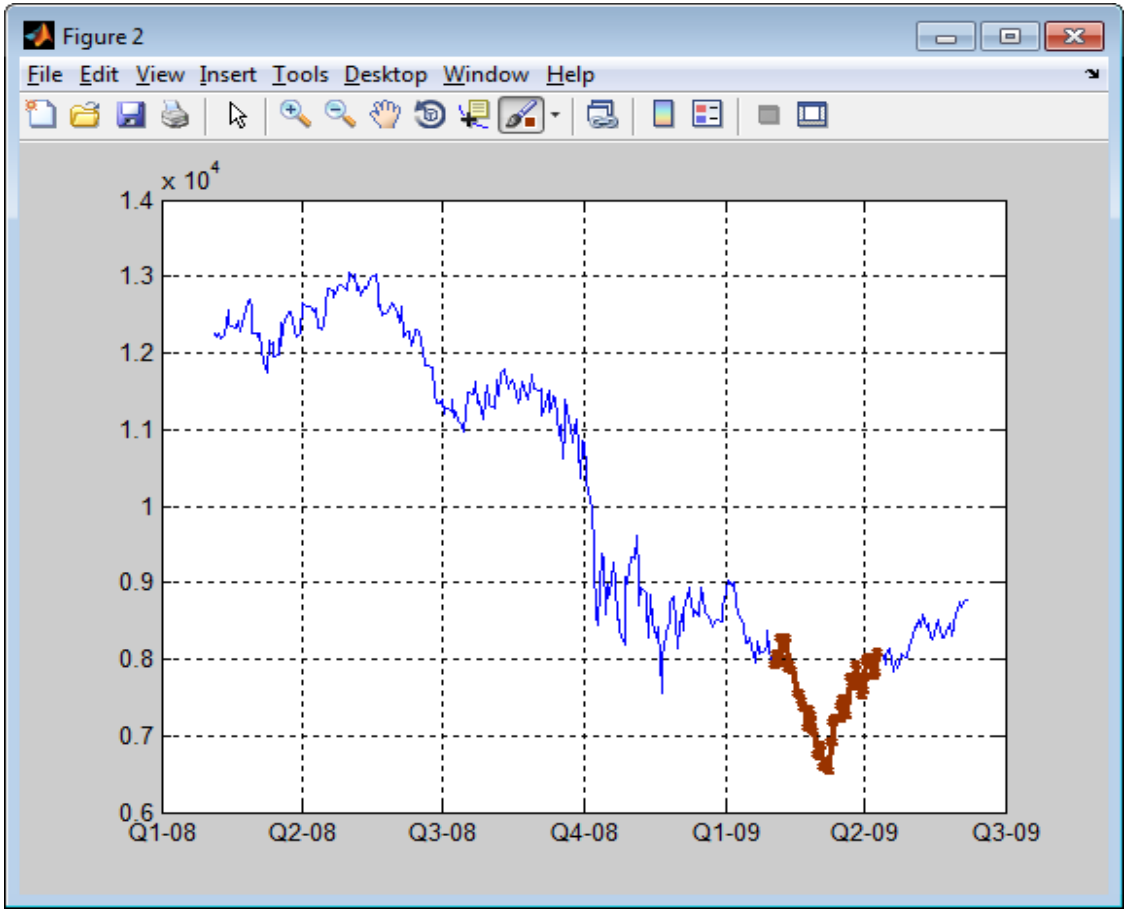
- Változási pontok



Forrás: A Unifying Framework for Detecting Outliers and Changepoints

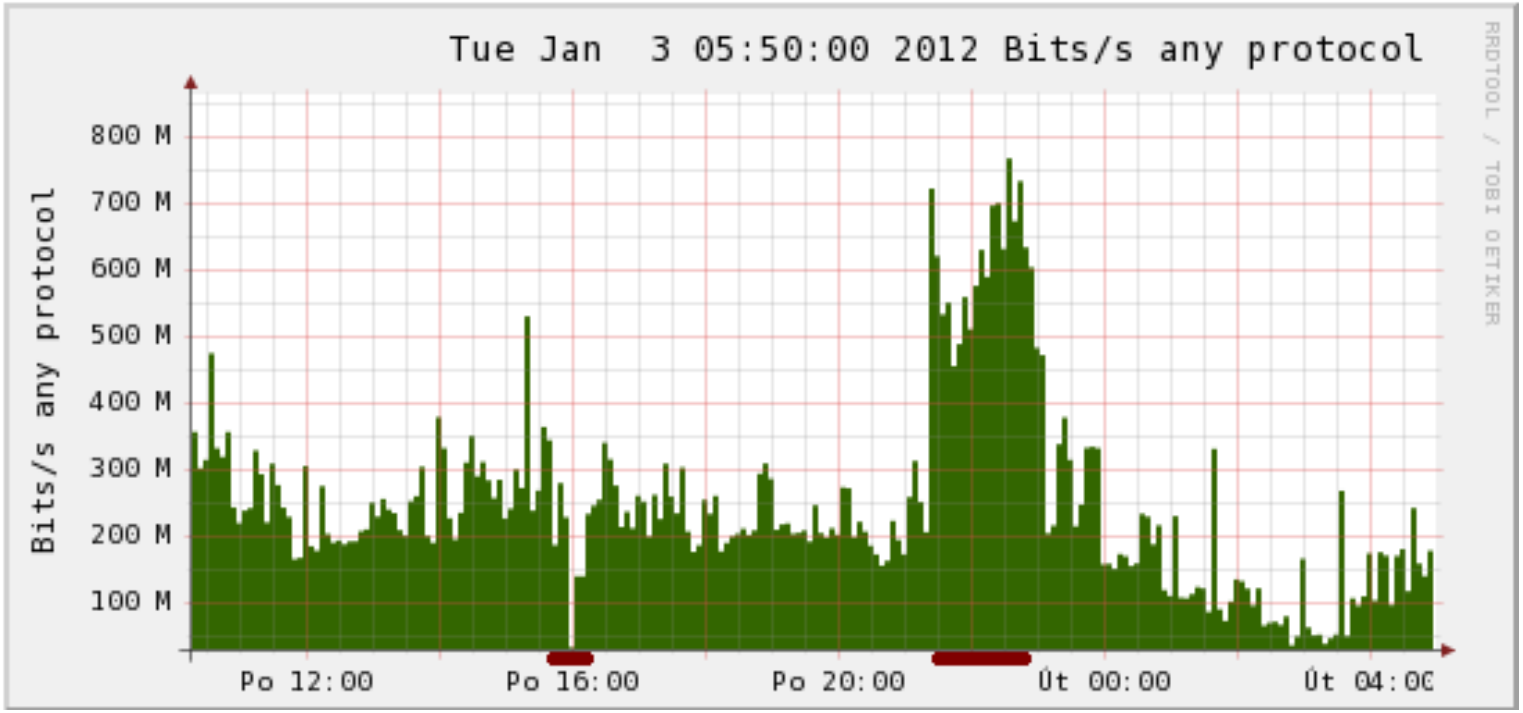
Anomáliák

- Kilógó érték



Anomáliák

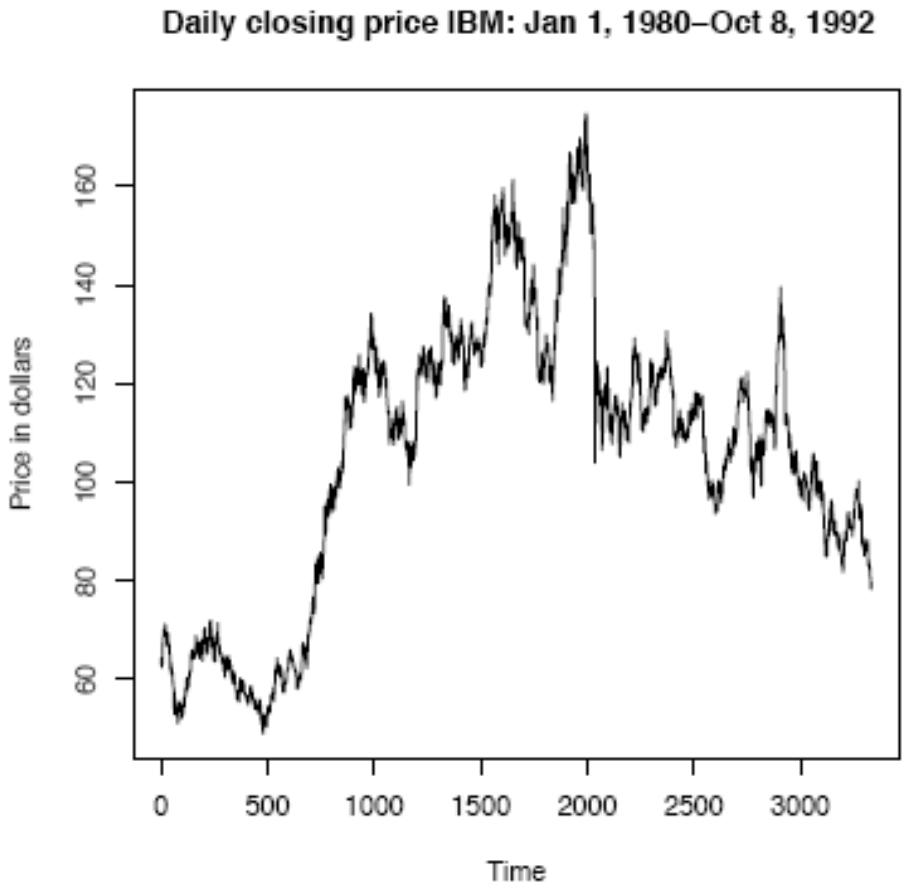
- Hálózati forgalom



Forrás: muni.cz

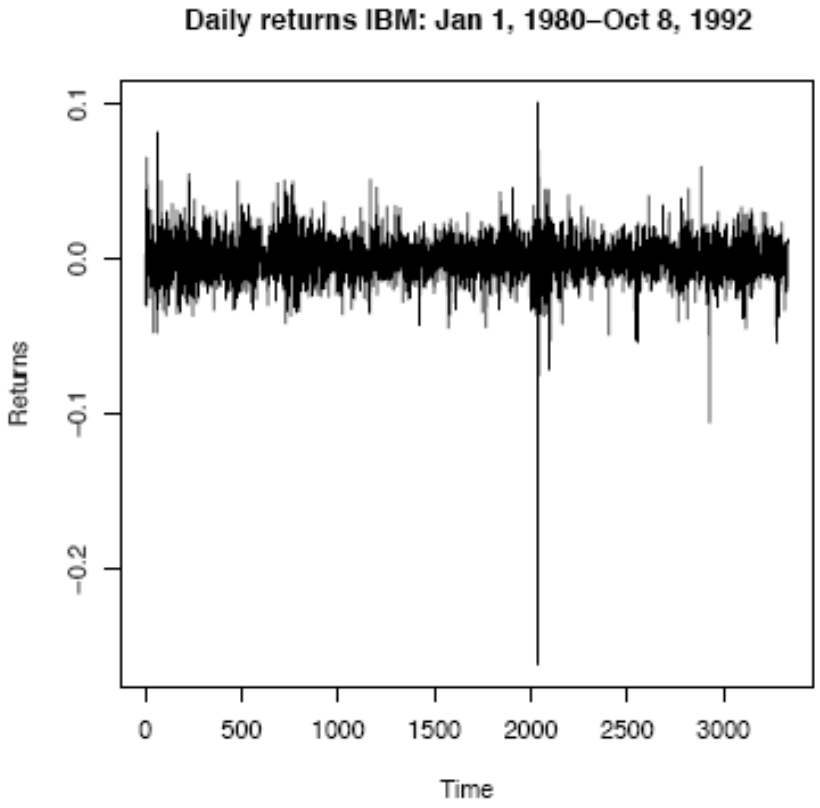
Anomáliák

- IBM tőzsde napi záróérték



Anomáliák

- IBM tőzsde napi százalékos változás



Anomáliák

- Megjegyzés:
  - Kilógó értékek és változási pontok további csoportokra bonthatók:
  - Kilógó érték:
    - Additív és innovatív
  - Változási pont:
    - Level change, variance change

Anomáliák

- Anomáliák detektálása előrejelzés céljából:
  - Az anomáliák komoly problémákat okozhatnak az idősor előrejelzésénél
    - Félrevihetik a modellt
  - Érdemes ezért az anomáliákat detektálni az előrejelzéshez
  - A detektált anomáliákat sokszor egy „átlagszerű” értékkel helyettesítik vagy ha lehetséges szimplán törlik az idősorból
  - Másik módszer lehet a detektált anomáliák kisebb súlyú figyelembevétele a modell megalkotásánál

Anomáliák

- Anomáliák detektálása elemzés céljából:
  - Klasszikus adatbányászati értelemben:
    - Nem várt érdekes tudást, információkat tárhat fel

Anomáliák

- Anomáliák detektálására módszerek:
  - Nem felügyelt módszerek
    - Nincs információnk arról, hogy mi normális és mi nem (=anomália)
    - Idősoroknál ez a jellemző
  - Felügyelt módszerek
    - Az adatok jelölve vannak (normális vagy nem)
    - Ezek alapján egy osztályozót tanítunk (fontos eltérés az általános osztályozási problémáktól, hogy itt erősen nem kiegyenlített az adathalmaz)
  - Félig felügyelt módszerek
    - Létrehoz egy modellt a normális adatok alapján, majd ehhez viszonyítva teszteli a többi adatot (valószínűséget ad meg)

Anomáliák

- Néhány módszer:
  - Statisztikai alapon (példa)
    - A statisztikai jellemzőktől (átlag, szórás…) való eltérések vizsgálatával
    - Pl. illesztünk egy AR modellt, majd a sűrűségfüggvények alapján valamilyen távolság definíció segítségével adjuk meg az anomáliavalószínűséget
  - Klaszterezéssel
    - Pl. a klaszterközéppontoktól azonos klaszterben lévő egy bizonyos távolságtól távolabb levő pontok
    - Pl. klaszterek határai
    - Idősoroknál az általános módszerek nem a legjobbak

Anomáliák

- Mozgó átlag
  - Pl. Az eredeti idősor és a mozgó átlag különbsége
- …

Anomáliák

- Például visszaéléssel, vagy hálózati behatolással kapcsolatban az anomáliának tekintett adatok sokszor nem ritka adatok, hanem burst-szerűek
- Az ilyen anomália típus nem felel meg annak az általános definíciónak, hogy az anomáliának ritkának „kell” lennie
- Ebben az esetben sok általános detektáló módszer nem működik jól
  - De! klaszterező algoritmus felismerheti ezeket a burstöket

Tartalom

- Bevezetés
- Modellezés
- Szegmentálás
- Anomáliák

További témakörök

- *Előrejelzés*
  - *Pl. Regresszió, ARIMA, neurális hálók…*
- Klaszterezés
- Osztályozás
  - Pl. jelnyelv: kézmozdulatok sorozatán keresztül felismerni a szavakat
- Idősorok összehasonlítása
- Mintafelismerés

Köszönöm a figyelmet!